

Proceedings of

COMPSTAT 2014

21st International Conference on **Computational Statistics**

hosting the **5th IASC World Conference**



Geneva, Switzerland

August 19–22, 2014



Manfred Gilli
Gil Gonzalez-Rodriguez
Alicia Nieto-Reyes (Eds.)

ISBN: 978-2-8399-1347-8

Proceedings of COMPSTAT 2014

Manfred Gilli
Geneva School of Economics and Management
University of Geneva
Switzerland
Manfred.Gilli@unige.ch

Gil González-Rodríguez
Department of Statistics
University of Oviedo
Spain
gil@uniovi.es

Alicia Nieto-Reyes
Department of Mathematics, Statistics and Computer Science
University of Cantabria
Spain
alicia.nieto@unican.es

ISBN 978-2-8399-1347-8

19th August 2014

Université de Genève – 1211 Genève, Switzerland

©2014 – The International Statistical Institute/International Association for Statistical Computing

All rights reserved. No part of this CD may be reproduced, stored in a retrieval system, or transmitted, in any other form or by any means without the prior permission from the publisher.

Preface

The 21st International Conference on Computational Statistics (COMPSTAT 2014) is held in Geneva. This year the Conference also hosts the 5th IASC World Congress. The Geneva edition coincides with the 40th anniversary of this biennial event which started in 1974 in Vienna and has been organized all over Europe. In the preface of the 1974 proceedings we can read: *'If we succeed in making statisticians aware of the great possibilities of modern computing facilities, which at any rate go beyond simple numerical computations, the Symposium serves its purpose.'* This goal has since been reached with certainty, as by now statisticians fully integrate computational tools in their work.

The Geneva edition seems to pursue 'the success story' with more than 400 participants and 370 presentations. The electronic Book of Proceedings includes a selection of 84 papers covering 700 pages, all peer reviewed.

Keynote lectures are addressed by Peter Bühlmann from the Swiss Federal Institute in Zurich, Anthony Davison from the Swiss Federal Institute in Lausanne and Xuming He from University of Michigan, USA. Two tutorials are offered, one by Dietmar Maringer, University of Basel, Switzerland and one by Stefan Van Aelst from KU Leuven, Belgium.

The editors thank the contributing authors, the referees and the members of the scientific program committee, and most importantly, all participants who are the soul of the conference.

The next edition of COMPSTAT will take place in Oviedo, Spain on August 23-26, 2016 and will be organized by Prof. Ana Colubi. We wish her the best success.

COMPSTAT 2014 Editors:

Manfred Gilli, University of Geneva, Switzerland.

Gil González-Rodríguez, University of Oviedo, Spain.

Alicia Nieto-Reyes, University of Cantabria, Spain.

Scientific Program Committee

Ex-officio:

COMPSTAT 2014 organiser and Chairperson of the SPC: Manfred Gilli, University of Geneva, Switzerland.

Past COMPSTAT organiser: Erricos John Kontoghiorghes, Cyprus University of Technology, Cyprus.

Next COMPSTAT organiser: Ana Colubi, University of Oviedo, Spain.

IASC-ERS Chairman: Vincenzo Esposito Vinzi, ESSEC Business School, France.

Members:

Alessandra Amendola, University of Salerno, Italy.

Ivette Gomes, University of Lisbon, Portugal.

Sandra Paterlini, European Business School, Wiesbaden, Germany.

Anne Philippe, University of Nantes, France.

Elvezio Ronchetti, University of Geneva, Switzerland.

Marieke Timmerman, University of Groningen, The Netherlands.

Consultative Members:

Representative of the IFCS: Anuska Ferligoj, University of Ljubljana, Slovenia.

Representative of the ARS of IASC: Jung Jin Lee, Soongsil University, Korea.

Representative of ERCIM WG CMS: Stefan Van Aelst, KU Leuven, Belgium.

COMPSTAT 2014 Proceedings Management Committee:

Manfred Gilli, University of Geneva, Switzerland.

Gil González-Rodríguez, University of Oviedo, Spain.

Alicia Nieto-Reyes, University of Cantabria, Spain.

Additional Referees:

Kohei Adachi, Ana Maria Aguilera, Marco Alfo, Andres M. Alonso, Tomas Aluja, Daniel Baier, Simona Balbi, Jose R. Berrendero, Sotiris Bersimis, Lucio Bertoli Barsotti, Patrice Bertrand, Concha Bielza, Angela Blanco-Fernandez, Xavier Bry, Carmen Cadarso, Daniela Calo, M. Angeles Carnero, Philippe Castagliola, Jose E. Chacon, Chun-Houh Chen, Victor Chepoi, Christophe Chesneau, Vartan Choulakian, Claudio Conversano, Mauro Costantini, Antonio Cuevas, Guglielmo D'Amico, Pierpaolo D'Urso, Michel Delecroix, Pedro Delicado, Marta Di Lascio, John Einmahl, Alesio Farcomeni, Arturo J. Fernandez, Silvia Ferrari, Peter Filzmoser, David Fletcher, Marta Garcia-Barzana, Luis Angel Garcia-Escudero, Laurent Gardes, Stelios D. Georgiou, Tomasz Gorecki, Sergei Grudsky, Bettina Grun, Serge Guillas, Armelle Guillou, Hwang Heungsun, Xianzheng Huang, Stephan Huckemann, Marie Huskova, Salvatore Ingrassia, Antonio Irpino, Yoshihide Kakizawa, George Karabatsos, Hyoungmoon Kim, Worapan Kusakunnirun, Agnes Lagnoux, Marc Lavielle, Michael Lechner, Anne Leucht, Christophe Ley, Gaorong Li, Xiang Liming, Chu-An Liu, Ann Maharaj, Dietmar Maringer, Marco Marozzi, Pablo Martinez-Camblor, Antonello Maruotti, Jorge Mateu, Agustin Mayo, Stefan Mittnik, Domingo Morales, Brenda Murphy, Fionn Murtagh, Matthew Nunes, Daniel Oberski, M. Carmen Pardo, Fulvia Pennoni, Carlos Perez-Gonzalez, Davide Pigoli, Ana Belen Ramos-Guajardo, Giovanna Ranalli, Philip Reiss, Holger Reulen, Havard Rue, Silvia Ruiz-Velasco Acosta Giorgio Russolillo, Luigi Salmaso, Antonio Salmeron, Theofanis Sapatinas, Gilbert Saporta, Johan Segers, Ana Sipols, Alwin Stegeman, Fabio Tardella, Andrew C. Titman, Valentin K. Todorov, Inmaculada Torres-Castro, Nickolay Trendafilov, M. Dolores Ugarte, Zidong Wang, Jinfang Wang, Keming Yu.

Contents

Jan Kalina, Zdeněk Valenta and Jurjen Duintjer Tebbens	
Computation of Regularized Linear Discriminant Analysis	1
Paul Fischer and Astrid Hilbert	
Fast Detection of Structural Breaks	9
Anthony C. Atkinson, Marco Riani, Andrea Cerioli and Domenico Perrotta	
Random Start Forward Searches for Detecting Mixtures of Regression Models	17
M. Helena Gonçalves and M. Salomé Cabral	
Incomplete longitudinal binary responses in marginal model	25
Grzegorz Konczak	
On the modification of the non-parametric test for comparing locations of two populations	35
Joan del Castillo, Maria Padilla and Isabel Serra	
Comparison of techniques for extreme values using financial data	45
Paulo C. Rodrigues, Andreia Monteiro and Vanda Lourenço	
New insights into the usefulness of robust singular value decomposition in statistical genetics	53
Borja Lafuente–Rego and Jose Antonio Vilar	
Time series clustering based on quantile autocovariances	61

Frederick Kin Hing Phoa	
A Graphical User Interface Platform of the Stepwise Response Refinement Screener for Screening Experiments	69
Helmut Vorkauf	
Unravel: A Method and a Program to Analyze Contingency Tables, Unveiling Confounders.	81
Juan Eloy Ruiz-Castro	
Preventive maintenance in a complex warm standby system. A transient analysis	89
Pranesh Kumar and Faramarz Kashanchi	
Linear Regression Models Using L_1, L_2 and L_∞-Norms	97
Simon Wilson et al.	
Using Storm for scaleable sequential statistical inference	103
M. Salomé Cabral and M. Helena Gonçalves	
A simulation study to assess statistical approaches for longitudinal count data	111
Matthieu Marbac, Christophe Biernacki and Vincent Vandewalle	
Mixture model of Gaussian copulas to cluster mixed-type data	119
Miguel Casquilho and Elisabete Carolino	
Sampling inspection by (Gaussian) variables via estimation of the lot fraction defective: a computational approach	127
José Antonio Roldan-Nofuentes	
Estimation of the weighted kappa coefficient subject to case-control design	135
Leyla Azarang and Jacobo de Uña-Álvarez	
The jackknife estimate of variance for transition probabilities in the non-Markov illness-death model	143

Ana M. Aguilera and M. Carmen Aguilera-Morillo	
Linear discriminant analysis based on penalized functional PLS	151
D. Ferrari, M. Giuzio and S. Paterlini	
A generalized Description Length approach for Sparse and Robust Index Tracking	157
Paolo Ghisletta, Stephen Aichele, Patrick Rabbitt	
Longitudinal data mining to predict survival in a large sample of adults	167
Fastrich, Paterlini and Winker	
Penalized Least Squares for Optimal Sparse Portfolio Selection	177
Alessandra Amendola and Giuseppe Storti	
Combining information at different frequencies in multivariate volatility prediction	187
Kohei Adachi and Nickolay T. Trendafilov	
Penalty-free sparse PCA	197
Ali Charkhi, Gerda Claeskens and Bruce E. Hansen	
Weight choice by minimizing MSE for general likelihood averaging	205
Jean-Baptiste Durand and Yann Guédon	
Quantifying and localizing state uncertainty in hidden Markov models using conditional entropy profiles	213
S.K. Ng and G.J. McLachlan	
Mixture of regression models with latent variables and sparse coefficient parameters	223
Souleyman Sahnoun and Pierre Comon	
Tensor polyadic decomposition for antenna array processing	233

Caren Hasler and Alina Matei	
Adjustment for nonignorable nonresponse using latent homogeneous response groups	241
Nobuhiro Taneichi, Yuri Sekiya and Jun Toyama	
Bartlett adjustment of deviance statistic for three types of binary response models	249
Yuichi Mori, Masahiro Kuroda, Masaya Iizuka and Michio Sakakihara	
Performance of acceleration of ALS algorithm in nonlinear PCA	257
Niklas Ahlgren and Paul Catani	
Finite-Sample Multivariate Tests for ARCH in Vector Autoregressive Models	265
Miguel Casquilho and Fátima Rosa	
Behaviour of the quality index in acceptance sampling by variables: computation and Monte Carlo simulation	273
Sara Fontanella, Nickolay T. Trendafilov and Kohei Adachi	
Sparse exploratory factor analysis	281
M. Ivette Gomes and Frederico Caeiro	
Efficiency of partially reduced-bias mean-of-order-p versus minimum-variance reduced-bias extreme value index estimation	289
J.R. Wishart	
Data-driven wavelet resolution choice in multichannel box-car deconvolution with long memory	299
Hirohito Sakurai and Masaaki Taguri	
Comparison of block bootstrap testing methods of mean difference for paired longitudinal data	309
M. Arisido	
Functional data modeling to measure exposure to ozone	319

Raquel Caballero-Águila, Aurora Hermoso-Carazo and Josefa Linares-Pérez	
Estimation based on covariances from multiple one-step randomly delayed measurements with noise correlation	327
C. Mante	
Density and Distribution Function estimation through iterates of fractional Bernstein Operators	335
Craig Anderson, Duncan Lee, Nema Dean	
Bayesian cluster detection via adjacency modelling	343
Jan Amos Visek	
Robust test of restricted model	351
Isabelle Charlier, Davy Paindaveine and Jérôme Saracco	
Conditional quantile estimation using optimal quantization: a numerical study	361
Robert M. Kunst	
A combined nonparametric test for seasonal unit roots	369
Hannah Frick, Carolin Strobl and Achim Zeileis	
To Split or to Mix? Tree vs. Mixture Models for Detecting Subgroups	379
Massimo Cannas and Bruno Arpino	
Propensity score matching with clustered data: an application to birth register data	387
Fernanda Figueiredo, M. Ivette Gomes and Adelaide Figueiredo	
Monitoring the shape parameter of a Weibull distribution	395
Mingfei Qiu, Vic Patrangenaru and Leif Ellingson	
How far is the Corpus Callosum of an Average Individual from Albert Einstein's?	403

Kosuke Okusa and Toshinari Kamakura	
Statistical Registration of Frontal View Gait Silhouette with Application to Gait Analysis	411
Pavel Mozgunov	
Application of Kalman Filter with alpha-stable distribution	419
Bernard Fichet	
Combining sub(up)-approximations of different type to improve a solution	427
Elvira Pelle <i>et al.</i>	
Log-linear multidimensional Rasch model for capture-recapture	435
Adelaide Figueiredo and Fernanda Figueiredo	
Monitoring the process variability using STATIS	443
Stefano M. Iacus and Lorenzo Mercuri	
Estimation of Lévy CARMA models in the yuima package: application on the financial time series	451
Christian Derquenne	
Modelling multivariate time series by structural equations modelling and segmentation approach	459
Martin Schindler, Jan Picek and Jan Kysely	
Study on the choice of regression quantile threshold in a POT model	467
Ryo Takahashi	
Reduced K-means with sparse loadings	475
Antonio Irpino, Antonio Balzanella and Rosanna Verde	
Spatial dependence monitoring over distributed data streams	483

F. Marta L. Di Lascio, Simone Giannerini and Alessandra Reale	
Imputation of complex dependent data by conditional copulas: analytic versus semiparametric approach	491
Antonio Abbruzzo, Luigi Augugliaro, Angelo M. Mineo and Ernst C. Wit	
Cyclic coordinate for penalized Gaussian graphical models with symmetry restrictions	499
Sujung Kim, Kuniyoshi Hayashi and Koji Kurihara	
The optimal number of lags in variogram estimation in spatial data analysis	507
F. Giordano, S.N. Lahiri and M.L. Parrella	
GRID for variable selection in high dimensional regression	515
Sakyajit Bhattacharya and Vaibhav Rajan	
Unsupervised Learning using Gaussian Mixture Copula Model	523
Francesco Bartolucci, Giorgio E. Montanari and Silvia Pandolfi	
A comparison of some estimation methods for latent Markov models with covariates	531
Fernández-Pascual, R. Espejo, R and Ruiz-Medina, M.D.	
Estimation of spatially correlated ocean temperature curves including depth dependent covariates	539
Frederico Caeiro and M. Ivette Gomes	
On the bootstrap methodology for the estimation of the tail sample fraction	545
Marco Di Marzio <i>et al.</i>	
Local likelihood estimation for multivariate directional data	553
Pierre Fernique, Jean-Baptiste Durand and Yann Guédon	
Estimation of Discrete Partially Directed Acyclic Graphical Models in Multitype Branching Processes	561

Manuela Neves and Clara Cordeiro	
Statistical modelling in time series extremes: an overview and new steps	569
Luca Frigau, Claudio Conversano and Francesco Mola	
A bivariate cost-sensitive classifier performance index	577
Ronald Hochreiter and Christoph Waldhauser	
Effects of Sampling Methods on Prediction Quality. The Case of Classifying Land Cover Using Decision Trees.	585
Stasi <i>et al.</i>	
β models for random hypergraphs with a given degree sequence	593
A.Wawrzynczak, P.Kopka, M.Jaroszynski and M.Borysiewicz	
Efficiency of Sequential Monte Carlo and Genetic algorithm in Bayesian estimation of the atmospheric contamination source	601
Muhammad-Anas Knefati and Farid Beninel	
Transfer of semiparametric single index model in binary classification	609
Pierre Michel and Badih Ghattas	
Clustering ordinal data using binary decision trees	617
Katrin Illner <i>et al.</i>	
Bayesian blind source separation applied to the lymphocyte pathway	625
Manuela Cattelan	
Maximum simulated likelihood estimation of Thurstonian models	633
Manuela Souto de Miranda, Conceição Amado and Margarida Silva	
Robust profiling of Site Index	641
Charalampos Chaniavidis, Ludger Evers and Tereza Neocleous	
Bayesian density regression for count data	649

Zdeněk Fabián

Score Function of Distribution and Heavy-tails **657**

Ayca Yetere Kursun, Cem Iyigun and Inci Batmaz

Consensus Clustering of Time Series Data **665**

Yusuke Matsui and Masahiro Mizuta

SDA for mixed-type data and its application to analysis of environmental radio activity level data **673**

Pasquale Dolce, Vincenzo Esposito Vinzi and Carlo Lauro

Predictive Component-based Multi-block Path Modeling **681**

Sadika Rjiba, Mireille Gettler Summa and Saloua Benammou

Joint analysis of closed and open-ended questions in a survey about the Tunisian revolution **685**

Computation of Regularized Linear Discriminant Analysis

Jan Kalina, *Institute of Computer Science AS CR*, kalina@cs.cas.cz

Zdeněk Valenta, *Institute of Computer Science AS CR*, valenta@cs.cas.cz

Jurjen Duintjer Tebbens, *Institute of Computer Science AS CR*, duintjertebbens@cs.cas.cz

Abstract. This paper is focused on regularized versions of classification analysis and their computation for high-dimensional data. A variety of regularized classification methods has been proposed and we critically discuss their computational aspects. We formulate several new algorithms for shrinkage linear discriminant analysis, which exploits a shrinkage covariance matrix estimator towards a regular target matrix. Numerical linear algebra considerations are used to propose tailor-made algorithms for specific choices of the target matrix. Further, we arrive at proposing a new classification method based on L_2 -regularization of group means and the pooled covariance matrix and accompany it by an efficient algorithm for its computation.

Keywords. Classification analysis, Regularization, Matrix decomposition, Shrinkage eigenvalues, High-dimensional data

1 Introduction

Classification analysis methods have the aim to construct (learn) a decision rule based on a training data set, which is able to automatically assign new data to one of K groups. Linear discriminant analysis (LDA) is a standard statistical classification method. In the whole paper, we consider n observations with p variables, observed in K different samples (groups) with $p > K \geq 2$,

$$X_{11}, \dots, X_{1n_1}, \dots, X_{K1}, \dots, X_{Kn_K}, \quad (1)$$

where $n = \sum_{k=1}^K n_k$.

LDA assumes the data in each group to come from a Gaussian distribution, while the covariance matrix Σ is the same across groups. Its pooled estimator will be denoted by S . LDA in its standard form assumes $n > p$ and is unsuitable for high-dimensional data with a number of variables exceeding the number of observations (large p /small n problem). In case where $n < p$, the matrix S of size p is singular and computing its inverse must be replaced by an appropriate alternative. Available approaches in this context are based e.g. on pseudoinverse

matrices, which are however unstable due to a small n [4]. Other proposals are based on the generalized SVD decomposition or on elimination of the common null space of the between-group and within-group covariance matrices [2].

Various authors suggested to use a regularized version of LDA for $n \ll p$ [3, 2, 4, 5]. Suitable regularized estimators of the covariance matrix are guaranteed to be regular and positive definite even for $n \ll p$. They have become established e.g. in image analysis, chemometrics, molecular genetics, or econometrics, while their fast computation and numerical stability remains to be an important issue [4, 7]. We will describe the most important approaches and critically discuss their possible computation.

The first approach to a regularized discriminant analysis by [3] is based on a shrinkage covariance matrix with two parameters, which are searched for in a grid search minimizing the classification error. Later, the computation was criticized as computationally intensive in [8], where a linear shrinkage estimator of the covariance matrix was proposed and the asymptotically optimal value of the regularization parameter was derived. The method is implemented in the *corpcor* package of *R* software; however, its computation for a large p is very slow.

Habitually used regularized versions of LDA are based either on regularizing only Σ using one of approaches of [8] or on a double shrinkage applied on the covariance matrix as well as means of each group. The latter approach was proposed by Guo et al. [4], who performed shrinking of the covariance matrix towards an identity matrix and at the same time shrinking of the mean of each group to zero. The method is implemented in the *rda* package of *R* software. For specific values of the parameters, the computation is based on the SVD algorithm, without applying methods of numerical linear algebra to decrease computational costs. The optimal values of shrinkage parameters are optimized in a cross-validation over a 2-dimensional grid, which has been described as tedious [4]. Moreover, there are many possible tuning parameters giving the same cross-validation error rate. The computational effectivity and stability of habitually used algorithms is not investigated even in the recent monograph [7] on covariance matrix estimation for high-dimensional data.

This paper studies efficient algorithms for computing various regularized versions of LDA. Section 2 of this paper formulates several algorithms for shrinkage LDA, which exploits a shrinkage covariance matrix estimator towards a regular target matrix. The computational effectivity of the algorithms is inspected using arguments of numerical linear algebra. For a specific choice of the target matrix, we are able to propose a tailor-made algorithm with a lower computational cost compared to algorithms which are formulated for a general context. Besides, we arrive at proposing new versions of classification methods and accompany them by efficient algorithms for their computation in Section 3. The classification performance of the methods is illustrated on real data in Section 4.

2 Algorithms for Regularized Linear Discriminant Analysis

This section is devoted to proposing and comparing new algorithms for a habitually used version of the regularized LDA [4]. We use suitable matrix decompositions to propose efficient algorithms either for a general choice of T or for its specific choices. To the best of our knowledge, tailor-made algorithms for a specific T have not been described. We compare the new algorithms in terms of their computational costs as well as numerical stability.

We will describe one of habitually used regularized versions of LDA. This will be denoted as LDA* to avoid confusion, because the concept of regularized discriminant analysis encompasses several different methods [4]. A given target matrix T will be used, which must be a regular symmetric positive definite matrix of size $p \times p$. Its most common choices include the identity matrix I_p or a diagonal (non-identity) matrix; other target matrices have been considered by [8].

Let us denote the mean of the observed values in the k -th group ($k = 1, \dots, K$) by \bar{X}_k . LDA* assigns a new observation $Z = (Z_1, \dots, Z_p)^T$ to group k , if $l_k^* > l_j^*$ for every $j \neq k$, where the regularized linear discriminant score for the k -th group has the form

$$l_k^* = \bar{X}_k^T (S^*)^{-1} Z - \frac{1}{2} \bar{X}_k^T (S^*)^{-1} \bar{X}_k + \log p_k, \quad k = 1, \dots, K, \quad (2)$$

where p_k is a prior probability of observing an observation from the k -th group and

$$S^* = \lambda S + (1 - \lambda) T \quad (3)$$

for $\lambda \in [0, 1]$ denotes a shrinkage estimator of the covariance matrix across groups. The situation with $l_k^* = l_{k'}^*$ for $k' \neq k$ does not need a separate treatment, because it occurs with a zero probability for data coming from a continuous distribution. Equivalently, LDA* assigns a new observation Z to group k , if

$$(\bar{X}_k - Z)^T S^{*-1} (\bar{X}_k - Z) = \min_{j=1, \dots, K} \{(\bar{X}_j - Z)^T S^{*-1} (\bar{X}_j - Z)\}. \quad (4)$$

First, the standard approach for computing LDA* may be improved by employing the eigendecomposition of S^* for a fixed λ . A suitable value of λ is found by a cross-validation in the form of a grid search over all possible values of $\lambda \in [0, 1]$.

Algorithm 2.1.

LDA for the general regularization (3) based on eigendecomposition.*

Step 1 Compute the matrix

$$A = [\bar{X}_1 - Z, \dots, \bar{X}_K - Z] \quad (5)$$

of size $p \times K$ whose k -th column is $\bar{X}_k - Z$.

Step 2 Compute S^* according to (3) with a fixed $\lambda \in [0, 1]$.

Step 3 Compute and store the eigenvalues of S^* in the diagonal matrix D_* , and compute and store the corresponding eigenvectors of S^* in the orthogonal matrix Q_* .

Step 4 Compute the matrix

$$B = D_*^{-1/2} Q_*^T A \quad (6)$$

and assign Z to group k if the column of B with largest Euclidean norm is the k -th column.

Step 5 Repeat steps 2 to 4 with different values of λ and find the classification rule with the best classification performance.

The main computational costs are in step 3; the eigendecomposition costs about $9 \cdot p^3$ floating point operations. Note that we need not (and should never) compute the inverse of S^* , thus

avoiding additional computations of the Mahalanobis distance, which is expensive of order p^3 and numerically rather unstable. The group assignment (4) is done by using

$$(\bar{X}_j - Z)^T S^{*-1} (\bar{X}_j - Z) = (\bar{X}_j - Z)^T Q_* D_*^{-1} Q_*^T (\bar{X}_j - Z) = \|D_*^{-1/2} Q_*^T (\bar{X}_j - Z)\|^2. \quad (7)$$

The algorithm can be made cheaper by replacing the eigendecomposition of S^* with its Cholesky decomposition

$$S^* = L_* L_*^T, \quad (8)$$

where L_* is a nonsingular lower triangular matrix. The costs of Cholesky decomposition are about $1/3 \cdot p^3$ floating point operations. On the other hand, Cholesky decomposition will suffer from instability when S^* is not positive definite.

Algorithm 2.2.

LDA for the general regularization (3) based on Cholesky decomposition.*

Step 1 Compute the matrix

$$A = [\bar{X}_1 - Z, \dots, \bar{X}_K - Z] \quad (9)$$

of size $p \times K$ whose k -th column is $\bar{X}_k - Z$.

Step 2 Compute S^* according to (3) with a fixed $\lambda \in [0, 1]$.

Step 3 Compute the Cholesky factor L_* of S^* .

Step 4 Compute the matrix

$$B = L_*^T A \quad (10)$$

and assign Z to group k if the column of B with largest Euclidean norm is the k -th column.

Step 5 Repeat steps 2 to 4 with different values of λ and find the classification rule with the best classification performance.

For specific target matrices, we can further reduce computational costs by using the following algorithm for LDA*. The pooled estimator S can be written in the form

$$S = Y^T Y, \quad Y = [X_{11} - \bar{X}, \dots, X_{1n_1} - \bar{X}, \dots, X_{K1} - \bar{X}, \dots, X_{Kn_K} - \bar{X}]^T \quad (11)$$

where Y is of size $n \times p$. Then using the singular value decomposition (SVD) of Y in the form

$$Y = P \Sigma Q^T, \quad (12)$$

we can express the eigendecomposition of S as

$$S = Y^T Y = (P \Sigma Q^T)^T P \Sigma Q^T = Q \Sigma^2 Q^T. \quad (13)$$

The costs will be about $4 \cdot n p^2$ floating point operations, thus with $p \gg n$ the gain is considerable. Moreover, if

$$S^* = \lambda S + (1 - \lambda) I_p, \quad \lambda \in [0, 1], \quad (14)$$

we immediately obtain the needed eigendecomposition of S^* as

$$S^* = \lambda S + (1 - \lambda) I_p = Q (\lambda \Sigma^2 + (1 - \lambda) I_p) Q^T. \quad (15)$$

The SVD can be computed in a backward stable way with all singular values accurate up to machine precision level [1]. For the special case (14), which is commonly denoted as Tikhonov or ridge regularization of S , a more efficient computation can be performed as follows.

Algorithm 2.3.

LDA* for the ridge regularization (14).

Step 1 Compute the matrix

$$A = [\bar{X}_1 - Z, \dots, \bar{X}_K - Z] \quad (16)$$

of size $p \times K$ whose k -th column is $\bar{X}_k - Z$ and compute the matrix Y in (11).

Step 2 Compute the singular value decomposition of Y as

$$Y = P\Sigma Q^T, \quad (17)$$

with singular values $\{\sigma_1, \dots, \sigma_n\}$ and complement these singular values with $p - n$ zero values $\sigma_{n+1} = \dots = \sigma_p = 0$.

Step 3 For a fixed $\lambda \in [0, 1]$, compute

$$D_* = \text{diag}\{\lambda\sigma_1^2 + (1 - \lambda), \dots, \lambda\sigma_p^2 + (1 - \lambda)\}. \quad (18)$$

Step 4 Compute the matrix

$$B = D_*^{-1/2} Q^T A \quad (19)$$

and assign Z to group k if the column of B with largest Euclidean norm is the k -th column.

Step 5 Repeat steps 2 to 4 with different values of λ and find the classification rule with the best classification performance.

Eigenvalues of the regularized covariance matrix forming the matrix D^* in (18) can be interpreted as shrinkage eigenvalues.

In an analogous manner, algorithms for a regularized quadratic discriminant analysis (QDA) can be obtained, using a regularized estimator of the covariance matrix in each group separately.

3 L_2 -regularized linear discriminant analysis

Disadvantages of SCRDA [4] include a computational intensity as well as an inconsistent approach to shrinkage. The means are namely modified by an L_1 -norm regularization and the covariance matrix in the sense of the L_2 -norm. As an alternative, this section proposes a new regularized version of LDA denoted as L_2 -LDA together with an efficient algorithm for its computation. It employs a shrinkage estimator of Σ and shrunken means towards the overall mean across groups. As a unique feature, both shrinkage approaches have the form of an L_2 -norm regularization.

The classification rule of L_2 -LDA assigns a new observation Z to the k -th group, if $l_k^\dagger > l_j^\dagger$ for every $j \neq k$, where

$$l_k^\dagger = \bar{X}_k'^T (S^*)^{-1} Z - \frac{1}{2} \bar{X}_k'^T (S^*)^{-1} \bar{X}_k' + \log p_k \quad (20)$$

and \bar{X}_k' denotes the shrunken mean of the k -th group towards the overall mean computed across groups. The method can be interpreted as based on a L_2 regularized Mahalanobis distance. As another contrast with the habitually used algorithm of SCRDA [4], we will estimate the

parameter λ in a straightforward way using an asymptotically optimal value minimizing the mean square error [8]. To avoid confusion, the asymptotically optimal value of λ will be denoted by λ^\dagger and the corresponding shrinkage covariance matrix by

$$S^\dagger = \lambda^\dagger S + (1 - \lambda^\dagger)T. \quad (21)$$

Algorithm 3.1.

L₂-LDA.

Step 1 Compute λ^\dagger as

$$\lambda^\dagger = \frac{2 \sum_{i=2}^p \sum_{j=1}^{i-1} \widehat{\text{var}}(S_{ij})}{2 \sum_{i=2}^p \sum_{j=1}^{i-1} S_{ij}^2 + \sum_{i=1}^p (S_{ii} - 1)^2}, \quad (22)$$

where $\widehat{\text{var}}(S_{ij})$ is the maximum likelihood estimator of the variance of values S_{ij} for a fixed i and j .

Step 2 Compute and store the eigenvalues of S^\dagger in the diagonal matrix D_* , and compute and store the corresponding eigenvectors of S^\dagger in the orthogonal matrix Q_* .

Step 3 For a fixed $\delta \in [0, 1]$, compute $\bar{X}'_k = \delta \bar{X}'_k + (1 - \delta)\bar{X}$, $k = 1, \dots, K$.

Step 4 Assign Z to group k , if

$$\|D_*^{-1/2} Q_*^T (\bar{X}'_k - Z)\| = \min_{j=1, \dots, K} \|D_*^{-1/2} Q_*^T (\bar{X}'_j - Z)\|. \quad (23)$$

Step 5 Repeat steps 3 and 4 for various δ and find the optimal classification rule yielding the best classification performance.

Algorithm 3.1 is formulated for a general target matrix T . For a specific choice of T , a computationally cheaper method can be obtained in an analogous way as Algorithms 2.2 and 2.3 from the general algorithm 2.1.

Another possibility is to regularize the within-group covariance matrix instead of regularizing S , which is however computationally more intensive.

4 Examples

We present two examples on real molecular genetic data sets in order to illustrate the behavior of the newly proposed L_2 -LDA method.

Example 1 contains data from a cardiovascular genetic study of the Center of Biomedical Informatics in Prague performed in 2006–2011. The data contain expressions of $p = 38\,590$ gene transcripts measured on 24 patients having a cerebrovascular stroke and 24 control persons.

In Example 2, a prostate cancer metabolomic data set [9] is analyzed, which contains $p = 518$ metabolites measured over two groups of patients, namely those with a benign prostate cancer (16 patients) and with other cancer types (26 patients). The task in both examples is to learn a classification rule allowing to discriminate between the two classes of individuals.

In both examples, we computed the classification methods described in this paper using the algorithms of Sections 2 and 3. For comparison, we computed also other available classification

Method	S^*	R Package	Function	Youden's index	
				Example 1	Example 2
SVM	-	<i>e1071</i>	<i>svm</i>	1.00	1.00
Classification tree	-	<i>tree</i>	<i>tree</i>	0.94	0.97
Self-organizing map	-	<i>kohonen</i>	<i>som</i>	0.88	0.93
Multilayer perceptron	-	<i>nnet</i>	<i>nnet</i>	Infeasible	Infeasible
LDA	-	<i>MASS</i>	<i>lda</i>	Infeasible	Infeasible
SCRDA	(14)	<i>rda</i>	<i>rda</i>	1.00	1.00
LDA*	(14)	-	-	1.00	1.00
LDA*	(24)	-	-	1.00	1.00
L_2 -LDA	(14)	-	-	1.00	1.00
L_2 -LDA	(24)	-	-	1.00	1.00
PCA \implies LDA	-	-	-	0.54	0.90
PCA \implies SCRDA	(14)	-	-	0.71	0.92
PCA \implies LDA*	(14)	-	-	0.63	0.81
PCA \implies LDA*	(24)	-	-	0.63	0.81
PCA \implies L_2 -LDA	(14)	-	-	0.71	0.92
PCA \implies L_2 -LDA	(24)	-	-	0.71	0.92
PCA \implies MWCD-LDA	-	-	-	0.69	0.90

Table 1: Results of Example 1 and Example 2. LDA* was computed using Algorithm 2.3 for the choice (14) and Algorithm 2.2 for (24). L_2 -LDA was computed using Algorithm 3.1. PCA uses 20 principal components.

methods, including the support vector machines (SVM), a classification tree, Kohonen's self-organizing map, a multilayer perceptron with 2 hidden layers, or the highly robust classification method MWCD-LDA of [6]. Various regularized versions of LDA include the most common choice $T = I_p$ or another choice

$$S^* = \lambda S + (1 - \lambda)sI_p, \quad \lambda \in [0, 1], \quad s = \sum_{i=1}^p S_{ii}/p. \quad (24)$$

We used the default settings to compute them in R software packages, which are listed also in Table 1. The classification performance is measured by means of the Youden's index, which is defined as sensitivity + specificity - 1. The dimensionality reduction was performed by the principal component analysis (PCA) with 20 principal components.

The results performed on raw data as well as after a dimensionality reduction reveal that the regularized versions of LDA perform quite similarly. The newly proposed method L_2 -LDA with an efficient algorithm seems to perform comparably with the available regularized methods with less efficient computation. Besides, the choice of the target matrix T does not seem to play an important role.

Further, we investigated the reduction in classification performance after reducing the dimensionality to 20 principal components in both examples. The approach of Algorithm 3.1 (PCA \implies L_2 -LDA) yields improved results compared to its standard counterpart (PCA \implies LDA). The results of regularized methods do not greatly differ from the robust MWCD-LDA procedure, which indicates that regularized versions of LDA do not greatly suffer by the presence of

outlying measurements in the data. Nevertheless, the robustness of regularized methods with respect to outliers has not been systematically investigated [5].

To conclude the paper, several new algorithms for shrinkage LDA are proposed, exploiting a shrinkage covariance matrix estimator towards a regular target matrix. Some algorithms are tailor-made for a specific choice of the target matrix and their computational costs are discussed. A new regularized classification method L_2 -LDA is proposed and accompanied by an efficient algorithm. An analysis of two real data sets reveals its classification performance to be comparable to available regularized classification methods for high-dimensional data.

Acknowledgement

The work of J. Kalina was financially supported by the Neuron Foundation for Supporting Science. The work of Z. Valenta was supported by the institutional support RVO:67985807. The work of J. Duintjer Tebbens was supported by the grant GA13-06684S of the Czech Science Foundation.

Bibliography

- [1] Barlow, J.L., Bosner, N. and Drmač, Z. (2005) *A new stable bidiagonal reduction algorithm*. Linear Algebra and its Applications, **397**, 35–84.
- [2] Duintjer Tebbens, J. and Schlesinger P. (2007) *Improving implementation of linear discriminant analysis for the high dimension/small sample size problem*. Computational Statistics & Data Analysis, **52**, 423–437.
- [3] Friedman, J.H. (1989) *Regularized discriminant analysis*. Journal of the American Statistical Association, **84**, 165–175.
- [4] Guo, Y., Hastie, T. and Tibshirani, R. (2007) *Regularized discriminant analysis and its application in microarrays*. Biostatistics, **8**, 86–100.
- [5] Kalina, J. (2014) *Classification methods for high-dimensional data*. Biocybernetics and Biomedical Engineering, **34** (1), 10–18.
- [6] Kalina, J. (2012) *Highly robust statistical methods in medical image analysis*. Biocybernetics and Biomedical Engineering, **32** (2), 3–16.
- [7] Pourahmadi, M. (2013) *High-dimensional covariance estimation*. Wiley, New York.
- [8] Schäfer, J. and Strimmer K. (2005) *A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics*. Statistical Applications in Genetics and Molecular Biology, **4**, Article 32.
- [9] Sreekumar, A. et al. (2009) *Metabolomic profiles delineate potential role for sarcosine in prostate cancer progression*. Nature, **457**, 910–914.

Fast Detection of Structural Breaks

Paul Fischer, *Technical University of Denmark*, pafi@dtu.dk

Astrid Hilbert, *Linnaeus University, Sweden*, astrid.hilbert@lnu.se

Abstract. A fundamental task in the analysis of time series is to detect structural breaks. A break indicates a significant change in the behaviour of the series. One method to formalise the notion of a break point, is to fit statistical models piecewise to the series. To find break points, the endpoints of the pieces are varied as is their number. A structural break is indicated by a significant change of the model parameters in adjacent pieces. Both, varying the pieces and repeatedly fitting models to them, are usually computationally very expensive. By combining genetic algorithms with a preprocessing of the time series we design a very fast algorithm for structural break detection. It reduces the time for model-fitting from linear to logarithmic in the length of the series. We show how this method can be used to find structural breaks for time series which are piecewise generated by AR(p)-models. Moreover, we introduce a non-parametric model for which the speed-up can also be achieved. Additionally we briefly present simulation results which demonstrate the manifold applications of these methods. A reference implementation is available at <http://www2.imm.dtu.dk/~pafi/StructBreak/index.html>

Keywords. Structural breaks, parametric and non-parametric models, efficient algorithms, range trees.

1 Introduction

We consider the problem to detect structural breaks in time series. A structural break is a point in time, where the behaviour of the time series changes. What precisely a “change of behaviour” (also called change of *regime*) is depends on the application. It might be a change in the level of the observed data or a change of the magnitude of the local variance (the volatility in terms of econometrics).

Often the times series is assumed to be generated by a known stochastic process. In this case, a structural break is defined as a change of the type of the underlying model or of its parameters. For example if an autoregressive (AR-) model is assumed, a structural break can be a change of the order (number of numerical parameters) or a significant change of the values of these parameters. See [4] for a recent overview.

From the above discussion there clearly cannot be a single algorithm for the detection of structural breaks. Very often visual inspection of the plot of the time series by an expert does

it. We provide a generic framework for the design of such algorithms, where the user has to supply the application specific knowledge, basically a procedure to evaluate how good a set of points is as structural breaks.

In practice, background information to guide the search for break points is in most cases not available and exhaustive search is not an option already for moderately long time series and few break points. The problem can be formulated as a black box optimisation problem where evolutionary algorithms are an obvious choice for a heuristic. For finding break points, the algorithm starts with a number of sets of candidate break points. In the course of the algorithm these sets are modified by moving, deleting or adding break points. The goodness of a set of break points is evaluated and better sets are kept while worse ones are deleted, details can be found in [3]. The efficiency of evolutionary algorithms has been proved also in other areas of statistics, see [5].

For series generated by AR-Models, Davis et. al. [1] propose an evolutionary algorithm for break point detection. The resulting algorithm requires repeatedly fitting AR-models to parts of the time series, making it computationally quite demanding and limiting its use to relatively short time series of a few thousands observations.

We present a generic framework which allows AR- and other parametric and non-parametric models to be used and which is computationally much more efficient. With our approach we can efficiently handle time series with millions of observations. The work was motivated by an industrial application where very long time series (some millions of observations) had to be analysed. Also the non-parametric method described in Section 3 has been designed for this application because it simultaneously finds structural breaks and outliers in this type of series.

After introducing the notation in Section 2, the general framework for finding structural breaks with evolutionary algorithms is introduced in Section 3. In Section 4 the central data structure is introduced, which speeds up the algorithm. It is shown how different statistical models can be adapted to profit from the speed-up. Simulation results on the running times and precision of the algorithms are presented in Section 5.

A reference implementation which can be used on user-supplied time-series is available on the net at <http://www2.imm.dtu.dk/~pafi/StructBreak/index.html>.

2 Notation and Problem Description

Formally a univariate time series is a sequence $Y = (y_0, \dots, y_{T-1})$ of real numbers, where y_t is the observation at time $t \in \{0, \dots, T-1\}$. For notational convenience, we assume that the observations are equidistant in time, though this is not necessary. Non-equidistant observations might however have an influence on the complexity of fitting the model.

The (indices of the) *break points* constitute an integer sequence (b_0, \dots, b_k) , where $b_j \in \{0, \dots, T\}$, and $b_j < b_{j+1}$, for $j = 0, \dots, k-1$. We assume that $b_0 = 0$, as it is the starting point of the first regime, and we set $b_k = T$, the first index after the end of the time series. In practice, a minimum distance m between successive break points might be assumed ($b_j + m < b_{j+1}$) in order to avoid unreasonable short regimes. However, for the use in outlier detection with the rectangle model, it is essential that break points can be close. Associated with each break point b_j might be a statistical model M_j valid in the interval $[b_j; b_{j+1} - 1]$ with parameter set B_j .

3 A Solution based on Evolutionary Algorithms

The general framework of our approach for break point detection, which uses evolutionary algorithms, is as follows: Initially a number of candidate solutions is created, for example randomly. Each solution is a set of indices and model parameters (potential break points). Then the evolutionary algorithm creates a new solution set by: moving the positions of some points, deleting some, creating new ones, or changing model parameters. The new solution is then evaluated, and compared to the already existing solutions. If the new solution is better than the worst old one, the latter is replaced by the new one. The actual creation of a new solution is performed by applying random crossover operations or mutations to existing solutions. In order to make the implementation space-efficient, we use an implicit representation of the break points. The algorithm is run multiple times. Then the best solutions of each run are used as start solutions for a final run.

The evolutionary algorithm is implemented in a generic manner, which means that the user has to supply the “modules” that guide the execution. The essential module is the so called *fitness function*, which associates a real number (the fitness) to a given solution, measuring how good a given solution is. Informally, we will define the fitness functions as follows: We are given a time series (y_0, \dots, y_{T-1}) , a sequence (b_0, \dots, b_k) of candidate break points, and a class of statistical models, which are assumed to generate disjoint pieces of the series. We then fit a model M_j to every interval $[b_j, b_{j+1} - 1]$, $j = 0, \dots, k - 1$. For every model M_j we determine the *goodness of fit* g_j on the corresponding interval. This can, for example, be the sum of the absolute (or squared) residuals. Then the total fitness f associated to the sequence of break points is the sum of individual ones. The fitness function can also contain a term $p(k)$ controlling the number k of break points, normally penalising a high number:

$$g_j = \sum_{i=b_j}^{b_{j+1}-1} |y_i - M_j(i)| \quad \text{and} \quad f = f(b_0, \dots, b_k) = \left(\sum_{j=0}^{k-1} g_j \right) + p(k). \quad (1)$$

The idea is that the best fit will be achieved when the candidate break points are the “true” break points, i.e., when f is minimal. The other essential module the user has to supply is a (randomised) method to generate the model data for a new break point, e.g., the order of the AR-model to be used in the next interval, see [3] for a detailed description of the algorithm.

Subsequently we present two approaches how to evaluate the quality of a given solution into a real numbered fitness value. The first one, called *axes-aligned rectangles*, does not assume that the time series is piecewise generated by a certain process. The second one, named *piecewise AR-models*, assumes that the time series is piecewise generated by AR-models.

Axes-aligned Rectangles. In this model we want to cover the time series with few axes-aligned rectangles having a small total area. Clearly a minimum number of rectangles and a minimal total area of them are conflicting aims. The fitness function must realise a balance between them. A minimum number of rectangles would be achieved by using only one, the bounding box of the whole time series, giving no internal break point. A minimum area would be achieved by using a single zero-area rectangle at each observation of the time series, making all points break points. Both are not desirable solutions.

Intuitively, the rectangle method detects large consecutive parts where the time series is almost constant and small parts where it is rapidly in- or decreasing. Instead of applying the

method to the original time series, it often yields better results when applied to a derived series such as the series of moving variances.

For a formal definition, let (b_0, \dots, b_k) be a sequence of candidate break points. Between break points b_j, b_{j+1} , we use the minimum axes-aligned rectangle (bounding-box) which contains the observations $y_{b_j}, \dots, y_{b_{j+1}-1}$. That is the rectangle R_j defined by

$$R_j = [b_j; b_{j+1} - 1] \times [\min\{y_i \mid i = b_j, \dots, b_{j+1} - 1\}; \max\{y_i \mid i = b_j, \dots, b_{j+1} - 1\}] .$$

An example is shown in Figure 1. All information stored at the break point is its index b_j for this non-parametric model. Let R denote the minimum axes-aligned rectangle containing all observations y_i , i.e., R is the bounding box of the whole time series. We define the fitness

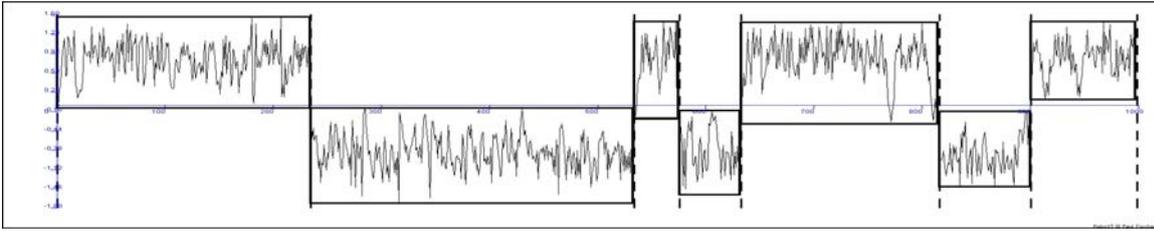


Figure 1: Example of a time series and a cover by rectangles.

function, to be maximised in this case, as follows. Given the time series of observations y_i , the fitness function f depends only on the break points (b_0, \dots, b_k) and is composed of two terms. As in (1), the first term f_a is responsible for minimising the area of the R_j (goodness of fit), by maximising the area of R not covered by the R_j and normalising to $[0; 1]$:

$$f_a = f_a(b_0, \dots, b_k) = \frac{\text{area}(R) - \sum_{j=0}^{k-1} \text{area}(R_j)}{\text{area}(R)} .$$

The second term $f_r = f_r(k)$ is responsible for minimising the number k of break points. We would like also to normalise f_r to $[0; 1]$, for which we propose two approaches: One uses a decreasing function of k , for example $1/k$, $1/\sqrt{k}$, or $1/\ln(e + k - 1)$. The first one decreases fastest and thus prefers few break points, the last one decreases slowest thus allowing more break points. Without any a priori knowledge, $f_r(k) = 1/\sqrt{k}$ proved to be a good choice in our experiments. If one roughly knows how many break points to expect, other choices for f_r are meaningful. The fitness function is then defined as

$$f(b_0, \dots, b_k) = f_a(b_0, \dots, b_k) + \alpha f_r(k) \quad (2)$$

for some choice of f_r . The parameter $\alpha \in [0; 1]$ controls the balance between minimising the area (goodness of fit) and minimising the number of intervals. Values in the range $[0.10; 0.75]$ give good results in experiments on artificial and real world time series.

The identification of break points in a real-world time series is always subjective. For our empirical evaluation we therefore used time series, where the experts we asked agreed on the positions of the break points. Additionally we produced a number of artificial time series, where the break points are clearly defined as level changes.

The computational complexity to compute the fitness of a solution is dominated by finding the maxima and minima for all intervals $[b_j, b_{j+1} - 1]$. In Section 4 we will see that this can be performed in time $O(\log(T))$ for every interval, thus in time $O(k \log(T))$ for $k - 1$ intervals.

Piecewise AR-models. In this setting we assume that there are indices $0 = s_0, \dots, s_m = T$ (the “true” break points indices), such that the time series is generated by a particular $\text{AR}(p_j)$ -model for every interval $[s_j, s_{j+1} - 1]$. The orders p_j and parameters of the models may vary. The task is to identify the break points s_j . As a side effect, the algorithms also produces estimations for the AR-models, that is the orders p_j and the parameters of the models.

As fitness function we used the sum of absolute (or squared) residuals to be minimised. Let M_j be the AR-Model, say of order p_j , fitted to the interval $[b_j; b_{j+1} - 1]$

$$y_i = c + \phi_1 y_{i-1} + \dots + \phi_{p_j} y_{i-p_j} + \varepsilon_j \quad (3)$$

where c is a constant, and the residual noise ε_i is $\mathcal{N}(0, \sigma)$ distributed for some $\sigma > 0$. The model is evaluated at indices $b_j + p_j$ through $b_{j+1} - 1$ using the observations y_i of the time series. Let $M_j(i)$ be the value provided by the model at index i . We compute

$$f_j = \sum_{i=b_j+p}^{b_{j+1}-1} |M_j(i) - y_i| \quad \text{and} \quad f = \sum_{j=0}^{k-1} f_j, \quad (4)$$

where k is the number of break points, and use f as one term in the fitness function. For an alternative, the sum of squares, we could not observe a significant change in the location of the break points found.

The evolutionary algorithm starts again by randomly allocating a number of candidate break points (b_0, \dots, b_k) . The AR-model M_j associated with break point b_j is specified by a set B_j of parameters. (The set B_k at the end of the series irrelevant.) Then the evolutionary algorithm modifies the locations b_j and the orders p_j . The user has to supply a rule for modifying the model order. A straightforward way, used in our implementation, is to choose a maximum model order p_{max} and select the p_j randomly from $0, 1, \dots, p_{max}$. We chose not to use more sophisticated methods like analysing the partial autocorrelation function, Akaike’s information criterion (AIC) or the Bayesian information criterion (BIC), see for example [7], because we want to support randomised nature of evolutionary algorithms.

The fitness function again consists of two parts: 1) For every interval $[b_j, b_{j+1} - 1]$, an $\text{AR}(p_j)$ -model is fitted (by setting up and solving the Yule-Walker equations). The fitness value is evaluated as in Equation (4). 2) A term minimising the number of break points. These are again conflicting aims. However, we observed that the second term (number of break points) is of much less importance than for the rectangle model. For many time series the results do not change when the first term (goodness of fit) is given very high (or even all) weight in the fitness function. The reason is, that many break points give rise to shorter intervals. Fitting AR-models to short intervals results in a worse fit, because the noise is not filtered well. This effect implicitly reduces the number of break point. There are, however, cases where the second term is essential for finding the right number and locations of the break points, e.g., a series which is composed of few AR-models having all the same order and only slightly differing coefficients.

Setting up the Yule-Walker equations is by far the computationally most demanding sub-task. We shall see below that, with an appropriate preprocessing, we can use a range tree to fit an $\text{AR}(p)$ -model to any interval $[b_j, b_{j+1} - 1]$ in time $O(\log(T))$, for $p \leq p_{max}$ and p_{max} constant. If p_{max} is not assumed constant then the time is bounded by $O(p_{max}^3 \log(T))$.

4 Range Trees

Range trees are a general data structure which support multiple queries on intervals of indexed data. A description of the general concept of a range tree may be found in [2]. We restrict the presentation to the situation where the data is a time series (y_0, \dots, y_{T-1}) , that is, we consider range trees for one-dimensional numerical data. A *range query* receives two indices a, b (the range), $0 \leq a \leq b \leq T - 1$, as inputs and returns as *answer* a quantity $q(a, b) = q(y_a, \dots, y_b)$ which is determined by the data y_a, \dots, y_b . We first describe the concept for the case where the query asks for the maximum value in a range (interval) $[a, b]$: $q(a, b) = \max\{y_i \mid a \leq i \leq b\}$. For a single query on range $[a, b]$, the most efficient way is to compute the answer $q(a, b)$ directly from the data, which takes time $O(b - a) = O(T)$. If multiple queries with different ranges have to be performed, a preprocessing might pay off. A straightforward preprocessing is to compute the maxima for all $T(T + 1)/2$ ranges $[a, b]$ in advance and store them in a table. Then a query $[a, b]$ can be answered by a look-up in the table in constant time. The time for the preprocessing is $\Theta(T^2)$. The quadratic preprocessing time and, especially, the quadratic space requirement make this approach infeasible already for medium data sizes of around 10,000.

The idea behind range trees is to compute the maxima only for a few ranges and then combine this information to determine the maximum for any other range. For example, if one knows the maxima for two adjacent ranges, $\max(a, c - 1) = \max\{y_i \mid a \leq i \leq c - 1\}$ and $\max(c, b) = \max\{y_i \mid c \leq i \leq b\}$, then the maximum for the range $[a, b]$ can be computed by a single addition $\max(a, b) = \max(a, c - 1) + \max(c, b)$. It is this *merging property* which allows the use of range trees. A range tree is a rooted, binary tree where every node covers a range $[a, b]$ and the left and right child, respectively, cover the ranges $[a, (a + b)/2]$ and $[(a + b)/2 + 1, b]$ (here and in the following we omit the details for handling the case that the division by 2 gives a remainder). The root covers all the data, i.e., range $[0, T - 1]$. The range tree is constructed bottom up, starting with T leaves formed by singleton ranges $[a, a]$ for which the maximum trivially is a . Then pairs of adjacent nodes are merged and the common maximum is stored in a new node which is the parent of two. For the maximum problem the preprocessing time is $\Theta(T)$ and the query time is $O(\log(T))$.

Preparing Rectangle Models for Range Trees

In order to apply the rectangle method, we have to be able to find the maximum $\max(a, b) = \max\{y_i \mid a \leq i \leq b\}$ for every interval $[a, b]$ and likewise the minimum $\min(a, b)$. To achieve this, we store at all nodes of the range tree the minimum and maximum of the range the node covers. The preprocessing time is $\Theta(T)$ and the query time is $O(\log(T))$.

Preparing AR-Models for Range Trees

In order to fit an AR-model of order p to an interval $[a, b]$ of the time series, $0 \leq a < b \leq T - 1$, using the Yule-Walker equations, we have to know the sample autocovariances r_ℓ for $\ell = 0, \dots, p$. In order to use range trees, one has to be able to compute the autocovariances for interval $[a, b]$ from those of the intervals $[a, (a + b)/2]$ and $[(a + b)/2 + 1, b]$. Instead of storing the sample autocovariances r_ℓ at the nodes of the range tree, we store a number of values from which the r_ℓ 's can be computed. These are the sum of the values of the part of the time series covered by the range and the sum of products of values with lags 0 and ℓ :

$$\mathbb{U}_\ell = \sum_{i=a}^{b-\ell} y_{i+\ell}, \quad V_\ell = \sum_{i=a}^{b-\ell} y_i y_{i+\ell}, \quad \text{for } \ell = 0, 1, \dots, p. \quad (5)$$

Also the size $S = b - a + 1$ of the range is stored. Then the autocovariance r_ℓ for lag ℓ becomes

$$r_\ell = \frac{V_\ell}{(S - \ell + 1)} - \frac{\mathbb{U}_0 \mathbb{U}_\ell}{(S - \ell + 1)^2}. \quad (6)$$

Now we are in a position to describe the merging step which is crucial for using the range tree. Let $a < c < b$ be three indexes of the time series. Let I' denote the interval $[a, c]$, and I'' denote $[c + 1, b]$. Let $S' = c - a + 1$, $S'' = b - c$, and $S = b - a + 1$. Let $I = I' \cup I'' = [a, b]$. Let $(\mathbb{U}_\ell, V_\ell)$ (resp. $(\mathbb{U}'_\ell, V'_\ell)$ and $(\mathbb{U}''_\ell, V''_\ell)$) be the aforementioned parameters for I (resp. I' and I'').

The parameters for I can be computed from those of I' and I'' by

$$S = S' + S'', \quad \mathbb{U}_\ell = \mathbb{U}'_\ell + \mathbb{U}''_\ell, \quad V_\ell = V'_\ell + V''_\ell, \quad \text{for } \ell = 0, \dots, p.$$

In addition, some values at the merging point c have to be computed. Since, for example, V'_ℓ only contains products $y_i y_{i+\ell}$ where both i and $i + \ell$ are in I' . Hence those with $i \in I'$ and $i + \ell \in I''$ have to be added to V . Alternatively, this data can be precomputed and stored in the range tree. When the maximum order p of an AR-model is fixed then the construction of the range tree can be done in time $O(T)$ and the time to fit an AR-model to a given range is $O(\log(T))$.

5 Simulation results

We only mention the most important results of our simulations, a much more detailed description can be found at the website below. The tests have been performed on artificial and real-world time series. <http://www2.imm.dtu.dk/~pafi/StructBreak/pfah-simulations.pdf>.

Using range trees for the AR-model pays off for series longer than 400. For series with 50,000 observations, fitting an AR-model to a randomly selected range is on average 40 times faster using range trees than a direct fit and 500 times faster for series with 1 million observations.

In order to evaluate how well break points are found by our methods, we tested them on artificially generated time series with well defined break points, but also on real world time series, where the “true” break points had been determined by human experts. In all cases, the break points were found with high precision, for short series ($T \leq 2000$) most of the time perfectly. For the rectangle model applied to series like the one in Figure 1 almost always the true break points were recovered precisely. For artificial series constructed of 2 to 5 AR-models of order 1 to 5, the break points (changes between models) were found within maximal ± 5 indices.

One has to remark, that using the residuals to measure the goodness of fit for the AR-model requires linear time $O(T)$, however this is at least one order of magnitude less than the time for naively fitting the model. In order to overcome using linear time, one can instead measure the goodness of fit by using the Minimum Description Length principle of Rissanen [6], see Davis [1] for details, which can be computed in constant time for fixed model order p . In our tests, however, the break points found, did not match the true ones as good as when using the residuals.

6 Conclusion

We have shown how evolutionary algorithms and an efficient data structure can be combined into very efficient and effective algorithms for detecting structural breaks. The method applies to time series which are piecewise generated by statistical models, which meet the “merging condition”, i.e., that the information for adjacent ranges can be merged into the information of the union of the ranges. For two such models, the rectangle model and AR-models, it is shown how the algorithm is used and that it performs very well on artificial and real world data.

The method generalises to higher dimensional data. For d -dimensional data, the complexity depends on the type of range queries. If every query uses the same range in all dimensions, then the time and space requirements are only increased by a factor d . Otherwise, the preprocessing time for the range tree is $O(T \log^{d-1}(T))$ and the query time is $O(\log^d(T))$. An implementation of the rectangle method for multi-dimensional data is under construction. We are also working on an R-callable Java implementation.

In an industrial application, we have used the rectangle method with great success to detect outliers in the series. It is a challenge to find further statistical models which meet the merging condition and thus allow the speed-up by using range tree.

Acknowledgement We would like to thank the reviewers for their very helpful comments.

Bibliography

- [1] R.A. Davis, T. Lee, and G. Rodriguez-Yam. Structural break estimation for nonstationary time series models. *J. of the American Statistical Association*, 101 (473):223–239, 2006.
- [2] M. de Berg, M. van Krefeld, M. Overmars, and O. Schwarzkopf. *Computational Geometry: Algorithms and Applications*. Springer, second edition, 2000.
- [3] Benjamin Doerr, Paul Fischer, Astrid Hilbert, and Carsten Witt. Evolutionary algorithms for the detection of structural breaks in time series. In *Proceeding of the fifteenth annual conference companion on Genetic and evolutionary computation conference companion*, pages 119–120. ACM, 2013.
- [4] N. Haldrup, R. Kruse, T. Timo Teräsverta, and R.T. Varneskov. Unit roots, nonlinearities and structural breaks. In N. Hashimzade and M.A. Thornton, editors, *Handbook of Research Methods and Applications in Empirical Macroeconomics*, chapter 4, pages 61–94. Edward Elgar Publishing Ltd, 2013.
- [5] Robin Nunkesser and Oliver Morell. An evolutionary algorithm for robust regression. *Computational Statistics & Data Analysis*, 54(12):3242–3248, 2010.
- [6] J Rissanen. *Stochastic Complexity in Statistical Inquiry*. World Scientific Series in Computer Science. World Scientific, Singapore, 1989.
- [7] Petre Stoica and Yngve Selén. Model-order selection. *IEEE Signal Processing Magazine*, 21(4):36–47, 2004.

Random Start Forward Searches for Detecting Mixtures of Regression Models

Anthony C. Atkinson, *London School of Economics*, a.c.atkinson@lse.ac.uk

Marco Riani, *Università di Parma, Italy*, mriani@unipr.it

Andrea Cerioli, *Università di Parma, Italy*, andrea.cerioli@unipr.it

Domenico Perrotta, *Joint Research Centre, Ispra, Italy*, domenico.perrotta@ec.europa.eu

Abstract. To detect outliers from a single regression model requires one, perhaps robust, fit to the data. But if the “outlying observations” are other regression models, it may be necessary to fit several different linear models in order to reveal the structure. We illustrate the diagnostic use of random start forward searches to reveal mixtures of regression models.

Keywords. Graphics, Least trimmed squares, Outliers, Regression diagnostics, Robust regression, Trade data

1 Introduction

The international trade data that inspired this study come from several linear regression models that need to be distinguished. To detect outliers from a single regression model requires one, perhaps robust, fit to the data. But if the “outlying observations” are other regression models, it may be necessary to fit several different linear models in order to reveal the structure. In this paper we illustrate the use of random start forward searches in exploring such mixtures of regression models.

The Forward Search (FS) for a robust, diagnostic fit to a single regression model proceeds by fitting subsets of the data of increasing size. The details are in **S2**. In **S3** the random start FS is illustrated on an example of 180 observations arising from international trade. Forward plots of aspects of the data as the subset size increases clearly reveal the structure. In **S4** we compare our results with those obtained by robust fitting under the assumption of a single model. The proposed method involving random starts, coupled with the graphical monitoring of residuals during the FS, provides a powerful diagnostic method for detecting data coming from a mixture of regression models.

2 The Forward Search for Regression Data

The forward search achieves robustness by fitting the model to subsets of the data of increasing size, where the subsets are sequentially chosen to be as close as possible to the fitted model. The introduction of outliers into the subset is diagnostically revealed by plots of residuals against subset size as well as formally by statistically tuned tests.

In the regression model $y = X\beta + \epsilon$, y is the $n \times 1$ vector of responses, X is an $n \times p$ full-rank matrix of known constants, with i th row x_i^T , and β is a vector of p unknown parameters. The normal theory assumptions are that the errors ϵ_i are i.i.d. $N(0, \sigma^2)$.

The least squares estimator of β is $\hat{\beta}$. Then the vector of n least squares residuals is $e = y - \hat{y} = y - X\hat{\beta} = (I - H)y$, where $H = X(X^T X)^{-1}X^T$ is the ‘hat’ matrix, with diagonal elements h_i and off-diagonal elements h_{ij} . The residual mean square estimator of σ^2 is $s^2 = e^T e / (n - p) = \sum_{i=1}^n e_i^2 / (n - p)$.

FS fits subsets of observations of size m to the data, with $m_0 \leq m < n$. Let $S^*(m)$ be the subset of size m found by FS, for which the matrix of regressors is $X(m)$. Least squares on this subset of observations yields parameter estimates $\hat{\beta}(m)$ and $s^2(m)$, the mean square estimate of σ^2 on $m - p$ degrees of freedom. Residuals can be calculated for all observations including those not in $S^*(m)$. The n resulting least squares residuals are

$$e_i(m) = y_i - x_i^T \hat{\beta}(m). \quad (1)$$

The search moves forward with the augmented subset $S^*(m + 1)$ consisting of the observations with the $m + 1$ smallest absolute values of $e_i(m)$. To start we take $m_0 = p$ and search over subsets of p observations to find the subset that yields the least median of squares (LMS, Rousseeuw, 1984) estimate of β . However, this initial estimator is not important, provided masking of outliers is broken.

To test for outliers the deletion residual is calculated for the $n - m$ observations not in $S^*(m)$. These residuals, which form the maximum likelihood tests for the outlyingness of individual observations, are

$$r_i(m) = \frac{y_i - x_i^T \hat{\beta}(m)}{\sqrt{s^2(m)\{1 + h_i(m)\}}} = \frac{e_i(m)}{\sqrt{s^2(m)\{1 + h_i(m)\}}}, \quad (2)$$

where the leverage $h_i(m) = x_i^T \{X(m)^T X(m)\}^{-1} x_i$. Let the observation nearest to those forming $S^*(m)$ be i_{\min} where

$$i_{\min} = \arg \min_{i \notin S^*(m)} |r_i(m)|.$$

To test whether observation i_{\min} is an outlier we use the absolute value of the minimum deletion residual, namely $|r_{i_{\min}}(m)|$, as a test statistic. If the absolute value is too large, the observation i_{\min} is considered to be an outlier, as well as all other observations not in $S^*(m)$.

In **S3** we use diagnostic plots of the evolution of all $r_i(m)$ with m in order to reveal the structure of the data. In **S4** we contrast this diagnostic approach with formal testing for outliers, for which we need a reference distribution for $r_i(m)$ in (2). If we estimated σ^2 from all n observations, the statistics would have a t distribution on $n - p$ degrees of freedom. However, in the search we select the central m out of n observations to provide the estimate $s^2(m)$, so that the variability is underestimated. To allow for estimation from this truncated distribution, let the variance of the symmetrically truncated normal distribution containing the central m/n

portion of the full distribution be $\sigma_T^2(m)$. See Riani *et al.* (2009) for a derivation from the general method of Tallis (1963). We take as our approximately unbiased estimate of variance $s_T^2 = s^2(m)/\sigma_T^2 = s^2(m)/c(m, n)$. In the robustness literature $c(m, n)$ is called a consistency factor (Maronna *et al.*, 2006).

For each m , the distribution of the minimum deletion residual $|r_{imin}(m)|$ can be found by repeated simulation of samples of size n . However, in the formal approach to testing for outliers of **S4** we repeatedly superimpose envelopes for varying values of n to establish the number of outliers. For this we use the order-statistics arguments of Riani *et al.* (2009).

3 The Random Start Forward Search for Regression Data

The international trade data record the transaction value and amount of imports of individual goods into the EU. For any individual supplier there should be a straight line relationship between value and quantity, although the relationship may be different for different suppliers. There may also be numerous outliers due to misrecording of the values of the two variables, or due to erroneous coding of goods. Interest is in detecting price-quantity relationships that are consistently anomalous; these may indicate money laundering or tax fraud. We analyse 180 observations with such a structure.

Random start forward searches have been used as a diagnostic tool to indicate the number of clusters in multivariate data and to suggest cluster membership. The analysis of Atkinson *et al.* (2004, **S3.4**) of the Swiss banknote data of Flury (1997), in which there are two clusters and some outliers, shows that the structure revealed by the search depends on whether the search starts in one or other of the clusters, or with some units from both. For clustering such data, Atkinson and Riani (2007) suggest running several hundred searches from random starting points. Many of the forward searches are attracted to clusters in the data and the structure is revealed.

We now exemplify this idea for the analysis of regression data. To run random start forward searches we select 500 random subsets of $m_0 = 2$ observations, and run a forward search from each. Although the search moves forward by incrementing the value of m , the new subset is chosen by ranking all n residuals using parameter estimates from the previous subset. Thus observations that become outlying can be dropped from the subset, which is attracted towards central observations from whichever is the nearest model. Such a process continues until all observations near to the particular model have been used in fitting. Then outliers, or observations close to other regressions, are included in the subset, and the parameter estimates may change appreciably. Once two random starting points have converged to the same subset $S^*(m)$, for some m , the searches cannot diverge again. Thus a forward plot of the minimum deletion residuals shows trajectories that converge to a few potentially informative curves. In Figure 1, from around $m = 100$, the forward plots of the minimum deletion residuals are reduced to only two trajectories. We now interrogate the plot to find out what structure is being revealed.

The darker trajectory in the figure lies above the 99% pointwise envelope from around $m = 130$ and continues to increase until $m = 133$. The lower left-hand panel of Figure 2 shows, by circles, the 133 units that are included in the subset at this point, which clearly form one line. The other units equally clearly fall on a second line; there are no outliers from these two structures. The second trajectory in Figure 1 goes outside the upper envelope slightly earlier. The bottom right-hand panel of Figure 2 shows a scatterplot of the units when this branch of

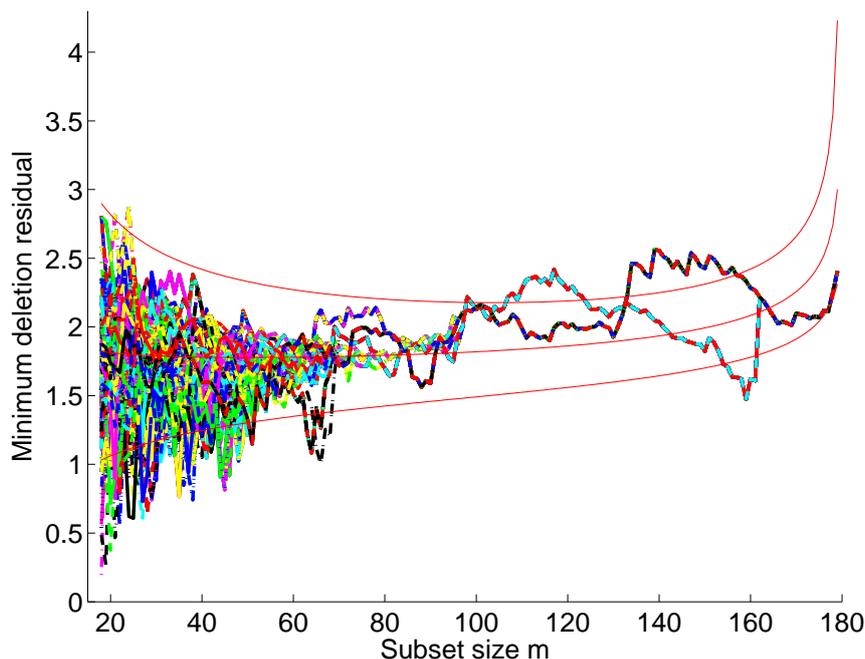


Figure 1: Trade data: forward plots of minimum deletion residuals from 500 random starts with pointwise 1% and 99% limits. There appear to be two distinct groups (regression lines)

the search is interrogated at $m = 107$. The structure of two lines is again revealed; as well as units not in the subset lying below those in $S^*(m)$ for larger values of x , there are also four units above those in the subset for the lowest values of x .

The curves in the upper panels of Figure 2 show the 500 trajectories divided into those that give one of the two peaks and those that give the other. We select one initial subset from each panel and follow the residuals generated during these searches. The left-hand panel of Figure 3 shows the scaled residuals for all units; those that are included in the subset at $m = 133$ are shaded grey and plotted with broken lines, whereas the remaining units are plotted with a continuous black line. The units included in the subset have, for most of the search, residuals that approximately lie between -1 and 1 . The residuals plotted in black all have positive values, as we would expect from the lower left-hand panel of Figure 2, with values between 1 and 4 for much of the search. The two groups are quite distinct until around $m = 150$ when the increasing presence of the observations from the upper line starts to influence the fitted slope. The effect of merging of the two lines shows more dramatically in the right-hand panel of the figure where, now, the residuals from observations not in the subset at $m = 107$ mostly lie below the majority. However, the four observations for low x have the highest residuals. In this plot the dramatic change for m between 158 and 162 comes from the interchange of units between those in the subset and those not (in these five steps 20 new units join the subset and 15 leave it). The fits in the left-hand and right-hand panels of the figure are now identical, although the vertical scales of the panels are different. The identity of the fits is also clear at the ends of the searches in the two upper panels of Figure 2, where the upwards jump in the right-hand panel signals the interchange.

Our analysis indicates that the data fall on two straight lines. Out of 180 observations, we

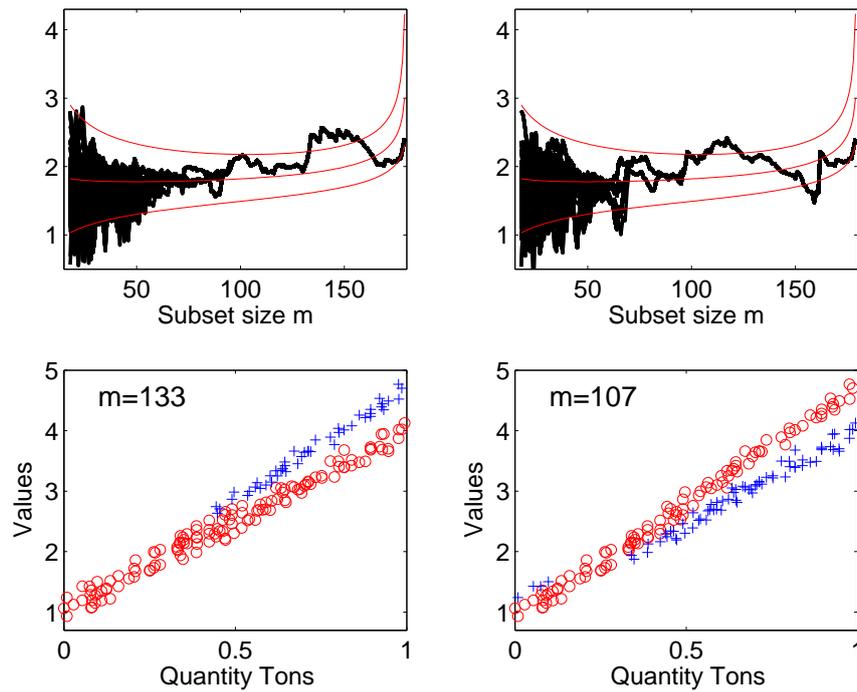


Figure 2: Trade data divided according to the two peaks in Figure 1. Upper panels, forward plots of minimum deletion residuals. Lower panels, scatterplots of observations; \circ observations included in the subset $S^*(m)$

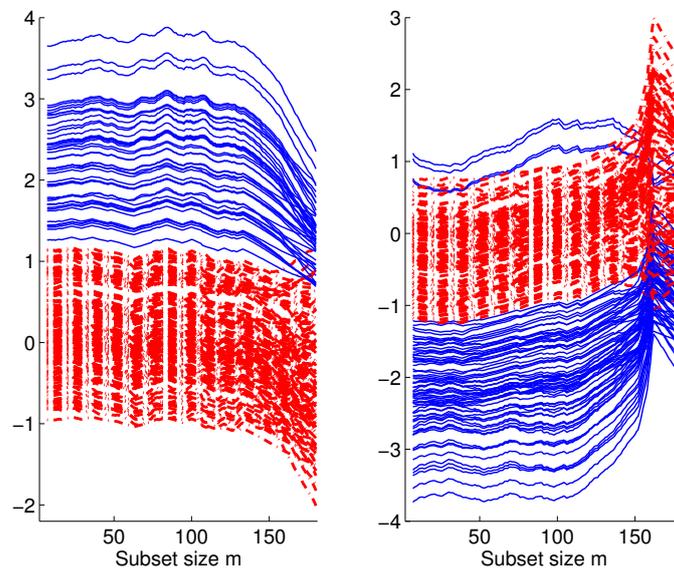


Figure 3: Trade data divided according to the two peaks in Figure 1: forward plots of scaled residuals. Black lines (blue in the .pdf); observations not in the subsets, plotted as + in Figure 2

have 133 that seem to lie on one line and 107 that may well lie on the other. These figures are a reminder that it is impossible to classify with any certainty those units that lie where the lines overlap. However, for the analysis of the trade data, the diagnostic evidence of the existence of two lines is the important outcome. The next stage is to return to the data and to identify the units according to country of origin and importer.

4 Robust Analyses

It is very well-known that multiple outliers in regression models may mask each other, so that they are not revealed by a least squares fit. A robust fit, using a single model, is needed. We now investigate numerically how well such robust fits perform for the trade data which arise from a mixture of regression models.

To calculate the confidence level for the observed value of $|r_{\min}(m)|$ in the FS we used the results of Riani *et al.* (2009) to obtain the confidence level γ as

$$\gamma = 1 - F_{2(n-m), 2(m+1)} \left\{ (m+1) \left[\frac{1}{2T_{m-p}\{r_{\min}(m)\sigma_T(m)\}} - 1 \right] \frac{1}{n-m} \right\}, \quad (3)$$

for $m = m_0, m_0 + 1, \dots, n - 1$. Here F and T are the c.d.f.s of the F and T distributions. As the envelopes in Figures 1 and 2 show, there is appreciable curvature in the plots as $m \rightarrow n$; the envelopes increase rapidly, as, in the absence of outliers, large residuals occur at the end of the search. To clarify visual presentation we now introduce a pointwise normal-score transformation of the envelopes, and of the observed distances, in order to give plots with horizontal envelopes. The plot in normal coordinates uses $\Phi^{-1}(\gamma)$, which, of course, does not change the rule.

To avoid the problem of multiple testing (one outlier test for each value of m) we adapt the rule of Riani *et al.* (2009) for multivariate data to obtain a procedure with a samplewise size of around 1%, replacing Mahalanobis distances by the absolute value of the minimum deletion residual. Now the FS starts from a single, carefully chosen subset using LMS. Outlier detection follows a two-stage procedure. In the first we use envelopes for all n observations. If outliers are present we receive a signal of an outlier at some value m^\dagger . Succeeding observations may be outliers. However, the envelopes depend on the value of n . If we reject some observations as outliers we need new envelopes for a smaller value of n , in general n^* . In the second stage of the procedure we superimpose envelopes for values of n from this point until the first time we introduce an observation we recognise as an outlier.

In the trade data the procedure from a robust starting subset yields a forward plot of values of $|r_{\min}(m)|$ which is of the form of those in the upper left-hand panel of Figure 2. The rule yields a signal at $m^\dagger = 135$ because of three successive values above the 99.99% threshold (the values 133 and 107 in Figure 2 were chosen as the first of such three consecutive values for searches leading to the respective peaks). We then superimpose envelopes for a series of values of n^* . Figure 4 illustrates this point for $n^* = 138, 148, 156$ and 158 . For the two lower values of n^* the plot of residuals lies below the 95% envelope. That when $n^* = 156$ lies within the 99% envelope (as does that for $n^* = 157$ which is not shown). However when $n^* = 158$ one value of the statistic lies above the 99.9% envelope. The conclusion is that there are 157 observations, out of 180, that can be used for estimating the regression model by least squares. There are therefore 23 outliers.

We now compare this analysis of the trade data with that obtained using the Least Trimmed Squares (LTS) and reweighted LTS (LTSr) algorithms described by Verboven and Hubert (2005).

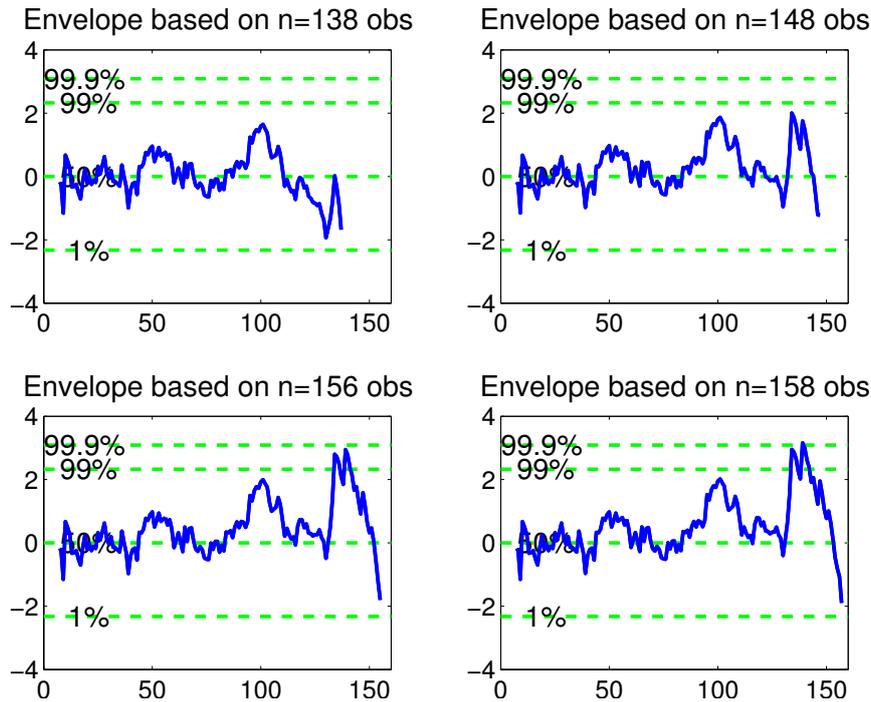


Figure 4: Trade data: resuperimposed envelopes for steps 138, 148, 156 and 158. In this process the first outlier is detected at $n^* = 158$, so there are 157 observations available for estimation

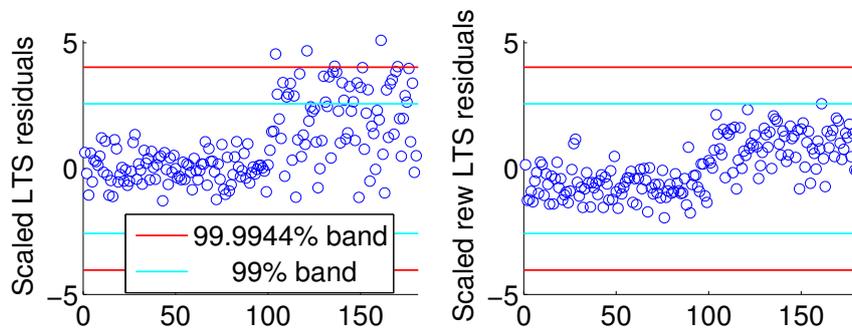


Figure 5: Trade data: left-hand panel, scaled residuals from LTS, right-hand panel scaled residuals from reweighted LTS. Inner bands, pointwise 99% region, outer bands, sample-wise 99% region. There is no indication of the structure evident in Figure 3

Because the FS algorithm is designed to have size α of declaring an outlier free sample to contain at least one outlier, we use a Bonferroni correction for simultaneity when identifying the outliers found by these methods. In LTSr the outliers identified by LTS are removed from the fit and the parameters re-estimated from the remaining observations. We also use the Bonferroni correction to identify the outliers in this intermediate step.

The left-hand panel of Figure 5 shows a plot of the scaled LTS residuals against observation number. In interpreting these plots it is important that the observations are numbered consecutively for one importer and then the other. The structure of different variances, particularly

evident in the left-hand panel, disappears when the observations are permuted.

The inner bands in the plot provide pointwise 99% tests for outliers. There are a large number of outliers at this pointwise level whereas we might expect two if the data followed a single regression model. However the outer, Bonferroni adjusted, bands indicate only three outliers and so no suggestion of the two lines that are the structure of the data. The plot of the scaled residuals from LTSr in the right-hand panel shows no evidence of any outliers at either level.

The further analysis of this section illustrates that robust methods designed to fit a single model whilst detecting outliers may not be effective in detecting departures from a single model if two or more models are present. In the random start forward searches of **S3** our diagnostic procedure detected linear fits with 47 or 73 observations excluded. The statistically controlled fitting of a single model using the FS, on the other hand, only revealed 23 observations as not coming from one of the models. Although scatterplots like those in Figure 2 would reveal a pattern of outliers from which the existence of two lines can be inferred, we need a procedure which, like the random start FS, reveals the presence of alternative models. The second conclusion is that use of very robust methods, such as LTS, designed to reveal up to 50% of outliers in the data, can fail if the major model and that for the outliers are close together. Several different fits to the data, combined with diagnostic plotting of residuals, provide a surer way of detecting data from mixtures of regression models.

Bibliography

- [1] Atkinson, A. C. and Riani, M. (2007). Exploratory tools for clustering multivariate data. *Computational Statistics and Data Analysis*, 52, 272–285. doi:10.1016/j.csda.2006.12.034.
- [2] Atkinson, A. C., Riani, M., and Cerioli, A. (2004). *Exploring Multivariate Data with the Forward Search*. Springer–Verlag, New York.
- [3] Flury, B. (1997). *A First Course in Multivariate Statistics*. Springer-Verlag, New York.
- [4] Maronna, R. A., Martin, R. D., and Yohai, V. J. (2006). *Robust Statistics: Theory and Methods*. Wiley, Chichester.
- [5] Riani, M., Atkinson, A. C., and Cerioli, A. (2009). Finding an unknown number of multivariate outliers. *Journal of the Royal Statistical Society, Series B*, 71, 447–466.
- [6] Rousseeuw, P. J. (1984). Least median of squares regression. *Journal of the American Statistical Association*, 79, 871–880.
- [7] Tallis, G. M. (1963). Elliptical and radial truncation in normal samples. *Annals of Mathematical Statistics*, 34, 940–944.
- [8] Verboven, S. and Hubert, M. (2005). LIBRA: a MATLAB library for robust analysis. *Chemometrics and Intelligent Laboratory Systems*, 75, 127–136. doi:10.1016/j.chemolab.2004.06.003.

Incomplete longitudinal binary responses in marginal model

M. Helena Gonçalves, *CEAUL and FCT, Universidade do Algarve, Portugal*, mhgoncal@ualg.pt

M. Salomé Cabral, *CEAUL and DEIO, Faculdade de Ciências da Universidade de Lisboa, Portugal*, salome@fc.ul.pt

Abstract. In the analysis of binary longitudinal data a frequent problem is the presence of missing data since it is difficult to have complete records of all individuals. Another feature in these studies is to take into account the autocorrelation structure present in successive observations, taken over time on each individual and associated with a certain response variable. In this paper we discuss the performance of the marginal models implemented in the R package `bird` when missing values are present in data provided that they are missing at random (MAR). In those marginal models inference is based on likelihood approach and serial dependence is regulated by a binary Markov chain mechanism. A simulation study is carried out and a real data set is also used to illustrate that behaviour.

Keywords. binary longitudinal data, marginal model, exact likelihood, Markov chain, missing data.

1 Introduction

Longitudinal binary data studies are a powerful design and they have become increasingly popular in a wide range of applications in clinical research. In these studies repeated observations of a response variable are taken over time on each subject in one or more treatment groups. In such cases the repeated measures of each vector of responses are likely to be correlated and the autocorrelation structure for the repeated data plays a significant role in the estimation of regression parameters. Although most longitudinal studies are designed to collect data on every subject in the sample at each time of follow-up, many studies have missing data since it is difficult to have complete records of all subjects for a wide variety of reasons. When longitudinal binary data are incomplete, there are important implications for their analysis and several methods have been proposed [1, 2, 7, 8]. A review of this topic is given by [6].

In the context of marginal model to binary longitudinal data, [3] proposed a methodology based on likelihood approach and used a binary Markov chain model to accommodate serial

dependence and odds-ratio to measure dependence between successive observations in the same individual. This methodology has been implemented in the R package `bird` [4, 5] and allows missing values on the response, provided they are missing at random (MAR) in the standard terminology of [9].

The goal of this paper is to study the performance of that methodology for analysing incomplete binary longitudinal responses. A simulation study is presented where complete and missing data are both considered. Data from the Muscatine Coronary Risk Factor [10] is analysed to illustrate the objective of this paper. In Section 2 we give a summary of the approach used. In Section 3 we report a small simulation study to examine the performances of the procedure. Complete and incomplete data cases are considered. In Section 4 we present the results of applying the approach to the aforementioned real data set. Finally, in Section 5 we draw some overall conclusions.

2 Models for binary data

Suppose that n independent individuals are observed at times $t = 1, \dots, T_i$, which need not be the same for all n individual, and denote by $y_{it} \in \{0, 1\}$ the binary response value at time t from individual i ($i = 1, \dots, n$), and by Y_{it} its generating random variable whose mean value is $\Pr(Y_{it} = 1) = \theta_{it}$. Associated with each observation time and each subject, a set of p covariates is available, denoted by x_{it} and β as the p -vector of unknown parameters. We shall refer collectively to the sequence (y_{i1}, \dots, y_{it}) as the i th individual profile.

A logistic regression model is assumed for the marginal mean of Y_{it} and the probability of success is

$$\text{logit } \theta_{it} = x_{it}^\top \beta. \quad (1)$$

For the first order Markov chain (MC1), the serial dependence is modeled using $\psi_1 = OR(Y_t, Y_{t-1})$ where

$$OR(Y_t, Y_{t-1}) = \frac{\Pr(Y_{t-1} = Y_t = 1) \Pr(Y_{t-1} = Y_t = 0)}{\Pr(Y_{t-1} = 0, Y_t = 1) \Pr(Y_{t-1} = 1, Y_t = 0)} = \frac{p_1/(1-p_1)}{p_0/(1-p_0)}$$

where p_j are the one-step transition probabilities given by

$$p_j = \Pr(Y_t = 1 | Y_{t-1} = j), \quad j = 0, 1. \quad (2)$$

For the second order Markov chain (MC2) the joint distribution of three successive components of the process at time, (Y_{t-2}, Y_{t-1}, Y_t) , is considered and the constraints,

$$\begin{aligned} OR(Y_{t-1}, Y_{t-2}) &= \psi_1 = OR(Y_{t-1}, Y_t) \\ OR(Y_{t-2}, Y_t | Y_{t-1} = 0) &= \psi_2 = OR(Y_{t-2}, Y_t | Y_{t-1} = 1) \end{aligned}$$

are imposed, with ψ_1 and ψ_2 two positive parameters. The two-steps transition probabilities are given by

$$p_{hj} = \Pr(Y_t = 1 | Y_{t-2} = h, Y_{t-1} = j), \quad h, j = 0, 1, \quad (3)$$

see [3] for a full account.

The serial dependence for MC2 models is regulated by $\lambda = (\lambda_1, \lambda_2) = (\log \psi_1, \log \psi_2)$, which are assumed to be constant across time and subjects. When, $\lambda_2 = 0$, the Markov chain reduces to MC1 models and the serial dependence is regulated by λ_1 .

The estimation of parameters is based on the likelihood approach. The contribution from a generic individual to the log-likelihood for the parameters (β, λ) is under MC1 model,

$$\ell_i(\beta, \lambda) = y_1 \text{logit}(\theta_1) + \log(1 - \theta_1) + \sum_{t=2}^{T_i} [y_t \text{logit}(p_j) + \log(1 - p_j)] \quad (4)$$

and under MC2 model,

$$\ell_i(\beta, \lambda) = [y_1 \text{logit}(\theta_1) + \log(1 - \theta_1)] + [y_2 \text{logit}(p_j) + \log(1 - p_j)] + \sum_{t=3}^{T_i} [y_t \text{logit}(p_{hj}) + \log(1 - p_{hj})] \quad (5)$$

where the three blocks on the right-hand side represent the contribution to the log-likelihood from y_1 , y_2 , and (y_3, \dots, y_T) , respectively, where p_{hj} is given by (3) and p_j by (2).

For both models, MC1 and MC2, the overall log-likelihood functions are obtained as the sum of the n logarithmic individual contributions given by (4) and (5), respectively. Numerical maximisation of the log-likelihood is required and the derivatives of the functions are supplied to improve the efficiency of the optimisation algorithms. Given the algebraic work required to obtain explicit expressions of the gradient of the log-likelihood it is completely unfeasible to develop analogous results for the Hessian matrix. Therefore, the observed information matrix for (β, λ) must be computed via numerical differentiation of the first derivatives. For a full account see [3].

Missing data

In this approach missing values are allowed on the response, provided they are MAR. If missing data occur at the beginning or at the end of an individual profile, this poses no problems, since this case is equivalent to an unbalanced design in the length profile T_i for that individual. Some restrictions exist for the presence of missing data when they occur in the middle of the profile. The precise description of the missingness patterns follows next, where p_j in (2) will be denoted here by $p_{t:j}$ and p_{hj} in (3) by $p_{t:hj}$.

If MC1 model is considered and we have a missing value at time point $t - 1$ it is required that there are observations at time points $t - 2$ and t . Expression (4) is modified as follows: y_{t-1} is replaced by y_t and p_j is replaced by $p_{t:j} = \Pr(Y_t = 1 | Y_{t-2} = j) = (1 - p_{t-1:j}) p_{t:0} + p_{t-1:j} p_{t:1}$.

If MC2 model is considered and we have a missing value at time point $t - 2$, it is required that there are observations at time points $t - 4, t - 3, t - 1$ and t . The modifications in (5) are:

1. If y_2 is missing, and there are observations at the two adjacent time points (y_3 and y_4), then (5) is modified as follows: y_2 is replaced by y_3 ; p_j is replaced by $\Pr(Y_3 = 1 | Y_1 = j) = (1 - p_{2:j}) p_{3:0} + p_{2:j} p_{3:1}$ using the one-step transition probabilities $p_{t:j}$ as used in (4); the contribution from Y_4 is obtained from $\Pr(Y_4 = 1 | Y_1 = h, Y_3 = j) = (1 - p_{2:h}) p_{4:0j} + p_{2:h} p_{4:1j}$.
2. If y_t is observed and y_{t-1} is missing, the last term of type p_{hj} in (5) is replaced by $\Pr(Y_t = 1 | Y_{t-3} = h, Y_{t-2} = j) = (1 - p_{t-1:hj}) p_{t:j0} + p_{t-1:hj} p_{t:j1}$.
3. The more common case refers to a missing datum in the middle of the observation period. Let us say that the missing value is at time point $t - 2$ and that there are observations at

time points $t, t-1, t-3, t-4$. The joint contribution from (y_{t-1}, y_t) is

$$\begin{aligned} \Pr(Y_{t-1} = r, Y_t = 1 | Y_{t-4} = h, Y_{t-3} = j) &= \\ &= (1 - p_{t-2:hj}) \left[(1 - p_{t-1:j0}) + (2p_{t-1:j0} - 1)r \right] p_{t:0r} \\ &\quad + p_{t-2:hj} \left[(1 - p_{t-1:j1}) + (2p_{t-1:j1} - 1)r \right] p_{t:1r}. \end{aligned}$$

This approach is implemented in the R package `bild` [4].

3 A simulation study

A simulation study was conducted when we had a serial dependence MC1 or MC2 with the aim to examine the impact of intermittent missingness status in the estimation parameters in terms of relative bias and mean square error.

In the simulation, we have considered the following model

$$\Pr(Y_{it} = 1 | t) = \frac{\exp(\beta_0 + \beta_1 t)}{1 + \exp(\beta_0 + \beta_1 t)} \quad (6)$$

where the fixed effect coefficients were set at $\beta_0 = -1$ and $\beta_1 = 0.5$. Each data set contains $I = 50$ subjects of size $T = 13$, with $t = -1.5, -1.25, -1, -0.75, -0.5, -0.25, 0, 0.25, 0.5, 0.75, 1, 1.25, 1.5$.

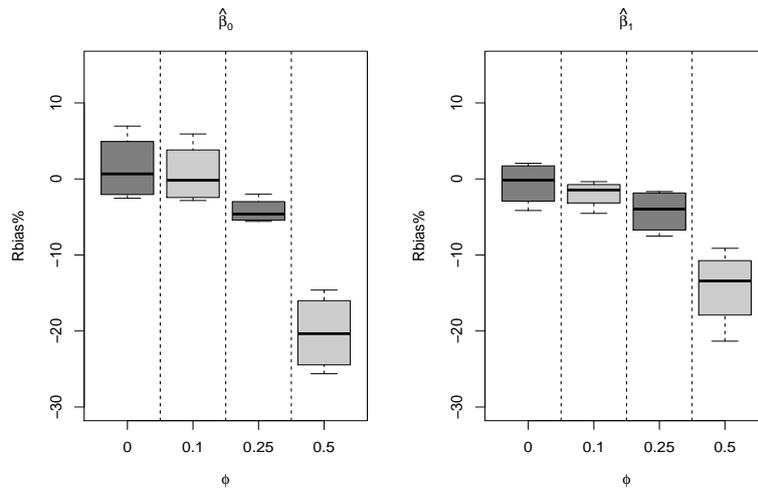
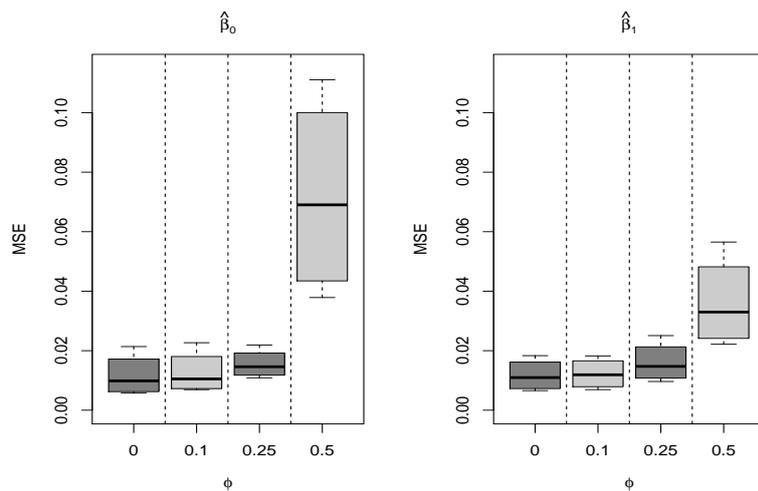
On each run we have generated T binary correlated data under the i th subject following a first order serial dependence with constant λ_1 or a second order serial dependence with constants (λ_1, λ_2) . Under MC1 we considered for λ_1 the values -2, -1, 1 and 2. Under MC2 we have considered for the pair (λ_1, λ_2) the combinations (-1,-1), (-1,1), (1,-1), (1,1).

In both situations an intermittent missing-data mechanism MAR was considered, taking into account the missingness restrictions described in Section 2. In this mechanism it is assumed that the binary response on the first occasion is always observed, $R_{i1} = 1$, here $\mathbf{R}_i = 1$ denote a $T \times 1$ vector of indicator variables for the i th subject, where $R_{it} = 1$ if Y_{it} is observed, and $R_{it} = 0$ if Y_{it} is missing. The binary response for the i th subject at time t (R_{it}) is generated with probability of success given by $(1 - \phi)^{1 - y_{it-1}}$, where ϕ is the nonresponse parameter [2]. To each serial dependence the missing-data mechanism was applied with $\phi = 0, 0.1, 0.25, 0.5$ ($\phi = 0$ corresponding to complete data). The whole estimation procedure was repeated for 1000 runs and the sample mean of the estimates (Mean), the sample mean of percent relative bias (Rbias%) and the sample mean square error (MSE) were computed. The estimates of the parameters were obtained through the function `bild` in the R package `bild` [4].

The results of our simulation are displayed from Tables 1-2. Each table lists the following: Mean, Rbias% and MSE over the 1000 simulations runs. From Figures 1-4 we present the boxplots of Rbias% and MSE as a summary of those results.

Taking into account that our goal is to study the performance of the methodology with intermittent missing data, the main conclusions of our simulation can be summarize as follows:

1. To the first group of data generated from a MC1 serial dependence (Table 1 and Figures 1 and 2) the three main conclusions are: (i) the MSE of both β_0 and β_1 is small for values of ϕ until 0.25 but when ϕ equal 0.5 the MSE of β_0 is greater than the MSE of β_1 ; (ii) the Rbias% of β_0 is greater than the Rbias% of β_1 to all situations considered; (iii) the Rbias% of both β_0 and β_1 increase with larger values of ϕ .

Figure 1: Rbias% of $\hat{\beta}_0$ and $\hat{\beta}_1$ for several λ_1 values and MC1 model.Figure 2: MSE of $\hat{\beta}_0$ and $\hat{\beta}_1$ for several λ_1 values and MC1 model.

2. To the second group of simulations generated from a MC2 serial dependence (Table 2 and Figures 3 and 4) the two main conclusions are: (i) the MSE of both β_0 and β_1 is very small for all values of ϕ ; (ii) the Rbias% of β_1 is greater than the Rbias% of β_0 .
3. When we compare the simulations for the two groups, the behaviour of Rbias% and MSE is very similar for all parameters when we have complete data cases ($\phi = 0$). When missing data is present the impact of intermittent missingness is smaller, both in terms of Rbias% and MSE, under the MC2 serial dependence and for all parameters.

Based on the previous conclusions we may say that the results of the simulation study suggest that the methodology implemented in R package `bild` has a satisfactory degree of robustness to

	λ_1	ϕ	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\lambda}_1$
Mean	-2	0.0	-0.975	0.510	-1.752
Rbias%			-2.537	2.063	-12.404
MSE			0.006	0.007	0.147
Mean		0.1	-0.972	0.505	-1.754
Rbias%			-2.838	1.123	-12.314
MSE			0.007	0.007	0.169
Mean		0.25	-0.944	0.492	-1.762
Rbias%			-5.584	1.658	-11.895
MSE			0.011	0.010	0.183
Mean		0.5	-0.854	0.454	-1.778
Rbias%			-14.616	-9.124	-11.125
MSE			0.038	0.022	0.365
Mean	-1	0.0	-0.984	0.507	-0.923
Rbias%			-1.561	1.378	-7.529
MSE			0.007	0.008	0.063
Mean		0.1	-0.979	0.498	-0.904
Rbias%			-2.051	-0.361	-9.564
MSE			0.008	0.009	0.073
Mean		0.25	-0.947	0.490	-0.945
Rbias%			-5.258	-2.026	-5.455
MSE			0.012	0.012	0.090
Mean		0.5	-0.826	0.438	-0.957
Rbias%			-17.408	-12.366	-4.261
MSE			0.050	0.026	0.184
Mean	1	0.0	-1.029	0.492	0.910
Rbias%			2.887	-1.686	-8.960
MSE			0.013	0.014	0.050
Mean		0.1	-1.017	0.491	0.898
Rbias%			1.706	-1.790	-10.194
MSE			0.013	0.015	0.055
Mean		0.25	-0.960	0.470	0.878
Rbias%			-3.974	-5.906	-12.157
MSE			0.016	0.017	0.074
Mean		0.5	-0.767	0.428	0.783
Rbias%			-23.293	-14.472	-21.748
MSE			0.089	0.040	0.160
Mean	2	0.0	-1.070	0.479	1.831
Rbias%			6.952	-4.150	-8.444
MSE			0.021	0.018	0.078
Mean		0.1	-1.059	0.477	1.827
Rbias%			5.916	-4.530	-8.664
MSE			0.023	0.018	0.085
Mean		0.25	-0.980	0.462	1.797
Rbias%			-1.988	-7.518	-10.139
MSE			0.022	0.025	0.104
Mean		0.5	-0.744	0.393	1.693
Rbias%			-25.625	-21.335	-15.367
MSE			0.111	0.057	0.216

Table 1: Results of the simulation study for $\lambda_1 = -2, -1, 1$ and 2 .

intermittent missing data status.

4 An illustrative example

To illustrate the results we have used a subset of data from the Muscatine Coronary Risk Factor Study, a longitudinal study of coronary risk factors in school children from Muscatine (Iowa, USA) [10]. The dataset contains records on 1014 children who were 7-9 years old in 1977 and were examined in 1977, 1979 and 1981. The binary response of interest is whether the child is obese (1) or not (0). Since one of the objectives of the study was to determine the effects of sex

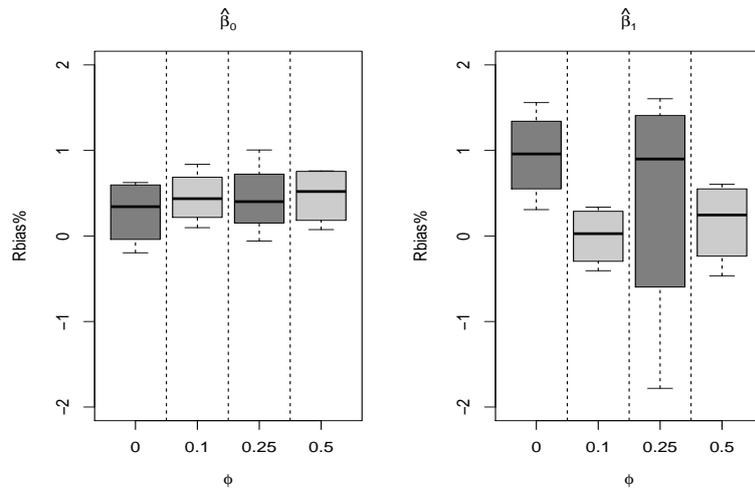


Figure 3: Rbias% of $\hat{\beta}_0$ and $\hat{\beta}_1$ for several (λ_1, λ_2) values and the MC2 model.

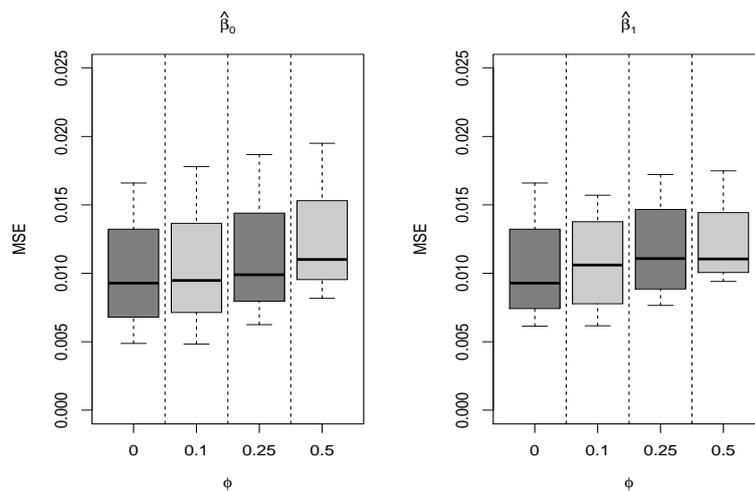


Figure 4: MSE of $\hat{\beta}_0$ and $\hat{\beta}_1$ for several (λ_1, λ_2) values and the MC2 model.

and age on risk of obesity a marginal model is appropriate. Many data records are incomplete, since not all children have participated in all the surveys, creating, as [1] said, a "genuine" missing data problem. We have considered these data, available in the R package `bild` [4], as an illustrative example to an easy comparison with findings of other authors [2, 1, 8] when there are missing values.

For comparison with the results of [2] and [1] we have fitted to data the same three models for the marginal probability of the event, namely:

Model I: $\text{logit}(\theta_{it}) = \beta_0 + \beta_1 G + \beta_2 A(L) + \beta_3 A(Q) + \beta_4 GA(L) + \beta_5 GA(Q)$

Model II: $\text{logit}(\theta_{it}) = \beta_0 + \beta_1 G + \beta_2 A(L) + \beta_3 A(Q)$

	(λ_1, λ_2)	ϕ	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\lambda}_1$	$\hat{\lambda}_2$
Mean	(-1, -1)	0.0	-0.998	0.502	-1.022	-1.053
Rbias%			-0.196	0.308	2.225	5.300
MSE			0.005	0.006	0.047	0.066
Mean		0.1	-1.003	0.498	-1.029	-1.038
Rbias%			0.337	-0.406	2.932	3.858
MSE			0.005	0.006	0.057	0.079
Mean		0.25	-1.004	0.503	-1.025	-1.035
Rbias%			0.362	0.587	2.517	3.530
MSE			0.006	0.008	0.073	0.095
Mean		0.5	-1.003	0.503	-1.024	-1.088
Rbias%			0.291	0.604	2.360	8.836
MSE			0.008	0.009	0.112	0.149
Mean	(-1, 1)	0.0	-1.006	0.508	-1.034	0.970
Rbias%			0.569	1.559	3.397	-3.036
MSE			0.009	0.010	0.089	0.047
Mean		0.1	-1.001	0.502	-1.043	0.982
Rbias%			0.097	0.337	4.270	-1.787
MSE			0.010	0.009	0.108	0.055
Mean		0.25	-1.010	0.508	-1.044	0.998
Rbias%			1.005	1.604	4.399	-0.249
MSE			0.010	0.010	0.141	0.065
Mean		0.5	-1.007	0.502	-1.052	0.950
Rbias%			0.074	0.497	5.203	-4.970
MSE			0.011	0.011	0.166	0.093
Mean	(1, -1)	0.0	-1.006	0.506	0.990	-1.044
Rbias%			0.625	1.124	-1.038	4.420
MSE			0.010	0.011	0.028	0.076
Mean		0.1	-1.008	0.502	0.981	-1.037
Rbias%			0.837	0.239	-1.870	3.727
MSE			0.009	0.012	0.035	0.085
Mean		0.25	-0.999	0.491	0.984	-1.053
Rbias%			-0.059	-1.781	-1.638	5.316
MSE			0.010	0.012	0.046	0.098
Mean		0.5	-1.007	0.500	0.966	-1.035
Rbias%			0.761	-0.006	-3.442	3.545
MSE			0.011	0.011	0.072	0.126
Mean	(1, 1)	0.0	-1.001	0.504	0.991	0.986
Rbias%			0.116	0.794	-0.865	-1.439
MSE			0.017	0.015	0.060	0.049
Mean		0.1	-1.005	0.499	0.972	0.958
Rbias%			0.536	-0.186	-2.796	-3.126
MSE			0.018	0.016	0.080	0.054
Mean		0.25	-1.004	0.506	0.966	0.958
Rbias%			0.441	1.212	-3.384	-4.197
MSE			0.019	0.017	0.096	0.071
Mean		0.5	-1.007	0.498	0.970	0.967
Rbias%			0.749	-0.466	-2.977	-3.332
MSE			0.020	0.017	0.134	0.092

Table 2: Results of the simulation study for several (λ_1, λ_2) values.

Model III: $\text{logit}(\theta_{it}) = \beta_0 + \beta_1 A(L) + \beta_2 A(Q)$

where G indicates gender (female=1, male=0) and $A(L)$, $A(Q)$ are orthogonal polynomial contrasts for linear and quadratic component of age effect, respectively. A fourth and simpler model was also fitted to data:

Model IV: $\text{logit}(\theta_{it}) = \beta_0 + \beta_1 A(L)$

In all four models a serial dependence MC2 has been considered, instead of the MC1 serial dependence used by [1]. The analysis was performed using the `bild` function in the R package `bild`.

For all the models fitted to data, the estimated values of the parameters, as well as their standard errors, t-ratio and corresponding p-values are given in Table 3. The estimates of the regression parameters and the corresponding standard errors are in close agreement with those of [2].

Model	LogL	Parameter	Estimate	SE	t-ratio	<i>p</i> - value
I	-949.8621	β_0	-1.346	0.097	-13.929	≈ 0
		β_1	0.042	0.138	0.303	0.762
		β_2	0.126	0.066	1.918	0.055
		β_3	0.020	0.035	0.570	0.568
		β_4	0.175	0.095	1.852	0.064
		β_5	-0.092	0.049	-1.899	0.058
		λ_1	3.146	0.200	15.728	≈ 0
		λ_2	1.900	0.329	5.770	≈ 0
II	-953.0435	β_0	1.360	0.097	-14.057	≈ 0
		β_1	0.069	0.137	0.507	0.612
		β_2	0.214	0.047	4.515	≈ 0
		β_3	-0.027	0.024	-1.129	0.259
		λ_1	3.104	0.197	15.722	≈ 0
		λ_2	1.867	0.324	5.756	≈ 0
III	-953.1726	β_0	1.326	0.069	-19.291	≈ 0
		β_1	0.214	0.047	4.511	≈ 0
		β_2	-0.027	0.024	-1.112	0.266
		λ_1	3.105	0.197	15.729	≈ 0
		λ_2	1.861	0.324	5.739	≈ 0
IV	-953.7947	β_0	1.325	0.069	-19.297	≈ 0
		β_1	0.209	0.047	4.483	≈ 0
		λ_1	3.103	0.198	15.710	≈ 0
		λ_2	1.863	0.323	5.760	≈ 0

Table 3: Log-likelihood, Parameters estimates, Standard errors, t-ratio and *p*-value for models I, II, III and IV.

As [2] reported the results of this analysis suggest that there is a linear increase (on the logit scale) in the rate of obesity over time, with no statistically discernible difference between males and females. As effect the decrease in deviance between the models I and IV is $\Delta D = 2 \times (953.7947 - 949.8621) = 7.865$ on four degrees of freedom (*p*-value = 0.09664) and thus the model IV is not rejected at the level of significance 5%.

MC1 serial dependence was used by [1] which leads to the differences between his models and ours. In this case we can ask which serial dependence is more appropriate. Despite of the fact that the several summaries presented in Table 3 point out to a strong correlation of second order, we have fitted to data two models with MC1 serial dependence. The first one (Mode II) with the marginal probability given by Model I and the second one (Mode IV1) with the marginal probability given by Model IV. In the first case the decrease in deviance between the models I and II is $\Delta D = 2 \times (966.5612 - 949.8621) = 33.398$ on one degree of freedom (*p*-value ≈ 0). In the second case the decrease in deviance between the models IV and IV1 is $\Delta D = 2 \times (970.3821 - 953.7947) = 33.175$ on one degree of freedom (*p*-value ≈ 0). And, as expected, in both situations the MC1 serial dependence is rejected at the level of significance 5%.

5 Conclusion

This paper is concerned with the impact of MAR data in binary longitudinal studies when the marginal models described along Section 2 and implemented in the R package `bild` [4] are considered. A simulation study was carried out and we conclude that the approach performs

quite well to intermittent missing data status in both situations of serial dependence (MC1 and MC2), as well as, when complete data sets are considered. Finally, an example using data from Muscatine Coronary Risk Factor Study set was analyzed. This allows us to compare our results with those obtained by [2]. Based on that comparison we can say that this methodology is a suitable alternative to the one presented by those authors.

Acknowledgements

The authors thank the anonymous reviewers for their comments and suggestions and Ivette Gomes for her helpful comments. Research partially sponsored by national funds through the Fundação Nacional para a Ciência e Tecnologia, Portugal-FCT under the project (PEst-OE/MAT/UI0006/2014).

Bibliography

- [1] Azzalini, A. (1994) *Logistic regression for autocorrelated data with application to repeated measures*. *Biometrika*, **81**(4), 767–775.
- [2] Fitzmaurice, G.M., Laird, N.M. and Lipsitz, S.R. (1994) *Analyzing incomplete longitudinal binary responses: A likelihood-based approach*. *Biometrics*, **50**, 601–612.
- [3] Gonçalves, M.H. and Azzalini, A. (2008) *Using Markov chains for marginal modelling of binary longitudinal data in an exact likelihood approach*. *Metron*, **LXVI**, 157–181.
- [4] Gonçalves, M.H., Cabral, M.S. and Azzalini, A. (2012) *bild: A package for BInary Longitudinal Data*. R foundation for statistical computing, version 1.1. url = <http://CRAN.R-project.org/package=bild>.
- [5] Gonçalves, M.H., Cabral, M.S. and Azzalini, A. (2012) *The R package bild for the analysis of binary longitudinal data*. *Journal of statistical software*, **46**(9), 1-17.
- [6] Jansen, I., Beunckens, C., Molenberghs, G., Verbeke, G. and Mallinckrodt, C. (2006) *Analyzing incomplete discrete longitudinal clinical trials data*. *Statistical science*, **21**, 52–69.
- [7] Laird, N.M. (1988) *Missing data in longitudinal studies*. *Statistics in medicine*, **7**, 305–315.
- [8] Lipsitz, S.R., Molenberghs, G., Fitzmaurice, G.M. and Ibrahim, J. (2000) *GEE with Gaussian estimation of the correlations when data are incomplete*. *Biometrics*, **56**, 528–536.
- [9] Little, R.J.A. and Rubin, D.B. (1987) *Statistical analysis with missing data*. John Wiley & Sons, New York.
- [10] Woolson, R.F. and Clarke, W.R. (1984) *Analysis of categorical incomplete longitudinal data*. *Journal of the royal statistical society, Serie A*, **147**, 87–99.

On the modification of the non-parametric test for comparing locations of two populations

Grzegorz Konczak, *University of Economics in Katowice*, grzegorz.konczak@ue.katowice.pl

Abstract. Classical methods for monitoring the average level of the process in quality control procedures are based on the normality assumption. The construction of the well known Shewhart's control charts is based on the sequence of parametric tests. The sample characteristics are compared to the theoretical distribution or to the reference sample taken from the stable process. To do this parametric tests are used. These tests could be used if the population is normally distributed and observations are independent of each other. In the case of non-normal distribution non-parametric tests (for example the Wilcoxon-Mann-Whitney test) can be used. The paper presents a proposal of a modification of the L. Hao and D. Houser adaptive test for comparing the locations of two distributions (HH test). The modification is based on the Hao L. and Houser D. paper (see [2]). In the mentioned paper due to the values of the robust asymmetry and shape characteristics, the test statistic is chosen. In the paper the method of continuous modification of the test statistic is described. The properties of the proposed procedure are analysed in the Monte Carlo study.

Keywords. adaptive test, quality control, process monitoring, non-normal process, permutation tests

1 Introduction

Classical methods for monitoring the average level of the process in quality control procedures are based on the normality assumption. The Shewhart's control charts are based on the sequence of parametric tests (see [9]). The main assumptions in these tests are that the population is normally distributed and observations are independent of each other. In many real-world applications the data are often non-normally distributed. Instead of the parametric tests, the non-parametric methods can be used. The two-phase of nonparametric control charts are presented in [10] and applications of a powerful nonparametric test for heavy-tailed and/or highly-skewed data are presented in [4]. The non-parametric tests often have less power than the parametric

tests. To increase the power of the non-parametric tests, the adaptive procedures can be used. Adaptive tests (see [11]) use the sample data to adjust the test procedure.

The paper presents a proposal of a modification of the L. Hao and D. Houser adaptive test presented in the paper [2]. In the mentioned paper due to the values of the robust asymmetry and shape parameters the form of the test statistic is chosen. In the paper the method of continuous modification of the statistic is described. The properties of the proposed method are analysed in the Monte Carlo study.

The main idea of the adaptive tests is to select the test statistic. The selection in the HH test is based on the asymmetry and kurtosis from the combined samples. The proposed modification is based on changing the weights in the statistic in dependence of the asymmetry and kurtosis of the combined sample. Adaptive tests are very flexible and can be modified in various ways also in contexts other than the location problem, see eg [7] which proposed a modification of an adaptive test for scale which uses Hogg tailweight measure.

Let us consider two samples X_1, X_2, \dots, X_n (the reference sample) and Y_1, Y_2, \dots, Y_m (the sample taken from the monitored process). Let us assume that the samples were taken from distributions $F(x)$ and $F(x + \theta)$ where θ is the shift of the location parameter. The hypothesis that the samples were taken from the same distribution will be considered. Formally, the null hypothesis can be written as follow

$$H_0 : \theta = 0$$

versus the alternative hypothesis

$$H_A : \theta = \delta \neq 0$$

2 Adaptive tests

The adaptive procedure for comparing two distributions has been presented by Hogg et al. in [3]. This procedure is based on calculating the asymmetry and kurtosis of combined samples. For determining the asymmetry (Q_3) and the kurtosis (Q_4) characteristics following robust estimators are used:

$$Q_3 = \frac{\bar{U}_{0.05} - \bar{M}_{0.50}}{\bar{M}_{0.50} - \bar{L}_{0.05}} \quad (1)$$

$$Q_4 = \frac{\bar{U}_{0.05} - \bar{L}_{0.05}}{\bar{U}_{0.50} - \bar{L}_{0.50}} \quad (2)$$

where $\bar{U}_{0.05}$, $\bar{L}_{0.05}$, $\bar{U}_{0.50}$ and $\bar{M}_{0.50}$ are the averages of the upper 5%, lower 5%, upper 50% and middle 50% of data (order statistic of the combined sample).

L. Hao and D. Houser in [2] have presented the modification of the adaptive test procedure, first proposed by Hogg et al in [3]. The test statistic used in this procedure has the following form

$$S = \sum_{i=1}^m a(R_i) \quad (3)$$

where $1 \leq R_i \leq N$ denotes the rank of the observation Y_i ($i = 1, 2, \dots, m$) in the combined sample of $n + m = N$ observations and the system of weights $a(R_i)$ depends on the robust measures of asymmetry and tailweight.

L. Hao and D. Houser in [2] consider three possible statistics. The selection of the statistic depends on the type of the distribution. Three variants of the distributions have been considered: the symmetric heavier-tailed distributions, the symmetric light-tailed distributions and the right-skewed distributions. The details of these models can be described as follows:

- Symmetric Heavier-Tailed Model.

This model is selected when Q_3 is less than 2.1 and Q_4 is greater than 2.1. In this case

$$a(R_i) = R_i. \quad (4)$$

Let us denote the statistic for this case by T_1 . The weights (4) lead to the Wilcoxon-Mann-Whitney statistics. Formally, the test statistic can be written as follows

$$T_1 = \sum_{i=1}^m R_i$$

- Right-Skewed Model.

This model is selected when Q_3 is greater than 2.1. In this case bottom ranks begin near the median and consequently are more informative about differences in medians. L. Hao and D. Houser in [2] choose a modified rank test. This test uses only the bottom 50% of the data. Let us denote the statistic for this case T_2 . The scoring function has the following form

$$a(R_i) = \begin{cases} R_i - \text{floor}[25\%(N + 1)] - 0.5 & \text{if } R_i \leq 25\%(N + 1) \\ R_i - \text{ceiling}[75\%(N + 1)] - 0.5 & \text{if } R_i \geq 75\%(N + 1) \\ 0 & \text{otherwise} \end{cases} \quad (5)$$

where $\text{floor}(x)$ rounds x down to the nearest integer and $\text{ceiling}(x)$ rounds x up to the nearest integer.

- Symmetric Light-Tailed Model.

This model is chosen when Q_3 and Q_4 are both less than or equal to 2.1. In this case the modified rank test is used. The extreme ranks are more informative about location shifts than the central ones. Only the data from the bottom 25% and top 25% of the combined samples are used in calculating the statistic. Let us denote the statistic for this case by T_3 . The scoring function has the form of:

$$a(R_i) = \begin{cases} R_i - \text{floor}[25\%(N + 1)] - 1 & \text{if } R_i \leq (N + 1)/2 \\ 0 & \text{otherwise} \end{cases} \quad (6)$$

The HH test leads to the use of the proper statistic (T_1, T_2 or T_3) based on the Q_3 and Q_4 values from the combined samples. The idea of the test statistic selection is illustrated in Fig 1.

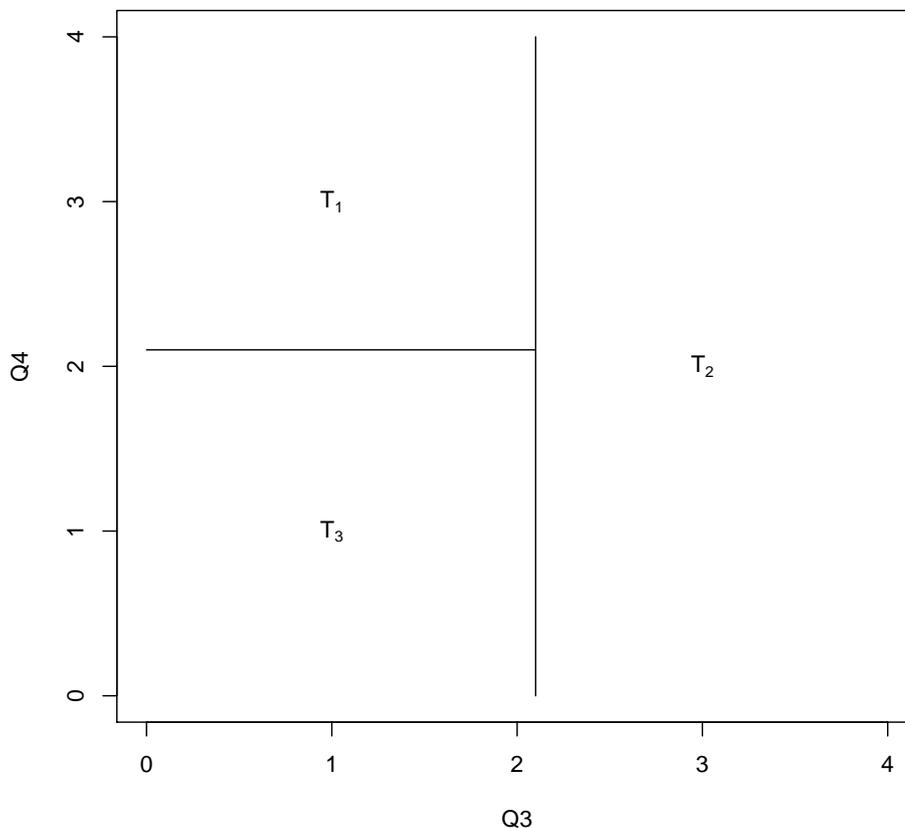


Figure 1: Model selection scheme for the HH adaptive test.

3 Modification of the HH test

The main idea of the proposed modification (mHH) of the HH test is not to select the proper test statistic but to change the weights of the combined statistics.

Let Q_3 and Q_4 be the robust asymmetry and the kurtosis statistics given by (1) and (2). Then (q_3, q_4) is a point on the Q_3/Q_4 plane as in Fig. 1. Let d_i for $i = 1, 2, 3$ be the euclidean distances of (q_3, q_4) to the border of the area of the use the statistic T_i in the HH test. Formally, it can be written as follows:

$$d_1 = \begin{cases} \sqrt{(q_3 - 2.1)^2 + (q_4 - 2.1)^2} & q_3 > 2.1 \text{ and } q_4 \leq 2.1 \\ q_4 - 2.1 & q_3 \leq 2.1 \text{ and } q_4 > 2.1 \\ q_3 - 2.1 & q_3 > 2.1 \text{ and } q_4 \leq 2.1 \\ 0 & \text{otherwise} \end{cases}$$

$$d_2 = \begin{cases} 2.1 - q_3 & q_3 < 2.1 \\ 0 & \text{otherwise} \end{cases}$$

$$d_3 = \begin{cases} \sqrt{(q_3 - 2.1)^2 + (q_4 - 2.1)^2} & q_3 > 2.1 \text{ and } q_4 > 2.1 \\ q_4 - 2.1 & q_3 \leq 2.1 \text{ and } q_4 > 2.1 \\ q_3 - 2.1 & q_3 > 2.1 \text{ and } q_4 \leq 2.1 \\ 0 & \text{otherwise} \end{cases}$$

Let us consider the test statistic

$$T = \alpha_1 T_1 + \alpha_2 T_2 + \alpha_3 T_3 \quad (7)$$

where α_i for $i = 1, 2, 3$ are weights given by $\alpha_i = \frac{w_i}{\sum_{i=1}^k w_i}$ and $w_i = \exp^{-d_i}$.

The distribution of the T statistic is unknown. To test the hypothesis that the samples were taken from the same distributions, a permutation test was used (see [11], [1]). The permutation procedure maintains the level of the significance of the test provided that under H_0 observations are exchangeable.

4 Monte Carlo study

Data process generation

Generalized lambda distribution (GLD) is a very useful means to test and fit data to well known distributions. This family of distribution can be used to generate random numbers from a distribution with a specified mean, variance, skewness and kurtosis. It is interesting because of the wide variety of distributional shapes it can take on (see [11]). Since the GLD is defined by its quantile function, it can provide a simple and effective algorithm for generating random variates. It can be used to generate random numbers with a specified mean, variance, skewness and kurtosis.

The generalized lambda distribution family GLD is a four-parameter family. The parameters are denoted by $\lambda_1, \lambda_2, \lambda_3, \lambda_4$ and the distribution usually by $GLD(\lambda_1, \lambda_2, \lambda_3, \lambda_4)$. The distribution is most easily specified in terms of its percentile function

$$Q(y) = \lambda_1 + \frac{y^{\lambda_3} - (1 - y)^{\lambda_4}}{\lambda_2} \quad (8)$$

where

λ_1 - location parameter,

λ_2 - scale parameter,

λ_3 - asymmetry parameter and

λ_4 - kurtosis parameter.

From (8) the probability density function can be written as follows

$$f(x) = f(Q(y)) = \frac{\lambda_2}{\lambda_3 y^{\lambda_3 - 1} - \lambda_4 (1 - y)^{\lambda_4 - 1}} \quad (9)$$

The generalized lambda distribution was used to generate data in the Monte Carlo study. The packages *gb*, *GLDEX*, *gld* and *gldist* from <http://www-r-project.org> were used.

Simulation procedure

In the Monte Carlo study the properties of the mHH test were compared to the properties of the HH test and the t test. Due to the construction of the T test statistic (7) the important area of the comparison is the one where q_3 and q_4 are close to 2.1 (see Fig. 1). The data in the study were generated from generalized lambda distribution. Computer simulations included the following steps

Step 1 Two samples are generated. The size of the first sample is n and the size of the second sample is m ($n = m = 25$).

Step 2 For the combined sample, the values of q_3 and q_4 are determined.

Step 3 The HH test (the test statistic is selected on the basis of Q_3 and Q_4 values), the t test and the proposed mHH test were performed.

Step 1 - 3 are repeated $N = 50,000$ times. The probabilities of the rejection of H_0 were estimated for the considered tests. For mHH test $N_p = 1,000$ number of permutations were considered.

See [8] for the assessment of the simulation error in estimating size and power of the tests which has two sources: an inner one (which corresponds to the inner permutation loop for p -values computation) and an outer one (which corresponds to the outer Monte Carlo loop for size/power computation).

Parameters $\lambda_1, \lambda_2, \lambda_3$ and λ_4 were established in such a way that the values Q_3 and Q_4 were close to 2.1 (see fig 1). To establish the values of the parameters $\lambda_1, \lambda_2, \lambda_3$ and λ_4 tables from [12] were used. Parameters $\lambda_1, \lambda_2, \lambda_3, \lambda_4$ were as follows: $\lambda_1 = 0, \lambda_2 = 1, \lambda_3 = 0.95, \lambda_4 = 2.25$ (mean = -0.21, variance=0.18). There were analysed two variants of the shift δ . The first one for the true H_0 where $\delta = 0$ and the second one for the false H_0 where $\delta = 0.1$.

Results

It is important for the comparison of the proposed mHH modification, that the HH and the t tests are near the borders of the three areas in Fig. 1. Random values were generated from the generalized lambda distribution.

For each sample, the values of q_3 and q_4 were calculated using formulas (3) and (4). Four regions were defined:

$$R_1: Q_3 \in (1.6, 2.1] \text{ and } Q_4 \in (1.6, 2.1]$$

$$R_2: Q_3 \in (2.1, 2.6] \text{ and } Q_4 \in (1.6, 2.1]$$

$$R_3: Q_3 \in (2.1, 2.6] \text{ and } Q_4 \in (2.1, 2.6]$$

$$R_4: Q_3 \in (1.6, 2.1] \text{ and } Q_4 \in (2.1, 2.6]$$

The analysed regions are presented in Fig. 2.

The results of testing the H_0 hypothesis for the proposed mHH test are presented in Fig. 3. In this figure only the results for the first 500 samples are presented. White dots denote "no rejection H_0 " and black dots denote "rejection H_0 ". Complete test results for the H_0 hypothesis for the HH test and the proposed modification are presented in Table 1.

The empirical size of the mHH test and the HH test are similar. Due to the non-normality of the distribution the size of the t test could not be maintained (see Table 1), but in analysed

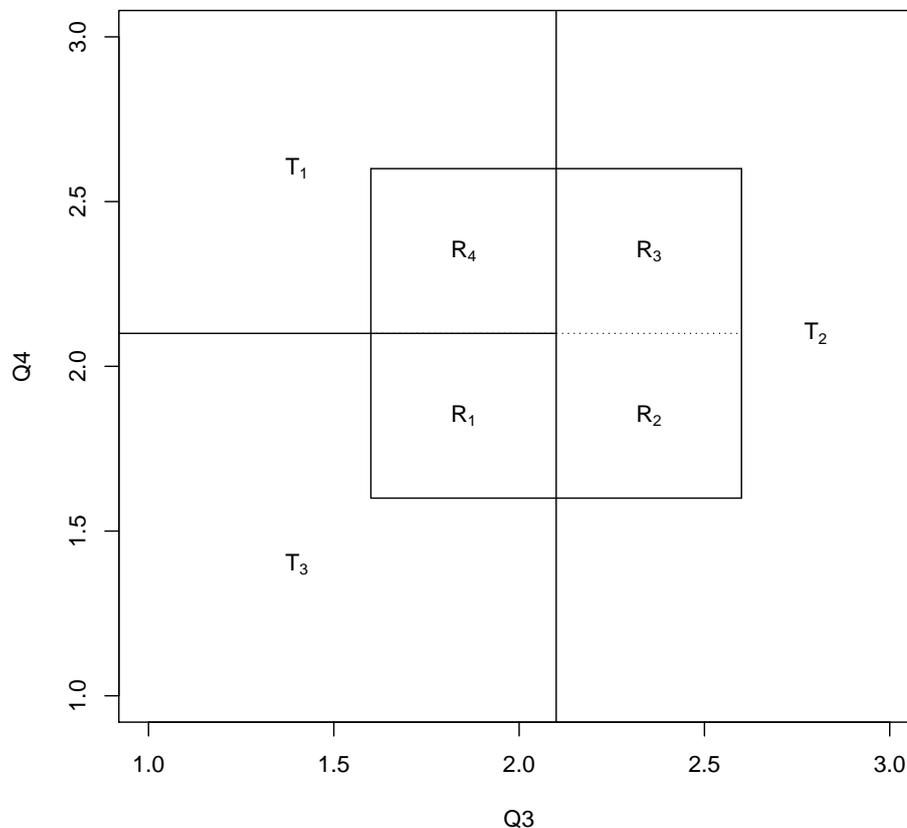


Figure 2: Considered regions of computer simulations.

Region	True H_0			False H_0		
	Test t	Test HH	Test mHH	Test t	Test HH	Test mHH
R_1	0.0520	0.0527	0.0538	0.3491	0.2318	0.3367
R_2	0.0481	0.0540	0.0538	0.3889	0.6087	0.5485
R_3	0.0473	0.0511	0.0489	0.3636	0.5707	0.5212
R_4	0.0530	0.0540	0.0527	0.3221	0.3473	0.3354

Table 1: Estimated probabilities of H_0 rejection in specified regions.

Regions $R_1 - R_4$ the size of this test is close to 0.05. The power of the HH test and the proposed mHH test is usually greater than the power of the t test.

Additionally, the power of these tests was analysed in the Monte Carlo study for three distributions. The symmetric distribution, skewed distribution and high kurtosis distribution were considered as in [11]:

D_1 - symmetric distribution - normal distribution $N(10, 1)$

D_2 - skewed distribution - GLD (mean=0, variance=1, skewness=1, kurtosis=4.2)

D_3 - high kurtosis distribution - GLD (mean=0, variance=1, skewness=2, kurtosis=15.6)

The power study was performed for three equal group sizes $n_1 = n_2 = 10, 15$ and 20 . The shift $\delta = 0.0, 0.2$ and 0.4 was considered. The power of the mHH test was compared to the power of t test, *Wilcoxon – Mann – Whitney* (WMW) test and *Kolmogorov – Smirnov* (KS) test. Zhang and Wu proposed omnibus test based on the likelihood ratio for location and shape (see [13]). If the distributions of populations are different in location only this test is as powerful as the old tests. The results of the size and the power Monte Carlo study are presented in Table 2.

The adaptive test mHH maintain its level of significance because it uses permutations method. Test t maintain its significance level only in the symmetric distribution (D_1) case. The size of the *Kolmogorov – Smirnov* test doesn't fit to the assumed significance level. It is possible fairly compare the power of the considered tests only if they maintain their significance levels. The power of the mHH test is similar to the power of t test in the symmetric distribution case.

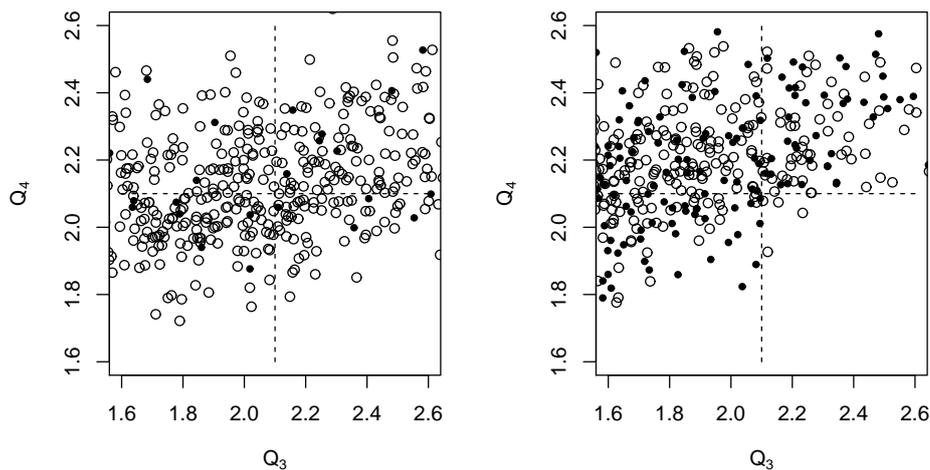


Figure 3: Results of testing H_0 - first 500 samples (left - true H_0 and right - false H_0).

5 Conclusions

The problem of comparing distributions based on two samples is often taken into consideration in quality control procedures, for example if the sample is compared to the reference sample taken from stable process. If the sample is taken from the normal population then the Shewhart control charts could be used. These tools are based on the sequence of the parametric tests. For the non-normal samples, non-parametric tests or adaptive tests could be used. The proposal of the adaptive test is presented in the paper. This test is based on ranks. It is a modification of L. Hao and D. Houser's procedure. The proposed adaptive test (mHH) is based on the

Test	$n_1 = n_2 = 10$			$n_1 = n_2 = 15$			$n_1 = n_2 = 20$		
	$\delta = 0$	$\delta = 0.2$	$\delta = 0.4$	$\delta = 0$	$\delta = 0.2$	$\delta = 0.4$	$\delta = 0$	$\delta = 0.2$	$\delta = 0.4$
D_1 - symmetric distribution									
<i>mHH</i>	0.0531	0.0739	0.1283	0.0497	0.0832	0.1847	0.0538	0.0963	0.2275
<i>WMW</i>	0.0445	0.0625	0.1188	0.0433	0.0753	0.1740	0.0504	0.0920	0.2190
<i>t</i>	0.0487	0.0694	0.1334	0.0490	0.0646	0.1902	0.0498	0.0939	0.2298
<i>KS</i>	0.0133	0.0179	0.0408	0.0247	0.0407	0.1016	0.0373	0.0607	0.1424
D_2 - skewed distribution									
<i>mHH</i>	0.0496	0.0671	0.1177	0.0544	0.0780	0.1548	0.0495	0.0827	0.1993
<i>WMW</i>	0.0439	0.0669	0.1393	0.0473	0.0841	0.2011	0.0480	0.1031	0.2692
<i>t</i>	0.0459	0.0676	0.1359	0.0479	0.0823	0.1906	0.0461	0.0918	0.2406
<i>KS</i>	0.0112	0.0207	0.0465	0.0288	0.0439	0.1184	0.0331	0.0678	0.1836
D_3 - high kurtosis distribution									
<i>mHH</i>	0.0520	0.0778	0.1484	0.0513	0.0862	0.2153	0.0498	0.1028	0.2654
<i>WMW</i>	0.0425	0.0738	0.1717	0.0443	0.0929	0.2662	0.0502	0.1176	0.3425
<i>t</i>	0.0404	0.0689	0.1556	0.0415	0.0828	0.2174	0.0461	0.0967	0.2674
<i>KS</i>	0.0127	0.227	0.0640	0.0245	0.0579	0.1812	0.0339	0.0809	0.2619

Table 2: Estimated probabilities of H_0 rejection.

permutation method. The size and the power of the adaptive test were analysed in the Monte Carlo study.

The Monte Carlo study has shown that the sizes of the tests in each analysed region in the case of *HH* test and the proposed *mHH* test are similar. Due to the permutation analysis, the proposed test maintains its significance level. The power of this test is close to the power of the *HH* test. The power of *t* test in the two cases is less than the power of the *HH* test and the *mHH* test. A direction of future research may be assessment of the scale problem (see [6]) and the location and scale problem (see [5]) which often arise in quality control and process monitoring.

Acknowledgement

The research was supported by Polish National Science Centre grant Nr DEC-2011/03/B/HS4/05630.

Bibliography

- [1] Efron, B., Tibshirani, R. (1993) *An Introduction to the Bootstrap*. Science Business Media, Inc.
- [2] Hao, L., Houser, D. (2012) *Adaptive Procedures for Wilcoxon-Mann-Whitney Test: Seven Decades of Advances*. <http://comp.uark.edu/~lhao/adaptive.pdf>. Communications in Statistics: Theory and Methods (forthcoming)

- [3] Hogg, R. V., Fisher, D. M. and Randles, R. H. (1975) *A two-sample adaptive distribution-free test*. Journal of the American Statistical Association, **70**, 656 – 661.
- [4] Marozzi, M. (2003) *Applications in Business, Medical and Industrial Statistics of Bi-Aspect Nonparametric Tests for Location Problems*. Statistical Methods and Applications, **12**, 187-194.
- [5] Marozzi, M. (2009) *Some Notes on the Location-Scale Cucconi Test*. Journal of Nonparametric Statistics **21**, 5, 629-647.
- [6] Marozzi, M. (2012) *A Combined Test for Differences in Scale Based on the Interquartile Range*. Statistical Papers **53**, 1, 61-72.
- [7] Marozzi, M. (2012) *A Modified Hall-Padmanabhan Test for the Homogeneity of Scales*. Communication in Statistics â Theory and Methods, **41**, 16-17, 3068-3078.
- [8] Marozzi, M. (2014) *The Multisample Cucconi Test*. Statistical Methods and Applications, DOI 10.1007/s10260-014-0255-x.
- [9] Montgomery, D. C., (2009) *Introduction to Statistical Quality Control*. John Wiley and Sons, New Jersey.
- [10] Mukherjee, A., Chakraborti, S. (2012) *A Distribution-free Control Chart for the Joint Monitoring of Location and Scale*. Quality and Reliability Engineering International, **28**, 335 - 352.
- [11] O’Gorman, T. W., Houser, D. (2012) *Adaptive Tests of Significance Using Permutations of Residuals with R and SAS*. John Wiley and Sons, Jefferson City.
- [12] Ramberg, J. S., Dudewicz, E. J., Tadikamalla P. R., Mykytka E. F. (1979) *A Probability Distribution and Its Uses in Fitting Data*. Technometrics vol. 21, no. 2, 201–214.
- [13] Zhang, J., Wu, Y. (2007) *k-Sample tests based on the likelihood ratio*. Computational Statistics & Data Analysis, **51**, 9, 4682-4691.

Comparison of techniques for extreme values using financial data

Joan del Castillo, *Universitat Autònoma de Barcelona*, castillo@mat.uab.cat

Maria Padilla, *Universitat Autònoma de Barcelona*, mpadilla@mat.uab.cat

Isabel Serra, *Universitat Autònoma de Barcelona*, iserra@mat.uab.cat

Abstract. In this article classical techniques of extreme value theory and two new statistical tools are compared through Monte Carlo techniques and on the daily log-returns of financial data extensively studied. The data sets predate the current economic crisis and so it is possible to evaluate retrospectively the quality of market risk estimates.

Keywords. Heavy tails, Exponential tails, Statistics of extremes, Value at risk, Tail index

1 Introduction

The extreme value theory (EVT) has two main approaches: Block maxima models and Threshold exceedance models. The financial markets provide many data sets where the two approaches may be compared estimating high quantile. The main objective of this paper is to compare the estimator of extreme value index using parametric, semi-parametric and non-parametric approach. Some semi-parametric models based on bias reduction techniques for heavy tails through the use of an adequate bias-corrected tail index estimator are considered. A new non-parametric tool based on the residual coefficient of variation is also analyzed, see [2]. This paper focuses on value-at-risk for log-returns arising in modeling extremes of four datasets in the field of finance, widely documented and studied. Applying extreme value methods in finance requires accurate estimators on extreme value index that can be around zero. New parametric models can still being of high interest for the analysis of extreme events, if associated with appropriate statistical inference methodologies, for instance, the full-tails gamma (FTG) distribution, see [3].

2 Techniques for extreme values

The generalized extreme value distribution (GEV) is the family $H(x; \xi, \mu, \phi) = H((x - \mu)/\phi; \xi)$ where $\mu \in \mathbb{R}$ and $\phi > 0$ are the localization and scale parameter and H corresponds to the

standard GEV defined by

$$H(x; \xi) = \begin{cases} \exp(-(1 + \xi x)^{-1/\xi}), & \xi \neq 0, \\ \exp(-e^{-x}), & \xi = 0, \end{cases} \quad (1)$$

where $1 + \xi x > 0$.

The cumulative distribution function of the generalized Pareto distribution (GPD) is given by

$$G(x; \xi, \psi) = 1 - (1 + \xi x/\psi)^{-1/\xi} \quad (2)$$

where $\psi > 0$ and ξ are scale and shape parameters. For $\xi > 0$ the range of x is $x > 0$ and the GPD is just one of several forms of the usual Pareto family of distribution often called the Pareto distribution. For $\xi < 0$ the range of x is $0 < x < \psi/|\xi|$, then GPD has bounded support. The limit case $\xi = 0$ corresponds to the exponential distribution (EXP). The opposite inverse of the shape parameter ξ is the tail index.

The methodology for modelling extreme values uses the peaks over threshold (PoT) approach. PoT is based on the theorem of Pickands-Balkema-DeHaan, see McNeil, *et al.* (2005) [9]. From this result, PoT is used by many authors for modelling exceedances in several fields such as finance and environmental science, see for instance Coles (2001) [4]. Several techniques have been developed to search for the optimal threshold to link a GPD, such as Hill-plot or ME-plot. This theoretical methodology shows some surprises in practical applications. For instance, Dutta and Perry (2006) [6] observed, in an empirical analysis of operational risk, that even when Pareto distribution fits correctly the data may result in unrealistic capital estimates (sometimes more than 100% of the asset size).

In order to contribute to solve these problems it is necessary to use alternative models to the GPD, but it requires certain properties that allow them to be treated as queuing models. In this way, Castillo *et al.* (2012) [3] introduce the FTG distribution with probability density function given by

$$f(x; \nu, \sigma, \theta) = \theta^\nu (x + \sigma)^{\nu-1} \exp(-\theta(x + \sigma)) / \Gamma(\nu, \sigma\theta) \quad (3)$$

where $\Gamma(\nu, \rho)$ is the upper incomplete gamma function, see Abramowitz y Stegun (1972) [1], the range of x is $(0, \infty)$ and $\nu \in \mathbb{R}, \theta > 0, \sigma > 0$. Remark that for σ fixed, if θ tends to zero, the FTG distribution corresponds to Pareto distribution and then ν is the tail index. The reason why FTG is more appropriate is because the financial data has heavy tails but they have some finite moments, see Shyriaev (1999)[10]. The existence of at least three moments allows us to develop new techniques for more satisfactory extreme values in practice. Furthermore, it is also interesting to consider the exponential distribution as the most basic tails model.

The coefficient of variation (CV) can be used also as a measure of non normality. The most popular measure of non normality nowadays is the kurtosis, defined for distributions with four finite moments. The next Lemma shows that the kurtosis can be obtained with the coefficient of variation.

Lemma 2.1. *Given a symmetric random variate X with respect to zero, the excess kurtosis is*

$$ku[X] + 3 = \frac{E[X^4]}{E[X^2]^2} = 1 + cv[X^2]^2,$$

therefore the kurtosis is a function of coefficient of variation of X^2 .

The coefficient of variation of X will be used as a measure of non normality, because it is defined using only the two first moments of the distribution. Hence, it is more stable and more widely applicable than kurtosis.

Let X be a continuous non-negative random variable (r.v.) with distribution function $F(x)$. For any threshold, $t > 0$, the r.v. of the conditional distribution of threshold exceedances $X - t$ given $X > t$, denoted by $X_t = (X - t | X > t)$, is called the *residual distribution* of X over t . The quantity $M(t) = E(X_t)$ is called the *residual mean* and $V(t) = \text{var}(X_t)$ the *residual variance*. The *residual coefficient of variation* is given by

$$CV(t) \equiv CV(X_t) = \sqrt{V(t)}/M(t), \quad (4)$$

like the usual CV, the function $CV(t)$ is independent of scale. Gupta and Kirmani (2000) [8] proved that the residual CV also characterizes the distribution. The residual CV for GPD, provided $\xi < 1/2$, is a constant given by

$$CV^2(t) = 1/(1 - 2\xi) \quad (5)$$

Hence, from Gupta and Kirmani (2000) [8] it follows that if $CV(t)$ is constant then the distribution of X is a *GPD*.

Castillo *et al.* (2014) [2], use these ideas to introduce a new graphical method. Given a sample $\{x_k\}$ of size n of positive numbers, we denote by $\{x_{(k)}\}$ the ordered sample, so that $x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)}$. A *CV-plot* is a representation of the empirical CV of the conditional exceedance (4), given by

$$k \rightarrow cv(x_{(k)}). \quad (6)$$

With this tool a non-parametric methodology can be used to estimate the tail index searching the value of the coefficient of variation that minimizes the distance between its confidence interval under hypothesis of constant tail index and the CV-plot. This non-parametric methodology provides both tail index and optimal threshold for computing high quantiles with PoT. This methodology combined with GPD as the model for the tail is denoted by CVm and some examples are showed in Table 1. Finally, the last methodology here considered is denoted by cHm and it consists in a semi-parametric method for high quantiles estimation based on the parametric model from Pareto and with a non-parametric techniques of bias-corrected Hill-estimator, see Gomes and Pestana (2007) [7], based on an adequate consistent estimator of the second order parameters, see Degen and Embrechts (2008) [5].

3 Numerical studies

In this section the two new methodologies, FTG and CVm, are compared with standard approaches, GPD and GEV, and the methodology based on second order corrections of the Hill estimator, cHm. First, the methods are compared using the daily log-returns of financial data extensively studied. Secondly, the behaviour of the techniques are studied when the simulated data are not really heavy tails, only semi-heavy. It can be observed that some methods presuppose heavy tails and do not consider that the tails can be exponential, what also happens in financial data.

Data analysis of log-returns

To compare the different techniques four sets of finance data are considered, collected over the same period: from January 4,1999 through November 17,2005. Those sets of data were the Euro-USA dollar (EUSD) daily exchange rates and the daily closing values of the Dow Jones Industrial Average In (DJI), Microsoft Corp. (MSFT), and International Business Machines Corp. (IBM) stocks.

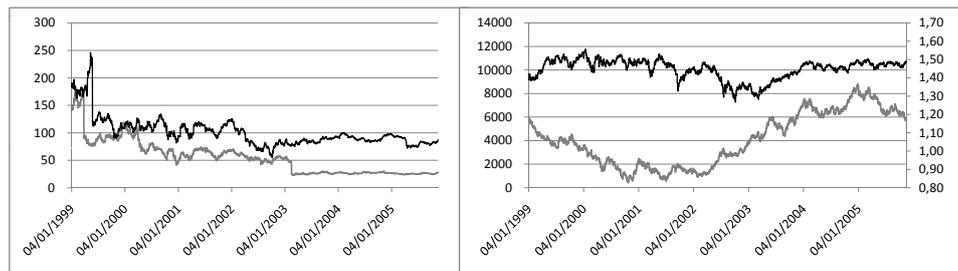


Figure 1: In the left, daily closing values of IBM data (dark line) and MSFT data (grey line). In the right, daily closing values of DJI (dark line and left axis) and EUSD daily exchange rates (grey line and right axis).

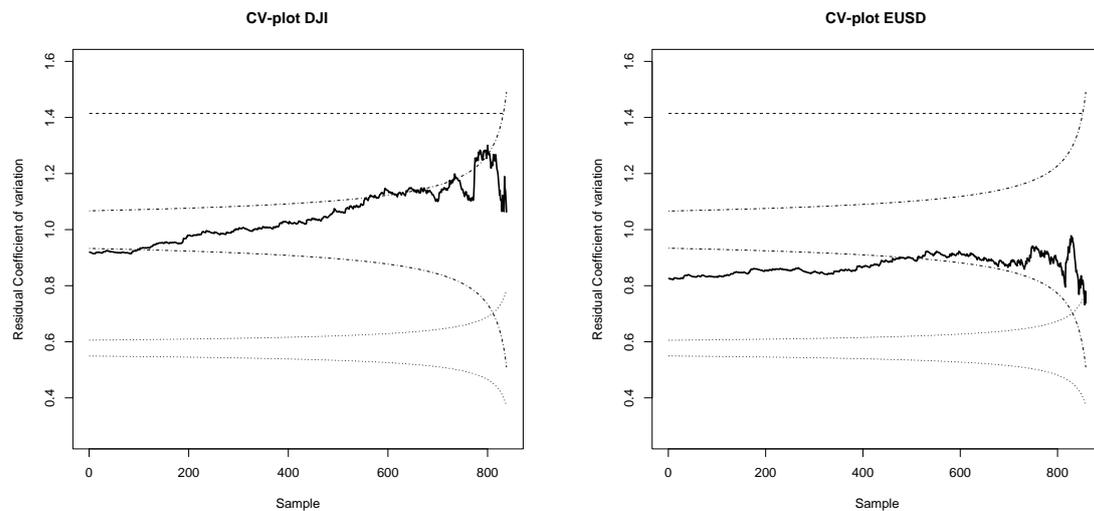


Figure 2: CV-plot of the absolute value of negative tail of log-returns. In the left, DJI data, in the right EUSD data. Dashed constant line corresponds to the residual coefficient of variation of a GPD with shape parameter 0.25, dotdash line corresponds to the 95% confidence interval of an exponential distribution and dotted line corresponds to the 95% confidence interval of a uniform distribution.

The assumption that financial data have heavy tail can lead to conclusions far removed from reality, in Figure 2 the CV-plot of EUSD shows that the shape parameter can be negative, since a residual CV less than 1 correspond to a negative shape parameter, see equation (5). From

Equation (5) it follows a residual CV is only possible with a negative shape parameter, so a heavy tail is not the best option. It is also necessary to be careful with the splits, due to the fact that a data can completely change the general behavior. Figure 3 shows the CV-plot of the same data with the difference that one contains the split but no the other in the IBM and MSFT cases. For example, the absolute value of the log-return the day that the split appears is 71% in the IBM case, in Figure 1 this value appears between 1999 and 2000.

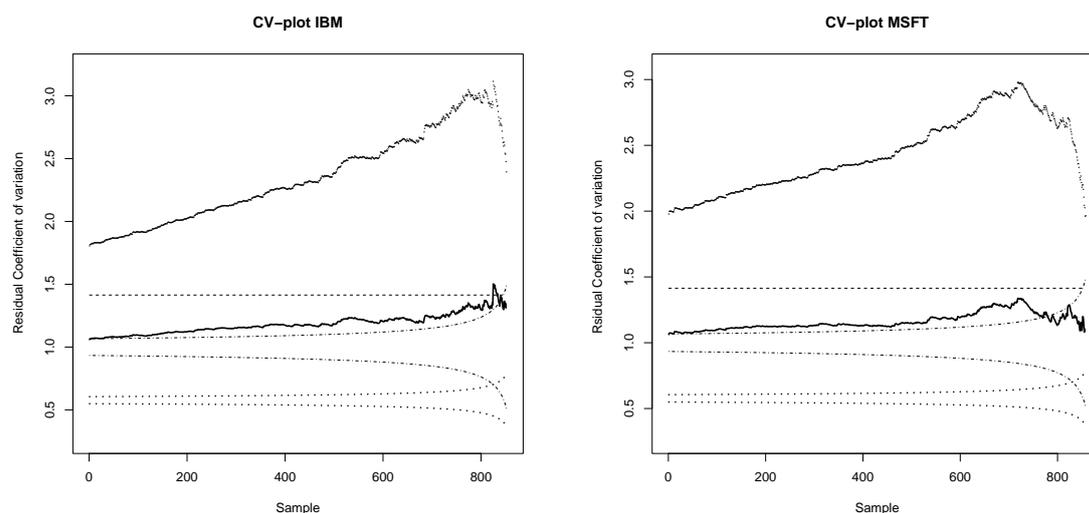


Figure 3: CV-plot of the absolute value of negative tail of log-returns. In the left IBM data with (dashed line) and without split (black line), in right, MSFT data with (dashed line) and without split (black line). Dashed constant line corresponds to the residual coefficient of variation of a GPD with shape parameter 0.25, dotdash line corresponds to the 95% confidence interval of an exponential distribution and dotted line corresponds to the 95% confidence interval of a uniform distribution.

Table 1 shows a brief of the results of the study. The cases *EXP*, *GPD*, *FTG* correspond to model the whole data as the corresponding parametric model and *GEV* to model the month maximums. To consider the new methodology *CVm* and the alternative *cHm*. The MSFT data do not appear in the table because it is very similar to IBM results. It can be observed some differences between the results using different methods. This differences are more significative when the split are included in the data. From applied point of view, more interesting results are obtained using POT with this advanced methodologies to search optimal threshold and improved parametric models for tails, for instance, the FTG.

Simulation study on the calculation of VaR

Monte Carlo simulations have been used to compare the previous methodologies for VaR 99,9% and tail index estimation. In Tables 2 and 3 the mean square error (MSE) obtained for each method is shown for 10,000 simulations performed for each of the sample sizes $n=150$, 250 and 500 of exponential distribution with scale 1. *EXP*, *GPD* and *FTG* denote the results of considering parametric models, exponential GPD and FTG, respectively. *cHm* and *CVm* denote

	99,9%	ξ		99,9%	ξ
DJI			EUSD		
GEV	0,104	0,17	GEV	0,026	-0,23
EXP	0,059	0	EXP	0,035	0
GPD	0,055	0,00	GPD	0,033	0,00
FTG	0,050	0	FTG	0,026	0
CVm	0,040	0,04	CVm	0,016	-0,16
cHm	0,068	0,30	cHm	0,027	0,26
IBM			IBMs		
GEV	0,660	0,42	GEV	0,304	0,43
EXP	0,110	0	EXP	0,104	0
GPD	0,144	0,11	GPD	0,104	-0,00
FTG	0,147	0	FTG	0,126	0
CV	4,857	2,58	CV	0,103	0,14
cHm	0,181	0,39	cHm	0,161	0,36

Table 1: A high quantile and the shape value ξ corresponding to the opposite inverse of the tail index for some simplified methodology and four sets of data: DJI, EUSD, IBM, and IBMs, the last corresponds to IBM data without splits.

the two semi-parametric models considered. Naturally, the EXP model provides the best results as well as the GPD because it contains the exponential and the FTG provides improved results since it is a model between them. CVm provides better results than cHm, since the underlying distribution is not a heavy tail. Remark that, it is important to consider that the nature of the data show the parametric and semiparametric methodologies can not be compared with this simulation results.

n	EXP	GPD	FTG	cHm	CVm
150	0.315	1.307	1.107	35.064	11.24
250	0.203	0.738	0.662	12.974	9.46
500	0.100	0.383	0.307	8.658	7.72

Table 2: MSE for the VaR 99.9% obtained for each method and for different sample sizes.

n	EXP	GPD	FTG	cHm	CVm
150	0	0.006	0	0.153	0.011
250	0	0.003	0	0.134	0.008
500	0	0.001	0	0.117	0.003

Table 3: MSE for the tail index, ξ , obtained for each method and for different sample sizes.

4 Conclusions

After analyzing the data sets of this study the following conclusions arise.

1. Given that EVT is very sensitive to outliers one must be very careful to analyze market data. It is repeatedly observed that the maximum values are outliers due to splits of corporations.
2. In practical applications since extrapolate for high quantiles is really difficult it is recommended to consider the data from different points of view and not be limited to a single technique.
3. The market data (once corrected for splits), is well fitted by models with semi-heavy tails that has few finite moments, as certain authors claim, see Shyriaev (1999) [10].
4. When evaluating risks, it is better to study separately the positive and negative tails of the distribution and not doing it together. Thus the coefficient of variation is a more appropriate tool than the kurtosis to assess the weight of the tails.

Acknowledgement

This work has been (partially) supported by Ministerio de Educación y Ciencia (MEC) of Spain under Grant Procesos Estocásticos Aplicados, MTM 2012-31118.

Bibliography

- [1] Abramowitz, M. and Stegun, I. A. (1972). *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*. New York: Dover.
- [2] Castillo, J. D., Daoudi, J. and Lockhart, R. (2014) *Methods to Distinguish Between Polynomial and Exponential Tails*. Scandinavian Journal of Statistics, **41**: 382–393. doi:10.1111/sjos.12037.
- [3] Castillo, J.D., Daoudi, J. and Serra, I (2012) *The full-tails gamma distribution applied to model extreme values*. arXiv:1211.0130.
- [4] Coles, S. (2001) *An Introduction to statistical of Extremes Values*. Springer, London.
- [5] Degen, M. and Embrechts, P. (2008) *EVT-based estimation of risk capital and convergence of high quantiles*. Advances in Applied Probability **40**, 696-715.
- [6] Dutta, K. and Perry, J. (2006) *A tale of tails: An empirical analysis of loss distribution models for estimating operational risk capital*. Working Papers 06-13
- [7] Gomes, M. I. and Pestana, D. (2007) *A sturdy reduced-bias extreme quantile (VaR) estimator*. Journal of the American Statistical Association Vol. **102**, No. 477.
- [8] Gupta, R. and Kirmani, S. (2000) *Residual coefficient of variation and some characterization results*. J. Stat. Plan. Infer. **91**, 23–31.

- [9] McNeil, A.J., Frey, R., and Embrechts, P. (2005) *Quantitative Risk Management: Concepts, Techniques, and Tools*. Princeton Series in Finance.
- [10] Shiryaev, Albert N (1999) *Essentials of stochastic finance: facts, models, theory*. Vol. **23**. Singapore: World scientific.

New insights into the usefulness of robust singular value decomposition in statistical genetics

Paulo Canas Rodrigues, *CMA-FCT-UNL, Nova University of Lisbon, Portugal; Federal University of Bahia, Brazil, paulocanas@gmail.com*

Andreia Monteiro, *CMA-FCT-UNL, Nova University of Lisbon, Portugal, andreiaforte50@gmail.com*

Vanda Lourenço, *CMA-FCT-UNL, Nova University of Lisbon, Portugal, vmml@fct.unl.pt.com*

Abstract. The distribution of continuous real life variables is usually not normal and plant phenotypes are no exception to the rule. These distributions often show heavy tails which are sometimes asymmetric. In such scenarios, the classical approach whose likelihood-based inference leans on the normality assumption may be inappropriate, having low statistical efficiency. Moreover, association tests may also be underpowered. Robust statistical methods are designed to accommodate for certain data deficiencies, allowing for reliable results under various conditions. They are designed to be resistant to influent factors as outlying observations, non-normality and other model misspecifications. Additionally, if the model verifies the classical assumptions, robust methods provide results close to the classical ones. Therefore, a new methodology where robust statistical methods replace the classic ones to model, structure and analyse genotype-by-environment interactions in the context of multi-location plant breeding trials, is presented. Here interest lies in the development of a robust version of the additive main effects and multiplicative interaction model whose performance is compared with its classical version. This is achieved through Monte Carlo simulations where one particular contamination scheme is considered.

Keywords. AMMI model, Robust statistics, Singular value decomposition, Statistical genetics

1 Introduction

Multi-environment trials (MET), which comprise experiments across multiple environments, are important tools for testing both broad and narrow genotype adaptation. Here, when two different genotypes show a differential response to a prototypic trait (e.g. yield) across environments, it

is said that genotype-by-environment interaction (GEI) is present. Data from MET are often summarized in two-way tables of means with genotypes in the rows and environments in the columns.

The additive main effects and multiplicative interaction (AMMI) model [4] is one of the most widely used tools for MET analysis. This tool works under a fixed-model framework and is conducted in two stages. First, the main effects of the model are estimated using the additive two-way analysis of variance (ANOVA) by least squares. Then, the singular value decomposition (SVD) is applied to the interaction residuals to obtain the estimates for the multiplicative terms of the AMMI model. The AMMI model in its standard form also implicitly assumes equal weights for all entries of the two-way data set and that no outlier is present in the data. However, field data such as data resulting from MET is prone to contamination and thus outlying observations are often found. As a consequence, the results from the analysis may be biased leading to possible misinterpretations which in turn may result in bad practical decisions. It is therefore important to improve the performance of the AMMI model in the cases where contamination is present in the data. For that reason we introduce in this work a robust AMMI model where the linear fit is replaced by a robust fit (M-regression) and the use of the standard SVD by a robust SVD approach. We underline that the choice of M-regression was based on the fact that in this kind of analysis contamination is only seen at the response variable level and not also at the explanatory variables level, in which case high breakdown and efficient MM-regression should be considered.

The proposed robust AMMI model is also useful in other studies where data contamination is inevitable, e.g., in QTL (quantitative trait loci) detection and QTL-by-environment interaction (QEI) studies. Here, the robust AMMI model will be used to calculate more accurate predicted values for GEI analysis. These predicted values can then be subject to a QTL analysis in a two stage procedure, similar to the ones described in [5, 11].

We present a Monte Carlo Simulation study to assess the performance of the proposed robust AMMI model, which is compared with the classical one under a particular contamination scheme.

2 Materials and methods

AMMI model

The AMMI model combines the features of ANOVA and SVD as follows: first the ANOVA estimates the additive main effects; then the SVD applied to the residuals from the additive ANOVA model, estimates the interaction with $N \leq \min(I - 1, J - 1)$ interaction principal components (IPC) axes. Here, I represents the number of genotypes (rows) and J the number of environments (columns) considered in the study and described in the two-way data table. Assuming for simplicity a completely randomized design for individual trials, the model can be written as [4]:

$$y_{i,j,k} = \mu + \alpha_i + \beta_j + \sum_{n=1}^N \lambda_n \gamma_{n,i} \delta_{n,j} + \rho_{i,j} + \epsilon_{i,j,k}, \quad (1)$$

where: $y_{i,j,k}$ is the phenotypic trait (yield or some other quantitative trait of interest) of the i th genotype in the j th environment for replicate k ; μ is the grand mean; α_i are the genotype deviations from μ ; β_j are the environment deviations from μ ; λ_n is the singular value of the IPC analysis axis n ; $\gamma_{n,i}$ and $\delta_{n,j}$ are the i th and j th genotype and environment IPC scores (i.e., the left and right singular vectors) for axis n , respectively; $\rho_{i,j}$ is the residual containing all multiplicative terms not included in the model; $\epsilon_{i,j,k}$ is the experimental error; N is the number of principal components retained in the model.

Robust AMMI model

We consider the following matrix formulation of the AMMI model:

$$\mathbf{Y} = \mathbf{1}_I \mathbf{1}_J^T \mu + \alpha_I \mathbf{1}_J^T + \mathbf{1}_I \beta_J^T + \mathbf{U} \mathbf{D} \mathbf{V}^T + \varepsilon, \quad (2)$$

where \mathbf{Y} is the $(I \times J)$ two-way data table of means, $\mathbf{1}_I \mathbf{1}_J^T \mu$ is a $(I \times J)$ matrix with the grand mean μ in all positions, $\alpha_I \mathbf{1}_J^T$ is a $(I \times J)$ matrix of genotype main effects (equal rows), $\mathbf{1}_I \beta_J^T$ is a $(I \times J)$ matrix of environmental main effects (equal columns).

The interaction part of the model $\mathbf{Y}^* = \mathbf{Y} - \mathbf{1}_I \mathbf{1}_J^T \mu - \alpha_I \mathbf{1}_J^T - \mathbf{1}_I \beta_J^T$ is approximated by the product of matrices $\mathbf{U} \mathbf{D} \mathbf{V}^T$, with \mathbf{U} an $(I \times N)$ matrix whose columns contain the left singular vectors of the interaction, \mathbf{D} a $(N \times N)$ diagonal matrix containing the singular values of \mathbf{Y}^* , and \mathbf{V} a $(J \times N)$ matrix whose columns contain the right singular vectors of \mathbf{Y}^* . The residual term in equation (2), the $(I \times J)$ matrix ε , includes both the lack of fit term and the error term of the model in equation 1.

We suggest that a robust AMMI model can be obtained in two stages as follows: (i) use the robust regression based on the M-Huber estimator [8] to replace the ANOVA model; (ii) use a robust SVD [6] to replace the standard SVD.

The robust methods described are available in the R software in packages MASS and pcaMethods via functions `rlm()` and `robustSVD()`, respectively.

3 Simulation study

We use Monte Carlo simulations to study the impact that contaminated data has in the results of the classical AMMI model and to assess the improvement that can be gained when using the proposed robust methodology. We discuss only a particular contamination setting with a fixed percentage of outliers to illustrate the advantage of the proposed methodology. Further complete studies are being carried but will not be presented here.

In this particular case we simulate 1000 two-way data tables with 100 rows/genotypes and 8 columns/environments each, where the interaction is explained by two multiplicative terms (i.e., two IPCs). The number of multiplicative terms was confirmed by the cross-validation procedure proposed by [3] for principal component analysis and then generalized by [2, 1] for the AMMI model. In each run of the simulation, the AMMI and robust AMMI models were used to analyse these data. After the AMMI is applied to the data, the biplots are constructed and the singular

values obtained. Having the good non-contaminated data, 5% of contamination is introduced in the two-way original data table so as to be consistent with the known shift-outlier case [10]: (i) 5% positions are randomly selected in the two-way table thus assigning contamination positions in different environments for distinct genotypes; (ii) the 5% bad data is generated from a $N(\mu_j+k, \sigma_j^2)$ (pure shift outliers; $k = 4\sigma_j$ units) where μ_j and σ_j^2 are taken as the sample phenotypic mean and sample phenotypic variance according to the correspondent environment j , $j = 1, \dots, 8$; and (iii) the bad data replaces the 5% of the good data from the two-way table at the positions assigned in (i). Method comparison is achieved using the mean squared error (MSE) [9]:

$$MSE(\hat{\lambda}_j) = \frac{1}{1000} \sum_{l=1}^{1000} (\hat{\lambda}_j^{(l)} - \lambda_j)^2, \quad (3)$$

where λ_j , $j = 1, 2$, are the true singular values of the two-way data tables, and the $\hat{\lambda}_j^{(l)}$, $j = 1, 2$, are the estimated singular values for each of the 1000 replications, using the robust AMMI model for the raw uncontaminated data, and both the AMMI and robust AMMI models for the contaminated data.

4 Results and Discussion

Simulation study

Figure 1 shows the biplots obtained for the AMMI and robust AMMI models with two principal components (AMMI2), with and without contamination, for one random simulation run. The component loadings (for the environments) are similar for all four models. As for the scores of the genotypes, a similar behaviour is seen for both models without contamination (top two plots of Figure 1), as expected. However, when 5% contamination is considered, the display of genotypes shows a completely different behaviour in the AMMI2 biplot (bottom left plot in Figure 1). This shows that the AMMI model is not appropriated when the data is contaminated. When comparing with the biplot for the robust AMMI model with 5% contaminated data (bottom right plot in Figure 1), the scores for the genotypes show strong similarities with the biplots for the data without contamination, showing the usefulness of the robust AMMI model for contaminated data. With the use of the robust AMMI model, the impact of the outlier observations is reduced and the position of the scores becomes similar to the “true” position given by the AMMI2 model without contamination. Consequently, the use of this robust version of the AMMI model will allow practitioners to make better strategic decisions.

The MSE obtained for the robust AMMI model applied to both the contaminated and uncontaminated data, and the MSE obtained for the AMMI model applied to the contaminated data, are presented in Table 1. As expected, the MSE between the AMMI model and the robust AMMI model is small when considering the data without contamination. However, when we consider the data with 5% contamination, the robust AMMI model provides a MSE 6.21 times lower for the IPC1 and 4.61 times lower for the IPC2, when compared with the AMMI model. This result confirms the usefulness of the robust AMMI model when dealing with contaminated data.

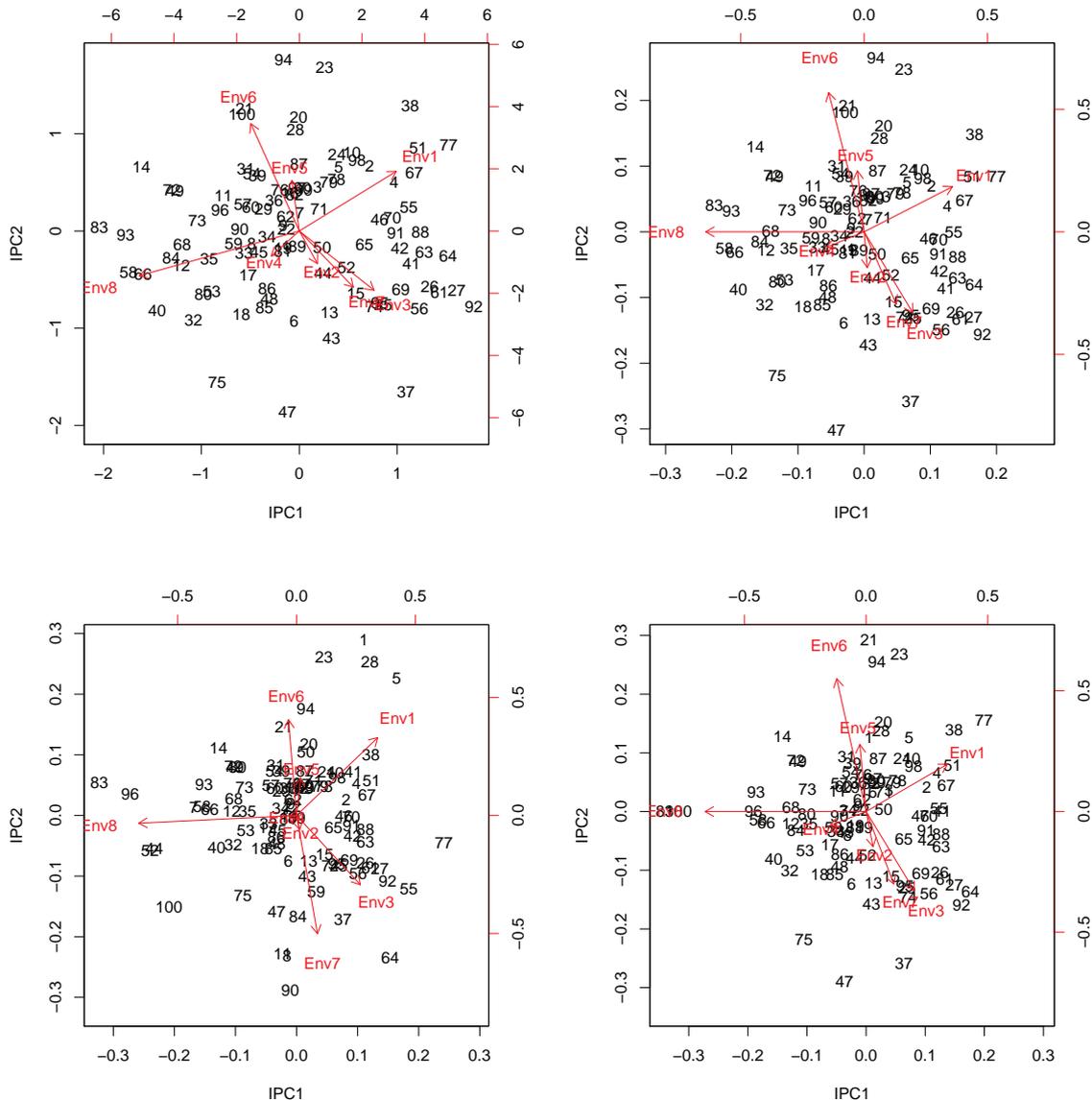


Figure 1: Biplots for: AMMI2 of data without contamination (top left); robust AMMI2 of data without contamination (top right); AMMI2 of data with 5% contamination (bottom left); robust AMMI2 of data with 5% contamination (bottom right).

To conclude, this preliminary simulation study outlined the fragility of the classical AMMI model in the presence of contaminated data. Moreover, the use of the robust methodologies proposed, not only provided results similar to the classical ones when there was no contamination but also proved to provide better results when the data was in fact contaminated. It is therefore important to study further the “robustification” of the AMMI model so as to account for what is in practice data reality: data contamination. This need is more than justified for the wide use of this technique in multi-environmental studies upon which many important decisions are

Model	IPC1	IPC2
Robust AMMI	26.00	34.03
AMMI (5% contaminated data)	1159.22	1268.81
Robust AMMI (5% contaminated data)	186.57	275.33

Table 1: Mean square errors for singular values of the the AMMI and robust AMMI models.

made.

Real data example

The real data set used for illustration is the Steptoe x Morex (SxM) barley mapping population [7]. Figure 2 shows the biplots obtained for the AMMI and robust AMMI models with two principal components. In the AMMI2 biplot for the classic model (left hand side of Figure 2) the environment OR1 shows a dominant effect over the biplot being non-correlated with most of other environments and presents an overlap in the direction of many of the loadings for the environments. This makes this biplot difficult to analyse. When considering the robust AMMI model (right hand side of Figure 2) the interpretation of the biplot seems easier, with the environments more spread and with different angles between their component loadings and without such dominant influence of the environment OR1. These results are consistent with [5] where environment OR1 was considered to be an outlying environment. Moreover, the use of the robust AMMI model made it possible for this particular environment to be included in the analysis without distorting the final results, which was achieved by reducing its influence on the final model. To conclude, this example application further reinforces the usefulness of the proposed methodology.

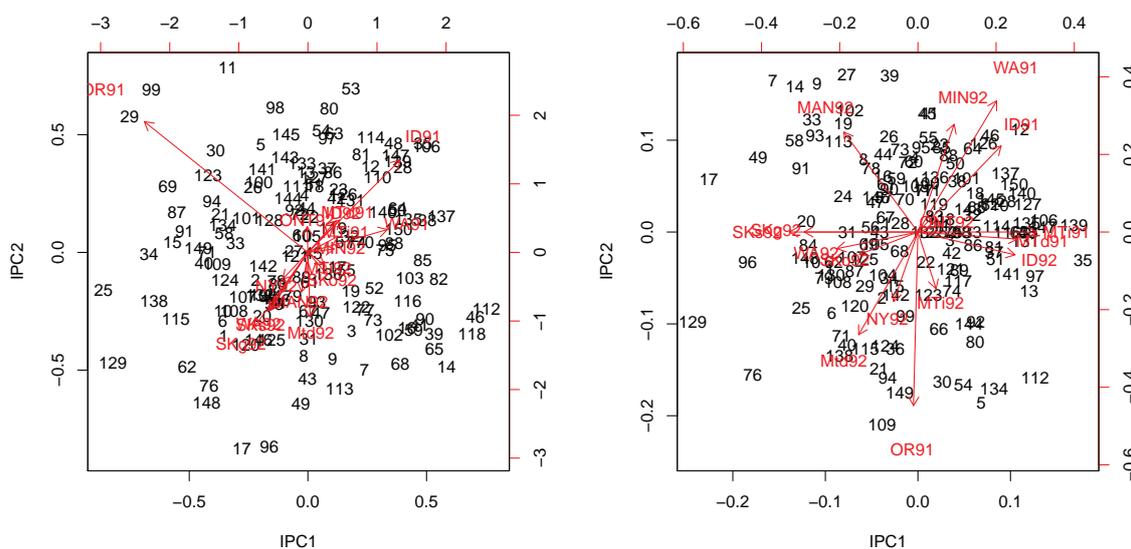


Figure 2: Biplots for: AMMI2 of SxM data (left); and robust AMMI2 of SxM data (right).

Acknowledgements

The authors acknowledge financial support from the Fundação para a Ciência e a Tecnologia (Portuguese Foundation for Science and Technology) through the projects PTDC/MAT-STA/0568/2012 and PEst-OE/MAT/UI0297/2014 (CMA).

Bibliography

- [1] Dias, C.T.S., and Krzanowski, W.J. (2006). *Choosing components in the additive main effect and multiplicative interaction (AMMI) models*. *Scientia Agricola* **63**, 169–175.
- [2] Dias, C.T.S., and Krzanowski, W.J. (2003). *Model Selection and Cross Validation in Additive Main Effect and Multiplicative Interaction Models*. *Crop Sci.* **43**, 865–873.
- [3] Eastment, H.T. and Krzanowski, W.J. (1982). *Cross-Validatory Choice of the Number of Components from a Principal Component Analysis*. *Technometrics* **24**, 73–77.
- [4] Gauch, H. G. (1992). *Statistical analysis of regional yield trials: AMMI analysis of factorial designs*. Elsevier, Amsterdam.
- [5] Gauch, H. G., Rodrigues, P. C., Munkvold, J. D., Heffner, E. L. and Sorrells, M. (2011). *Two New Strategies for Detecting and Understanding QTL x Environment Interactions*. *Crop Science* **51**: 96-113.
- [6] Hawkins, D. M., Liu, L. and Young, S. (2001). *Robust Singular Value Decomposition*. National Institute of Statistical Sciences, Technical Report Number 122.
- [7] Hayes, P.M., Liu, B.H., Knapp, S.J., Chen, F., Jones, B., Blake, T., Frankowiak, J., Rasmusson, D., Sorrells, M., Ullrich, S.E., Wesenberg, D., and Kleinhofs, A. (1993). *Quantitative trait locus effects and environmental interaction in a sample of North American barley germplasm*. *Theor. Appl. Genet.*, **87**, 392–401.
- [8] Huber, P. J. (1981). *Robust Statistics*. New York: John Wiley and Sons.
- [9] Hubert, M., Rousseeuw, P.J., and Branden, K.V. (2005). *Robpca: a new approach to robust principal component analysis*. *Technometrics*, **47(1)**, 64–79.
- [10] Rocke, D. M. and Woodruff, D. L. (1996). *Identification of Outliers in Multivariate Data*. *J. Am. Stat. Assoc.*, **91**, 1047–1061.
- [11] Rodrigues, P.C., Malosetti, M., Gauch, H. and Eeuwijk, F.A. (2014). *Weighted AMMI to study genotype-by-environment interaction and QTL-by-environment interaction*. doi:10.2135/cropsci2013.07.0462

Time series clustering based on quantile autocovariances

Borja Lafuente–Rego, *Universidade da Coruña*, borja.lafuente@udc.es
Jose Antonio Vilar, *Universidade da Coruña*, jose.vilarf@udc.es

Abstract. Time series clustering is an active research topic with applications in many fields. Unlike conventional clustering on static data, time series are inherently dynamic and hence the similarity searching must be governed by the behavior of the series over their observation periods. A dissimilarity aimed to compare quantile autocovariance functions is proposed to perform clustering. Results from an extensive simulation study show that the proposed metric outperforms a range of alternative dissimilarities reported in the literature. Estimation of the optimal number of clusters is also discussed. A prediction-based resampling algorithm proposed by Dudoit and Fridlyand [2] is adjusted to be applied in clustering based on quantile autocovariances. Several criteria to select the number of clusters are examined in new simulations.

Keywords. Time series, Clustering, Quantile autocovariances, Clest.

1 Introduction

Time series clustering is a central problem in many application fields and it is nowadays an active research area in a vast range of fields (finance and economics, medicine, engineering, pattern recognition, among many others). Comprehensive surveys can be seen in Liao [10] and more currently in Fu [5]. A crucial point is to determine the similarity notion between time series. Unlike conventional clustering on static data objects, time series are inherently dynamic, with underlying autocorrelation structures, and therefore the similarity searching must be governed by the behavior of the series over their periods of observation. Many dissimilarity measures have been proposed in the literature. The R package `TSclust` [15] presents a large set of well-established peer-reviewed time series dissimilarity measures, including measures based on raw data, extracted features, underlying parametric models, complexity levels, and forecast behaviors. We focus on the feature-based approach, where the raw data are replaced by a reduced number of extracted features and then dissimilarity between these representations is assessed. Some authors have considered measures based on comparing estimated simple or partial autocorrelations and cepstral coefficients (see [1, 3, 14]). We propose to measure dissimilarity by comparing quantile autocovariance functions [8]. For a given time series X_t , the quantile autocovariance function (QAF) consists of the cross-variances $cov(I(X_t \leq x), I(X_{t+r} \leq y))$, where $I(\cdot)$ denotes the indicator function. The quantile autocovariances examine the so-called serial

dependence structure, i.e. the joint distribution of (X_t, X_{t+r}) , for all t and r , so accounting for sophisticated serial features that simple autocovariances are unable to detect. To the best of our knowledge, QAF has not been considered to perform clustering, even though it satisfies suitable properties to carry out this task, such as light computational complexity and robustness inherent to quantile methods. Unlike the usual autocovariance function, QAF is robust to the non-existence of moments. This way, a QAF-based dissimilarity should take advantage to discriminate between series generated from processes with different heavy-tailed marginal distributions or presenting different conditional heteroscedasticity models. Many financial time series (log-return series of stock indices, share prices, exchange rates, etc) are known to exhibit this kind of properties. In such cases, usual feature-based dissimilarities are unable to capture differences between dynamic behaviours. For instance, similar correlograms and flats spectra are exhibited by both an ARCH(1) and a Gaussian white noise process. Theoretical properties of the quantile autocovariances and the quantile spectral density have been established in [11, 8, 9]. In this work, the behavior in time series clustering of a QAF-based dissimilarity is examined on different simulation scenarios and compared with other dissimilarities.

The problem of estimating the number of clusters K underlying the database is also addressed by adjusting the prediction-based resampling algorithm (so-called Clest) proposed by Dudoit and Fridlyand [2]. Clest is aimed to select the value of K providing the strongest evidence against $H_0 : K = 1$. For each value of K , Clest evaluates the amount of reproducibility, say R_K , of the K -cluster solution combining ideas from supervised and unsupervised learning, and then examines whether the value of R_K is significantly larger than the expected one under the null hypothesis of no clusters. In the original procedure, the expected value for R_K under H_0 is approximated resampling a multivariate uniform distribution. Nevertheless, this is not reasonable when dependent data are considered. To overcome this drawback, the uniformity assumption is marginally considered for each quantile autocovariance, i.e. the reference datasets are generated from univariate uniform distributions. This modified version of Clest algorithm was examined and compared with other alternative procedures in a new simulation study. All the simulations and the analysis of real data have been carried out using the R language [18].

2 Clustering procedure

Consider a set of p time series $S = \{\mathbf{X}^{(1)}, \dots, \mathbf{X}^{(p)}\}$, with $\mathbf{X}^{(j)} = (X_1^{(j)}, \dots, X_T^{(j)})$ being a T -length partial realization from a real valued process $\{X_t^{(j)}, t \in \mathbb{Z}\}$. Our goal is to perform cluster analysis on S to group the series into K homogeneous clusters. First, a dissimilarity measure between two series is introduced in terms of sequences of estimated quantile covariances.

Given a strictly stationary time series $\{X_t\}$, the *quantile covariance function* is defined by $\gamma_r(q, q') = cov\{I(X_t \leq q), I(X_{t+r} \leq q')\} = P(X_t \leq q, X_{t+r} \leq q') - P(X_t \leq q)P(X_{t+r} \leq q')$, with $(q, q') \in \mathbb{R}^2$. Function $\gamma_r(q, q')$ can be estimated from a T -length stretch (X_1, \dots, X_T) by

$$\hat{\gamma}_r(q, q') = \frac{1}{T} \sum_{t=1}^{T-r} Z_t(q)Z_{t+r}(q'), \quad (1)$$

where Z_t is the centered variable $Z_t(q) = I(X_t \leq q) - \hat{F}_T(q)$, with $\hat{F}_T(q) = \frac{1}{T} \sum_{i=1}^T I(X_i \leq q)$.

Each series $\mathbf{X}^{(j)}$ in S is characterized by an ordered set $\Gamma^{(j)}$ of quantile autocovariances estimated according to (1). Specifically, for a prefixed range L of lags l_1, \dots, l_L and r quantiles $(q_{\tau_1}, \dots, q_{\tau_r})$, with $q_{\tau_i} = F^{-1}(\tau_i)$, $\Gamma^{(j)}$ is given by

$$\Gamma^{(j)} = \left(\Gamma_{l_1}^{(j)}, \dots, \Gamma_{l_L}^{(j)} \right), \quad (2)$$

with $\Gamma_{l_i}^{(j)} = (\hat{\gamma}_{l_i}(q_{\tau_1} q_{\tau_1}), \dots, \hat{\gamma}_{l_i}(q_{\tau_1} q_{\tau_r}), \hat{\gamma}_{l_i}(q_{\tau_2} q_{\tau_2}), \dots, \hat{\gamma}_{l_i}(q_{\tau_2} q_{\tau_r}), \dots, \hat{\gamma}_{l_i}(q_{\tau_r} q_{\tau_r}))$, $i = 1, \dots, L$.

In practice, the quantiles q_{τ_i} are unknown and must be estimated by the empirical quantiles \hat{q}_{τ_i} . Then the dissimilarity between a pair of series $\mathbf{X}^{(i)}$ and $\mathbf{X}^{(j)}$ is defined as the squared Euclidean distance between $\Gamma^{(i)}$ and $\Gamma^{(j)}$, and it is denoted by $d_{QAF}(\mathbf{X}^{(i)}, \mathbf{X}^{(j)})$. Computing these distances for all pairs of series in S allows us to set a pairwise dissimilarity matrix, which is taken as starting point to develop a conventional agglomerative hierarchical clustering algorithm.

3 Simulation study: Part I

A first set of simulations was conducted to assess the behaviour of d_{QAF} in time series clustering. Different processes were considered to examine robustness, and comparisons with other model-free and model-based dissimilarities were carried out. In this abstract, results from two particular classification setups are shown, namely classification of nonlinear models and classification of different structures of conditional heteroscedasticity. The specific models are presented below.

- **Scenario 1:** Non-linear processes classification. The studied models are:

Model 1:	TAR	$X_t = 0.5X_{t-1}I(X_{t-1} \leq 0) - 2X_{t-1}I(X_{t-1} > 0) + \epsilon_t$
Model 2:	EXPAR	$X_t = (0.3 - 10\exp(-X_{t-1}^2))X_{t-1} + \epsilon_t$
Model 3:	MA	$X_t = -0.4\epsilon_{t-1} + \epsilon_t$
Model 4:	NLMA	$X_t = -0.5\epsilon_{t-1} + 0.8\epsilon_{t-1}^2 + \epsilon_t$

- **Scenario 2:** Conditional heteroscedastic processes classification. Consider $X_t = \mu_t + a_t$, with $\mu_t \sim \text{MA}(1)$ and $a_t = \sigma_t \epsilon_t$, $\epsilon_t \sim \text{IID}(0, 1)$. Then, the following structures for the varying conditional variance are considered:

Model 1:	ARCH(1)	$\sigma_t^2 = 0.1 + 0.8a_{t-1}^2$
Model 2:	GARCH(0,1)	$\sigma_t^2 = 0.1 + \sigma_{t-1}^2$
Model 3:	GJR–GARCH	$\sigma_t^2 = 0.1 + (0.25 + 0.3N_{t-1})a_{t-1}^2 + 0.5\sigma_{t-1}^2$; $N_{t-1} = I(a_{t-1} < 0)$
Model 4:	EGARCH	$\ln(\sigma_t^2) = 0.1 + \epsilon_{t-1} + 0.3[\epsilon_{t-1} - \mathbb{E}(\epsilon_{t-1})] + 0.4\ln(\sigma_{t-1}^2)$

In all cases, the error process ϵ_t consisted of i.i.d. $\mathcal{N}(0, 1)$ variables. Five series of length $T = 200$ were generated from each model over $N = 100$ trials. The considered dissimilarity measures and proper references are briefly summarized below.

- *Periodogram-based distances* [1, 16]. Euclidean distance between periodograms (d_P), log–periodogram (d_{LP}), normalized periodograms (d_{NP}) and log–normalized periodograms (d_{LNP}).
- *Autocorrelation-based distances* [1, 16]. Euclidean distance between simple (d_{ACF}) and partial (d_{PACF}) autocorrelations using a number of significant lags. Versions d_{ACFG} and d_{PACFG} including geometric weights decaying with the lag $\omega_i = \pi(1 - \pi)^i$, with $0 < \pi < 1$, were also considered. In our study, ten lags and $\pi = 0.5$ were used.
- *Model-based distances*. AR distances proposed by Piccolo (d_{PIC}) [17] and Maharaj (d_M) [12].
- *Nonparametric dissimilarities in the frequency domain*. A spectral dissimilarity measure based on local linear fits of log-spectra using maximum likelihood (d_{WLK}) [20, 16], and a dissimilarity measure based on the integrated squared difference between estimated log-spectra (d_{ISD}) [16].
- *The proposed metric d_{QAF}* . Results presented here were obtained with $r = 3$ empirical quantiles given by $(\hat{q}_{0.1}, \hat{q}_{0.5}, \hat{q}_{0.9})$ and only one lag, that is $L = 1$, with $l_1 = 1$.

Assuming that the clustering is governed by similarity between underlying models, the “true” cluster partition is given by the four clusters involving the five series generated from

the same model. The experimental 4-cluster solutions are compared with the true partition using three agreement measures based on known “ground-truth”: the Gavrilov index [6], the adjusted Rand index and the one-nearest-neighbour classifier evaluated by leave-one-out cross-validation (loo1NN) [7]. In all cases, the closer to 1 is the index, the higher is the agreement between the true and experimental partitions. The obtained indexes, averaged over 100 trials, are shown in Table 1.

Measure	Scenario 1			Scenario 2		
	Gavrilov	Adj. Rand	loo1NN	Gavrilov	Adj. Rand	loo1NN
<i>Periodograms</i>						
d_P	0.402	0.081	0.429	0.441	0.113	0.497
d_{LP}	0.713	0.501	0.694	0.689	0.441	0.653
d_{NP}	0.488	0.145	0.366	0.468	0.151	0.428
d_{LNP}	0.486	0.115	0.373	0.570	0.248	0.457
<i>Autocorrelations</i>						
d_{ACFG}	0.592	0.310	0.554	0.604	0.313	0.599
d_{PACFG}	0.667	0.397	0.613	0.641	0.358	0.589
d_{PACF}	0.610	0.306	0.550	0.625	0.327	0.541
<i>Model-based</i>						
d_{PIC}	0.674	0.443	0.751	0.560	0.291	0.615
d_M	0.680	0.453	0.746	0.632	0.374	0.653
<i>Non-parametric</i>						
d_{WLK}	0.914	0.821	0.920	0.733	0.530	0.764
d_{ISD}	0.916	0.826	0.919	0.740	0.541	0.765
<i>Quantile autocov.</i>						
d_{QAF}	0.961	0.917	0.980	0.908	0.800	0.919

Table 1: Clustering on nonlinear (Scenario 1) and heteroscedastic (Scenario 2) processes: cluster evaluation indexes averaged through all the 4-cluster hierarchical solutions for several dissimilarity measures. Series length $T = 200$. Number of trials $N = 100$. Complete linkage procedure.

Results in Table 1 allows us to conclude that the quantile autocovariance dissimilarity d_{QAF} produced the best results in both scenarios. Except for the adjusted Rand index in Scenario 2, d_{QAF} always led to indexes above 0.9, and sometimes very close to 1 in Scenario 1. As expected, metrics based on ARMA models (d_{PIC} and d_M) were strongly affected by model misspecification and produced poor results. The nonparametric dissimilarities work fairly well in Scenario 1, with results above 0.9 and close to the best ones attained by d_{QAF} . These measures take advantage of being free of the linearity restriction, and hence their good performance. Nevertheless, their behaviour substantially worsened by classifying heteroscedastic models. In fact, d_{QAF} noticeably outperforms both d_{WLK} and d_{ISD} in Scenario 2. In addition, the nonparametric measures employed computing times significantly higher than d_{QAF} , which is very important in time series clustering where huge databases with long series are often used. Distance d_{QAF} obtained excellent scores for loo1NN index and this is also remarkable because this criterion directly evaluates the efficacy of the dissimilarity measure regardless of the considered clustering algorithm. The remaining metrics produced the poorest results, corresponding the worst classification to the periodograms-based measures. In particular, the Euclidean distance between periodograms (d_P) worked really bad. Simple ACF and PACF were not able to separate correctly the considered models such as quantiles autocovariances did. Additional simulations were carried out

using different clustering algorithms and alternative scenarios. In all cases, d_{QAF} led to very good results, attaining competitive results even to cluster ARMA models.

4 Determining the number of clusters

A prediction–based resampling algorithm (called Clest) introduced by Dudoit and Fridlyand [2] to estimate the optimal number of clusters K is here adjusted to: (i) use d_{QAF} in the clustering process involved by the algorithm, and (ii) overcome the dependence underlying the classifying variables $\Gamma^{(j)}$ given in (2). Clest is aimed to select \hat{K} , $2 \leq \hat{K} \leq M$, with $M \leq p$ denoting the maximum possible of clusters, that provides the strongest evidence against the null hypothesis $H_0 : K = 1$. The version of Clest including the proposed adjustments is outlined below.

Algorithm 4.1.

For each k , $2 \leq k \leq M$, perform steps 1–4 below.

Step 1 Repeat B times:

- i. Randomly split the set of series S into two groups, a learning set \mathcal{L}^b and a test set \mathcal{T}^b .
- ii. Using the clustering procedure described in Section 2, based in d_{QAF} , obtain partitions $\mathcal{P}(\cdot; \mathcal{L}^b)$ and $\mathcal{P}(\cdot; \mathcal{T}^b)$ of the sets \mathcal{L}^b and \mathcal{T}^b , respectively.
- iii. Classify each series of the test set \mathcal{T}^b into the closest cluster (according to d_{QAF}) of $\mathcal{P}(\cdot; \mathcal{L}^b)$, thus obtaining the new partition $\mathcal{C}(\cdot; \mathcal{T}^b)$.
- iv. Evaluate an index of agreement $s_{k,b}$ between partitions $\mathcal{C}(\cdot; \mathcal{T}^b)$ and $\mathcal{P}(\cdot; \mathcal{T}^b)$.

Step 2 Compute $R_k = \text{median}(s_{k,1}, \dots, s_{k,B})$.

Step 3 Generate B_0 resamples of the quantile autocovariances matrix under $H_0 : K = 1$. As the columns of this matrix are dependent, the resamples of each column are separately generated from an uniform distribution with support determined by the range of the column. Then, Steps 1 and 2 are repeated for each resample obtaining $R_{k,1}, \dots, R_{k,B_0}$.

Step 4 Compute $R_k^0 = \frac{1}{B_0} \sum_{b=1}^{B_0} R_{k,b}$, $d_k = R_k - R_k^0$, and $p_k = \frac{1}{B_0} \#\{R_{k,b} \geq R_k : 1 \leq b \leq B_0\}$.

Define $K^- = \{2 \leq k \leq M : p_k \leq p_{max}, d_k \geq d_{min}\}$, where p_{max} and d_{min} are preset thresholds. If K^- is empty, take $\hat{K} = 1$. Otherwise, take $\hat{K} = \text{argmax}_{\{k \in K^-\}} d_k$.

5 Simulation Study: Part II

Second part of our experiments was conducted to examine the behaviour of Clest compared with other methods for estimating the optimal number of clusters. Besides Scenarios 1 and 2, where $K = 4$ clusters lie behind data, two new scenarios under $H_0 : K = 1$ (“no clusters”) are considered. Specifically, 50 realizations of length $T = 100$ were generated from each of the following processes:

- **Scenario 3:** $X_t = (0.3 - 10 \exp(-X_{t-1}^2)) X_{t-1} + \epsilon_t$, with ϵ_t iid $\mathcal{N}(0, 1)$.
- **Scenario 4:** $X_t = \mu_t + a_t$, with $\mu_t \sim \text{MA}(1)$ and $a_t = \sigma_t \epsilon_t$, where $\sigma_t^2 = 0.1 + 0.8a_{t-1}^2$ and ϵ_t iid $\mathcal{N}(0, 1)$.

The parameters required by Clest were: $M = 7$; $B = B_0 = 25$; size of learning set, $2/3p$; $p_{max} = 0.05$, $d_{min} = 0.05$ and the agreement between partitions in Step 1.iv was the index of Fowlkes and Mallows. Besides Clest algorithm, five commonly used criteria were considered:

maximization of the average silhouette width and of the indexes proposed by Calinski–Harabasz and Krzanowski–Lai; minimization of the Hartigan index and the gap method proposed by Tibshirani *et al.* [19]. A brief review of these indexes can be seen in, e.g., [19]. Figure 1 illustrates the behaviour of the tested methods based on $N = 100$ trials. Under the alternative

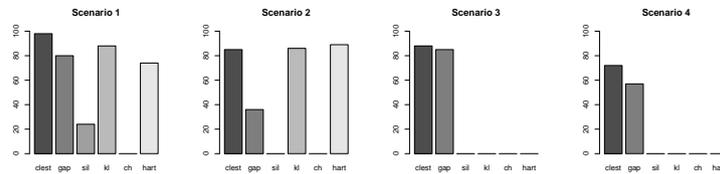


Figure 1: Percentage of trials where the number of clusters was correctly estimated.

hypothesis (Scenarios 1 and 2), the Clest procedure produced very good results, being the winner method in the nonlinear framework and clearly competitive together with the Krzanowski–Lai and Hartigan indexes in the heteroscedastic setup. Good results were also obtained by the gap method in Scenario 1, although this criterion performed poorly in Scenario 2. Graphs for Scenarios 3 and 4 show that only Clest and gap were able to detect the lack of clustering structure. In short, Clest algorithm has shown a good performance regardless of the considered scenario and only the gap procedure seems to show similar robustness. Krzanowski–Lai and Hartigan indexes worked well under alternative but clearly failed under the null. Silhouette and Calinski–Harabasz indexes do not work in any scenario.

6 A real data example

For illustrative purposes, the proposed metric was used to cluster dailies' returns of Euro exchanges rates against 28 international currencies (sample period: January 2009 -February 2014, $T = 1885$). Series in study can be adequately modeled by GARCH models and our clustering approach should work properly. This same example (with shorter observation period) was also considered by [4] to illustrate the merits of their fuzzy clustering approaches based on GARCH models. The dendrogram obtained with d_{QAF} and the complete linkage is shown in Figure 2.

Three clusters seem to be determined. One of them groups 18 Euro exchange rates against the major international currencies and those linked to the US dollar (US dollar -USD-, Canadian dollar -CAD-, Great Britain pound -GBP-, among others). The other two clusters are formed by 5 memberships. The cluster grouping {South African rand (ZAR), Russian rubel (RUB), Argentine peso (ARS), South Korean won (KRW) and Hong Kong dollar (HKD)} is the most heterogeneous by including Euro exchange rates against Asian, European, South American and African currencies.

The Clest algorithm was also executed and $\hat{K} = 3$ was obtained, which is according to the intuitive solution derived from the dendrogram.

7 Concluding remarks

In time series clustering, the identification of proper models is not *per se* the objective. The real challenge is to find out an effective dissimilarity measure to deal with different generating processes and detect structural similarities to form representative clusters. This is a central problem in many real applications and the main motivation behind this work. With this objective in mind, a metric based on quantile autocovariance functions is proposed. These functions

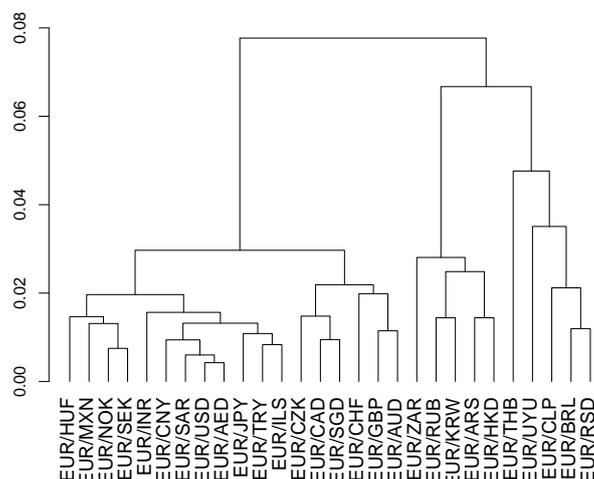


Figure 2: Complete linkage dendrogram based on d_{QAF} for the returns of the exchange rates.

account for important dynamic features of time series and are well-defined for a broad class of processes, including nonlinear and heteroscedastic processes. In particular, clustering of heteroscedastic models is still a little explored topic (see works on fuzzy clustering by [13, 4]). An extensive numerical study shows that the proposed dissimilarity produces excellent results in clustering regardless the kind of processes subjected to cluster. In complex scenarios including conditional heteroscedastic processes, our proposal clearly leads to the best results compared with alternative metrics introduced in the literature. Furthermore, our metric also outperforms metrics specifically designed to tackle nonlinear series, and, although not all results are presented here, it was highly competitive to classify linear models. In short, the quantile-autocovariance-based metric shows a very interesting robustness property with respect to the kind of processes and presents an efficient implementation at a very low cost in terms of computing time.

Estimation of the optimal number of clusters is also addressed in the present work. An adaptation of the Clest algorithm to cover the metric based on quantile autocovariances is proposed and promising results are observed in a broad simulation study, where the modified Clest is clearly the winner procedure compared with other classic alternatives.

Bibliography

- [1] Caiado J., Crato N. and Peña D. (2006) *A periodogram-based metric for time series classification* Computational Statistics and Data Analysis, **50**, 2668–2684.
- [2] Dudoit S. and Fridlyand J. (2002) *A prediction-based resampling method for estimating the number of clusters in a dataset*. Genome Biology, **3(7)**, 1–21.
- [3] D’Urso P. and Maharaj E.A. (2009) *Autocorrelation-based fuzzy clustering of time series*. Fuzzy Sets and System, **160**, 3565–3589.
- [4] D’Urso P., Cappelli C., Di Lallo D. and Massari R. (2013) *Clustering of financial time series*. Physica A: Statistical Mechanics and its Applications, **392(9)**, 2114–2129.

- [5] Fu T-ch. (2011) *A Review on time series data mining*. Engineering Applications of Artificial Intelligence, **24(1)**, 164–181.
- [6] Gavrilov M., Anguelov D., Indyk P. and Motwani R. (2000) *Mining the Stock Market: Which measure is best?* Proceedings of the 6th International Conference on Knowledge Discovery and Data Mining, 487–496.
- [7] Keogh E. and Kasetty S. (2003) *On the need for time series data mining benchmarks: A survey and empirical demonstration*. Data Mining and Knowledge Discovery, **7(4)**, 349–371.
- [8] Lee J. and Rao S.S. (2012) *The quantile spectral density and comparison based tests for nonlinear time series*. (arXiv:1112.2759v2). *ArXiv e-prints*.
- [9] Li T-H.(2014) *Quantile periodograms*. Journal of the American Statistical Association, **107(498)**, 765–776.
- [10] Liao T.W. (2005) *Clustering of time series data: A survey*. Pattern Recognition, **38(11)**, 1857–1874.
- [11] Linton O. and Whang Y-J. (2007) *The quantilogram: With an application to evaluating directional predictability*. Journal of Econometrics, **1**, 250–282.
- [12] Maharaj E.A. (1996) *A significance test for classifying ARMA models*. Journal of Statistical Computation and Simulation, **54**, 305–331.
- [13] Maharaj E.A., D’Urso P. and Galagedera, D.U.A. (2010) *Wavelet-based Fuzzy Clustering of Time Series*. Journal of Classification, (**27**), 231–275.
- [14] Maharaj E.A. and D’Urso P. (2011) *Fuzzy clustering of time series in the frequency domain*. Information Sciences, **181(7)**, 1187–1211.
- [15] Montero P. and Vilar J.A. (2014) *TSclust: Time series clustering utilities*. R package version 1.2.1 <http://CRAN.R-project.org/package=TSclust>.
- [16] Pérttega S. and Vilar J.A. (2010) *Comparing several parametric and nonparametric approaches to time series clustering: A simulation study*. J. of Classification, **27(3)**, 333–â362.
- [17] Piccolo D. (1990) *A distance measure for classifying ARIMA models*. Journal of Time Series Analysis, **11**, 153–164.
- [18] R Core Team (2014) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, <http://www.R-project.org>
- [19] Tibshirani R., Walther G. and Hastie T. (2001) *Estimating the number of clusters in a dataset via the Gap Statistic*. J. of the Royal Statistical Society. Ser. B, **63(2)**, 411–â423.
- [20] Vilar J.A. and Pérttega S. (2004) *Discriminant and cluster analysis for Gaussian stationary processes: local linear fitting approach*. Nonparametric Statistics, **16(3–4)**, 443–â462.

A Graphical User Interface Platform of the Stepwise Response Refinement Screener for Screening Experiments

Frederick Kin Hing Phoa, *Academia Sinica*, `fredphoa@stat.sinica.edu.tw`

Abstract. Supersaturated designs (SSDs) are useful in investigating a large number of factors with few experimental runs, particularly in screening experiments. The Stepwise Response Refinement Screener (SRRS) method is a new analysis introduced to screen important effects in the experiments using both a SSD and a general factorial design with the consideration of interactions. The cross-platform package **SRRS** is developed in R and the interface is built using the Tck/Tk bindings provided by the `tcltk` package included with R. The users are required to input the data and responses in the form of text files and the significant factors are suggested as an output. In addition, users are allowed to specify the threshold values, the selection criterion and whether the two-factor interactions are considered in the function setting panel.

Keywords. Supersaturated Design, Graphical User Interface, Screening Experiments, Model Selection

1 Introduction

As science and technology have advanced, scientific researchers and industrial practitioners are capable of studying large-scale systems. Typically the initial stage of these systems contain a large number of potentially important factors and interactions among these factors, but the probing and studying of a large-scale system is commonly expensive. Under the condition of factor sparsity, it might be useful to run experiments with fewer runs than there are factors to try to identify a small number of factors that appear to have dominant effects, and a supersaturated design (SSD) is suggested in such cases for run-size economy.

SSDs were first constructed in the discussion of the papers by [1] and [2]. [3] presented the first SSDs, but no more works were published until the papers by [4] and [5]. A comprehensive list of early works are referred in [6] and [7]. Traditionally, SSDs are employed primarily for screening main effects, discarding the possibility of interactions. Even the analysis considers main effects only, usual regression methods using all candidate factors cannot be used. Some

refined analysis methods were developed since [4] and a brief list of these works are provided in [8], [9], [10], [11], [12] and many others, and thus omitted here.

Recently, [8] introduced a new method, called the Stepwise Response Refinement Screener (SRRS), for analyzing the results of experiments using supersaturated designs. The method can further be extended in [13] to the experiments using a general factorial design, with the consideration of interaction effects. The SRRS method is a two-step procedure: Factor Screening and Model Searching. The first step aims at selecting a pool of potentially important effects from all factors in the experiments and the second step aims at searching the best model, under a given criterion, built among the selected effects in the first step.

Traditionally, Akaike information criterion (AIC) is used for model selection. For linear models, $AIC = n \log(RSS/n) + 2p$, where $RSS = \sum_{i=1}^n (y_i - \hat{y}_i)^2$ is the residual sum of squares, n is the number of runs, and p is the number of parameters in the model. It is known that AIC tends to overfit the model when the sample size is small. [14] imposed a heavy penalty on the model complexity and proposed a modified version of AIC for automatic variable selection procedure of the Dantzig selector (DS) method, $mAIC = n \log(RSS/n) + 2p^2$. The mAIC typically chooses a smaller model than AIC. This new criterion works well in both the DS method in [14] and the SRRS method in [8, 13].

To the best of our knowledge, there is no R package oriented to the variable selection problem in SSDs. Hence, we implement the SRRS method and introduce the package **SRRS** in this paper. The scope of **SRRS** is to allow any person with knowledge on variable selection and/or the SRRS method to start using **SRRS** for their everyday work without having to learn anything about the R syntax. The cross-platform package **SRRS** is developed in R for statistical computing and the graphical user interface (GUI) is built using the Tcl/Tk bindings provided by the `tcltk` packages included in R [15, 16]. The user only needs to type in the function name and the setting panel pops up in a window mode. After loading the data and selecting the required settings, the analysis is performed via the RUN button and the outputs are directly exported in the result panel when it is finished. The R package described in this paper is available from the Comprehensive R Archive Network at “<http://CRAN.R-project.org/package=SRRS>”.

The remainder of this paper proceeds as follows: the methodology of SRRS is reviewed in Section 2, the functions of **SRRS** is described in Section 3 and the examples on the analysis of real data are presented in Section 4. Finally, some concluding remarks and future possible extensions of the package are given in Section 5.

2 A review on Stepwise Response Refinement Screener (SRRS)

The SRRS, proposed in [8], is used for analyzing the experiments using SSDs. Consider a linear regression model $y = X\beta + \epsilon$, where y is an $n \times 1$ vector of observations, X is an $n \times k$ model matrix, β is a $k \times 1$ vector of unknown parameters, and ϵ is an $n \times 1$ vector of random errors. Assume that $\epsilon \sim N(\mathbf{0}, \sigma^2 I_n)$. In addition, X is assumed to be supersaturated, $n < k$. We let m be the number of potentially important effects (PIEs) and S_{inf} be the influential set of PIEs found in the process. It proceeds as follows.

Algorithm 2.1.

SRRS - Factor Screening

- Step 1.** Standardize data so that y_0 has mean 0 and the columns of X have equal lengths.
- Step 2.** Compute the correlation $\rho(X_i, y_0)$ for all factors X_i , $i = 1, \dots, k$.
- Step 3.** Choose E_0 such that $|\rho(E_0, y_0)| = \max_{X_i} |\rho(X_i, y_0)|$ and include E_0 as the first PIE in S_{inf} .
- Step 4.** Obtain the estimate β_{E_0} by regressing y_0 on E_0 .
- Step 5.** For the next m PIEs E_j , $j = 1, \dots, m$, $m < n - 2$,
- (a) compute the refined response $y_j = y_{j-1} - E_{j-1}\beta_{E_{j-1}}$;
 - (b) compute the marginal correlation $\rho(X_i, y_j)$ for all X_i , $i = 1, \dots, k$;
 - (c) choose T_j such that $|\rho(T_j, y_j)| = \max_{X_i} |\rho(X_i, y_j)|$;
 - (d) obtain the estimate β_{T_j} by regressing y_j on E_0, \dots, E_{j-1}, T_j ;
 - (e) if $|\beta_{T_j}| \geq \gamma$ and T_j has not been included in S_{inf} , put $E_j = T_j$ and include it in S_{inf} ;
 - (f) repeat (a) to (e) up to m^{th} step, where $E_j = E_m$ is not included in S_{inf} , m determined by either $m < n - 2$ or the threshold condition $|\beta_{T_j}| \geq \gamma$, or both.

SRRS - Model Searching

- Step 6.** Perform an all-subset search for all E_j , from models with one to m factors, where m is minimum of $n/3$ and the number of E_j in S_{inf} .
- Step 7.** Compute the objective function for each model and choose the final model as the one with optimal objective function; all E_j included in the final model are considered to be significant to the response y_0 .

Demonstrated in [13], SRRS can also be used to analyze the experiments using a general factorial design, with the consideration of interactions. Consider a nonregular FFDs with k_1 main effects and n runs, where $n < m$. There are $k_2 = k_1(k_1 + 1)/2$ interactions between two different main effects. If all two-factor interactions are considered together with all main effects, it is possible that $k_2 > m$, then the design matrix is supersaturated. Traditionally, the analysis of nonregular FFDs is based on two assumptions: the factor sparsity principle and the effect heredity principle. The first assumption has been embedded in the SRRS method, but the second assumption does not.

In order to implement the heredity principle into the SRRS method, some procedures are slightly modified: (1) Step 2: Due to the heredity principle, two-factor interactions are never be selected as the first PIE, so only the marginal correlations of all main effects are compared for selecting the first PIE. (2) Step 5: During the search of the j^{th} PIE, not all two-factor interactions are considered in the comparison of marginal correlation. According to the heredity principle, a two-factor interaction X_{ij} is considered in Step 5(b) if and only if either X_i or X_j or both parents main effects have been included in S_{Inf} in the previous searches. Therefore, the modifications in Step 5 take away a subset of two-factor interactions that none of their corresponding parent main effects have been PIEs. (3) Step 6: The reduced models built in this step must follow the heredity principle in order to avoid the situation that some significant two-factor interactions are included in the reduced model but none of their parent main effects have been included.

A detailed discussion on the main idea of each step in SRRS is out of the scope of this paper and we omit it here. Readers who are interested in these details are referred to [8].

3 SRRS interface and function

As is typical plug-ins, the SRRS package can either be loaded directly or by the command `library("SRRS")`.

Data files preparation. In practical applications, the user needs to prepare for two data files to run the program SRRS. The first file is a design matrix file that corresponds to the design matrix X in the SRRS method. Only main effect columns are required in this file because two-factors interaction effects are considered as an option in the procedure. The second file is a response file that corresponds to the response variable Y in the SRRS method. Only one column of response values is needed. Both data files should be saved as text files, preferably with `.txt` or `.dat` extensions. If either files contain headers, one should put them in the first row of the files.

SRRS input panel. To start the program, after loading the library, one only need to type `SRRS()` in the command prompt window, and a GUI panel like Figure 1 will pop up as a separate window. This panel is roughly divided into four parts from top to bottom. The first part is

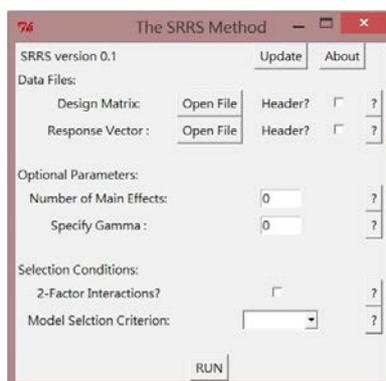


Figure 1: An empty SRRS input panel.

the information about the SRRS method. The current version is 0.1. The first button labeled “Update” brings the users to the official website of this program via a R command `browseURL`. It provides a convenient link for the users to update their SRRS library when a new version is available. The second button labeled “About” provides some basic information about SRRS, including the current version with date, authors’ names and affiliations (with maintenance personnel and email) and some legal claims. In addition, the question-mark buttons on the right of the panel provide short hints about the functions in the following three parts of the panels.

The second part of the panel aims at loading the data files into SRRS. Two buttons labeled “Open File” open the directory window for the users to select their prepared files. The checkboxes on the right of two “Open File” buttons are indicators if the files have headers or not. For example, if a user prepares a design matrix file with header names located in the first row, then the first checkbox should be clicked after loading the design matrix file in order to avoid program crash. Notice that if one wrongly clicks the checkbox for a file without headers, the program

will not crash but the analysis will be incorrect for the missing of the first row. However, if one forgets to click the checkbox for a file with headers, the program will crash because all entries of the matrix are not numeric anymore. In addition, if the design matrix file does not contain a row of header names, the program will automatically assign the names for all columns in the design matrix with the syntax “X1”, “X2”, etc.

The third part of the panel allows the users to enter some optimal parameters in **SRRS**. The first textbox is for the number of main effects. By default, if a user enters 0 or any number that is greater than the number of columns in the design matrix, the number of main effects is automatically set to be the number of columns in the design matrix. For example, assume that there are 8 columns in the design matrix, the users can provide optimal setting on the number of main effects by entering any integers from 1 to 8. Entering 0 or any numbers greater than 8 will be equivalent to entering 8 in the textbox. Furthermore, entering a negative number is equivalent to entering 0, and entering a number with decimal points is equivalent to entering that number without decimal points (i.e., entering 4.5 is equivalent to entering 4).

The second textbox is for the specification of γ . It is a tuning parameter in the factor screening procedure of the **SRRS** method, mainly for the termination of the screening when the magnitude of the potential important effect is too small when compared to noise. By default, if a user enter 0 in this checkbox, the suggestion of [8] is followed and γ will be automatically set to be 1/10 of the magnitude of the first potential important effect. Similar to the first checkbox, entering a negative number is equivalent to entering 0. However, it is different from the first checkbox that there is no upper limit to this checkbox, and a number with decimal point is allowed. Details on how γ is set appropriately are referred to [8].

The fourth part of the panel relates to the selection conditions in **SRRS**. The checkbox provides an option to users to consider two-factor interaction effects in the analysis. If there are two-factor interactions that have significant impact to the response but the analysis excludes them, serious biases to the estimates of main effects may lead to inconclusive results. [20] has an extensive discussion on this manner. If the user clicks this checkbox, **SRRS** will consider the significance of two-factor interactions under heredity principle. The combobox provides an option to users to choose which model selection criterion is used in the analysis. There are two choices, **mAIC** and **AIC** criteria, in the current version. **AIC** is a standard measure in many traditional analyses, but a problem of overfitting is observed when the sample size is small. **mAIC**, first proposed in [14], is a new measure with heavier penalty on the model dimensions.

SRRS result panel. Once the data loading and parameter setting are finished, the user clicks the “**RUN**” button to run the analysis via **SRRS**. A result panel will pop out after the analysis is done and a sample result panel is provided in Figure 5. The panel always includes five suggested models and their ranks are given in the first column. The second column provides the important effects in these five models and the syntax needs some explanations. For example, the second rank model in Figure 5 is “ $D + F + F : G$ ”. This means that the main effects D , F and the interaction effect FG are important and a model with these three effects are selected as the second rank model. The “+” sign are used to separate effects and the “:” sign indicates the interaction effect. The model is ranked via the user-selected criterion. In Figure 5, the model rank is determined by the **mAIC** criterion, and the smaller the **mAIC**, the better the model. The last row provides the threshold value γ used in the analysis. If the user enters γ as an optional parameter in the input panel, the number will be the same in the result panel. If the user skip the optional setting, the γ will be automatically determined in **SRRS** and reported here.

4 Some illustrations on real-data examples

A classic supersaturated design example. In this example, we apply SRRS to a classic supersaturated design example demonstrated by [4]. The original dataset has 24 factors but two factors (13 and 16) are identical. As [17], [14] and [8], we delete factor 13 and rename factors 14-24 as 13-23. The design matrix and response files are found in the SRRS package named `Lin_Dx.txt` and `Lin_Yx.txt` respectively.

Figure 2 is the input panel for this example. Since both data files do not have headers, we

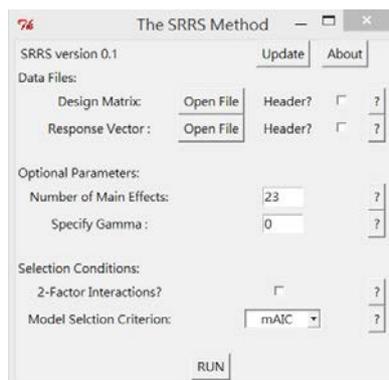


Figure 2: The input panel for the classic supersaturated design example.

do not click the checkboxes in the data files part. We consider all columns as main effects, so we enter “23” in the first textbox. Alternatively, we can leave the textbox unchanged (i.e., “0”) and SRRS still recognizes the number of main effects as “23”. We do not specify the threshold γ and let SRRS determine it automatically. Since the design matrix is supersaturated (number of factors exceeds number of runs), there is not enough degree of freedom to estimate interaction effects and we leave the corresponding checkbox unclicked. We choose mAIC as the model selection criterion. Finally, we click “RUN” to analyze this dataset.

Figure 3 is the result panel for this example. The analysis via SRRS suggests that a model

Rank	Model	mAIC
1	X14	105.7253
2	X4+X12+X14+X19	106.3701
3	X12+X14	106.8446
4	X12+X14+X19	107.0885
5	X14+X19	107.7655
gamma	5.3214	

Figure 3: The result panel for the classic supersaturated design example.

with factor 14 only has the minimum mAIC, and thus it is the best model among all models. The same analysis result were reported in several literature, see [14] and [8]. Some additional observations are highlighted in this result panel. First, the threshold γ is reported in the result panel and it is 5.3214, which is 10% of the magnitude of the first potential important effect. Second, since there is no headers in the design matrix, all factors are assigned with a generated header, namely “X1”, ..., “X23”. Finally, since we do not click the checkbox for two-factor interactions, none of these models include interaction effects. Readers who are interested in its mathematical analysis are referred to [8].

A classic factorial design example with consideration of interaction effects. In this example, we apply SRRS to the cast fatigue experiment, a real data set consisting of seven two-level factors. [18] demonstrated the difference in the analysis result of factorial experiment with and without considering the interaction effects. The result is confirmed in later literature, see [14]. The design matrix and response files are found in the SRRS package named `cast_Dh.txt` and `cast_Yh.txt` respectively.

Figure 4 is the input panel for this example. Opposite to the previous example, both

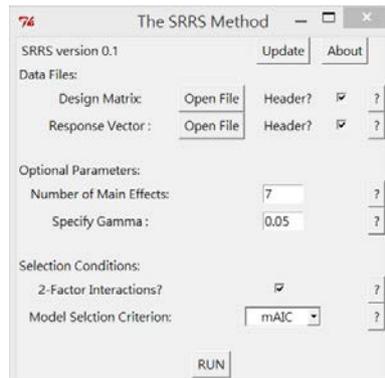


Figure 4: The input panel for the cast fatigue example.

data files have their header names in the first row, so we click the checkboxes in the data files part. Although the data matrix consists of 11 columns, which are the columns of a 12-run Plackett-Burman design, we only use the first 7 columns in this example. Therefore, we specify the number of main effects as “7” in the first textbox. In addition, assume that there is a prior information that the threshold is around 0.05, so we specify this value in the second textbox. We consider the possibility that there may exist some significant two-factor interactions, so we click the corresponding checkbox. We choose mAIC again as the model selection criterion. Finally, we click “RUN” to analyze this dataset.

Figure 5 is the result panel for this example. The analysis via SRRS suggests that a model

Rank	Model	mAIC
1	F+F:G	-27.8169
2	D+F+F:G	-21.2074
3	F+E:F+F:G	-19.7512
4	E+F:F:F:G	-19.5627
5	C+F:F:F:G	-18.0597
gamma	0.05	

Figure 5: The result panel for the cast fatigue example

with main effect F and interaction effect FG has the minimum mAIC, and thus it is the best model among all models. The same analysis result were reported in several literature, see [18], [14] and [13]. In this result panel, the threshold γ is exactly the same value as we entered in the input panel. In addition, interaction effects are considered under heredity principle in the model searching procedure, and some good models reported in the result table contain interaction effects. Readers who are interested in its mathematical analysis are referred to [13].

Another factorial design example with different analysis result. In this example, we apply SRRS to the HPLC experiment. [19] performed this experiment with 8 two-level factors via a 12-run Plackett-Burman design, but they consider main effects only and ignore the importance of interactions. The experiment is reanalyzed in [20], which a better model is suggested with a significant interaction effect. The design matrix and response files are found in the SRRS package named `HPLC_Dh.txt` and `HPLC_Yh.txt` respectively.

Figures 6 and 7 are the input and result panels for this example. In the input panel, we

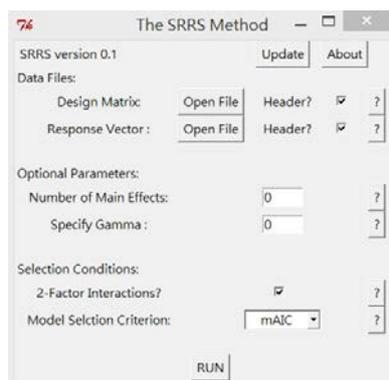


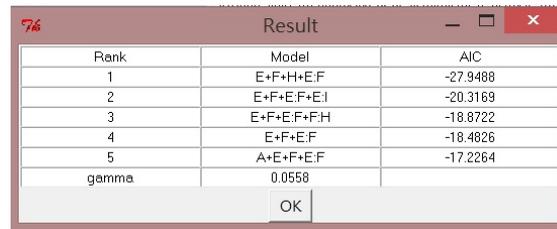
Figure 6: The input panel for the HPLC example.

Rank	Model	mAIC
1	E+F+EF	-6.4826
2	E+EF	-5.4846
3	E+F+H+EF	-3.9488
4	F+EF	-1.8045
5	E+EF+EI	-0.1999
gamma	0.0558	

Figure 7: The result panel for the HPLC example with mAIC criterion.

click the header checkboxes because both design matrix and response files contain headers. We do not need to enter optional parameters because all columns in the design matrix are main effects and we have no prior information on the threshold γ . We consider two-factor interaction effects and choose mAIC as the model selection criterion. In the result panel, the analysis via SRRS suggests that a model with main effects E , F and interaction effect EF has the minimum mAIC, and thus it is the best model among all models. The same analysis result was reported in [13].

However, the result is found to be different in [20], where the analysis is based on AIC criterion. Figure 8 is the result panel that AIC is chosen as the model selection criterion in the input panel, and all other settings remain unchanged. The analysis suggests the best model that includes an additional main effect H . The same model is ranked third when mAIC is the model selection criterion in Figure 7. The discrepancy comes from the different penalty on the additional dimension in the model. This example demonstrates that the result provided by SRRS is a suggestion and it is necessary to perform a follow-up experiment for further confirmation.



Rank	Model	AIC
1	E+F+H+E.F	-27.9488
2	E+F+E.F+E.I	-20.3169
3	E+F+E.F+F.H	-18.8722
4	E+F+E.F	-18.4826
5	A+E+F+E.F	-17.2264
gamma	0.0558	

Figure 8: The result panel for the HPLC example with AIC criterion.

5 Conclusion and future development

In this paper we have presented the R package **SRRS** for analyzing the experiments using SSDs and/or a factorial design with consideration of interactions. All features of the GUI in **SRRS** have been demonstrated through several real-data examples in Section 4. The latest version of **SRRS** can be downloaded at “<http://www.stat.sinica.edu.tw/fredphoa>”.

Further development in the **SRRS** method and its R package **SRRS** will focus on the following directions.

1. The current version of **SRRS** allows users to consider the potentially important two-factors interaction effects, but it is possible to have some higher-order interaction effects that have significant impacts to the response. A future version of **SRRS** will allow the search of potentially important higher-order interaction effects under the principle of effect heredity.
2. The current version of **SRRS** provides AIC and mAIC as two choices of model selection criteria. There are many other criteria that are commonly used in the model selection, like BIC, cAIC, Mallows’s C_p , etc. A future version of **SRRS** will provide additional choices on model selection criteria. Furthermore, it is desired to allow users to specify their own criterion as the objective function. In such case, the best possible way is to allow loading the criterion as a text file and read it in R environment.
3. The current version of **SRRS** utilizes the all-subset search in the model searching part. Although it is possibly the best method among all model searching methods, the complexity of this method is relatively higher than many other methods in the literature. For example, it is possible to substitute it via the DS method proposed in [14], which is known as an efficient linear programming method. A future version of **SRRS** will include a new choice on the search methods in the setting panel, which allows the users to select their preferred model searching method.

Appendix: Computational details

All computations and graphics in this paper have been obtained using the R version 3.0.0. Several utility packages have been created to help in the analyzing process. For examples, package **tc1tk2** [21] creates the input and result panels and package **gregmisc** [22] enumerates the possible combinations of columns from all available choices of main effects and/or interactions via its command **combinations**.

Acknowledgements

This work was supported by the National Science Council (Grant Number: 100-2118-M-001-002-MY2 and 102-2628-M-001-002-MY3) of Taiwan, Republic of China.

Bibliography

- [1] F.E. Satterthwaite (1959) *Random Balance Experimentation*. Technometrics, **1**, 111–137.
- [2] T.A. Budne (1959) *The Application of Random Balance Designs*. Technometrics, **1**, 139–155.
- [3] K.H.V. Booth and D.R. Cox (1962) *Some Systematic Supersaturated Designs*. Technometrics, **4**, 489–495.
- [4] D.K.J. Lin (1993) *A New Class of Supersaturated Designs*. Technometrics, **35**, 28–31.
- [5] C.F.J. Wu, (1993) *Construction of Supersaturated Designs through Partially Aliased Interactions*. Biometrika, **80**, 661–669.
- [6] Y. Liu and M.Q. Liu (2011) *Construction of Optimal Supersaturated Design with Large Number of Levels*. Journal of Statistical Planning and Inference, **141**, 2035–2043.
- [7] F. Sun, D.K.J. Lin and M.Q. Liu (2011) *On Construction of Optimal Mixed-Level Supersaturated Designs*. Annals of Statistics, **39**, 1310–1333.
- [8] F.K.H Phoa (2014) *Stepwise Response Refinement Screener (SRRS)*. Statistica Sinica, **24**, 197–210.
- [9] D. Slanzi and I.Poli (2014) *Evolutionary Bayesian Network Design for High Dimensional Experiments*. Chemometrics and Intelligent Laboratory Systems, **135**, 172–182.
- [10] P.R. Scintoa, R.G. Wilkinson, Z. Wanga and A. Rose (2014) *Comment: Need for Guidelines on Appropriate Screening Designs for Practitioners*. Technometrics, **56**, 23–24.
- [11] S.D. Georgiou (2014) *Supersaturated Designs: A Review of Their Construction and Analysis*. Journal of Statistical Planning and Inference, **144**, 92–109.
- [12] U.Das, S. Gupta and Shuva Gupta (2014) *Screening Active Factors in Supersaturated Designs*. Computational Statistics and Data Analysis, **77**, 223–232.
- [13] F.K.H Phoa (2013) *The Stepwise Response Refinement Screener (SRRS) and Its Applications to Analysis of Factorial Experiments*. Pattern Recognition - Applications and Methods, **204**, 13–22.
- [14] F.K.H. Phoa, Y.H. Pan and H. Xu (2009a) *Analysis of Supersaturated Designs via the Dantzig Selector*. Journal of Statistical Planning and Inference, **139**, 2362–2372.
- [15] P. Dalgaard (2001) *A Primer on the R-Tcl/TkPackage*. R News, **1/3**, 27–31.
- [16] P. Dalgaard (2002) *Changes to the R-Tcl/TkPackage*. R News, **2/3**, 25–27.

- [17] S.D. Beattie, D.K.H. Fong and D.K.J. Lin (2002) *A Two-Stage Bayesian Model Selection strategy for Supersaturated Designs*. *Technometrics*, **44**, 55–63.
- [18] C.F.J. Wu and M. Hamada (2000) *Experiments: Planning Analysis and Parameter Design Optimization*. Wiley, New York.
- [19] Y. Vander Heyden, M. Jimidar, E. Hund, N. Niemeijer, R. Peeters, J. Smeyers-Verbeke, D.L. Massart and J. Hoogmartens (1999) *Determination of System Suitability Limits with a Robustness Test*. *Journal of Chromography A*, **845**, 145–154.
- [20] F.K.H. Phoa, W.K. Wong and H. Xu (2009b) *The Need of Considering the Interactions in the Analysis of Screening Designs*. *Journal of Chemometrics*, **23**, 545–553.
- [21] P. Grosjean (2013) *tcltk2: Tcl/TkAdditions*. R-package version 1.2-5, url=["http://CRAN.R-project.org/package=tcltk2"](http://CRAN.R-project.org/package=tcltk2).
- [22] G.R. Warnes (2013) *gregmisc: Greg's Miscellaneous Functions*. R-package version 2.1.5, url=["http://CRAN.R-project.org/package=gregmisc"](http://CRAN.R-project.org/package=gregmisc).

Unravel: A Method and a Program to Analyze Contingency Tables, Unveiling Confounders.

Helmut Vorkauf, *Bern, retired*, helmut@vorkauf.ch

Abstract. An information theoretic approach to analyze multidimensional contingency tables to find the important relations between dependant and independent variables, uncovering confounding effects in a straightforward manner

Keywords. Multiple Associations, Multidimensional contingency tables, Confounding, Simpson's paradox, Effect size.

1 Method

When planning a study of cause and effect, one primarily selects an effect Y and a probable cause X , and then designs a study that lets one find out whether the presumed cause X actually has a significant influence on the effect Y . One is always aware that other variables might also have an effect on Y or X , therefore the study almost always includes measurements of further variables Z_i that might need to be controlled. In experimental designs one can control such further variables that might also have an effect through direct control or randomization, but in survey studies, observational by design, such a control becomes impossible.

In a paper [8] we introduced two measures of dependance based on information theory:

1. Gamma, the dependability γ_y or γ_x , is an unsymmetrical indicator of how reproducible or non-random the outcome of Y is, how unequivocally Y is determined by the other variables. This is a multivariate extension of Theil's uncertainty coefficient [6].
2. Zeta, the diagonality¹ or conciseness ζ , is a symmetrical coefficient of the closeness of relations between variables. This coefficient determines how free of slackness a logical link between variables is. It is defined for tables with any number of dimensions.

¹We had coined the name terseness for this coefficient when searching for a proper translation for the German *Prägnanz* and pregnancy seemed unfitting. Now the term diagonality seems more appropriate to us, indicating the accumulation of cases along the diagonal of a table with categories appropriately ordered. Conciseness might be an alternative term.

Both γ and ζ are normalized to 1, independent of the base of the logarithm and, especially, independent of the sample size N . That is why they are comparable for tables of different size and dimensionality, a quality that the usual measures do not achieve, especially not χ^2 . This comparability led us to choose them for an analysis of multivariate tables, where the many sub-tables of different size and dimensionality of a high-dimensional table have to be compared. An additional benefit stems from the clear quantitative meaning of the dependability². These qualities make the measures ideal instruments for quantifying strength of effect.

We should not be discouraged by the numerically small values we find working with dependabilities, we should not fall into the trap of interpreting them like e.g. correlation coefficients. The following informal but useful thresholds may serve as guidelines:

- 0.10** is enormous,
- 0.01** is very considerable,
- 0.001** and below may be regarded as negligible.

Thus we introduce a method of analyzing the dependence of Y on X , given the multitude of relations with and between the controlling variables Z_i . Failure to recognize a disturbing effect of such confounding relations may easily lead to erroneous conclusions concerning the primary purpose of a study. Our approach, we hope, may help to direct ones attention to problems that can occur in non-orthogonal designs when one or more of the control variables Z have disturbing effects and turn out to be aptly named *confounding*, i.e., they start to make the researcher confused.

2 Radelet's Data on the Death Penalty in Florida

A simple (in the sense of a low dimensionality only) problem is Radelet's study on Florida death penalties influenced by the race of the defendant when controlling for the victim's race (cited from [1]) as shown in table 1.

If the victim was white, black defendants received a death penalty more often than white defendants (22.9 % vs. 11.3 %), and this was also the case when the victim was black (2.8 % vs. 0). Yet, when collapsing the table by ignoring the victim's race (summing out the victim's race), in the total white defendants received the death penalty more often (11 % vs. 7.9 %).

The primary question is: "How strongly does the color of the defendant determine the penalty?", and we get two conflicting answers when we compare the total result with the result

²The clear quantitative meaning of the dependability γ shall be briefly demonstrated with artificial data:

	a	b	c	Total
A			20	
B		40		120
C	35			
D	25			
E	30	20	10	60
Total	90	60	30	

Row E with 60 (a third of all cases) was deliberately constructed to have the same distribution as the column totals. For this third of cases one can say nothing more about the column than was known beforehand from the column totals. In the other two thirds of the table the column is determined clearly without ambiguity. Consequently the dependability γ_y is $\frac{2}{3}$. What coefficient to describe the contingency has this nice quantitative interpretability?

Race of ...		Y=Penalty		Percent Death Penalty	
Z=Victim	X=Defendant	Death	No death		
White	White	53	414	11.3	↓
	Black	11	37	22.9	
Black	White	0	16	0.0	↓
	Black	4	139	2.8	
Total	White	53	430	11.0	↑
	Black	15	176	7.9	

Table 1: Frequencies of Death Penalties in Florida.

within victim’s race. The puzzling reversal of trend in the collapsed table is known as Simpson’s paradox. It is a phenomenon less known to statisticians trained in the analysis of experimental designs where all variables are orthogonal to each other.

Such orthogonality is rare, however, in surveys where Y and Z may vary freely, and in this case the free variation has led to an enormous linking of X and Z as shown in table 2. Black

		X=Defendant	
		White	Black
Z=Victim	White	467	48
	Black	16	143

Table 2: Frequencies of Victims and Defendants, $\gamma_{\text{Defendant}} = .4736$.

defendants tend to have killed black victims and white defendants tend to have killed white victims, and this non-orthogonality produces the baffling paradox.

A way to treat this annoying interdependence of X and Z is a technique (first introduced by Preuss [7]) which he called *uncoupling*³; the interdependence is eliminated by combining the cross-tabulated values of X and Z into a composite variable: instead of analyzing the effect of the 2 × 2 contingency table of X= race of defendant and Z= race of victim on Y= penalty, Preuss analyzes the effect of the composite variable [victim+defendant] with four values, namely [victim/defendant = W/W, W/B, B/W, B/B] on Y=penalty. This in effect removes any dependence between X and Z.

Our measure of diagonality ζ , which is $\zeta = .257726$ for the complete 2 × 2 × 2 table, is reduced to just $\zeta = .014110$ for the 2 × [2 × 2] table in which X and Z are uncoupled. The majority (95%) of the diagonality of the complete 2 × 2 × 2 table thus vanishes when the relation between X and Z is eliminated by uncoupling; this relation between X and Z has the enormous size of $\zeta = .328238$ (table 2).

We should revise our original question and ask: “How strongly is the sentence determined by the composite of victim’s and defendant’s race?”. The dependability γ_{sentence} of predicting the death sentence using the composite variable of the victim’s and the defendant’s race is .0505,

³Only now (March 2014), on re-reading Goodman and Kruskal[7], Preuss found that they had described the same technique on p.761, without following it through, however.

and this is the answer to the newly formulated question. We might go on to look at white and black defendants only and find that γ_{sentence} is a rather small .0113 for white defendants versus a strong .1612 for black defendants. The black defendant is the poor bugger, his sentence is strongly influenced by the race of his victim. This finding is rarely mentioned in published analyses.

It is our conviction that the summing out of the control variable Z =victim, in an effort to produce a summary, amounts to an illegal act that produces Simpson's paradox which leaves us confounded. In this extreme case, where summing out produced Simpson's paradox, you will probably agree, but we would like to propose a general rule banning the summing out of control variables, even when their effect seems marginal. The error involved in collapsing tables when an effect is insignificant, routinely done in parsing log-linear models, is only gradually less severe than when a very large X - Z -relationship produces Simpson's paradox.

3 Byssinosis, a higher-dimensional example

Years employed	Smoking	Gender	Race	Dustiness of Workplace								
				most dusty			medium dusty			least dusty		
				No	Yes	$p(\text{Yes})$	No	Yes	$p(\text{Yes})$	No	Yes	$p(\text{Yes})$
< 10 yrs	Yes	M	white	37	3	0.075	74	0	0.000	258	2	0.008
			other	139	25	0.152	88	0	0.000	242	3	0.012
	F	white	other	5	0	0.000	93	1	0.011	180	3	0.016
			other	22	2	0.083	145	2	0.014	260	3	0.011
	No	M	white	16	0	0.000	35	0	0.000	134	0	0.000
			other	75	6	0.074	47	1	0.021	122	1	0.008
F	white	other	4	0	0.000	54	1	0.018	169	2	0.012	
		other	24	1	0.040	142	3	0.021	301	4	0.013	
10 – 20 yrs	Yes	M	white	21	8	0.276	50	1	0.020	187	1	0.005
			other	30	8	0.211	5	0	0.000	33	0	0.000
	F	white	other	0	0	??	33	1	0.029	94	2	0.021
			other	0	0	??	4	0	0.000	3	0	0.000
	No	M	white	8	2	0.200	16	1	0.059	58	0	0.000
			other	9	1	0.100	0	0	??	7	0	0.000
F	white	other	0	0	??	30	0	0.000	90	1	0.011	
		other	0	0	??	4	0	0.000	4	0	0.000	
> 20 yrs	Yes	M	white	77	31	0.287	141	1	0.007	495	12	0.024
			other	31	10	0.244	1	0	0.000	45	0	0.000
	F	white	other	1	0	0.000	91	3	0.032	176	3	0.017
			other	1	0	0.000	0	0	??	2	0	0.000
	No	M	white	47	5	0.096	39	0	0.000	182	3	0.016
			other	15	3	0.167	1	0	0.000	23	0	0.000
F	white	other	2	0	0.000	187	3	0.016	340	2	0.006	
		other	0	0	??	2	0	0.000	3	0	0.000	

Table 3: Frequencies of Byssinosis by Length of Employment, Smoking, Gender and Race.

Let us now turn to a complex data set with six variables [4] as shown in table 3. The complete $3 \times 3 \times 2 \times 2 \times 2 \times 2$ table is difficult to read. When one tries to find the main factors leading to byssinosis, a lung disease caused by exposure to cotton dust, one has to take into account some very strong interrelations between the possibly illness-inducing variables. Higgins and Koch have devised a laborious χ^2 -based set of rules designed to find the important factors; they concluded that dustiness of the workplace is the most important determinant of illness,

gender of employee is next important, and smoking is in 3rd place. From the content of the study, it seems curious that the length of employment and therefore the length of exposure to dust came in 4th place only. Could it be that some disturbing X - Z -relation has suppressed the Z - Y -relation between length of exposition and byssinosis?

Our analysis starts by looking at dependability γ of byssinosis which is $\gamma = .2077$, a very strong reproducibility. Summing out of single variables results in varying degrees of loss of dependability. This loss of dependability of byssinosis we interpret as the importance of the variable summed out, it is its contribution to the dependability (table 4).

When summing out...	Dependability γ	% Loss by summing out
Dustiness of workplace	.0639	71
Length of employment	.1935	11
Smoking	.2011	8
Gender	.2046	6
Race	.2089	4

Table 4: Dependability when summing out Variables.

May we say that as a first impression we like the ordering of variables better than the ordering of Higgins and Koch? Exposure to dust, length of the exposure and smoking as the major determinants of a lung disease seem plausible to us, gender in 2nd place does not.

Next we investigate the full cross-tabulation, uncoupling in turn 15 pairs, 20 triples, 15 quadruples and 6 quintuples of variables, looking for effects of uncoupling on the diagonality of the full table. The diagonality ζ for the full table is .098369. The following partial table lists only pairs and triples of variables uncoupled that produced larger losses (table 5).

Uncoupled variables	Diagonality ζ	% Loss due to uncoupling
Race Employment	.049793	49
Dustiness Gender	.084665	14
Smoking Gender	.088207	10
Race Employment Dustiness	.040804	59
Race Employment Gender	.042130	57
Race Employment Smoking	.046762	52
Race Employment Byssinosis	.048688	51

Table 5: Diagonality when uncoupling Variables.

By far the strongest bivariate relation found in the data is the longer employment of white employees, nonwhites have a much higher turnover, $\gamma_{\text{Employment}} = .1664$ (table 6).

When uncoupling race and length of employment, this reduces ζ by 49 %. This reduction, due to the strong association of race and employment, repeats in the triples with larger losses.

The disproportionality has the effect that the clear increase of byssinosis with length of employment and thus exposure seen within race (table 7) is greatly reduced when race is summed out; $\gamma_{\text{Byssinosis}}$ is reduced to .0069 when summing out race, while it is .0191 for whites and .0255 for others (table 6).

		Race	
		white	other
Employment	< 10 years	1071	1658
	10 to 19 years	604	108
	20+ years	1841	137

Table 6: Frequencies Length of Employment by Race of Employee.

Race	Employment	Percentage of Byssinosis
White	< 10	1.12
	10 to 19	2.81
	20+	3.42
Other	< 10	3.08
	10 to 19	8.33
	20+	9.49
All Races	< 10	2.31
	10 to 19	3.65
	20+	3.84

Table 7: Summing out Race.

Here, the collapsing of the table by summing out race was not yet an error that produced a reversal of trend as in Simpson's paradox, but it is an error that led Higgins and Koch to underestimate the strong effect of length of exposure on developing a byssinosis. There is a continuum of degree of error that summing out may produce, and Simpson's paradox simply is a more severe error. We might say that race and employment in the Higgins and Koch data produced not an outright Simpson's paradox, but an attenuated one.

This error of summing out will affect any of the statistical models we usually apply in the analysis of data, as in the last resort they all use summaries of partially collapsed tables to arrive at their estimates of main effects. Fortunately, collapsing of tables by summing out minor variables is not needed; Preuss' method of uncoupling can successfully replace it without producing confounding results, as it does not discard data but merely rearranges them.

4 An Application to Two Meta-Analyses

A Chinese study on the effect of smoking on lung cancer by Liu [5], serves as the first example (table 8).

The γ for predicting cancer for the whole table reaches a sizeable .0267. Ignoring smoking by summing it out reduces γ to .0019 (a reduction by 93 %), so smoking is a very important determinant for cancer, whereas summing out the studies reduces γ to .0236, a loss of only 11%. Trying to gain oversight in this meta-analytic study seems not to be hampered by too much variation among the eight local studies. Summing out the eight localities to arrive at a 2×2 table linking smoking to cancer seems justified with not too much error.

Smokers		Non Smokers		Study
Cancer	No Cancer	Cancer	No Cancer	
126	100	35	61	Beijing
908	688	497	807	Shanghei
913	747	336	598	Shenyang
235	172	58	121	Nanjing
402	308	121	215	Harbin
182	156	72	98	Zhengzhou
60	99	11	43	Taiyuan
104	89	21	36	Nanchang

Table 8: Data of a Chinese Meta-Analysis of Smoking and Cancer.

Smokers		Non Smokers		Study
Cancer	No Cancer	Cancer	No Cancer	
83	72	3	14	01_D_Muller 1939
90	227	3	43	02_D_Schairer&Schoniger 1943
129	81	7	19	03_NL_Wassink 1948
70	397	12	125	04_US_Schrek 1950
412	299	32	131	05_US_Mills&Porter 1950
597	666	8	114	06_US_Wynder&Graham 1950
88	174	5	12	07_GB_McConnel 1952
1350	1296	7	61	08_GB_Doll&Hill 1952
60	106	3	27	09_US_Wynder&Cornfield 1953
459	534	18	81	10_US_Sadowsky 1953
724	246	4	54	11_SF_Koulumies 1953
499	462	19	56	12_US_Breslow 1954
451	1729	39	636	13_US_Levin 1954
260	259	5	28	14_US_Watson&Conte 1954

Table 9: Data of an International Meta-Analysis of Smoking and Cancer.

Dorn [2] compiled 14 international studies, published between 1939 and 1954, on the association of smoking and lung cancer (table 9). Our analysis reveals a picture differing from the Chinese data: γ for predicting cancer reaches an enormous .1189, and by summing out the studies factor, γ is reduced to .0431, a dramatic loss of 64 %. The diagonality of the $2 \times 2 \times 14$ table is $\zeta = .0319$; if we take out the cancer \times study relation by uncoupling we lose 50 % of this diagonality, half of the conciseness is due to the great differences in cancer prevalence of the 14 studies. Thus, in this collection of data, it seems to make no sense to gain a 2×2 table of cancer by smoking, which was the signal the study was aiming at. The signal is drowned in the noise of the far too different studies.

5 A program for the analysis

A program *Unravel* running under Windows is available that computes the separabilities with each variable in turn regarded as the dependant variable when each single variable or pair or higher dimensional tuples of variables is summed out. Likewise, the diagonalities are computed with pairs and tuples of variables uncoupled. In short, all tuples of variables are analyzed like in CFA, Lienert's Analysis of Configuration Frequencies (in fact, the program started with a CFA program I wrote decades ago). You may use the output for CFA interpretation, looking for types and antitypes. The main program is written in Delphi Pascal and expects input in the form of dBase tables. These can be either raw data of one case per row or contingency tables with one row per cell. You might want to use Excel data and save them as dBase, or you can enter dBase data directly with an additional routine written in FoxPro. A further addition, also written in FoxPro, is only needed for getting error estimates by bootstrap sampling.

Bibliography

- [1] Agresti, Alan (2002) *Categorical Data Analysis*.
- [2] Dorn, Harold F. (1954) *The Relationship of Cancer of the Lung and the Use of Tobacco*. The American Statistician, Vol 8, No 5
- [3] Goodman, L.A. and Kruskal, W.H. *Measures of Association for Cross Classifications*. JASA, December 1954, **49**, 732–764.
- [4] Higgins, J.E. and Koch, G.G. (1977) *Variable Selection and Generalized Chi-Square Analysis of Categorical Data applied to a Large Cross-Sectional Occupational Health Survey*. International Statistical Review, **45**, 51–62
- [5] Liu, Z. (1992) *Smoking and Lung Cancer in China: combined analysis of eight case-control studies*. Internat. J. Epidemiol., 21(2), 197–201
- [6] Press, William H., Flannery, Brian P., Teukolsky, Saul A., Vetterling, William T. (1992) *Numerical Recipes: the Art of Scientific Computing* 3rd ed., Cambridge University Press, p. 761.
- [7] Preuss, Lucien (1980) *A class of statistics based on the information concept*. Communications in Statistics – Theory and Methods, Volume 9, Issue 15
- [8] Preuss, Lucien and Vorkauf, Helmut (1997) *The Knowledge Content of Statistical Data*. Psychometrika, Vol 62, No 1, 133–161

Preventive maintenance in a complex warm standby system. A transient analysis

Juan Eloy Ruiz-Castro, *University of Granada*, jeloy@ugr.es

Abstract. Preventive maintenance plays an important role in the reliability field. Fatal failures with the corresponding damage associated can be avoided by considering preventive maintenance. A complex warm standby system that evolves in discrete time is modeled in transient regime. The system is composed of one online unit and the rest in warm standby. All units can undergo repairable failures due to wear. Besides, the online unit is subject to external shocks, which can produce a repairable failure. If any unit suffers a repairable failure, this one goes to the repair facility for corrective repair. The corrective repair time depends on the type of failure (online or warm standby unit). Preventive maintenance is introduced as response to random inspections over the online unit. When one inspection occurs, two possible degradation levels of the online unit can be observed: minor or major. In latter case preventive maintenance is carried out. The system is modeled and some interesting measures such as reliability, availability and some conditional probability of failure or preventive maintenance are worked out in transient regime. The modeling and the measures have been calculated in an algorithmic form through matrix algebraic expressions. The results have been implemented computationally with Matlab.

Keywords. Preventive maintenance, warm standby system, Phase type distribution, Matlab

1 Introduction

Redundancy and preventive maintenance are methods that are widely applied to improve reliability and availability in system design. They are necessary in order to improve overall reliability, prevent system failures and reduce costs. Classical texts on reliability have examined such techniques. Recently, valuable contributions have been made to maintenance policies in the area of reliability theory. Nakagawa (2005) [2] considered standard and advanced problems of maintenance policies for system reliability models, analysing topics such as repair, age, block and periodic replacement and preventive maintenance. Preventive maintenance is of special interest

when system degradation and non-repairable failures, due to wear and/or external shocks, are present.

One of the main problems encountered when analyzing standby systems with three or more units is that it may not be feasible to build the model and its associated measures. It is desirable to model complex reliability systems in an algorithmic and well structured form. If phase-type distributions are assumed for the embedded times in the system, then this objective is reached. This class of distribution was introduced by Neuts (1981) [3] and has been applied in fields such as queuing and reliability theory. Recently, PH distributions and block-structured stochastic models are analysed in He (2014) [1]. This class of distributions has been considered to model complex redundant systems ([5], [6]) and systems with preventive maintenance ([4]) in an algorithmic form. When PH distributions are considered in modeling reliability systems, their construction and the measures associated with the system are achieved in an algorithmic matrix form which simplifies the computational implementation.

The main objective of this paper is modelling and analysing, in an algorithmic form, the effect of preventive maintenance introduced in a complex warm standby system, with an indeterminate number of units, where the online unit can go through degradation levels. The online unit is subject to internal failures and external shocks that produce failure. Any warm standby unit can fail at any time with probability p . All failures are repairable and only one repairperson is assumed in the repair facility. We assume that the events which produce the external shocks occur statistically independently of the performance of the device, and that they may occur even if the device is currently under repair. Two types of repair are considered: corrective repair, which is carried out when repairable failures occur, and preventive maintenance. The preventive maintenance is performed in response to random inspections. When an inspection takes place, if any internal damage is observed, the device is sent for preventive maintenance, unless the damage is trivial.

The paper is organized as follows. The system is described in Section 2 and it is modelled through a vector Markov process in Section 3. The transient distribution is built in an algorithmic form in Section 4. Some interesting reliability measures, such as availability, reliability and conditional probability of failures are developed in Section 5. A numerical example shows the versatility of the model in Section 6. The results have been implemented computationally with Matlab.

2 The warm standby system

We assume a warm standby system that evolves in discrete time. There are K units, the online one and the rest disposed in warm standby. All units are subject to repairable failures, but the online unit is also subject to external shocks, which can produce a repairable failure. Each time that the online unit undergoes a repairable failure, this one goes to the repair facility for corrective repairing and one warm standby, if any, unit occupies the online place. There is a repairperson. The corrective repair time is different depending on the failure is from the online unit or from a warm standby. The internal behavior of the online unit one can go through two degradation levels: minor and major. While there is at least one operational unit, a random inspection can occur. When it happens, the degradation level of the online unit is observed. If this degradation level is major, the unit goes to the repair facility for preventive maintenance only if there is at least one warm standby. In other case, the unit continues working. The

following assumptions are considered.

Assumption 1. The internal operational time of the online unit is PH distributed with representation (α, T) with order n . The states are partitioned in two degradation levels: minor and major. The minor degradation level is composed of the states $1, \dots, n_1$, and the major is composed of the rest of the states.

Assumption 2. The accidental failure of the online unit is produced by external shocks. The external shocks happen according to a PH renewal process where the time between two consecutive shocks is PH distributed with representation (γ, L) with order t .

Assumption 3. The time between two consecutive inspections of the online unit is PH distributed with representation (η, M) with order ϵ .

Assumption 4. Each warm standby unit fails at any time with probability p .

Assumption 5. There are two different corrective repairs depending on the type of failure: internal failure of the online unit and the failure of a warm standby unit. The corrective repair distribution is PH in both cases with representations (β_0, S_0) with order z_0 and (β_2, S_2) with order z_2 respectively.

Assumption 6. If one inspection is produced and the degradation level of the online unit is major, then the unit goes to preventive maintenance if there is at least unit in warm standby. The preventive maintenance distribution is also PH distributed with representation (β_1, S_1) with order z_1 .

Assumption 7. Preventive maintenance and any type of failure of the online unit have preference among the warm standby failures at same time for the repair facility.

State space

The system can pass through $K + 1$ macro-states. The macro-state space is given by $S = \{S_0, S_1, \dots, S_K\}$, where S_l contains the phases when there are l units in repair for $l = 0, \dots, K$. If we denote as $\theta_\nu = 0, 1, 2$ to the kind of failure of the ν -th unit in the repair facility (0: internal failure, 1: preventive maintenance, 2: warm standby failure), then the phases of these macro-states are given by

$$S_0 = \{(i, j, s); 1 \leq i \leq n, 1 \leq j \leq t, 1 \leq s \leq \epsilon\},$$

$$S_l = \{E_{\theta_1, \theta_2, \dots, \theta_l}; \theta_\nu = 0, 1, 2, \nu = 1, \dots, l, l = 1, \dots, K, \text{ where}$$

$$E_{\theta_1, \theta_2, \dots, \theta_l} = \{(i, j, s, r); 1 \leq i \leq n, 1 \leq j \leq t, 1 \leq s \leq \epsilon, 1 \leq r \leq z_{\theta_1}\}, l = 1, \dots, K - 1, \text{ and}$$

$$E_{\theta_1, \theta_2, \dots, \theta_K} = \{(j, r); 1 \leq j \leq t, 1 \leq r \leq z_{\theta_1}\}.$$

The phases of the macro-states can be interpreted in the following way. For instance, the macro-state S_l contains the phases when there are l units in the repair facility. The composition of the repair queue is given by $E_{\theta_1, \theta_2, \dots, \theta_l}$. It indicates that the unit that is being repaired is type θ_1 , and the units in queue are types $\theta_2, \dots, \theta_l$ (in order). If $l = 1, \dots, K - 1$, then this macro-state, $E_{\theta_1, \theta_2, \dots, \theta_l}$, contains the phases (i, j, s, r) where i indicates the state of the internal operational time, j the state of the external shock time, s the state of the inspection time and r the state of the repair time. Throughout the paper, given a matrix A , the column vector A^0 is defined as $A^0 = e - Ae$ where e is a column vector of ones with appropriate order.

3 The model

The system described above is modeled through a vector Markov process with state space S . The transition probability matrix is given by

$$P = \begin{pmatrix} B_{00} & B_{01} & B_{02} & \dots & B_{0,K-2} & B_{0,K-1} & B_{0,K} \\ B_{10} & B_{11} & B_{12} & \dots & B_{1,K-2} & B_{1,K-1} & B_{1,K} \\ \mathbf{0} & B_{21} & B_{22} & \dots & B_{2,K-2} & B_{2,K-1} & B_{2,K} \\ \vdots & \ddots & \ddots & \ddots & \vdots & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} & \ddots & B_{K-2,K-3} & B_{K-2,K-2} & B_{K-2,K-1} & B_{K-2,K} \\ \mathbf{0} & \mathbf{0} & \dots & \mathbf{0} & B_{K-1,K-2} & B_{K-1,K-1} & B_{K-1,K} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \dots & \mathbf{0} & B_{K,K-1} & B_{K,K} \end{pmatrix},$$

where the block B_{ij} contains the transition probabilities between the macro-states $S_i \rightarrow S_j$. For instance, the matrix blocks from the macro-state S_0 are shown.

Block B_{00}

This block contains the transition probabilities from the macro-states $S_0 \rightarrow S_0$. All units are working and at next time all units continues working. It occurs because the internal operational time, the external shock time and the inspection time change of state, and then they are not taken place ($T \otimes L \otimes M$). Also, one inspection can occur and it observes a minor degradation level, the external shock time follows changing of state ($U_1 T \otimes L \otimes M^0 \eta$). The auxiliary matrix U_1 is given by

$$U_1(i, j) = \begin{cases} 1 & ; \quad i = j \leq n_1 \\ 0 & ; \quad otherwise \end{cases}$$

Respect to the warm standby, none of them fails and it occurs with probability $(1 - p)^{K-1}$. Therefore,

$$B_{00} = (1 - p)^{K-1} T \otimes L \otimes M + U_1 T \otimes L \otimes M^0 \eta.$$

Block B_{01}

This block contains the transition probabilities from the macro-states $S_0 \rightarrow S_1$. All units are operational and one failure or preventive maintenance occurs. This failure can be due to an internal operational failure, to an external shock or to preventive maintenance of the online unit after inspection; or one warm standby unit fails. This matrix can be partitioned depending on the type of failure (0: failure of the online unit, 1: preventive maintenance, 2: failure of a warm standby unit) as $B_{01} = \{B_{01}(0), B_{01}(1), B_{01}(2)\}$,

$$\begin{aligned} B_{01}(0) &= (1 - p)^{K-1} [T^0 \alpha \otimes (L + L^0 \gamma) \otimes (M^0 \eta + M) \\ &\quad + (e - T^0) * \alpha \otimes L^0 \gamma \otimes (M^0 \eta + M)] \otimes \beta^0, \\ B_{01}(1) &= (1 - p)^{K-1} U_2 (e - T^0) \alpha \otimes L \otimes M^0 \eta \otimes \beta^1, \\ B_{01}(2) &= (K - 1) p (1 - p)^{K-2} [T \otimes L \otimes M + U_1 T \otimes L \gamma \otimes M^0 \eta] \otimes \beta^2, \end{aligned}$$

being $U_2(i, j) = \begin{cases} 1 & ; \quad i = j \geq n_1 + 1 \\ 0 & ; \quad otherwise. \end{cases}$

Blocks B_{0l} for $l = 2, \dots, K$

This block contains the transition probabilities from the macro-states $S_0 \rightarrow S_l$. All units are operational and l failures (or one preventive maintenance) occur. These failures can be due to an internal operational failure, to an external shock or to preventive maintenance of the online unit after inspection and $l - 1$ warm standby fail; or l warm standby unit fails. This matrix can be partitioned depending on the transitions between the macro-states $S_0 \rightarrow E_{\theta_1, \theta_2, \dots, \theta_l}$ in alphabetic order. The blocks different to zero are the following ones.

$$B_{0l}(0, 2, 2, \dots, 2) = \binom{K-1}{l-1} p^{l-1} (1-p)^{K-l} [T^0 \alpha \otimes (L + L^0 \gamma) \otimes (M^0 \eta + M) + (e - T^0) \alpha \otimes L^0 \gamma \otimes (M^0 \eta + M)] \otimes \beta^0,$$

$$B_{0l}(1, 2, 2, \dots, 2) = \binom{K-1}{l-1} p^{l-1} (1-p)^{K-l} U_2 (e - T^0) \alpha \otimes L \otimes M^0 \eta \otimes \beta^1,$$

$$B_{0l}(2, 2, 2, \dots, 2) = \binom{K-1}{l} p^l (1-p)^{K-l-1} [T \otimes L \otimes M + U_1 T \otimes L \otimes M^0 \eta + I_{\{l=K-1\}} p^{K-1} U_2 T \alpha \otimes L \otimes M^0 \eta] \otimes \beta^2,$$

where the function I is the indicator function. Finally,

$$B_{0K}(0, 2, 2, \dots, 2) = p^{K-1} [T^0 \otimes (L + L^0 \gamma) \otimes e + (e - T^0) \otimes L^0 \gamma \otimes e] \otimes \beta^0.$$

4 Transient distribution

The transient distribution is obtained from the transition probability matrix by considering the matrix blocks. The probability of occupying the different phases at time ν is worked out by blocks as $P^{(\nu)} = P^\nu = (B_{ij}^{(\nu)})_{i,j=0,\dots,K}$. The blocks have been expressed in a recursive form as

$$B_{ij}^{(1)} = B_{ij},$$

$$B_{ij}^{(\nu)} = \sum_{k=0}^{\min\{2+j,K\}} B_{ik}^{(\nu-1)} B_{kj}; \nu \geq 2.$$

This recursive method can be developed obtaining that

$$B_{ij}^{(1)} = B_{ij},$$

$$B_{ij}^{(\nu)} = \sum_{k_{\nu-1}=0}^{\min\{2+j,K\}} \sum_{k_{\nu-2}=0}^{\min\{2+k_{\nu-1},K\}} \dots \sum_{k_1=0}^{\min\{2+k_2,K\}} B_{i,k_1} B_{k_1,k_2} \dots B_{k_{\nu-2},k_{\nu-1}} B_{k_{\nu-1},j}.$$

The initial distribution is also expressed by considering the macro-states defined. Let ω_l be the initial probability vector of having l units in the repair facility. Then, the initial distribution is $\omega = (\omega_0, \omega_1, \dots, \omega_K)$.

Therefore, the probability that the system occupies the corresponding phases of the macro-state l (l units broken) at time ν is

$$p_l^\nu = \sum_{i=0}^K \omega_i B_{il}^{(\nu)}.$$

If the system is new initially then $\omega_0 = \alpha \otimes \gamma \otimes \eta$, $\omega_l = 0$ for $l = 0, \dots, K$. In this case

$$p_l^\nu = \omega_0 B_{0l}^{(\nu)}.$$

5 Transient Reliability Measures

Several measures of interest associated to the system are shown in this section in an algorithmic form.

Availability

The availability is the probability that at time ν the system is operational (at least one unit is working). This probability is equal to

$$A(\nu) = 1 - p_K^{(\nu)} e = 1 - \omega_0 B_{0K}^{(\nu)} e.$$

Times up to a determinate macro-state

In this section the first visit time distribution for a determinate macro-state is calculated. We define P_{-j} and ω_{-j} to the matrix P and the vector ω without the probabilities corresponding to the phases of the macro-state j . The first visit time up to macro-state j distribution follows a phase-type distribution with representation (ω_{-j}, P_{-j}) .

The mean time up to first time that the system visits the macro-state j is given by

$$-\omega_{-j} (I - P)^{-1} e.$$

Therefore, if the reliability function is defined as the time up to first time that the system has all units in the repair facility, it is given by

$$R(\nu) = \omega_{-K} (I - P_{-K})^{-1} P_{-K}^\nu P_{-K}^0.$$

Conditional Probability of Failure

Three different conditional probabilities of failure are defined in this section depending on the types of failures and preventive maintenance.

- a The system is in macro-state l and only h warm standby units fail

We assume that the device is working with l units in the repair facility at time $\nu - 1$, and h warm standby units fail at next time, $h = 1, \dots, K - l - 1$. This probability is equal to

$$\begin{aligned} \psi_{2,l,h}(\nu) &= \binom{K-l-1}{h} p^h (1-p)^{K-l-1-h} \\ &\cdot p_l^{(\nu-1)} \left[(e - T^0) \otimes (e - L^0) \otimes (e - M^0) + U_1 (e - T^0) \otimes (e - L^0) \otimes M^0 \right] \otimes e. \end{aligned}$$

- b The system is in macro-state l and the online undergoes an internal failure and h warm standby units fail

We assume that the device is working with l units in the repair facility at time $\nu - 1$, and the online unit undergoes an internal failure and h warm standby units fail at next time, $h = 0, \dots, K - l - 1$. This probability is

$$\psi_{in,l,h}(\nu) = \binom{K-l-1}{h} p^h (1-p)^{K-l-1-h} p_l^{(\nu-1)} [T^0 \otimes (e - L^0) \otimes e].$$

- c The system is in macro-state l and the online undergoes an accidental failure and h warm standby units fail

We assume that the device is working with l units in the repair facility at time $\nu - 1$, and the online unit undergoes an accidental failure and h warm standby units fail at next time, $h = 0, \dots, K - l - 1$. It is given by

$$\psi_{acc,l,h}(\nu) = \binom{K-l-1}{h} p^h (1-p)^{K-l-1-h} p_l^{(\nu-1)} [(e - T^0) \otimes L^0 \otimes e].$$

- d The system is in macro-state l and the online undergoes an internal or external accidental failure and h warm standby units fail

$$\psi_{0,l,h}(\nu) = \psi_{in,l,h}(\nu) + \psi_{acc,l,h}(\nu) + \binom{K-l-1}{h} p^h (1-p)^{K-l-1-h} p_l^{(\nu-1)} [T^0 \otimes L^0 \otimes e].$$

- e The system is in macro-state l and the online undergoes a preventive maintenance and h warm standby units fail

Finally, while the unit is working on macro-state l at time $\nu - 1$, one inspection occurs by observing major degradation level at next time, h warm standby units fail at same time. The probability of occurrence is given by

$$\psi_{1,l,h}(\nu) = \binom{K-l-1}{h} p^h (1-p)^{K-l-1-h} p_l^{(\nu-1)} [U_2(e - T^0) \otimes (e - L^0) \otimes M^0 \otimes e].$$

6 A numerical example

A numerical example shows the versatility of the model. The measures described throughout the paper have been implemented computationally with Matlab and they have been applied for analyzing the behavior of a system with and without preventive maintenance. Any warm standby can fail with probability $p = 0.001$ and the embedded times in the model are PH distributed. The Table 1 shows the corresponding representations. Some transient measures for both systems, with and without preventive maintenance, are shown in Table 2.

Acknowledgement

This paper is partially supported by the Junta de Andalucía, Spain, under the grant FQM-307 and by the Ministerio de Ciencia e Innovación, España, under Grant MTM2010-20502.

Internal operational time	External shock time	Inspection time
$\alpha = (1, 0, 0)$	$\gamma = (1, 0)$	$\eta = (1, 0)$
$T = \begin{pmatrix} 0.994 & 0.002 & 0.0025 \\ 0 & 0.999 & 0.0002 \\ 0 & 0 & 0.998 \end{pmatrix}$	$L = \begin{pmatrix} 0.95 & 0.03 \\ 0.98 & 0.01 \end{pmatrix}$	$M = \begin{pmatrix} 0.65 & 0.14 \\ 0.44 & 0.41 \end{pmatrix}$
Corrective repair time	Preventive maintenance time	Warm standby repair time
$\beta_0 = (1, 0)$	$\beta_1 = (1, 0)$	$\beta_2 = (1, 0)$
$S_0 = \begin{pmatrix} 0.5 & 0.45 \\ 0.8 & 0.15 \end{pmatrix}$	$S_1 = \begin{pmatrix} 0.1 & 0.06 \\ 0.05 & 0.10 \end{pmatrix}$	$S_2 = \begin{pmatrix} 0.6 & 0.2 \\ 0.3 & 0.65 \end{pmatrix}$

Table 1: Embedded time distributions

ν	CPCR		MTCR		OMT	
	PM	No-PM	PM	No-PM	PM	No-PM
10	0.0212	0.0211	0.9532	0.9533	10.9975	10.9975
50	0.0207	0.0206	12.1341	12.1425	50.5504	50.5727
100	0.0203	0.0202	30.6766	30.7204	98.9240	99.0241
200	0.0202	0.0201	70.4531	70.6266	194.5930	194.8694
1000	0.0202	0.0201	335.7359	394.4028	958.1201	959.7753

Table 2: Conditional probability of corrective repair (CPCR), mean time working corrective repair (MTCR) and operational mean time (OMT), up to a certain time ν

Bibliography

- [1] He, Q.M. (2014) *Fundamentals of Matrix-Analytic Methods*. Springer Science+Business Media New York.
- [2] Nakagawa, T. (2005) *Maintenance Theory of Reliability*. Springer Series in Reliability Engineering series. Springer-Verlag London Limited.
- [3] Neuts, M.F. (1981) *Matrix-Geometric Solutions in Stochastic Models*. An Algorithmic Approach, Baltimore, MD: John Hopkins University Press.
- [4] Ruiz-Castro, J.E. (2013) *A preventive maintenance policy for a standby system subject to internal failures and external shocks with loss of units*. International Journal of Systems Science, (in press) DOI: 10.1080/00207721.2013.827258.
- [5] Ruiz-Castro, J.E. and Fernández-Villodre, G. (2012) *A complex discrete warm standby system with loss of units*. European Journal of Operational Research, **218**, 2, 456–469.
- [6] Ruiz Castro, J.E. and Li, Q.L. (2011) *Algorithm for a general discrete k-out-of-n: G system subject to several types of failure with an indefinite number of repairpersons*. European Journal of Operational Research, **211**, 1, 97-111.

Linear Regression Models Using L_1 , L_2 and L_∞ -Norms

Pranesh Kumar, *Department of Mathematics and Statistics, University of Northern British Columbia, Prince George, BC, V2N 4Z9, Canada, Pranesh.Kumar@unbc.ca*

Faramarz Kashanchi, *Department of Mathematics and Statistics, University of Northern British Columbia, Prince George, BC, V2N 4Z9, Canada, Faramarz.Kashanchi@northernhealth.ca*

Abstract. The L_2 - norm based linear regression or the least-squares estimation (LSE) models often perform relatively well under conditions such as the model errors follow normal or approximately normal distributions, are free of large size outliers and satisfy the Gauss-Markov assumptions. Under these conditions, LSE is optimal and provides the best linear unbiased estimators of the linear regression model parameters. However, there are often situations wherein the LSE based linear regression may not meet one or others of these assumptions and hence fails to be optimal. We have considered for some experimental data sets the L_1 , L_2 and L_∞ -norm estimation based linear models and noted that the LSE based models do not always perform best. We discuss results of the L_p -norm based estimations by describing types of data sets varying in size and probability distributions, model fit, residual analysis and residual plots.

Keywords. Linear models, Least-squares estimation, L_p -norm estimation, Prediction, Forecasting.

1 Introduction

The least-squares estimation (*LSE*) technique, first published by Legendre in 1805, is used for estimating the linear regression models. The linear regression models based on *LSE* technique perform well provided the errors follow a normal or approximately normal distribution, do not possess large size outliers and follow Gauss-Markov assumptions. Under these conditions, the *LSE* is optimal and provides the best linear unbiased estimators of the model parameters. A number of alternatives to the *LSE* which are more robust to departures from the usual least squares assumptions have been studied [Gauss (1809,1820), Laplace (1812), Stigler (1990), Farebrother (1999)]. In this paper, we have investigated the performance of the L_1 , L_2 and L_∞ - linear regression models. We consider experimental data from the applications which are of small to large size and follow different types of bivariate probability distributions. We have evaluated fitted models using error based measures and residual analysis. Numerical calculations are carried out using Matlab codes.

2 The L_p -norm Linear Regression Models

Suppose that data be available on n cases, y_i is the observed response and x_{i1}, \dots, x_{ik} are the values of k independent variables of the i th case. The values of k independent variables are treated as fixed constants, however, responses are subjected to variation. A general linear regression model for a single response variable y given k independent variables is

$$y = X\beta + \epsilon, \quad (1)$$

where y is the vector of n response observations, X is the $n \times k$ matrix of values of k independent variables, β is the vector of $k+1$ model parameters and ϵ is the vector of n residual values. Residuals ϵ in the model are assumed to follow a multivariate normal distribution, i.e., $\epsilon \sim N(0, \sigma^2 I)$. Similarly, $y \sim N(X\beta, \sigma^2 I)$.

Definition 2.1. The L_p -norm of the residual vector ϵ is

$$\|\epsilon\|_p = \begin{cases} (\sum_{i=1}^n |\epsilon_i|^p)^{1/p}, & \text{for } p \in [1, \infty), \\ \max |\epsilon_i|, & \text{for } p \rightarrow \infty. \end{cases} \quad (2)$$

An estimator minimizing a L_p - norm of the residual vector is called an L_p - norm estimator.

Measuring the size of ϵ in (1) using the L_p - norm, we arrive at the L_p -regression problem. In regression analysis, goal is to find β that attains the minimum L_p -norm for the difference between y and $X\beta$. Thus, the L_p -regression problem is to determine β such that

$$\min_{\beta} \|X\beta - y\|_p. \quad (3)$$

2.1. L_1 -norm regression model

Setting $p = 1$ in (3), the L_1 -norm regression problem becomes $\min_{\beta} \|X\beta - y\|_1$, which can be written as the linear programming (LP) problem

$$\min_{t, \beta} \sum_{i=1}^n t_i : -t_i \leq x_i^T \beta - y_i \leq t_i, i = 1, 2, \dots, n, \quad (4)$$

where X^T is the transpose of X . Methodology of estimating unknown parameters in L_1 -norm regression model was first introduced by Boscovich (1757). He proposed to estimate parameters according to the minimum of a function of the measurement errors. The proposed function was the sum of absolute measurement errors. This method is known as the minimum absolute deviations (MAD) or least absolute error (LAE) or minimum sum of absolute errors (MSAE) or least first power or L_1 -norm estimator. The estimating method was computationally complicated [Nyquist (1980)].

2.2. L_2 -norm (Least Square) regression model

Case $p = 2$ in (3) results in the L_2 -norm regression problem which is $\min_{\beta} \|X\beta - y\|_2$. This is equivalent to minimize

$$\sum_{i=1}^n \left(y_i - \sum_{j=1}^k x_{ij} \beta_j \right)^2 \quad (5)$$

with respect to β . The explicit formula for estimation of β is $\hat{\beta} = (X^T X)^{-1} X^T y$. This is the formula for L_2 -norm regression and is commonly known as the least square estimation (LSE) or least square regression estimators. It may be noted from the works of Legendre (1805) and Gauss (1809) that they proposed to minimize the sum of the squares of the measurement errors and, thereafter, the method of least square became the most popular estimating technique. The main reason for LSE's popularity is presumably due to easy computation and due to the fact that when the residuals are independent and identically normally distributed, the least squares estimators of a L_2 -norm regression model are also the best linear unbiased estimator as well as equivalent to the maximum likelihood estimator, implying the inference to be easily performed [Nyquist (1980)]. However, it has been noted that the least squares estimates are sensitive to departures from the assumptions, for example, normally distributed errors.

2.3. L_∞ -norm regression model

the L_∞ -norm regression problem translates to $\min_{\beta} \|X\beta - y\|_\infty$, which can be written as the linear programming (LP) problem

$$\min_{t, \beta} t : -t \leq x_i^T \beta - y_i \leq t, i = 1, 2, \dots, n. \quad (6)$$

This minimization problem is often referred to as the Chebyshev approximation. Laplace (1818) and Edgeworth (1887) have shown that the L_p -norm estimator is preferable to the least squares, when estimating a simple linear regression model with fat-tailed distributed residuals. Nyquist [1980] has investigated the L_p -norm estimators of linear regression models. In particular, he discussed results on the existence, uniqueness and asymptotic distributions of L_p -norm estimators and gave geometrical interpretations of L_p -norm estimation.

3 Numerical Applications

We first describe six bivariate data sets from the wide range of application areas [Abraham and Ledolter(2005)]. We focus only on the comparisons of estimated model parameters using various L_p -norms.

3.1. Descriptive Summaries of Data Sets

Data set A originates from a company which builds custom electronic instruments and computer components. The firm wants to investigate the association between overhead cost and the total direct labor hours. Data set A have a smaller number of degrees of freedom equals to 15 only. In data set B, iron contents of crushed blast furnace slag is of interest. Two methods, one chemical analysis in the laboratory which is time-consuming and expensive and other magnetic test on-site which is cheaper, are available. We investigate the extent to which the chemical tests of iron content can be predicted from a magnetic tests of iron contents. Measurements on 53 consecutive slags are available resulting in large number of degrees of freedom. Data set C is from a study on effects of environmental pollutants upon animals excluding man. An industrial pollutant, Polychlorinated biphenyl (PCB) is thought to have harmful effects on the thickness of egg shells. To investigate the relationship between the thickness of the egg shell in millimeters and the amount of PCB in parts per million in Pelican eggs, data are collected from 65 Anacapa pelican eggs. Data set D refers to the energy requirements in Mcal/day for a sample of 64 grazing merino sheep together with their body weights in kg. Objective is to fit a model that explains the energy requirements as a linear function of body weight. Data

set E is from a research study on advances in oxygen equivalence equations for predicting the properties of Titanium welds. Data on oxygen content in parts per million and strength in ksi for 29 welds are recorded to study their relationship. Data set F is from a research study on establishing a relationship between the erythrocyte adenosine triphosphate, ATP levels in the youngest and oldest sons in the families. The ATP level determines the ability of blood to carry energy to cells of the body. The data for the oldest and youngest sons are extracted from the 17 sampled families. ATP levels are expressed as micromoles per gram of hemoglobin and we estimate regression line for predicting ATP level of youngest son from that of the oldest son. It is noted that three data sets A,C and F follow approximately bivariate normal distributions, however, remaining three data sets B, D and E do not represent bivariate normal populations.

3.2. Checking Model Adequacy

The principle of analysis of variance partitions the total response variance into two components: the variance explained by the model and the variance that remained unexplained. For assessing model adequacy, one commonly used measure calculated from estimated residuals is the well known coefficient of determination R^2 which is defined as the proportion of the total response variance that is explained by the model:

$$R^2 = 100 \times \left[1 - \frac{\sum \epsilon_i^2}{\sum (y_i - \bar{y})^2} \right]. \quad (7)$$

We define a new measure denoted by $\| R^2 \|_1$ based on estimated residuals for checking model accuracy as

$$\| R^2 \|_1 = 100 \times \left[1 - \frac{\sum |\epsilon_i|}{\sum |y_i - \bar{y}|} \right]. \quad (8)$$

It may be noted that the numerator ϵ_i and denominator $(y_i - \bar{y})$ terms of $\| R^2 \|_1$ are L_1 -norm while in case of R^2 these are L_2 -norm. Either measure provides an overall measure of how well the model fits. A higher value of $\| R^2 \|_1$ or R^2 indicates a better fit.

3.3. Estimated Model Parameters and Error Measures

Estimated L_1 , L_2 and L_∞ -norm based linear regression model parameters along with model adequacy measures $\| R^2 \|_1$ and R^2 defined in (7) and (8) are presented in Table 1. It may be noted that for the data sets A, B and D, L_2 -norm based estimated model have the maximum $\| R^2 \|_1$ and R^2 values respectively as [39.38,62.62], [32.50, 53.72] and [34.72, 56.31]. Thus, for three populations A, B and D, L_2 -norm based estimated models are expected to perform better than the L_1 and L_∞ -norm based linear regression models. Referring to the data sets C, E and F, model accuracy measures $\| R^2 \|_1$ and R^2 do not lead to a consensus about the best estimated model. For data set C, we note that the L_1 -norm results in the best model, however, $L_2 \approx L_1$ using measure $\| R^2 \|_1$ and L_2 -norm is the best model according to R^2 criterion. In data sets E and F, we notice that the L_1 -norm is the best model, however, $L_2 \approx L_1$ using measure $\| R^2 \|_1$ and L_2 -norm is the best model according to R^2 criterion, however, $L_1 \approx L_2$. Thus interestingly, both estimated models L_1 - and L_2 -norm based linear regression models are good competitors for the populations represented by data set E and F. It may further be noted that data set E represents a non-normal population whereas data set F represents a normal population.

3.4. Residual Analysis

Residual analyses of the fitted linear models are presented in Table 2. Plots of residuals against fitted values and residual lag plots are not included (because of space limitations) however

Data Set	L_p Norm	Regression Coefficient	Intercept	Error Measure 1	Error Measure 2	Error Measure 3
A Ex 2.16 p.60	L_1	12.2518	15201	70.48	95.72	60.98
	L_2 (LSE)	10.9820	16310	73.06	95.50	62.62
	L_∞	8.8329	17423	93.11	94.04	38.59
B Ex 2.21 p.62	L_1	0.6154	7.6923	755.63	86.52	51.79
	L_2 (LSE)	0.5866	8.9565	793.36	86.41	53.72
	L_∞	0.6522	8.0495	819.23	86.03	52.23
C Ex 2.24 p.64	L_1	-0.0003	0.3714	1365.39	80.11	5.93
	L_2 (LSE)	-0.0003	0.3749	1395.64	80.03	2.95
	L_∞	-0.0002	0.3767	1611.22	78.47	9.68
D Ex 2.25 p.65	L_1	0.0463	0.0475	1073.01	83.96	55.95
	L_2 (LSE)	0.0434	0.1329	1062.09	83.94	56.31
	L_∞	0.0667	-0.8135	1364.30	79.53	36.03
E Ex 7.16 p.517	L_1	20.8393	43.8264	181.21	93.76	27.80
	L_2 (LSE)	16.9229	49.7796	183.75	93.67	29.37
	L_∞	29.8039	29.6033	199.49	93.15	11.55
F Ex 9.3 p.350	L_1	0.8254	1.0495	135.22	92.47	35.57
	L_2 (LSE)	0.8337	0.9867	135.50	92.42	35.69
	L_∞	1.2097	-0.5095	191.75	89.24	13.32

Table 1: L_p -norm Based Linear Regression Models and Model Adequacy Measures.

indicative conclusions are discussed. The characteristics of a well-behaved residual versus fitted values plots and residual lag plots, what they suggest about the appropriateness of the simple linear regression model, are described. (i) Linear relationship: In all cases, residuals are more or less spread randomly about the zero line. This suggests that the assumption that the relationship is linear is reasonable. (ii) Error Variance: The residuals have no increasing or decreasing trend and roughly form a horizontal band around the zero line. This suggests that the variances of the residuals are constant. (iii) Independence of Residuals: The residual lag plot by plotting residual (i) against lag residual ($i-1$) indicates the dependency of the residual terms. A random pattern in a lag plot suggests that the residuals are independent. This assumption appears to hold good for all models. (iv) Normality and Outlier Detection: The Shapiro-Wilk statistic and probability values given in Table 2 indicate that residuals have normal distributions. Since normality assumption holds, approximately 95 percent of the standardized residuals will fall between -2 and +2. It is seen from Table 2 that it is true for all models. Also from residual vs. fits plot, no one residual falls out from a random pattern of residuals. This suggests that there are no outliers.

4 Concluding Remarks

The least squares estimation (LSE) although is simple and algebraically highly developed, studies have shown that LSE based linear regression may not be the optimal model when one or others of its assumptions fail. For bivariate populations representing small to large size and normal and non-normal distributions, We have estimated L_1 , L_2 (LSE) and L_∞ -norm based linear regression models. Our findings are in agreement with those in some earlier studies. Our study also raises questions on the distributional properties of the L_p -norm based linear regression estimated models. The effects of deviating from the assumptions of LSE on the L_p -norm linear regression models. The statistical inference issues, like interval estimation, hypothesis testing and prediction bands etc., for the L_p -norm models. For given application data, how to determine optimal choice of the L_p -norm.

Data: L_p Norm	Min	Max	Q ₁	Q ₂	Q ₃	Skew	Kurtosis	Shapiro-Wilk	
								Statistic	Prob.
A: L_1	-2.01	1.85	-.856	.141	.782	-.307	-.093	.974	.894
L_2 (LSE)	-2.02	1.72	-.961	.1493	.748	-.350	-.297	.975	.908
L_∞	-1.94	1.43	-.694	.052	.672	-.312	-.552	.965	.748
B: L_2	-1.85	1.60	-.950	-.620	.007	.889	.837	.916	.145
L_2 (LSE)	-1.92	1.51	-.97-	-.648	.032	.824	.702	.923	.187
L_∞	-1.76	1.70	-.934	-.582	-.025	.967	1.018	.906	.099
C: L_2	-1.63	1.87	-1.00	-.219	.654	.286	-.700	.963	.720
L_2 (LSE)	-1.62	1.87	-.997	-.223	.654	.286	-.692	.963	.717
L_∞	-1.70	1.85	-.943	-.275	.650	.246	-.533	.977	.937
D: L_1	-2.10	3.47	-.821	.152	.687	.757	1.433	.942	.371
L_2 (LSE)	-.197	3.61	-.845	.080	.611	.980	1.915	.922	.179
L_∞	-.260	2.18	-.572	.398	1.027	-.518	.315	.958	.628
E: L_1	-1.79	1.44	-.969	.016	.902	-.349	-1.001	.938	.325
L_2 (LSE)	-1.75	1.32	-1.1	.129	.991	-.285	-1.240	.930	.247
L_∞	-1.87	1.62	-.96	.056	.752	-.278	-.702	.944	.399
F: L_1	-1.83	2.23	-.284	.041	.305	.203	.868	.910	.115
L_2 (LSE)	-1.83	2.23	-.286	.046	.294	.209	.870	.909	.113
L_∞	-1.41	2.30	-.558	.074	.581	.427	.415	.949	.477

Table 2: Residual Analysis.

Bibliography

- [1] Abraham, B. and Ledolter, J. (2005). *Introduction to Regression Modeling*. Cengage Learning: Duxbury Applied.
- [2] Adrien-Marie Legendre (1806) *Nouvelles méthodes pour la détermination des orbites des comètes 2014*. Encyclopædia Britannica Online. <http://www.britannica.com/EBchecked/topic/420949/Nouvelles-methodes-pour-la-determination-des-orbites-des-cometes>.
- [3] Boscovich, R.J. 1757. *De literaria expeditione per pontificiam ditionem et synopsis amplioris operis*. Bononiensi Scientiarum et Artum Instituto atque Academia Commentarii, **4**, 353–396. Reprinted with a Croatian translation by N. Cubranic, Institute of Higher Geodesy, Zagreb, 1961.
- [4] Edgeworth, F.Y. 1887. *On observations relating to several quantities*. Phil. Mag., 5th Series, **24**, 222–223.
- [5] Farebrother, R.W. 1999. *Fitting Linear Relationships: A History of the Calculus of Observations 1750-1900*. Springer-Verlag, New York Inc.
- [6] Gauss, C.F. 1809. *Theoria motus corporum coelestium in sectionibus conicis solem ambientium*. Hamburg.
- [7] Gauss, C.F. 1820. *Theoria combinationum observationum erroribus minimis obnoxiae, pars prior*. Printed in Werke, (Göttingen, 1880), IV: 6-7.
- [8] Laplace, P.S. 1812. "Théorie analytique des probabilités". Paris.
- [9] Legendre, A.M. 1805. *Nouvelles méthodes pour la détermination des orbites des comètes*. 72-75, Paris.
- [10] Nyquist, H. 1980. *Recent Studies on L_p -norm Estimation*. Doctoral thesis, University of Umea.
- [11] Stigler, S.M. 1990. *The History of Statistics: The Measurement of Uncertainty Before 1900*. Harvard University Press.

Using Storm for scaleable sequential statistical inference

Simon Wilson, *Trinity College Dublin*, simon.wilson@tcd.ie

Tiep Mai, *Bell Laboratories, Dublin*, maik@tcd.ie

Peter Cogan, *Amdocs, Dublin*, peter.cogan@gmail.com

Arnab Bhattacharya, *Trinity College Dublin*, bhattaca@tcd.ie

Oscar Robles Sánchez, *Universidad Rey Juan Carlos*, oscardavid.robles@urjc.es

Louis Aslett, *University of Oxford*, louis.aslett@stats.ox.ac.uk

Seán O’Ríordáin, *Trinity College Dublin*, seoriord@tcd.ie

Gernot Roetzer, *Trinity College Dublin*, roetzer@tcd.ie

Abstract. This article describes Storm, an environment for doing streaming data analysis. Two examples of sequential data analysis — computation of a running summary statistic and sequential updating of a posterior distribution — are implemented and their performance is investigated.

Keywords. Storm, sequential inference, streaming data

1 Introduction

In sequential statistical inference, data arrive as a stream and inference is an iterative process that updates as new data are available. Numerous examples and applications exist, starting with the Kalman filter and its generalisations such as the dynamic state space model [4]. Approaches to implement inference in this setting are the subject of much current work e.g. sequential Monte Carlo [3]. The challenge is not only to work with data sources that require sophisticated analyses, but also for scaleable inference algorithms that can cope with increased data dimension and arrival rates.

Computational capabilities for the collection, management and analysis of large volumes of data continue to increase at a fast rate. Most of the well known internet companies have developed storage and processing systems that adopt the MapReduce paradigm [2], where scaleability is achieved by exploiting the availability of many processing units that can work in parallel on independent tasks, and fault tolerance is achieved by managing these tasks so that they can be re-assigned to a different processor if a fault is detected. MapReduce implementations of

algorithms are now relatively easy to code with software libraries such as Hadoop [9]. These are batch computations i.e. a single computation with a pre-defined set of data.

However, analysis of streaming data is becoming another important challenge, for which Hadoop has not been designed; it treats a sequential analysis as a sequence of batch analyses. This will typically involve writing data to memory after each batch and then reading it again which can be very inefficient. To address this, environments such as Storm have been developed. They aim to permit the programming of analyses of streams of data in a scaleable and reliable manner that is analogous to MapReduce in many ways.

In the context of statistical analysis, it is natural then to ask what are the advantages of using a streaming data environment such as Storm to implement sequential statistical inference algorithms, and for which algorithms are these advantages greatest. In this paper, we describe a programming environment called Storm [5]. This is one of several such environments for the processing of streaming data in a distributed manner. It is applied to two examples: computation of running summary statistics and a grid-based approximation. The performance of these algorithms is evaluated and discussed with respect to these examples.

2 What is Storm?

Storm is an example of an open source, distributed, fault tolerant framework for the processing of streaming data. This is achieved via the concept of *topologies*, a directed acyclic graph which, at an abstract level, represents both the computation to be performed and the flow of data through the system. Each datum in the data stream is known as a *tuple*. Data are introduced into the topology via *spouts*, processed by *bolts* and data flows between them according to *stream groupings*. Simply, spouts are sources of data, bolts are functions in the code that have input variables and produce an output, and the topology shows how the inputs and outputs of each propagate through the computation according to the stream groupings. Parallelisation is achieved by setting the number of replications (referred to as tasks) of each spout and bolt. Storm manages the computational load across the available processors; see [1] for more details.

Storm was initially developed in 2011 by a company called BackType which had been founded in 2008. BackType was acquired by Twitter in July 2011, and Twitter made Storm open-source later in September 2011. In September 2013 Storm became an Apache incubation project; this ensures that the code base of Storm will not be abandoned.

One interesting aspect of the way that Storm manages the data stream concerns guaranteeing that every tuple that is input into the system, as well as any new tuples that are created from it during the computation, has been fully processed. This guarantee is implemented by assigning a unique message id to each tuple generated within a spout. Once it and any tuple generated from it have been processed then the acknowledgement function `ack()` is called by the originating spout. If that does not happen then a `fail()` function is called and the tuple is reprocessed. The `ack()` function can be used for temporal synchronization of ordered data, i.e. the spout can send the next data tuple when the previous tuple has been fully processed. However, such usage induces a strong bottleneck in the system as the computation will then move at the rate of the slowest bolt to process any part of a tuple in each temporal step.

3 Performance Assessment

The performance of a streaming data processing algorithm can be evaluated in several ways, the most common of which are:

Throughput: This is the average number of tuples processed per unit time.

Latency: This is the average time it takes for a tuple to be processed. Latency may also be defined for parts of a computation, such as a bolt or combinations of bolts. A special case is *execute latency* which is the time taken by the bolts in the topology to process a tuple, ignoring communication time and other overheads in managing the computation.

Capacity: This is a measure of the proportion of time that Storm spends in processing tuples with the bolts in the topology, defined as

$$\text{Capacity} = \frac{\text{Execute latency} \times \text{No. of observations processed}}{\text{Total computation time}}.$$

A capacity of 1 usually indicates that bolts are overloaded and unable to process data as quickly as it can be streamed.

These statistics play an important role in scaling the streaming system, and so Storm has a user interface that allows one to monitor performance of each bolt, spout and processor being used. A capacity near to 1 indicates a bottleneck of the current system which could be improved with more computational bolts or cluster machines. Ideally, when scaling an algorithm to make use of a larger number of processors, one should be able to increase throughput close to linearly with the number of processors while both latency and capacity remain steady.

4 Example: Computing running summary statistics

In this first example, a stream of bivariate normal observations $(x_1, y_1), (x_2, y_2), \dots$ is generated and the goal is to output the running sample correlation:

$$r_n = \frac{n \sum_{i=1}^n x_i y_i - \sum_{i=1}^n x_i \sum_{i=1}^n y_i}{\sqrt{n \sum_{i=1}^n x_i^2 - (\sum_{i=1}^n x_i)^2} \sqrt{n \sum_{i=1}^n y_i^2 - (\sum_{i=1}^n y_i)^2}}, \quad n = 2, 3, \dots \quad (1)$$

Figure 1 shows the topology. On the left, one or more spouts called bvn data simulate bivariate normal observations. More than one spout may be needed if we are testing the performance limits of the algorithm because the generation of the data requires more computation than the computation of the correlation. The data are streamed in groups of size k , with each group transmitted to only one summary bolt. This assignment of a group to a particular replication of the summary bolt is done using one of Storm's standard transmission options called shuffle stream grouping, where the bolt is chosen at random.

The m th set of k observations $D_m = \{(x_i, y_i) \mid i = (m-1)k+1, \dots, mk\}$ is sent to a summary bolt, which computes the five summary statistics

$$S_m = \sum_{i=(m-1)k+1}^{mk} (x_i, y_i, x_i^2, y_i^2, x_i y_i)$$

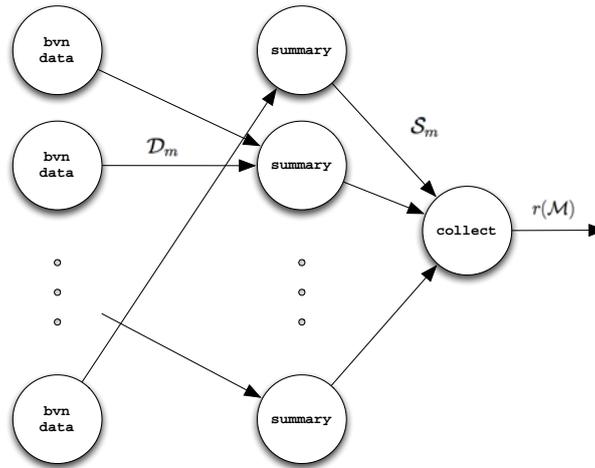


Figure 1: The topology for computing the running correlation of a stream of bivariate observations.

needed to compute the correlation, and then transmits S_m to the collect bolt. The collect bolt updates the running sum of the summary statistics and uses them to compute the sample correlation. Defining $M = \{m \mid S_m \text{ transmitted to collect}\}$, collect will compute and store the 5 summary statistics over all transmitted sets:

$$\mathbb{S} = \sum_{m \in M} S_m,$$

from which it can output the sample correlation, as defined in Equation 1, by

$$r(M) = \frac{|M|k\mathbb{S}_5 - \mathbb{S}_1\mathbb{S}_2}{\sqrt{|M|k\mathbb{S}_3 - (\mathbb{S}_1)^2} \sqrt{|M|k\mathbb{S}_4 - (\mathbb{S}_2)^2}}.$$

This example illustrates the issue of synchronisation. There is no guarantee that if M sets of statistics S_m have arrived to the collect bolt then they are S_1, \dots, S_M . However as can be seen above, the indices m of the sets that have been transmitted to collect can also be transmitted if needed, so that at least one knows which data have been used in the computation of the correlation.

This topology was implemented on a cluster of 6 machines with a total of 32 cores using observation groups of size $k = 50$. Thus for every 50 observations generated, one correlation value should be transmitted by collect. The throughput of observations and correlations for different numbers of bvn data spouts and summary bolts was explored. It was observed that peak throughput occurred when between 8 and 16 bvn data spouts were used per summary bolt, and so the experiments kept to that ratio. With the ratio of bolts to spouts constant, in principle the capacity of the algorithm to process observations is constant, and so changes in performance are due to the overhead involved in managing different numbers of spouts and bolts. The algorithm was allowed to run for several minutes to eliminate any initialization effects, and then data were recorded for 6 minutes; throughput is reported as the average output per minute. Figure 2 shows that, for this cluster, performance begins to deteriorate when more than about 250 spouts are replicated. Having more bolts does give better performance, but having twice as many (runs with 8 spouts per bolt) does not give twice the throughput.

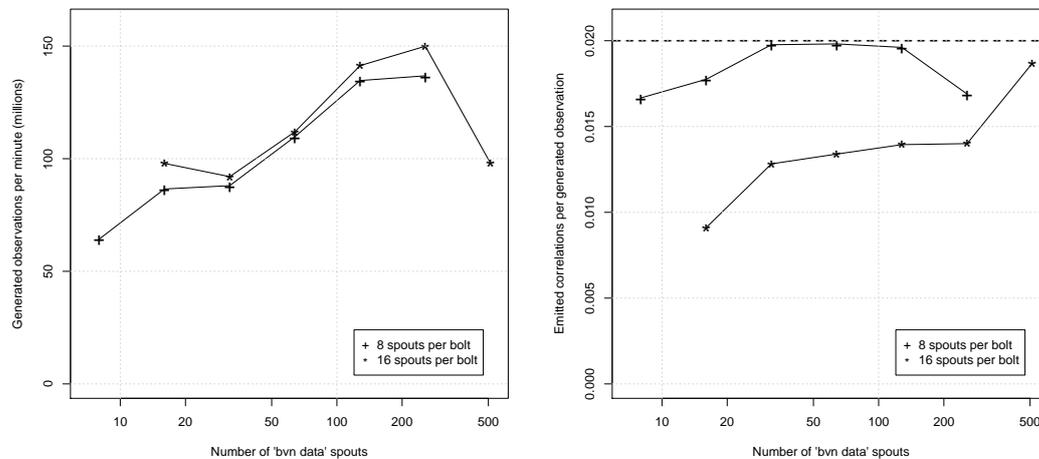


Figure 2: Summary of experiments with different numbers of bvn data spouts with a fixed ratio of spouts to summary bolts. Left: observation throughput as a function of the number of bvn data spouts. Right: number of correlations emitted per observation generated as a function of the number of bvn data spouts; the dashed line shows where 1 correlation is emitted for every $k = 50$ data points e.g. all data points are being processed.

5 Example: Sequential posterior computation

A stream of observations x_1, x_2, \dots is to be fitted to a parametric probability model $p(x|\theta)$. It is assumed that θ is of small enough dimension so that it is possible to compute the posterior distribution of the parameters on a discrete grid of points Θ . The goal is to sequentially update the posterior; when x_{n+1} arrives, the posterior is updated via the Bayes recursion:

$$p(\theta | x_{1:n+1}) \propto p(\theta | x_{1:n}) p(x_{n+1} | \theta),$$

where $x_{1:n} = \{x_1, \dots, x_n\}$. The output is a stream of sets of posterior distribution values $p(\theta | x_{1:n}), \theta \in \Theta$ for $n = 1, 2, \dots$

A parallel implementation of this computation is to partition Θ and assign the computation of the unnormalized log posterior

$$l(\theta) = \log(p(\theta)) + \sum_{i=1}^n \log(p(x_i | \theta))$$

over each part of the partition to bolt replications, where $p(\theta)$ is a prior. Let M be the degree of parallelization available for the computation and let $\Theta_1, \dots, \Theta_M$ be a partition of Θ ; load balancing considerations imply that the Θ_m should be of similar size.

Figure 3 shows the topology. There are M instances of the logpost bolt; each is assigned a different subset of the grid Θ_m over which to store the unnormalized log posterior values $P_m = \{l(\theta) | \theta \in \Theta_m\}$. When a new observation x_{n+1} arrives, the transmit bolt transmits it to all M instances of the logpost bolt; this is an all stream grouping, in contrast to the first example, where data was transmitted to only one summary bolt. The replication that is responsible for

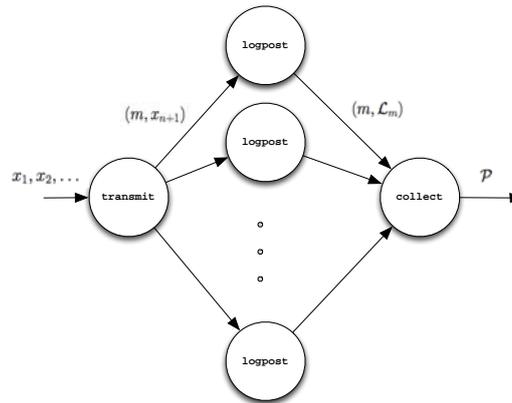


Figure 3: The topology for sequential posterior computation.

Θ_m computes $\log(p(x_{n+1} | \theta))$, $\theta \in \Theta_m$, and adds it to the corresponding element of P_m . After every K observations have been processed by the logpost bolts, they transmit P_m to the collect bolt that then exponentiates and normalises the values to derive the posterior density over the grid.

An important distinction between this example and the previous one is that the logpost bolts have state; they must store the current value of the log posterior. If a bolt dies then that state is lost and can be recovered only by computing the log posterior from scratch on its partition. Alternatively, the state could be stored and read from memory, but that again implies an overhead to the computation.

We illustrate this idea for Gaussian data with unknown mean μ and precision τ , so that $\theta = (\mu, \tau)$ and $p(x | \theta) = (\tau/2\pi)^{0.5} \exp(-0.5\tau(x - \mu)^2)$. For this example we assume independent non-informative Gaussian (zero mean, large variance) and gamma (scale and shape are 0.5) priors on μ and τ .

This topology was implemented on a cluster of 5 identical machines, each with four 3.4 GHz cores. One million Gaussian observations were generated and stored to a file; the file was streamed and processed using 4, 8, 12, 16 and 20 logpost bolts. The posterior density was computed by the collect bolt every $K = 50,000$ observations. This value of K was used because of the large size of the output, given the rate at which data can be processed; with a smaller K then the input-output time begins to dominate the processing time in the system. A small grid of size $76 \times 86 = 6,536$ and a larger one of $376 \times 426 = 160,176$ points were used, with points distributed as evenly as possible between the bolts. Further, this problem was implemented in two ways, which we label as ack and nack: with ack, the transmit spout acknowledges that each observation has been completely processed successfully. When a fail() is called, Storm will automatically replay the tuple. With nack, no acknowledgement is made.

Figure 4 shows results from these experiments. The left plot shows the median data throughput over 6 runs as a function of the number of logpost bolts for 3 cases: the small grid with ack, the small grid with nack and the large grid with nack. As it involves more computation per observation, the larger grid has a lower data throughput than the smaller grid, hence the data throughput curves of two datasets are not comparable. Still, they are plot together in Figure 4a for convenience and for the progression of data throughput over number of bolts. There is a considerable cost to using ack, which grows larger as the number of logpost bolts increases.

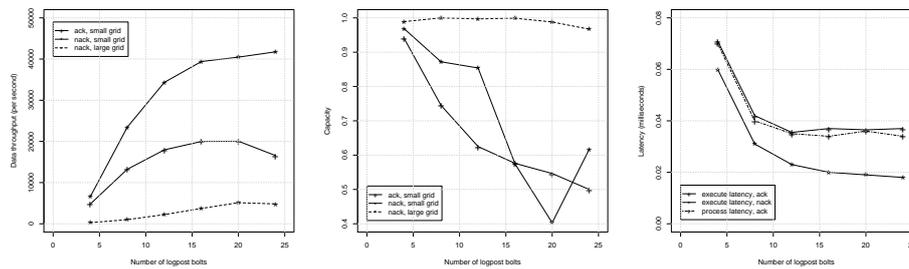


Figure 4: Performance of the sequential computation of the posterior density of the mean and precision of a Gaussian distribution as a function of the number of logpost bolts over 6 runs. From left to right: median data throughput, median latency and median capacity.

Performance worsens considerably in one case from 20 to 24 bolts; the cluster has 20 cores, and so managing 20 or 24 bolts means 2 or more bolts running on some cores and a computation overhead results. The capacity plot shows that the larger grid is more efficient in that it spends more time in computing log likelihoods (the dominant computation in the bolts) rather than in communication. In the nack small grid case, the capacity is around 0.97 when there are 4 log-post bolts, meaning that each bolt is very busy. This high capacity implies a bottleneck in a system but, unlike the throughput measurement, it does not measure how fast the system is. In the nack-small-grid case, when the capacity value is from 0.85 to 1, the system throughput can be improved significantly by adding more processing power (bolts). In the nack large grid case, the capacity is almost 1, which implies that a larger cluster would lead to a faster computation. Finally, latencies are plotted for 3 cases, all with the small grid: execute latency for ack, execute latency for nack and process latency for ack. The latency of the big grid is not drawn as it follows the same pattern but on a different scale (from 1.4ms down to 0.4ms). It can be seen that the execute latency is slightly longer than the process latency. As with throughput, there is a considerable overhead in using ack that grows with the number of bolts, and performance does not improve significantly with more than 16 bolts.

6 Concluding Remarks

In this paper we have introduced Storm and illustrated its use in 2 examples of sequential data analysis. The topology of the second example, where a function is evaluated on all data at each point in a discrete grid, is a common scenario. In Bayesian inference, it is often the computationally most demanding step of the integrated nested Laplace approximation [8]. Another example where this topology could be used is the griddy Gibb's sampler [7].

Sequential Monte Carlo methods, such as the particle filter, have a similar structure to the second example but where the fixed grid is the set of particles. However they have an important distinction in that the topology has a cycle; results of processing one datum, such as particle weights, are needed to process the next. While Storm can implement such topologies, it introduces potentially difficult issues of synchronization. This has spurred the development of systems for iterative computation e.g. [6]. For sequential statistical methods like the particle filter, an interesting question is which will be more effective.

The examples demonstrate the typical properties of a parallel algorithm, with a trade off between increasing parallelization and the overhead of managing a larger number of processors. In terms of Storm and its alternatives for streaming computation, we see advantages in terms of ease of coding, easy scaleability, reliability and the development of interfaces with higher level languages such as R. It is faster than R, much better suited to streaming data applications than OpenMP and OpenMPI and much easier to program than a GPU through CUDA.

Acknowledgement

This work was supported by the STATICA project, contract number 08/IN.1/I1879, and the Insight Centre for Data Analytics, contract number 12/RC/2289. Both are funded by Science Foundation Ireland.

Bibliography

- [1] Bedini, I., S. Sakr, B. Theeten, A. Sala, and P. Cogan (2013). Modeling performance of a parallel streaming engine: bridging theory and costs. In *Proceedings of the International Conference on Performance Engineering*, pp. 173–184.
- [2] Dean, J. and S. Ghemawat (2008). MapReduce: simplified data processing on large clusters. *Communications of the ACM* 51, 107–113.
- [3] Doucet, A., J. de Freitas and N. Gordon (2001). An introduction to sequential Monte Carlo methods. In A. Doucet, J. de Freitas, and N. Gordon (Eds.), *Sequential Monte Carlo methods in practice*. New York: Springer-Verlag.
- [4] Durbin, J. and S. J. Koopman (2001). *Time series analysis by state space methods*. Oxford University Press.
- [5] Marz, N. (2013). Storm: Distributed and fault-tolerant realtime computation. <http://storm-project.net>.
- [6] Murray, D. G., F. McSharry, R. Isaacs, M. Isard, P. Barham and M. Abadi (2013). Naiad: a timely dataflow system. *Proceedings of the 24th ACM Symposium on Operating Systems Principles*, 439–455. New York: ACM.
- [7] Ritter, C. and M. Tanner (1992). Facilitating the Gibbs sampler: the Gibbs stopper and the griddy-Gibbs sampler. *Journal of the American Statistical Association* 87, 861–868.
- [8] Rue, H., S. Martino, and N. Chopin (2009). Approximate Bayesian inference for latent Gaussian models using integrated nested Laplace approximations. *Journal of the Royal Statistical Society, Series B* 71(2), 319–392.
- [9] White, T. (2012). *Hadoop, the Definitive Guide* (Third ed.). Yahoo Press, O’Reilly.

A simulation study to assess statistical approaches for longitudinal count data

M. Salomé Cabral, *CEAUL and DEIO, Faculdade de Ciências da Universidade de Lisboa, Portugal*, salome@fc.ul.pt

M. Helena Gonçalves, *CEAUL and FCT, Universidade do Algarve, Portugal*, mhgoncal@ualg.pt

Abstract. In this paper is studied the performance of statistical methods used to analyze longitudinal count data when the target of inference is the population. The goal of this study is to give a statistical assessment of marginal approaches in terms of properties such as efficiency and coverage probability, as well as, to give some guidelines for the choice of the statistical approach to an applied researcher. Two approaches are considered: the generalized estimating equations (GEE) and the maximum likelihood estimation with a serial dependence of Markovian type (MML). A simulation study was carried out and the results indicate to a better performance of the MML approach when the correlation among response variable for a given subject increases.

Keywords. count longitudinal data, marginal model, exact likelihood, generalized estimating equations, Markov chain.

1 Introduction

Longitudinal count data are commonly encountered in both experimental and observational studies across all disciplines. In these studies repeated measurements are made on the same subject across occasions in one or more treatment groups. In order to make correct inferences, the correlation among response variable for a given subject must be take into account. In the context of marginal model, this is, when the target of inference is the population, several models have been proposed. [6] proposed the generalized estimation equations (GEE) method. [9] proposed an estimation equation method for regression analysis with a time series of counts analogous to the one used by [6]. [5] used generalized estimating equations to model longitudinal count data with overdispersion. [1] proposed an approach based on maximum likelihood estimation where the serial dependence is assumed to be of Markovian type (MML).

The goal of this paper is to give information to the practitioners about which of the two procedures, GEE or MML, is more appropriate to use for their data at hand. To achieved that goal a simulation study is carried out to compare the two aforementioned approaches in terms of properties such as efficiency and coverage probability.

For GEE approach the estimates were obtained through the function `geeglm` in the R package `geepack` [3]. The function `cold` in the R package `cold` [2] is used to obtain the MML estimates.

The paper is organized as follows: Section 2 gives a summary of the models used. Section 3 reports a small simulation study to assess the performance of the procedures. Section 4 concludes the paper.

2 Parametric models

Consider count responses y_{it} ($t = 1, \dots, T_i$) at time t from subject i ($i = 1, \dots, n$), a set of p explanatory variables, x_{it} , associated with each observation time and each subject, and Y_{it} its generating random variable which has a Poisson distribution with $E(Y_{it}) = \theta_{it}$. The Poisson regression which links the covariates and the probability distribution of the response, is given by

$$\ln(\theta_{it}) = x_{it}^\top \beta, \quad (1)$$

where β is the p -vector of unknown parameters.

Maximum likelihood estimation

The approach based on maximum likelihood estimation proposed by [1] is implemented in the R package `cold` and is summarized in this section. In this approach is made use of the idea of self-decomposable probability distribution following [7] and the serial dependence is assumed to be of Markovian type. To simplify notation, the subscript i is dropped temporarily.

$$Y_t = \rho \circ Y_{t-1} + \varepsilon_t, \quad (t = 2, 3, \dots, T), \quad (2)$$

where for any given t , $E(Y_t) = \theta_t$ assuming that $E(Y_1) = \theta_1$, ε_t is a Poisson random disturbance, $\rho \in (0, 1)$ and $\rho \circ Y_{t-1}$ [7] is defined by

$$\rho \circ Y_{t-1} = \sum_{h=1}^{Y_{t-1}} Z_h, \quad (3)$$

where Z_1, Z_2, \dots is a sequence of independent Bernoulli variables with common probability of success ρ , $\Pr(Z_h = 1) = 1 - \Pr(Z_h = 0) = \rho$. See [4] for details.

The response variable Y_t , as given in (2), is the sum of two independent random variables; one of which has Poisson distribution with expected value equal to $\theta(1 - \rho)$, and the other has binomial distribution with probability of success equal to ρ . The m -step transition probabilities are

$$\Pr(Y_t = j | Y_{t-m} = i; \theta) = \sum_{k=0}^{\min(i,j)} \binom{i}{k} \rho^{mk} (1 - \rho^m)^{i-k} \frac{\exp(-v_{t,m}) v_{t,m}^{j-k}}{(j-k)!}. \quad (4)$$

The contribution from a generic individual to the likelihood for the parameters (β, ρ) is

$$L_i(\beta, \rho) = \frac{\exp(-\theta_1) \theta_1^{y_1}}{y_1!} \prod_{t=m+1}^T \Pr(Y_t = y_t | Y_{t-m} = y_{t-m}; \theta). \quad (5)$$

The overall log-likelihood function is obtained as the sum of the n logarithmic individual contributions of type (5).

Generalized estimating equations

The generalized estimating equations (GEE) presented in [6] are an extension of the quasi-likelihood of [8] to the case when the second moment cannot be fully specified in terms of expectation but rather additional correlation parameters must be estimated, what differs is the way to choose the variance-covariance matrix. This approach is implemented in the R package `geepack` and can be summarized as follows. Consider,

$$\text{var}(Y_{it}) = \phi V(\theta_{it}), \quad (6)$$

where ϕ is a common scale parameter and $V(\theta_{it})$ is a known variance function.

The GEE for β are

$$U_{\beta}(\beta, \alpha) = \sum_{i=1}^n D_i^{\top} V_i^{-1} (Y_i - \theta_i) = 0,$$

where $D_i = \frac{\partial \theta_i}{\partial \beta}$, $Y_i = (Y_{i1}, \dots, Y_{iT_i})^{\top}$, θ_i the vector of the mean of Y_i and V_i is now called a “working” variance-covariance matrix. For the i th subject

$$V_i = \phi A_i^{1/2} R_i(\alpha) A_i^{1/2},$$

where A_i is the diagonal matrix with entries $V(\theta_{it})$ and $R_i(\alpha) = \text{corr}(Y_i)$ is a $T_i \times T_i$ “working” correlation matrix.

3 A simulation study

A brief simulation study was carried out to study the performance of both methodologies. The marginal Poisson model with a first order autocorrelation between two successive observations of the same subject was considered. The model included a dichotomous treatment, a linear effect time and an interaction between time and treatment and is given by

$$\theta_{it} = \exp(\beta_0 + \beta_1 t + \beta_2 x_i + \beta_3 (t \times x_i)), \quad (7)$$

where $x_i = 0$ for half the population and 1 for the remainder. The regression coefficients were set at $\beta_0 = 1, \beta_1 = 0.5, \beta_2 = 1.5$ and $\beta_3 = 0.10$.

To reflect the range of experimental data encountered in practice several designs were considered. The number of subjects was set to either small ($n = 20$) or large ($n = 50$). The length of profile on each subject was short ($T = 5$) or long ($T = 13$). The correlation between successive observations of the same subject was set at $\rho = 0.25, 0.5$ and 0.75 (low, moderate or high, respectively). On each run were generated T correlated Poisson observations under the i th subject following the AR(1) model given by (2). The time points were set for $T = 5$ at $t = -1, -0.5, 0, 0.5, 1$, and for $T = 13$ at $t = -1.5, -1.25, -1, -0.75, -0.5, -0.25, 0, 0.25, 0.5, 0.75, 1, 1.25, 1.5$. The whole estimation procedure was repeated for 1000 runs and the sample mean of estimate parameter (Mean), the sample mean of percent relative bias (Rbias%) and the sample mean square

error (MSE) were computed, as well as, the coverage probabilities of nominal 95% confidence intervals.

For each simulated dataset the estimated 95% confidence interval of each parameter in the model was computed based on the sample normal approximation. To GEE approach the sandwich standard error was used. When the MML approach was considered the standard error was based on the Fisher information matrix. The coverage probabilities of nominal 95% confidence intervals were computed as the proportion of simulated intervals that cover the true parameter used to generate the simulated data. The relative efficiency (RE) of MML estimators to GEE estimators was computed, as usual, by the ratio of the respective MSE. $RE > 1$ means MML estimator is preferred.

The estimates of the parameters using the MML approach were obtained through the function `cold` in the R package `cold`. When the GEE approach was considered the function `geeglm` in the R package `geepack` was used.

The simulation results are given from Figures 1 to 3 and in Table 1. In Figures 1 and 2 are display the graphics of the coverage probabilities of nominal 95% confidence intervals of β to both approaches in all the situations considered. Figures 3 gives the relative efficiency of the MML estimators to GEE estimators of $\hat{\beta}$.

In Table 1 are displayed the simulation results to ρ parameters. To each approach the table lists the following: Mean, Rbias% and MSE over the 1000 simulations runs and, in parentheses, the coverage probabilities of nominal 95% confidence intervals.

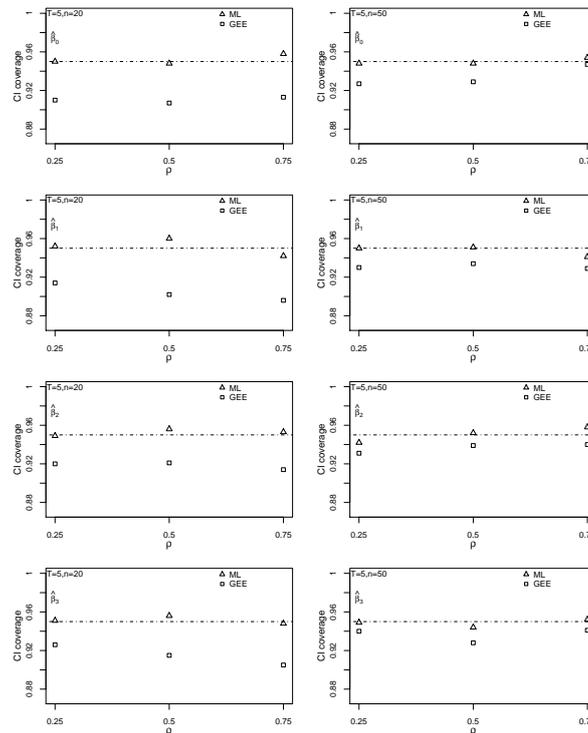


Figure 1: Coverage probabilities of nominal 95% confidence intervals for β to MML and GEE approaches when $T = 5$.

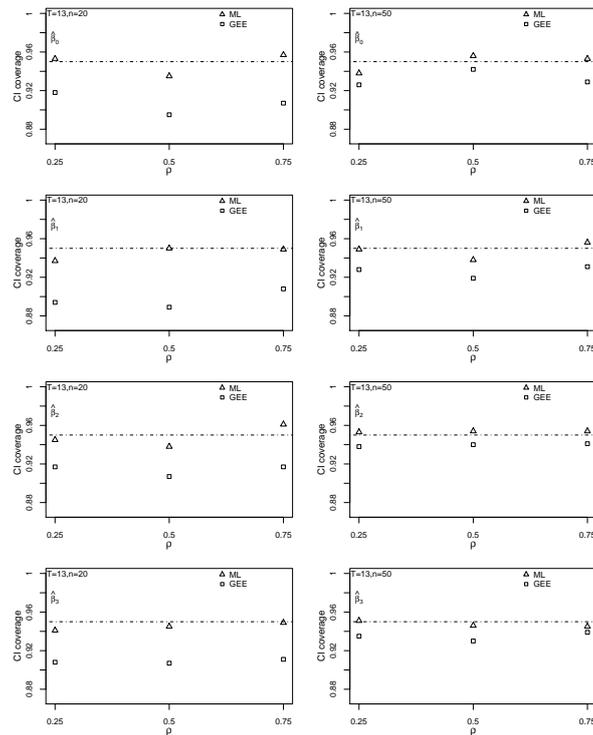


Figure 2: Coverage probabilities of nominal 95% confidence intervals for β to MML and GEE approaches when $T = 13$.

	ρ		0.25		0.5		0.75	
	T	n	MML	GEE	MML	GEE	MML	GEE
Mean	5	20	0.225 (0.960)	0.182 (0.809)	0.481 (0.936)	0.393 (0.758)	0.744 (0.952)	0.620 (0.751)
Rbias%			-9.930	-27.217	-3.759	-21.409	-0.781	-17.348
MSE			0.017	0.020	0.012	0.026	0.004	0.028
Mean		50	0.241 (0.946)	0.205 (0.872)	0.490 (0.937)	0.415 (0.731)	0.747 (0.943)	0.639 (0.597)
Rbias%			-3.800	-18.089	-1.944	-17.056	-0.344	-14.756
MSE			0.006	0.008	0.004	0.014	0.002	0.017
Mean	13	20	0.238 (0.962)	0.213 (0.856)	0.491 (0.940)	0.439 (0.770)	0.747 (0.944)	0.674 (0.713)
Rbias%			-4.691	-14.948	-1.732	-12.286	-0.576	-10.187
MSE			0.004	0.006	0.003	0.010	0.001	0.012
Mean		50	0.245 (0.938)	0.224 (0.881)	0.498 (0.944)	0.457 (0.787)	0.748 (0.945)	0.690 (0.685)
Rbias%			-2.076	-10.220	-0.461	-8.568	-0.301	-8.042
MSE			0.002	0.003	0.001	0.004	0.0004	0.006

Table 1: Results of the simulation study for ρ . Coverage probabilities of nominal 95% confidence intervals given in parentheses.

Taking into account the goal of the simulation study the main conclusions can be summarize as follows.

- (1) To all β parameters: (i) the coverage probabilities are closer to nominal for the MML approach than for the GEE approach; (ii) the MML estimators are more efficient than the GEE

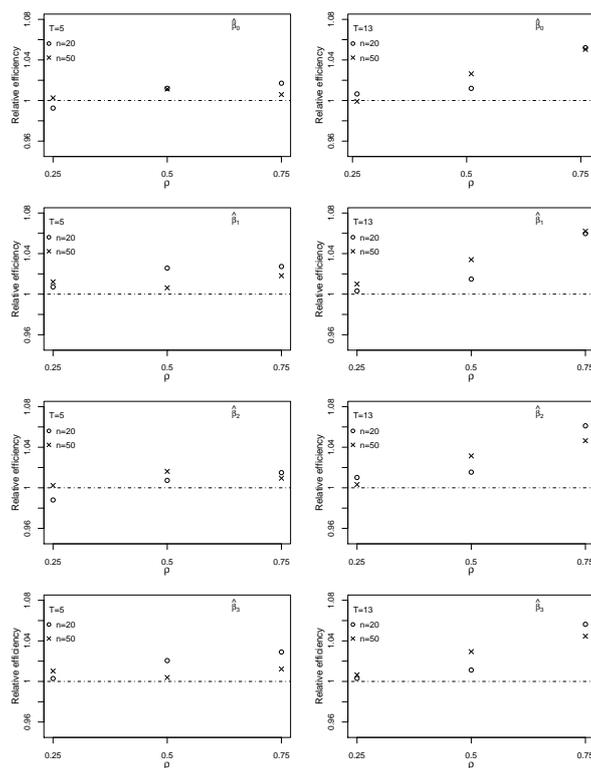


Figure 3: Relative efficiency of $\hat{\beta}$ for different correlations ρ and MML and GEE approaches.

estimators to higher values of ρ . This is so much better applied as the length of the profile of each subject increases. (2) To the estimator of ρ and in all situations consider: (iii) the MML approach gives lesser values of $Rbias\%$ and MSE than GEE approach. The coverage probabilities are closer to nominal for MML approach than for the GEE approach.

4 Conclusion

This paper is concerned with the asses of performance of the MML approach implemented in R package `cold` and GEE approach implemented in R package `geepack` for the analysis of longitudinal count data in the context of marginal model. The results of the simulation study point out that the MML approach seems to be preferable to GEE in all situations considered by checking that its performance is so much better the higher the correlation between observations of the same subject, regardless of the number of subjects involved in the study or the length of their profile.

Acknowledgements

Research partially sponsored by national funds through the Fundação Nacional para a Ciência e Tecnologia, Portugal-FCT under the project (PEst-OE/MAT/UI0006/2014).

Bibliography

- [1] Azzalini, A. (1994) *Logistic regression and other discrete data models for serially correlated observations*. J. ital. statist. soc., **2** 169–179.
- [2] Gonçalves, M.H. and Cabral, M.S. (2013) *cold: A package for count longitudinal data* R foundation for statistical computing, version 1.0-3. url = <http://CRAN.R-project.org/package=cold>.
- [3] Højsgaard, S., Halekoh, U. and Yan, J. (2006) *The R package geepack for generalized estimating equations*. Journal of statistical software, **15**(2), 1–11.
- [4] Gonçalves, M.H., Cabral, M.S., Ruiz de Villa, M.C., Escrich, E. and Solanas, M. (2007) *Likelihood approach for count data in longitudinal experiments*. Computational statistics and data analysis, **51**(12), 6511–6520.
- [5] Jowaheer, V. and Sutradhar, B.C. (2002) *Analysing longitudinal count data with overdispersion*. Biometrika, **89**(2), 389–399.
- [6] Liang, K.Y. and Zeger, S.L. (1986). *Longitudinal data analysis using generalized linear models*. Biometrika, **73**(1):13–22.
- [7] F. W. Steutel and K. van Harn. (1979) *Discrete analogues of self-decomposability and stability*. The annals of probability, **7**(5):893–899.
- [8] Wedderburn, R.W.M. (1974) *Quasi-likelihood functions, generalized linear models and the Gauss-Newton method*. Biometrika, **61**:439–447.
- [9] Zeger, S.L. (1988) *A regression model for time series of counts*. Biometrika, **75**(4):621–629.

Mixture model of Gaussian copulas to cluster mixed-type data

Matthieu Marbac, *DGA & Inria Lille & University Lille 1*,
matthieu.marbac-lourdelle@inria.fr

Christophe Biernacki, *University Lille 1 & CNRS & Inria Lille*,
christophe.biernacki@math.univ-lille1.fr

Vincent Vandewalle, *EA 2694 University Lille 2 & Inria Lille*,
vincent.vandewalle@univ-lille2.fr

Abstract. A mixture model of Gaussian copulas is proposed to cluster mixed data. This approach allows to straightforwardly define simple multivariate intra-class dependency models while preserving classical distributions for the one-dimensional margins of each component in order to facilitate the model interpretation. Moreover, the intra-class dependencies are taken into account by the Gaussian copulas which provide one robust correlation coefficient per couple of variables and per class. This model generalizes different existing models defined for homogeneous or mixed variables. The Bayesian inference is performed via a Metropolis-within-Gibbs sampler. The model is illustrated by a real data set clustering.

Keywords. Clustering, Gaussian copula, Gibbs sampler, Mixed data, Mixture models.

1 Introduction

With the informatics advent, multivariate data sets become more complex. Particularly, they often contain mixed data (variables of different kinds). *Clustering* provides an efficient solution to extract the main information from the data by grouping the individuals into few characteristic classes. It can be performed by probabilistic methods modelling the data generation whose the most popular one uses finite mixture models of parametric components [12]. In such a case, a class gathers together the individuals drawn by the same distribution. Obviously, the choice of the component distributions depends on the kind of the variables at hand. However, few distributions exist to model mixed data and their margin distributions are often complex [8].

The simplest way to cluster mixed variables consists in approaching the data distribution with a finite mixture model assuming independence conditionally on the class membership of each individual. This model, called *locally independent model*, obtains good results in many real clustering problems [11, 6], especially when few individuals are described by several variables.

Indeed, when its one-dimensional margins of each component follow classical distributions, this model provides a meaningful summary of the data by its margin parameters. However, this model leads to biases when its assumption of conditional independence is violated.

The aim of this paper is to present a model-based clustering for mixed data of any kinds of variables admitting a cumulative distribution function. This model has a double objective: to preserve *classical distributions* for *all* its margin distributions of each component and to *model the intra-class dependencies*. This objective can naturally be achieved by the use of copulas [9] since these objects allow to build a multivariate model by setting, on the one hand, the one-dimensional *margins*, and, on the other hand, the *dependency model* between variables. More precisely, the data distribution is approached by a full parametric *mixture model of Gaussian copulas* whose the margin distributions of each component are classical and whose the Gaussian copulas [7] model the intra-class dependencies. The new mixture model is meaningful since each class is summarized by its proportion, by the parameters of each marginal distributions and by the correlation matrix of the Gaussian copula providing one coefficient per couple of variables measuring the intra-class dependency. In addition, a principal component analysis (PCA) computed per class is a straightforward by-product of the model. Indeed, it is computed on the correlation matrix of the class and it can be used to summarize the main intra-class dependencies and to provide a scatter-plot of the individuals according to the class parameters.

This paper is organized as follows. Section 2 presents the mixture model of Gaussian copulas for clustering, its links with the existing models and its contribution to the visualization of mixed variables. Section 3 is devoted to the parameter estimation in a Bayesian framework. Section 4 illustrates the model by a real data set clustering. Section 5 concludes this work.

2 Mixture model of Gaussian copulas

Finite mixture model

Let the vector of e mixed variables $\mathbf{x} = (x^1, \dots, x^e) \in \mathbb{R}^c \times \mathcal{X}$, whose the first c elements are the set of the continuous variables further denoted by \mathbf{x}^c , and whose the last d elements are the set of the discrete variables (integer, ordinal or binary) further denoted by \mathbf{x}^d , with $e = c + d$. Note that if x^j is an ordinal variable with m_j modalities, then it uses a numeric coding $\{1, \dots, m_j\}$. Data \mathbf{x} are supposed to be drawn by the mixture model of g parametric distributions whose the probability distribution function (pdf) is written as

$$p(\mathbf{x}; \boldsymbol{\theta}) = \sum_{k=1}^g \pi_k p(\mathbf{x}; \boldsymbol{\alpha}_k), \quad (1)$$

where $\boldsymbol{\theta} = (\boldsymbol{\pi}, \boldsymbol{\alpha})$ and where $\boldsymbol{\pi} = (\pi_1, \dots, \pi_g)$ groups the proportions of each class k denoted by π_k , and respects the following constraints $0 < \pi_k \leq 1$ and $\sum_{k=1}^g \pi_k = 1$, while $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_1, \dots, \boldsymbol{\alpha}_g)$ groups the parameters of each class k denoted by $\boldsymbol{\alpha}_k$.

One-dimensional margins of the components

The margin distribution of x^j , for the component k , belongs to the exponential family and has $p(x^j; \boldsymbol{\beta}_{kj})$ for pdf and $P(x^j; \boldsymbol{\beta}_{kj})$ for cumulative distribution function (cdf). More precisely, the margin distribution of each component is a *Gaussian* (if x^j is *continuous*), *Poisson* (if x^j is *integer*) or *multinomial* (if x^j is *ordinal*) distribution where $\boldsymbol{\beta}_{kj}$ denotes the usual parameters.

Dependency model of the components

The model assumes that each component k follows a Gaussian copula whose the correlation matrix is $\mathbf{\Gamma}_k$. We note $\Phi_e(\cdot; \mathbf{\Gamma}_k)$ the cdf of the e -variate centred Gaussian distribution with correlation matrix $\mathbf{\Gamma}_k$, and $\Phi_1^{-1}(\cdot)$ the inverse cumulative distribution function of univariate Gaussian variable $\mathcal{N}_1(0, 1)$. Thus, the cdf of the component k is written as

$$P(\mathbf{x}; \boldsymbol{\alpha}_k) = \Phi_e(\Phi_1^{-1}(u_k^1), \dots, \Phi_1^{-1}(u_k^e); \mathbf{0}, \mathbf{\Gamma}_k), \tag{2}$$

where $u_k^j = P(x^j; \boldsymbol{\beta}_{kj})$, $\boldsymbol{\alpha}_k = (\boldsymbol{\beta}_k, \mathbf{\Gamma}_k)$ and $\boldsymbol{\beta}_k = (\boldsymbol{\beta}_{k1}, \dots, \boldsymbol{\beta}_{ke})$.

Remark 2.1 (Standardized coefficient of correlation per class).

The Gaussian copula provides a robust coefficient of correlation per couple of variables. Indeed, when both variables are continuous, it is equal to the upper bound of the coefficient of correlation obtained by all the monotonic transformations of the variables [10]. Furthermore, when both variables are discrete, it is equal to the polychoric coefficient of correlation [13].

Remark 2.2 (Two latent variables).

The mixture model of Gaussian copulas involves two latent variables: a categorical one using a condense coding $z \in \{1, \dots, g\}$ denoting the class membership and an e -variate Gaussian one $\mathbf{y} = (y^1, \dots, y^e) \in \mathbb{R}^e$. Indeed, if $\mathbf{y}|z = k \sim \mathcal{N}_e(\mathbf{0}, \mathbf{\Gamma}_k)$ and if $x^j = P^{-1}(\Phi_1(y^j); \boldsymbol{\beta}_{kj})$, $\forall j = 1 \dots, e$, then the component k is a Gaussian copula whose the cdf is defined in (2). Thus, we deduce the following generative model

- *Class membership sampling: $z \sim \mathcal{M}_g(\pi_1, \dots, \pi_g)$*
- *Gaussian copula sampling: $\mathbf{y}|z = k \sim \mathcal{N}_e(\mathbf{0}, \mathbf{\Gamma}_k)$*
- *Observed data deterministic computation of \mathbf{x} as such $x^j = P^{-1}(\Phi_1(y^j); \boldsymbol{\beta}_{kj})$.*

Probability distribution function of the components

We introduce the function $\Psi(\mathbf{x}^c; \boldsymbol{\alpha}_k) = \left(\frac{x^j - \mu_{kj}}{\sigma_{kj}}; j = 1, \dots, c\right)$ and the space of the antecedents of \mathbf{x}^d in the class k , by $\mathcal{S}_k = \mathcal{S}_k^{c+1} \times \dots \times \mathcal{S}_k^e$, where \mathcal{S}_k^j is the interval defined by $\mathcal{S}_k^j =]b_k^\ominus(x^j), b_k^\oplus(x^j)]$, for $j = c + 1, \dots, e$, whose the bounds are $b_k^\ominus(x^j) = \Phi_1^{-1}(P(x^j - 1; \boldsymbol{\beta}_{kj}))$ and $b_k^\oplus(x^j) = \Phi_1^{-1}(P(x^j; \boldsymbol{\beta}_{kj}))$. The pdf of the component k is written as

$$p(\mathbf{x}; \boldsymbol{\alpha}_k) = p(\mathbf{x}^c; \boldsymbol{\alpha}_k)p(\mathbf{x}^d|\mathbf{x}^c; \boldsymbol{\alpha}_k) \tag{3}$$

$$= \frac{\phi_c(\Psi(\mathbf{x}^c; \boldsymbol{\alpha}_k); \mathbf{0}, \mathbf{\Gamma}_{kCC})}{\prod_{j=1}^c \sigma_{kj}} \int_{\mathcal{S}_k} \phi_d(\mathbf{u}; \boldsymbol{\mu}_k^d, \boldsymbol{\Sigma}_k^d) d\mathbf{u}, \tag{4}$$

where $\mathbf{\Gamma}_k = \begin{bmatrix} \mathbf{\Gamma}_{kCC} & \mathbf{\Gamma}_{kCD} \\ \mathbf{\Gamma}_{kDC} & \mathbf{\Gamma}_{kDD} \end{bmatrix}$ is decomposed into sub-matrices, for instance $\mathbf{\Gamma}_{kCC}$ is the sub-matrix of the first c rows and columns of $\mathbf{\Gamma}_k$, where $\boldsymbol{\mu}_k^d = \mathbf{\Gamma}_{kDC} \mathbf{\Gamma}_{kCC}^{-1} \Psi(\mathbf{x}^c; \boldsymbol{\alpha}_k)$ is the conditional mean of \mathbf{y}^d and where $\boldsymbol{\Sigma}_k^d = \mathbf{\Gamma}_{kDD} - \mathbf{\Gamma}_{kDC} \mathbf{\Gamma}_{kCC}^{-1} \mathbf{\Gamma}_{kCD}$ is its conditional covariance matrix.

Heteroscedastic and homoscedastic versions of the model

The trade off between the bias and the variance of the model may be improved by adding some constraints on the parameter space. Thus, we propose an homoscedastic version of the mixture model of Gaussian copulas by assuming the equality between the correlation matrices, so

$$\mathbf{\Gamma}_1 = \dots = \mathbf{\Gamma}_g. \tag{5}$$

The heteroscedastic (resp. homoscedastic) mixture model of Gaussian copulas requires ν_{He} (respectively ν_{Ho}) parameters where

$$\nu_{\text{He}} = (g - 1) + g \left(\frac{e(e + 1)}{2} + d \right) \text{ and } \nu_{\text{Ho}} = (g - 1) + \frac{e(e - 1)}{2} + g(e + d). \quad (6)$$

Related models

The mixture model of Gaussian copulas allows to generalize many classical mixture models, among them one can cite the four followers.

- Obviously, if the correlation matrices are diagonal (*i.e.* $\mathbf{\Gamma}_k = \mathbf{I}, \forall k = 1, \dots, g$), then the mixture model of Gaussian copulas is equivalent to the *locally independent mixture model*.
- If all the variables are continuous (*i.e.* $c = e$ and $d = 0$), then both versions of the heteroscedastic and homoscedastic mixture models of Gaussian copulas are equivalent to the heteroscedastic and homoscedastic *multivariate Gaussian mixture models* [1].
- The mixture model of Gaussian copulas is linked to the *binned Gaussian mixture model*. For instance, it is equivalent, when data are ordinal, to the mixture model of [5]. In such a case and under the true model assumption, this model is stable by fusion of modalities.
- When the variables are continuous and ordinal, the mixture model of Gaussian copulas is a new parametrization of *model proposed by Everitt* [4] which directly estimates the space S_k containing the antecedents of \mathbf{x}^{D} and not the margin parameters. The maximum likelihood inference is performed via a simplex algorithm dramatically limiting the number of ordinal variables. Note that our approach detailed in Section 3 avoids this drawback.

Data visualization per class: a by-product of Gaussian copulas

We can use the model parameters to perform a *visualization* of the individuals *per class* and to bring out the main intra-class dependencies. Thus, for the class k , we firstly compute the coordinates $\mathbb{E}[\mathbf{y}|\mathbf{x}, z = k; \boldsymbol{\alpha}_k]$ and we secondly project them on the principal component analysis space of the Gaussian copula of the component k , obtained by the spectral decomposition of $\mathbf{\Gamma}_k$. The individuals drawn by the component k follow a centred Gaussian distribution in the factorial map (so they are close to the origin) while the other ones have an expectation different to zero (so they are farther from the origin). Finally, the correlation circle summarizes the intra-class correlations. The application given in Section 4 illustrates this phenomenon.

3 Bayesian inference

We observe a sample $\mathbf{x} = (\mathbf{x}_1, \dots, \mathbf{x}_n)$ composed by n individuals $\mathbf{x}_i \in \mathbb{R}^c \times \mathcal{X}$ assumed to be independently drawn by a mixture model of Gaussian copulas. We assume the independence between the prior distributions and we select the classical conjugate prior distributions for each parameters. The following Gibbs sampler allows to perform the inference, in a Bayesian framework, since its stationary distribution is $p(\boldsymbol{\theta}, \mathbf{z}|\mathbf{x})$. Thus, it samples a sequel of parameters according to the marginal posterior distribution $p(\boldsymbol{\theta}|\mathbf{x})$. This algorithm relies on two instrumental variables: the class membership of the individuals of \mathbf{x} denoted by $\mathbf{z} = (z_1, \dots, z_n)$ and the Gaussian vector of the individuals denoted by $\mathbf{y} = (\mathbf{y}_1, \dots, \mathbf{y}_n)$.

Algorithm 3.1 (The Gibbs sampler).

Starting from an initial value $\theta^{(0)}$, its iteration (r) is written as

$$\mathbf{z}^{(r)}, \mathbf{y}^{(r-1/2)} \sim \mathbf{z}, \mathbf{y} | \mathbf{x}, \theta^{(r-1)} \quad (7)$$

$$\beta_{kj}^{(r)}, \mathbf{y}_{[rk]}^{j(r)} \sim \beta_{kj}, \mathbf{y}_{[rk]}^j | \mathbf{x}, \mathbf{y}_{[rk]}^{\uparrow j(r)}, \mathbf{z}^{(r)}, \beta_k^{\uparrow j(r)}, \Gamma_k^{(r-1)} \quad (8)$$

$$\pi^{(r)} \sim \pi | \mathbf{z}^{(r)} \quad (9)$$

$$\Gamma_k^{(r)} \sim \Gamma_k | \mathbf{y}^{(r)}, \mathbf{z}^{(r)}, \quad (10)$$

where $\mathbf{y}_{[rk]} = \mathbf{y}_{\{i: z_i = k\}}$, $\mathbf{y}_i^{\uparrow j(r)} = (y_i^{1(r)}, \dots, y_i^{j-1(r)}, y_i^{j+1(r-1/2)}, \dots, y_i^{e(r-1/2)})$ and $\beta_k^{\uparrow j(r)} = (\beta_{k1}^{(r)}, \dots, \beta_{kj-1}^{(r)}, \beta_{kj+1}^{(r-1)}, \dots, \beta_{ke}^{(r-1)})$.

Remark 3.2 (Twice sampling of the Gaussian variable).

The Gaussian variable \mathbf{y} is twice generated during one iteration of the Gibbs sampler but, obviously, its stationary distribution stays unchanged. This twice sampling is mandatory because of the strong dependency between \mathbf{y} and \mathbf{z} , and between $\mathbf{y}_{[rk]}^j$ and β_{kj} .

Remark 3.3 (On the Metropolis-within-Gibbs sampler).

If the samplings from (9) and (10) are classical, the two other ones are more complex. Indeed, the sampling from (7) involves to compute the conditional probabilities of the class memberships, so to compute the integral defined in (4). If the number of discrete variables is large, this computation is time consuming. However, the sampling from (7) can be efficiently performed by one iteration of a Metropolis-Hastings algorithm having $p(z_i, \mathbf{y}_i | \mathbf{x}_i, \theta^{(r-1)})$ as stationary distribution. Concerning the sampling according to (8), it is performed in two steps. Firstly, the margin parameter is sampled by one iteration of a Metropolis-Hastings algorithm having $p(\beta_{kj} | \mathbf{x}, \mathbf{y}_{[rk]}^{\uparrow j(r)}, \mathbf{z}^{(r)}, \beta_k^{\uparrow j(r)}, \Gamma_k)$ as stationary distribution. Secondly, the latent Gaussian vector is sampled from its full conditional distribution.

Remark 3.4 (Initialization of the algorithm).

The algorithm is initialized on the maximum likelihood estimate of the locally independent model. Thus, it is initialized in a point close to the maximum of the posterior distribution if the variables are not strongly intra-class correlated.

4 Application: clustering of Portuguese wines

The data The data set [3] contains 6497 variants of the Portuguese ‘‘Vinho Verde’’ wine (1599 red wines and 4898 white wines) described by eleven physiochemical continuous variables (fixed acidity, volatile acidity, citric acidity, residual sugar, chlorides, free sulfur dioxide, total density dioxide, density, pH, sulphates, alcohol) and one integer variable (quality of the wine evaluated by experts). The kinds of the wines (red or white) are hidden and we cluster the data set by excluding of the study one white wine (number 4381) since it is an outlier.

Model selection We estimate the three mixture models (locally independent one, the heteroscedastic and homoscedastic versions of the mixture model of Gaussian copulas) for different numbers of classes. The estimate is obtained by taking the mean of the sampled parameters computed after 1000 iterations. The model selection is performed by using two information criteria (BIC criterion [14], ICL criterion [2]) computed on the maximum *a posteriori* estimate.

We present the values of both used information criteria in Table 1 which distinctly select the bi-component heteroscedastic mixture model of Gaussian copulas.

	g	1	2	3	4	5	6
BIC	loc. indpt.	-63516	-61069	-61010	-55967	-60250	-57163
	hetero.	-44675	-34520	-39724	-44692	-44484	-48349
	homo.	-44675	-39372	-38289	-45209	-43217	-42417
ICL	loc. indpt.	-63516	-61229	-61365	-56310	-60726	-58138
	hetero.	-44675	-34688	-40176	-44933	-44758	-48959
	homo.	-44675	-39607	-38791	-45380	-43345	-42667

Table 1: Values of the BIC and ICL criteria for the three mixture models estimated.

Partition comparison Table 2 presents the values of the adjusted Rand index and the confusion matrices in order to compare the relevance of the estimated partitions according to the true one (wine color). These results confirm that the bi-component heteroscedastic Gaussian copula mixture model is the best one among the competing models since its partition is the closest to the true one.

	white	red		white	red		white	red
class 1	4359	9	class 1	2441	12	class 1	2547	1561
class 2	538	1590	class 2	1911	7	class 2	2007	35
(a) Adj. Rand.: 0.68			class 3	545	1580	class 3	275	3
			(b) Adj. Rand.: 0.30			class 4	68	0
						(c) Adj. Rand.: 0.00		

Table 2: Adjusted Rand indices and confusion matrices related to: (a) the bi-component heteroscedastic Gaussian copula mixture; (b) the tri-component homoscedastic Gaussian copula mixture; (c) the four-component locally independent mixture.

Visualization Figure 1 displays the individuals in a PCA map of both classes estimated by the bi-component free mixture model of Gaussian copulas. According to these scatter-plots, classes are well-separated.

Interpretation of the best model The following interpretation is based on the margin parameters and on the intra-class correlation matrices summarized in Figure 2. The majority class ($\pi_1 = 0.59$) is principally composed by white wines. This class is characterized by lower rates of acidity, pH, chlorides and sulphites than them of the minority class ($\pi_2 = 0.41$) which is principally composed by red wines. The majority class has larger values for both sulfur dioxide measures and the alcoholic rate. Note than the wine quality of both classes is similar ($\beta_{1\text{quality}} = 5.96$ and $\beta_{2\text{quality}} = 5.58$). The majority class is characterized by a strong correlation between both sulfur measures opposite to a strong correlation between the density and acidity measures. The minority class underlines that the wine quality is dependent with a larger alcoholic rate and small values for the chlorides and acidity measures.

Conclusion On this data set, the mixture model of Gaussian copulas overcomes the locally independent model (reduction of the number of classes, better values of the information criteria,

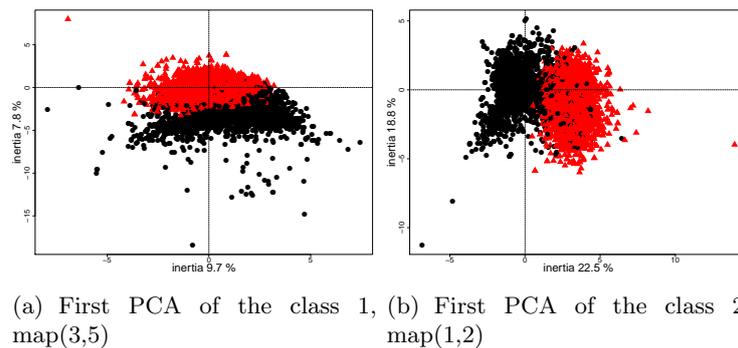


Figure 1: Visualization of the partition by the bi-component heteroscedastic mixture model of Gaussian copulas (Class 1 is drawn by black circles and Class 2 by red triangles).

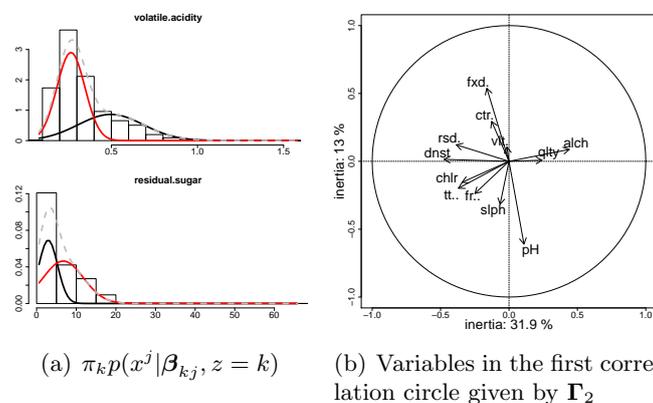


Figure 2: Summary of the bi-component heteroscedastic mixture model of Gaussian copula. Class 1 is drawn in black and Class 2 in red. (fixed acidity: fxd., volatile acidity: vlt., citric acidity: ctr., residual sugar: rsd., chlorides: chl., free sulfur dioxide: fr., total density dioxide: tt., density: dnst., pH, sulphates: slph., alcohol: alch., quality: qlty.).

estimated partition closest to the true one). Based on the individual scatter-plots in the model PCA, the estimated classes are relevant since they are well-separated. Finally, the estimation of the intra-class dependencies helps the interpretation since it underlines the link between the wine quality of the minority class and its physiochemical properties.

5 Conclusion and future extensions

The proposed model uses the properties of copulas: independent choice of the margin distributions and of the dependency relations. Thus, the mixture model of Gaussian copulas allows to fix classical margins belonging to the exponential family for the component margin distributions and takes into account the intra-class dependencies. An approach based on a PCA per class of the Gaussian latent variable allows to summarize the main intra-class dependencies and to visualize the data by using the model parameters. The application points out that this model

is sufficiently flexible to efficiently fit data and that it can reduce the biases of the locally independent model (for instance the reduction of the number of classes). The number of parameters increases with the number of classes and variables especially because of the correlation matrices of the Gaussian copulas. To avoid this drawback, we propose an homoscedastic version of the model assuming the equality between the correlation matrices. This model may better fit the data than the heteroscedastic Gaussian mixture models.

Bibliography

- [1] J.D. Banfield and A.E. Raftery. Model-based Gaussian and non-Gaussian clustering. *Biometrics*, pages 803–821, 1993.
- [2] C. Biernacki, G. Celeux, and G. Govaert. Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(7):719–725, 2000.
- [3] P. Cortez, A. Cerdeira, F. Almeida, T. Matos, and J. Reis. Modeling wine preferences by data mining from physicochemical properties. *Decision Support Systems*, 47(4):547–553, 2009.
- [4] B.S. Everitt. A finite mixture model for the clustering of mixed-mode data. *Statistics & Probability Letters*, 6(5):305–309, 1988.
- [5] C. Gouget. *Utilisation des modèles de mélange pour la classification automatique de données ordinales*. PhD thesis, Université de Technologie de Compiègne, 2006.
- [6] D.J. Hand and K. Yu. Idiot’s Bayes - Not So Stupid after All? *International Statistical Review*, 69(3):385–398, 2001.
- [7] P.D. Hoff. Extending the rank likelihood for semiparametric copula estimation. *The Annals of Applied Statistics*, pages 265–283, 2007.
- [8] L. Hunt and M. Jorgensen. Clustering mixed data. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 1(4):352–361, 2011.
- [9] H. Joe. *Multivariate models and multivariate dependence concepts*, volume 73. CRC Press, 1997.
- [10] C.A.J. Klaassen and J.A. Wellner. Efficient estimation in the bivariate normal copula model: normal margins are least favourable. *Bernoulli*, 3(1):55–77, 1997.
- [11] D.D. Lewis. Naive (Bayes) at forty: The independence assumption in information retrieval. In *Machine learning: ECML-98*, pages 4–15. Springer, 1998.
- [12] G.J. McLachlan and D. Peel. *Finite mixture models*. Wiley Series in Probability and Statistics: Applied Probability and Statistics, Wiley-Interscience, New York, 2000.
- [13] U. Olsson. Maximum likelihood estimation of the polychoric correlation coefficient. *Psychometrika*, 44(4):443–460, 1979.
- [14] G. Schwarz. Estimating the dimension of a model. *Annals of Statistics*, 6:461–464, 1978.

Sampling inspection by (Gaussian) variables via estimation of the lot fraction defective: a computational approach

Miguel Casquilho, *Department of Chemical Engineering, Instituto Superior Técnico, Universidade de Lisboa (University of Lisbon), mcasquilho@tecnico.ulisboa.pt*
Elisabete Carolino, *Escola Superior de Tecnologia da Saúde de Lisboa, Instituto Politécnico de Lisboa (Polytechnic Institute of Lisbon), lizcarolino@gmail.com*

Abstract. Quality Control has lost impetus in the last decades toward managerial features that evade the intricacies of Statistics, but these can, in the computer age, be made comfortable, namely through the Internet. In Quality Control, acceptance sampling (AS) *by variables* (as opposed to *by attributes*) often assumes, as we do here, that the quality characteristic is a Gaussian variable, and has, as decision criterion on the lot, the comparison of the quality index with the acceptance constant (Form 1). This criterion is simple and applies only to the case, addressed here, of a single specification limit, but can be confronted with another (Form 2), mathematically equivalent, to which attention is drawn in this paper. In this latter, the decision is based on the comparison of the estimated "lot percent defective" with its maximum, critical value. Transforming the former criterion into the latter is done by the incomplete beta ratio function, for the computing of which we prepared a computer program and an open webpage. So nowadays either criterion becomes easy to be adopted by the decision maker, with the advantage going to the latter, Form 2, which presents intuitive results.

Keywords. Quality Control, acceptance sampling, inspection by variables, Gaussian variable, international standards, "Form 2".

1 Fundamentals and scope

Quality is currently a general concern in every productive activity, but in the last few decades it has lost impetus toward other managerial features that evade the rigorous facets and intricacies of Statistics, as acutely observed, *e.g.*, by Gunter in a blunt article ([8]). Otherwise, there is

no motive why nowadays the harder, computation-based aspects of Quality should not be more easily made available to the users, as we propose in this study, namely through the Internet.

From a statistical standpoint, Quality Control is usually divided in two broad categories, acceptance sampling (AS), and statistical process control (SPC), the former to be applied in the frontiers of the production system and the latter inside of the system. In this regard, an argument used against AS is its uselessness due to the stable interest in SPC, together with the cooperation with the suppliers, both of which indeed reduce the need for AS. Nevertheless, the fact that AS proper continues to be necessary is attested, not just by the many classical studies, *e.g.*, [13], [15], but by the recent update (in 2013) of the successor to the original Mil-Std 414 ([12]), the corresponding ISO standard ([10]).

Acceptance sampling decides on the quality of a lot from the observation of a random sample taken from it, and deals with variables that can be discrete (control *by attributes*, counting nonconformities) or continuous (control *by variables*). In this paper, we address the control by variables of continuous, Gaussian variables with a single specification limit, as treated in the applicable international standards for AS by variables ([2]).

The standards establish two mathematically equivalent decision criteria on the lot for a single specification limit, the so-called “Form 1” and “Form 2”. In the former, a comparison is made between the *quality index*, Q , and the *acceptance constant*, k , acceptance occurring iff $Q \geq k$; and in the latter (mandatory for double specification limits), a comparison is made between the *estimated lot percent defective* (fraction nonconforming), ϖ , and its *maximum*, M , acceptance occurring iff $\varpi \leq M$. The procedure in Form 1 is simple to apply, but can be confronted with the richer information yielded by Form 2, to which attention is drawn in this paper. This becomes computationally accessible, as will be seen, and is generally advantageous, namely, to non-specialized decision makers.

2 Sampling plan

Underlying an AS procedure is a certain *sampling plan*, which gives, as is well known, the size of the random sample to be drawn, n , and the critical value of the test statistic, k , leading to the criterion to accept or reject the lot of given size, under inspection. In order to try to avoid the rejection of “good” lots (Type I error), and the acceptance of “bad” lots (Type II error), under Form 1, the calculation of n and k results from (*e.g.*, [3], [6]) the resolution of the classical system of inequalities

$$\begin{cases} P_{ac}(\varpi = \text{AQL}) > 1 - \alpha \\ P_{ac}(\varpi = \text{LTPD}) < \beta, \end{cases} \quad (1)$$

where: P_{ac} is the probability of acceptance (a function of n and k); AQL, “Acceptance Quality Limit”, is the maximum fraction defective (nonconforming) that corresponds to the producer’s risk, α ; and LTPD, “Lot Tolerance Percent Defective”, is the maximum fraction defective that corresponds to the consumer’s risk, β . Note that n has, of course, to be integer. (The nomenclature in Quality Control bears the tradition of informal terms, such as “percent” instead of “fraction”, used since its inception in the early 20.th century, to make it easier for the laymen to apply it.) The following parameters are stipulated according to the situation (example values): AQL = 1.5%, α = 5%, LTPD = 12%, and β = 10%.

Regarding Form 1, exemplified here for the (arbitrarily chosen) lower specification limit, the lot acceptance criterion is given by the following condition, in which \bar{X} and S are the sample average and standard deviation, respectively.

$$Q_L = \frac{\bar{X} - L}{S} \geq k \quad (2)$$

where Q_L is the quality index, *i.e.*, Q (general) referred to the lower specification limit, L , and k the acceptance constant. In this equation, the *equals* sign—which is meaningless in terms of probability—is important because, upon application, the comparison is made with Q rounded according to the significant figures in k . For the upper specification limit, the numerator is changed to $U - \bar{X}$, for practical convenience (so that a higher, positive, quality index always means better quality).

In order to transform the quality index, Q (typically used to decide lot acceptance when the quality characteristic is a Gaussian variable) into the estimated lot percent (proportion) defective, ϖ , the equation presented below is used, leading to Form 2. While, as mentioned, in Form 1, the lot is accepted iff $Q \geq k$, in Form 2 it is accepted iff $\varpi \leq M$, with M a critical value, defined below, Form 2 being more intuitive to the non-specialized decision makers.

Form 2, optional in the case of a single specification limit, whether lower, L , or upper, U (as mentioned, mandatory for two limits), comes from the transformation ([14]) of both terms of the comparison in Eq. 2 into an estimate of the lot fraction defective, ϖ_L , depending on n , and its critical value, M , *i.e.*, maximum acceptable fraction, compatible with AQL and n . The transformation comes from the application to each side of Eq. 2 ([9], [14]) of

$$\varpi = F\left(x, \frac{n}{2} - 1, \frac{n}{2} - 1\right) \quad (3)$$

where F represents the “incomplete beta ratio function”, *i.e.*, the cumulative distribution function (‘cdf’), with $n > 2$ ($n = 2$ making the two parameters 0 in Eq. 3), and with (equal) parameters $\frac{n}{2} - 1$, and x is

$$x = \max\left(0, \frac{1}{2} - \frac{1}{2}Q \frac{\sqrt{n}}{n-1}\right) \quad (4)$$

From Eqs. 2 and 3 will come the alternative criterion of acceptability, for a single specification limit (in this case, the lower one),

$$\varpi_L \leq M_L. \quad (5)$$

When the two specification limits are present (not addressed in this article), the acceptance criterion becomes the following three simultaneous conditions ([9]):

$$\begin{aligned} \varpi_L &\leq M_L \\ \varpi_U &\leq M_U \\ \varpi_L + \varpi_U &\leq \max(M_L, M_U) \end{aligned} \quad (6)$$

In the criterion in Eq. 5 or the set of conditions in Eq. 6 (besides having the third condition), the decision on the lot is intuitive, as comparisons are simply between percent (fraction)

TABLE B-5—Continued
Table for Estimating the Lot Percent Nonconforming Using Standard Deviation Method¹

Q_U or Q_L	Sample Size														
	3	4	5	7	10	15	20	25	30	35	50	75	100	150	200
1.50	0.00	0.00	3.80	5.28	5.87	6.20	6.34	6.41	6.46	6.50	6.55	6.60	6.62	6.64	6.65
1.51	0.00	0.00	3.61	5.13	5.73	6.06	6.20	6.28	6.33	6.36	6.42	6.47	6.49	6.51	6.52
1.52	0.00	0.00	3.42	4.97	5.59	5.93	6.07	6.15	6.20	6.23	6.29	6.34	6.36	6.38	6.39
1.53	0.00	0.00	3.23	4.82	5.45	5.80	5.94	6.02	6.07	6.11	6.17	6.21	6.24	6.26	6.27
1.54	0.00	0.00	3.05	4.67	5.31	5.67	5.81	5.89	5.95	5.98	6.04	6.09	6.11	6.13	6.15
1.55	0.00	0.00	2.87	4.52	5.18	5.54	5.69	5.77	5.82	5.86	5.92	5.97	5.99	6.01	6.02
1.56	0.00	0.00	2.69	4.38	5.05	5.41	5.56	5.65	5.70	5.74	5.80	5.85	5.87	5.89	5.90
1.57	0.00	0.00	2.52	4.24	4.92	5.29	5.44	5.53	5.58	5.62	5.68	5.73	5.75	5.78	5.79
1.58	0.00	0.00	2.35	4.10	4.79	5.16	5.32	5.41	5.46	5.50	5.56	5.61	5.64	5.66	5.67
1.59	0.00	0.00	2.19	3.96	4.66	5.04	5.20	5.29	5.34	5.38	5.45	5.50	5.52	5.54	5.56
1.60	0.00	0.00	2.03	3.83	4.54	4.92	5.09	5.17	5.23	5.27	5.33	5.38	5.41	5.43	5.44
1.61	0.00	0.00	1.87	3.69	4.41	4.81	4.97	5.06	5.12	5.16	5.22	5.27	5.30	5.32	5.33
1.62	0.00	0.00	1.72	3.57	4.30	4.69	4.86	4.95	5.01	5.04	5.11	5.16	5.19	5.21	5.23
1.63	0.00	0.00	1.57	3.44	4.18	4.58	4.75	4.84	4.90	4.94	5.01	5.06	5.08	5.11	5.12
1.64	0.00	0.00	1.42	3.31	4.06	4.47	4.64	4.73	4.79	4.83	4.90	4.95	4.98	5.00	5.01
1.65	0.00	0.00	1.28	3.19	3.95	4.36	4.53	4.62	4.68	4.72	4.79	4.85	4.87	4.90	4.91

Figure 1: Excerpt from the table in [10] to transform Q (or k) into ϖ (or M).

defectives instead of values of Q and k , all the more dependent on n . These calculations are shown below and made available on our website.

3 Computation

The transformation of values of the quality index, Q , into ϖ (and the corresponding critical value of Q , the acceptance constant, k , into M) is made available in the standard through a quite extensive table (ten pages), of which a small excerpt is shown in Figure 1.

For further verification of the computation done in our website, some values taken from the complete table are shown in Table 1.

Q	$n = 5$	$n = 10$	$n = 35$
1.50	3.80	5.87	6.50
1.65	1.28	3.95	4.72
1.75	0.19	2.93	3.72
1.80	0	2.49	3.35
2.00	0	1.17	2.02
2.50	0	0.04	0.45

Table 1: Values (%) from the table in [10] for verification.

In order to verify the values in Table 1, the computation of ϖ as a function of Q for a given n can be done at our dedicated webpage ([7]), through a computer program of ours that implements Eq. 3: the values are thoroughly confirmed.

Notice that the values of M , critical values of ϖ , are themselves the transformation of the acceptance constants, k (given in the standard). These constants are, of course, critical values of Q , the quality index, which, in the form $Q\sqrt{n}$, follows a noncentral t-distribution. This is not addressed here, but can be computed in one of our webpages ([5]) by Monte Carlo simulation and directly.

The “incomplete beta ratio function” (Eq. 3) is usually denoted by $I_x(\alpha, \beta)$, and is given by the following expression,

$$I_x(\alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} \int_0^x t^{\alpha-1}(1-t)^{\beta-1} dt \quad (7)$$

where α and β are parameters. The integral in Eq. 7 must be computed numerically, but becomes easy as it benefits from some peculiarities: (a) the Γ function, in this application, has an integer or half-integer argument (as $\alpha = \beta = \frac{n}{2} - 1$), so its computing is straightforward (factorials or multiples of $\sqrt{\pi}$); (b) the integrand is “well behaved”; and, (c) thus, a “simple Euler” or Simpson’s rule integration can be used. As the computing of the function is necessary for many (successive) values of x , a progressive form of the numerical integration is computationally convenient, which (generally overlooked in the common literature) was done according to our previous practice ([4]). This progressive form was precisely necessary to make Figure 2.

The webpage mentioned ([7]) is open to anyone wishing to do the transformation, showing how acceptance sampling by variables Form 2 can be easily used.

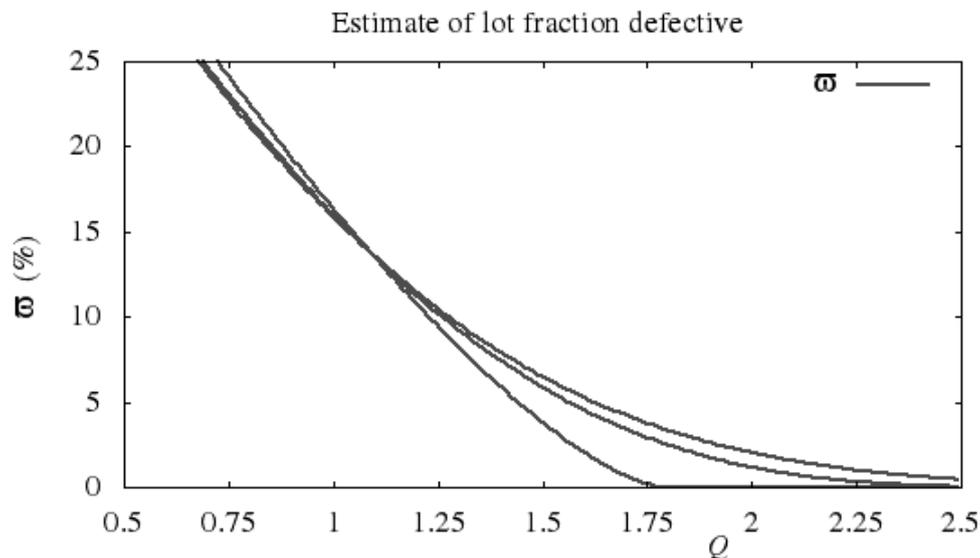


Figure 2: Variation of ϖ with Q for $n = 5, 10, 35$ (respectively, right hand side curves upwards).

Conclusions

Quality Control (QC) has lost impetus towards many current managerial directions, which try to avoid the intricacies of Statistics. The current availability of computing power, namely through the Internet, makes QC accessible, even to non-specialists. Thus, of the two branches

of statistical QC, acceptance sampling and statistical process control, the former can now be more approachable.

The application of AS *by variables* to the typical Gaussian random variable, according to the generally adopted international standards, was shown in its more intuitive and informative “Form 2”, where clear, simple percentages are made available to the decision maker. The underlying computations were mentioned, and we prepared an open website available to anyone wishing to transform a quality index, Q , into ϖ , an estimate of the lot percent defective (fraction nonconforming).

Acknowledgement

We thank: (M. Casquilho) the Department of Chemical Engineering, Instituto Superior Técnico (IST), and CERENA, “Centro de Recursos Naturais e Ambiente” (*Centre for Natural Resources and the Environment*), which has included “Centro de Processos Químicos” (*Centre for Chemical Processes*), at IST, Universidade de Lisboa (*University of Lisbon*), Lisbon, Portugal; and (E. Carolino) ESTeSL, Escola Superior de Tecnologia da Saúde de Lisboa (*Lisbon School of Health Technology*), Polytechnic Institute of Lisbon, Lisbon, Portugal. We also thank CIIST (*Computing Centre of IST*), and “Milipeia” (Laboratory for Advanced Computing), University of Coimbra. The anonymous referees’ comments were also beneficial to the clarity of the text.

Bibliography

- [1] ANSI/ASQC Z1.9-2008 (2008) *Sampling procedures and tables for inspection by variables for percent nonconforming*. ASQ, Milwaukee, WI (USA).
- [2] Carolino, E., Casquilho, M., and Barão, M. I. (2007) *Amostragem de aceitação para uma variável assimétrica, a exponencial*. (Acceptance sampling for an asymmetric variable, the Exponential.) Proceedings of the “XIV Congresso Anual da Sociedade Portuguesa de Estatística (XIV Annual Congress of the Portuguese Statistical Society), Covilhã (Portugal), 281–292.
- [3] Casquilho, M. (2010) *Numerical integration in tabular form*. ICEE-2010, Proceedings of the “International Conference on Engineering Education”, Gliwice (Poland).
- [4] Casquilho, M. (2012) *Convergence to non-central t curve.*, <http://web.tecnico.ulisboa.pt/~mcasquilho/compute/qc/F-tncConverg.php>, accessed 2014-Jan-15.
- [5] Casquilho, M. (2014) *Transformation of Q into ϖ in acceptance sampling by variables*, <http://web.tecnico.ulisboa.pt/mcasquilho/compute/or/Fx-ASvarQttoLotFD.php>, accessed 2014-Jan-15.
- [6] Gomes, M. I., Figueiredo, F., and Barão, M. I. (2010) *Controlo Estatístico da Qualidade* (Statistical control of Quality), 2.nd ed. revised and expanded, Ed. SPE (Portuguese Statistical Society), Lisbon (Portugal).
- [7] Gunter, B. (1998) *Farewell fusillade*. Quality Progress, **31**, No. 4, 111–114.

- [8] ISO 3951-2:2013 (2013) *Sampling procedures for inspection by variables – Part 2: General specification for single sampling plans indexed by acceptance quality limit (AQL) for lot-by-lot inspection of independent quality characteristics*. ISO, International Organization for Standardization, Geneva (Switzerland).
- [9] Lieberman, G. J., and Resnikoff, G. J. (1957) *Tables of the non-central t-distribution: density function, cumulative distribution function, and percentage points.*, Stanford University Press, Redwood City, CA (USA).
- [10] MIL-STD-414 (1957) *Sampling procedures and tables for inspection by variables for percent defective*. Office of the Assistant Secretary of Defense, Washington, D.C. (USA).
- [11] Owen, D. B. (1969) *Summary of recent work on variables acceptance sampling with emphasis on non-normality*. *Technometrics*, **11**, 4, 631–637.
- [12] Schilling, E. G., Neubauer, D. V. (2009) *Acceptance sampling in Quality Control*, 2.nd ed., Chapman and Hall / CRC, New York, NY (USA).
- [13] Wetherill, G. B., and Chiu, W. K. (1975) *A review of acceptance sampling schemes with emphasis on the economic aspect*. *International Statistical Review / Revue Internationale de Statistique*, **43**, 2, 191–210.

Estimation of the weighted kappa coefficient subject to case-control design

José Antonio Roldan-Nofuentes, *University of Granada*, jaroldan@ugr.es

Abstract. Assessment of the accuracy of a binary diagnostic test subject to a case-control sample is frequent in clinical practice. The estimation of the sensitivity and the specificity of the likelihood ratios of the diagnostic test is easily carried out as it consists of the estimation of binomial proportions and of ratios of binomial proportions respectively. Nevertheless, the estimation of parameters that depend on the disease prevalence is more complex and requires, from a frequentist perspective, knowledge of the disease prevalence. In this article, we study the estimation of the weighted kappa coefficient of a binary diagnostic test subject to a case-control sample. The weighted kappa coefficient is a parameter that depends on the sensitivity and the specificity of the diagnostic test, on the disease prevalence and the relative importance between the false negatives and the false positives. The estimation of this parameter requires knowledge of a value of the disease prevalence. Two confidence intervals are proposed which are based on the asymptotic normality of the estimator of the parameter: a Wald-type interval and another one based on the logit transformation. Simulation experiments were carried out to study the asymptotic coverage of these intervals. The results obtained were applied to a real example.

Keywords. Binary diagnostic test, Case-control design, Weighted kappa coefficient

1 Introduction

The most common parameters to assess the accuracy of a binary diagnostic test are the sensitivity and specificity, the likelihood ratios and the positive and negative predictive values. Moreover, when the losses of an erroneous classification with the diagnostic test are considered, the accuracy of the diagnostic test is measured in terms of the weighted kappa coefficient [1, 2]. The weighted kappa coefficient depends on the sensitivity (Se) and the specificity (Sp) of the diagnostic test, on the disease prevalence (p) and the weighting index (c). The weighting index c is a measure of the relative importance between the false negatives and the false positives. In a case-control design, the estimation of the sensitivity (specificity) is made from the sample of diseased (non-diseased)

individuals applying methods for binomial proportions. The positive and the negative likelihood ratios are estimated from both samples applying methods to estimate the ratio of independent binomial proportions. Nevertheless, the estimation of the positive and the negative predictive value requires, from a frequentist perspective, knowledge of the disease prevalence [3]. Mercaldo et al [3] studied the estimation of the predictive values of a binary diagnostic test subject to this type of sampling. . In this article, we study the estimation of the weighted kappa coefficient subject to case-control design, assuming that the disease prevalence is known. We have studied two asymptotic confidence intervals for the weighted kappa coefficient: a Wald-type interval and another interval based on the logit transformation. This study is organized as follows. In Section 2, we describe the weighted kappa coefficient. In Section 3, the two confidence intervals to be studied are presented, simulation experiments are carried out to study the asymptotic coverage of these intervals subject to case-control design and we describe a programme in R to solve this problem of estimation. In Section 4, the results are applied to a real example, and in Section 5 the results obtained are discussed.

2 Weighted kappa coefficient

Let L be the loss that occurs when for a diseased individual the result of the diagnostic test is negative, and let L' be the loss that occurs when for a non-diseased individual the result of the diagnostic test is positive. Loss L is associated with a false negative and loss L' is associated with a false positive. Losses L and L' are equal to zero if all of the individuals are classified correctly by the diagnostic test. For example, let us consider the diagnosis of breast cancer using as a diagnostic test a mammogram. If the mammogram is positive for a woman who does not have breast cancer, the woman will undergo a biopsy which will finally be negative. Loss L' will be determined from the economic costs of the diagnosis and also taking into account the risks, stress, etc, caused for the woman. If the mammogram is negative for a woman who has breast cancer, the woman may be diagnosed at a later stage, but the cancer may spread, reducing the possibility of successful treatment. In this situation, the cancer may spread and the chances of successful treatment will be reduced. Loss L will be determined from these considerations. Therefore, these losses are not only measured in economic terms but also with reference to other considerations, and for this reason in clinical practice it is not possible to determine the value of such losses [1]. Let $c = L/(L + L')$ be the weighting index, then the weighted kappa coefficient is expressed as [1, 2]

$$\kappa(c) = \frac{p(1-p)Y}{p(1-Q)c + (1-p)Q(1-c)}, \quad (1)$$

where $Q = pSe + (1-p)(1-Sp)$ and $Y = Se + Sp - 1$ is the Youden index. The weighting index is a measure of the relative loss between the false positives and the false negatives and varies between 0 and 1. If $L = 0$ then $c = 0$ and the weighted kappa coefficient is

$$\kappa(0) = \frac{Sp - (1 - Q)}{Q}. \quad (2)$$

If $L' = 0$ then $c = 1$ and the weighted kappa coefficient is

$$\kappa(1) = \frac{Se - Q}{1 - Q}. \quad (3)$$

The coefficients $\kappa(1)$ and $\kappa(0)$ are the chance-corrected sensitivity and the chance-corrected specificity respectively. If $L = L'$ then $c = 0.5$ and the weighted kappa coefficient (called the Cohen kappa coefficient) is

$$\kappa(0.5) = \frac{2\kappa(0)\kappa(1)}{\kappa(0) + \kappa(1)}. \tag{4}$$

The weighted kappa coefficient can be written as

$$\kappa(c) = \frac{p(1 - Q)c\kappa(1) + (1 - p)Q(1 - c)\kappa(0)}{p(1 - Q)c + (1 - p)Q(1 - c)}. \tag{5}$$

In practice, losses L and L' cannot be determined, and therefore the clinician usually allocates values to the weighting index depending on their knowledge of the relative importance of false positives and false negatives. Thus, for example, if the clinician decides that the false positives are twice as important as the false negatives, then the clinician will allocate the value $1/3$ to the weighting index c . The values of the weighted kappa coefficient vary between -1 and 1 . If the value of the weighted kappa coefficient is lower than 0 , then the results of the diagnostic test must be interchanged and therefore the analysis must be limited to positive values of the weighted kappa coefficient.

3 Estimation subject to case-control sampling

Let us consider a binary diagnostic test which is applied to two random samples, one of n_1 diseased individuals (case sample) and another one of n_2 non-diseased individuals (control sample). In Table 1 we can see the frequencies obtained when applying the diagnostic test to two samples.

Sample	Positive Test	Negative Test	Total
Case	s_1	s_0	n_1
Control	r_1	r_0	n_2

Table 1: Observed frequencies.

The estimators of sensitivity and specificity of the diagnostic test are

$$\hat{S}e = \frac{s_1}{n_1}, \tag{6}$$

and

$$\hat{S}p = \frac{r_0}{n_2}. \tag{7}$$

Assuming that the disease prevalence p is known, the estimator of the weighted kappa coefficient is [4]

$$\hat{\kappa}(c) = \frac{p(1 - p)\hat{Y}}{p(1 - \hat{Q})c + (1 - p)\hat{Q}(1 - c)}, \tag{8}$$

where $\hat{Y} = \hat{S}e + \hat{S}p - 1$ and $\hat{Q} = p\hat{S}e + (1-p)(1-\hat{S}p)$. Applying the delta method, the estimated variance of $\hat{\kappa}(c)$ is

$$\hat{V}ar(\hat{\kappa}(c)) = \frac{1}{[c\{1-\hat{S}p+p(\hat{Y}-1)\}-(1-p)\{1-\hat{S}p+p\hat{Y}\}]^4} \times \left[\frac{(1-p)^2 p^2 \{(1-p)(1-\hat{S}p)+c(\hat{S}p-(1-p))\}^2 \hat{S}e(1-\hat{S}e)}{n_1} + \frac{(1-p)^2 p^2 \{c(p-\hat{S}e)+\hat{S}e(1-p)\}^2 \hat{S}p(1-\hat{S}p)}{n_2} \right]$$

We now propose two confidence intervals (CIs) for the weighted kappa coefficient of a binary diagnostic test subject to case-control sampling.

Wald-type confidence interval

Based on the asymptotic normality of the weighted kappa coefficient, the Wald-type CI is

$$\hat{\kappa}(c) \pm z_{1-\alpha/2} \times \sqrt{\hat{V}ar(\hat{\kappa}(c))}, \quad (9)$$

where $z_{1-\alpha/2}$ is the $100(1-\alpha/2)$ th percentile of the normal standard distribution.

Logit confidence interval

In Statistics, it is common for a parameter not to be studied directly but instead one of its transformations is studied. Thus, for example, for a binomial proportion we can obtain a confidence interval based on the logit transformation [5]. Since the values of the weighted kappa coefficient are limited to values between 0 and 1 (as is explained in Section 2), a logit transformation can be used to obtain a confidence interval for this parameter. Based on the asymptotic normality of $\hat{\kappa}(c)$, its logit transformation, $logit(\hat{\kappa}(c))$, has a normal distribution with mean $logit(\kappa(c))$. Then the $100(1-z_{1-\alpha/2})\%$ confidence interval for the logit is

$$logit(\hat{\kappa}(c)) \pm z_{1-\alpha/2} \times \sqrt{\hat{V}ar(logit(\hat{\kappa}(c)))}, \quad (10)$$

and applying the delta method, the estimator of the variance of $logit(\hat{\kappa}(c))$ is

$$\hat{V}ar(logit(\hat{\kappa}(c))) = \frac{1}{[\hat{Y}\{(1-p)(1-c)\hat{S}p-cp(1-\hat{S}e)-(1-p)(1-c)\}]^2} \times \left[\frac{\{(1-p)(1-c)+c-(1-p)\}\hat{S}p\}^2 \hat{S}e(1-\hat{S}e)}{n_1} + \frac{\{(c-(1-p))\hat{S}e-cp\}^2 \hat{S}p(1-\hat{S}p)}{n_2} \right].$$

Finally, the logit CI for the weighted kappa coefficient is

$$\frac{\exp[logit(\hat{\kappa}(c)) \pm z_{1-\alpha/2} \sqrt{\hat{V}ar(logit(\hat{\kappa}(c)))}]}{1 + \exp[logit(\hat{\kappa}(c)) \pm z_{1-\alpha/2} \sqrt{\hat{V}ar(logit(\hat{\kappa}(c)))}]}, \quad (11)$$

Simulation experiments were carried out to study the asymptotic coverage of these two CIs. In order to do so, 10000 binomial samples were generated, both of case samples and control samples, with different sample sizes and from the different values of sensitivity and specificity. As prevalence, different values were taken ($p = 0.10, 0.25, 0.50$) and as the weighting index

the values $c = 0.1, 0.5, 0.9$ were taken. . In Table 2, the results are shown for $Se = 0.90$, $Sp = 0.80$ and $p = 0.10$, and in Table 3 the results are shown for $Se = 0.70$, $Sp = 0.90$ and $p = 0.25$. The simulation experiments showed that the CI logit has a better average coverage and average width than the Wald CI. The coverage of the logit interval fluctuates around the coverage of 95%, whereas that of the Wald interval is usually lower than 95%. A programme in R has been written, called “ewkcccs” (Estimation of the Weighted Kappa Coefficient subject to a Case Control Study), in order to solve this problem of estimation. The programme is available at the following website: “<http://www.ugr.es/~bioest/software.htm#Potros>”. The programme runs with the command “ewkcccs($s_1, s_0, r_1, r_0, cindex, p$)” when the confidence intervals are calculated to 95% of confidence, and where s_i and r_i are the frequencies observed, $cindex$ is the value of the weighting index ($0 \leq cindex \leq 1$) and p is the disease prevalence; and the programme runs with the command “ewkcccs($s_1, s_0, r_1, r_0, cindex, p, conflevel$)” when the intervals are calculated to $100conflevel\%$.

n_1	n_2	c	Coverage Wald CI	Length Wald CI	Coverage Logit CI	Length Logit CI
50	50	0.1	0.937	0.301	0.956	0.293
50	50	0.5	0.931	0.339	0.955	0.327
50	50	0.9	0.934	0.274	0.953	0.269
50	100	0.1	0.938	0.211	0.950	0.209
50	100	0.5	0.944	0.245	0.952	0.241
50	100	0.9	0.940	0.226	0.952	0.223
100	50	0.1	0.940	0.299	0.956	0.291
100	50	0.5	0.940	0.333	0.956	0.322
100	50	0.9	0.941	0.248	0.954	0.244
100	100	0.1	0.942	0.209	0.950	0.207
100	100	0.5	0.945	0.239	0.952	0.234
100	100	0.9	0.948	0.194	0.954	0.193

Table 2: Results of the simulation experiments (I).

4 Example

The results were applied to the study made by Li et al [6] on the diagnosis of Alzheimers disease using as a diagnostic test the genotype ApoE.e4. In order to do so, the authors applied the diagnostic test to a sample of 418 individuals with Alzheimers disease (the test was positive for 240 of them), and they also applied the diagnostic test to a sample of 375 individuals who did not have Alzheimers disease (the test was negative for 288 of them). Assuming that the prevalence of Alzheimers disease is 50% [3], in Table 4 we can see the estimations of the weighted kappa coefficient and the CIs for different values of the weighting index. When the weighting index is higher than 0.5, the beyond chance agreement between the diagnostic test and the disease takes a mediocre value (at 95% confidence). When the weighting index is lower than 0.5, beyond chance agreement between the diagnostic test and the disease takes a mediocre to moderate value (at 95% confidence) depending on the value assigned to the c index.

n_1	n_2	c	Coverage Wald CI	Length Wald CI	Coverage Logit CI	Length Logit CI
50	50	0.1	0.922	0.427	0.957	0.407
50	50	0.5	0.942	0.327	0.959	0.317
50	50	0.9	0.935	0.301	0.957	0.293
50	100	0.1	0.930	0.320	0.951	0.310
50	100	0.5	0.937	0.273	0.954	0.267
50	100	0.9	0.939	0.291	0.953	0.283
100	50	0.1	0.920	0.418	0.950	0.399
100	50	0.5	0.935	0.294	0.951	0.286
100	50	0.9	0.947	0.227	0.957	0.223
100	100	0.1	0.932	0.308	0.953	0.299
100	100	0.5	0.939	0.234	0.950	0.230
100	100	0.9	0.940	0.213	0.949	0.210

Table 3: Results of the simulation experiments (II).

c	$\hat{\kappa}(c)$	95% Wald CI	95% Logit CI
0.1	0.41	0.33-0.48	0.33-0.48
0.2	0.39	0.31-0.46	0.32-0.46
0.3	0.37	0.30-0.44	0.31-0.44
0.4	0.36	0.29-0.42	0.29-0.42
0.5	0.34	0.28-0.41	0.28-0.41
0.6	0.33	0.27-0.39	0.27-0.39
0.7	0.32	0.26-0.38	0.26-0.38
0.8	0.31	0.25-0.37	0.25-0.37
0.9	0.30	0.24-0.35	0.24-0.36

Table 4: Results from the study of Li et al.

5 Conclusions

The estimation of the parameters of a diagnostic test that depend on the disease prevalence is conditioned by the type of sampling. When case-control sampling is used, it is necessary to know the value of the disease prevalence, since this cannot be estimated from the study itself. In this article, we have studied two approximate confidence intervals for the weighted kappa coefficient of a diagnostic test subject to this type of sampling, and it is obtained that the logit interval performs better than the Wald type interval in terms of coverage and width.

Acknowledgement

This research was supported by the Spanish Ministry of Science, Grant Number MTM2012-35591. The author would like to thank the two referees for their comments which have helped to improve the quality of the paper.

Bibliography

- [1] Kraemer, H.C. (1992) **Evaluating medical tests**. SAGE Publications, Newbury Park.
- [2] Roldan Nofuentes, J.A., Luna del Castillo, J.D. and Montero Alonso, M.A. (2009) *Confidence intervals of weighted kappa coefficient of a binary diagnostic test*. Communications in Statistics - Simulation and Computation, **38**, 1562–1578.
- [3] Mercaldo, N.D., Lau, K.F. and Zhou, X.H. (2007) *Confidence intervals for predictive values with an emphasis to casecontrol studies*. Statistics in Medicine, **26**, 2170–2183.
- [4] Kraemer, H.C. and Bloch, D.A. (1990) *A note on case-control sampling to estimate kappa coefficients*. Biometrics, **46**, 45–49.
- [5] Bubin, D.B. and Schenker, N. (1987) *Logit-based interval estimation for binomial data using the Jeffreys prior*. Sociological Methodology, **17**, 131–144.
- [6] Li, Y. et al. (2009) *Association of late-onset Alzheimers disease with genetic variation in multiple members of the GAPD gene family*. Proceedings of the National Academy of Sciences, **101**, 15688–15693.

The jackknife estimate of variance for transition probabilities in the non-Markov illness-death model

Leyla Azarang, *University of Vigo*, leyla.azarang@uvigo.es
Jacobo de Uña-Álvarez, *University of Vigo*, jacobode@uvigo.es

Abstract. Multi-state models are often used to represent the individuals' progress along a certain disease. The estimation of transition probabilities is an important goal in such a setting. The progressive illness-death model is an important multi-state model which has many applications in medical research. Non-parametric estimators of transition probabilities for the non-Markov illness-death model were recently introduced as an alternative to the Aalen-Johansen estimator, which may be inconsistent when the Markov assumption is violated. In this work, the problem of estimating the variance of these transition probabilities is discussed. The jackknife approach is considered to this end. A consistency result is established, and the finite-sample performance of the jackknife estimator is investigated through simulations. A real medical dataset is included for illustration purposes.

Keywords. Censored data, Illness-death model, Jackknife estimator, Kaplan-Meier.

1 Introduction

Multi-state models are models for stochastic processes which represent the states possibly visited by an individual along time, and the allowed transitions among them. They often involve assumptions on the joint distribution of the successive transition times, the influence of covariates on transition intensities, and so on. Multi-state models have become a key tool for data analysis and inferences in medical research; existing reviews include Commenges (1999), Hougaard (1999), Andersen and Keiding (2002), or Meira-Machado et al. (2009). One important target in applications is the estimation of transition probabilities. Nonparametric estimation of transition probabilities in a general multi-state model goes back to Aalen and Johansen (1978). The Aalen-Johansen estimator is consistent for Markov models; however, in practice, Markov assumption may be violated, and the Aalen-Johansen estimator may show a systematic bias ([3], [8]).

Meira-Machado et al. (2006) introduced an alternative estimator of the transition matrix

for the progressive illness-death model, which does not require the Markov condition. The progressive illness-death model (or disability model, cfr. Hougaard, 2000) is a very specific multi-state model, but with many practical applications. It involves three states: 'Healthy' (state 1), 'Diseased' (state 2), and 'Dead' (state 3), and three possible transitions among them: $1 \rightarrow 2$, $2 \rightarrow 3$, and $1 \rightarrow 3$. In this model, states 1 and 2 are transient, while state 3 is absorbing; note also that 'recovery' (i.e. transition $2 \rightarrow 1$) is not allowed. Let Z denote the sojourn time in state 1, and let T denote the absorption time (time to reach state 3 from state 1); thus, the relevant transition probabilities are, with $s < t$,

$$p_{11}(s, t) = P(Z > t | Z > s),$$

$$p_{12}(s, t) = P(Z \leq t < T | Z > s),$$

and

$$p_{22}(s, t) = P(T > t | Z \leq s < T).$$

If (Z, T) are observable, obvious non-Markov estimators for these curves are given by sampling proportions. The presence of censored trajectories demands however for a more sophisticated structure; Meira-Machado et al. (2006)'s estimators involve the computation of two Kaplan-Meier curves: the one pertaining to Z , and that corresponding to T . These estimators are consistent (regardless the Markov condition) provided that the potential censoring time C is independent of the process (i.e. of the pair (Z, T)), and that the support of C contains that of T . See [3] for related estimators and comparative results.

To be specific, and to introduce the main ideas and novelties of this work as soon as possible, we focus on the transition probability $p_{22}(s, t)$. Let $(\tilde{Z}_i, \tilde{T}_i, \delta_i, \Delta_i)$, $1 \leq i \leq n$, be a random sample of $(\tilde{Z}, \tilde{T}, \delta, \Delta)$, where (\tilde{Z}, \tilde{T}) are the (possibly) censored versions of (Z, T) , and (δ, Δ) are the corresponding censoring indicators. Let $\hat{S}(t)$ be the Kaplan-Meier estimator of $S(t) = P(T > t)$ (computed from the (\tilde{T}_i, Δ_i) 's) and let W_{ni} be the jump of $\hat{S}(t)$ at $t = \tilde{T}_i$, that is, $W_{ni} = \frac{\Delta_i}{(n-i+1)} \prod_{j=1}^{i-1} [\frac{n-j}{n-j+1}]^{\Delta_j}$, where we assume that the sample is ordered with respect to the variable \tilde{T} (in the uncensored case these weights simply reduce to $W_{ni} = 1/n$, $1 \leq i \leq n$) Meira-Machado et al. (2006) introduced as a suitable estimator for $p_{22}(s, t)$ the empirical

$$\hat{p}_{22}(s, t) = \frac{\sum_{i=1}^n W_{ni} I(\tilde{Z}_i \leq s, t < \tilde{T}_i)}{\sum_{i=1}^n W_{ni} I(\tilde{Z}_i \leq s < \tilde{T}_i)} = \frac{\sum_{i=1}^n W_{ni} \varphi_{s,t}(\tilde{Z}_i, \tilde{T}_i)}{\sum_{i=1}^n W_{ni} \varphi_{s,s}(\tilde{Z}_i, \tilde{T}_i)},$$

where $\varphi_{s,t}(u, v) = I(u \leq s, t < v)$. This estimator is a quotient of two multivariate Kaplan-Meier integrals, in the sense of Stute (1993). As such, available asymptotics for multivariate Kaplan-Meier integrals ([11], [12]) apply. This results in the consistency and the asymptotic normality of $\hat{p}_{22}(s, t)$ under a number of conditions; in particular, by using the delta method, the asymptotic variance of $\hat{p}_{22}(s, t)$ is given by

$$AVar(\hat{p}_{22}(s, t)) = \frac{\sigma^2}{n}$$

where

$$\sigma^2 \equiv \sigma^2(s, t) = \sigma_{11} \frac{1}{S(\varphi_{s,s})^2} + \sigma_{22} \frac{S(\varphi_{s,t})^2}{S(\varphi_{s,s})^4} - 2\sigma_{12} \frac{S(\varphi_{s,t})}{S(\varphi_{s,s})^3}, \quad (1)$$

$S(\varphi) = E[\varphi(Z, T)]$, and $\sigma_{ij} \equiv \sigma_{ij}(s, t)$ stands for the limit covariance between two Kaplan-Meier integrals $S_n(\varphi_i)$ and $S_n(\varphi_j)$ of the general form $S_n(\varphi) = \sum_{i=1}^n W_{ni}\varphi(\tilde{Z}_i, \tilde{T}_i)$; here we put $\varphi_1 = \varphi_{s,t}$ and $\varphi_2 = \varphi_{s,s}$.

In practice, estimation of the limit variance σ^2 is required for (e.g.) the computation of confidence limits for $p_{22}(s, t)$. In (1), the quantities $S(\varphi_{s,t})$ and $S(\varphi_{s,s})$ may be replaced by the corresponding Kaplan-Meier integrals, $S_n(\varphi_{s,t})$ and $S_n(\varphi_{s,s})$ respectively. Regarding the estimation of the σ_{ij} , Azarang et al. (2013) established the consistency of the jackknife approach (cfr. Shao and Tu, 1995) in the general setting of censored data with multiple covariates, thus extending previous results in Stute (1996b) for the univariate setting; here, by considering Z as a 'covariate' of the absorption time T , their result applies. More specifically, let $S_n^{(k)}(\varphi)$, $1 \leq k \leq n$, be the pseudovalues of $S_n(\varphi)$; $S_n^{(k)}(\varphi)$ is computed like $S_n(\varphi)$ but deleting the k -th datum $(\tilde{Z}_k, \tilde{T}_k, \delta_k, \Delta_k)$ from the initial sample. The jackknife estimate of covariance between $S_n(\varphi_i)$ and $S_n(\varphi_j)$, $1 \leq i, j \leq 2$, is defined as

$$n\widehat{Cov}_{ij} = (n-1) \sum_{k=1}^n (S_n^{(k)}(\varphi_i) - S_n^{(\bullet)}(\varphi_i))(S_n^{(k)}(\varphi_j) - S_n^{(\bullet)}(\varphi_j))$$

where $S_n^{(\bullet)}(\varphi)$ denotes the average of the $S_n^{(k)}(\varphi)$'s. In the definition of $n\widehat{Cov}_{ij}$, when the largest datum $\tilde{T}_{(n)}$ is uncensored but the second largest $\tilde{T}_{(n-1)}$ is censored, we artificially set $\tilde{T}_{(n)}$ to be censored; see [4] for discussion. Introduce the estimator

$$\hat{\sigma}^2 = n\widehat{Cov}_{11} \frac{1}{S_n(\varphi_{s,s})^2} + n\widehat{Cov}_{22} \frac{S_n(\varphi_{s,t})^2}{S_n(\varphi_{s,s})^4} - 2n\widehat{Cov}_{12} \frac{S_n(\varphi_{s,t})}{S_n(\varphi_{s,s})^3}.$$

Put $\gamma_0(t) = \exp\left\{\int_0^{t-} (1-H)^{-1} dH^0\right\}$ where $H(t) = P(\tilde{T} \leq t)$ and $H^0(t) = P(\tilde{T} \leq t, \Delta = 0)$. We have the following result.

Theorem 1.1. *Under condition*

$$E[-\log(1 - \sqrt{H(\tilde{T})})\gamma_0(\tilde{T})^2\Delta] < \infty$$

we have with probability one $\hat{\sigma}^2 \rightarrow \sigma^2$ as $n \rightarrow \infty$.

Proof. The result is a consequence of the SLLN for multivariate Kaplan-Meier integrals in Stute (1993) and the Theorem in Azarang et al. (2013), up to noting that both $\varphi_{s,t}$ and $\varphi_{s,s}$ are bounded functions. \square

Remark. The condition in Theorem 1.1 above ensures that the censoring effects do not dominate at the right tail of the distribution of T . In the particular case in which both T and C are exponentially distributed, the condition holds provided that the expected proportion of censored data is below 0.5.

Theorem 1.1 above suggests to approximate the variance of $\hat{p}_{22}(s, t)$ by $\hat{\sigma}^2/n$ when n is large. The finite-sample accuracy of this approximation is explored through simulations in Section 2,

while in Section 3 we illustrate the jackknife method with real medical data. Since the bootstrap is a popular method to approximate the variance of a given statistic, we include it in our study for comparison purposes. Note however that, for the best of our knowledge, the consistency of the bootstrap variance has not been formally established for the setting considered in this paper. The jackknife method may be used to introduce estimators for the other relevant transition probabilities in Meira-Machado et al. (2006) too. Since $p_{11}(s, t)$ only involves the marginal distribution of Z , the consistency of the jackknife approach for this transition probability will immediately follow from existing results for (univariate) Kaplan-Meier integrals (Stute, 1996b). The theory for $p_{12}(s, t)$ is not so easily obtained, since the estimator pertaining to this transition probability depends on two different Kaplan-Meier curves (the ones corresponding to Z and T); new technical results are required in this case.

2 Simulation study

In this section we investigate the performance of the jackknife estimate of variance for $\hat{p}_{22}(s, t)$ and we compare the jackknife and bootstrap methods through simulations.

To simulate the data in the illness-death model, the procedure is as follows:

Step 1 $V_1 \sim U(0, 1)$, $V_2 \sim U(0, 1)$ are independently generated

Step 2 $U_1 = V_1$, $U_2 = C^{-1}(V_2|U_1)$; where $C^{-1}(y|x) = -\frac{1}{\theta} \log[1 + \frac{y(e^{-\theta} - 1)}{y + (1 - y)e^{-\theta x}}]$

Step 3 $Z = -\log(U_1)$, $T_{23} = -\log(U_2)$

Step 4 $\rho \sim Ber(p)$ is generated independently of Z

Step 5 $T = Z + \rho T_{23}$

This corresponds to Frank's copula model for the dependence between the (exponentially distributed) sojourn times in state 1 and 2, Z and T_{23} respectively, for those individuals visiting the latter state ($\rho = 1$). We take $\theta = 12$ which implies a positive association between Z and T_{23} (Kendall's Tau is 0.71). Also, an independent exponential censoring time C is generated, according to $Exp(0.59)$, $Exp(0.20)$, $Exp(0.10)$ models, which correspond to 50%, 26%, and 15% of censoring respectively. The simulated models are non-Markov due to the dependence between Z and T_{23} . In each simulation $M = 1000$ samples are generated, and sample sizes 50, 150, and 250 are considered. The proportion of individuals going through state 2 is $p = 0.7$. The true variance of $\hat{p}_{22}(s, t)$, denoted by σ_{MC}^2 , is approximated using Monte Carlo simulation.

In Tables 1-3 we report the mean values of the jackknife variance estimator ($\hat{\sigma}_J^2 = \hat{\sigma}^2/n$) and the bootstrap variance estimator ($\hat{\sigma}_B^2$) along the $M = 1000$ simulations, for the cases $(s, t) = (0.2231, 1.6094)$, $(s, t) = (0.5108, 0.9163)$, and $(s, t) = (0.9163, 1.6094)$, which correspond to the 0.2, 0.4, 0.6 and 0.8 quantiles of the $Exp(1)$ model. The bootstrap variance estimator is defined as the variance of the bootstrap values of $\hat{p}_{22}(s, t)$ along $B = 999$ bootstrap resamples; the simple bootstrap which resamples each datum (with replacement) with probability $1/n$ is used to this end. In Tables 1-3 we also give n times the bias ($n.Bias$), n times the standard deviation ($n.SD$), and n^2 times the mean square error ($n^2.MSE$) of $\hat{\sigma}_J^2$ and $\hat{\sigma}_B^2$. These three

<i>CP:</i>	<i>n</i> =50			<i>n</i> =150			<i>n</i> =250		
	50%	26%	15%	50%	26%	15%	50%	26%	15%
σ_{MC}^2	0.00010	0.00118	0.00027	0.00015	0.00010	0.00008	0.00010	0.00005	0.00004
$\hat{\sigma}_B^2$	0.00006	0.00012	0.00021	0.00010	0.00009	0.00008	0.00007	0.00005	0.00004
$\hat{\sigma}_J^2$	0.00255	0.00184	0.00028	0.00010	0.00008	0.00007	0.00008	0.00005	0.00004
<i>n.BiasB</i>	-0.00219	-0.05297	-0.00316	-0.00703	-0.00179	-0.00078	-0.00695	-0.00059	-0.00000
<i>n.BiasJ</i>	0.12249	0.03298	0.00038	-0.00665	-0.00276	-0.00142	-0.00633	-0.00105	-0.00066
<i>n.SDB</i>	0.08515	0.13864	0.20454	0.28335	0.21932	0.14351	0.34126	0.13420	0.11000
<i>n.SDJ</i>	3.92920	2.40963	0.20442	0.28988	0.20198	0.13556	0.34612	0.12847	0.10266
$n^2.MSEB$	0.00726	0.02203	0.04185	0.08033	0.04810	0.02059	0.11650	0.01801	0.01210
$n^2.MSEJ$	15.45364	5.80739	0.04179	0.08407	0.04080	0.01838	0.11984	0.01651	0.01054

Table 1: Results of the simulation study for the case $(s, t) = (0.2231, 1.6094)$.

<i>CP:</i>	<i>n</i> =50			<i>n</i> =150			<i>n</i> =250		
	50%	26%	15%	50%	26%	15%	50%	26%	15%
σ_{MC}^2	0.07466	0.04755	0.04353	0.02078	0.01351	0.01179	0.00739	0.00696	0.00655
σ_B^2	0.04070	0.03755	0.03636	0.01999	0.01372	0.01237	0.00794	0.00710	0.00737
σ_J^2	0.24713	0.10072	0.03062	0.01849	0.01292	0.01171	0.00772	0.00690	0.00715
<i>n.BiasB</i>	-1.69802	-0.50028	-0.35847	-0.11912	0.03185	0.08692	0.13788	0.03702	0.20532
<i>n.BiasJ</i>	8.62323	2.65831	-0.64543	-0.34432	-0.08818	-0.01229	0.08315	-0.01452	0.14974
<i>n.SDB</i>	2.13836	1.64609	1.4795	1.65376	0.86834	0.72824	0.56648	0.47532	0.49615
<i>n.SDJ</i>	41.5393	14.34684	1.21307	1.40787	0.75497	0.64130	0.53077	0.44442	0.45989
$n^2.MSEB$	7.45586	2.95990	2.31742	2.74911	0.75504	0.53788	0.33991	0.22730	0.28832
$n^2.MSEJ$	1799.873	212.8985	1.88812	2.10065	0.57775	0.41142	0.28863	0.19772	0.23392

Table 2: Results of the simulation study for the case $(s, t) = (0.5108, 0.9163)$

<i>CP:</i>	<i>n</i> =50			<i>n</i> =150			<i>n</i> =250		
	50%	26%	15%	50%	26%	15%	50%	26%	15%
σ_{MC}^2	0.07065	0.03485	0.03179	0.02440	0.01131	0.00975	0.01426	0.00723	0.00552
σ_B^2	0.03890	0.03266	0.03000	0.02171	0.01134	0.00958	0.01334	0.00674	0.00567
σ_J^2	0.24833	0.08349	0.02613	0.02055	0.01086	0.00923	0.01300	0.00658	0.00555
<i>n.BiasB</i>	-1.58723	-0.10980	-0.08947	-0.40312	0.00456	-0.02520	-0.23041	-0.12358	0.03545
<i>n.BiasJ</i>	8.88433	2.43169	-0.28254	-0.57685	-0.06648	-0.07827	-0.31419	-0.16370	0.00647
<i>n.SDB</i>	2.30719	1.35810	1.12683	2.22722	0.66445	0.52867	1.48568	0.45833	0.36264
<i>n.SDJ</i>	44.78929	16.01074	0.92105	1.99408	0.59817	0.47153	1.46354	0.43768	0.34248
$n^2.MSEB$	7.84240	1.85649	1.27775	5.12303	0.44152	0.28013	2.26034	0.22534	0.13276
$n^2.MSEJ$	2085.012	262.2567	0.92815	4.30911	0.36223	0.22847	2.24066	0.21836	0.11733

Table 3: Results of the simulation study for the case $(s, t) = (0.9163, 1.6094)$.

quantities should converge to zero for an increasing sample size, provided that the estimators are consistent (see Theorem 1.1 above for the jackknife).

We see in Tables 1-3 that the bias, the SD, and the MSE of the jackknife estimator decreases as the sample size increases, revealing its consistency. The only exception is the case with 50% of censoring in Table 2, where both the SD and the MSE increases when moving from $n = 150$ to $n = 250$. The same holds true for the bootstrap estimator. Larger sample sizes could be needed

to replicate the convergence result in Theorem 1.1. Bias is of a smaller order of magnitude compared to SD. On the other hand, the error in estimation increases with the censoring degree, as expected. The estimator based on the jackknife performs better than that based on the bootstrap in most of the cases; however, when the sampling information is very scarce ($n = 50$, moderate to large censoring degree), the bootstrap may report more accurate results. Regarding the influence of the particular (s, t) values, the results suggest that large values of variance are less accurately estimated (something expected). Other scenarios have been simulated and the results and full discussion will be reported elsewhere.

3 Colon cancer data

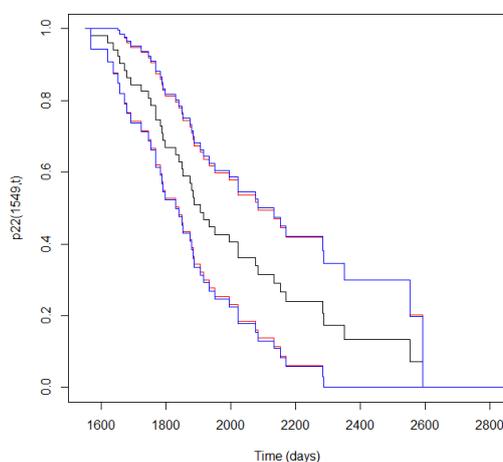


Figure 1: Estimated transition probabilities of $p_{22}(s, t)$ for $s = 1549$ (black line) with 95% jackknife confidence bands (red lines) and bootstrap confidence bands (blue lines). Colon cancer data.

For illustration, we apply the jackknife method to data from a large clinical trial on Duke's stage III patients, affected by colon cancer (Moertel et al., 1990). This data set is freely available as a part of the R *survival* package. These data come from one of the first successful trials of adjuvant chemotherapy for colon cancer. In this study, from the total of 929 patients that underwent a curative surgery for colorectal cancer, 423 patients remained alive at the end of the follow-up; 468 patients developed recurrence and among them 414 died; and 38 patients died without recurrence. Here, recurrence is considered as state 2. Using the progressive illness-death model, there are three states: 'Alive and disease-free', 'Alive with recurrence', and 'Dead'. Figures 1, 2, and 3 depict the estimator $\hat{p}_{22}(s, t)$ proposed by Meira-Machado et al. (2006) for $s = 1549$ days along t , with 95% pointwise confidence limits obtained by the jackknife and bootstrap methods, for the whole colon cancer data, levamisole treatment group, and levamisole plus 5-FU treatment group respectively. From the figures we conclude that both the jackknife and bootstrap methods report similar results, and that the jackknife confidence intervals are often slightly narrower than those of the bootstrap (particularly true for levamisole and levamisole plus 5-FU treatment groups).

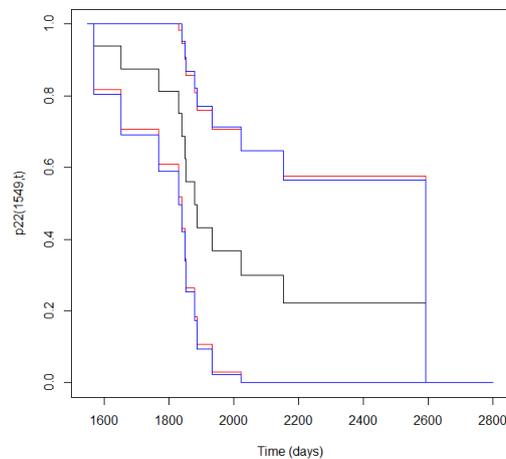


Figure 2: Estimated transition probabilities of $p_{22}(s, t)$ for $s = 1549$ (black line) with 95% jackknife confidence bands (red lines) and bootstrap confidence bands (blue lines). Levamisole treatment group, colon cancer data.

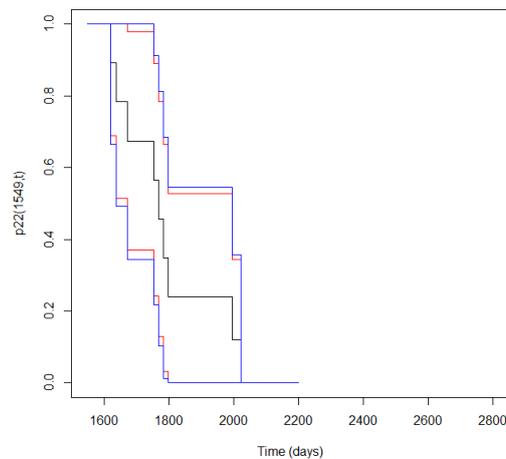


Figure 3: Estimated transition probabilities of $p_{22}(s, t)$ for $s = 1549$ (black line) with 95% jackknife confidence bands (red lines) and bootstrap confidence bands (blue lines). Levamisole plus 5-FU treatment group, colon cancer data.

Acknowledgement

This work was supported by funding from the European Community's Seventh Framework Programme FP7/2011: Marie Curie Initial Training Network MEDIASRES ("Novel Statistical Methodology for Diagnostic/Prognostic and Therapeutic Studies and Systematic Reviews"; www.mediasres-itn.eu) with the Grant Agreement Number 290025. Second author was sup-

ported by the Grant MTM2011-23204 of the Spanish Ministry of Science and Innovation (FEDER support included).

Bibliography

- [1] Aalen, O. and Johansen, S.(1978) *An empirical transition matrix for nonhomogeneous Markov chains based on censored observations*. Scandinavian Journal of Statistics, **5**, 141–150.
- [2] Andersen, PK. and Keiding, N.(2002) *Multi-state models for event history analysis*. Statistical Methods in Medical Research, **11**, 91–115.
- [3] Allignol, A. , Beyersmann J. , Gerds, T. and Latouche, A.(2013) *A competing risks approach for nonparametric estimation of transition probabilities in a non-Markov illness-death model*. Lifetime Data Analysis, DOI: 10.1007/s10985-013-9269-1.
- [4] Azarang, L. , de Uña-Álvarez, J. and Stute, W. (2013) *The jackknife estimate of covariance under censorship when covariables are present*. Report 13/04, Discussion Papers in Statistics and OR, University of Vigo. Available online at: http://webs.uvigo.es/depc05/reports/13_04.pdf.
- [5] Commenges, D. (1999) *Multi-state models in epidemiology*. Lifetime Data Analysis, **5**, 315–327.
- [6] Hougaard, P. (1999) *Multi-state models: a review*. Lifetime Data Analysis, **5**,239–264.
- [7] Hougaard, P. (2000) *Analysis of Multivariate Survival Data*. Springer, New York.
- [8] Meira-Machado, L. , de Uña-Álvarez, J. and Cadarso-Suárez, C.(2006) *Nonparametric estimation of transition probabilities in a non-Markov illness-death model*. Lifetime Data Analysis, **12**, 325–344.
- [9] Moertel C.G., Fleming T.R., Macdonald J.S. et al. (1990) *Levamisole and fluorouracil for adjuvant therapy of resected colon carcinoma*. New England Journal of Medicine **322**, 352–358.
- [10] Shao, J. and Tu, D. (1995) *The Jackknife and Bootstrap*. Springer-Verlag, New York.
- [11] Stute, W. (1993) *Consistent estimation under random censorship when covariables are present*. Journal of Multivariate Analysis, **45**, 89–103.
- [12] Stute, W. (1996a) *Distributional convergence under random censorship when covariables are present*. Scandinavian Journal of Statistics, **23**,461–471.
- [13] Stute, W. (1996b) *The jackknife estimate of variance of a Kaplan-Meier integral*. Annals of Statistics, **24**, 2679–2704.

Linear discriminant analysis based on penalized functional PLS

Ana M. Aguilera, *University of Granada*, aaguiler@ugr.es

M. Carmen Aguilera-Morillo, *University Carlos III of Madrid*, maguiler@est-econ.uc3m.es

Abstract. The aim is to classify a set of functional data according to a categorical variable with more than two categories. To this end, functional linear discriminant analysis (LDA) is considered to classify the curves. Two ways to achieve functional linear discriminant analysis based on different penalized estimation of the PLS components are proposed. Both are based on a two-step algorithm: first the data set is projected into a reduced number of functional PLS components, and after that LDA is carried out on the original response variable. In order to show the good performance of these penalized functional classification approaches, they have been compared with the non-penalized version in an application to classify spectral data.

Keywords. Functional data analysis, Linear discriminant analysis, Partial least squares regression, P-splines, NIR spectra.

1 Overview

The aim of this work is to classify a set of functional data according to a categorical variable with more than two categories. In fact, we are interested in functional data which are affected by some noise or contamination. Therefore, in order to get a good classification of the samples curves and an accurate interpretability, reduction dimension techniques and regularization must be considered. In that sense, LDA is a consolidate technique for classification widely used in chemometric studies. A solution to the high dimension problem is to decompose and project the sample curves onto a small number of orthogonal components given by principal component analysis (PCA) or PLS regression. PCA was applied to classify NIR spectral data of vegetable oils in [16]. In [12], PLS analysis was applied as a discriminant as well as a quantitative tool in the analysis of edible fats and oils by Fourier transform near-infrared (FT-NIR) spectroscopy. Once the dimension is reduced, multivariate classification techniques such as LDA, quadratic discriminant analysis (QDA) and non-linear regression $\{0, 1\}$ were applied in [10].

But sometimes functional data are not smooth and therefore some penalty or regularization is needed. In fact, a new method called regularized discriminant analysis (RDA) which is a penalized alternative to the classical maximum likelihood estimates for the covariance matrices was

developed in [8]. Another regularized classification method consisting of discriminant analysis with shrunken covariances (DASCO) was proposed in [7], providing superior performance than the old favorites. An alternative to these penalized methods is the penalized LDA proposed by Hastie and Tibshirani [9]. A general overview of regularized techniques in discriminant analysis, for continuous and discrete response variables, can be seen in [13]. In that context, our contributions are two different functional versions for penalized discriminant analysis based on penalized functional PLS regression. The first one introduces a P-spline penalty in the initial estimation of the sample curves. After that, functional LDA on the smoothed sample curves is carried out. The second one introduces the penalty directly in the definition of the norm involved in the PLS algorithm.

In order to show the good performance of the proposed penalized methods, they are compared with functional LDA on non-penalized PLS components in an application to classify spectral data.

2 Functional LDA based on functional PLS regression

In this work we focus on LDA when the predictor $X = \{X(t) : t \in T\}$ is functional (continuous and second order stochastic process whose sample paths take values in the Hilbert space of squared integrable functions $L_2([0, T])$) and the response is a categorical variable Y with K categories. The aim of functional LDA is to find linear combinations $\Phi(X) = \int_0^T X(t)\beta(t)dt$ so that the between class variance is maximized with respect to the total variance

$$\max_{\beta} \frac{V(E[\Phi(X)|Y])}{V(\Phi(X))}.$$

Due to the infinite dimension of the functional predictor, the estimation of $\beta(t)$ by LDA is an ill-posed problem. In order to reduce the dimension of the data and to estimate the discriminant coefficient functions, functional LDA based on PLS regression on functional data was proposed in [14] taking into account the equivalence between LDA and canonical correlation analysis. In [6] PLS regression was used in functional data classification problems.

By considering the basis representation of the functional data, a B-spline approach for functional PLS regression was proposed in [1]. These functional approaches were applied to estimate the quality of cookies from the resistance of dough during the kneading process. The case of functional LDA for irregularly sampled curves was studied in [11]. Following the ideas developed in these works, in order to improve the estimation and the classification ability of functional LDA different penalized versions of functional LDA are proposed in this paper.

Denoting by $\{Y_i \in (0, 1) : i = 1, \dots, K-1\}$ the dummy variables associated to the categorical response Y , the functional LDA-PLS model consists of performing the classical LDA of Y on a reduced set of PLS components obtained from the PLS regression of the vector (Y_1, \dots, Y_{K-1}) on the functional predictor X .

Penalized functional PLS

Non penalized PLS components $t = \int_T X(t)w(t)dt$ are estimated by solving the following maximization problem

$$\max_{w,c} Cov^2 \left(\int_T X(t)w(t)dt, \sum_{i=1}^{K-1} c_i Y_i \right)$$

restricted to $\|w\| = \|c\| = 1$, with $\|\cdot\|$ representing the usual norms in the spaces $L_2[0, T]$ and \mathbb{R}^{K-1} where the component weights belong to, respectively. By considering a basis representation of the functional predictor given by

$$X(t) = \sum_{j=1}^p \alpha_j \phi_j(t),$$

it can be concluded that FPLS is equivalent to an ordinary PLS on the vector of variables $(\Psi^{1/2})'\alpha$, where $\Psi^{1/2}$ is the squared root of the matrix of inner products between basis functions and α is the vector of basis coefficients of the functional predictor X [1].

When we are working with noisy functional data some penalization is required to get a smooth estimation of the partial PLS weight functions. In that sense, two different penalized functional PLS regressions are considered.

In order to smooth the estimation of the functional linear model, two different PCR and PLSR approaches for functional data were proposed in [15]. The main difference between our penalized PLS versions and the mentioned above, is that these penalized estimation approaches did not consider the functional form of the sample paths and they are based on multivariate linear regression of the response in terms of the matrix of discrete-time observations of the sample curves.

The first version consists of introducing a P-spline penalty [5] in the initial smoothing of the sample curves by considering their basis representation (see [2] for more details). This type of penalty was used in [3] for functional LDA and functional logit regression when the response variable has only two categories $\{0, 1\}$. The other version introduces a P-spline penalty in the definition of the norm in the functional space given by $\|w\|^2 + \lambda PEN_d(w)$, with λ being the smoothing parameter and $PEN_d(w)$ a d-order discrete penalty. This penalized version of PLS is equivalent to an ordinary PLS on the vector of variables $L^{-1}\Psi\alpha$, where $L = (\Psi + \lambda P_d)^{1/2}$ with $P_d = (\Delta^d)^T \Delta^d$ and Δ^d being the matrix of d-order differences between adjacent basis coefficients.

As stated before, once the functional PLS regression is computed, classical LDA is carried out on a reduced set of PLS components. For penalized methods, both the smoothing parameter and the optimal number of PLS components are jointly estimated by a 10-fold-cross-validation algorithm.

3 Results

The aim is to classify mayonnaise sauce spectra according to a categorical response variable that represents the type of oil from which the samples of mayonnaise were made. Exactly we have 162 NIR spectra observed in 351 equally spaced wavelengths in the 1100-2500nm area based on six types of vegetable oils (soybean oil, sunflower oil, canola oil, olive oil, corn oil and grapeseed oil). The sample paths have been displayed in Figure 1.

In Table 1 the miss-classification rates (MCR) for the compared methods are shown. The good performance of the penalized versions is proved, being LDA on functional PLS regression penalizing the norm which provides the lowest miss-classification rate. From the results on the mayonnaise spectra we can conclude that the penalized functional classification approaches considered in this paper significantly improves the classification with respect to the non-penalized approach.

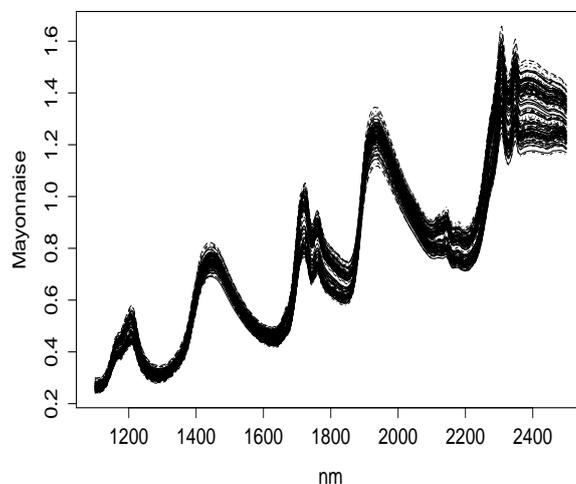


Figure 1: 162 NIR spectra observed in 351 equally spaced wavelengths in the 1100-2500nm.

	LDA-FPLS	LDA-Pspl FPLS	LDA-Penalized-Norm FPLS
MCR	14%	10%	5%

Table 1: Miss-classification rates (MCR) for LDA on a set of PLS components obtained by non-penalized functional PLS (LDA-FPLS), functional PLS on P-splines (LDA-Pspl FPLS) and functional PLS penalizing the norm (LDA-Penalized-Norm FPLS).

Acknowledgement

This research was supported by Project FQM-08068 from Consejería de Innovación, Ciencia y Empresa de la Junta de Andalucía Spain.

Bibliography

- [1] Aguilera, A.M., Escabias, M., Preda, C., and Saporta, G. (2010) *Using basis expansion for estimating functional pls regression. applications with chemometric data*. Chemometrics and Intelligent Laboratory Systems, **104**, 289–305.
- [2] Aguilera, A.M. and Aguilera-Morillo, M.C. (2013) *Comparative study of different B-spline approaches for functional data*. Mathematical and Computer Modelling, **58**, 1568–1579.
- [3] Aguilera-Morillo, M.C., Aguilera, A.M., Escabias, M. and Valderrama, M.J. (2013) *Penalized spline approaches for functional logit regression*. TEST, **22**, 251–277.

- [4] Aguilera-Morillo, M. C. and Aguilera, A. M. (2013) *P-spline estimation of functional classification methods for improving the quality in the food industry*. Communications in Statistics - Simulation and Computation, **in press**.
- [5] Eilers, P.H.C. and Marx, B.D. (1996) *Flexible smoothing with B-splines and penalties*. Statistical Science, **11**, 89–121.
- [6] Delaigle, A. and Hall, P. (2012) *Achieving near perfect classification for functional data*. Journal of the Royal Statistical Society. Series B, **74**, 267–286.
- [7] Frank, I. and Friedman, J.H. (1989) *Classification: Oldtimers and newcomers*. Journal of Chemometrics, **3**, 463–475.
- [8] Friedman, J.H. (1989) *Regularized Discriminant Analysis*. Journal of the American Statistical Association, **84**, 165–175.
- [9] Hastie, T., Buja, A. and Tibshirani, R. (1995) *Penalized Discriminant Analysis*. The Annals of Statistics, **23**, 73–102.
- [10] Indahl, U.G., Sahni, N.S., Kirkhus, B. and Naes, T. (1999) *Multivariate strategies for classification based on NIR-spectra—with application to mayonnaise*. Chemometrics and Intelligent Laboratory Systems, **49**, 19–31.
- [11] James, G.M. and Hastie, T.J. (2001) *Functional discriminant analysis for irregularly sampled curves*. Journal of the Royal Statistical Society. Series B, **63**, 533–550.
- [12] Li, H., van de Voort, F.R., Ismail, A.A., Sedman, J. Cox, R., Simard, C. and Buijs, H. (2000) *Discrimination of Edible Oil Products and Quantitative Determination of Their Iodine Value by Fourier Transform Near-Infrared Spectroscopy*. Journal of the American Oil Chemists’s Society, **77**, 29–36.
- [13] Mkhadri, A., Celeux, G. and Nasroallah, A. (1997) *Regularization in discriminant analysis: an overview*. Computational Statistics and Data Analysis, **23**, 403–423.
- [14] Preda, C., Saporta, G., and Lévêder, C. (2007) *Pls classification for functional data*. Computational Statistics, **22**, 223–235.
- [15] Reiss, P. T. and Ogden, R. T. (2007) *Functional principal component regression and functional partial least squares*. Journal of the American Statistical Association, **102**, 984–996.
- [16] Sato, T. (1994) *Application of principal-component analysis on near-infrared spectroscopic data of vegetable oils for their classification*. Journal of the American Oil Chemists’s Society, **71**, 293–298.

A generalized Description Length approach for Sparse and Robust Index Tracking

Davide Ferrari, *University of Melbourne*, davide.ferrari@unimelb.edu.au

Margherita Giuzio, *EBS Universität für Wirtschaft und Recht*, margherita.giuzio@ebs.edu

Sandra Paterlini, *EBS Universität für Wirtschaft und Recht*, sandra.paterlini@ebs.edu

Abstract. We develop a new minimum description length criterion for index tracking, which deals with two main issues affecting portfolio weights: estimation errors and model misspecification. The criterion minimizes the uncertainty related to data distribution and model parameters by means of a generalized q -entropy measure, and performs model selection and estimation in a single step, by assuming a prior distribution on portfolio weights. The new approach results in sparse and robust portfolios in presence of outliers and high correlation, by penalizing observations and parameters that highly diverge from the assumed data model and prior distribution. The Monte Carlo simulations and the empirical study on financial data confirm the properties and the advantages of the proposed approach compared to state-of-art methods.

Keywords. q -entropy, penalized least squares, sparsity, index tracking

1 Introduction

Since Markowitz [1], an optimal portfolio in asset allocation is determined by first considering the risk/return performance of each asset, in terms of mean and variance, and then selecting the portfolio with the best trade-off. Portfolio weights result then to be very sensitive to changes in parameter estimates, especially in presence of model misspecification and high dimensionality of the problem. Thus, estimation bias may heavily affect the optimization process resulting in suboptimal and unsatisfactory performance ([2], [3], [4]). Typically, asset returns are highly correlated with a leptokurtic distribution, which is largely contaminated by outliers [5]. If these statistical regularities are not properly considered, the misspecification of the data model may result in imprecise parameter estimates. To deal with these issues, several methods have been proposed in the financial literature, i.e. robust estimation methods, minimum divergence models and penalized least squares. We formulate a new criterion for portfolio selection that is able to

deal with both estimation errors and model misspecification, and develop a general algorithm to obtain robust and sparse portfolios, i.e. with a low number of active positions.

In particular, we propose a description length criterion that codes the uncertainty about the data and the model parameters through a q -entropy, a generalized information measure [6] that accounts for the divergence from the assumed data model and the target prior distribution. It enhances the robustness of the portfolio to model misspecifications by assigning a lower weight to observations and parameter estimates that are not consistent with the assumed models. The whole criterion performs model selection and estimation in a single step and depends on the choice of two tuning parameters, q and λ . The former manages the trade-off between accuracy and stability of parameter estimates [7], while the latter controls the penalization of portfolio weights.

Section 2 introduces the description length criterion. Section 3 describes the re-weighting algorithm for portfolio selection and the special cases in which data are assumed to follow a Normal or a t-Student distribution, while the prior distribution on the parameters is a Laplace function. Section 4 presents the simulation study comparing the performance of our method to the main state-of-art benchmark. Section 5 illustrates the behaviour of our portfolio selection method in an index tracking framework with real-world financial data. Section 6 concludes.

2 Description Length Criterion

Let a financial portfolio return be defined as $Y = \beta^T \mathbf{X}$, where \mathbf{X} is a p -dimensional random vector of asset returns with unknown multivariate distribution and β is the vector of asset weights. Given observations $x_i, i = 1, \dots, n$, let μ and σ^2 be the portfolio expected return and variance. Then, the true probability density function of the standardized portfolio return $g(z)$, can be modelled through the function f , which may be for example the standard Normal or the t-Student distribution. Given a mean target value $\mu = \mu^*$, we can then compute portfolio weights $\hat{\beta}_{q,\lambda}$, by minimizing the following description length criterion:

$$\widehat{D}_{q,\lambda}(\beta, \sigma) = - \sum_{i=1}^n L_q \left\{ f \left(\frac{\mathbf{x}_i^T \beta - \mu^*}{\sigma} \right) \right\} - \sum_{j=1}^p L_q \{ \pi(\beta_j; \lambda) \}, \quad (1)$$

for fixed tuning constants $\lambda \geq 0$ and $q \leq 1$. In (1), $L_q(\cdot)$ is the generalized q -logarithm

$$L_q(u) = \begin{cases} (u^{1-q} - 1)/(1 - q), & q \neq 1, \\ \log(u), & q = 1, \end{cases} \quad (2)$$

and $\pi(\beta_j; \lambda)$ is a symmetric distribution for β_j with zero mean and variance depending on λ . In the general framework, no restrictions are placed on the vector of portfolio weights β . We notice that when $q \rightarrow 1$, criterion (1) is equal to maximum a posteriori (MAP) estimation of β , where $\pi(\beta_j; \lambda)$ represent a prior probability density function on β_j . The penalty function $\pi(\beta_j; \lambda)$ controls the model selection and sparsity by shrinking to zero the weights of the assets that do not contribute to obtain a mean target value μ^* . From now on, $\pi(\cdot)$ is assumed to be a Laplace function and then $L_q(\pi)$ results in a non-convex function. In (1), the first term represents the information provided by the data \mathbf{x}_i given a model, while the second term encodes the information about the model itself, given by the prior distributions $\pi(\beta_j; \lambda)$. Minimizing this criterion results in the most efficient description of the data, including the description of

the model itself [8]. Differentiating function (1) with respect to parameters $(\boldsymbol{\beta}, \sigma)^T$, we get the following estimating equations:

$$\mathbf{0} = \nabla \hat{D}_{q,\lambda}(\boldsymbol{\beta}, \sigma) = \sum_{i=1}^n w_q(\mathbf{x}_i, \boldsymbol{\beta}, \sigma) \nabla \log f(\sigma^{-1}(\mathbf{x}_i^T \boldsymbol{\beta} - \mu^*)) + \sum_{j=1}^p v_q(\beta_j, \lambda) \nabla \log \pi(\beta_j; \lambda), \quad (3)$$

where

$$w_q(\mathbf{x}_i, \boldsymbol{\beta}, \sigma) = f(\sigma^{-1}(\mathbf{x}_i^T \boldsymbol{\beta} - \mu^*))^{1-q}, \quad v_q(\beta_j, \lambda) = \pi(\beta_j; \lambda)^{1-q} \quad (4)$$

are the vectors of weights applied to the observations and the parameters, respectively. The weights w_q downweight observations \mathbf{x}_i that diverge from the assumed data model f , while v_q downweights the $|\hat{\beta}_j|$ that diverge from the assumed prior distribution π . For example, when $q < 1$, the linear combinations $\mathbf{x}_i^T \boldsymbol{\beta}$ that are far away from the target mean μ^* are assigned a small w_q . If $q \rightarrow 1$, $f(z)$ is the normal density function and $\pi(\beta; \lambda)$ is the Laplace function, we recover the popular Lasso method [9], in which $w_i = v_j = 1$. However, as shown by [10], since the weights in Lasso do not affect the optimization process, we may obtain unstable and inaccurate results in presence of large coefficients. Our approach proposes a remedy to such problem.

3 Re-weighting algorithms

The following section describes the weighting algorithm we introduce to estimate optimal portfolios in the general case in which data are assumed to follow a generic distribution f , and then focus on the specific cases in which f is a Normal or a t-Student distribution. The aim of the optimization process is to obtain the parameter estimates $\hat{\boldsymbol{\beta}}_{q,\lambda}$ by minimizing criterion (1). Since the L_q terms are typically non-convex in $\boldsymbol{\beta}$, we divide the whole process in several convex optimization steps. In particular, if we fix q , the vectors of weights w_q and v_q become $w_i, i = 1, \dots, n$ and $v_j, j = 1, \dots, p$, and the criterion results in a penalized likelihood problem that we can solve with an iteratively re-weighted scheme: given the weights w_i and v_j , we estimate $\hat{\boldsymbol{\beta}}_{q,\lambda}$ by solving equation (3) and then update the weights using the new parameter estimates. We call this process a doubly re-weighted (2RE) algorithm as the re-weighting is applied to both data and penalty scores.

Algorithm 3.1.

Given the tuning constants $q \leq 1$, $\lambda \geq 0$, and a target portfolio return μ^* , the algorithm consists of the following steps:

Step 0 At Iteration $s = 0$, compute the parameter estimates $\hat{\boldsymbol{\beta}}^{(s)}$ and $\hat{\sigma}^{(s)}$.

Step 1 Set $s = s + 1$, and update the vector of weights as

$$\hat{w}_i^{(s)} = f((\mathbf{x}_i^T \hat{\boldsymbol{\beta}}^{(s-1)} - \mu^*) / \hat{\sigma}^{(s-1)})^{1-q}, \quad \hat{v}_j^{(s)} = \pi(\hat{\beta}_j^{(s-1)}; \lambda)^{1-q}. \quad (5)$$

Step 2 Compute the parameter estimates $\tilde{\boldsymbol{\beta}}$ and $\tilde{\sigma}$ by minimizing

$$\sum_{i=1}^n \hat{w}_i \log f((\mathbf{x}_i^T \boldsymbol{\beta} - \mu^*) / \sigma) + \sum_{j=1}^p \hat{v}_j \log \pi(\beta_j; \lambda). \quad (6)$$

Step 3 Update $\widehat{\beta}^{(s)}$ and $\widehat{\sigma}^{(s)}$ by solving $f(\mathbf{x}_i^T \widehat{\beta} - \mu^*)/\widehat{\sigma}^q$ for β and σ .

Step 4 Repeat Steps 1 and 2 until a stopping criterion is satisfied.

In Step 3, a re-scaling operation re-centers the estimates to correct the bias arising from the weights $w_q(\mathbf{x}_i, \beta, \sigma)$, as suggested by [7].

The parameter λ controls the penalty term on the β coefficients and regulates the sparsity of the portfolio. The literature suggests to choose such tuning parameters by information criteria like the AIC and BIC. As [11], given a certain level of q , we select the optimal values of λ by minimizing the robust Bayesian Information Criterion defined as below, where $k \leq p$ is the number of active positions:

$$\text{BIC}_q = -2 \sum_{i=1}^n L_q \left\{ f \left(\frac{\mathbf{x}_i^T \widehat{\beta}_{q,\lambda} - \mu^*}{\widehat{\sigma}_{q,\lambda}} \right) \right\} + \log(n)k. \quad (7)$$

Normal portfolios

If we assume that data follow a p -variate normal distribution and $\pi(\beta_j; \lambda)$ is a Laplace function, then $Y \sim N(\mu, \sigma^2)$. In this case, the 2RE algorithm can be adapted as follows.

Algorithm 3.2.

Given $q \leq 1$, $\lambda \geq 0$, and a target return μ^* :

Step 0 At Iteration $s = 0$, initialize $w_i^{(s)}$, $v_j^{(s)}$ and $\sigma^{(s)}$.

Step 1 Set $s = s + 1$, and obtain $\widehat{\beta}^{(s)}$ by solving

$$\widehat{\beta}^{(s)} = \underset{\beta}{\text{argmin}} \left\{ \sum_{i=1}^n \widehat{w}_i^{(s-1)} \frac{1}{2} \left(\frac{\mu^* - \mathbf{x}_i^T \beta}{\widehat{\sigma}^{(s-1)}} \right)^2 + \lambda \sum_{j=1}^p \widehat{v}_j^{(s-1)} |\beta_j| \right\}, \quad (8)$$

Step 2 Update the vectors of weights as

$$\widehat{w}_i^{(s-1)} = \left[\frac{1}{\sqrt{2\pi\widehat{\sigma}^{2(s-1)}}} \exp \left\{ -\frac{\left(\mu^* - \mathbf{x}_i^T \widehat{\beta}^{(s-1)} \right)^2}{2\widehat{\sigma}^{2(s-1)}} \right\} \right]^{1-q}, \quad \widehat{v}_j^{(s-1)} = \left[\frac{\lambda}{2} \exp \left\{ -\lambda |\widehat{\beta}_j^{(s-1)}| \right\} \right]^{1-q}. \quad (9)$$

Step 3 When the portfolio variance is a fixed target σ^{*2} , we set $\widehat{\sigma}^{2(s)} = \sigma^{*2}$, for all $s \geq 0$; otherwise

$$\widehat{\sigma}^{2(s)} = \frac{\sum_{i=1}^n \widehat{w}_i^{(s-1)} \left(\mu^* - \mathbf{x}_i^T \widehat{\beta}^{(s-1)} \right)^2}{q \sum_{i=1}^n \widehat{w}_i^{(s-1)}}. \quad (10)$$

Step 4 Repeat Steps 1 to 3 until a stopping criterion is satisfied.

The optimization function in (8) is a weighted L_1 -penalized least squares problem that we solve by applying the gradient projection algorithm developed by [12]. Other algorithms, like coordinate wise and quadratic optimization ([9]), could be used to efficiently estimate $\widehat{\beta}^{(s)}$. However, as the gradient projection is faster and updates parameters and solutions by using the optimal values of the previous iteration as warm-start points ([13]), we rely on it for solving the penalized least squares problem.

t-portfolio

If we assume the portfolio Y to be a non standardized t-Student distribution with mean μ , variance σ and number of degrees of freedom $\nu > 1$ (i.e. $Y \sim f_\nu(\mu, \sigma)$), then Step 2 of the 2RE algorithm computes $\{\hat{\beta}^{(s)}, \hat{\sigma}^{(s)}\}$ as

$$\operatorname{argmin}_{\beta, \sigma} \left\{ -\left(\frac{\nu+1}{2}\right) \sum_{i=1}^n \hat{w}_i^{(s-1)} \log \left\{ 1 + \frac{(\mathbf{x}_i \beta^T - \mu^*)^2}{\nu \sigma^2} \right\} + \lambda \sum_{j=1}^p \hat{v}_j^{(s-1)} |\beta_j| \right\}, \quad (11)$$

where $\sigma > 0$. While the penalty weights \hat{v}_j are updated as in (9), the data weights \hat{w}_i are obtained as

$$\hat{w}_i^{(s-1)} = \left[f_\nu \left(\mathbf{x}_i^T \hat{\beta}^{(s-1)}; \mu, \hat{\sigma}^{(s-1)} \right) \right]^{1-q}, \quad i = 1, \dots, n. \quad (12)$$

When data are assumed to follow the nonstandardized t-Student distribution and $\lambda \rightarrow 0$, equation (11) results in biased estimates for β and σ . Thus, according to Proposition 1 in [7], we solve this issue by adjusting the degrees of freedom parameter: we use $\nu_q = q\nu + (q - 1)$ instead of ν . Also, the optimization function (11) represents a non-convex problem, which results in imprecise estimates if solved directly. Therefore, by writing a t-Student observation as a scale mixture of normals $Y_i \sim N(\mu, \sigma^2 Z_i^{-1})$, where Z_i follows a Gamma distribution $Z_i \sim \text{Ga}(\nu/2, \nu/2)$, we derive an EM algorithm, which efficiently estimates the optimal solutions as follows.

Algorithm 3.3.

For any $s > 0$, we set the initial weights $\hat{z}_i = 1/n$, $i = 1, \dots, n$ and estimate $\hat{\beta}^{(s)}$ and $\hat{\sigma}^{(s)}$ through the expectation-maximization steps:

M-Step Estimate β and σ as

$$\beta' = \operatorname{argmin}_{\beta} \left\{ \sum_{i=1}^n \hat{w}_i^{(s-1)} \hat{z}_i^{(s-1)} \frac{1}{2} \left(\frac{\mathbf{x}_i^T \beta - \mu}{\hat{\sigma}^{(s-1)}} \right)^2 + \lambda \sum_{j=1}^p \hat{v}_j^{(s-1)} |\beta_j| \right\}, \quad (13)$$

$$\sigma'^2 = \frac{\sum_{i=1}^n \hat{w}_i^{(s-1)} \hat{z}_i^{(s-1)} (\mathbf{x}_i^T \hat{\beta} - \mu)^2}{\sum_{i=1}^n \hat{w}_i^{(s-1)} \hat{z}_i^{(s-1)}} \times \frac{\nu}{(\nu+1)q-1}. \quad (14)$$

E-Step Update the mixing constants \hat{z}_i , such that

$$\hat{z}_i = \frac{(\nu_q + 1)\sigma'^2}{\nu_q \sigma'^2 + \hat{w}_i^{(s-1)} (\mathbf{x}_i^T \beta' - \mu)^2}, \quad i = 1, \dots, n, \quad (15)$$

4 Simulation study

In the following simulation study, we evaluate and compare the behaviour of the 2RE algorithm, for both normal (GDL_N) and t-Student portfolios (GDL_t), with respect to the Lasso penalization model. In particular, we want to test the robustness of the proposed methods in presence of outliers and correlated assets \mathbf{X} . We simulate data from a multivariate t-Student distribution with ν degrees of freedom: $t_p(\mu, \Sigma, \nu)$, where $\mu_j = 1$, if $j \leq k$, and $\mu_j = 0$, if $j > k$, and the covariance matrix has diagonal elements $\Sigma_{jj} = 1$, $j = 1, \dots, p$, and off-diagonal elements

$\Sigma_{jk} = \rho$, $0 \leq \rho < 1$ $j \neq k$. We construct four settings by considering four different levels of correlation between assets, $\rho = 0.2, 0.4, 0.6, 0.8$. For each setting, we generate $B = 50$ samples with $n = 500$, $p = 50$, $k = 10$.

We evaluate the average performance of the B portfolios in terms of sparsity, model selection performance and risk/return characteristics with respect to a specific target $\mu^* = k$. In particular, we compute (i) the number of active positions as $\hat{k} = \sum_{j=1}^p I(|\hat{\beta}_j| > \tau)$, where $\tau = 0.005$ is a threshold value, below which the estimated weights are set equal to zero; (ii) the F-measure to assess whether the portfolios select the "correct" assets, which in our model are the ones in the first k positions; (iii) the Monte Carlo mean squared error to compare the risk/return performance to the specified target:

$$\widehat{MSE} = \frac{1}{B} \sum_{b=1}^B \left(\frac{\boldsymbol{\mu}^T \hat{\boldsymbol{\beta}}_b - \mu^*}{\sqrt{\hat{\boldsymbol{\beta}}_b^T \boldsymbol{\Sigma} \hat{\boldsymbol{\beta}}_b}} \right)^2, \quad F\text{-measure} = 2 \frac{|\text{supp}(\boldsymbol{\beta}^*)| \cap |\text{supp}(\hat{\boldsymbol{\beta}})|}{|\text{supp}(\boldsymbol{\beta}^*)| + |\text{supp}(\hat{\boldsymbol{\beta}})|}, \quad (16)$$

where, given a vector $\boldsymbol{\beta}$, the support is equal to $\text{supp}(\boldsymbol{\beta}) = \{j : |\beta_j| \geq \tau\}$, and $\boldsymbol{\beta}^*$ represents the vector of weights whose first k positions are equal to 1.

For each setting, we set $q = 0.9$ and select from a grid of values the λ associated to the model with the lowest BIC. We then compare the average portfolio performances with the ones obtained using Lasso. As specified in Section 3, we handle the optimization problem by using the DC-programming as proposed by [13]. As the EM algorithm is very sensitive to the initialization of β , we initialize the Lasso and the GDL_N algorithms with the OLS β estimates, while the GDL_t approach uses instead the optimal estimates obtained by the GDL_N. Finally, the initial vectors of weights w_i and v_j are set equal to $w_i = 1/n$ and $v_j = 1/p$.

Figure 1 shows from left to right the boxplots of the average number of active positions \hat{k} estimated by the GDL methods and Lasso (a), and the relative F-measure (b) and MSE (c) obtained in 50 simulations for different values of correlation $\rho = 0.2, 0.4, 0.6, 0.8$ on the x-axis. We can compare the performance of the three methods in terms of sparsity and selection ability, and analyse their robustness in presence of correlated data.

First of all, we notice that the GDL criteria estimate much sparser portfolios than Lasso for each value of ρ . The number of active positions is very close to the optimal value of 10 and it is not influenced by the level of correlation between assets (Panel (a)). The stability of the GDL criteria represents a clear advantage when comparing with Lasso, whose performance becomes worse when ρ increases: on average it selects approximately 17 assets when $\rho = 0.2$ and 27 assets when $\rho = 0.8$, against the 8 and 11 assets selected by the GDL_t with ρ equal to 0.2 and 0.8, respectively. In terms of F-measure, the GDL approaches obtain better performance than Lasso as closer to 1, showing very good model selection properties. However, for all the methods, the average value of F-measure highly depends on the level of ρ (Panel (b)): when data exhibit low correlation, Lasso obtains a value of 0.74 while GDL_N and GDL_t are closer to the maximum of 1, that represents the case in which we select the correct vector of assets $\boldsymbol{\beta}^*$; when data are highly correlated, Lasso presents a value of 0.52, while the GDL methods obtain approximately 0.6. The GDL_N and GDL_t algorithms show similar results in terms of sparsity and F-measure since they both select the same active positions and their estimated weights differ only in magnitude. Finally, we analyse the overall performance of the three methods with respect to the return target μ^* by comparing their MSE. Though the two GDL criteria slightly differ in their results, they both outperform the Lasso, whose performance get much worse when

data show high correlation (i.e. with $\rho = 0.8$ the MSE is twice the value obtained with $\rho = 0.2$). As expected, given that the true model is a t-Student one, the GDL_t obtains the lowest MSE in all settings, indicating very good performance. However, this advantage might also result from the initialization of the vector of β as the optimal solution of the GDL_N algorithm.

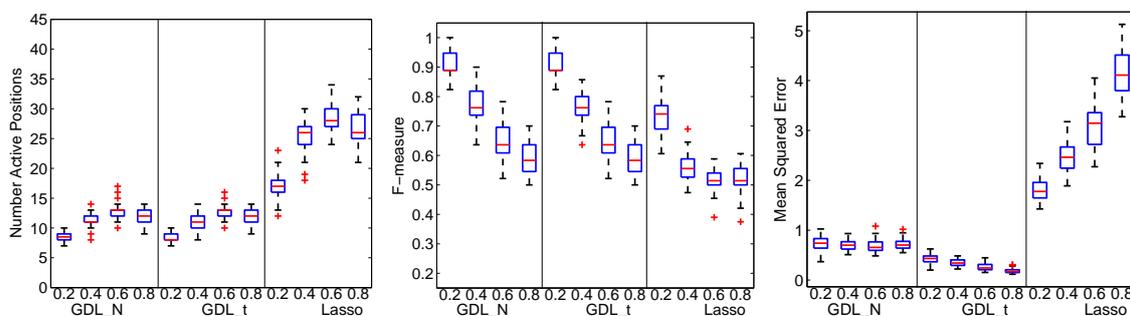


Figure 1: Average number of estimated active positions \hat{k} , F-measure and Mean Squared Error for different levels of correlation ρ in 50 simulations, using GDL for Normal and t-Student, and Lasso methods.

Further simulations considering different set-ups support the main reported findings. Results are available upon request. This study points out the main advantages of the proposed approach with respect to a well-known benchmark: (i) the sparsity of the selected portfolios obtained by penalizing and weighting the vector of asset weights β ; (ii) the high robustness of the estimates in presence of correlation between assets, which is ensured by weighting the observations according to their divergence from an assumed distribution.

5 Sparse and Robust Index Tracking

In this section, we test our approach in an index tracking framework, where we try to reproduce the performance obtained by a certain index by selecting a vector of active weights only for some of its components, in order to limit transaction and managing costs. The optimization problem can be described as a regression problem, where the dependent variable \mathbf{y} represents the vector of index returns and \mathbf{X} is the return matrix of its components.

Using a penalized technique may help to obtain good out-of-sample performance with respect to the index by optimally selecting a small number of components. In order to evaluate the behaviour of the proposed GDL criteria, we focus on three financial indexes by using $n = 1401$ daily return observations of the Fama & French 100, the S&P 200 and the S&P 500, with different number of constituents p , equal to 100, 200 and 500, respectively. For each index, we compare the performance of three strategies: the GDL for Normal and t-Student portfolios, and the Lasso.

We estimate the optimal portfolios using a rolling window sample of 250 observations, and compute the excess return of the first out-of-sample observation with respect to the index. For the GDL criteria, we set $q = 0.9$ and select the λ in each window as described in Section 3. First, we evaluate the risk/return performance of the optimal portfolios through the Information Ratio (IR), which is computed dividing the excess return by the tracking error volatility (TEV). Then,

we check sparsity by means of the number of estimated active positions \hat{k} and finally, we test the tracking ability computing the correlation with respect to the index.

Strategy	ER (%)	TEV (%)	IR	\bar{k}	TO	Cor
PANEL A: F&F 100						
GDL N	0.338	0.624	0.542	37.749	0.068	0.999
GDL t	0.170	0.492	0.346	32.241	0.066	0.999
Lasso	1.030	2.117	0.486	65.939	0.017	0.990
PANEL B: S&P 200						
GDL N	0.319	4.500	0.071	36.950	0.399	0.963
GDL t	-2.421	4.897	-0.494	28.431	0.520	0.933
Lasso	4.760	7.267	0.655	66.532	0.037	0.950
PANEL C: S&P 500						
GDL N	2.906	6.966	0.417	44.564	0.605	0.932
GDL t	1.192	9.018	0.132	27.770	0.811	0.872
Lasso	2.986	10.315	0.289	66.407	0.053	0.926

Table 1: Out-of-sample statistics of each tracking portfolio: strategy (column 1), annualized excess return ER (column 2), tracking error volatility TEV (column 3), Information Ratio IR (column 4), average number of active components \bar{k} (column 5), turnover TO (column 6), correlation w.r.t. index Cor (column 7).

Table 1 shows the out-of-sample statistics of each tracking strategy. In terms of IR (Column 4), the GDL_N has the best performance for F&F 100 and S&P 500, while the Lasso outperforms the other strategies in the second dataset, S&P 200. However, the GDL criteria always obtain a lower out-of-sample TEV (Column 3), which is a characteristic already underlined in the simulation study, where the GDL showed smaller MSE than Lasso. This result is even more important if we consider that the GDL strategies select very sparse solutions for each dataset (Column 5). While the Lasso always uses approximately 66 positions, the GDL strategies select 35% of the available assets for the first index, less than 25% for the second index and less than 10% for the third index. In terms of tracking ability, the GDL_N portfolios achieve values of Cor near 1 and outperform the Lasso by closer tracking the indexes, especially in small dataset, where the TEV is lower.

6 Conclusion

In this paper we propose a generalized description length criterion to obtain sparse and robust portfolios in presence of estimation errors and model misspecification. By relying on a q -entropy measure, the approach minimizes the uncertainty about the distributions of data and model parameters by assigning a lower weight to observations and parameters that diverge from the assumed models. After deriving the general estimation algorithm, we specify two interesting cases, in which data are assumed to follow a Normal or a t-Student distribution, and develop the corresponding algorithms, GDL_N and GDL_t. The simulation study supports the theoretical properties of the GDL criterion and shows that it achieves better performance in terms of sparsity, stability and robustness of the estimates with respect to the well-known Lasso benchmark, especially when data exhibit high correlation. The empirical results presented for the index

tracking framework show that the GDL criterion is able to obtain good out-of- sample estimates and reproduce the performance of an index by using only a small number of its components in order to limit transaction and managing costs.

Bibliography

- [1] H. Markowitz, Portfolio selection, *The Journal of Finance* 7 (1952) 77–91.
- [2] M. J. Best, R. R. Grauer, On the sensitivity of mean-variance efficient portfolios to changes in asset means: some analytical and computational results, *The Review of Financial Studies* 4(2) (1991) 315–342.
- [3] R. Jagannathan, T. Ma, Risk reduction in large portfolios: Why imposing the wrong constraints helps, *Journal of Finance* LVIII (2003) 1651–1683.
- [4] V. De Miguel, F. J. Nogales, Portfolio selection with robust estimation, *Operations Research* 57 (2009) 560–577.
- [5] R. Cont, Empirical properties of asset returns: stylized facts and statistical issues, *Quantitative Finance* 1 (2001) 223–236.
- [6] J. Havrda, F. Charvát, Quantification method of classification processes: Concept of structural entropy, *Kibernetika* 3 (1967) 30–35.
- [7] D. Ferrari, D. La Vecchia, On robust estimation via pseudo-additive information, *Biometrika* 99 (1) (2012) 238–244.
- [8] P. Grünwald, *The Minimum Description Length Principle (Adaptive Computation and Machine Learning)*, The MIT Press, 2007.
- [9] R. Tibshirani, Regression shrinkage and selection via the lasso, *Journal of the Royal Statistical Society* 58 (1996) 267–288.
- [10] J. Fan, R. Li, Variable selection via nonconcave penalized likelihood and its oracle properties, *Journal of American Statistical Association* 96 (2001) 1348–1360.
- [11] E. Ronchetti, Robustness aspects of model choice, *Statistica Sinica* 7 (1997) 327–338.
- [12] M. Figueiredo, R. Nowak, S. Wright, Gradient projection for sparse reconstruction: application to compressed sensing and other inverse problems, *IEEE Journal of Selected Topics in Signal Processing: Special Issue on Convex Optimization Methods for Signal Processing* 1, 4 (2007) 586–598.
- [13] G. Gasso, A. Rakotomamonjy, S. Canu, Recovering sparse signals with a certain family of nonconvex penalties and dc programming, *IEEE Transactions on Signal Processing* 57, 12 (2009) 4686–4698.

Longitudinal data mining to predict survival in a large sample of adults

Paolo Ghisletta, *University of Geneva*, paolo.ghisletta@unige.ch
Stephen Aichele, *University Geneva*, stephen.aichele@unige.ch
Pat Rabbitt, *University of Oxford*, patrick.rabbitt@psy.ox.ac.uk

Abstract. We applied data mining techniques to explore survival in a sample of 6'203 adults (age range 42-93 years), living in the Manchester and Newcastle-upon-Tyne (UK.) areas. We were particularly interested in the relations between cognitive performance and mortality prediction. Participants were assessed up to four times over 20 years on several psychological and health-related variables and were also administered an extensive battery of cognitive tasks. We applied linear mixed models to estimate level of cognitive decline and change (mostly decline) therein for each individual. We then utilized Cox proportional-hazards modeling to predict time to death based on levels of and changes in cognitive performance, and on demographic and social predictors. Next, to gain further insight into the survival process, we used recently developed induction trees and ensemble methods. These models allow studying complex and asymmetric interactions and non-additive functions of model predictors. Particularly relevant to our theoretical purposes, the random forest approach allowed us to identify a set of demographic and cognitive variables that strongly influenced survival. We conclude that induction trees and ensemble methods are a useful extension to more classical models in that they are not limited by common modeling assumptions and can reveal complex patterns of relation.

Keywords. longitudinal data mining, survival analysis

1 Introduction

The prediction of selected outcomes is of central importance to scientific inquiry, and statistical modeling is essential to this endeavor. When the outcome is a specific event, such as death, and the investigative aim is to predict the amount of time preceding this event, the preferred approach is generally survival analysis [1]. In survival analysis, the dependent variable is indicative of the occurrence of the event of interest, contingent upon the amount of time elapsed before the event. A survival model can also accommodate data from observations for which the event has not occurred (i.e., right-censored data).

In survival analysis the specification of predictors is not limited to their main (direct) effects. Interaction (indirect) effects, generated via multiplications of predictors, can also be tested. Such interactions are additive, in that they add to (or subtract from) the main effect the amount of influence contingent upon levels of one or more additional predictors. There are four limitations of this conceptualization of interactions. First, such interactions are typically specified by the analyst a priori, and this excludes the exploration of all indirect effects of a predictor. Second, statistical reliability requires the interactions to be defined over the entire data space. However, in practice the data are often too sparse to include all instantiations of an interaction. Third, although testing higher-order interactions is possible, for instance by multiplying more than two variables, the interpretation of such interactions is often arduous. Therefore, analysts typically limit their analyses to include two-way interactions. Fourth, the statistical complexity inherent to nonlinear interactions often prohibits researchers from examining them.

One approach to address these shortcomings, and thereby gain further insight from classical survival regression models, is to employ Induction Trees (ITs; also called Classification and Regression Trees, [2]), a family of data mining techniques that originated from the machine learning literature. ITs have gained much interest in genetics, epidemiology, and medicine, where oftentimes the analyst faces the so-called “small n , big p ” problem, in which data from a large number of variables is obtained from relatively small samples of observations. For an excellent introduction to ITs and extensions thereof, see [19].

Here, we will use survival trees to explore demographic, social, and cognitive performance variables as predictive of survival in a large sample of British adults who were tested repeatedly across a span of 20 years. Although our application of ITs does not fall within the “small n , big p ” situation, we apply survival trees to extend our knowledge about the survival process gained from classical survival models. In particular, we explore complex, asymmetric interactions among predictors, as well as non-additive functions of the predictors. Finally, we apply recently developed ensemble methods, to examine the robustness of the survival tree results.

2 Sample and Measures

Sample

The data come from the University of Manchester Longitudinal Study [17], a large-scale 20-year longitudinal examination of cognitive performance in a large sample of cognitively healthy adult individuals, who initially ranged in age from 42 to 96 years. The original researchers tracked changes in a large number of variables related to participants’ demographics, cognitive functioning, social functioning, health, etc. Overall, the sample includes 6203 volunteering participants. The majority (70.6%) were female, and overall 45.5% came from the Greater Manchester (UK.) area, while the remaining 54.5% from the Newcastle-upon-Tyne (UK.) area. Participants did not suffer from major visual or auditory handicaps and could wear corrective aids during assessments. The Registrar General’s Scale of Occupational Categories [14] was used to classify participants according to six levels of socio-economic status: professional (4.7%), intermediate (31.6%), non-manual (26.8%), manual skilled (21.6%), partly skilled (7.4%), and unskilled (0.8%) - for 7.1% this information was unknown.

The most recent mortality update by the researchers at the University of Manchester took place in August 2012. At that point, 1906 of the initial participants were still alive, 4085 were deceased, and information were not available for the remaining 212 individuals.

Cognitive assessment

Two cognitive batteries were administered to participants. Both batteries contained tasks assessing perceptual speed (the speed at which simple, abstract information is processed), fluid intelligence (basic cognitive abilities such as reasoning, independent of prior learning and acquisition), crystallized intelligence (higher-order abilities to use cultural, educational, and vocational knowledge and experience to learn new information), and memory (verbal and visual types). Table 1 outlines these tasks, classified according to the cognitive domain assessed, and accompanied by the abbreviations, to which we refer hereafter. Participants were tested in groups of 5-20 by two trained experimenters in well-lit, comfortable and quiet rooms. At each testing occasion, tasks were administered across two sessions of about 90 minutes each. Further detail is available in [17] and [7].

Domain	Task	Abbreviation
Perceptual speed	Visual search	vs
Perceptual speed	Alphabet coding task	act
Perceptual speed	Semantic reasoning	sr
Fluid intelligence	Heim intelligence test 1	aha
Fluid intelligence	Heim intelligence test 2	ahb
Fluid intelligence	Cattell's culture fair test	cft
Crystallized intelligence	Raven Mill Hill vocabulary A	mha
Crystallized intelligence	Raven Mill Hill vocabulary B	mhb
Crystallized intelligence	Wechsler's Adult Intelligence Scale - vocabulary	waisv
Verbal memory	Verbal free recall	vfr
Verbal memory	Cumulative verbal recall	cvr
Verbal memory	Immediate verbal free recall	ivfr
Verbal memory	Propositions about people	pap
Verbal memory	Memory objects	mo
Visual memory	Picture recognition	pr
Visual memory	Shape + spatial locations	shspl

Table 1: Cognitive tasks assessed in the University of Manchester Longitudinal Study.

We performed all analyses in the open source and freely available R language and environment. In a series of preliminary analyses we applied linear mixed-effects models to analyze cognitive performance as a function of age. This is typical in developmental psychology, where an age-appropriate description of phenomena is of major theoretical interest. This analytical approach allowed us describing both the sample average trajectory (fixed effects) of each cognitive task and the individuals deviations (random effects) from the sample average trajectory. The analyses thus characterized individuals with respect to their overall average performance and also their rate of linear change (typically decline) in performance across the repeated measures (i.e., as individuals aged). From those analyses we estimated each participant's intercept and linear slope score, to be used here as markers of cognitive performance to predict survival. The association between cognitive decline and mortality is a theme of long-lasting interest in the psychological literature (for a recent review see [6, 7]). Note that for two tasks (*mhb* and *waisv*) there were no reliable interindividual differences in change, hence no estimate of linear slope

could be obtained. The usual assumptions of linear mixed-effects models (normality of random effects and of residuals, homoskedasticity of residuals; [15]) were not violated (clearly the large sample size was advantageous for estimation).

3 Survival analysis

Next, we estimated survival using the well-known Cox proportional-hazards model [4]. Given that we had no a-priori hypothesis concerning interactions between predictor variables, we only tested their main effects. In a first model we included participants' (a) initial age upon entry into the study, (b) sex, (c) city of origin, (d) socio-economic status, and (e) cognitive performance, in terms of intercept and linear slope scores for each cognitive variable. We used the package `survival` (version 2.37-7) of the R (version 3.1.0) language and environment [16].

When we checked the proportional-hazards assumption of this first model, it was clear that the hazards were non-proportional for initial age, sex, and city. We thus specified a second survival model that included interactions of survival time with age, sex, and city. The proportional-hazards assumption of this second model was met for all predictors (except, as expected, for the interactions with time). Furthermore, there were no particularly influential observations, nor evidence of nonlinearity (for a full description of model diagnostics see [5]).

Table 2 shows the parameter estimate of each predictor of the second survival model. For space reasons, we only include predictors whose parameter estimate appears, according to the z-test, to be different from zero.

Predictor	estimate	exp(est.)	lower 95%	upper 95%	z-value	p-value(> z)
AgeFirst_Aprox	1.654e+00	5.227e+00	4.771e+00	5.726e+00	35.534	< 2e-16
AgeFirstLast	-1.972e-02	9.805e-01	9.794e-01	9.815e-01	-36.773	< 2e-16
Female	-2.005e+00	1.346e-01	2.794e-02	6.485e-01	-2.500	0.012427
FemaleAgeLast	2.255e-02	1.023e+00	1.004e+00	1.042e+00	2.386	0.017040
Newcastle	3.901e+00	4.944e+01	9.020e+00	2.710e+02	4.494	7.00e-06
NewcastleAgeLast	-4.897e-02	9.522e-01	9.333e-01	9.715e-01	-4.792	1.65e-06
LinearSlope.aha	8.161e-01	2.262e+00	1.191e+00	4.295e+00	2.494	0.012631
LinearSlope.ahb	-6.256e-01	5.349e-01	3.119e-01	9.175e-01	-2.273	0.023046
Intercept.mha	5.197e-02	1.053e+00	1.020e+00	1.088e+00	3.145	0.001659
LinearSlope.mha	-3.258e+00	3.846e-02	2.290e-03	6.459e-01	-2.264	0.023593
Intercept.mhb	-2.207e-02	9.782e-01	9.587e-01	9.980e-01	-2.154	0.031256
Intercept.vfr	3.340e-02	1.034e+00	1.005e+00	1.063e+00	2.338	0.019398
Intercept.cvr	-2.250e-02	9.778e-01	9.595e-01	9.963e-01	-2.346	0.018955
LinearSlope.cvr	1.117e+00	3.057e+00	1.602e+00	5.833e+00	3.389	0.000702
LinearSlope.cft	-1.757e-01	8.389e-01	7.617e-01	9.239e-01	-3.567	0.000361
Intercept.shspl	-7.982e-02	9.233e-01	8.555e-01	9.965e-01	-2.050	0.040316

Table 2: Results from a survival analysis (only significant estimates are shown). These estimated effects correspond to the “hazard” of death as an outcome (i.e., the log-odds change in probability of death, with smaller values indicative of relatively longer projected life span).

As expected, initial age influences the hazard of death (for each additional year, the hazard

increases by over 400%). This effect, however, diminishes with age (by about 2% per year). Females have a hazard of dying that is 87% that of males. This proportion, however, diminishes by 2% per year. The hazard of death of residents of Newcastle-upon-Tyne is nearly 5000% that of Mancunians, and this disadvantage diminishes by 4% each year. Finally, performance on several cognitive variables, both in the intercept and in the linear slope, is related to survival.

Although the diagnostics of this model were satisfying and no estimation issues appeared, the very high estimates of a age and city may indicate the fragility of this solution. This may be due, in part, to a bad specification with respect to interactions among predictors [19]. We thus turn to induction trees and ensemble methods to verify these doubts.

4 Survival tree

To test for complex, asymmetric interactions we computed a survival tree on the same data, using the package `party` (version 1.0-13)[9]. This package computes survival trees based on accelerated failure time models, rather than Cox proportional-hazards model, mainly because of four specific issues: (a) to handle both censored and uncensored data within the same model; (b) to remove restrictive model assumptions (in particular that hazards may not be proportional but rather accelerate linearly or nonlinearly with time); (c) to deal with problems of high dimensionality (when a large number of predictors is tested); and (d) to accommodate selection and evaluation of models with strict statistical criteria (for more details see [10]).

A survival tree classifies observations into groups based on the proximity of their survival information, contingent upon the predictors of the model. For each predictor, the cutoff value that maximally discriminates the survival information into two groups is found. This procedure is dictated by well-defined statistical criteria such as entropy measures (e.g., Gini Index, Shannon Entropy). A predictor may intervene multiple times in the overall tree, and may thus interact with other predictors that also produce a separation into two groups.

The resulting survival tree is shown graphically in Figure 1. In the end, 19 groups are distinguished based on their survival information. These are arranged from left to right in ascending survival order. As can be seen, these groups are defined based on complex interactions among several predictors. For instance, the group with the lowest survival is composed of 59 individuals younger than 67 years, males, with a linear slope score of `cft` lower than or equal to -3.265, and a linear slope score of `sr` lower than or equal to -0.427. The group with the highest survival is composed of 18 individuals older than 85 years (the package also displays this information in a text output, not shown for space reasons). This analysis shows that relatively older individuals at the start of the study were less likely to die at an earlier age than more youthful participants (a manifest selectivity effect).

Sex appears three times as an important discriminant variable, in interaction with initial age and with several cognitive predictors. Five cognitive variables also appear as discriminant, mainly with respect to their change information (linear slope rather than intercept). This indeed confirms that individuals with accelerated rates of cognitive decline have lower probabilities of survival into old age than those with shallower rates of cognitive loss.

the forest (estimated at 34.75%) is minimized after about 900 trees, which indicates that 1'000 trees are sufficient to obtain a stable solution. Moreover, we see that by far initial age is the most important variable, followed by several markers of cognitive change.

In a follow-up analysis we recomputed a random survival forest but excluded initial age. We observed that the most important cognitive predictors remained unchanged. Also, to check the robustness of the initial survival forest, we computed a number of additional analyses, in which we systematically altered the number of variables randomly sampled from all predictors at each branch. As suggested by Breiman (see [13]), we estimated forests, which used either twice the default or half the default number of predictors for each branch. In all cases, the error rate was subject to minor changes (less than 1%) and the relative importance of the predictors remained virtually the same. Moreover, 1'000 trees always resulted in stable estimates of error rate and variable importance.

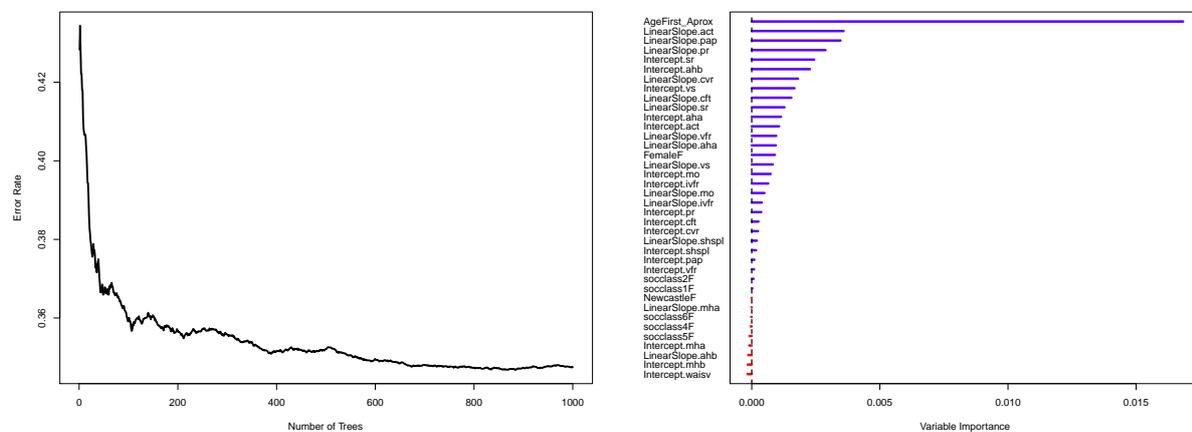


Figure 2: Results from a survival tree

6 Conclusions

In this application, we wanted to explore the robustness of the results of a Cox proportional-hazards model, despite the fact that the model's diagnostics were reassuring (indicating that the assumptions were probably met). Moreover, we wanted to avoid overfitting the model to our data. The statistical approach illustrated here allowed us obtaining further evidence in favor of the psychological hypothesis stating that individuals with steeper cognitive decline are more likely to die at an earlier age than individuals with age-resistant trajectories of cognitive performance. Furthermore, this predictive effect appears pervasive across multiple cognitive domains, rather than specific to a given domain.

Modern computational means allow implementing easily random survival forests even on basic, inexpensive portable computers. This statistical procedure, which relies heavily on re-sampling techniques, can thus be used to complement classical "one-shot" predictive analyses or even replace them when adequate. Finally, the R language and environment allows implementing induction trees and random forests on a wide variety of operating systems. Examples and tutorials are available on the internet and will certainly become more numerous in the near future.

For all these reasons we think that induction trees and random forests, and, more generally, ensemble methods are a readily available opportunity that should not be ignored by modern data analysts.

Bibliography

- [1] Allison, P. D. (1984). *Event history analysis*. Beverly Hills, CA: Sage.
- [2] Breiman, L., Friedman, J., Olshen, R. A., and Stone, C. J. (1984). *Classification and regression trees*. New York, NY.: Chapman & Hall.
- [3] Breiman, L. (2001). *Random forests*. Machine Learning, **1**, 5–32.
- [4] Cox, D. R. (1972). *Regression models and life tables*. Journal of the Royal Statistical Society Series B (Methodological), **34**, 187-220.
- [5] Fox, J. (2002). *Cox proportional-hazards regression for survival data*. <http://cran.r-project.org/doc/contrib/Fox-Companion/appendix-cox-regression.pdf>.
- [6] Ghisletta, P. (2008). *Application of a joint multivariate longitudinal-survival analysis to examine the terminal decline hypothesis in the Swiss Interdisciplinary Longitudinal Study on the Oldest Old*. Journal of Gerontology: Psychological Sciences, **63B**, P185-P192.
- [7] Ghisletta, P. (2014). *Recursive partitioning to study terminal decline in the Berlin Aging Study*. In McArdle, J. J. and Ritschard, G. (Eds.), *Contemporary issues in exploratory data mining in the behavioral sciences* (pp. 405–428). New York, NY.: Routledge Academic.
- [8] Ghisletta, P., Rabbitt, P., Lunn, M. and Lindenberger, U. (2012). *Two thirds of the age-based changes in fluid and crystallized intelligence, perceptual speed, and memory in adulthood are shared*. Intelligence, **40**, 260–268.
- [9] Hothorn, T., Hornik, K., and Zeileis, A. (2006). *Unbiased recursive partitioning: A conditional inference framework*. Journal of Computational and Graphical Statistics, **15**, 651–674 <http://cran.r-project.org/web/packages/party/index.html>.
- [10] Hothorn, T., Bühlmann, P., Dudoit, S., Molinaro, A., and Van Der Laan, M. J. (2006). *Survival ensembles*. Biostatistics, **7**, 355-373.
- [11] Ishwaran, J. and Kogalur, U. B. (2013) *Random Forests for Survival, Regression and Classification (RF-SRC)*. <http://cran.r-project.org/web/packages/randomForestSRC/index.html>.
- [12] Ishwaran, J. and Kogalur, U. B. (2007). *Random survival forests for R*. R News, **7**, 25–31 http://CRAN.R-project.org/doc/Rnews/Rnews_2007-2.pdf.
- [13] Liaw, A. and Wiener, M. (2002) *Classification and regression by randomForest*. R News, **2**, 18–22 http://cran.r-project.org/doc/Rnews/Rnews_2002-2.pdf.
- [14] Office of Population Censuses and Surveys (1980). *Classification of occupations 1980*. London, UK.: Her Majesty’s Stationery Office.

- [15] Pinheiro, J. C., and Bates, D. M. (2000). *Mixed-effect models in S and S-PLUS*. New York, NY.: Springer.
- [16] R Core Team (2014). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. <http://www.R-project.org/>.
- [17] Rabbitt, P. M. A., McInnes, L., Diggle, P., Holland, F., Bent, N., Abson, V., Pendleton, N. and Horan, M *The University of Manchester longitudinal study of cognition in normal healthy old age, 1983 through 2003*. *Aging, Neuropsychology, and Cognition*, **11**, 245–279
- [18] Strobl, C., Boulesteix, A.-L., Kneib, T., Augustin, T., and Zeileis, A. (2008). *Conditional variable importance for random forests*. *BMC Bioinformatics*, **9** (307) <http://www.biomedcentral.com/1471-2105/9/307>.
- [19] Strobl, C., Malley, J. and Tutz, G. (2009) *An introduction to recursive partitioning: Rationale, application and characteristics of regression trees, bagging, and random forests*. *Psychological Methods*, **14**, 323–348.

Penalized Least Squares for Optimal Sparse Portfolio Selection

Bjoern Fastrich, *University of Giessen*, Bjoern.Fastrich@wirtschaft.uni-giessen.de
Sandra Paterlini, *EBS Universität für Wirtschaft und Recht*, Sandra.Paterlini@ebs.edu
Peter Winker, *University of Giessen*, Peter.Winker@wirtschaft.uni-giessen.de

Abstract. Markowitz portfolios often result in an unsatisfying out-of-sample performance, due to the presence of estimation errors in inputs parameters, and in extreme and unstable asset weights, especially when the number of securities is large. Recently, it has been shown that imposing a penalty on the 1-norm of the asset weights vector not only regularizes the problem, thereby improving the out-of-sample performance, but also allows to automatically select a subset of assets to invest in. Here, we propose a new, simple type of penalty that explicitly considers financial information and consider several alternative non-convex penalties, that allow to improve on the 1-norm penalization approach. Empirical results on U.S.-stock market data support the validity of the proposed penalized least squares methods in selecting portfolios with superior out-of-sample performance with respect to several state-of-art benchmarks.

Keywords. Penalized Least Squares, Regularization, LASSO, Non-convex penalties, Minimum Variance Portfolios

1 Introduction

The Markowitz mean-variance portfolio model [1] is the cornerstone of modern portfolio theory. Given a set of assets with expected return vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$, Markowitz's model aims to find the optimal asset weight vector that minimizes the portfolio variance, subject to the constraint that the portfolio exhibits a desired portfolio return. Since $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are unknown, some estimates $\hat{\boldsymbol{\mu}}$ and $\hat{\boldsymbol{\Sigma}}$ must be obtained from a finite sample of data to compute the optimal asset allocation vector. As financial literature has largely shown, using sample estimates can hardly provide reliable out-of-sample asset allocations in practical implementations [2],[3],[4],[5],[6]. [7], [8], [2], and [9] already provided strong empirical evidence that estimates of the expected portfolio return and variance are very unreliable. Here, we focus on the minimum-variance portfolio (MVP), which relies solely on the covariance structure and neglects the estimation of expected returns altogether [10],[11],[12],[13],[14],[15],[16]. Somewhat surprisingly, MVPs are usually found to perform better out-of-sample than portfolios that consider asset

means [17, 11, 6], because the (co)variances can be estimated more accurately than the means. A superior performance also prevails when performance measures consider both portfolio means and variances. Nevertheless, MVPs still suffer considerably from estimation errors [10],[11],[12].

One stream of research has recently focused on shrinking asset allocation weights by using penalized least squares methods. Among the first contributors, [18] and [19] use ℓ_1 -penalization to obtain stable and sparse (i.e. with few active weights) portfolios, which is an adaptation of the Least Absolute Shrinkage and Selection Operator (LASSO) by [20]. The LASSO relies on imposing a constraint on the ℓ_1 -norm the regression coefficients $\beta \in \mathbb{R}^K$, where $\ell_1 = |\beta_1| + \dots + |\beta_K|$. Recently, [14] provide both theoretical and empirical evidence supporting the use of ℓ_1 -penalization to identify sparse and stable portfolios by limiting the gross exposure, showing that this causes no accumulation of estimation errors, the result of which is an outperformance compared to standard Markowitz portfolios. Further examples of penalised methods applied in the Markowitz framework are [21, 22, 23], and [15].

Despite the appeal of using ℓ_1 -penalization in portfolio optimization to estimate (numerically stable) asset weights and select the portfolio constituents in a single step by solving a convex optimization problem, [24] show that the ℓ_1 -penalty, as a linear function of absolute coefficients, tends to produce biased estimates for large (absolute) coefficients. As a remedy, they suggest using penalties that are singular at the origin, just like the ℓ_1 -penalty, in order to promote sparsity, but non-convex, in order to countervail bias. Ideally, a good penalty function should result in an estimator with three properties: unbiasedness, sparsity, and continuity. Then, new non-convex penalties such as the so-called Smoothly Clipped Absolute Deviation (SCAD), the Zhang-penalty, the Log-penalty and the ℓ_q -penalties with $0 < q < 1$ were introduced (e.g. see [25] for a comparison). The seemingly nice properties of non-convex penalties come at the cost of posing a difficult optimization challenge, which, however, can nowadays be solved quite efficiently by using a dual-convex approach, as suggested by [25]. An alternative to non-convex approaches, which can still retain the oracle property, has been suggested by [26]. His approach is now known as the adaptive LASSO and has proven to be able to prevent bias while preserving convexity of the optimization problem, and thus clearly alleviates the optimization challenge as compared to the non-convex approaches.

This work contributes to the literature on portfolio regularization by proposing a new, simple type of convex penalty, which is inspired by the adaptive LASSO and explicitly considers financial information to optimally determine the portfolio composition. Moreover, we are the first to apply non-convex penalties in the Markowitz framework to identify sparse and stable portfolios with desirable out-of-sample properties, when dealing with a large number of assets.

2 Penalized Approaches for Minimum Variance Portfolios

Given a set of K assets and a penalty function $\rho(\cdot)$, the regularized minimum-variance problem can be stated as:

$$\mathbf{w}^* = \underset{\mathbf{w} \in \mathbb{R}^K}{\operatorname{argmin}} \left\{ \mathbf{w}' \Sigma \mathbf{w} + \lambda \sum_{i=1}^K \rho(w_i) \right\} \quad (1)$$

$$\text{subject to} \quad \mathbf{1}'_K \mathbf{w} = 1, \quad (2)$$

where \mathbf{w}^* is the optimal (and potentially sparse) $(K \times 1)$ -vector of asset weights, $\mathbf{1}_K$ is a $(K \times 1)$ -vector of ones and λ is the regularization parameter that controls the intensity of the penalty and

thereby the sparsity of the optimal portfolio. The optimization problem (1) can be re-written as a penalized least square problem.

Assuming we estimate Σ by $\widehat{\Sigma}$ and we set $\lambda=0$, the solution to problem (1)-(2) is the MVP, where the optimized portfolio weights vector \mathbf{w}^* is (over)fitted to the correlation structure in $\widehat{\Sigma}$, thereby assuming absence of estimation error and unlimited trust in the precision of the estimate $\widehat{\Sigma}$, which is obviously very naive. On the contrary, whenever $\lambda > 0$, the penalty term $\sum_{i=1}^K \rho(w_i)$ will allow to control for the estimation error by selecting only few active weights. The larger λ , the smaller the number of active weights and the total amount of shorting. The optimal solution \mathbf{w}^* is thus determined by a trade-off between the estimated portfolio risk and the corresponding penalty term, whose magnitude is controlled by λ .

In this work, we focus on penalty functions $\rho(\cdot)$ that are singular at the origin and thus allow a shrinkage of the components in \mathbf{w} to exactly zero. Hence, the corresponding approaches not only stabilize the problem to improve the out-of-sample performance, but simultaneously also conduct the asset selection step. Table 1 reports the definition of the six penalties functions we consider.

The Least Absolute Shrinkage and Selection Operator (LASSO) has already received considerable attention in the portfolio optimization context and therefore we choose it as a benchmark to test the validity of the newly proposed approaches. Due to the budget constraint, the minimum value that $\|\mathbf{w}\|_1$ can be shrunk to is one. This is possible only when the portfolio weights are shrunk towards zero until they are all non-negative, identifying the so-called no-shortsale portfolio. Increasing values of λ cause the construction of portfolios with less shorting, or more precisely, with a shrunken ℓ_1 -norm of the portfolio weight vector. This prevents the estimation errors contained in $\widehat{\Sigma}$ from entering unhindered in the portfolio weight vector. Note that while the intensity of shrinkage is controlled by the value of λ , the decision as to which assets to shrink and to which relative extent is determined by the estimated correlation structure.

The weighted Lasso approach, henceforth *w8Las*, was proposed in its statistical formulation by [26] to countervail the difficulties of the LASSO that are related to potentially biased estimates of large true coefficients [24]. The idea is to replace the equal penalty that is applied to all coefficients (here portfolio weights) with a penalization-scheme that can vary among the K portfolio weights. This can be achieved by introducing a weight ω_i for each of the absolute portfolio weights $|w_i|$. In general, the intuition is to over- or underweight some assets in comparison to the LASSO in order to improve performance. Specifically, this intuition depends on the method used to determine the ω_i , for which no “blueprint” exists in a portfolio optimization context. We suggest determining the (individual) regularization weights λ_i by considering specific financial time series properties that are ignored when many, e.g. $T=250$, historical observations are used to estimate one (constant) covariance matrix. In particular, we focus on comparing short-term and long-term estimates of the volatilities to extract some signals, such that if the short term volatility is below the long-term volatility estimate, a smaller penalty λ_i is applied and, consequently, a larger portfolio weight in comparison to the LASSO. Due to space limitations, we refer to [27] for a detailed description of the implementation of the *w8Las* penalty.

While LASSO and *w8Las* are convex penalties, as Figure 1 shows, the remaining four penalties (i.e. SCAD, Zhang, Log and ℓ_q with $0 < q < 1$) are non-convex and allow to deal with the potentially biased LASSO estimates of large absolute coefficients. The economic intuition behind the non-convex penalties is as follows: if the true correlation of assets is high, shorting can reduce the risk, since it accounts for true similarities of the assets instead of being the result of overfitting. Analogously, large portfolio weights tend to be appropriate if the true correlations

Table 1: Penalties

penalty	$\lambda\rho(w_i)$	domains
LASSO =	$\lambda w_i $	all
$w8Las$ =	$\lambda w_i w_i $	all
SCAD =	$\begin{cases} \lambda w_i & w_i \leq \lambda \\ \frac{- w_i ^2 + 2a\lambda w_i - \lambda^2}{2(a-1)} & \lambda < w_i \leq a\lambda \\ \frac{(a+1)\lambda^2}{2} & a\lambda < w_i \end{cases}$	
Zhang =	$\begin{cases} \lambda w_i & w_i < \eta \\ \lambda\eta & \eta \leq w_i \end{cases}$	
L_q =	$\lambda w_i ^q, 0 < q < 1$	all
Log =	$\begin{cases} \lambda \ln(w_i + \phi) \\ -\lambda \ln(\phi) \end{cases}$	all

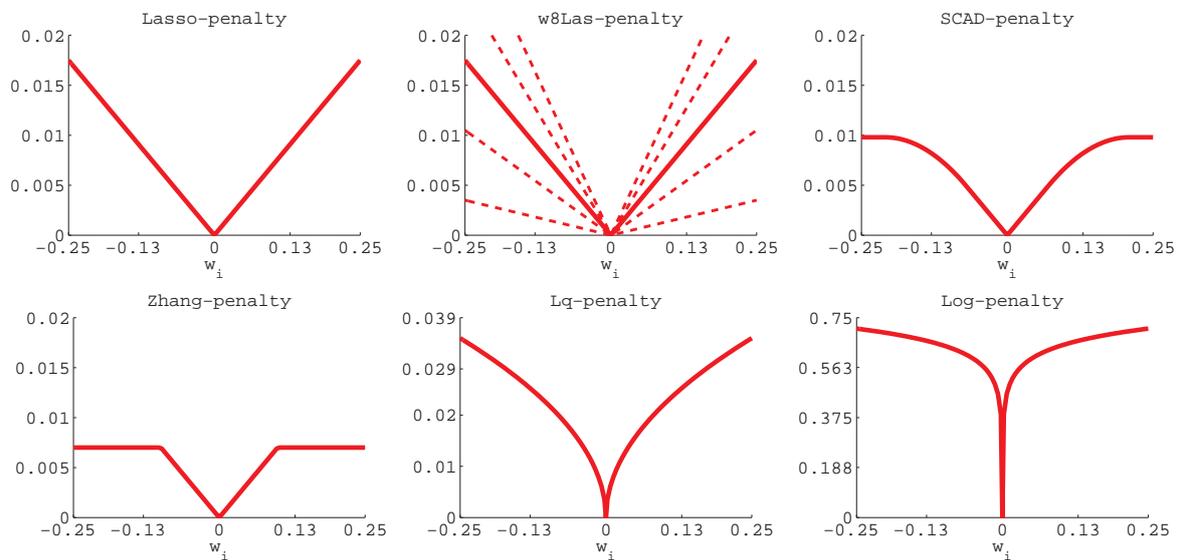


Figure 1: The six (non-)convex penalty functions under consideration in this work.

Table 2: U.S. stock market datasets for the period 23.08.02 to 27.03.08

	dataset	source	obs	K	\bar{r}	$\hat{\sigma}$	\hat{S}	\hat{K}
S&P200:	largest firms (w.r.t. ME)	Datastream	1401	200	6.57	14.79	0.0487	5.32
S&P500:	largest firms (w.r.t. ME)	Datastream	1401	500	6.57	14.77	0.0410	5.13
S&P1036:	largest firms (w.r.t. ME)	Datastream	1401	1036	6.39	14.88	0.0380	4.99

Table 2 reports the datasets under consideration, the source of the data, the number of assets (K), and the number of observations (obs) in each dataset. For the S&P datasets, value weighted indices are computed whose return distributions are characterized by the mean p.a. \bar{r} , the standard deviation p.a. ($\hat{\sigma}$), the skewness (\hat{S}), and the kurtosis (\hat{K}) given in the last four columns. The S&P indices are market value weighted. The weighting schemes are updated daily and applied the following day.

are small. Now, if a correlation structure is “strong enough” to grow absolute portfolio weights – against the counteracting penalty – large enough, it is considered reliable and should therefore enter the portfolio to a greater extend. The main differences between them, as pointed out by Figure 1 is on the intensity on penalizing the different asset weights. The ℓ_q - and the Log-penalty provide a particularly strong incentive to avoid small and presumably dispensable positions in favor of selecting a small subset of presumably indispensable assets. This tendency to construct very sparse and less diversified portfolios coincides with the suggestion of [28] to use the ℓ_q -norm as a diversity measure for portfolios.

3 Empirical Analysis

Data and Experimental Set-Up

We consider daily observations of five different datasets shown in Table 2 that represent the U.S. stock market at different levels of aggregation. Datasets are characterized by a different number of constituents, which include the 200, 500, and 1036 largest individual firms (with respect to the market value on March 27, 2008) of the S&P 1500, which we label as *large* datasets. We refer to [27] for results also on the 48 industry portfolios and the 98 firm portfolios provided by Kenneth French, which could be considered as *small* dataset.

We backtest the out-of-sample performance of the proposed methods with a moving time window procedure, where $\tau = 250$ in-sample observations (corresponding to one year of market data) are used to form a portfolio. The optimized portfolio allocations are then kept unchanged for the subsequent 21 trading days (corresponding to one month of market data) and the out-of-sample returns are recorded. After holding the portfolios unchanged for one month, the time window is moved forward, so that the formerly out-of-sample days become part of the in-sample window and the oldest observations drop out. The updated in-sample window is then used to form a new portfolio, according to which the funds are reallocated. The $T = 1401$ observations allow for the construction of $\Gamma = 54$ portfolios with the corresponding out-of-sample returns.

Table 3 shows the different measures we use to evaluate the out-of-sample performance and the composition of the portfolios, where $F_r^{-1}(p)$ is the value of the inverse cumulated empirical distribution function of the daily out-of-sample returns at point p .

For comparative evaluations, we also implement the following standard benchmarks: (i) the shortsale-unconstrained MVP, denoted MVP_{ssu}, the shortsale-constrained MVP, denoted

Table 3: Portfolio evaluation measures

Measures based on the out-of-sample portfolio returns		
Portfolio variance (s^2) $\frac{1}{T-\tau-1} \sum_{t=\tau+1}^T (r_t - \bar{r})^2$	Sharpe ratio (SR) $\frac{\bar{r}}{\sqrt{s^2}}$	95% Value-at-Risk (VaR) $ F_r^{-1}(0.05) $
Measures based on the portfolio composition		
No. active positions (No. act.) $\frac{1}{\Gamma} \sum_{\gamma=1}^{\Gamma} \{i \mid w_{i,\gamma} \neq 0 \forall i\} $	Shorting (Short) $\frac{1}{\Gamma} \sum_{j=\{i \mid w_{i,\gamma} < 0 \forall i\}} -w_{j,\gamma}$	Turnover (TO) $\frac{1}{\Gamma-1} \sum_{\gamma=2}^{\Gamma} \sum_{i=1}^K w_{i,\gamma} - w_{i,\gamma-1} $

MVPssc, the market value weighted portfolio, denoted mvw, and the equally weighted portfolio, denoted 1oK.

To determine the optimal minimum variance portfolio, we choose to focus on three types of frequently used covariance matrix estimators: (i) the sample estimator, (ii) a three-factor model estimator [10] and (iii) the Ledoit-Wolf estimator [12]. However, we report in the following results related to the three-factor model and refer the reader to [27] for a complete empirical analysis.

Determining the Regularization Parameter

Prior to optimizing problem formulation (1)-(2) for any of the six penalization approaches, a value of the regularization parameter λ must be chosen. Since the optimal values λ^* for the various penalties are unknown, we try for each approach a set of 30 ascending values starting from zero. The largest element in each set is chosen such that the resulting portfolios exhibit only few active positions and a high out-of-sample portfolio variance. In this manner, it is most likely that the intervals spanned by zero and the largest regularization parameters cover λ^* .

Each of the 30 regularization parameters corresponds to one specific (optimized) portfolio, which demands a decision about in which one to eventually invest. This difficult decision is the reason we split the empirical experiments into two setups: (i) we keep track of *all* 30 portfolios that correspond to the *entire spectrum* of 30 regularization parameters over all periods; (ii) we invest in only *one* portfolio by applying ten-fold cross-validation to choose a suited value of λ prior to the investment decision in each period. While procedure (ii) is more realistic from an investment perspective,⁴ procedure (i) provides valuable insights into the potential benefit of regularization and how different values of λ affect the portfolio performance. However, due to space limitations, we refer the reader to [27] for results related to the entire spectrum of regularization parameters and we focus in the next section on results related to the cross-validation procedure.

⁴The cross-validation procedure is as follows: 21 observations are randomly picked from the in-sample data, portfolios are optimized on the remaining 229 observations for all 30 regularization parameters, and the portfolio variance is computed using the 21 picked observations. This is done ten times and the λ is chosen that corresponds to smallest average portfolio variance.

Table 4: Three-factor model covariance matrix (cross-validation experiment)

	MVPssu	MVPssc	mvw	1oK	Lasso	w8Las	Log	ℓ_q	Zhang	SCAD
Panel A: S&P 200 individual firms										
$s^2 \cdot 10^5$	3.007	3.162	6.023	6.524	2.843	2.808	3.017	3.009	2.777	2.942
$VaR \cdot 10^2$	0.885	0.898	1.312	1.348	0.828	0.824	0.893	0.916	0.843	0.881
SR	0.054	0.062	0.018	0.050	0.049	0.050	0.054	0.048	0.049	0.054
<i>No. act.</i>	200.0	54.9	200.0	200.0	82.6	91.1	66.1	65.6	93.9	64.8
<i>Short</i>	0.75	0.00	0.00	0.00	0.26	0.29	0.38	0.38	0.32	0.39
<i>TO</i>	0.57	0.52	0.04	0.00	0.59	0.68	0.96	0.98	0.73	0.90
Panel B: S&P 500 individual firms										
$s^2 \cdot 10^5$	2.883	3.796	6.081	6.799	2.529	2.495	2.617	2.601	2.538	2.643
$VaR \cdot 10^2$	0.923	1.071	1.335	1.385	0.834	0.835	0.794	0.814	0.847	0.842
SR	0.031	0.042	0.018	0.045	0.043	0.043	0.043	0.049	0.042	0.036
<i>No. act.</i>	500.0	278.6	500.0	500.0	131.9	147.6	102.8	108.1	151.6	101.0
<i>Short</i>	0.83	0.00	0.00	0.00	0.20	0.24	0.33	0.35	0.24	0.33
<i>TO</i>	0.61	0.22	0.04	0.00	0.69	0.75	1.11	1.04	0.80	1.09
Panel C: S&P 1036 individual firms										
$s^2 \cdot 10^5$	2.649	4.593	6.254	9.001	2.382	2.379	2.343	2.356	2.485	2.369
$VaR \cdot 10^2$	0.833	1.166	1.352	1.566	0.802	0.792	0.775	0.789	0.819	0.754
SR	0.031	0.031	0.016	0.028	0.054	0.050	0.041	0.045	0.050	0.044
<i>No. act.</i>	1036.0	572.4	1036.0	1036.0	276.7	308.3	179.6	153.8	298.7	161.3
<i>Short</i>	0.84	0.00	0.00	0.00	0.26	0.30	0.33	0.31	0.28	0.31
<i>TO</i>	0.65	0.22	0.04	0.00	0.84	0.89	1.30	1.13	0.87	1.26

Table 4 shows results of the four benchmarks and the six regularization approaches for the three large datasets and the three-factor model covariance matrix.

Empirical Results

Table 4 shows that the cross-validation approach works well for the considered large datasets. The out-of-sample variances of the penalized approaches are always lower than the constrained minimum variance approach (MVPssc) and the equally weighted (mvw) and often also than the unconstrained minimum variance portfolio (MVPssu). This shows that the possibility of having a stronger shrinkage in some periods but not in others is beneficial. The only exception is for the S&P 200 dataset in Panel A, where the Log- and the ℓ_q -regularized portfolios exhibit even higher risks than the MVPssu. However, this fits the picture that the non-convex approaches perform the better the larger the number of constituents compared to the number of observations, which corresponds to a window size of 250. The *w8Las* reaches the smallest variance for both S&P200 and S&P500, while the Log-penalty outperforms for S&P1036. In terms of Sharpe Ratio, the equally weighted portfolio is a tough benchmark, especially for S&P500, where only the ℓ_q -penalty allows to reach a slightly larger value by using just an average subset of 108.1 active components. Lasso, *w8Las* and Zhang penalty reach the largest Sharpe Ratios values for S&P1036, while still investing in an average number of assets much larger than the Log, ℓ_q and SCAD penalties. Clearly, as the non-convex penalties lead often to sparser solutions than other methods, they end up paying a price in terms of turnover rates and identify optimal portfolios with larger shorting amounts, while the extreme risks, as captured by VaR and ES, are still often smaller than the MVPssu, MVPssc and Mvw portfolios.

4 Conclusions

Introducing a penalty in the Markowitz minimum variance framework can allow to determine optimal portfolios that better control for estimation error and have superior out-of-sample performances than the unconstrained approach and the equally weighted benchmark. In particular, we propose a new type of a (convex) penalty whose construction allows for easy processing of all kinds of signals to optimized portfolios, may they be gained from (time series) econometrics, fundamental or technical analysis, or expert knowledge. Moreover, we consider four non-convex penalty functions that have not yet been examined in a portfolio optimization context. It turned out that these approaches perform very well when dealing with very large datasets, where they not only outperformed standard benchmarks but also the (convex) “state-of-the-art” LASSO approach. The success of these approaches stems from their ability to maintain relevant assets in the portfolio with large absolute weights, while only the weights of the remaining assets are shrunk. This allows for a better exploitation of the higher potential to diversify portfolio risk in larger datasets. Further research aims to further develop the underlying signal extraction that could be operationalized in the *w8Las* approach and investigate alternative cross-validation criteria, which likely will allow for a further improvement of the results.

Bibliography

- [1] H. Markowitz, Portfolio selection, *Journal of Finance* 7 (1) (1952) 77–91.
- [2] J. Jobson, R. Korkie, Estimation for Markowitz efficient portfolios, *Journal of the American Statistical Association* 75 (371) (1980) 544–554.
- [3] M. Best, J. Grauer, On the sensitivity of mean-variance-efficient portfolios to changes in asset means: Some analytical and computational results, *The Review of Financial Studies* 4 (2) (1991) 315–342.
- [4] M. Broadie, Computing efficient frontiers using estimated parameters, *Annals of Operations Research* 45 (1) (1993) 21–58.
- [5] M. Britten-Jones, The sampling error in estimates of mean-variance efficient portfolio weights, *Annals of Operations Research* 54 (2) (1999) 655–671.
- [6] V. DeMiguel, J. Garlappi, R. Uppal, Optimal versus naive diversification: How inefficient is the $1/n$ portfolio strategy?, *Review of Financial Studies* 22 (5) (2009) 1915–1953.
- [7] G. Frankfurter, H. Phillips, J. Seagle, Portfolio selection: The effects of uncertain means, variances, and covariances, *Journal of Financial and Quantitative Analysis* 6 (5) (1971) 1251–1262.
- [8] J. Dickinson, The reliability of estimation procedures in portfolio analysis, *Journal of Financial and Quantitative Analysis* 9 (3) (1974) 447–462.
- [9] P. Frost, J. Savarino, For better performance: Constrain portfolio weights, *Journal of Portfolio Management* 15 (1) (1988) 29–34.

- [10] L. Chan, J. Karceski, J. Lakonishok, On portfolio optimization: Forecasting covariances and choosing the risk model, *The Review of Financial Studies* 12 (5) (1999) 937–974.
- [11] R. Jagannathan, T. Ma, Risk reduction in large portfolios: Why imposing the wrong constraints helps, *The Journal of Finance* 58(4) (2003) 1651–1683.
- [12] O. Ledoit, M. Wolf, Improved estimation of the covariance matrix of stock returns with an application to portfolio selection, *Journal of Empirical Finance* 10 (5) (2003) 603–621.
- [13] V. DeMiguel, F. Nogales, Portfolio selection with robust estimation, *Operations Research* 57 (3) (2009) 560–577.
- [14] J. Fan, J. Zhang, K. Yu, Vast portfolio selection with gross exposure constraints, *Journal of the American Statistical Association* 107 (498) (2012) 592–606.
- [15] M. Fernandes, G. Rocha, T. Souza, Regularized minimum-variance portfolios using asset group information, Available from http://webpace.qmul.ac.uk/tsouza/index_arquivos/Page497.htm (2012) 1–28.
- [16] P. Behr, A. Guettler, F. Truebenbach, Using industry momentum to improve portfolio performance, *Journal of Banking and Finance* 36 (5) (2012) 1414–1423.
- [17] P. Jorion, Bayes-Stein estimation for portfolio analysis, *Journal of Financial and Quantitative Analysis* 21 (3) (1986) 279–292.
- [18] J. Brodie, I. Daubechies, C. DeMol, D. Giannone, D. Loris, Sparse and stable Markowitz portfolios, *Proceedings of the National Academy of Science USA* 106 (30) (2009) 12267–12272.
- [19] V. DeMiguel, L. Garlappi, J. Nogales, R. Uppal, A generalized approach to portfolio optimization: Improving performance by constraining portfolio norms, *Management Science* 55 (5) (2009) 798–812.
- [20] R. Tibshirani, Regression shrinkage and selection via the Lasso, *Royal Statistical Society* 58 (1) (1996) 267–288.
- [21] Y.-M. Yen, A note on sparse minimum variance portfolios and coordinate-wise descent algorithms, Available from http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1604093 (2010) 1–27.
- [22] M. Carrasco, N. Noumon, Optimal portfolio selection using regularization, Working Paper University of Montreal; available from <http://www.unc.edu/maguilar/metrics/carrasco.pdf>.
- [23] Y.-M. Yen, T.-J. Yen, Solving norm constrained portfolio optimizations via coordinate-wise descent algorithms, Available from http://personal.lse.ac.uk/yen/sp_090111.pdf (2011) 1–41.
- [24] J. Fan, R. Li, Variable selection via nonconcave penalized likelihood and its oracle properties, *Journal of the American Statistical Association* 96 (456) (2001) 1348–1360.

- [25] G. Gasso, A. Rakotomamonjy, S. Canu, Recovering sparse signals with a certain family of nonconvex penalties and DC programming, *IEEE Transactions on Signal Processing* 57 (12) (2009) 4686–4698.
- [26] H. Zou, The adaptive lasso and its oracle properties, *Journal of the American Statistical Association* 101 (476) (2006) 1418–1429.
- [27] B. Fastrich, S. Paterlini, P. Winker, Constructing optimal sparse portfolios using regularization methods, Working paper; available from http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2169062.
- [28] R. Fernholz, R. Garvy, J. Hannon, Diversity weighted indexing, *Journal of Portfolio Management* 24 (2) (1998) 74–82.

Combining information at different frequencies in multivariate volatility prediction

Alessandra Amendola, *University of Salerno*, alamendola@unisa.it
Giuseppe Storti, *University of Salerno*, storti@unisa.it

Abstract. In the last two decades the literature has been focusing on the development of dynamic models for predicting conditional covariance matrices from daily returns and, more recently, on the generation of co-volatility forecasts by means of dynamic models directly fitted to realized measures. Despite the number of contributions on this topic some open issues still arise. First, are dynamic models based on realized measures able to produce more accurate forecasts than standard MGARCH models based on daily returns? Second, which is the impact of the choice of the volatility proxies on forecasting accuracy? Is it possible to improve the forecasts accuracy by combining forecasts from MGARCH and models for realized measures? Finally, can combining information observed at different frequencies help to improve over the performance of single models? In order to gain some insight about these research questions, in this paper we perform an extensive forecast comparison of different multivariate volatility models considering both MGARCH models and dynamic models for realized covariance measures. Furthermore, we investigate the possibility of increasing predictive accuracy by combining forecasts generated from these two classes of models, using different combination schemes and mixing forecasts based on information sets observed at different frequencies.

Keywords. Forecast combination, multivariate GARCH, realized covariance, model confidence set.

1 Introduction

The literature on multivariate volatility prediction from vector time series of daily returns is relatively recent, originating at the end of the 80s with the paper by Bollerslev et al. (1988) proposing the VEC model. The general version of their model is very flexible but, even for moderately large dimensions, it is characterized by a large number of parameters. The subsequent research on multivariate generalizations of the standard GARCH model (MGARCH) has

focused on two main issues. First, the need for more parsimonious specifications allowing the analysis of large dimensional datasets without paying a too high price in terms of model's flexibility. Second, substantial efforts have been dedicated to the derivation of parameter constraints inducing well defined, positive definite and stable, sequences of estimated covariance matrices. A comprehensive review of the literature on MGARCH models can be found in Bauwens et al. (2006).

More recently, the increasing availability of high-frequency data on financial transactions has stimulated a new stream of research proposing to use dynamic models, directly fitted to time series of realized covariance matrices, in order to predict future conditional variances and covariances. Bauwens et al. (2012) provide a review of these contributions. The predictive performances of these two sets of approaches have been recently empirically compared by Boudt et al. (2014) considering an application to Value at Risk (VaR) estimation. Their results provide evidence in favour of the hypothesis that dynamic models for realized covariance measures, henceforth RC models, can be more accurate than standard MGARCH models in predicting conditional variance and covariance matrices.

One of the main drawbacks of the approach based on RC models is related to the choice of the discretization frequency used for computing the realized covariance estimator and, more generally, to the choice of the realized estimator used for approximating the volatility matrix. This issue has been recently addressed in the paper by Varneskov and Voev (2013) who also find that substantial accuracy gains can be obtained moving from a plain approach based on simple daily returns to the use of high-frequency information.

Our aim in this paper is, first, to compare the predictive performances of MGARCH models and RC models estimated at different frequencies. Second, and more important, we are interested in assessing the profitability, in terms of forecasting accuracy, of a forecast combination scheme merging forecasts from models estimated at different frequencies. Our approach extends the algorithm discussed by De Pooter et al. (2010) to the prediction of conditional covariance matrices. We compute the combined predictor averaging the forecasts generated by the models included in the time-varying set of optimal models which is identified applying the Model Confidence Set (MCS) approach of Hansen and Lunde (2011) over a rolling window. The results show that the combined predictor can improve over each of the single models separately considered. The paper is structured as follows. Section 2 describes the MGARCH and RC models used for our analysis while the forecast combination strategy is illustrated in Section 3. The results of an empirical application to a portfolio of U.S. stocks are presented in Section 4 while section 5 concludes.

2 Candidate Models

Forecast combinations require the take up of two important decisions related to which forecasts should be included in the analysis and to the approach that should be adopted for determining the weights assigned to the included models. The first task is, therefore, related to what is often called the design of the model universe. The models that have been considered in this paper can be classified into two groups. The first group includes MGARCH models that do not exploit intra-daily information and are fitted to time series of daily returns. Namely, we consider two different variants of the Dynamic Conditional Correlation (DCC) model of Engle (2002) and a scalar version of the BEKK model proposed by Engle and Kroner (1995), the RiskMetrics (RM)

model (J.P.Morgan, 1996) and a Moving Covariance (MC) estimator. Differently, the models included into the second group are directly fitted to time series of realized covariance matrices. In particular, these include the Conditionally Autoregressive Wishart, CAW, model proposed by Golosnoy et al. (2012) and *realized* versions of the RM and MC estimators.

MGARCH models

The DCC model, in the original formulation of Engle (2002), is defined as:

$$\begin{aligned} H_t &= D_t R_t D_t \\ D_t &= \text{diag}(h_t) \quad h_{i,t} = \sqrt{H_{ii,t}} \\ H_{ii,t} &= a_{0,i} + a_{1,i} r_{i,t-1}^2 + b_{1,i} H_{ii,t-1} \\ R_t &= (\text{diag}(Q_t))^{-1/2} Q_t (\text{diag}(Q_t))^{-1/2} \\ Q_t &= (1 - \alpha - \beta) \bar{Q} + \alpha (\epsilon_{t-1} \epsilon'_{t-1}) + \beta Q_{t-1} \end{aligned}$$

where $\epsilon_t = D_t^{-1} r_t$ is the $(n \times 1)$ vector process of standardized residuals and $\bar{Q} = \hat{S} = (1/T) \sum_{t=1}^T \epsilon_t \epsilon'_t$ is the sample covariance matrix of ϵ_t . Aielli (2013) points out that, for consistent targeting, \bar{Q} should be (asymptotically) equal to $E(Q_t)$ which is not the case in the Engle's formulation. This motivates his corrected DCC (cDCC) model that differs from the basic DCC in the specification of the dynamic updating equation for Q_t that is defined as

$$Q_t = (1 - \alpha - \beta) \Psi + \alpha (\eta_{t-1} \eta'_{t-1}) + \beta Q_{t-1}$$

where

$$\eta_t = \text{diag}(Q_t)^{1/2} \epsilon_t$$

and

$$\Psi = E(\eta_t \eta'_t).$$

One point to note is that Ψ depends on the correlation parameters $(\alpha, \beta)'$. So, at the estimation stage, the log-likelihood function must be simultaneously maximized with respect to $(\alpha, \beta)'$ and Ψ . This makes the estimation unfeasible for vast dimensional models. To deal with applications to large datasets, Aielli (2013) proposes to use a generalized profile Quasi Log-Likelihood estimator. Simulation results show that parameter estimates for both the DCC and cDCC models can be severely biased in large dimensional systems. To reduce this bias, Engle et al. (2008) propose to use a Gaussian Composite Quasi Maximum Likelihood (CQML) estimator. Their simulation results show that the CQML estimator outperforms the standard Gaussian QML in large dimensional systems.

An alternative approach to bias reduction in the estimation of DCC models in large dimensions is proposed by Hafner and Reznikova (2012) who derive an alternative formulation of the standard DCC model in which the targeting matrix \bar{Q} is obtained by shrinkage estimators

$$\bar{Q} = \delta M + (1 - \delta) \hat{S}$$

where \hat{S} is the sample covariance matrix of standardized residuals ϵ_t , M is the targeting matrix and δ denotes the shrinkage intensity. M can be chosen in different ways. If we set $M = I_n$, an identity matrix of order n , and $\delta = \delta_I \mu$, where $\mu = \text{trace}(\hat{S})/n$ is the Frobenius inner product,

the resulting estimator is called a *shrinkage to identity* estimator. Alternatively, we will get a *shrinkage to equicorrelation* estimator if we set $\delta = \delta_E$ and $M = E$, a $(n \times n)$ matrix such that $E_{ij} = \bar{\rho}\sqrt{\hat{S}_{ii}\hat{S}_{jj}}$ with

$$\bar{\rho} = \frac{1}{2n(n-1)} \sum_{i=1}^{n-1} \sum_{j=i+1}^n \rho_{ij}$$

where $\rho_{ij} = \hat{S}_{ij}/\sqrt{\hat{S}_{ii}\hat{S}_{jj}}$. Finally, Hafner and Reznikova (2012) suggest a single-index factor models in order to estimate the shrinkage targets. From a Monte Carlo simulation study it arises that i) for problems of small to medium dimension, the shrinkage to equicorrelation estimator outperforms the QML and CQML estimators of the standard DCC model ii) for large dimensional problems, the most accurate estimator is that based on Gaussian CQML.

Finally, we consider the Dynamic Equicorrelation (DECO) model proposed by Engle and Kelly (2008) as an alternative to the standard DCC model for the estimation of the conditional covariance matrices of large dimensional portfolios. The DECO model differs from the standard DCC in the specification of the conditional correlation matrix R_t which is defined as

$$R_t = (1 - \rho_t)I_n + \rho_t J_n$$

where J_n is a $(n \times n)$ matrix of ones and ρ_t is the dynamic equicorrelation coefficient given by the average of the off-diagonal elements of the DCC conditional correlation matrix

$$\bar{\rho} = \frac{1}{2n(n-1)} \sum_{i=1}^{n-1} \sum_{j=i+1}^n \frac{Q_{t,ij}}{\sqrt{Q_{t,ii}Q_{t,jj}}}$$

Estimates of correlation parameters can be easily obtained by maximizing a Gaussian QML function.

As a simple alternative to DCC estimators, we consider a scalar BEKK model. In the BEKK model proposed by Engle and Kroner (1995), assuming homogeneous dynamics, the dynamic equation for the conditional covariance matrix is given by:

$$H_t = (1 - \alpha^2 - \beta^2)\bar{H} + \alpha^2 r_{t-1} r'_{t-1} + \beta^2 H_{t-1}$$

where $\bar{H} = (1/T) \sum_{t=1}^T r_t r'_t$. Estimates of α and β can be obtained by Gaussian QML. As for the DCC model, Engle et al. (2008), however, show that these estimates are severely biased in large dimensional models. By Monte Carlo simulations they also show that this bias does not affect CQML estimators of the parameters of BEKK models.

The RM estimator can be derived as a special case of an integrated scalar BEKK model in which $\beta^2 = 1 - \alpha^2 = 0.94$. Finally, the h -days MC estimator is a simple tool used by practitioners for obtaining a quick and preliminary estimated of the conditional covariance matrix of returns and can be defined as:

$$H_t = \frac{1}{h} \sum_{i=1}^h r_{t-i} r'_{t-i} \quad \text{for } i > m.$$

Dynamic RC models

Let us denote by Σ_t , $t = 1, \dots, T$ a time series of realized covariance matrices. CAW models (Golosnoy et al., 2012) are based on the assumption that, conditional on past information I_{t-1} ,

the matrix Σ_t follows a n -dimensional central Wishart distribution:

$$\Sigma_t | I_{t-1} \sim W_n(\nu, H_t/\nu), \quad (1)$$

where $\nu > n - 1$ is the degrees of freedom parameter, H_t/ν is a $n \times n$ symmetric positive definite scale matrix. It follows that

$$E(\Sigma_t | I_{t-1}) = H_t$$

where H_t can be interpreted as the latent conditional covariance matrix of returns. The dynamic updating equation for H_t is specified using a BEKK formulation with covariance targeting:

$$H_t = (1 - \alpha^2 - \beta^2)\bar{\Sigma} + \alpha^2\Sigma_{t-1} + \beta^2H_{t-1} \quad (2)$$

where $\alpha^2 + \beta^2 < 1$ and $\bar{\Sigma} = 1/T \sum_{t=1}^T \Sigma_t$. QML estimates of the parameters in (2) can be obtained by maximization of a Wishart QL function. Furthermore, Bauwens and Storti (2013) have derived an alternative CQML estimator that allows for computationally efficient estimation of the model parameters in large dimensional problems.

In addition, we consider the Realized RiskMetrics (RRM) estimator

$$H_t = 0.06\Sigma_{t-1} + 0.94H_{t-1} \quad (3)$$

and a Realized Moving Covariance (RMC) estimator given by:

$$H_t = \frac{1}{h} \sum_{i=1}^h \Sigma_{t-i}.$$

3 The forecast combination approach

Assume that r_t , $t = 1, \dots, T$ is a time series of returns generated by the model

$$r_t = S_t z_t \quad t=1, \dots, T$$

where $z_t \stackrel{iid}{\sim} (0, I_n)$ and S_t is any $(n \times n)$ positive definite (p.d.) matrix such that $S_t = S(I_{t-1}, \theta)$. From the above specification it follows that $H_t = S_t S_t'$ is the conditional covariance matrix of returns given past information I_{t-1} . The shape of the dynamic process generating S_t , which is the shape of $S(\cdot)$, is unknown.

Also assume that k candidate models for the prediction of H_t are available and denote by $H_t^{(j)}$ the forecasts, symmetric and p.d., of the covariance matrix of r_t , conditional on I_{t-1} , generated by the j -th candidate model. In general a *combined predictor* based on the available k candidate models is defined as

$$\tilde{H}_t = C(H_t^{(1)}, \dots, H_t^{(k)}; w_t)$$

where $C(\cdot)$ is an appropriately chosen *combination function* and w_t is a vector of combination parameters. Different combination functions $C(\cdot)$ can in principle be used and there is no *a priori* valid procedure for selecting the optimal function. Among all the possible choices of $C(\cdot)$, the most common is the *linear* combination function

$$\tilde{H}_t = w_{t,1}H_t^{(1)} + \dots + w_{t,k}H_t^{(k)} \quad w_{t,j} \geq 0 \quad (4)$$

MMM	ABT	AA	AAPL
ALL	MO	AXP	AIG
AMGN	T	BAC	BK

Table 1: Symbols identifying the 12 NYSE stocks included in the analyzed portfolio.

where w_t coincides with the vector of combination weights. The assumption of non-negative weights is required in order to guarantee the positive definiteness of \tilde{H}_t .

The approach we pursue in this paper is based on the use of a linear combination function where the weights are determined by the MCS approach. In practice, the combined predictor is defined as a simple average of the candidate models included in the MCS while a weight equal to 0 is assigned to all the other models excluded from the MCS.

Namely, our approach is based on a fixed-rolling window forecasting scheme. Let us denote by T_{in} the in-sample size for estimating the model parameters, our forecasting procedure is based on the following steps

1. Estimate all the candidate models over the window including observations from 1 to T_{in}
2. Conditional on the estimated parameters, generate static 1-step ahead forecasts of the conditional covariance matrix for the following m observations
3. Re-estimate the candidate models over the window including observations from $m + 1$ to $T_{in} + m$
4. Iterate 2 and 3 until the end of the series.

At each re-estimation we then compute the MCS including the best performing models according to some adequately chosen loss function. The combined predictor \tilde{H}_t is then computed as the equally weighted average of the models included in the MCS ⁵. It follows that our forecasting strategy allows the structure of the combined predictor to vary over time.

4 Empirical results

In this section we present the results of an application to a portfolio of 12 NYSE stocks (table 1).

Our raw data are composed of price quotations observed every minute, from 9.30 a.m. to 4.00 p.m., from May 12, 1997 to July 18, 2008 ⁶ for a total of 2780 observations. The raw returns have then been aggregated over intervals of 5, 10, 15 and 30 minutes, respectively, in order to compute the associated time series of daily realized covariance matrices. Our choice of using open-to-close returns follows the approach of Andersen et al. (2010) who argue that the overnight return can be interpreted as a deterministically occurring jump. Hence the open-to-close return can be considered as the daily return adjusted for the overnight jump.

⁵In order to initialize the computation of the MCS, in the first estimation rounds, in-sample estimates of the conditional covariance matrices are used. These are then gradually replaced by out-of-sample forecasts.

⁶The data are available online at www.tickdata.com.

Our aim is i) comparing the performances of the members of our set of candidate models in generating one-step-ahead predictions of the conditional covariance matrix of daily returns ii) evaluating the ability of forecast combinations to improve the performance of the single candidate models.

The cDCC and BEKK models have been estimated by means of a Gaussian QML estimator as well as a CQML estimator based on the whole set of feasible bivariate sub-systems. Similarly, the CAW model has also been estimated by both QML and CQML estimators. In this way our combined predictor allows to account for model as well as estimation uncertainty. The length of the moving window for the calculation of the MC and RMC estimator has been set equal to $m = 100$.

The RC based predictors have been computed for each of the above considered intradaily sampling frequencies. This gives an overall number of 24 candidate predictors to be used for forecast combination. The accuracy of each of these models in predicting the conditional covariance matrix is assessed using a loss function based on the Frobenius norm

$$L_F^{(\delta)} = Tr[(\Sigma_t^{(\delta)} - H_t^{(j)})'(\Sigma_t^{(\delta)} - H_t^{(j)})]$$

where $\Sigma^{(\delta)}$ is the realized covariance matrix computed from δ -minutes intradaily returns, with $\delta = 5, 10, 15, 30$ minutes. So the rolling MCS is performed four times, one for each value of δ , yielding four different combined predictors $H_t^{(\delta)}$.

Figure 1 reports, for each re-estimation step, the size of the MCS comparing it with \bar{h}_j

$$\bar{h}_j = \frac{1}{T_{in}} \sum_{t=T_{in}(j-1)+1}^{T_{in}j} \left(\frac{1}{12} \sum_{i=1}^{12} h_{t,i}^2 \right)$$

which is the average volatility of the assets included in our portfolio over the j -th estimation window. The plot shows that the size of the MCS is related to the average volatility level. This is particularly evident in the last part of the sample, approximately corresponding to the long low volatility period immediately preceding the financial crisis started in summer 2008. The composition of the MCS, under the four different volatility proxies considered, is summarized by the plots in figure 2. The analysis of these plots reveals that the composition of the MCS is not particularly sensitive to the intra-daily sampling frequency for $\delta > 5$ minutes.

Finally, in order to compare the predictive accuracy of the candidate models we have re-computed the MCS on the whole out-of-sample forecasting period (from observation 501 to observation 2780). The set of candidate predictors is now composed of 28 predictors, obtained from the merging of the initial set of 24 candidate models with the 4 combined predictors computed by the rolling-window MCS. In all cases it turns out that the estimated whole-period MCS includes only one predictor given by the combined predictor $H_t^{(30)}$. This result suggests that combining predictions generated from different models, possibly using information at different frequencies, can improve over the forecasting performance of single, misspecified, forecasting models.

5 Concluding remarks

In this paper we have compared the predictive accuracy of MGARCH and RC models estimated at different frequencies. Furthermore, we have investigated the possibility of improving the forecast accuracy of single misspecified models by using forecast combination techniques.

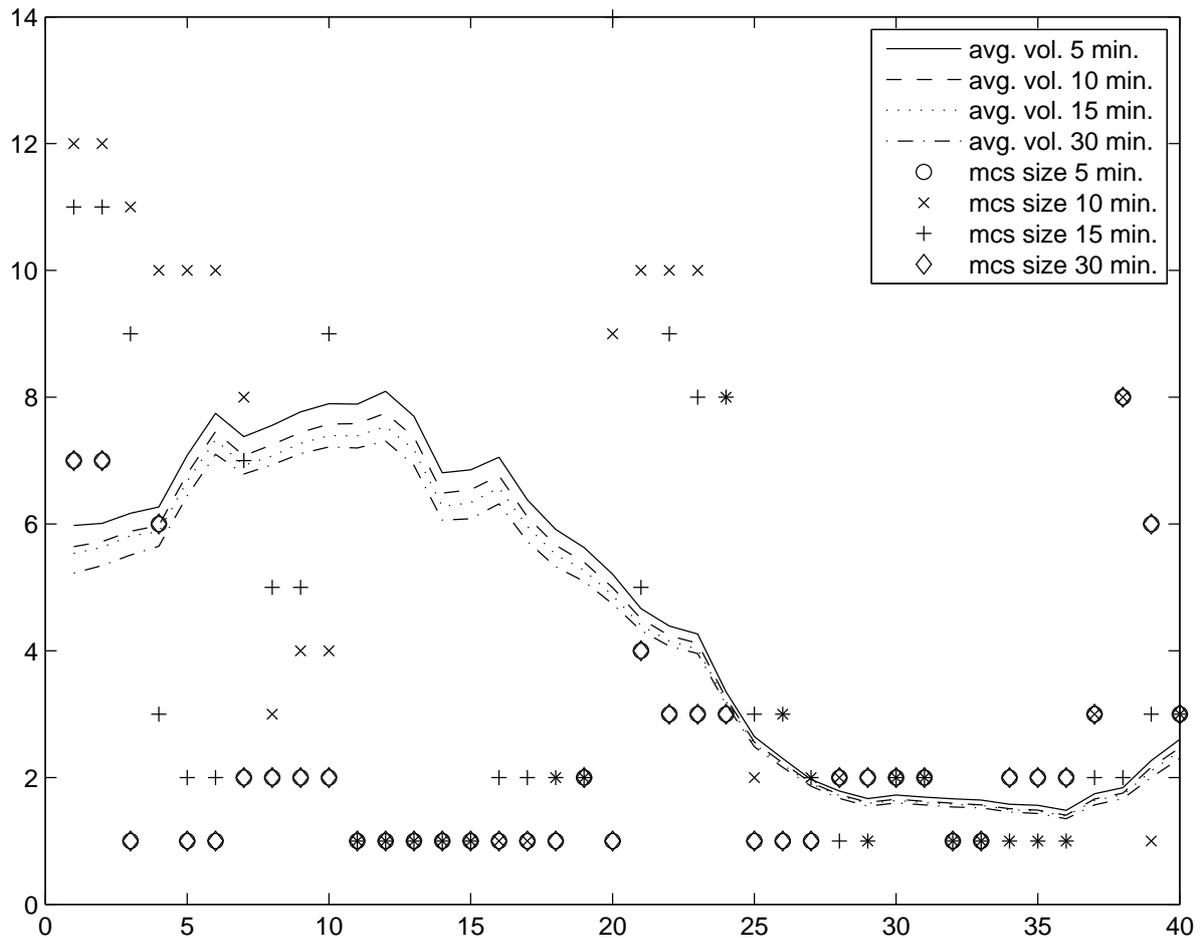


Figure 1: Size of the MCS computed over 40 re-estimations versus the average volatility level over the same period (\bar{h}).

The empirical results of our analysis suggest that it is not possible to identify a clearly winning approach between MGARCH and RC models. This is evident looking at the composition of the MCS which is not stable over time but is characterized by the alternance of models from the two different groups. These results appear to be quite robust to the choice of the sampling frequency of the RC matrix used for assessing the forecast accuracy.

The main finding achieved within the paper is that combining forecasts from models estimated at different frequencies can allow to improve over the predictive ability of single models.

Acknowledgement

The authors gratefully acknowledge financial support from MIUR within the PRIN project 2010-2011 (prot. 2010J3LZEN): *Forecasting economic and financial time series: understanding the complexity and modelling structural change*.

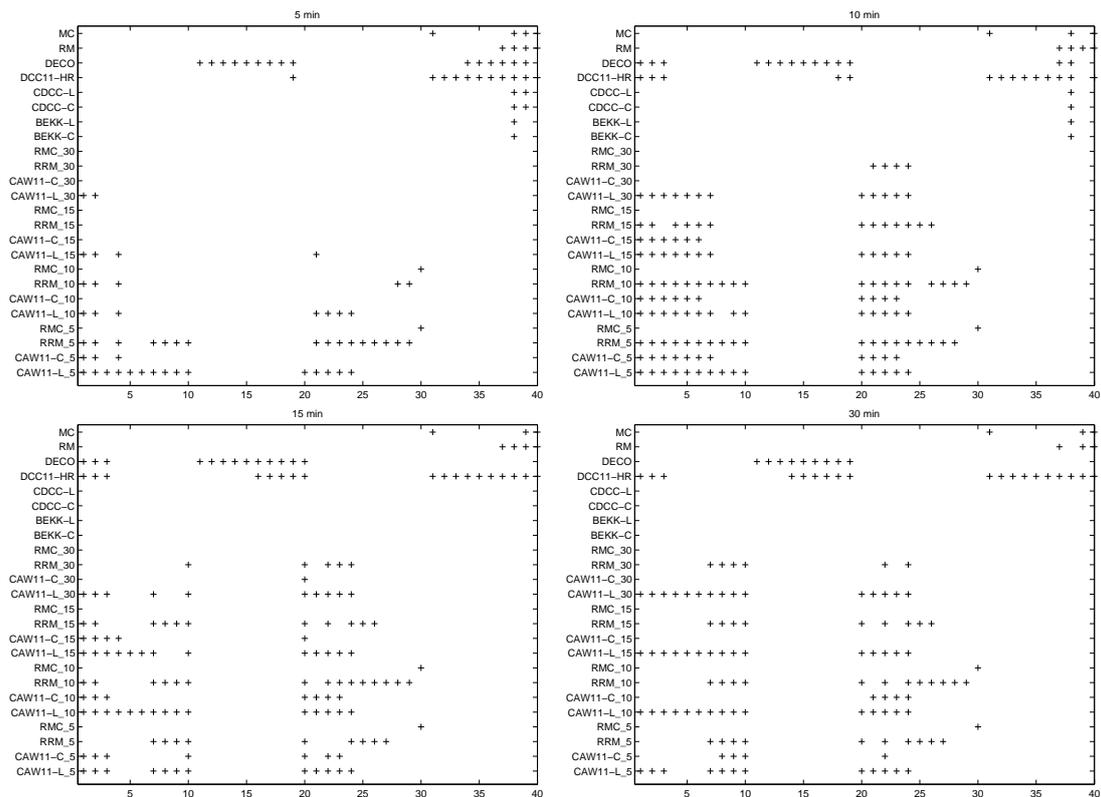


Figure 2: Composition of the MCS computed over 40 re-estimations under 4 different RC measures. From left to right and from top to bottom: $\Sigma_t^{(5)}$, $\Sigma_t^{(10)}$, $\Sigma_t^{(15)}$, $\Sigma_t^{(30)}$. Circles (o) indicate that a given model is included in the MCS at the corresponding estimation round.

Bibliography

- [1] Aielli, G. P. (2013) Dynamic Conditional Correlation: On Properties and Estimation, *Journal of Business & Economic Statistics*, 31(3), 282-299.
- [2] Andersen, T. G., Bollerslev, T., Frederiksen, P., Nielsen, O. (2010) Continuous-time models, realized volatilities, and testable distributional implications for daily stock returns, *Journal of Applied Econometrics*, vol. 25(2), 233–261.
- [3] Bauwens, L., Laurent, S. and Rombouts, J.V.K. (2006) Multivariate GARCH models: a survey, *Journal of Applied Econometrics*, 21, 79-109.
- [4] Bauwens, L., Storti, G., (2013) *Computationally efficient inference procedures for vast dimensional realized covariance models*, in *Complex Models and Computational Methods in Statistics*, 37-49, Springer, Berlin.
- [5] Bauwens, L., Storti, G., Violante, F. (2012) *Dynamic conditional correlation models for realized covariance matrices*, CORE Discussion Paper 2012/060, Universit Catholique de Louvain, Center for Operations Research and Econometrics (CORE).

- [6] Bollerslev, T., Engle, R.F. and Wooldridge, J.M. (1988). A Capital Asset Pricing Model with Time-Varying Covariances, *Journal of Political Economy*, 96, 116-131.
- [7] Boudt, K., Laurent, S., Lunde, A. and Quaedvlieg, R. (2014) *Positive Semidefinite Integrated Covariance Estimation, Factorizations and Asynchronicity*, CREATES Research Papers 2014-05, School of Economics and Management, University of Aarhus.
- [8] De Pooter, M., Ravazzolo, F. and van Dijk, D. (2010) *Term Structure Forecasting Using Macro Factors And Forecast Combination*. Board of Governors of the Federal Reserve System, Discussion Paper 993.
- [9] Engle, R.F. (2002) Dynamic Conditional Correlation - A Simple Class of Multivariate GARCH Models, *Journal of Business and Economic Statistics*, 20, 339-350.
- [10] Engle, R.F., Kelly, B. (2008) Dynamic Equicorrelation, *Journal of Business and Economic Statistics*, 30(2), 212-228.
- [11] Engle, R.F., Kroner, F. (1995) Multivariate Simultaneous Generalized ARCH, *Econometric Theory*, 11, 122-150.
- [12] Engle, R.F., Shephard, N., Sheppard, K. (2008) *Fitting vast dimensional time-varying covariance models*, Economics Series Working Papers 403, University of Oxford, Department of Economics.
- [13] Golosnoy V., Gribisch, B., Liesenfeld, R. (2012) The conditional autoregressive Wishart model for multivariate stock market volatility, *Journal of Econometrics*, 167 (1), 211-223.
- [14] Hafner, C., Reznikova, O. (2012) On the estimation of dynamic conditional correlation models, *Computational Statistics & Data Analysis*, 56(11), pages 3533-3545.
- [15] Hansen, P. R., Lunde, A. and Nason, J. M. (2011) The Model Confidence Set, *Econometrica*, 79, 453-497.
- [16] J.P. Morgan (1996) *Riskmetrics Technical Document*. 4th ed. J.P.Morgan, New York.
- [17] Varneskov, R. and Voev, V. (2013) The role of realized ex-post covariance measures and dynamic model choice on the quality of covariance forecasts, *Journal of Empirical Finance*, 20, 83-95.

Penalty-free sparse PCA

Kohei Adachi, *Osaka University*, adachi@hus.osaka-u.ac.jp

Nickolay Trendafilov, *Open University*, Nickolay.Trendafilov@open.ac.uk

Abstract. A drawback of the sparse principal component analysis (PCA) procedures using penalty functions is that the number of zeros in the matrix of component loadings as a whole cannot be specified in advance. We thus propose a new sparse PCA procedure in which the least squares PCA loss function is minimized subject to a pre-specified number of zeros in the loading matrix. The procedure is called unpenalized sparse matrix PCA (USMPCA), as it does not use a penalty function and obtains component loadings matrix-wise, i.e., simultaneously rather than sequentially. The key point of USMPCA is to use the fact that the PCA loss function can be decomposed into sum of two terms, one of them irrelevant to loadings, and another one being a function easily minimized under the considered cardinality constraint. This decomposition makes it possible to construct an efficient alternate least squares algorithm for USMPCA. Another useful feature is that the PC score matrix is column-orthonormal, which helps to define naturally the percentage of explained variance by the sparse PCs. USMPCA is illustrated with real data examples.

Keywords. Sparse component loadings, loss function decomposition, constrained matrix complexity.

1 Introduction

For an n -observations \times p -variables column-centered data matrix X , principal component analysis (PCA) can be formulated as minimizing

$$f(F, A) = \|X - FA^\top\|^2 \tag{1}$$

over an $n \times m$ PC score matrix F and a $p \times m$ component loading matrix A , with $\|\cdot\|^2$ indicating the squared Frobenius norm and the number of components $m \leq \min(n, p)$. The resulting solution is interpreted by noting the loadings in A which quantify the relationships between the p variables and m components. It is desired for A to be sparse, i.e., to have a number of zero elements, since a sparse matrix is easily interpreted by focusing only on the variables and components linked with nonzero elements. However, such sparse A cannot be obtained by the standard PCA. For this reason, a number of modified PCA procedures have been proposed in the last decade, which produce sparse solutions [8]. Such procedures are called sparse PCA.

Almost all existing sparse PCA procedures are using penalized approaches: they are formulated by combining a PCA objective function with penalty functions that penalize A to have nonzero elements. Such examples are SCoTLASS [3], SPCA [10], and sPCA-rSVD [11], where the relative importance of penalty functions is controlled by tuning parameters. That is, they control the number of nonzero elements, which is called cardinality. Though a number of other penalized procedures have been developed for improving the preceding ones [4, 14, 8], they are formulated by the same format.

A common drawback of the penalized sparse PCA is that the appropriate value of the tuning parameter which corresponds to the desired cardinality is not obvious. Thus, the penalized sparse PCA is not convenient for users who wish to have a loading matrix with a specified number of zero elements. The procedures studied in [1] and [5] avoid such a difficulty. Their authors presented efficient heuristic algorithms called "greedy" search to find component loadings sequentially with direct cardinality constraint. In this paper, we also propose a directly constrained cardinality procedure without using a penalty function. However, our proposed procedure differs from the "greedy" search approaches in that all component are extracted simultaneously (not sequentially), i.e., F and A are obtained matrix-wise (not column-wise). We, thus, refer to our proposed procedure as unpenalized sparse matrix PCA (USMPCA). Moreover, the resulting PC scores are uncorrelated, which helps to define naturally the percentage of explained variance as described in Section 4.

2 Unpenalized Sparse Matrix PCA

In USMPCA, the PCA loss function (1) is minimized subject to the column-orthonormality condition for $n^{-1/2}F$ and the constraint on $\text{card}(A)$ which denotes the cardinality of A . That is, USMPCA is formulated as

$$\min_{F,A} f(F, A) = \|X - FA^\top\|^2, \text{ subject to } \frac{1}{n}F^\top F = I_m \text{ and } \text{card}(A) = c \quad (2)$$

with I_m denoting the $m \times m$ identity matrix and c being a specified integer.

The key point of USMPCA is to use the fact that the orthonormality $\frac{1}{n}F^\top F = I_m$ allows the loss function (1) to be decomposed as

$$\|X - FA^\top\|^2 = \|X - FB^\top + FB^\top - FA^\top\|^2 = \|X - FB^\top\|^2 + n\|B - A\|^2, \quad (3)$$

with B being the cross-product matrix of p -variables $\times m$ -components:

$$B = \frac{1}{n}X^\top F. \quad (4)$$

The decomposition (3), which is derived from $(X - FB^\top)(FB^\top - FA^\top)$ being the zero matrix, shows that a simple function $\|B - A\|^2$ is only relevant to A , which allows us to easily attain the cardinality constrained minimization of (1) as found in the next section.

3 Algorithm

The USMPCA problem (2) can be solved by alternately performing the two steps:

A-step minimizing (1) over A subject to $\text{card}(A) = c$ with F being kept fixed;

F-step minimizing (1) over F subject to $\frac{1}{n}F^\top F = I_m$ with A kept fixed.

First, let us consider the A-step, which is equivalent to minimizing $g(A) = \|B - A\|^2$ under $\text{card}(A) = c$, since of (3). Using $A = (a_{ij})$ and $B = (b_{ij})$, we can rewrite $g(A)$ as

$$g(A) = \|B - A\|^2 = \sum_{(i,j) \in O} b_{ij}^2 + \sum_{(i,j) \in O^\perp} (a_{ij} - b_{ij})^2 \geq \sum_{(i,j) \in O} b_{ij}^2. \tag{5}$$

Here, O denotes the set of the $q = pm - c$ indexes (i, j) 's indicating the locations of the loadings a_{ij} to be zero, while the complement set O^\perp contains the c (i, j) 's of nonzero a_{ij} . The inequality in (5) shows that $g(A)$ attains its lower limit $\sum_{(i,j) \in O} b_{ij}^2$ when the non-zero loadings a_{ij} with $(i, j) \in O^\perp$ are taken equal to the corresponding b_{ij} . Moreover, the limit $\sum_{(i,j) \in O} b_{ij}^2$ is minimal, when O contains the indexes for the q smallest b_{ij}^2 among all squared elements of B . Thus, $g(A)$ is minimized for $A = (a_{ij})$ being

$$a_{ij} = \begin{cases} 0 & \text{if } b_{ij}^2 \leq b_{[q]}^2 \\ b_{ij} & \text{otherwise} \end{cases}, \tag{6}$$

with $b_{[q]}^2$ the q th smallest value among all b_{ij}^2 .

Next, let us consider the minimization in F-step. It is attained for

$$F = \sqrt{n}KL^\top = XAL\Lambda^{-1}L^\top, \tag{7}$$

where K and L are given by the singular value decomposition (SVD) of XA defined as

$$\frac{1}{\sqrt{n}}XA = K\Lambda L^\top \tag{8}$$

with $K^\top K = L^\top L = I_p$ and Λ a diagonal matrix. However, it is shown in the next paragraph that the update of F by (7) can be skipped.

Using $\frac{1}{n}F^\top F = I_m$ and (2), the loss function (1) can be expanded as

$$f(F, A) = \text{tr}X^\top X + \text{tr}AF^\top FA^\top - 2\text{tr}X^\top FA = n\text{tr}S + n\text{tr}A^\top A - 2n\text{tr}B^\top A, \tag{9}$$

with $S = \frac{1}{n}X^\top X$. Noting that (9) is a function of B and the use of (7) in (2) leads to

$$B = \frac{1}{n}X^\top XAL\Lambda^{-1}L^\top = SAL\Lambda^{-1}L^\top, \tag{10}$$

we can find that (1) or (9) is minimized for B given by (10) and this B is also used for (6): F may not be obtained in F-step. Moreover, the original data matrix X may not be available and only the sample covariance matrix S suffices for minimizing (1), since $L\Lambda^{-1}L^\top$ in (10) can be obtained through the eigenvalue decomposition (EVD)

$$A^\top SA = L\Lambda^2L^\top, \tag{11}$$

following from (8): X is found to vanish in (9), (10), and (11).

It should be noted that the A resulting in (6) satisfies $\text{tr}A^\top A = \text{tr}B^\top A$. We can use it in (9) to find that the value of loss function (1) after the update (6) is expressed as

$$f(A) = n\text{tr}S - n\text{tr}A^\top A = n\text{tr}S \times f_N(A). \tag{12}$$

Here, $f_N(A) = 1 - \text{tr}A^\top A/\text{tr}S$ is normalized so as to take a value within $[0, 1]$, thus convenient for checking convergence. Thus, the USMPCA algorithm can be formed as follows:

1. Initialize A .
2. Perform EVD (11) to obtain B with (10).
3. Obtain A with (6).
4. Finish if $f_N(A) \leq \varepsilon$; otherwise go back to 2.

Here, $f_N(A)$ denotes the change in $f_N(A)$ from the previous round. In this paper, $\varepsilon = 0.1^7$ and the algorithm is repeated fifty times with random initialization. Among the resulting solutions, we select the one with the lowest $f_N(A)$ value as the optimal solution, in order to avoid local minimizers. After those procedures, F can be obtained using (7).

4 Percentages of Explained Variances

The loss function value (12) allows us to define the goodness of the resulting A as

$$\text{PEV} = 100\text{tr}A^\top A/\text{tr}S, \quad (13)$$

with $\text{tr}A^\top A = \frac{1}{n}\|FA^\top\|^2$ following from $\frac{1}{n}F^\top F = I_m$. The statistic (13) can be called total percentage of explained variance (PEV), since $\text{tr}S$ in (13) is the total variance of the variables, while $\text{tr}A^\top A = \frac{1}{n}\|FA^\top\|^2$ is the total variance of FA^\top , since (7) shows that F is column-centered as X is so.

The total PEV (13) can be decomposed as the sum of

$$\text{PEV}(j) = 100a_j^\top a_j/\text{tr}S, \quad (14)$$

over $j = 1, \dots, m$. It serves as the PEV index for each component. On the other hand, the PEV for each variable is derived from the fact that (12) can be rewritten as $n \sum_{i=1}^p (s_{ii} - \|\tilde{a}_i\|^2) = n \sum_{i=1}^p s_{ii}(1 - \|\tilde{a}_i\|^2/s_{ii}) \geq 0$, with \tilde{a}_i^\top the i th row of A and s_{ii} the variance of variable i . It gives the percentage of $\|\tilde{a}_i\|^2 = \frac{1}{n}\|F\tilde{a}_i^\top\|^2$ to s_{ii} ,

$$\text{PEV}[i] = 100\|\tilde{a}_i\|^2/s_{ii}. \quad (15)$$

In the same forms as (13), (14), and (15), PEV indices are defined for the standard PCA, which is formulated as minimizing (1) with $\frac{1}{n}F^\top F = I_m$ and $A^\top A$ being a diagonal matrix. The same forms of definitions facilitate the comparison of solutions between USMPCA and PCA in goodness-of-fit. Since PCA is the best rank m approximation of X , the value of the total PEV (13) for USMPCS cannot exceed the one for PCA. However, if the former value is not substantially less than the latter, the USMPCS solution can be considered to be acceptable. It should be noted that USMPCS can be superior to PCA in (14) and (15), as illustrated in Section 6.1.

5 Nonzero Loadings as Covariances

The matrix B defined in (2) contains the covariances of p variables to m components, since X and F are column-centered. By taking this fact into account in (6), the nonzero loadings in A are found to equal the corresponding covariances in B : nonzero a_{ij} equals the covariance between variable i and component j . It implies that the nonzero loadings equal the correlation coefficients of variables to components, when the columns of X have unit variances or S is a correlation matrix, since of $\frac{1}{n}F^\top F = I_m$.

6 Two Examples

The first example is the Pitprop data set [2] given as the correlation matrix obtained from a 180×13 data matrix. We set $m = 6$ following the previous studies to perform USMPCA. The solution subject to $\text{card}(A) = pm/2 = 39$ is shown left in Table 1 with blank indicating zero loadings. There, the total PEV 86.7 is found to be almost equivalent to the PEV 87.0 for PCA: USMPCA approximated the data as well as PCA with a half of loadings vanishing in the former solution. We further performed USMPCA with decreasing $\text{card}(A)$ one by one, to find that the PEV for the solution with $\text{card}(A) = 17$ nearly exceeded 80, a benchmark percentage not being very lower than 87.0 for PCA. That solution is shown right in Table 1. Bold font is used for the PEV for variables and components which exceed the corresponding ones for PCA. One notes that the USMPCA components with $j = 4, 5, 6$ explain more variance than the PCA ones.

Vars	USMPCA: $\text{card}(A) = 39$						PEV	USMPCA: $\text{card}(A) = 17$						PEV	PCA PEV	
	1	2	3	4	5	6		1	2	3	4	5	6			
topdiam	.86	.40					90.4	.89							79.2	90.9
length	.90	.33					92.0	.91							82.9	92.5
moist		.98	-.10				97.5		.96						92.4	97.8
testsg		.90		-.40			97.5		.94						88.6	97.5
ovensg		-.17		-.93			88.7			.81					65.0	86.8
ringtop	.32	.19	.59	-.55	0.29		87.1	.37		.79					76.6	86.4
ringbut	.61		.61	-.41		-.14	93.1	.67		.62					83.4	92.7
bowmax	.54		.15		-0.60		67.8	.61					.51		63.4	68.4
bowdist	.75	.15			-0.20		62.9	.80							63.5	64.0
whoris	.66		.33		-0.38	-.42	86.8	.75			.44				75.1	87.2
clear		.15				.97	96.9				-.98				95.3	95.9
knots	-.11	.25		.25	0.80		77.5						-.92		85.5	80.4
diaknot	.15	.00	-.87	.10	0.31		88.6							-.96	91.6	90.7
PEV	25.8	17.1	12.5	12.0	10.5	8.8	86.7	29.1	13.9	12.8	8.8	8.6	7.0		80.2	87.0
PEV _{PCA}	32.5	18.3	14.5	8.5	7.0	6.3	87.0	32.5	18.3	14.5	8.5	7.0	6.3		87.0	

Table 1: USMPCA solutions for Pitprop data with PCA’s PEV in the final row and column.

The variables are well clustered with every variable loading only one or two components. It makes sense to compare the USMPCA solutions with the classic (subjective) interpretation of the Pitprop component loadings [2], which is summarized in Table 2. The adopted notations mean that the first component is determined by topdiam, length, ringbut, bowmax, bowdist and whoris, the second – by moist and testsg, and etc. The ringbut value for component four in [2, Table 4, p.229] seems incorrect, by inspecting the corresponding eigenvalue. The corrected ”classic” interpretation is given in [8], where ringbut is dropped off the fourth component.

Clearly, the USMPCA solution with $\text{card}(A) = 17$ suggests identical interpretation of the first three components as the one given in [2, p.230]. The fourth component is, indeed, a contrast, but between clear and whoris. The fifth component is also a contrast between knots and bowmax, and the sixth component is a direct measure of diaknot (the average diameter of the knots in inches).

The second example concerns the gene expression data matrix of $n = 17$ time points by $p = 384$ genes presented by [9] and available at <http://faculty.washington.edu/kayee/pca>. The 384 genes are categorized into five phases of cell cycles, with each phase containing 67, 135, 75, 52, and 55 genes, respectively. It suggests $m = 5$, but this choice yielded one trivial component in preliminary trials. We thus reduced m to 4. For $\text{card}(A)$, we first used the integer nearest to the one-third of pm , then increase $\text{card}(A)$ one by one to find that the total PEV of the solution with $\text{card}(A) = 538 \cong 0.35pm$ nearly exceeds a benchmark 70, which is not considerably lower than the PEV 81.2 for PCA with $m = 4$. The resulting A with $\text{card}(A) = 538$

Vars	1	2	3	4	5	6
topdiam	x					
length	x					
moist		x				
testsg		x	x			
ovensg			x			x
ringtop	x		x			
ringbut	x			x		
bowmax	x					
bowdist	x					
whoris	x					
clear				x		
knots					x	
diaknot						x

Table 2: Classic interpretation of the Pitprop component loadings [2, p.229-30].

are presented block-wise in Figure 1. There, the blocks correspond to the five phases, with the block for the second one divided into two, and positive/negative nonzero loadings represented as filled squares/triangles, respectively. The solution is considered to be reasonable, as each phase has a unique feature of loadings: [a] Phases 1, 2, and 4 are characterized by positive loadings for Components 1, 2, and 3, respectively; [b] Phases 5 are characterized by positive loadings for Component 4 and negative ones for 2; [c] Phases 3 consists of the genes positively loaded by Component 2 or 3 and by both.

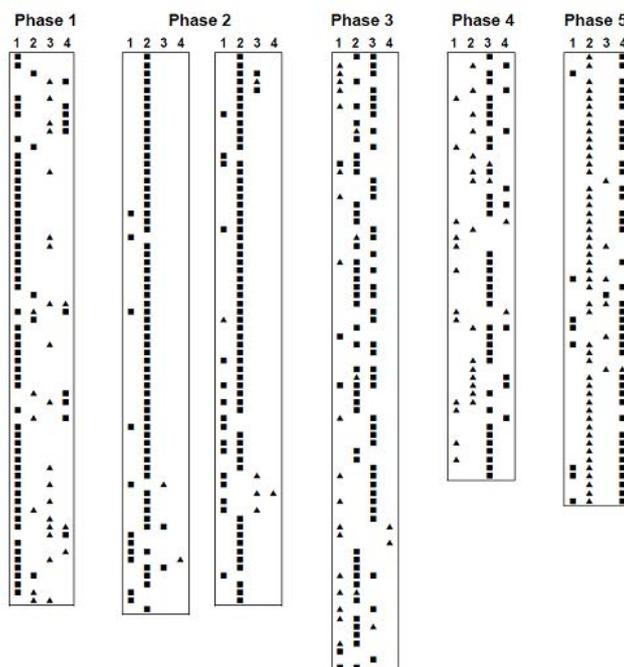


Figure 1: USMPCA solution for gene expression data with blank indicating zero

7 Final Remarks

In this paper, we proposed the penalty-free sparse PCA procedure USMSPCA and presented its alternate least squares algorithm. An advantage of USMSPCA over the penalized sparse PCA is that the cardinality of loadings can be set to a specified integer in advance. For that integer we can use the one conceived easily such as a half or the one-third of the number of loadings, which can be flexibly changed for finding a better solution, as illustrated in the examples. There, it was also illustrated that a solution obtained can be validated by comparing the PEV value with the corresponding one for the standard PCA. The reasonableness of this PEV comparison follows from that the PEV indices for USMSPCA are defined in the same manner as in PCA.

Acknowledgement

This work is supported by a grant RPG-2013-211 from The Leverhulme Trust, UK.

Bibliography

- [1] d'Aspremont, A., Bach, F., & Ghaoui, L. E. (2008) Optimal solutions for sparse principal component analysis, *Journal of Machine Learning Research*, 9, 1269-1294.
- [2] Jeffers, J. N. R. (1967). Two case studies in the application of principal component analysis. *Applied Statistics*, 16, 225-236.
- [3] Jolliffe, I. T., Trendafilov, N. T., & Uddin, M. (2003). A modified principal component technique based on the LASSO. *Journal of Computational and Graphical Statistics*, 12, 531-547.
- [4] Journée M, Nesterov Y, Richtárik P, & Sepulchre R (2010) Generalized power method for sparse principal component analysis, *Journal of Machine Learning Research*, 11, 517-553.
- [5] Moghaddam, B., Weiss, Y., & Avidan, S. (2006). Spectral bounds for sparse PCA: Exact and greedy algorithms. *Advances in Neural Information Processing System*, 18, 915-922.
- [6] Shen, H. & Huang, J. Z. (2008). Sparse principal component analysis via regularized low rank matrix approximation. *Journal of Multivariate Analysis*, 99, 1015 - 1034.
- [7] Qi, X., Luo, R., & Zhao, H. (2013). Sparse principal component analysis by choice of norm. *Journal of Multivariate Analysis*, 114, 127-160.
- [8] Trendafilov, N. T. (2013). From simple structure to sparse components: a review. *Computational Statistics*, published on line DOI:10.1007/s00180-013-0434-5.
- [9] Yeung, K. Y., & Ruzzo, W. L. (2001). Principal component analysis for clustering gene expression data. *Bioinformatics*, 17, 763-774.
- [10] Zou, D. M., Hastie, T., & Tibshirani, R. (2006). Sparse principal component analysis. *Journal of Computational and Graphical Statistics*, 15, 265-286.

Weight choice by minimizing MSE for general likelihood averaging

Ali Charkhi, *KU Leuven*, ali.charkhi@kuleuven.be

Gerda Claeskens, *KU leuven*, gerda.claeskens@kuleuven.be

Bruce E. Hansen, *University of Wisconsin*, behansen@wisc.edu

Abstract. In model averaging a weighted estimator is constructed based on a set of models, extending model selection where a single estimator is constructed from one selected model found via an information criterion. Several studies discuss the weight choice for linear models only and almost all studies assign weights to models by using optimization routines, specifically quadratic programming and nonlinear optimization. None of these studies worried about the unicity of the estimated weights, while in fact, with those methods the chosen weight is often non-unique, resulting in difficulties with interpretations of weighted averages. Our contribution is threefold: (1) We minimize an estimator for the mean squared error in a local misspecification framework from which unique weights can be assigned to a set of ‘linearly independent design matrix’ models. (2) The weight choice applies to a broad range of models including generalized linear models. (3) In linear models the computational complexity of averaging may be reduced since weighted predictions from nested and singleton models are equal. In a simulation study in Poisson regression the performance of our method of averaging is compared with other such methods. The simulation results show that the proposed method performs well.

Keywords. Model averaging, Likelihood, Mean squared error, Choice of weights, Smoothed AIC, Smoothed BIC.

1 Introduction

Model averaging is an alternative to model selection in which a new estimator of a population quantity is constructed based on a weighted average of estimators in each candidate model. Most of the model averaging literature considers the least squares framework, [4] proposed the Mallows criterion for model averaging which was extended by [9] for non nested models. Unlike most of the theoretical results for least squares model averaging with the homoscedasticity assumption, [5] challenged this assumption and defined a jackknife model averaging estimator with heteroskedastic errors and they proved the optimality of their estimator. [8] proposed a new

model averaging estimator in the local asymptotic framework for linear regression and derived the asymptotic distribution of a plug-in averaging estimator.

We consider the choice of weights for model averaging in likelihood regression models. Several models are available for the estimation of a population parameter μ . Various models, including parts, or all, of the covariates might be considered, each one coming with its own estimator of μ , say, $\hat{\mu}_S$ for a model indexed by S . So, the model averaging estimator is

$$\hat{\mu}_w = \sum_{j=1}^M w_j \hat{\mu}_{S_j}. \quad (1)$$

Using the likelihood framework, which has so far received scant attention, [6] studied the properties of the averaging estimator when random weights are used to construct a compromised estimator; [7] went one step further and used an estimator for the mean squared error (MSE) of a non-random weighted estimator which they minimized in a special class of random data dependent weights. Their method is also applicable for least squares estimations. Logistic regression was considered by [10] who minimized a plug-in estimator of the asymptotic squared error to define weights for ordered logit models.

In this paper we consider averaged estimators obtained by general maximum likelihood estimation, with an application to Poisson regression. The studied choice of the weights is by minimizing an estimated mean squared error of $\hat{\mu}_w$ under local misspecification. Our main contributions are (i) to define a method for averaging estimators in a general likelihood framework, (ii) to find a set of models for which we can assign unique weights for each model in that set and (iii) our proposed method is computationally attractive. Unlike other methods, we do not need quadratic programming for minimizing risk ([4, 5, 8]) nor heavy nonlinear optimization routines ([7]). We derive the theoretical formula for the weights in a general case. Also, (iv) we show the equality of prediction values for our method in models with linearly independent design matrices in linear regression. Hence singleton models (with only a single covariate in each model) perform as well as any other linearly independent design matrix models. This result is promising for high dimensional data. Moreover, our weights are not restricted to lie in the unit simplex set but the sum of weights should be equal to one which is a necessary assumption for consistency of the averaging estimator ([6]).

2 Notation and setting

In a regression setting, take Y_1, \dots, Y_n independent with density function $f_n(y; x) = f(y; x, \theta_0, \gamma_0 + \delta/\sqrt{n})$, where in a variable selection context, the p -vector θ is included in every model and components of the q -vector γ may or may not be relevant (e.g. these are coefficients corresponding to irrelevant covariates). The true values of (θ, γ) are $(\theta_0, \gamma_0 + \delta/\sqrt{n})$ under the local misspecification setting, and (θ_0, γ_0) under a narrow model when only θ is included in the model and $\gamma = \gamma_0$ a known value (e.g. zero).

We here phrase some further notation for the regression setting. In the case of i.i.d. data, the covariate vector x is not present and the averages reduce to a single term. Define the vector of first derivatives of the log-likelihood

$$\begin{pmatrix} U(y; x) \\ V(y; x) \end{pmatrix} = \begin{pmatrix} \partial \log f(y; x, \theta_0, \gamma_0) / \partial \theta \\ \partial \log f(y; x, \theta_0, \gamma_0) / \partial \gamma \end{pmatrix},$$

and let the information matrix

$$J(x) = \text{Var} \begin{pmatrix} U(Y;x) \\ V(Y;x) \end{pmatrix} \text{ and } J_n = \frac{1}{n} \sum_{i=1}^n J(x_i),$$

be partitioned according to the lengths of θ and γ as

$$J(x) = \begin{pmatrix} J_{00}(x) & J_{01}(x) \\ J_{10}(x) & J_{11}(x) \end{pmatrix}, \quad J_n = \begin{pmatrix} J_{n,00} & J_{n,01} \\ J_{n,10} & J_{n,11} \end{pmatrix}, \quad J_n^{-1} = \begin{pmatrix} J_n^{00} & J_n^{01} \\ J_n^{10} & J_n^{11} \end{pmatrix}.$$

In a regression context, the information matrix J_n (and submatrices thereof) are all averages over the different observations, assumed to converge to a matrix J when $n \rightarrow \infty$. Submatrices of the limit matrices J and its inverse J^{-1} are defined as above, though without the subscript n .

Let S be a subset of $\{1, \dots, q\}$, indicating a submodel of the full model. We wish to estimate a population quantity $\mu = \mu(\theta, \gamma)$ (a focus parameter), for which we assume that its derivatives with respect to θ and γ exist in a neighborhood of (θ_0, γ_0) . This case is more general than averaging the regression coefficients, we may also average predictions of the form $x^t\beta$ in this way.

Maximum likelihood estimators are used for estimation in each submodel. We then know that in each submodel estimators are normally distributed in the limit ([6])

$$\sqrt{n}(\hat{\mu}_S - \mu_{true}) \xrightarrow{d} \Lambda_S = \Lambda_0 + \omega^t(\delta - G_S D). \tag{2}$$

Here, $\omega = J_{10}J_{00}^{-1}\partial\mu/\partial\theta - \partial\mu/\partial\gamma$, $\Lambda_0 \sim N(0, \tau_0^2)$ with $\tau_0^2 = (\partial\mu/\partial\theta)^t J_{00}^{-1} \partial\mu/\partial\theta$, $D \sim N_q(\delta, Q)$ with $Q = J^{11}$ and $G_S = Q_S^0 Q^{-1} = \pi_S^t Q_S \pi_S Q^{-1}$ with $Q_S = (\pi_S Q^{-1} \pi_S^t)^{-1}$ and π_S is a $|S| \times q$ projection matrix selecting the rows with an index belonging to S .

Note that the matrices Q_S^0 are ‘partial’ inverses of Q^{-1} . The ‘0’ superscript denotes by definition the following construction. First select of Q^{-1} those rows and columns with indices in S , then invert that matrix. Next, place this in a full $q \times q$ matrix in the rows and columns indicated by S and set all other entries equal to zero. In particular, when $S = \{1, \dots, q\}$ (full model), $Q_S^0 = Q$. Some possibilities of sets of models to average over are all possible subsets ($M = 2^q$), nested models ($M = q + 1$) and singleton models ($M = q + 1$).

3 The mean squared error expression

In this section, we state the mean squared error (MSE) of the averaged estimator and its asymptotic distribution. From (2), it immediately follows that the MSE of a single submodel estimator is converging to

$$\text{MSE}(\hat{\mu}_S, \delta) = \tau_0^2 + \omega^t Q_S^0 \omega + \omega^t (I_q - G_S) \delta \delta^t (I_q - G_S)^t \omega.$$

For the weighted estimator (1), with M a finite number of models, not depending on the sample size, and with a non-random set of weights w_1, \dots, w_M that sums to 1, it follows that (see [6])

$$\sqrt{n}(\hat{\mu}_w - \mu_{true}) \xrightarrow{d} \sum_{j=1}^M w_j \Lambda_{S_j} = \sum_{j=1}^M w_j \{ \Lambda_0 + \omega^t (\delta - G_{S_j} D) \},$$

from which follows that $MSE(\hat{\mu}_w) = \tau_0^2 + R(\delta)$ with the same τ_0^2 as before (since the weights sum to 1) and

$$R(\delta) = \omega^t \left\{ (I_q - \sum_{j=1}^M w_j Q_{S_j}^0 Q^{-1}) \delta \delta^t (I_q - \sum_{j=1}^M w_j Q_{S_j}^0 Q^{-1})^t + (\sum_{j=1}^M w_j Q_{S_j}^0) Q^{-1} (\sum_{j=1}^M w_j Q_{S_j}^0)^t \right\} \omega, \quad (3)$$

or equivalently $R(\delta) = w^t F w$ where the (j, k) th entry of F is given by $(j, k = 1, \dots, M)$

$$F_{jk} = \omega^t \left\{ (I_q - Q_{S_j}^0 Q^{-1})^t \delta \delta^t (I_q - Q_{S_k}^0 Q^{-1}) + (Q_{S_j}^0 Q^{-1} Q_{S_k}^0) \right\} \omega. \quad (4)$$

Hence, the theoretically optimal weights that minimize the MSE are

$$w_{mse} = \underset{w \in \mathcal{H}}{\operatorname{argmin}} w^t F w. \quad (5)$$

where $\mathcal{H} = \{(w_1, \dots, w_M) : \sum_{j=1}^M w_j = 1\}$.

Some properties of the plug-in estimator

Calculating the weights in (5) requires to estimate all quantities in (3) or (4). Almost all unknown parameters in (4) can be estimated consistently except the δ . The unbiased estimator for δ which is $\hat{\delta} = \sqrt{n}(\hat{\gamma}_{full} - \gamma_0) \rightarrow_d D \sim N_q(\delta, Q)$ may be used in estimation of the MSE. By plugging in estimators, it follows that $(j, k = 1, \dots, M)$

$$\hat{F}_{jk} = \hat{\omega}^t (I_q - \hat{Q}_{S_j}^0 \hat{Q}^{-1})^t \hat{\delta} \hat{\delta}^t (I_q - \hat{Q}_{S_k}^0 \hat{Q}^{-1}) \hat{\omega} + \hat{\omega}^t (\hat{Q}_{S_j}^0 \hat{Q}^{-1} \hat{Q}_{S_k}^0) \hat{\omega}. \quad (6)$$

For interpretation purposes, it is important to know whether the obtained weights are unique or not. In [1] we obtain and prove the following useful results. Property 1 presents sufficient conditions for the unicity of the weights.

Property 1 *If Q is positive definite, ω is not equal to 0_M and the matrices $Q_{S_j}^0$ ($j = 1, \dots, M$) are linearly independent, then the $M \times M$ matrix \tilde{Q} with (j, k) th element $\omega^t Q_{S_j}^0 Q^{-1} Q_{S_k}^0 \omega$ is positive definite.*

Considering equation (6), the first term is always positive semi-definite and under the conditions of Property 1 it is positive-definite which results in a positive-definite matrix F . So, by solving (5)

$$\widehat{w}_{mse} = \underset{w \in \mathcal{H}}{\operatorname{argmin}} w^t \hat{F} w = \frac{1_M^t \hat{F}^{-1}}{1_M^t \hat{F}^{-1} 1_M}, \quad (7)$$

where 1_M denotes a vector of ones with length M .

The main conclusion of Property 1 is that the number of models for having unique weights cannot exceed $q + 1$ where q is the number of potential covariates, plus one for the narrow model. Several sets of models with at most $q + 1$ independent design matrices can be considered such as nested models and singleton models. The main challenging part of using nested models is the order of the regressors, whereas singleton models are independent of regressor orders, see also the simulation study.

It should be noted that our method can also consider any set of models like all possible models, but the weights are not longer unique. In these cases, our method is similar to other

proposed model averaging methods in linear regression, e.g., [5], [8] and [10] where they used quadratic programming. There are some other methods like [7] in which they used nonlinear constrained optimization method to define the weights in a specific class of weights. In their method, the weights may not be unique even not when considering only $q + 1$ models. The main problem of non unique weights is that the prediction values are not unique, so with the same method one can get different predictions for population parameter.

It can be shown that the matrix F in (6) converges in distribution to F^* for which the (j, k) th element ($j, k = 1, \dots, M$) is equal to

$$F_{j,k}^* = \omega^t (I_q - Q_{S_j}^0 Q^{-1}) D D^t (I_q - Q_{S_k}^0 Q^{-1})^t \omega + \omega^t (Q_{S_j}^0 Q^{-1} Q_{S_k}^0) \omega,$$

with $D \sim N(\delta, Q)$.

While the explicit form of the weights in (7) is useful for direct computation, it hints at a complicated limiting distribution. Using that $\widehat{F} \rightarrow_d F^*$, we get a limiting distribution of \widehat{w}_{mse} in terms of F^* too, see [1].

Property 2 *Assume that \widehat{F} and F^* are invertible. Let $\widehat{w}_{mse} = \operatorname{argmin}_{w \in \mathcal{H}} w^t \widehat{F} w$ and $w^* = \operatorname{argmin}_{w \in \mathcal{H}} w^t F^* w$. Then $\widehat{w}_{mse} \rightarrow_d w^*$. Also, by using the joint convergence in distribution of all $\sqrt{n}(\widehat{\mu}_{S_j} - \mu_{true})$ and \widehat{w} to corresponding Λ_{S_j} and w^* , the model averaging estimator has a limiting distribution*

$$\sqrt{n}(\widehat{\mu}_{\widehat{w}_{mse}} - \mu_{true}) \xrightarrow{d} \sum_{j=1}^M w_j^* \Lambda_{S_j}.$$

Note that by the randomness of the weights w^* the limiting distribution is not normal. For deterministic weights, the limiting distribution is normal.

The third result, see [1], states that in linear regression the prediction values for the mean of the response vector ($E(Y) = X\beta$) for averaging over nested models and over singleton models with our method are equal.

Property 3 *If $p \geq 1$ and $q \geq 2$, then the prediction values in linear models for nested model averaging and singleton model averaging are equal when MSE optimal weights (\widehat{w}_{mse}) for weighted prediction are used.*

This has promising consequences for models with a large number of covariates where an all subsets model averaging would be time consuming, while singleton models are much easier to fit.

4 Simulation Study for Poisson Regression

We now investigate the finite sample performance of the proposed plug-in estimator of the MSE (PMSE) via a Monte Carlo simulation in nested and singleton models. Our goal for this simulation is twofold: (i) compare the MSE estimator model averaging scheme with other model averaging schemes, (ii) examine the effect of the number of non-zero coefficients for the auxiliary regressors.

Four estimators are considered to be compared with our estimator: AIC and BIC post-model selection methods and model averaging estimators corresponding to their smoothed estimator, SAIC and SBIC, with the weights for the m th model

$$w_m^{saic} = \frac{\exp\left(-\frac{1}{2}AIC_m\right)}{\sum_{j=1}^{q+1} \exp\left(-\frac{1}{2}AIC_j\right)}, \quad w_m^{sbic} = \frac{\exp\left(-\frac{1}{2}BIC_m\right)}{\sum_{j=1}^{q+1} \exp\left(-\frac{1}{2}BIC_j\right)}.$$

The response values Y_i have a Poisson distribution with mean $\mu_i = \exp(x_i^t \delta)$ with the following specifications: $p = 0$ (i.e. no core regressors), $q = 8$ with $(x_{1i}, \dots, x_{8i}) \sim N_{p+q}(0, \Omega)$ in which $\Omega_{ij} = 1$ for $i = j$ and $\Omega_{ij} = \rho$ for $i \neq j$. The value of ρ varies in the set $\{0, 0.25, 0.5, 0.75\}$. The value of γ_0 is set to zero and δ values are considered according to the following scenarios:

$$\text{scenario 1A: } \delta_1 = (-1, -4, 3, -4, 0.6, 4, 0, 5)$$

$$\text{scenario 2A: } \delta_2 = (-1, -4, 3, -4, 0, 0, 0, 0)$$

For nested models, the order in which the variables enter the model is important. There are two other scenarios, scenarios 1B and 2B. In the construction of the weighted estimator in scenarios 1B and 2B we use the same random samples as above, but construct the design matrix for these scenarios (and take the implied order for constructing nested models) as $X = (x_5, x_6, x_7, x_8, x_1, x_2, x_3, x_4)$ and the δ values corresponding each regressor stay the same. The sample sizes are varying in the set $\{100, 500, 1100\}$. All Monte Carlo simulations are based on 2000 replications. We generate $n + 1$ observations in which the last observation is used as test data set. The focus parameter is the mean of the response value for the test data set. Each method is assessed based on the median of the squared prediction error for the test data sets over 2000 replications which can be written as

$$\text{MSPE} = \text{median}\{(\hat{\mu}_{\hat{w},i} - \mu_i)^2 : i = 1, \dots, 2000\}$$

where for each test data set with values x ,

$$\hat{\mu}_{\hat{w}} = \exp\left\{\sum_{j=1}^M \hat{w}_j x^t \hat{\delta}_j\right\}.$$

Table 1 presents the results for the simulations. For singleton models, the AIC and BIC values are identical (the penalty does not have an effect in singleton models), hence, we show the results for AIC, SAIC and the PMSE methods. As Table 1 shows, the order of the regressors for nested models is important and within the same data set, the results are varying from one order of regressors to another order and determining an in some sense ‘optimal’ order of regressors is challenging. So, it is not that informative to interpret the results for nested averaging in general when the regressors are not naturally ordered. In contrast, singleton models are independent of regressor order and the PMSE method performs the best for singleton averaging. Moreover, PMSE singleton averaging for scenarios 1A, 1B and 2B in almost all situations performs better than all other methods in singleton and nested averaging. Another interesting property of singleton averaging by PMSE is that it performs quite independently of correlation between regressors. Other things being equal, changing the ρ reduces the MSPE for other methods in singleton models, while PMSE works equally well for all considered values of ρ . For example, for $n = 500$ in scenario 2A, increasing the ρ from 0 to 0.75 cause a reduction in the MSPE from 0.023 to 0.006 for the SAIC method, whereas in our method the MSPE is always around 0.005.

ρ	n	Scen.	nested					singleton		
			AIC	BIC	SAIC	SBIC	PMSE	AIC	SAIC	PMSE
0	100	1A	0.019	0.019	0.019	0.019	0.030	0.283	0.257	0.019
		1B	0.018	0.019	0.019	0.019	0.031			
		2A	0.011	0.009	0.012	0.010	0.025	0.108	0.091	0.019
		2B	0.020	0.023	0.022	0.025	0.029			
	500	1A	0.005	0.006	0.005	0.006	0.006	0.052	0.049	0.005
		1B	0.005	0.006	0.006	0.006	0.006			
		2A	0.003	0.002	0.003	0.003	0.006	0.026	0.023	0.005
		2B	0.005	0.011	0.006	0.011	0.006			
	1100	1A	0.003	0.003	0.003	0.003	0.003	0.025	0.023	0.002
		1B	0.003	0.003	0.003	0.003	0.003			
		2A	0.001	0.001	0.002	0.001	0.003	0.012	0.011	0.003
		2B	0.003	0.007	0.003	0.006	0.003			
0.25	100	1A	0.019	0.020	0.019	0.020	0.027	0.198	0.190	0.017
		1B	0.019	0.021	0.020	0.023	0.028			
		2A	0.011	0.010	0.012	0.010	0.024	0.073	0.067	0.020
		2B	0.022	0.028	0.023	0.028	0.028			
	500	1A	0.005	0.006	0.005	0.006	0.006	0.043	0.041	0.005
		1B	0.005	0.006	0.005	0.007	0.005			
		2A	0.003	0.003	0.003	0.003	0.005	0.020	0.018	0.005
		2B	0.005	0.012	0.005	0.012	0.005			
	1100	1A	0.003	0.003	0.003	0.004	0.003	0.019	0.019	0.003
		1B	0.003	0.004	0.003	0.004	0.003			
		2A	0.001	0.002	0.001	0.002	0.002	0.009	0.008	0.002
		2B	0.003	0.008	0.003	0.008	0.003			
0.5	100	1A	0.025	0.031	0.027	0.033	0.030	0.140	0.132	0.023
		1B	0.026	0.035	0.028	0.035	0.031			
		2A	0.014	0.015	0.015	0.016	0.029	0.067	0.056	0.026
		2B	0.028	0.045	0.030	0.045	0.031			
	500	1A	0.005	0.009	0.006	0.009	0.005	0.030	0.028	0.005
		1B	0.005	0.013	0.006	0.011	0.006			
		2A	0.003	0.004	0.003	0.004	0.005	0.013	0.011	0.005
		2B	0.006	0.018	0.006	0.016	0.005			
	1100	1A	0.002	0.007	0.003	0.006	0.002	0.013	0.012	0.002
		1B	0.003	0.009	0.003	0.008	0.003			
		2A	0.002	0.003	0.002	0.003	0.003	0.006	0.005	0.003
		2B	0.003	0.011	0.003	0.010	0.003			
0.75	100	1A	0.028	0.056	0.028	0.046	0.026	0.071	0.070	0.021
		1B	0.028	0.061	0.028	0.052	0.026			
		2A	0.014	0.021	0.014	0.018	0.025	0.035	0.027	0.023
		2B	0.030	0.059	0.028	0.047	0.028			
	500	1A	0.006	0.019	0.007	0.017	0.006	0.016	0.015	0.005
		1B	0.007	0.019	0.007	0.016	0.006			
		2A	0.004	0.007	0.003	0.006	0.005	0.007	0.006	0.005
		2B	0.007	0.015	0.007	0.014	0.005			
	1100	1A	0.003	0.009	0.003	0.008	0.003	0.007	0.006	0.003
		1B	0.003	0.009	0.003	0.008	0.003			
		2A	0.002	0.004	0.002	0.003	0.003	0.003	0.003	0.003
		2B	0.003	0.008	0.004	0.007	0.003			

Table 1: MSPE of nested and singleton models based on AIC, BIC, SAIC, SBIC and PMSE.

Bibliography

- [1] Charkhi, A., Claeskens, G. and Hansen, B.E. (2014). On the unicity of model averaging weights in likelihood models. Technical report, KU Leuven, Belgium.
- [2] Claeskens, G. and Hjort, N. L. (2003) *The focused information criterion*. Journal of the American Statistical Association, **98**, 900–916. With discussion and a rejoinder by the authors.
- [3] Dostal, Z. (2009). *Optimal Quadratic Programming Algorithms, with Applications to Variational Inequalities*. Springer US, New York.
- [4] Hansen, B. (2007). *Least squares model averaging*. Econometrica, **75**, 1175–1189.
- [5] Hansen, B. E. and Racine, J. S. (2012). *Jackknife model averaging*. Journal of Econometrics, **167**, 38–46.
- [6] Hjort, N. L. and Claeskens, G. (2003). *Frequentist model average estimators*. Journal of the American Statistical Association, **98**, 879–899. With discussion and a rejoinder by the authors.
- [7] Liang, H., Zou, G., Wan, A. T. K., and Zhang, X. (2011). *Optimal weight choice for frequentist model average estimators*. Journal of the American Statistical Association, **106(495)**, 1053–1066.
- [8] Liu, C.-A. (2013). *Distribution theory of the least square averaging estimator*. National University of Singapore, Working paper.
- [9] Wan, A., Zhang, X., and Zou, G. (2010). *Least squares model averaging by Mallows criterion*. Journal of Econometrics, **156**, 277–283.
- [10] Wan, A. T. K., Zhang, X., and Wang, S. (2013). *Frequentist model averaging for multinomial and ordered logit models*. International Journal of Forecasting, **30**, 118–128.

Quantifying and localizing state uncertainty in hidden Markov models using conditional entropy profiles

Jean-Baptiste Durand, *Univ. Grenoble Alpes, LJK and Inria, Mistis, F-38000 Grenoble, France,*
jean-baptiste.durand@imag.fr

Yann Guédon, *CIRAD, UMR AGAP and Inria, Virtual Plants, F-34095 Montpellier, France,*
guedon@cirad.fr

Abstract. A family of graphical hidden Markov models that generalizes hidden Markov chain (HMC) and tree (HMT) models is introduced. It is shown that global uncertainty on the state process can be decomposed as a sum of conditional entropies that are interpreted as local contributions to global uncertainty. An efficient algorithm is derived to compute conditional entropy profiles in the case of HMC and HMT models. The relevance of these profiles and their complementarity with other state restoration algorithms for interpretation and diagnosis of hidden states is highlighted. It is also shown that classical smoothing profiles (posterior marginal probabilities of the states at each time, given the observations) cannot be related to global state uncertainty in the general case.

Keywords. Hidden Markov models, State inference, Conditional entropy.

1 Introduction

Hidden Markov models (HMMs) have been used frequently in sequence analysis for modeling various types of latent structures, such as homogeneous zones or noisy patterns (Ephraim & Mehrav, 2002). They have been extended from sequences to more general structures, particularly tree structures. In HMMs, inference for model parameters can be distinguished from inference for the state process given parameters. This work focuses on state process inference.

State inference is particularly relevant in numerous applications where the unobserved states have a meaningful interpretation. In such cases, the state sequence has to be restored. The restored states may be used, typically, in prediction, in segmentation or in denoising (Ephraim

& Mehrav, 2002). Such use of the state sequence relies on the assumption that uncertainty on the state process given observations should be reasonably low. Not only is state restoration essential for model interpretation, it is generally used for model diagnostic and validation as well, for example by visualising some functions of the states – typically, to compare histograms with conditional densities given the states. The use of restored states in the above-mentioned contexts makes assessment of state sequence uncertainty a critical step of the analysis.

Global quantification of such uncertainty has been addressed by Hernando *et al.* (2005). However, this is insufficient for detailed state interpretation: knowledge of the distribution of that global uncertainty along the structure is also of primary importance. Quantification of local state uncertainty given observed sequence $\mathbf{X} = \mathbf{x}$ for a known HMC model has been addressed by either enumeration of state sequences, or by state profiles, which are state sequences summarised in a $K \times T$ array, T being the sequence length and K the number of states (Guédon, 2007).

We here address quantification of state uncertainty in an HMM with observed process $\mathbf{X} = (X_v)_{v \in \mathcal{V}}$ indexed by a fixed Directed Acyclic Graph (DAG) \mathcal{G} with vertex set \mathcal{V} and edge set \mathcal{E} . This family of HMMs is referred to as graphical hidden Markov models (GHMMs). This family contains hidden Markov chain (HMC) and tree (HMT) models. Let $\mathbf{S} = (S_v)_{v \in \mathcal{V}}$ denote the associated hidden state process, S_v taking values in the set $\{0, \dots, K-1\}$. Let \mathbf{x} be a possible realization of \mathbf{X} . Let $\text{pa}(v)$ denote the parent of vertex v and for any subset U of \mathcal{V} , let X_U (resp. \mathbf{x}_U) denote the family of random variables $(X_u)_{u \in U}$ (resp. observations $(x_u)_{u \in U}$). It is assumed that: \mathbf{S} satisfies the Markovian factorization property associated with DAG \mathcal{G} , where the vertex set \mathcal{V} is assimilated to the family of random variables $(S_v)_{v \in \mathcal{V}}$ (Lauritzen, 1996); the distribution of \mathbf{S} is parametrized by the transition probabilities $p_{\mathbf{s}_{\text{pa}(v)},k} = P(S_v = k | \mathbf{S}_{\text{pa}(v)} = \mathbf{s}_{\text{pa}(v)})$ and for the source vertices (vertices with no parent) u in \mathcal{G} , by the initial probabilities $(P(S_u = k))_k$; given \mathbf{S} , the random variables $(X_v)_v$ are independent and X_u is independent from $(S_v)_{v \neq u}$.

Usually, profiles of smoothed probabilities $(P(S_v = k | \mathbf{X} = \mathbf{x}))_{v \in \mathcal{V}}$ with $k = 0, \dots, K-1$ have been used for quantifying state uncertainty. This approach suffers from two main shortcomings: as will be shown later, perception of state uncertainty associated with those profiles leads to overestimating global uncertainty of \mathbf{S} given $\mathbf{X} = \mathbf{x}$. Moreover, visualization of those multidimensional profiles is made difficult by the graphical nature of arbitrary DAGs \mathcal{G} , provided that $K > 2$. In our approach, entropy H is considered as the canonical measure of uncertainty. Thus, $H(\mathbf{S} | \mathbf{X} = \mathbf{x})$ quantifies state process uncertainty given observations. This entropy can be decomposed into a sum of entropies. Every term of that sum is associated with one vertex in \mathcal{V} . Hence, these entropies can be interpreted as local contributions to global uncertainty. Since these profiles are unidimensional, they can be drawn whatever the graphical structure \mathcal{G} .

In what follows, this decomposition is made explicit. Then efficient algorithms are given in the HMC and HMT model cases to compute the elements of the decomposition. It is shown using synthetic and real-case data that the obtained local entropy profiles are relevant for state uncertainty diagnosis and state interpretation. These algorithms are complementary with approaches that enumerate the L most likely state restorations (so-called generalized Viterbi algorithm), and with approaches that compute profiles of alternative states to the most likely state process value. This so-called Viterbi forward–backward algorithm formally solves the optimization problem

$$(\arg) \max_{(s_u)_{u \neq v}} P((S_v = s_v)_{u \neq v}, S_v = k | \mathbf{X} = \mathbf{x}).$$

It is also shown that usual smoothed probability profiles are not relevant for quantifying global state uncertainty, due to their inherent marginalization property.

2 Conditional entropy profiles

Let \mathbf{X} be a GHMM as defined in Section 1. It is assumed that the associated hidden state process \mathbf{S} satisfies the factorization associated with the Markov property on \mathcal{G} :

$$\forall \mathbf{s}, P(\mathbf{S} = \mathbf{s}) = \prod_{v \in \mathcal{V}} P(S_v = s_v | \mathbf{S}_{\text{pa}(v)} = \mathbf{s}_{\text{pa}(v)}), \quad (1)$$

where $P(S_v = s_v | \mathbf{S}_{\text{pa}(v)} = \mathbf{s}_{\text{pa}(v)})$ refers to $P(S_s = s_s)$ if $\text{pa}(v) = \emptyset$.

The decomposition of entropy $H(\mathbf{S} | \mathbf{X} = \mathbf{x})$ comes from the conditional distribution of \mathbf{S} given $\mathbf{X} = \mathbf{x}$ also satisfying the factorization property of \mathcal{G} :

$$P(\mathbf{S} = \mathbf{s} | \mathbf{X} = \mathbf{x}) = \prod_v P(S_v = s_v | \mathbf{S}_{\text{pa}(v)} = \mathbf{s}_{\text{pa}(v)}, \mathbf{X} = \mathbf{x}),$$

with the same convention as before if $\text{pa}(v) = \emptyset$.

Proof. This property is proved by induction on the vertices of \mathcal{G} (as would be proved factorization (1)). The random variables (\mathbf{S}, \mathbf{X}) satisfy the Markov property on DAG \mathcal{G}' which edge set \mathcal{E}' is defined as $a \in \mathcal{E}' \Leftrightarrow \{[a = (S_u, S_v) \text{ and } (u \in \text{pa}(v))] \text{ or } a = (S_u, X_u)\}$. Let u in \mathcal{G} be a sink vertex (vertex without children): then S_u is separated from $(S_v)_{v \neq u, v \notin \text{pa}(u)}$ by $\mathbf{S}_{\text{pa}(u)}$ in the moral graph of \mathcal{G}' . Thus, the following factorization holds:

$$P(\mathbf{S} = \mathbf{s} | \mathbf{X} = \mathbf{x}) = P(S_u = s_u | \mathbf{S}_{\text{pa}(u)} = \mathbf{s}_{\text{pa}(u)}, \mathbf{X} = \mathbf{x}) P((S_v)_{v \neq u} = (s_v)_{v \neq u} | \mathbf{X} = \mathbf{x}).$$

□

The additive decomposition of entropy is obtained by applying the chain rule (Cover & Thomas, 2006, chap. 2)

$$H(\mathbf{S} | \mathbf{X} = \mathbf{x}) = \sum_v H(S_v | \mathbf{S}_{\text{pa}(v)}, \mathbf{X} = \mathbf{x}), \quad (2)$$

with the same convention as before if $\text{pa}(v) = \emptyset$. As a consequence, the global state process uncertainty is decomposed as a sum of conditional entropies $(H(S_v | \mathbf{S}_{\text{pa}(v)}, \mathbf{X} = \mathbf{x}))_{v \in \mathcal{V}}$, which define an entropy profile. Hence, each term of the sum is interpreted as a local uncertainty that contributes additively to global uncertainty.

In contrast, marginal entropies $(H(S_v | \mathbf{X} = \mathbf{x}))_{v \in \mathcal{V}}$ quantify uncertainty associated with smoothed probabilities $\xi_v(k) = P(S_v = k | \mathbf{X} = \mathbf{x})$ for $v \in \mathcal{V}$ and $0 \leq k < K$. These marginal entropies are upper bounds of the conditional entropies (Cover & Thomas (2006), chap. 2). Hence,

$$H(\mathbf{S} | \mathbf{X} = \mathbf{x}) \leq \sum_v H(S_v | \mathbf{X} = \mathbf{x}).$$

As a consequence, smoothed probability profiles do not represent uncertainty on the value of \mathbf{S} .

The particular case of HMC models is considered. Here \mathcal{G} is a linear graph with T vertices, and for any $t < T$, $X_0 = x_0, \dots, X_t = x_t$ is denoted by $X_0^t = x_0^t$. Here (2) can be rewritten as

$$H(\mathbf{S} | \mathbf{X} = \mathbf{x}) = H(S_0 | \mathbf{X} = \mathbf{x}) + \sum_{t=1}^{T-1} H(S_t | S_{t-1}, \mathbf{X} = \mathbf{x}),$$

with

$$H(S_t|S_{t-1}, \mathbf{X} = \mathbf{x}) = - \sum_{i,j} P(S_t = j, S_{t-1} = i | \mathbf{X} = \mathbf{x}) \log P(S_t = j | S_{t-1} = i, \mathbf{X} = \mathbf{x}).$$

This results from definition $H(S_t|S_{t-1}, \mathbf{X} = \mathbf{x}) = E[-\log P(S_t|S_{t-1}, \mathbf{X} = \mathbf{x})]$, where expectation is under $P(S_t, S_{t-1} | \mathbf{X} = \mathbf{x})$. The usual forward recursion computes $\alpha_t(j) = P(S_t = j | X_0^t = x_0^t)$ and $\gamma_t(j) = P(S_t = j | X_0^{t-1} = x_0^{t-1})$ for each time t and each state j and combine them in the backward recursion to yield the smoothed probabilities $\xi_t(j) = P(S_t = j | \mathbf{X} = \mathbf{x})$. Thus, computation of the conditional entropy profile $H(S_t|S_{t-1}, \mathbf{X} = \mathbf{x})$ with $0 < t \leq T - 1$ can be integrated in the backward recursion by computing $P(S_t = j | S_{t-1} = i, \mathbf{X} = \mathbf{x}) = \xi_t(j)p_{ij}\alpha_{t-1}(i)/\{\gamma_t(j)\xi_{t-1}(i)\}$ where $p_{ij} = P(S_t = j | S_{t-1} = i)$ is the transition probability. This approach can be seen as an alternative to the algorithm of Hernandez *et al.* (2005). It allows the computation of $H(\mathbf{S} | \mathbf{X} = \mathbf{x})$ with the same complexity in $\mathcal{O}(TK^2)$, but the advantage of our approach is to provide the conditional entropy profile.

In the case of HMTs indexed by tree $\mathcal{G} = \mathcal{T}$ the smoothed probabilities $\xi_v(k) = P(S_v = k | \mathbf{X} = \mathbf{x})$ are computed for $v \in \mathcal{T}$ by an upward-downward algorithm. A numerically stable iterative algorithm was proposed by Durand *et al.* (2004). It relies on an upward recursion, initialized at the leaf vertices of \mathcal{T} . The computed quantities are $\beta_v(k) = P(S_v = k | \bar{\mathbf{X}}_v = \bar{\mathbf{x}}_v)$ and $\beta_{\text{pa}(v),v}(k) = P(\bar{\mathbf{X}}_v = \bar{\mathbf{x}}_v | S_{\text{pa}(v)} = k) / P(\bar{\mathbf{X}}_v = \bar{\mathbf{x}}_v)$ for each vertex v and each state j , where $\bar{\mathbf{X}}_v$ denotes the subtree rooted in v . These quantities are computed as a function of β_u and $\beta_{\text{pa}(u),u}$ for the children u of v . The algorithm complexity is in $\mathcal{O}(K^2)$ per iteration. The smoothed probabilities are computed using a downward recursion initialized at the root vertex of \mathcal{T} . In this recursion, the $\xi_v(k)$ are computed as a function of $\xi_{\text{pa}(v)}$, β_v and $\beta_{\text{pa}(v),v}$. The complexity is in $\mathcal{O}(K^2)$ per iteration as well. Similarly to the HMC case, adding the computation of

$$H(S_v | S_{\text{pa}(v)}, \mathbf{X} = \mathbf{x}) = - \sum_{i,j} P(S_v = j, S_{\text{pa}(v)} = i | \mathbf{X} = \mathbf{x}) \log P(S_v = j | S_{\text{pa}(v)} = i, \mathbf{X} = \mathbf{x})$$

to the downward recursion, with $P(S_v = j | S_{\text{pa}(v)} = i, \mathbf{X} = \mathbf{x}) = \beta_v(j)p_{ij} / \{P(S_v = j)\beta_{\text{pa}(v),v}(i)\}$ and $p_{ij} = P(S_v = j | S_{\text{pa}(v)} = i)$, allows for extracting conditional entropy profiles, while keeping the complexity per iteration of the algorithm in $\mathcal{O}(K^2)$.

3 Applications

Synthetic examples

A two-state HMC family is considered. Its transition probability matrix is parametrized by $\varepsilon = P(S_t = 1 | S_{t-1} = 0) = P(S_t = 0 | S_{t-1} = 1)$, $\varepsilon \in [0, 0.5]$. The initial state distribution π is $P(S_0 = 0) = P(S_0 = 1) = 0.5$. The observation process takes values in $\{0, 1, 2\}$ and the emission distributions (conditional probabilities of observations given the states) are $P(X_t = 0 | S_t = 0) = 1 - p$; $P(X_t = 1 | S_t = 0) = p$; $P(X_t = 1 | S_t = 1) = p$; $P(X_t = 2 | S_t = 1) = 1 - p$ where $p \in [0, 1]$ is an additional parameter.

In a first experiment, p is fixed at 0.5 and the considered observed sequence is $x_t = 1$ for $t = 0, \dots, T - 1$. The smoothed probabilities are $\xi_t(0) = \xi_t(1) = 0.5$ for $t = 0, \dots, T - 1$. Thus, for any value of ε , marginal entropy is $\log 2$ and the sum of these entropies over t is $T \log 2$. In

contrast, global entropy of the hidden state sequence is a strictly increasing function of ε . Its minimum $\log 2$ is reached for $\varepsilon = 0$, whereas its maximum $T \log 2$ is reached for $\varepsilon = 0.5$.

Marginal and conditional entropy profiles are represented in Figure 1 a). For $\varepsilon = 0$, the conditional entropy profile is interpreted as follows: global uncertainty is $\log 2$, which corresponds to uncertainty concerning the first state only. Given this first state, every subsequent state is deterministic and does not contribute to global uncertainty. The marginal entropy profile highlights equiprobability of both states at each time t given the observations. The same statement would hold under an independent mixture assumption for $(X_t)_{t \geq 0}$. Marginal entropy results from uncertainty concerning state S_t due to observing $X_t = x_t$, but also to propagation of uncertainty from past states. As a consequence, marginal entropy cannot be interpreted in terms of local contributions to global uncertainty. In contrast, conditioning by the past state in entropy withdraws the effect of uncertainty propagation.

In a second experiment, the effect of p and ε on global state entropy is assessed by simulating 100 sequences of length $T = 300$ for each $p \in [0, 1]$ and each $\varepsilon \in [0, 0.5]$ on a regular grid with 40×40 points. The mean global entropy over the 100 sequences is represented in Figure 1 b). As expected, entropy increases with the emission distribution overlapping ($p \rightarrow 1$) and as the rows of the transition probability matrix tend to π ($\varepsilon \rightarrow 0.5$), so that maximal entropy $T \log 2$ is obtained in the independence case $\varepsilon = 0.5$ with full overlapping $p = 1.0$.

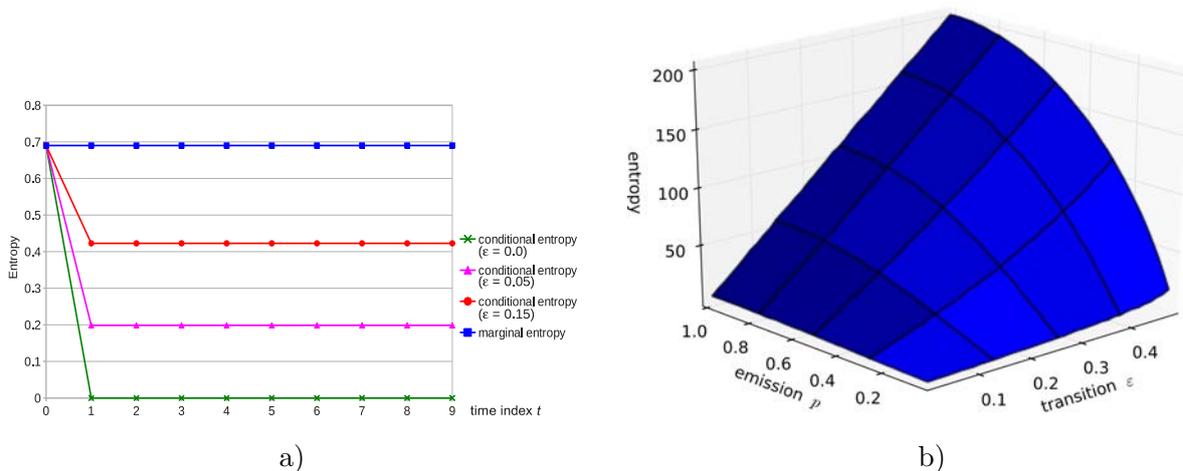


Figure 1: a) Marginal and conditional entropy profiles for a 2-state HMC model with transition probabilities $\varepsilon = 0.0$, $\varepsilon = 0.05$ and $\varepsilon = 0.15$. b) Mean global state entropy for simulated sequences as a function of transition probability ε and emission probability p .

Analysis of the structure of Aleppo pines

The aim of this study was to build a model of the architectural development of Aleppo pines. The dataset contained seven branches of Aleppo pines, issued from different individuals. They were described at the scale of annual shoots v (segment of stem established within a year). Each branch was assimilated with a (mathematical) tree. Each tree vertex v (shoot) was characterized through one observed 5-dimensional vector X_v composed of the: number of growth cycles (from 1 to 3), presence of male sexual organs (binary variable), presence of female sexual organs (binary variable), length in cm, number of branches per tier. The parameters were estimated by maximum likelihood using the EM algorithm. The number of states was chosen by the

ICL-BIC criterion (see Section 4), leading to selection of a 6-state HMT model. The Markov tree is initialized in state 0 with probability one. A summary of the state transitions and an interpretation of the hidden states are provided in Figure 2.

As a first step, profiles of conditional entropies were represented using a colormap (mapping between entropy values and color intensities) – see Figure 3 a). This step highlighted location of the vertices with least ambiguous states along the branch main axes, and location of the vertices with most ambiguous states at the peripheral parts of branches. Then, state profiles were drawn along paths extending from the root vertex to leaf vertices. These paths were chosen so as to contain vertices with high conditional entropies. On the one hand, a detailed analysis of state uncertainty along the paths were obtained by Viterbi upward–downward profiles. This provided local alternative state values to the most likely tree states given by the Viterbi algorithm. On the other hand, the generalized Viterbi algorithm was used to characterize how clusters of neighbor vertices had simultaneous state changes in alternative state configurations. These results highlighted that the paths with most ambiguous states were composed of successions of unbranched, sterile shoots with one single growth cycle.

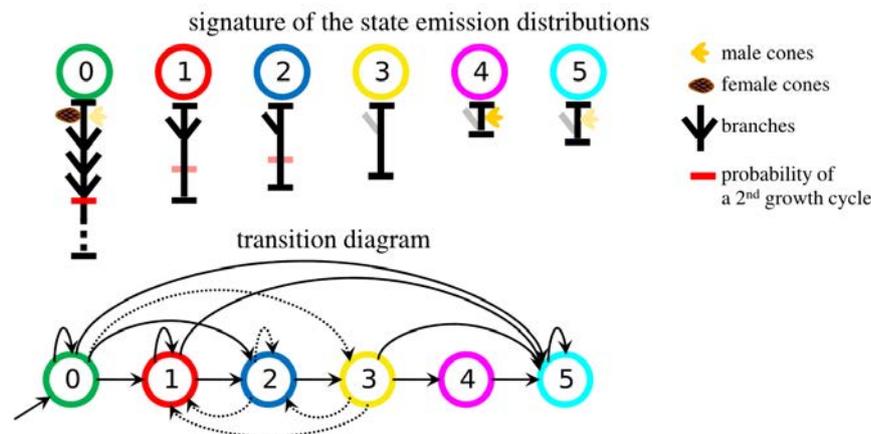


Figure 2: 6-state HMT model: transition diagram and symbolic representation of the state signatures (conditional mean values of the variables given the states, depicted by typical shoots). The separation between growth cycle is represented by a horizontal red segment, which intensity is proportional to the probability of occurrence of a second growth cycle. Dotted arrows correspond to transitions with associated probability < 0.1 . Mean shoot lengths given each state are proportional to segment lengths, except for state 0 (which mean length is slightly more than twice the mean length for state 1).

The application of this methodology is illustrated below on a path containing successive monocyclic, sterile shoots. This path belongs to the fourth individual (for which $H(\mathbf{S}|\mathbf{X} = \mathbf{x}) = 47.5$). It is composed by 5 vertices, referred to as $\{0, \dots, 4\}$. Shoots 0 and 1 are long and highly branched, and thus are in state 0 with probability ≈ 1 (also, shoot 0 is bicyclic). Shoots 2 to 4 are monocyclic and sterile. Shoots 2 and 3 bear one branch, and can be in states 1 or 2 essentially. Shoot 4 is unbranched and from the Viterbi profiles in Figure 3c), it can be in states 2, 3 or 5. This is summarized by the conditional entropy profile in Figure 3b).

This conditional entropy profile can be further interpreted, in relation with mutual information $I(S_u; S_{pa(u)}|\mathbf{X} = \mathbf{x})$. On the one hand, $I(S_1; S_2|\mathbf{X} = \mathbf{x}) = 0$. This results from state S_1 being known. Thus, conditioning by S_1 does not provide further information on its children state S_2 . On the other hand, $I(S_3; S_4|\mathbf{X} = \mathbf{x}) = 0.2$. Uncertainty associated with the posterior

distribution of S_4 is high, since $H(S_4|\mathbf{X} = \mathbf{x}) = 0.67$. However, knowledge of its parent state S_3 would reduce the uncertainty on S_4 : if $S_3 = 1$ then $S_4 = 5$; if $S_3 = 2$ then $S_4 = 2$ (or less likely, $S_4 = 3$) and if $S_3 = 3$ then $S_4 = 5$ (or less likely, $S_4 = 2$).

Using an extension of (2) to subgraphs of \mathcal{T} , the contribution of the vertices of the considered path \mathcal{P} to global state tree entropy can be computed as $\sum_{u \in \mathcal{P}} H(S_u | S_{\text{pa}(u)}, \mathbf{X} = \mathbf{x})$ and is equal to 1.41 in the above example (that is, 0.28 per vertex on average). The global state tree entropy for this individual is 0.24 per vertex, against 0.20 per vertex in the whole dataset. This is explained by the lack of information brought by the observed variables (several successive sterile monocyclic shoots, which can be in states 1, 2, 3 or 5).

The contribution of \mathcal{P} to the global state tree entropy corresponds to the sum of the heights of every point of the profile of conditional entropies in Figure 3b).

Note that the representation of state uncertainty using profiles of posterior state probabilities induces a perception of global uncertainty on the states along \mathcal{P} equivalent to that provided by marginal entropy profile in Figure 3b). The mean marginal state entropy for this individual is 0.37 per vertex, which strongly overestimates the global state tree entropy per vertex (0.24).

4 Concluding remarks

In this work, conditional entropy profiles are proposed to assess both local and global state uncertainty in GHMMs. As shown in the examples, these profiles allow deeper understanding of the local roles of the model parameters, the neighbouring states and the observed data, concerning state uncertainty. These profiles are a valuable tool to analyse alternative state restorations, which may involve zones of connected vertices. Such situations are characterised by high mutual information between connected vertices. Moreover, the examples highlight that the posterior state probability profiles introduce confusion between (i) local state uncertainty due to overlap of emission distributions for different states and (ii) mere propagation of uncertainty from past to future states. Contrary to conditional entropy profiles, they suggest strong local contributions to global state uncertainty in zones where such uncertainty is in fact far more limited.

In the perspective of model selection, entropy may also be useful. If irrelevant states or variables are added to GHMMs, global state entropy is expected to increase. This explains why several model selection criteria based on a compromise between log-likelihood and state entropy were proposed. Among these is the Normalised Entropy Criterion introduced by Celeux & Soromenho (1996) in independent mixture models, and ICL-BIC introduced by McLachlan & Peel (2000, chap. 6). Their generalization to GHMMs is rather straightforward. By favouring models with small state entropy and high log-likelihood, these criteria aim at selecting models such as the uncertainty of the state values is low, whilst achieving good fit to the data.

Bibliography

- [1] CELEUX, G., AND SOROMENHO, G. (1996) *An entropy criterion for assessing the number of clusters in a mixture model*. Classification Journal, **13**, 195–212.
- [2] COVER, T., AND THOMAS, J. (2006) *Elements of Information Theory, 2nd edition*. Hoboken, NJ: Wiley.

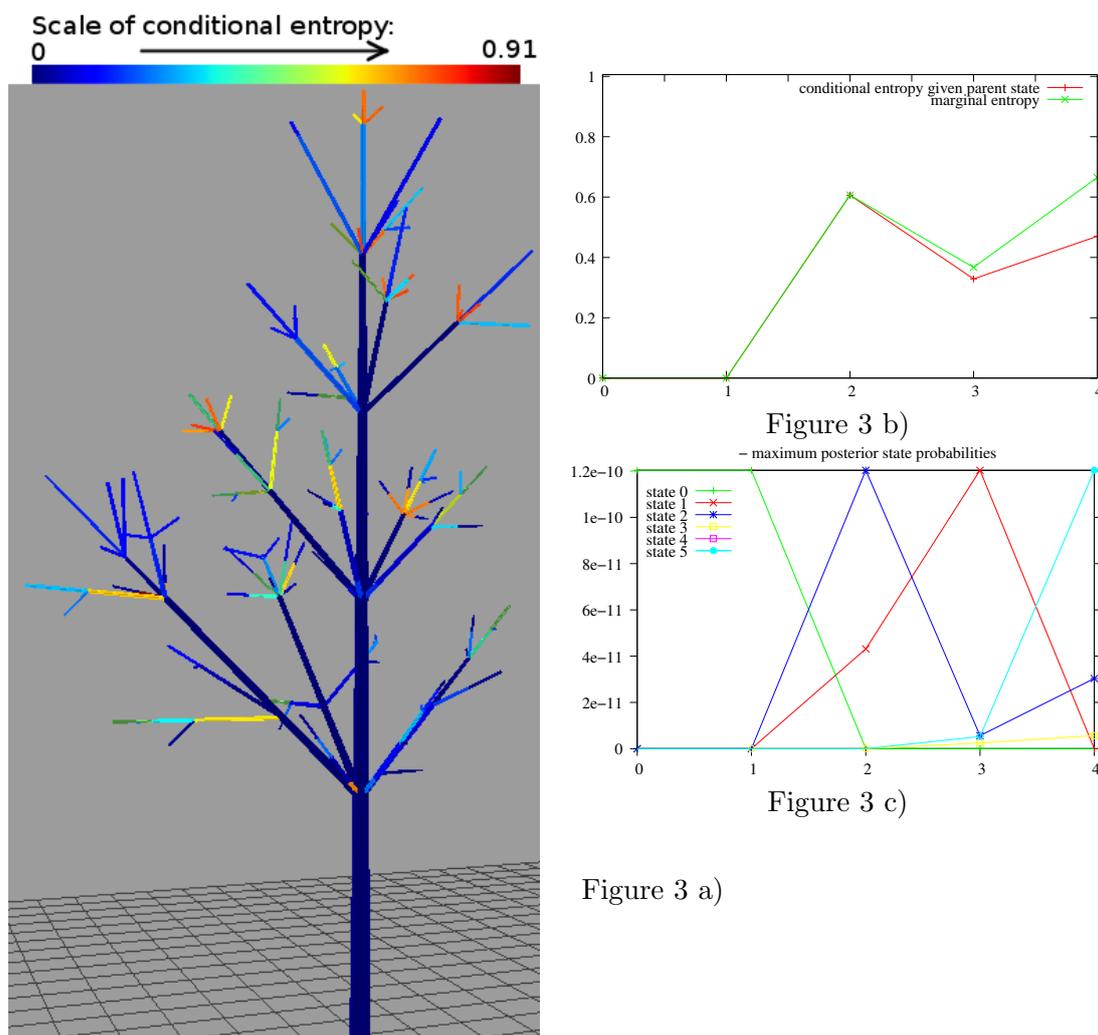


Figure 3: a) Conditional entropy profiles $H(S_u|S_{pa(u)}, \mathbf{X} = \mathbf{x})$ for vertices u associated with one of the seven branches. Blue corresponds to lowest and red to highest conditional entropies. b) Profiles of conditional and of marginal entropies along a path containing mainly sterile monocyclic shoots. c) State tree restoration with the Viterbi upward-downward algorithm.

- [3] DURAND, J.-B., GONÇALVÈS, P., AND GUÉDON, Y. (Sept. 2004) *Computational Methods for Hidden Markov Tree Models – An Application to Wavelet Trees*. IEEE Transactions on Signal Processing, **52**, 9, 2551–2560.
- [4] EPHRAIM, Y., AND MERHAV, N. (June 2002) *Hidden Markov processes*. IEEE Transactions on Information Theory, **48**, 1518–1569.
- [5] GUÉDON, Y. (2007) *Exploring the state sequence space for hidden Markov and semi-Markov chains*. Computational Statistics and Data Analysis, **51**, 5, 2379–2409.

- [6] HERNANDO, D., CRESPI, V., AND CYBENKO, G. (2005) *Efficient computation of the hidden Markov model entropy for a given observation sequence*. IEEE Transactions on Information Theory, **51**, 7, 2681–2685.
- [7] LAURITZEN, S. (1996) *Graphical Models*. Clarendon Press, Oxford, United Kingdom.
- [8] MCLACHLAN, G., AND PEEL, D. (2000) *Finite Mixture Models*. Wiley Series in Probability and Statistics. John Wiley and Sons.

Mixture of regression models with latent variables and sparse coefficient parameters

Shu Kay Ng, *School of Medicine, Griffith Health Institute, Griffith University, Meadowbrook Q4131, Australia, s.ng@griffith.edu.au*

Geoffrey J McLachlan, *Department of Mathematics, University of Queensland, Brisbane Q4072, Australia, g.mclachlan@uq.edu.au*

Abstract. Mixture models have been widely used in marketing research and epidemiology to capture heterogeneity in endogenous latent variables among individuals. However, when collinearity between endogenous latent variables at the component level is present, some component-specific path coefficients will be zero. In this paper, a systematic computational algorithm is developed to identify parameters that need to be constrained to be zero and to address other issues including the initialization procedure, the provision of standard errors of estimates, and the method to determine the number of components. The proposed algorithm is illustrated using simulated data and a real data set concerning emotional behaviour of preschool children.

Keywords. Mixture models, Latent variables, Regression models, Sparse coefficients, EM algorithm

1 Introduction

Regression models involving latent variables (or constructs) are very common in the marketing research and epidemiology [2, 3]. With this approach, simultaneous regression equations are adopted to model the relationships between multiple dependent (endogenous) latent variables and independent (exogenous) latent variables. Let $\boldsymbol{\eta}_j$ and $\boldsymbol{\xi}_j$ denote the vectors of endogenous and exogenous latent variables for the j th individual ($j = 1, \dots, n$), respectively. The “inner” model is specified in terms of q simultaneous regression equations as

$$\mathbf{B}\boldsymbol{\eta}_j + \mathbf{\Gamma}\boldsymbol{\xi}_j = \boldsymbol{\zeta}_j, \quad (1)$$

where \mathbf{B} is a $q \times q$ matrix with q being the number of endogenous latent variables, $\mathbf{\Gamma}$ is a $q \times p$ matrix where p is the number of exogenous latent variables, and $\boldsymbol{\zeta}_j$ is a random vector of residuals. The matrices \mathbf{B} and $\mathbf{\Gamma}$ represent the (path) coefficients relating to the endogenous

and exogenous latent variables, respectively, in the inner model. The relationships between the latent variables and the manifest variables, either reflective indicators or formative measures, are specified in the “outer” model [3]. Estimation of model parameters and values for latent variables can be proceed with two different approaches. The structural equation modelling (SEM) approach attempts to reproduce the covariance matrix of the observed measures, while the partial least squares (PLS) approach focuses on maximizing the variance of the endogenous variables explained by the exogenous variables.

In many real problems, the presence of heterogeneity among individuals in terms of different path coefficients is prevalence. Such kind of heterogeneity is due to different individual perception of latent variables and can be captured in the regression modelling via a finite mixture model approach [3, 10]. With the PLS approach to regression models with latent variables, it is assumed that the endogenous latent variables $\boldsymbol{\eta}_j$ ($j = 1, \dots, n$) come from a mixture of a finite number, say g of multivariate normal distributions in some unknown proportions π_1, \dots, π_g that sum to one:

$$f(\boldsymbol{\eta}_j; \boldsymbol{\Psi}) = \sum_{i=1}^g \pi_i \phi(\boldsymbol{\eta}_j; \boldsymbol{\mu}_{ij}, \boldsymbol{\Sigma}_i) \quad (j = 1, \dots, n), \quad (2)$$

where $\boldsymbol{\mu}_{ij} = (\mathbf{I} - \mathbf{B}_i)\boldsymbol{\eta}_j - \boldsymbol{\Gamma}_i\boldsymbol{\xi}_j$ is the mean vector of the i th component, where \mathbf{I} is an identity matrix, and $\boldsymbol{\Sigma}_i = \text{diag}(\boldsymbol{\sigma}_i^2)$ is a diagonal matrix constructed from the vector $\boldsymbol{\sigma}_i^2$, which represents the variance of the random residuals $\boldsymbol{\zeta}_{ij}$ ($i = 1, \dots, g$). In (2), $\boldsymbol{\Psi}$ is the vector of all the unknown parameters containing π_1, \dots, π_{g-1} and the free parameters in \mathbf{B}_i , $\boldsymbol{\Gamma}_i$, and $\boldsymbol{\Sigma}_i$ for $i = 1, \dots, g$. From (1), the conditional multivariate normal density is given by

$$\phi(\boldsymbol{\eta}_j; \boldsymbol{\mu}_{ij}, \boldsymbol{\Sigma}_i) = \frac{|\mathbf{B}_i|}{\sqrt{(2\pi)^q |\boldsymbol{\Sigma}_i|}} \exp\left\{-\frac{1}{2}(\mathbf{B}_i\boldsymbol{\eta}_j + \boldsymbol{\Gamma}_i\boldsymbol{\xi}_j)^T \boldsymbol{\Sigma}_i^{-1} (\mathbf{B}_i\boldsymbol{\eta}_j + \boldsymbol{\Gamma}_i\boldsymbol{\xi}_j)\right\}, \quad (3)$$

where the superscript T denotes vector transpose.

While mixtures of multivariate normal distributions are generically identifiable (that is, the model is unique up to a permutation of the component labels; see [4, 7]), mixtures of regression models with latent variables arisen from (1) and (2) are not identifiable unless some elements of matrices \mathbf{B}_i and $\boldsymbol{\Gamma}_i$ ($i = 1, \dots, g$) are constrained to zero [3]. In practice, the links between the latent variables represented by simultaneous regression equations in the inner model are usually hypothetical models pre-specified based on a researcher’s own experience. When collinearity between endogenous latent variables at the component level is present, some component-specific path coefficients will be zero. However, the setting up of such parameter constraints at present is somewhat arbitrary. There are also issues of initialization procedure, provision of standard errors of parameter estimates for statistical inference, and determination of the number of components g in the mixture model [7]. In this paper, we tackle these issues by developing a systematic computational algorithm for the implementation of mixtures of regression models with latent variables and sparse coefficient parameters as presented in (1) and (2).

The rest of the paper is organized as follows: Section 2 describes the expectation-maximization (EM) algorithm for the iterative computation of maximum likelihood (ML) estimates of the mixture model and the procedure to identify sparse coefficient parameters. Also, we show how to initialize the algorithm, to obtain standard errors using a bootstrap resampling approach, and to determine the value of g . In Section 3, we present simulation studies to illustrate the applicability of the proposed algorithm in terms of the accuracy of the final model derived and the corresponding estimate biases. We show in Section 4 the application of the proposed method to a real data set. Section 5 ends the paper with further discussion.

2 Algorithm for fitting mixture of sparse regression models

The proposed algorithm applies directly to the scores of endogenous and exogenous latent variables, $\boldsymbol{\eta}_j$ and $\boldsymbol{\xi}_j$, calculated using an iterative scheme of standard PLS on the observed manifest variables with specification based on the constraints of \mathbf{B} and $\mathbf{\Gamma}$ for all individuals ($j = 1, \dots, n$); see [3, 10]. The “aggregate” predictors of \mathbf{B} and $\mathbf{\Gamma}$ estimated in the PLS procedure may also be used to guide the initial estimates for \mathbf{B}_i and $\mathbf{\Gamma}_i$ ($i = 1, \dots, g$) in the mixture model.

Maximum likelihood estimation and parameter constraint

The fitting of the mixture model (2) to latent variables $\boldsymbol{\eta}_j$ and $\boldsymbol{\xi}_j$ ($j = 1, \dots, n$) obtained by PLS can be implemented using ML. An estimate $\hat{\boldsymbol{\Psi}}$ is obtained by solving the log likelihood equation iteratively via the EM algorithm [8]. An appealing property of the EM algorithm is that the likelihood is not decreased after each iteration. Within the EM framework, each individual is conceptualized to have arisen from one of the g components of the mixture model and the unobservable component-indicator vector \mathbf{z}_j is treated as missing data. Precisely, the i th element z_{ij} of \mathbf{z}_j is taken to be one or zero according as the j th individual does or does not come from the i th component ($i = 1, \dots, g$; $j = 1, \dots, n$). On the $(k+1)$ th iteration of the EM algorithm, the E-step computes the so-called Q -function, which is the conditional expectation of the complete-data log likelihood using the current fit for $\boldsymbol{\Psi}$:

$$Q(\boldsymbol{\Psi}; \boldsymbol{\Psi}^{(k)}) = \sum_{i=1}^g \sum_{j=1}^n \tau_{ij}^{(k)} \{\log \pi_i + \log \phi(\boldsymbol{\eta}_j; \boldsymbol{\mu}_{ij}, \boldsymbol{\Sigma}_i)\}, \quad (4)$$

where we simply have to calculate

$$\tau_{ij}^{(k)} = \frac{\pi_i^{(k)} \phi(\boldsymbol{\eta}_j; \boldsymbol{\mu}_{ij}^{(k)}, \boldsymbol{\Sigma}_i^{(k)})}{\sum_{h=1}^g \pi_h^{(k)} \phi(\boldsymbol{\eta}_j; \boldsymbol{\mu}_{hj}^{(k)}, \boldsymbol{\Sigma}_h^{(k)})} \quad (i = 1, \dots, g; j = 1, \dots, n), \quad (5)$$

which is the posterior probability that the j th individual belongs to the i th component of the mixture; see [7].

The M-step updates the estimate of $\boldsymbol{\Psi}$ by the new value $\boldsymbol{\Psi}^{(k+1)}$ of $\boldsymbol{\Psi}$ that maximizes the Q -function with respect to $\boldsymbol{\Psi}$. It can be seen from (4) that the maximization with respect to the mixing proportions and coefficient parameters can be obtained separately as follows:

$$\begin{aligned} \pi_i^{(k+1)} &= \sum_{j=1}^n \tau_{ij}^{(k)} / n & \boldsymbol{\Sigma}_i^{(k+1)} &= \frac{\sum_{j=1}^n \tau_{ij}^{(k)} (\mathbf{B}_i^{(k)} \boldsymbol{\eta}_j + \mathbf{\Gamma}_i^{(k)} \boldsymbol{\xi}_j)^T (\mathbf{B}_i^{(k)} \boldsymbol{\eta}_j + \mathbf{\Gamma}_i^{(k)} \boldsymbol{\xi}_j)}{\sum_{j=1}^n \tau_{ij}^{(k)}} \\ \mathbf{B}_i^{(k+1)} &= \sum_{j=1}^n \tau_{ij}^{(k)} \mathbf{\Gamma}_i^{(k)} \boldsymbol{\xi}_j \boldsymbol{\eta}_j^T \left[\sum_{j=1}^n \tau_{ij}^{(k)} \boldsymbol{\eta}_j \boldsymbol{\eta}_j^T \right]^{-1} & \mathbf{\Gamma}_i^{(k+1)} &= \sum_{j=1}^n \tau_{ij}^{(k)} \mathbf{B}_i^{(k)} \boldsymbol{\eta}_j \boldsymbol{\xi}_j^T \left[\sum_{j=1}^n \tau_{ij}^{(k)} \boldsymbol{\xi}_j \boldsymbol{\xi}_j^T \right]^{-1} \end{aligned} \quad (6)$$

In addition to the parameter constraints specified under the hypothetical model in (1) under (2), extra constraints at the component level may be required in the formulation of the final mixture model when collinearity between some endogenous variables is present. In this paper, we propose the following systematic scheme to determine which additional component-parameters in \mathbf{B}_i ($i = 1, \dots, g$) need to be constrained to be zero:

1. Perform model estimation without any additional constraints;
2. Monitor the log likelihood values at each iteration and the parameter estimates of \mathbf{B}_i ($i = 1, \dots, g$);
3. Determine if the algorithm converges or not (failure to convergence is indicated by either singularity of \mathbf{B}_i or decrease of log likelihood values due to estimate in \mathbf{B}_i , say b_{ilm} , being very close to zero, such as being less than 0.000001 in absolute value⁷);
4. Constrain the parameter b_{ilm} , if convergence fails to achieve in (3), to be zero and then rerun the model estimation;
5. Repeat (2) and (4) to constrain one parameter at a time⁸ until convergence of model estimation is achieved.

Initialization, computation of standard errors, and model selection

With applications where the log likelihood equation has multiple local maxima, the EM algorithm should be implemented from a wide choice of initial parameter values in an attempt to search for all local maxima [7, 8]. The proposed algorithm provides three options to initialize the EM algorithm, where the user can either (a) specify initial estimates of the unknown parameters (such as those guided by estimates obtained by the standard PLS); (b) use random groupings of the data to get initial estimates of the unknown parameters; or (c) run the EM algorithm from different random starts as in (b) and use the set of parameter estimates corresponding to the largest likelihood value as initial values for obtaining the final model.

With the proposed algorithm, the standard errors of the estimates of Ψ are obtained using the bootstrap resampling method with replacement, where the number of bootstrap replications is taken to be 100 [7].

In the absence of any prior information as to the number of components present in the data, we can monitor the increase in log likelihood function as the value of g increases in order to determine an appropriate value of g . At any stage, the choice of $g = g_0$ versus $g = g_0 + 1$ can be made by using some information-based criterion, such as the Bayesian Information Criterion (BIC) [9] or by a bootstrap resampling approach to assess the null distribution (and hence the p-value) of the likelihood ratio test statistic [7]; see also [5] and [6]. There is also the integrated classification likelihood (ICL) criterion [1]. Other criteria for the determination of g , including the Akaike Information Criterion (AIC), the consistent AIC (CAIC), and the entropy measure (EN), have been considered specifically within the marketing research [3, 10]. Comparison of these methods in the general context of mixture models has been reported [7].

3 Simulation experiments

In this section, we study the performance of the proposed computational algorithm for fitting mixtures of sparse regression models. We consider a marketing research setting with $p = 5$ exogenous and $q = 7$ endogenous variables. Let $\xi = (\xi_1, \dots, \xi_5)^T$ and $\eta = (\eta_1, \dots, \eta_7)^T$ be the scores of exogenous and endogenous variables, respectively, with the subscript j that indicates the j th individual dropped, the 7 simultaneous regression equations that define the path model are given by

⁷ Other thresholds close to zero may be used and the choice should not affect the results.

⁸ If constraints in multiple parameters are needed, sensitivity analysis may be used to determine the order.

$$\begin{aligned}
\eta_1 &= \gamma_{11}\xi_1 + \zeta_1; & \eta_5 &= \gamma_{54}\xi_4 + \zeta_5; \\
\eta_2 &= \gamma_{22}\xi_2 + \zeta_2; & \eta_6 &= \gamma_{65}\xi_5 + \zeta_6; \\
\eta_3 &= b_{32}\eta_2 + \zeta_3; & \eta_7 &= b_{74}\eta_4 + b_{75}\eta_5 + b_{76}\eta_6 + \zeta_7, \\
\eta_4 &= b_{41}\eta_1 + b_{43}\eta_3 + \gamma_{43}\xi_3 + \zeta_4;
\end{aligned} \tag{7}$$

which imply that the specifications for \mathbf{B}_i and $\mathbf{\Gamma}_i$ ($i = 1, \dots, g$) are:

$$\mathbf{B}_i = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & -b_{i32} & 1 & 0 & 0 & 0 & 0 \\ -b_{i41} & 0 & -b_{i43} & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & -b_{i74} & -b_{i75} & -b_{i76} & 1 \end{bmatrix} \quad \text{and} \quad \mathbf{\Gamma}_i = \begin{bmatrix} -\gamma_{i11} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & -\gamma_{i22} & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & -\gamma_{i43} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -\gamma_{i54} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & -\gamma_{i65} & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \tag{8}$$

In the simulation experiments, it is assumed that there are $g = 3$ groups of individuals and the total number of individuals is $n = 1000$. Each vector of the exogenous latent variable scores ξ_j ($j = 1, \dots, 1000$) was generated independently from a multivariate normal distribution with mean vector and covariance matrix as

$$\text{Mean} = \begin{pmatrix} -0.063 \\ -0.131 \\ -0.012 \\ 0.080 \\ -0.013 \end{pmatrix} \quad \text{and} \quad \text{Cov.} = \begin{bmatrix} 1.14 & 0.66 & 0.72 & 0.45 & 0.57 \\ 0.66 & 1.19 & 0.53 & 0.29 & 0.43 \\ 0.72 & 0.53 & 1.01 & 0.48 & 0.58 \\ 0.45 & 0.29 & 0.48 & 0.99 & 0.47 \\ 0.57 & 0.43 & 0.58 & 0.47 & 1.01 \end{bmatrix}. \tag{9}$$

The parameter values for Ψ with reference to (8) are given in Table 1; these parameter values are based on a fitted mixture model we have obtained on a real data set. Realizations of component membership were generated in which an individual has a probability of π_i to belong to the i th component ($i = 1, 2, 3$). Given the component membership, realizations of η_j were then generated from the corresponding component density $\phi(\eta_j | \mu_{ij}, \Sigma_i)$ as in (2) under (7).

To illustrate the proposed scheme presented in Section 2 for the constraint of additional component-parameters in \mathbf{B}_i ($i = 1, 2, 3$), we consider collinearity between the seventh η_7 and the fourth η_4 endogenous latent variables in (7) for the first component. This implies that both parameters b_{175} and b_{176} are zero, with a very small σ_{17}^2 ; see Table 1. Using a data set of $n = 1000$ scores generated as above, we first consider a mixture model without any additional constraints on parameters in \mathbf{B}_i (see Equation (8)). The algorithm fails to converge as the estimate of b_{175} has a value smaller than 0.000001. We then consider a model with an additional constraint of $b_{175} = 0$. The algorithm again fails to converge as the estimate of b_{176} has a value smaller than 0.000001. We thus constrain $b_{176} = 0$ as well. This final model with two additional constraints ($b_{175} = 0$ and $b_{176} = 0$) converges.

Ten independent simulation experiments were conducted to assess the generalization performance of the proposed algorithm for fitting mixtures of sparse regression models. Such evaluation is based on the accuracy of the final model derived, the misclassification rate, and the bias of estimates. In all ten experiments, the algorithm identifies the correct final model with two additional constraints in b_{175} and b_{176} (the rate of correctly identifying sparse coefficients is 100%).

Parameter	$i = 1$	$i = 2$	$i = 3$	Parameter	$i = 1$	$i = 2$	$i = 3$
π_i	0.42	0.46	0.12	γ_{i54}	0.63	0.50	0.17
b_{i32}	0.97	0.64	0.67	γ_{i65}	1.02	0.94	-0.80
b_{i41}	0.60	0.14	0.27	σ_{i1}^2	0.07	0.64	0.39
b_{i43}	0.32	0.40	0.32	σ_{i2}^2	0.09	0.89	0.68
b_{i74}	0.87	0.20	0.49	σ_{i3}^2	0.07	0.60	0.44
b_{i75}	0.00	0.27	0.11	σ_{i4}^2	0.02	0.47	0.35
b_{i76}	0.00	0.24	0.21	σ_{i5}^2	0.58	0.82	0.82
γ_{i11}	0.91	0.67	0.74	σ_{i6}^2	0.01	0.01	0.73
γ_{i22}	1.16	0.49	0.29	σ_{i7}^2	1E-6	0.86	0.87
γ_{i43}	0.03	0.39	0.22				

Table 1: Parameter values for a 3-component mixture model (Simulation experiments).

Parameter	$i = 1$	$i = 2$	$i = 3$	Parameter	$i = 1$	$i = 2$	$i = 3$
π_i	-0.001	0.010	-0.009	γ_{i54}	-0.003	0.008	-0.017
b_{i32}	0.001	0.018	-0.043	γ_{i65}	0.001	0.001	-0.048
b_{i41}	-0.002	0.003	0.003	σ_{i1}^2	-0.001	0.006	0.031
b_{i43}	-0.004	0.002	0.010	σ_{i2}^2	0.003	-0.005	0.022
b_{i74}	-0.001	-0.008	0.045	σ_{i3}^2	0.002	-0.003	-0.018
b_{i75}	—	-0.011	0.016	σ_{i4}^2	0.001	-0.005	-0.011
b_{i76}	—	0.001	0.040	σ_{i5}^2	0.006	-0.001	-0.033
γ_{i11}	0.001	-0.006	0.003	σ_{i6}^2	0.001	0.001	0.009
γ_{i22}	0.002	0.011	0.008	σ_{i7}^2	0.000	-0.006	0.067
γ_{i43}	0.001	-0.004	-0.011				

Table 2: Average bias of estimates for a 3-component mixture model (Simulation experiments).

The average misclassification rate is 0.0137. The average bias of estimates are presented in Table 2. It can be seen that no appreciable bias is observed in the estimation of Ψ .

4 Real example: Emotional behaviour of preschool children

This real example is based on the Early Head Start Research and Evaluation (EHSRE) project conducted from 1996 to 2001. The data set is available from the Inter-University Consortium for Political and Social Research (ICPSR) at <http://www.icpsr.umich.edu>. It contains data about 2977 children under 3 years who were randomized to receive designed Early Head Start (EHS) services or to seek their own early childhood care in their community; see, for example, [12].

In the current study, we considered $n = 1498$ individuals with complete observations in eight manifest variables and focus on the conceptual model described in [12] for hypothesized relationships among maternal mental health, parenting stress, parent-child routines, and child emotional development. The endogenous and exogenous latent variables of the hypothetical model are presented in Figure 1. In the inner model, there are $p = 1$ exogenous (maternal mental health) and $q = 3$ endogenous (parenting stress, parent-child routine and child emotional development) latent variables. The 3 simultaneous regression equations that define the path

model are given by

$$\begin{aligned}\eta_1 &= \gamma_{11}\xi_1 + \zeta_1; & \eta_3 &= b_{31}\eta_1 + b_{32}\eta_2 + \zeta_3. \\ \eta_2 &= b_{21}\eta_1 + \zeta_2;\end{aligned}\tag{10}$$

The standard PLS analysis is implemented using the “plspm” package in R [11] to obtain the scores for the 4 latent variables corresponding to the path model presented in Figure 1. The proposed algorithm is then used to fit mixtures of regression models to the scores of the latent variables with $g = 1$ to $g = 5$. No additional parameter constraints are necessary. Using the BIC, we identified two groups of individuals. The larger group ($i = 1$, $n_1 = 1434$) of individuals have all links in the hypothetical inner model significant; see Table 3 for the estimates of the path coefficients. Comparing to the majority, the smaller group ($n_2 = 64$) of individuals have smaller impact from maternal mental health on parenting stress (γ_{211}), and from parenting stress and parent-child routine on child emotional development (b_{231} and b_{232}). A post-hoc analysis finds that these two groups are significantly different in RACE (p -value = 0.001; see Table 3), but not in the program allocated, child gender, child overweight indicator, and maternal age at birth.

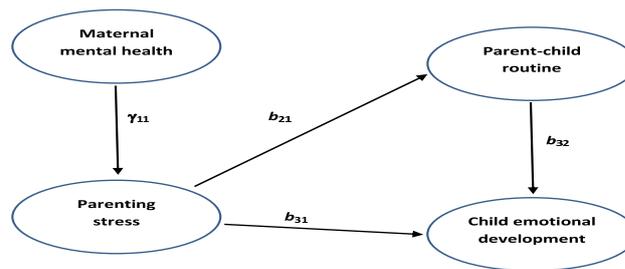


Figure 1: Hypothetical inner model relating maternal mental health, parenting stress, parent-child routines, and child emotional development

Group	b_{i21}	b_{i31}	b_{i32}	γ_{i11}	Race = Hispanic
$i = 1$	-0.178 (0.032)	-0.194 (0.033)	0.142 (0.028)	0.423 (0.021)	312/1357* (23.0%)
$i = 2$	-0.159 (0.039)	-0.074 (0.035)	0.071 (0.062)	0.237 (0.105)	26/61* (42.6%)

Table 3: Estimates (standard errors) of path coefficients for a 2-component mixture model and proportion of Hispanic children (* Missing data exist in both groups).

5 Discussion

We have developed a computational algorithm for fitting mixtures of regression models with latent variables and sparse coefficient parameters. The algorithm adopts a systematic scheme to determine which additional component-parameters in the matrices of path coefficients \mathbf{B}_i need to be constrained to be zero. Simulated and real data sets have been used to illustrate

the applicability of the proposed algorithm. The method can be readily adopted for component distributions that are not multivariate normal.

Acknowledgement

This work was supported by a grant from the Australian Research Council. The authors are grateful to the Inter-University Consortium for Political and Social Research (ICPSR) for providing the Early Head Start Research and Evaluation (EHSRE) project data (ICPSR ID 3804) for the illustration described in Section 4. The findings and views reported in this paper are those of the authors and should not be attributed to either the ICPSR or the EHSRE project.

Bibliography

- [1] Biernacki, C., Celeux, G. and Govaert, G. (1998) *Assessing a mixture model for clustering with the integrated classification likelihood*. IEEE Transactions on Pattern Analysis and Machine Intelligence, **22**, 719–725.
- [2] Dollman, J., Ridley, K., Magarey, A., Martin, M. and Hemphill, E. (2007) *Dietary intake, physical activity and TV viewing as mediators of the association of socioeconomic status with body composition: a cross-sectional analysis of Australian youth*. International Journal of Obesity, **31**, 45–52.
- [3] Hahn, C., Johnson, M.D., Herrmann, A. and Huber, F. (2002) *Capturing customer heterogeneity using a finite mixture PLS approach*. Schmalenbach Business Review, **54**, 243–269.
- [4] Jones, P.N. and McLachlan, G.J. (1992) *Fitting finite mixture models in a regression context*. Australian Journal of Statistics, **34**, 233–240.
- [5] Keribin, C. (2000) *Consistent estimation of the order of mixture models*. Sankhyā: The Indian Journal of Statistics, **62**, 49–66.
- [6] Leroux, B.G. (1992) *Consistent estimation of a mixing distribution*. Annals of Statistics, **20**, 1350–1360.
- [7] McLachlan, G.J. and Peel, D. (2000) *Finite Mixture Models*. New York: Wiley.
- [8] Ng, S.K. (2013) *Recent developments in expectation-maximization methods for analyzing complex data*. WIREs Computational Statistics, **5**, 415–431.
- [9] Ng, S.K. and McLachlan, G.J. (2014) *Mixture models for clustering multilevel growth trajectories*. Computational Statistics and Data Analysis, **71**, 43–51.
- [10] Ringle, C.M., Wende, S. and Will, A. (2010) *Finite mixture partial least squares analysis: Methodology and numerical examples*. In: Handbook of Partial Least Squares, V.E. Vinzi, W.W. Chin, J. Henseler and H. Wang (Eds.). Heidelberg: Springer, pp. 195–218.
- [11] Sanchez, G. (2013) *PLS Path Modeling with R*. Berkeley: Trowchez Editions. Available from: http://www.gastonsanchez.com/PLS_Path_Modeling_with_R.pdf.

- [12] Zajicek-Farber, M.L., Mayer, L.M. and Daughtery, L.G. (2012) *Connections among parental mental health, stress, child routines, and early emotional behavioral regulation of preschool children in low-income families*. *Journal of the Society for Social Work and Research*, **3**, 31–50.

Tensor polyadic decomposition for antenna array processing

Souleymen Sahnoun, *Gipsa-Lab, CNRS, Grenoble France*,
Pierre Comon, *Gipsa-Lab, CNRS, Grenoble France*, `firstname.lastname@gipsa-lab.fr`

Abstract. In the present framework, a tensor is understood as a multi-way array of complex numbers indexed by three (or more) indices. The decomposition of such tensors into a sum of decomposable (i.e. rank-1) terms is called “Polyadic Decomposition” (PD), and qualified as “canonical” (CPD) if it is unique up to trivial indeterminacies. The idea is to use the CPD to identify the location of radiating sources in the far-field from several sensor subarrays, deduced from each other by a translation in space. The main difficulty of this problem is that noise is present, so that the measurement tensor must be fitted by a low-rank approximate, and that the infimum of the distance between the two is not always reached.

Our contribution is three-fold. We first propose to minimize the latter distance under a constraint ensuring the existence of the minimum. Next, we compute the Cramér-Rao bounds related to the localization problem, in which nuisance parameters are involved (namely the translations between subarrays). Then we demonstrate that the CPD-based localization algorithm performs better than ESPRIT when more than 2 subarrays are used, performances being the same for 2 subarrays. Some inaccuracies found in the literature are also pointed out.

Keywords. multi-way array ; localization ; antenna array processing ; tensor decomposition ; low-rank approximation ; complex Cramer-Rao bounds

1 Introduction

The goal is to estimate the Directions of Arrival (DoA) of R narrow-band radiating sources, which impinge on an array of sensors, formed of L identical subarrays of K sensors each. Subarrays do not need to be disjoint, but must be distinct. The hypotheses are [1, 2, 3]: (H1) sources are in the far-field, so that waves are plane; (H2) taking one subarray as reference, every subarray is deduced from the reference one by an unknown translation in space, defined by some vector $\boldsymbol{\delta}_\ell$ of \mathbb{R}^3 , $1 < \ell \leq L$, $\boldsymbol{\delta}_1 \stackrel{\text{def}}{=} \mathbf{0}$; (H3) measurements are recorded on each sensor k of each subarray ℓ and for various time samples m , $1 \leq m \leq M$. In hypothesis (H2), the fact that translations are not exactly known is legitimate, if subarrays are arranged far away from each other, or when their location is changing with time [4]. With these hypotheses, the observation model below

can be assumed [2, 3, 4]:

$$\mathbf{Z}(k, \ell, m) = \sum_{r=1}^R A_{kr} B_{\ell r} C_{mr} + N(k, \ell, m) \quad (1)$$

where $N(k, \ell, m)$ is a measurement/background noise that will be assumed normally distributed. In addition, (H4) parameters $\{A_{kr}, 1 \leq k \leq K, 1 \leq r \leq R\}$ are complex numbers of unit modulus, and $B_{1r} = 1 = A_{1r}, \forall r$. In the present framework, we shall consider subarrays that are formed of equispaced sensors, so that the following can also be assumed (H5):

$$\exists \psi_r : A_{kr} = \exp(j\pi(k-1) \cos \psi_r) \quad (2)$$

where ψ_r is the so-called Direction of Arrival (DoA) of the r th source as illustrated in Figure 1, and $j = \sqrt{-1}$. Space is lacking to explain the physical context, but further details can be found in [1, 2, 3, 4]. The literature is abundant about DoA estimation, but most approaches have been based on second-order moments; see *e.g.* [5] and references therein. On the other hand, the use of space diversity via more than two subarrays is much more recent, and is due to [2]. The key originality therein is that the approach is not based on moments but proceeds by direct parameter estimation from the data.

A key ingredient in this problem is the use of complex random variables, which turn out to be very useful because the formalism is much simpler when working in baseband with complex envelopes of transmitted signals. Among the useful ingredients, we have at disposal matrix differentiation [6], Kronecker and tensor calculus [7, 8], complex differentiation and the derivation of complex Cramér-Rao bounds (see Section 3).

Notation. \mathbb{R} and \mathbb{C} designate the real and complex fields, respectively. Bold lower case letters, *e.g.* \mathbf{z} , always denote column vectors, whereas arrays with 2 indices or more are denoted by bold uppercase symbols, *e.g.* \mathbf{V} or \mathbf{Z} . Array entries are scalar numbers and are denoted in plain font, *e.g.* z_i, V_{ij} or Z_{klm} . The gradient of a p -dimensional function $\mathbf{f}(\mathbf{x})$ with respect to a n -dimensional variable \mathbf{x} is the $p \times n$ matrix $[\partial \mathbf{f} / \partial \mathbf{x}]_{ij} = \partial f_i / \partial x_j$.

2 Tensor formalism and constrained optimization

In this framework, what is meant by *tensor* is just a multi-way array of coordinates; this is not restrictive as long as the coordinate system is fixed [8]. The noiseless part of (1) is a sum of *decomposable tensors*, whose coordinates are of the form $D_{klm} = a_k b_\ell c_m$. Any tensor can be decomposed into a sum of decomposable tensors, and the minimal number of terms necessary to obtain an exact decomposition is called *tensor rank*. Hence nonzero decomposable tensors have a rank equal to 1. Because of the presence of noise, the best rank- R tensor approximate of \mathbf{Z} needs to be found, for instance in the sense of the Frobenius norm, which is consistent with the log-likelihood (11) in the presence of additive Gaussian noise. However, as pointed out in [3, 8, 4] and references therein, the infimum of

$$\Upsilon(\mathbf{A}, \mathbf{B}, \mathbf{C}) \stackrel{\text{def}}{=} \left\| \mathbf{Z} - \sum_{r=1}^R \mathbf{a}_r \otimes \mathbf{b}_r \otimes \mathbf{c}_r \right\|^2$$

may not be reached. Here, $\mathbf{a}_r, \mathbf{b}_r, \mathbf{c}_r$ denote the columns of matrices $\mathbf{A}, \mathbf{B}, \mathbf{C}$ defined in (1), respectively, and \otimes is the tensor outer product.

It has been proved in [3] that a sufficient condition ensuring existence of the best rank- R approximation is that

$$\mu_A \mu_B \mu_C < \frac{1}{R-1} \tag{3}$$

where $\mu_A = \sup_{k \neq \ell} |\mathbf{a}_k^H \mathbf{a}_\ell|$, $\mu_B = \sup_{k \neq \ell} |\mathbf{b}_k^H \mathbf{b}_\ell|$, $\mu_C = \sup_{k \neq \ell} |\mathbf{c}_k^H \mathbf{c}_\ell|$.

Therefore, the following *differentiable* constraint, proposed in [4], can be imposed:

$$\mathcal{C}_\rho \stackrel{\text{def}}{=} 1 - R + \mu(\mathbf{A}, \rho)^{-1} \mu(\mathbf{B}, \rho)^{-1} \mu(\mathbf{C}, \rho)^{-1} > 0, \quad \mu(\mathbf{A}, \rho) \stackrel{\text{def}}{=} \left(\sum_{p < q} |\mathbf{a}_p^H \mathbf{a}_q|^{2\rho} \right)^{1/2\rho} \tag{4}$$

In fact, the inequality between L^p norms

$$\|\mathbf{x}\|_\infty = \max_k \{x_k\} \leq \|\mathbf{x}\|_p \stackrel{\text{def}}{=} \left(\sum_k x_k^p \right)^{1/p}, \quad \forall x_k \in \mathbb{R}^+, \quad p \geq 1,$$

guarantees that constraint (4) implies condition (3). In practice, the following penalized objective function has been minimized in subsequent computer simulations:

$$\Upsilon(\mathbf{A}, \mathbf{B}, \mathbf{C}) + \eta \exp(-\gamma \mathcal{C}_\rho(\mathbf{x})) \tag{5}$$

with $\rho = 13$, $10^{-6} \leq \eta \leq 1$, $\gamma = 5$.

3 Complex Cramér-Rao bounds

When parameters are complex, expressions of Cramér-Rao bounds (CRB) depend on the definition of the complex derivative. Since a real function is never holomorphic (unless it is constant) [9], this definition is necessary; this has been overlooked in [10]. Originally, the derivative of a real function $\mathbf{h}(\boldsymbol{\theta}) \in \mathbb{R}^p$ with respect to a complex variable $\boldsymbol{\theta} \in \mathbb{C}^n$, $\boldsymbol{\theta} = \boldsymbol{\alpha} + j\boldsymbol{\beta}$, $\boldsymbol{\alpha}, \boldsymbol{\beta} \in \mathbb{R}^n$, has been defined as the $p \times n$ matrix [9]:

$$\frac{\partial \mathbf{h}}{\partial \boldsymbol{\theta}} \stackrel{\text{def}}{=} \frac{\partial \mathbf{h}}{\partial \boldsymbol{\alpha}} + j \frac{\partial \mathbf{h}}{\partial \boldsymbol{\beta}}$$

Even if the numerical results are independent of the definition assumed for theoretical calculations, we shall subsequently assume the definition proposed in [11], for consistency with [12]:

$$\frac{\partial \mathbf{h}}{\partial \boldsymbol{\theta}} \stackrel{\text{def}}{=} \frac{1}{2} \frac{\partial \mathbf{h}}{\partial \boldsymbol{\alpha}} - \frac{j}{2} \frac{\partial \mathbf{h}}{\partial \boldsymbol{\beta}} \tag{6}$$

With this definition, one has for instance that $\partial \boldsymbol{\alpha} / \partial \boldsymbol{\theta} = \frac{1}{2} \mathbf{I}$, and $\partial \boldsymbol{\beta} / \partial \boldsymbol{\theta} = -\frac{j}{2} \mathbf{I}$. This is a key difference with [9], where we had instead: $\partial \boldsymbol{\alpha} / \partial \boldsymbol{\theta} = \mathbf{I}$, and $\partial \boldsymbol{\beta} / \partial \boldsymbol{\theta} = j \mathbf{I}$. Assume that parameter $\boldsymbol{\theta}$ is wished to be estimated from an observation \mathbf{z} , of probability distribution $\mathcal{L}(\mathbf{z}; \boldsymbol{\theta})$, and denote $\mathbf{s}(\mathbf{z}; \boldsymbol{\theta})$ the score function. Then we have for any function $\mathbf{h}(\boldsymbol{\theta}) \in \mathbb{R}^p$:

$$\mathbb{E}\{\mathbf{h}(\mathbf{z}) \mathbf{s}(\mathbf{z}; \boldsymbol{\theta})^T\} = \frac{\partial}{\partial \boldsymbol{\theta}} \mathbb{E}\{\mathbf{h}(\mathbf{z})\}, \quad \text{with } \mathbf{s}(\mathbf{z}; \boldsymbol{\theta})^T \stackrel{\text{def}}{=} \frac{\partial}{\partial \boldsymbol{\theta}} \log \mathcal{L}(\mathbf{z}; \boldsymbol{\theta}) \tag{7}$$

This is a direct consequence of the fact that $\mathbf{E}\{\mathbf{s}\} = \mathbf{0}$, valid if derivation with respect to $\boldsymbol{\theta}$ and integration with respect to $\Re(\mathbf{z})$ and $\Im(\mathbf{z})$ can be permuted. Now let $\mathbf{t}(\mathbf{z})$ be an unbiased estimator of $\boldsymbol{\theta}$. Then, following [9], one can prove that $\mathbf{E}\{\mathbf{t}\mathbf{s}^\top\} = \mathbf{E}\{(\mathbf{t} - \boldsymbol{\theta})\mathbf{s}^\top\} = \mathbf{I}$ and $\mathbf{E}\{\mathbf{t}\mathbf{s}^H\} = \mathbf{0}$. Finally, by expanding the covariance matrix of the random vector $(\mathbf{t} - \boldsymbol{\theta}) - \mathbf{F}^{-1}\mathbf{s}^*$, one readily obtains that:

$$\mathbf{V} \geq \mathbf{F}^{-1}, \quad \text{with } \mathbf{V} \stackrel{\text{def}}{=} \mathbf{E}\{(\mathbf{t} - \boldsymbol{\theta})(\mathbf{t} - \boldsymbol{\theta})^H\} \text{ and } \mathbf{F} \stackrel{\text{def}}{=} \mathbf{E}\{\mathbf{s}^* \mathbf{s}^\top\} \quad (8)$$

Note that the definition of the Fisher information matrix is the complex conjugate of that of [9], because of a different definition of the complex derivation (and hence a different definition of the complex score function).

4 Cramér-Rao bounds of the localization problem

In the presence of R sources, the observations can be stored in a three-way array unfolded in vector form:

$$\mathbf{z} = \sum_{r=1}^R \mathbf{a}_r \boxtimes \mathbf{b}_r \boxtimes \mathbf{c}_r + \mathbf{n}, \quad \mathbf{z} \in \mathbb{C}^{KLM} \quad (9)$$

where \boxtimes denotes the Kronecker product, and the additive noise \mathbf{n} is assumed to follow a circularly-symmetric complex normal distribution. Let

$$\boldsymbol{\theta} = \underbrace{[\psi_1, \dots, \psi_R]}_{\boldsymbol{\psi}} \underbrace{[\bar{\mathbf{b}}_1^\top, \dots, \bar{\mathbf{b}}_R^\top, \mathbf{c}_1^\top, \dots, \mathbf{c}_R^\top]}_{\boldsymbol{\xi}} \underbrace{[\bar{\mathbf{b}}_1^H, \dots, \bar{\mathbf{b}}_R^H]}_{\boldsymbol{\xi}^*} \quad (10)$$

denote the unknown parameter vector, where $\bar{\mathbf{b}}_r \stackrel{\text{def}}{=} [B_{2,r}, \dots, B_{L,r}]^\top$. It is useful to include both $\boldsymbol{\xi}$ and $\boldsymbol{\xi}^*$, in case the distribution of the estimate $\hat{\boldsymbol{\xi}}$ is not circularly-symmetric, *i.e.* $\mathbf{E}\{\hat{\boldsymbol{\xi}}\hat{\boldsymbol{\xi}}^\top\} \neq \mathbf{0}$. The aim here is to derive the CRB of the parameters in $\boldsymbol{\theta}$. The CRB for factor matrices have been computed in [12]. However, it should be emphasized that, unlike [12], no assumption is needed on the elements of matrix \mathbf{C} to derive the CRB. In fact, assuming that the first row of \mathbf{A} and \mathbf{B} is fixed to $[1, \dots, 1]_{1 \times R}$ is sufficient. The latter assumption is satisfied in the considered array configuration (hypothesis: H4). The log-likelihood then takes the form:

$$\log \mathcal{L}(\mathbf{z}, \boldsymbol{\theta}) = -KLM \log(\sigma^2 \pi) - \frac{1}{\sigma^2} (\mathbf{z} - \boldsymbol{\mu}(\boldsymbol{\theta}))^H (\mathbf{z} - \boldsymbol{\mu}(\boldsymbol{\theta})) \quad (11)$$

where $\boldsymbol{\mu}(\boldsymbol{\theta})$ is the noise free part of \mathbf{z} . The CRB for unbiased estimation of the complex parameters $\boldsymbol{\theta}$ is equal to the inverse of the Fisher information matrix \mathbf{F} , defined in equation (8). Then, a straightforward calculation yields:

$$\mathbf{s}^\top = \frac{1}{\sigma^2} \left[\mathbf{n}^\top \frac{\partial \boldsymbol{\mu}^*}{\partial \boldsymbol{\theta}} + \mathbf{n}^H \frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{\theta}} \right] \quad (12)$$

where $\mathbf{n} = \mathbf{z} - \boldsymbol{\mu}$. By substituting the score function \mathbf{s} by its expression, and since $\mathbf{E}\{\mathbf{nn}^H\} = \sigma^2 \mathbf{I}_{KLM}$ and $\mathbf{E}\{\mathbf{nn}^\top\} = \mathbf{0}$, the Fisher information matrix can be written as:

$$\mathbf{F} = \frac{1}{\sigma^2} \left[\left(\frac{\partial \boldsymbol{\mu}^*}{\partial \boldsymbol{\theta}} \right)^H \left(\frac{\partial \boldsymbol{\mu}^*}{\partial \boldsymbol{\theta}} \right) + \left(\frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{\theta}} \right)^H \left(\frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{\theta}} \right) \right] \quad (13)$$

Since parameters in $\boldsymbol{\psi}$ are real and those in $\boldsymbol{\xi}$ are complex, a first writing of the derivatives in (13) is:

$$\frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{\theta}} = \begin{bmatrix} \frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{\psi}}, & \frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{\xi}}, & \mathbf{0} \end{bmatrix} \quad \text{and} \quad \frac{\partial \boldsymbol{\mu}^*}{\partial \boldsymbol{\theta}} = \begin{bmatrix} \left(\frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{\psi}}\right)^*, & \mathbf{0}, & \left(\frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{\xi}}\right)^* \end{bmatrix} \quad (14)$$

Therefore, the Fisher information matrix becomes:

$$\mathbf{F} = \frac{1}{\sigma^2} \begin{bmatrix} 2 \operatorname{Re} \{ \mathbf{G}_{11} \} & \mathbf{G}_{12} & \mathbf{G}_{12}^* \\ \mathbf{G}_{12}^H & \mathbf{G}_{22} & \mathbf{0} \\ \mathbf{G}_{12}^T & \mathbf{0} & \mathbf{G}_{22}^* \end{bmatrix} \quad (15)$$

where $\mathbf{G}_{ij} = \left(\frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{\theta}_i}\right)^H \left(\frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{\theta}_j}\right)$, $(i, j) \in \{1, 2\} \times \{1, 2\}$, $\boldsymbol{\theta}_1 = \boldsymbol{\psi}$ and $\boldsymbol{\theta}_2 = \boldsymbol{\xi}$. (16)

In view of (15), it is clear that the introduction of $\boldsymbol{\xi}^*$ in the parameter vector was not necessary. With a non circular complex gaussian noise, this would not have been the case. To complete the calculation of \mathbf{F} , it remains to give partial derivative expressions of $\boldsymbol{\mu}$ with respect to $\boldsymbol{\psi}$ and $\boldsymbol{\xi}$.

Derivatives of $\boldsymbol{\mu}$ with respect to $\boldsymbol{\psi}$

Using the chain rule we have

$$\frac{\partial \boldsymbol{\mu}}{\partial \psi_f} = \left(\frac{\partial \boldsymbol{\mu}}{\partial \mathbf{a}_f^T}\right) \left(\frac{\partial \mathbf{a}_f^T}{\partial \psi_f}\right) \quad (17)$$

and $[\partial \boldsymbol{\mu} / \partial \mathbf{a}_f^T]$ can be computed using complex derivative formulas. Then, we obtain:

$$\frac{\partial \boldsymbol{\mu}}{\partial \mathbf{a}_f^T} = \mathbf{I}_K \boxtimes \mathbf{b}_f \boxtimes \mathbf{c}_f \in \mathbb{C}^{KLM \times K}, \quad 1 \leq f \leq R. \quad (18)$$

To calculate $[\partial \mathbf{a}_f^T / \partial \psi_f]$, we use the expressions of the considered sensor array configuration, namely equation (2), which yields:

$$\frac{\partial \mathbf{a}_f^T}{\partial \psi_f} = -j\pi \sin \psi_f (\mathbf{a}_f \boxminus \mathbf{v}_K) \quad (19)$$

where $\mathbf{v}_K = [0, 1, \dots, K - 1]^T$. By substituting (18) and (19) in (17), we get

$$\frac{\partial \boldsymbol{\mu}}{\partial \psi_f} = -j\pi \sin \psi_f (\mathbf{I}_K \boxtimes \mathbf{b}_f \boxtimes \mathbf{c}_f) (\mathbf{a}_f \boxminus \mathbf{v}_K) \stackrel{\text{def}}{=} \boldsymbol{\phi}_{\psi_f} \quad (20)$$

and $\frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{\psi}} = [\boldsymbol{\phi}_{\psi_1}, \dots, \boldsymbol{\phi}_{\psi_R}] \in \mathbb{C}^{KLM \times R}$ (21)

Derivatives of $\boldsymbol{\mu}$ with respect to $\boldsymbol{\xi}$

Taking partial derivatives of $\boldsymbol{\mu}$ with respect to $\bar{\mathbf{b}}_f^\top$ and \mathbf{c}_f^\top , we obtain:

$$\frac{\partial \boldsymbol{\mu}(\boldsymbol{\theta})}{\partial \bar{\mathbf{b}}_f^\top} = (\mathbf{a}_f \boxtimes \mathbf{I}_{LM})(\mathbf{I}_L \boxtimes \mathbf{c}_f) \mathbf{J}_L \stackrel{\text{def}}{=} \boldsymbol{\phi}_{\bar{\mathbf{b}}_f} \in \mathbb{C}^{KLM \times (L-1)} \quad (22)$$

$$\frac{\partial \boldsymbol{\mu}(\boldsymbol{\theta})}{\partial \mathbf{c}_f^\top} = \mathbf{a}_f \boxtimes \mathbf{b}_f \boxtimes \mathbf{I}_M \stackrel{\text{def}}{=} \boldsymbol{\phi}_{\mathbf{c}_f} \in \mathbb{C}^{KLM \times M} \quad (23)$$

where $\mathbf{J}_L = [\mathbf{0}_{(L-1),1} \ \mathbf{I}_{L-1}]^\top \in \mathbb{C}^{L \times (L-1)}$ is a selection matrix. To sum up,

$$\frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{\xi}} = [\boldsymbol{\phi}_{\bar{\mathbf{b}}_1}, \dots, \boldsymbol{\phi}_{\bar{\mathbf{b}}_R}, \boldsymbol{\phi}_{\mathbf{c}_1}, \dots, \boldsymbol{\phi}_{\mathbf{c}_R}] \in \mathbb{C}^{KLM \times R(L+M-1)} \quad (24)$$

DoA Cramér-Rao bound

The CRB related to DoAs only is obtained as the first leading $R \times R$ block in matrix \mathbf{F}^{-1} , where \mathbf{F} is defined in (15). Doing this assumes that translations $\boldsymbol{\delta}_\ell$ are nuisance parameters, *i.e.* unknown but not of interest. This realistic context has been overlooked in the literature.

5 Computer results

To evaluate the efficiency of the proposed method, we compare its performances to two other algorithms, ESPRIT and MUSIC [13, 14]. The performance criterion is the *total* mean square error (total MSE) of the DoA: $\frac{1}{RN} \sum_{r=1}^R \sum_{n=1}^N (\hat{\psi}_{r,n} - \psi_r)^2$ where $\hat{\psi}_{r,n}$ is the estimated DoA at the n -th Monte-Carlo trial and N is the number of trials. The deterministic CRB computed in the previous section is reported as a benchmark. The considered scenario on which the proposed algorithm is tested can be of interest in numerous applications, where translations $\boldsymbol{\delta}_\ell$ are unknown. Note that the CRB of the DoA where locations of all sensors are known can be found in [13, 14]. The three examples we study in this section are reported in the table below:

	Subarrays	Translations	DoA
Example 1	$L = 2$	$\boldsymbol{\delta}_2 = [0, 25\lambda, 0]^\top$	$40^\circ, 64^\circ, 83^\circ$
Example 2	$L = 3$	$\boldsymbol{\delta}_2 = [0, 25\lambda, 0]^\top, \boldsymbol{\delta}_3 = [0, 37.5\lambda, 5\lambda]^\top$	$40^\circ, 64^\circ, 83^\circ$
Example 3	$L = 3$	$\boldsymbol{\delta}_2 = [0, 25\lambda, 0]^\top, \boldsymbol{\delta}_3 = [0, 37.5\lambda, 5\lambda]^\top$	$7^\circ, 64^\circ, 83^\circ$

where $\lambda = \omega/2\pi\varsigma$ is the wavelength and ς the wave celerity. In all examples, each subarray is an ULA array of 4-element with half-wavelength spacing (see Figure 1), and the narrowband source signals have the same power.

In all experiments, $M = 200$ time samples are used, and 200 Monte-Carlo simulations are run for each SNR level. Figures 2, 3 and 4 report the MSE of the DoA obtained in examples 1, 2 and 3, respectively.

Example 1. This experiment shows that: (i) the proposed CP algorithm exhibits the same performances as ESPRIT, which makes sense, (ii) MUSIC performs the best, but exploits more information, namely the exact knowledge of sensor locations, whereas this information is actually not available in the present scenario. Hence MUSIC performances just serve as a reference.

Example 2. This experiment shows that the proposed algorithm yields better results than ESPRIT. The reason is that ESPRIT uses at most two subarrays, whereas the proposed algorithm uses all of them. Again, MUSIC is reported just as a reference benchmark.

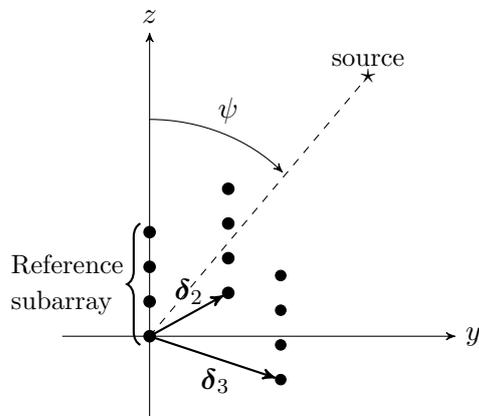


Figure 1: One source ($R = 1$) radiating on a sensor array with $L = 3$ subarrays.

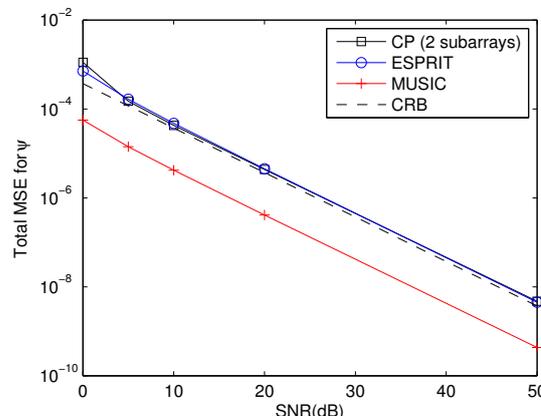


Figure 2: Total DoA error versus SNR, with $L = 2$ subarrays, $\psi = [40^\circ, 65^\circ, 83^\circ]$.

Example 3. This experiment shows the same results as in example 2, except for an increase in MSE at low SNR, which is due to the direction of arrival $\psi = 7^\circ$. Actually, for an ULA, the source localization accuracy degrades as the DoA come closer to the end-fire, so that the so-called *threshold region* (which always exists at low SNR) becomes visible.

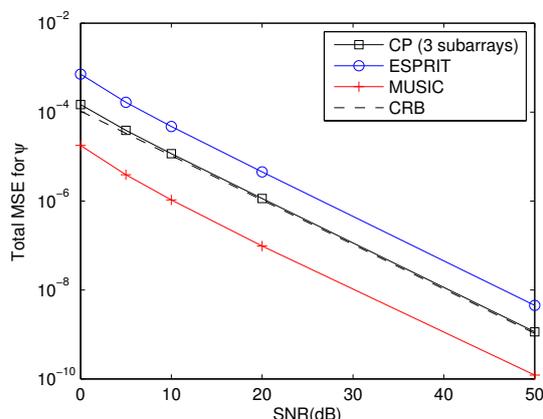


Figure 3: Total DoA error versus SNR, with $L = 3$ subarrays, $\psi = [40^\circ, 65^\circ, 83^\circ]$.

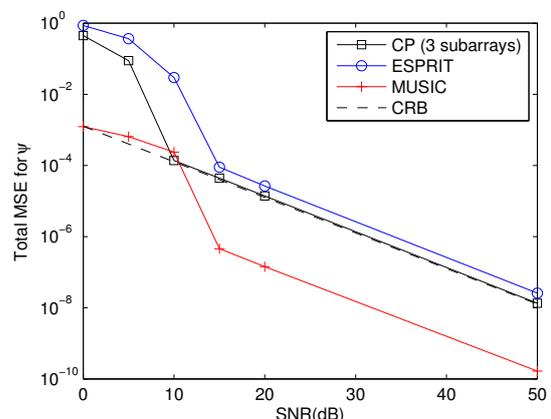


Figure 4: Total DoA error versus SNR, with $L = 3$ subarrays, $\psi = [7^\circ, 65^\circ, 83^\circ]$.

6 Conclusion

The source localization problem is taken as an illustration of the interest in resorting to CP tensor decomposition. We took the opportunity of this illustration to emphasize the usefulness of complex formalism when computing the CRB. Our contributions include the computation of CRB of DoAs when space translations are unknown (section 4), and an original algorithm to compute the CP decomposition under a constraint ensuring the existence of the best low-rank approximate (section 2).

Some inaccuracies on this subject may be found in the literature: (i) in [10], functions of the complex variable are assumed holomorphic, whereas real functions never are; (ii) in [12], CRB are derived, but without assuming that factor matrix \mathbf{A} is parameterized by angles of arrival; moreover, additional constraints have been added therein to fix permutation ambiguities, whereas they are not necessary; (iii) in [14], CRB are computed for the ESPRIT technique, but translations are assumed known whereas they are actually unknown nuisance parameters; if they are known, ESPRIT cannot perform better than MUSIC.

Acknowledgement This work has been funded by the European Research Council under the 7th Framework Programme FP7/2007–2013 Grant Agreement no. 320594.

Bibliography

- [1] R. Roy and T. Kailath, “ESPRIT - estimation of signal parameters via rotational invariance techniques,” *IEEE Trans. Acoust. Speech Signal Proc.*, vol. 37, pp. 984–995, July 1989.
- [2] N. D. Sidiropoulos, R. Bro, and G. B. Giannakis, “Parallel factor analysis in sensor array processing,” *IEEE Trans. Sig. Proc.*, vol. 48, no. 8, pp. 2377–2388, Aug. 2000.
- [3] L.-H. Lim and P. Comon, “Blind multilinear identification,” *IEEE Trans. Inf. Theory*, vol. 60, no. 2, pp. 1260–1280, Feb. 2014, open access.
- [4] S. Sahnoun and P. Comon, “Deterministic blind identification in antenna array processing,” in *8th IEEE SAM Workshop*, A Coruna, Spain, June 22-25 2014, hal-00957357.
- [5] H. Krim and M. Viberg, “Two decades of array signal processing research,” *IEEE Sig. Proc. Mag.*, pp. 67–95, July 1996.
- [6] W. J. Vetter, “Derivative operations on matrices,” *IEEE Trans. Auto. Control*, vol. 15, no. 2, pp. 241–244, 1970.
- [7] J. W. Brewer, “Kronecker products and matrix calculus in system theory,” *IEEE Trans. on Circuits and Systems*, vol. 25, no. 9, pp. 114–122, Sept. 1978.
- [8] P. Comon, “Tensors: a brief introduction,” *IEEE Sig. Proc. Magazine*, vol. 31, no. 3, May 2014, special issue on BSS. hal-00923279.
- [9] P. Comon, “Estimation multivariable complexe,” *Traitement du Signal*, vol. 3, no. 2, pp. 97–101, Apr. 1986, hal-00979476.
- [10] A. van den Bos, “A Cramér-Rao bound for complex parameters,” *IEEE Trans. Sig. Proc.*, vol. 42, no. 10, pp. 2859, Oct. 1994.
- [11] A. Hjørungnes and D. Gesbert, “Complex-valued matrix differentiation: Techniques and key results,” *IEEE Trans. Sig. Proc.*, vol. 55, no. 6, pp. 2740–2746, June 2007.
- [12] X. Liu and N. Sidiropoulos, “Cramér-Rao lower bounds for low-rank decomposition of multidimensional arrays,” *IEEE Trans. Sig. Proc.*, vol. 49, no. 9, pp. 2074–2086, 2001.
- [13] P. Stoica and A. Nehorai, “MUSIC, maximum likelihood, and Cramer-Rao bound,” *IEEE Trans. Acoust. Speech Sig. Proc.*, vol. 37, no. 5, pp. 720–741, 1989.
- [14] B. Ottersten, M. Viberg, and T. Kailath, “Performance analysis of the total least squares ESPRIT algorithm,” *IEEE Trans. Sig. Proc.*, vol. 39, no. 5, pp. 1122–1135, 1991.

Adjustment for nonignorable nonresponse using latent homogeneous response groups

Caren Hasler, *University of Neuchâtel, Switzerland*, caren.hasler@unine.ch

Alina Matei, *University of Neuchâtel and IRDP Neuchâtel, Switzerland*, alina.matei@unine.ch

Abstract. estimated response probabilities are used to compute a two-phase estimator of the population total. Simulations are performed in order to compare the proposed estimators with other estimators currently used. The advantages in terms of bias and variance of the proposed approaches are confirmed through these simulations. We consider a setup in which nonignorable nonresponse is present in the survey. In such a case, the unit response probabilities depend on the variable of interest. When the variable of interest follows a mixture distribution (a typical example of such a variable is income), it is possible to highlight latent homogeneous response groups based on the variable of interest and auxiliary information. Two approaches are discussed. In both approaches, response probabilities are estimated through logistic regression. The estimated response probabilities are then used to compute a two-phase estimator of the population total. Simulations are performed in order to compare the performance of the proposed estimators with that of other estimators currently used. The advantages in terms of reduction of nonresponse bias and variance of the proposed approaches are confirmed through these simulations.

Keywords. Survey sampling, Unit response probability, Two-phase estimation

1 Introduction

Reweighting procedures are commonly used to compensate for unit nonresponse in surveys. The main idea is to increase the sampling weights of each respondent in order to compensate for the nonrespondents. One refers to such procedures as nonresponse weighting adjustment (NWA) methods. Nonresponse can be viewed as a second phase of the survey. Theory of two-phase sampling hence suggests a two-phase estimator which extends the usual Horvitz-Thompson estimator by multiplying the sampling weights of the respondents by the inverse of their response probabilities. As the response probabilities are unknown, a preliminary step consists of estimating them. The sampling weights of the respondents are then multiplied by

the inverse of their estimated response probabilities and a two-phase estimator adjusted for nonresponse is obtained. In the literature, several approaches have been used to estimate the response probabilities, as for example response homogeneity groups, calibration, or parametric modelling as in [2] and [7]. Auxiliary information available at the sample or population level plays a central role in the estimation process. It can simultaneously decrease variance and nonresponse bias of estimators if it is adequately used in the response probabilities estimation. The reader may refer to [11] for an overview of NWA methods.

Nonignorable nonresponse refers to a nonresponse mechanism which depends on the variable of interest itself (see [9] for a formal definition). It is particularly difficult to handle as the process that leads to nonresponse is defined through characteristics of interest which are partially or completely missing. Sophisticated techniques must therefore be used to control for nonresponse bias and variance in this framework. The problem of nonignorable nonresponse in surveys has already been addressed as for instance in [6], [10], [1], and [4].

We propose two NWA procedures for handling nonignorable nonresponse, when the variable of interest follows a mixture distribution with different components. The goal is to reduce nonresponse bias and variance of estimators. Latent homogeneous response groups based on both auxiliary information and the variable of interest are highlighted for respondents and are imputed using auxiliary information for nonrespondents. In the presented procedures, the response probabilities are modelled through logistic regression including information about the groups (observed or imputed). The estimated response probabilities are then used in a two-phase estimator for the total of the variable of interest. The inclusion of information about the groups in the estimation of the response probabilities allows to control simultaneously nonresponse bias and variance of the two-phase estimator.

A typical example of application where the proposed methods can be used is a survey whose variable of interest is the income. Indeed, it is customary and sensible to suppose that the willingness to answer questions related to income depends on the income itself. On the other hand, income data typically shows heterogeneity and mixture distributions represent a powerful tool to model such data (see [5]). It follows that a natural assumption is the existence of homogeneous response groups depending on the underlying income mixture groups and auxiliary information.

The paper is organized as follows. Section 2 introduces the framework and notation. Section 3 discusses the response probabilities estimation for nonignorable nonresponse using logistic regression. The proposed procedures are presented in Section 4. Next, in Section 5, the performance of the proposed procedures is tested and compared to that of other NWA procedures through a simulation study. Finally, Section 6 closes the paper with brief concluding remarks.

2 Framework

Consider a finite population U of size N , indexed by i from 1 to N . Let $\mathbf{x}_i = (x_{i1}, x_{i2}, \dots, x_{iq})^\top$ be a vector of q auxiliary variables attached to unit i and suppose that the parameter of interest is the population total $Y = \sum_{i \in U} y_i$, for some continuous or categorical variable of interest y . In a first phase, a sample s of size n is selected from the population U using a sampling design $p(s)$. Let $\pi_i = \sum_{s: s \ni i} p(s)$ denote the first-order inclusion probability of unit i and suppose thereafter that $\pi_i > 0$ for all $i \in U$. The vector of auxiliary variables \mathbf{x}_i is assumed to be available for each population unit $i \in U$ or at least for each sampled unit $i \in s$. In the presence of unit

nonresponse, some selected units do not respond to the survey. This results in two subsets which form a partition of s : the survey *respondents* (the set r) and the survey *nonrespondents* (the set \bar{r}). The value y_i of the variable of interest is then observed for each respondent $i \in r$ but is missing for each nonrespondent $i \in \bar{r}$. For $i \in s$, let R_i be the response indicator of y_i which takes value 1 if unit i is a respondent (i.e. if $i \in r$) and 0 if unit i is a nonrespondent (i.e. if $i \in \bar{r}$). Let p_i be the response propensity of unit i , that is $p_i = \Pr(i \in r | s; i \in s)$. It is supposed that units respond independently from each other. The response indicator R_i is therefore generated from a Bernoulli random variable with parameter p_i . Moreover, it is thereafter assumed that $p_i > 0$ for all $i \in U$. In the ideal case of complete response, the Horvitz-Thompson estimator

$$\widehat{Y}_\pi = \sum_{i \in s} \frac{1}{\pi_i} y_i, \quad (1)$$

is a design unbiased estimator for Y . In the presence of nonresponse, however, this latter is intractable as the values y_i of the variable of interest are missing for nonrespondents $i \in \bar{r}$. Nonresponse can be viewed as a second phase of the survey. A subsample r of s is indeed selected according to a Poisson sampling design $q(r|s) = \prod_{i \in r} p_i \prod_{i \in \bar{r}} (1 - p_i)$. Theory of two-phase sampling proposes, in this case, the double expansion estimator $\widehat{Y}_{\pi,p} = \sum_{i \in r} \frac{1}{\pi_i} \frac{1}{p_i} y_i$, which extends the estimator in Expression (1). This estimator would be unbiased for Y if the response probabilities p_i were known. Unfortunately, this is never the case. A preliminary step therefore consists of estimating the response probabilities. Those are then replaced by the estimated response probabilities \widehat{p}_i in the previous estimator and the two-phase estimator adjusted for nonresponse

$$\widehat{Y}_{\pi,\widehat{p}} = \sum_{i \in r} \frac{1}{\pi_i} \frac{1}{\widehat{p}_i} y_i, \quad (2)$$

is obtained. If the response probabilities are parametrically modeled, then it is shown in [7] that estimator $\widehat{Y}_{\pi,\widehat{p}}$ is more efficient than estimator $\widehat{Y}_{\pi,p}$ when maximum likelihood is used to estimate the parameters. In Section 3, the question of the response probabilities estimation for nonignorable nonresponse is discussed.

3 Estimating response probabilities

Under nonignorable nonresponse, a solution to estimate the response probabilities consists of modelling them with logistic regression in which the variable of interest plays the role of a covariate. Hence, the following two models can be considered:

$$p_i = \mathbb{E}(R_i | y_i) = \frac{1}{1 + \exp[-(\beta_0 + \beta_1 y_i)]}, \quad (3)$$

$$p_i = \mathbb{E}(R_i | y_i, \mathbf{x}_i) = \frac{1}{1 + \exp[-(\beta_0 + \beta_1 y_i + \mathbf{x}_i^\top \boldsymbol{\alpha})]}, \quad (4)$$

where β_0 , β_1 , and $\boldsymbol{\alpha}$ are parameters. In the presence of nonresponse, however, these parameters can not be estimated as the values y_i of the variable of interest are missing for the nonrespondents.

A solution is proposed in [2] and is presented below. It consists of considering only the auxiliary variables as covariates. This results in the following model

$$p_i = \mathbb{E}(R_i | \mathbf{x}_i) = \frac{1}{1 + \exp[-(\beta_0 + \mathbf{x}_i^\top \boldsymbol{\alpha})]}, \quad (5)$$

where β_0 and $\boldsymbol{\alpha}$ are parameters. As the values \mathbf{x}_i of the auxiliary variables are known for each sampled unit $i \in s$, the parameters can now be estimated considering (R_i, \mathbf{x}_i) for $i \in s$. Consider $\widehat{\beta}_0$ and $\widehat{\boldsymbol{\alpha}}$ the maximum likelihood estimates of parameters β_0 and $\boldsymbol{\alpha}$, and estimate the response probabilities by replacing the parameters by their estimates in Expression (5), that is $\widehat{p}_i = 1 / \{1 + \exp[-(\widehat{\beta}_0 + \mathbf{x}_i^\top \widehat{\boldsymbol{\alpha}})]\}$. If the auxiliary variables are good predictors for the variable of interest or for the response probabilities, then this procedure provides protection against nonresponse bias (see [2]).

4 Latent homogeneous response groups

We assume that the variable of interest y follows a mixture distribution with t components $y_i \sim \sum_{\ell=1}^t \lambda_\ell f_\ell(y_i | \mathbf{x}_i, \theta_\ell)$, $\lambda_\ell \geq 0$, $\sum_{\ell=1}^t \lambda_\ell = 1$, where λ_ℓ is the prior probability of component ℓ (y_i is drawn from a mixture of densities of underlying groups or clusters or subpopulations in unknown proportions $\lambda_1, \dots, \lambda_t$) and θ_ℓ is the specific parameter vector for the density function f_ℓ in the ℓ th component. If f_ℓ is a univariate normal density and $\theta_\ell = (\mu_\ell, \sigma_\ell^2)'$, one describes a mixture of standard linear regression models, also called latent class regression or cluster-wise regression (see [3]). Other f_ℓ densities can also be used.

A typical example of such a variable y is income. Models based on mixed distributions better explain the income heterogeneity in different subpopulations. When nonresponse treatment is added, latent homogeneous response groups can be highlighted based on these subpopulations. These response groups depend on the variable of interest and the auxiliary information. An important gain in terms of reduction of nonresponse bias and variance can be derived from including information about these groups in the estimation of the response probabilities. In the presence of nonresponse, however, these groups are not fully observed as the values y_i of the variable of interest are unknown for nonrespondents. In the current section, a procedure to reconstruct these latent homogeneous response groups is presented. Then, two solutions to include them in the response probabilities estimation are proposed.

As stated above, homogeneous response groups are observed for respondents only. A procedure to reconstruct the group membership of the nonrespondents is provided here. The main idea is to impute the missing groups by nearest neighbor imputation. Suppose that k homogeneous groups are observed for the respondents. Moreover, let $c_i \in \{1, 2, \dots, k\}$ be the observed group membership value of respondent $i \in r$ and consider $c_i^* \in \{1, 2, \dots, k\}$ the reconstructed membership group value of a unit $i \in s$. As the membership group value is observed for each respondent, we set $c_i^* = c_i$ for $i \in r$. For a nonrespondent, however, the membership group value is unobserved and that one is reconstructed by nearest neighbor imputation using auxiliary information. Hence, for $i \in \bar{r}$, consider $c_i^* = c_{j(i)}$ where $j(i)$ satisfies $d(\mathbf{x}_i, \mathbf{x}_{j(i)}) = \min_{j \in r} d(\mathbf{x}_i, \mathbf{x}_j)$ for some distance measure $d(\cdot, \cdot)$. Therefore, observed group membership values are combined with imputed group membership values. This leads to a reconstructed group membership variable whose values c_i^* are available for every sampled unit $i \in s$.

Two different models can be constructed. In the first one, the reconstructed group membership variable (observed or imputed) is added as a categorical covariate. This results in the following model

$$p_i = \mathbb{E}(R_i | \mathbf{x}_i, c_i^*) = \frac{1}{1 + \exp[-(\beta_0 + \mathbf{x}_i^\top \boldsymbol{\beta}_1 + \beta_2 c_i^* + \mathbf{x}_i^\top \boldsymbol{\beta}_3 c_i^*)]}, \quad (6)$$

where $\beta_0, \beta_1, \beta_2$, and β_3 are parameters. The maximum likelihood estimation is then applied to fit this model considering $(R_i, \mathbf{x}_i, c_i^*)$ for $i \in s$. This leads to estimates $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3$, and $\hat{p}_i = 1 / \{1 + \exp[-(\hat{\beta}_0 + \mathbf{x}_i^\top \hat{\beta}_1 + \hat{\beta}_2 c_i^* + \mathbf{x}_i^\top \hat{\beta}_3 c_i^*)]\}$. If the auxiliary variables are good predictors of the variable of interest or good predictors of the response probabilities and moreover if the reconstructed groups are homogeneous with respect to the variable of interest or with respect to the response probabilities, then this procedure provides additional protection against nonresponse bias and variance compared to Model (5).

In the second proposed procedure, the missing values of the variable of interest are imputed in each reconstructed group. The response probabilities are estimated using logistic regression and the variable of interest (observed or imputed); see also [8]. Hence, let y_i^* denote a reconstructed value of the variable of interest of a unit $i \in s$. For a respondent $i \in r$, this value corresponds to the observed value of the variable of interest, that is $y_i^* = y_i$. Then, for the nonrespondents, the missing y_i 's are reconstructed by using regression imputation independently in each reconstructed group. Hence, for each nonrespondent $i \in \bar{r}$ we set $y_i^* = \left(\sum_{j \in r | c_j^* = c_i^*} \frac{1}{\pi_j} \mathbf{x}_j \mathbf{x}_j^\top\right)^{-1} \left(\sum_{\ell \in r | c_\ell^* = c_i^*} \frac{1}{\pi_\ell} \mathbf{x}_\ell y_\ell\right) \mathbf{x}_i$. Therefore, observed values of the variable of interest are combined with imputed values. This leads to a reconstructed variable of interest whose values y_i^* are available for every sampled unit $i \in s$. This variable then plays the role of covariate in the logistic regression used to estimate the response probabilities. Hence, the parameters δ_0 and δ_1 of the logistic regression model

$$p_i = \mathbb{E}(R_i | y_i^*) = \frac{1}{1 + \exp[-(\delta_0 + \delta_1 y_i^*)]}, \quad (7)$$

are estimated by maximum likelihood considering (R_i, y_i^*) for $i \in s$. This leads to estimates $\hat{\delta}_0, \hat{\delta}_1$, and $\hat{p}_i = 1 / \{1 + \exp[-(\hat{\delta}_0 + \hat{\delta}_1 y_i^*)]\}$. If the auxiliary variables are good predictors of the variable of interest within the reconstructed groups but not necessarily within the whole population, then this procedure provides additional protection against nonresponse bias and variance compared to Model (5). Even though y_i^* is essentially a linear combination of the outer product of the auxiliary variables and the (imputed) latent groups, Model (7) is different from the model including \mathbf{x}_i and c_i^* as covariates as in Expression (6), because it uses the original y_i for the respondents and performs closer to the assumed response model.

5 Simulations

A simulation study was conducted to evaluate the performance of the procedures proposed in Section 4. Two different settings were considered. In each setting, a population of size $N = 1000$ divided into two groups of equal size, a variable of interest y generated from a mixture distribution, and an auxiliary variable x were considered. A census was considered in both cases, which implies that we set $U = s$ and $\pi_i = 1$ for each $i \in s$. Ten thousand simulations were conducted.

For each setting, the simulations were conducted as follows. First, for each unit i , the response probabilities were obtained from the logistic function $p_i = 1 / \{1 + \exp[-(\beta_0 + \beta_1 y_i)]\}$, where β_0 and β_1 were fixed to obtain a mean response rate close to 65%. Then, 10000 response sets were created by generating 10000 response indicator vectors R . Each component $R_i, i \in U$ of R was generated from a Bernoulli distribution with parameter p_i . For each response set generated, the population total for the variable of interest was estimated through the two-phase

estimator adjusted for nonresponse of Expression (2) by considering different choices for the estimated response probabilities \hat{p}_i as follows:

1. $\hat{Y}_{\hat{p}(x)}$: estimator proposed in [2], i.e. response probabilities estimated through logistic regression with the auxiliary variables as covariates as in Model (5),
2. $\hat{Y}_{\hat{p}(x,c^*)}$: first proposed procedure, i.e. response probabilities estimated through logistic regression with the auxiliary variables and the reconstructed membership groups variable as covariates as in Model (6),
3. $\hat{Y}_{\hat{p}(y^*)}$: second proposed procedure, i.e. response probabilities estimated through logistic regression with the values of the variable of interest (observed or imputed through regression imputation in the reconstructed groups) as covariates as in Model (7),
4. $\hat{Y}_{\hat{p}(y^{nn})}$: response probabilities estimated through logistic regression with the vector of observed and imputed by nearest neighbor values of the variable of interest y^{nn} as covariate. The coefficients of y^{nn} are thus defined as $y_i^{nn} = y_i$ if $i \in r$ and $y_i^{nn} = y_{j(i)}$ where $|x_i - x_{j(i)}| = \min_{j \in r} |x_i - x_j|$ if $i \in \bar{r}$,
5. \hat{Y}_p : true response probabilities considered in the two-phase estimator.

The following comparison measures were considered for these five estimators, here generically denoted by \hat{Y} :

- The Monte Carlo relative bias: $RB = B/Y$, where $B = \mathbb{E}_{sim}(\hat{Y}) - Y$, $\mathbb{E}_{sim}(\hat{Y}) = \sum_{i=1}^M \hat{Y}_i / M$, \hat{Y}_i is the estimate of \hat{Y} obtained at the i -th simulation, and M is the number of simulations,
- The Monte Carlo variance: $VAR = \frac{1}{M-1} \sum_{i=1}^M [\hat{Y}_i - \mathbb{E}_{sim}(\hat{Y})]^2$,
- The Monte Carlo mean square error: $MSE = B^2 + VAR$.

Details and results from the two considered settings are presented below.

Setting 1: A single auxiliary variable $x = (x_i)_{i=1}^N$ was considered. Its coefficients were generated by independent draws of a uniform distribution with parameters 0 and 1 for units that belong to the first group, and by independent draws of a uniform random variable with parameters 2 and 3 for units that belong to the second group. Next, the variable of interest $y = (y_i)_{i=1}^N$ was generated as follows: $y_i = 5 + 5x_i + 3\varepsilon_i$ if i belongs to the first group and $y_i = 40 - (x_i - 5)^2 + 3\varepsilon_i$ if i belongs to the second group, where ε_i are independent draws of a normal random variable with mean 0 and variance 1. Simulations were then conducted according to the scheme described above. The results are presented in Table 1.

The two proposed estimators ($\hat{Y}_{\hat{p}(x,c^*)}$ and $\hat{Y}_{\hat{p}(y^*)}$) display a decrease in relative bias compared to estimators $\hat{Y}_{\hat{p}(x)}$ and $\hat{Y}_{\hat{p}(y^{nn})}$. The gap between the relative bias of $\hat{Y}_{\hat{p}(x,c^*)}$ and that of $\hat{Y}_{\hat{p}(y^{nn})}$ is not large and makes it difficult to clearly rank these two estimators. The proposed estimators, however, imply a clear decrease in variance compared to estimators $\hat{Y}_{\hat{p}(x)}$ and $\hat{Y}_{\hat{p}(y^{nn})}$. Estimator \hat{Y}_p is clearly the best in terms of bias, which is not surprising. Indeed, it uses the true response probabilities and is therefore unbiased for the total (the small relative bias is due to the simulation process). Finally, the four estimators with estimated probabilities imply a huge decrease in

Table 1: Comparison measures for five estimators in setting 1.

Estimator	RB ($\times 10^{-3}$)	Var ($\times 10^3$)	MSE ($\times 10^4$)
$\widehat{Y}_{\widehat{p}(x)}$	6.96	7.59	2.83
$\widehat{Y}_{\widehat{p}(x,c^*)}$	4.63	5.10	1.43
$\widehat{Y}_{\widehat{p}(y^*)}$	3.25	5.17	0.97
$\widehat{Y}_{\widehat{p}(y^{nn})}$	5.56	9.34	2.25
\widehat{Y}_p	-0.13	226.90	22.69

variance compared to the estimator with the true probabilities (\widehat{Y}_p), which confirms the result in [7].

Setting 2: The values y_i of the variable of interest $y = (y_i)_{i=1}^N$ were generated independently from a gamma distribution with parameters 10 and 1 for units that belong to the first group and from a gamma distribution with parameters 40 and 1 for units that belong to the second group. Next, values of an auxiliary variable $x = (x_i)_{i=1}^N$ were generated as follows. We set $x_i = 5 + \rho_1 y_i + \varepsilon_i$, where $\rho_1 = 0.7$ and where ε_i was drawn from a normal random variable with mean 0 and variance $10(1 - \rho_1^2)$ if i belongs to the first group. Moreover, we set $x_i = 5 + \rho_2 y_i + \varepsilon_i$, where $\rho_2 = 0.93$, and where ε_i was drawn from a normal random variable with mean 0 and variance $40(1 - \rho_2^2)$ if unit i belongs to the second group. Simulations were then conducted according to the scheme described above. The results are presented in Table 2. These results

Table 2: Comparison measures for five estimators in setting 2.

Estimator	RB ($\times 10^{-3}$)	Var ($\times 10^3$)	MSE ($\times 10^4$)
$\widehat{Y}_{\widehat{p}(x)}$	11.04	13.61	9.01
$\widehat{Y}_{\widehat{p}(x,c^*)}$	6.69	10.02	3.82
$\widehat{Y}_{\widehat{p}(y^*)}$	5.44	9.83	2.84
$\widehat{Y}_{\widehat{p}(y^{nn})}$	6.94	14.93	4.52
\widehat{Y}_p	0.02	267.78	26.78

follow a fairly similar pattern to those of setting 1. The two proposed estimators ($\widehat{Y}_{\widehat{p}(x,c^*)}$ and $\widehat{Y}_{\widehat{p}(y^*)}$) display a decrease in relative bias compared to estimators $\widehat{Y}_{\widehat{p}(x)}$ and $\widehat{Y}_{\widehat{p}(y^{nn})}$. However, the gap between the relative bias of $\widehat{Y}_{\widehat{p}(x,c^*)}$ and that of $\widehat{Y}_{\widehat{p}(y^{nn})}$ is very small and does not allow us to rank these two estimators. The proposed estimators again imply a clear decrease in variance compared to estimators $\widehat{Y}_{\widehat{p}(x)}$ and $\widehat{Y}_{\widehat{p}(y^{nn})}$. Finally, estimator \widehat{Y}_p also displays by far the smallest relative bias and the largest variance.

6 Conclusion

We have proposed two NWA procedures for handling nonignorable nonresponse when the variable of interest follows a mixture distribution. Homogeneous response groups can be constructed

based on the hidden structure of the variable of interest; they include information about the variable of interest and the auxiliary information. Benefits in terms of reduction of nonresponse bias and variance of the total estimator can be obtained if these groups are taken into account in the response probability estimation. Our results are confirmed through a simulation study. We have not considered the problem of variance estimation of the total estimator when the proposed methods are applied. This problem is currently under investigation.

Acknowledgement

The authors are grateful to a reviewer for his constructive comments and suggestions.

Bibliography

- [1] Beaumont, J.-F. (2000). An estimation method for nonignorable nonresponse. *Survey Methodology*, 26:131–136.
- [2] Cassel, C. M., Särndal, C.-E., and Wretman, J. H. (1983). Some uses of statistical models in connexion with the nonresponse problem. In Madow, W. G. and Olkin, I., editors, *Incomplete Data in Sample Surveys*, volume 3, pages 143–160. Academic Press, New York.
- [3] DeSarbo, W. S., and Cron, W. L. (1988). A maximum likelihood methodology for clusterwise linear regression. *Journal of Classification*, 5, 249–282.
- [4] Fang, F., Hong, Q., and Shao, J. (2010). Empirical likelihood estimation for samples with nonignorable nonresponse. *Statistica Sinica*, 20:263–280.
- [5] Flachaire, E. and Nuñez, O. (2007). Estimation of the income distribution and detection of subpopulations: An explanatory model. *Computational Statistics & Data Analysis*, 51:3368–3380.
- [6] Greenlees, J. S. and Reece, W. S., and Zieschang, K. (1982) Imputation of missing values when the probability of response depends on the variable being imputed. *Journal of the American Statistical Association*, 378:251–261.
- [7] Kim, J. K. and Kim, J. (2007). Nonresponse weighting adjustment using estimated response probability. *The Canadian Journal of Statistics / La Revue Canadienne de Statistique*, 35(4):501–514.
- [8] Laaksonen, S. and Chambers, R. (2006). Survey estimation under informative nonresponse with follow-up. *Journal of Official Statistics*, 22(1):81–95.
- [9] Little, R. J. A. (1982). Models for nonresponse in sample surveys. *Journal of the American Statistical Association*, 77(378):237–250.
- [10] Little, R. J. A. and Rubin, D. B. (1987). *Statistical Analysis with Missing Data*. Wiley, Canada.
- [11] Särndal, C.-E. and Lundström, S. (2005). *Estimation in Surveys with Nonresponse*. Wiley, New York.

Bartlett adjustment of deviance statistic for three types of binary response models

Nobuhiro Taneichi, *Kagoshima University*, taneichi@sci.kagoshima-u.ac.jp

Yuri Sekiya, *Hokkaido University of Education*, sekiya.yuri@k.hokkyodai.ac.jp

Jun Toyama, *The Institute for the Practical Application of Mathematics*, mandheling@nifty.com

Abstract. A logistic regression model, complementary log-log model and probit model are frequently used for a generalised linear model of binary data. We consider deviance (log likelihood ratio statistic) as a goodness-of-fit statistic. In this paper, using the continuous term of asymptotic expansion for the deviance under the null hypothesis that each model is correct, we obtain the Bartlett adjusted deviance statistic for each model that improves the speed of convergence to chi-square limiting distribution of deviance. Performance of each adjusted deviance statistic is also investigated numerically.

Keywords. Asymptotic expansion, Bartlett adjustment, Complementary log-log model, Deviance, Generalized linear model, Logistic regression model, Probit model

1 Introduction

We consider generalized linear models (Nelder and Wedderburn [5]) in which the response variables are measured on a binary scale. Let random variables Y_α , $\alpha = 1, \dots, S$ be the number of successes in S different subgroups, which are independent distributed according to binomial distributions $B(n_\alpha, \pi_\alpha)$, $\alpha = 1, \dots, S$. If we use a monotone and differentiable function $g(\cdot)$ as a link function, we obtain a generalized linear model for binary data as

$$g(\pi_\alpha) = \mathbf{x}'_\alpha \boldsymbol{\beta}, \quad \alpha = 1, \dots, S, \quad (1)$$

where $\mathbf{x}_\alpha = (x_{\alpha 1}, \dots, x_{\alpha p})'$, $\alpha = 1, \dots, S$, are covariate vectors and $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)'$ is an unknown parameter vector and ($p < S$). When $g(t)$ is a canonical link function, that is,

$$g(t) = \log \left(\frac{t}{1-t} \right),$$

model (1) is a logistic regression model. When

$$g(t) = g_P(t) = \Phi^{-1}(t),$$

where

$$\Phi(t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^t \exp\left(-\frac{x^2}{2}\right) dx,$$

model (1) is a probit model. When

$$g(t) = \log\{-\log(1-t)\},$$

model (1) is a complementary log-log model.

We consider the null hypothesis

$$H_0 : \pi_\alpha = \pi_\alpha(\boldsymbol{\beta}) = g^{-1}(\mathbf{x}'_\alpha \boldsymbol{\beta}), \quad \alpha = 1, \dots, S. \quad (2)$$

The deviance (log likelihood ratio statistic) is

$$D = 2 \sum_{\alpha=1}^S n_\alpha \left\{ \frac{Y_\alpha}{n_\alpha} \log \left(\frac{Y_\alpha}{n_\alpha \hat{\pi}_\alpha} \right) + \left(1 - \frac{Y_\alpha}{n_\alpha} \right) \log \left(\frac{1 - \frac{Y_\alpha}{n_\alpha}}{1 - \hat{\pi}_\alpha} \right) \right\},$$

where $\hat{\pi}_\alpha = \pi_\alpha(\hat{\boldsymbol{\beta}})$, $\alpha = 1, \dots, S$ and $\hat{\boldsymbol{\beta}} = (\hat{\beta}_1, \dots, \hat{\beta}_p)'$ is the maximum likelihood estimator of $\boldsymbol{\beta}$ under H_0 given by (2). Under the null hypothesis H_0 , it is known that the deviance D has a χ^2_{S-p} limiting distribution if

$$n_\alpha/n \rightarrow \mu_\alpha \quad (0 < \mu_\alpha < 1) \text{ for each } \alpha, \quad \text{as } n \rightarrow \infty, \quad (3)$$

where $n = \sum_{\alpha=1}^S n_\alpha$ and $\sum_{\alpha=1}^S \mu_\alpha = 1$. Usually, using large sample results, we test H_0 by using the statistic D for a goodness-of-fit test statistic of each model.

However, in the case in which all n_α , $\alpha = 1, \dots, S$ are not large enough, such an approximation by a χ^2_{S-p} limiting distribution to the distribution of D under H_0 becomes poor. So, there are risks that the hypothesis test based on large sample theory will give results opposite to those of an exact test. In this paper, in order to reduce the risks, we propose a new adjusted statistic \tilde{D}^B of D whose speed of convergence to a chi-square distribution is quicker than that of D . To construct \tilde{D}^B , we use the following procedure. First, we formally obtain the asymptotic expansion of the original statistic D assuming a continuous distribution of D . Next, we obtain adjusted statistic \tilde{D}^B by performing Bartlett adjustment to D on the basis of the asymptotic expansion assuming a continuous distribution of D .

2 An asymptotic approximation for the distribution of D under H_0

With regard to evaluation of the lower probability of the deviance D under H_0 , we obtain the following theorem (a special case of Taneichi *et al.* [13]). Here, we consider the following Assumption 2.1 instead of the assumption given by (3).

Assumption 2.1. $n_\alpha \rightarrow \infty$, $\alpha = 1, \dots, S$, as $n \rightarrow \infty$, with n_α depending on n in such a way that $n_\alpha/n = \mu_\alpha$, $\alpha = 1, \dots, S$, where $0 < \mu_\alpha < 1$ and $\sum_{\alpha=1}^S \mu_\alpha = 1$.

Theorem 2.1. When g^{-1} is a fourth time continuously differentiable function, under Assumption 2.1 and assuming that D is continuously distributed, the lower probability of the deviance D under H_0 is evaluated as

$$\Pr\{D \leq x|H_0\} = \Pr\{\chi_{S-p}^2 \leq x\} + \frac{1}{n} \sum_{j=0}^1 v_j \Pr\{\chi_{S-p+2j}^2 \leq x\} + O(n^{-2}),$$

where χ_f^2 denotes a chi-square random variable with degrees of freedom f ,

$$v_0 = -\frac{1}{24}(2A_1 - 6A_2 + 12A_3 - 3A_4 + 4B_1 - 12B_2 + 6B_3 - 3B_4),$$

$$v_1 = -v_0,$$

where

$$A_1 = \sum_{\alpha=1}^S \frac{1 - \pi_\alpha + \pi_\alpha^2}{\mu_\alpha \pi_\alpha (1 - \pi_\alpha)}, \quad A_2 = \sum_{\alpha=1}^S \frac{\mu_\alpha (1 - 3\pi_\alpha + 3\pi_\alpha^2)}{\pi_\alpha^3 (1 - \pi_\alpha)^3} G_1^4(\alpha) \sigma_{\alpha\alpha}^2,$$

$$A_3 = \sum_{\alpha=1}^S \frac{\mu_\alpha (1 - 2\pi_\alpha)}{\pi_\alpha^2 (1 - \pi_\alpha)^2} G_1^2(\alpha) G_2(\alpha) \sigma_{\alpha\alpha}^2, \quad A_4 = \sum_{\alpha=1}^S \frac{\mu_\alpha}{\pi_\alpha (1 - \pi_\alpha)} G_2^2(\alpha) \sigma_{\alpha\alpha}^2,$$

$$B_1 = \sum_{\alpha=1}^S \sum_{\gamma=1}^S \frac{\mu_\alpha (1 - 2\pi_\alpha) \mu_\gamma (1 - 2\pi_\gamma)}{\pi_\alpha^2 (1 - \pi_\alpha)^2 \pi_\gamma^2 (1 - \pi_\gamma)^2} G_1^3(\alpha) G_1^3(\gamma) \sigma_{\alpha\gamma}^3,$$

$$B_2 = \sum_{\alpha=1}^S \sum_{\gamma=1}^S \frac{\mu_\alpha}{\pi_\alpha (1 - \pi_\alpha)} \frac{\mu_\gamma (1 - 2\pi_\gamma)}{\pi_\gamma^2 (1 - \pi_\gamma)^2} G_1(\alpha) G_2(\alpha) G_1^3(\gamma) \sigma_{\alpha\gamma}^3,$$

$$B_3 = \sum_{\alpha=1}^S \sum_{\gamma=1}^S \frac{\mu_\alpha}{\pi_\alpha (1 - \pi_\alpha)} \frac{\mu_\gamma}{\pi_\gamma (1 - \pi_\gamma)} G_1(\alpha) G_2(\alpha) G_1(\gamma) G_2(\gamma) \sigma_{\alpha\alpha}^3 \sigma_{\alpha\gamma}^3,$$

$$B_4 = \sum_{\alpha=1}^S \sum_{\gamma=1}^S \frac{\mu_\alpha}{\pi_\alpha (1 - \pi_\alpha)} \frac{\mu_\gamma}{\pi_\gamma (1 - \pi_\gamma)} G_1(\alpha) G_2(\alpha) G_1(\gamma) G_2(\gamma) \sigma_{\alpha\alpha} \sigma_{\alpha\gamma} \sigma_{\gamma\gamma},$$

$$G_i(\alpha) = u^{(i)}(\mathbf{x}'_\alpha \boldsymbol{\beta}), \quad \alpha = 1, \dots, S, \quad i = 1, 2,$$

$$\sigma_{\alpha\gamma} = \mathbf{x}'_\alpha K^{-1} \mathbf{x}_\gamma,$$

$$K = \sum_{\lambda=1}^S \frac{\mu_\lambda}{\pi_\lambda (1 - \pi_\lambda)} G_1^2(\lambda) \mathbf{x}_\lambda \mathbf{x}'_\lambda,$$

where $u^{(i)}$ is the i th derivative of $u(x) = g^{-1}(x)$.

Evaluation for the logistic regression model is given by applying

$$g^{-1}(x) = \frac{\exp(x)}{1 + \exp(x)},$$

$$G_1(\alpha) = \pi_\alpha (1 - \pi_\alpha), \quad \alpha = 1, \dots, S,$$

and

$$G_2(\alpha) = \pi_\alpha (1 - \pi_\alpha) (1 - 2\pi_\alpha), \quad \alpha = 1, \dots, S$$

to Theorem 2.1. Similarly, evaluation for the probit model is given by applying

$$g^{-1}(x) = \Phi(x),$$

$$G_1(\alpha) = \frac{1}{\sqrt{2\pi}} \exp \left[-\frac{\{\Phi^{-1}(\pi_\alpha)\}^2}{2} \right], \quad \alpha = 1, \dots, S,$$

and

$$G_2(\alpha) = -\frac{1}{\sqrt{2\pi}} \Phi^{-1}(\pi_\alpha) \exp \left[-\frac{\{\Phi^{-1}(\pi_\alpha)\}^2}{2} \right], \quad \alpha = 1, \dots, S$$

and evaluation for the complementary log-log model is given by applying

$$g^{-1}(x) = 1 - \exp\{-\exp(x)\},$$

$$G_1(\alpha) = -(1 - \pi_\alpha) \log(1 - \pi_\alpha), \quad \alpha = 1, \dots, S,$$

and

$$G_2(\alpha) = -(1 - \pi_\alpha) \{\log(1 - \pi_\alpha)\} \{1 + \log(1 - \pi_\alpha)\}, \quad \alpha = 1, \dots, S$$

to Theorem 2.1, respectively.

We consider the appropriateness of using the Edgeworth approximation assuming a continuous distribution like Theorem 2.1. Yarnold [14] obtained an asymptotic expansion for the null distribution of X^2 (Pearson's chi-square statistic). The expansion consists of continuous and discontinuous terms. Yarnold [14] numerically examined the accuracy of approximations based on the expansion, χ^2 approximation, and Edgeworth approximation assuming a continuous distribution for the null distribution of X^2 and concluded that the Edgeworth approximation assuming a continuous distribution should never be used when random variable has a lattice distribution. In a similar fashion to X^2 statistic, approximations based on asymptotic expansions for null distributions of the log likelihood ratio test statistic and the Freeman-Tukey statistic were obtained by Siotani and Fujikoshi [9], that of the power-divergence statistics was obtained by Read [6] and that of the ϕ -divergence statistics was obtained by Menéndez et al. [4]. The numerical accuracy of the approximation was shown by Yarnold [14] for X^2 statistic and by Read [7] for power-divergence statistics. When the discontinuous term in the asymptotic expansion can be expressed in a simple form as the discontinuous term for the null distribution of above statistics, we must respect Yarnold's recommendation.

On the other hand, from the numerical results obtained by Yarnold [14], we notice that the χ^2 approximation rarely performs better than the Edgeworth approximation assuming a continuous distribution. Thus, the Edgeworth approximation assuming a continuous distribution appears to be an effective approximation when the discontinuous term in the asymptotic expansion cannot be expressed in a simple form. Unlike in the case of the null distribution of above statistics, it is very difficult to represent the discontinuous term in a simple form in the case of the distribution of statistics under alternative hypothesis and in the case of that for more general multinomial models such as contingency tables. The reason for the results are shown in Taneichi *et al.* [11] and Taneichi and Sekiya [12], mathematically. Edgeworth approximations of the distributions of some kinds of multinomial goodness-of-fit statistics under alternative hypotheses have been investigated Taneichi *et al.* [11, 10] and Sekiya and Taneichi [8]. Taneichi and Sekiya [12] discussed approximations for the distribution of statistics for the test of independence in $r \times s$ contingency tables. Based on numerical investigations, we found that an omission of the discontinuous term does not lead to a serious error.

3 Bartlett adjusted deviance statistic

In this section, we propose the Bartlett adjusted deviance statistic for improving small sample accuracy of χ^2 approximation of the distribution of a random variable.

Suppose that a nonnegative random variable T has an asymptotic expansion such that

$$\Pr\{T \leq x\} = \Pr\{\chi_f^2 \leq x\} + \frac{1}{n} \sum_{j=0}^1 a_j \Pr\{\chi_{f+2j}^2 \leq x\} + O(n^{-2}).$$

Also suppose that the coefficients a_j , ($j = 0, 1$) do not depend on the parameter $n (> 0)$ and must satisfy the relation $a_0 + a_1 = 0$.

In order to increase the accuracy of χ^2 approximation of a random variable T , we consider Bartlett adjustment of random variable T defined by T_B .

$$T_B = \left(1 + \frac{2a_0}{fn}\right) T. \quad (4)$$

Then, it holds that

$$\Pr\{T_B \leq x\} = \Pr\{\chi_f^2 \leq x\} + O(n^{-2}).$$

Lawley [3], Barndorff-Nielsen and Cox [1], and Barndorff-Nielsen and Hall [2] discussed Bartlett adjustment for the log likelihood ratio statistic. Applying Theorem 2.1 to T_B given by (4), we obtain the Bartlett adjusted deviance statistic D^B .

$$D^B = \left\{1 + \frac{2v_0}{n(S-p)}\right\} D.$$

Practically, we must use estimate \hat{v}_0 obtained by substituting the maximum likelihood estimate $\hat{\beta}$ for true value β in v_0 . Therefore, we propose the statistic \tilde{D}^B that is obtained by substituting \hat{v}_0 for v_0 in D^B .

4 Numerical studies

In this section, we compare the performance of the Bartlett adjusted deviance statistic \tilde{D}^B with that of the original deviance D using the Monte Carlo procedure.

We consider a generalized linear model given by (1) with $p = 2$ and $x_{\alpha,1} = 1$ and $x_{\alpha,2} = x_\alpha$, $\alpha = 1, \dots, S$. Let the true values of parameters β_1 and β_2 be β_1^* and β_2^* , respectively. Then, the true value of π_α is

$$\pi_\alpha^* = g^{-1}(\beta_1^* + \beta_2^* x_\alpha), \quad \alpha = 1, \dots, S.$$

As a link function $g(\cdot)$, we consider the logit link, complementary log-log link and probit link. We give a design matrix

$$\mathbf{X} = (\mathbf{1}, \text{vec}\{\mathbf{x}\})$$

and execute the following procedure.

For each α , we generate n_α , $\alpha = 1, \dots, S$ binomial random numbers that are distributed according to $B(1, \pi_\alpha^*)$. From them, we calculate the number of successes Y_α , $\alpha = 1, \dots, S$ and the maximum likelihood estimates $\hat{\beta}_1$ and $\hat{\beta}_2$ for the parameters β_1 and β_2 by Fisher scoring

method. Using the estimates, we calculate the values $\pi_\alpha(\hat{\beta})$, $\alpha = 1, \dots, S$, where $\hat{\beta} = (\hat{\beta}_1, \hat{\beta}_2)'$, and the observed values of the statistics D and \tilde{D}^B . This process is repeated J times.

Among the J times, let V be the number of times that the observed values of the statistic exceed the upper ε point of the χ^2 distribution with degrees of freedom $S - p$, that is, $\chi_{S-p}^2(\varepsilon)$. The performance of χ^2 approximation for the distribution of each statistic can be evaluated on the basis of the index

$$I = \frac{V}{J} - \varepsilon.$$

We consider the following two true parameters

(i) $\beta_1^* = -0.1, \beta_2^* = 0.1,$

(ii) $\beta_1^* = 0.1, \beta_2^* = -0.1,$

and investigate the performance of the following four cases of design matrix when $S = 8$.

(I) $\text{vec}\{\mathbf{x}\} = (2.7, 3.0, 3.3, 3.6, 3.9, 4.2, 4.5, 4.8)'$.

(II) $\text{vec}\{\mathbf{x}\} = (2.85, 3.05, 3.25, 3.45, 3.65, 3.85, 4.05, 4.25)'$.

(III) $\text{vec}\{\mathbf{x}\} = (\log(2.7), \log(3.0), \log(3.3), \log(3.6), \log(3.9), \log(4.2), \log(4.5), \log(4.8))'$.

(IV) $\text{vec}\{\mathbf{x}\} = (\log(2.85), \log(3.05), \log(3.25), \log(3.45), \log(3.65), \log(3.85), \log(4.05), \log(4.25))'$.

For each case, we consider the following two sample designs

(A) $n_1 = \dots = n_8 = n_A,$

(B) $n_1 = \dots = n_4 = n_B, n_5 = \dots = n_8 = 2n_B.$

We investigate the performance for all combinations of two true parameters (i) and (ii), four design matrices (I), (II), (III), and (IV), and sample design (A), where $n_A = 10, 20,$ and $30,$ and sample design (B), where $n_B = 10, 20,$ and $30.$ In the investigation, the number of repetitions is $J = 10^6.$ Figure 1 shows the absolute values of index I in the cases of true parameters (i) and (ii), design matrices (I)–(IV) and significance level $\varepsilon = 0.01, 0.05,$ and 0.10 when the model is given by the complementary log-log link, sample design is (A) and $n_A = 10, 20,$ and $30.$ Figure 2 shows those for the model that is given by the probit link in the same situation as that in Figure 1. When models are given by complementary log-log link and probit link with sample design (B) and when the model is given by logit link with sample designs (A) and (B), results of simulation are almost the same as those in Figure 1 and Figure 2.

From the results of our simulation, we find that the performance of the Bartlett adjusted deviance statistic \tilde{D}^B is better than that of the original deviance statistic $D.$

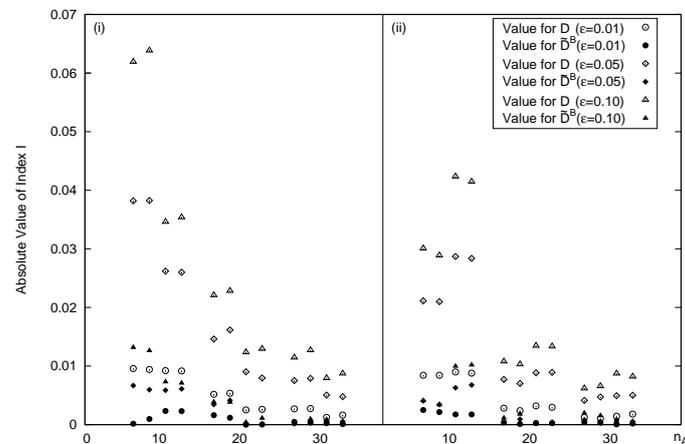


Figure 1: Absolute value of index I when the model is given by the complementary log-log link function for true parameters (i) and (ii) and sample design (A) with $n_A = 10, 20, 30$: \circ, \diamond and \triangle are the values for D when $\epsilon = 0.01, 0.05$ and 0.10 , respectively, and \bullet, \blacklozenge and \blacktriangle are the values for \tilde{D}^B when $\epsilon = 0.01, 0.05$ and 0.10 , respectively. The 1st column is for design matrix (I), the 2nd column is for design matrix (II), the 3rd column is for design matrix (III), and the 4th column is for design matrix (IV).

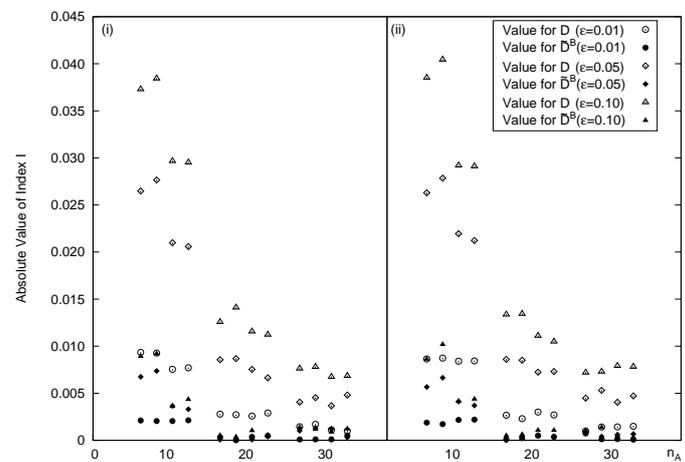


Figure 2: Absolute value of I when the model is given by the probit link function for true parameters (i) and (ii) and sample design (A) with $n_A = 10, 20, 30$: \circ, \diamond and \triangle are the values for D when $\epsilon = 0.01, 0.05$ and 0.10 , respectively, and \bullet, \blacklozenge and \blacktriangle are the values for \tilde{D}^B when $\epsilon = 0.01, 0.05$ and 0.10 , respectively. The 1st column is for design matrix (I), the 2nd column is for design matrix (II), the 3rd column is for design matrix (III), and the 4th column is for design matrix (IV).

Bibliography

- [1] Barndorff-Nielsen, O. E. and Cox, D. R. (1984) *Bartlett adjustments to the likelihood ratio statistic and the distribution of maximum likelihood estimator*. J. R. Statist. Soc., B, **46**, 483–495.
- [2] Barndorff-Nielsen, O. E. and Hall, P. (1988) *On the level-error after Bartlett adjustment of the likelihood ratio statistic*. Biometrika, **75**, 374–378.
- [3] Lawley, D. N. (1956) *A general method for approximating to the distribution of the likelihood ratio criteria*. Biometrika, **43**, 295–303.
- [4] Menéndez, M. L., Pardo, J. A., Pardo, L. and Pardo, M. C. (1997) *Asymptotic approximations for the distributions of the (h, ϕ) -divergence goodness-of-fit statistics: application to Renyi's statistic*. Kybernetes, **26**(4), 442–452.
- [5] Nelder, J. A. and Wedderburn, R. W. M. (1972) *Generalized linear models*. J. R. Statist. Soc. A, **135**, 370–384.
- [6] Read, T. R. C. (1984) *Closer asymptotic approximations for the distributions of the power divergence goodness-of-fit statistics*. Ann. Inst. Statist. Math., **36**, 59–69.
- [7] Read, T. R. C. (1984) *Small-sample comparisons for the power divergence goodness-of-fit statistics*. J. Am. Statist. Assoc., **79**, 929–935.
- [8] Sekiya, Y. and Taneichi, N. (2004) *Improvement of approximations for the distributions of multinomial goodness-of-fit statistics under nonlocal alternatives*. J. Multivariate Anal., **91**, 199–223.
- [9] Siotani, M. and Fujikoshi Y. (1984) *Asymptotic approximations for the distributions of multinomial goodness-of-fit statistics*. Hiroshima Math. J., **14**, 115–124.
- [10] Taneichi, N., Sekiya, Y. and Suzukawa, A. (2001) *An asymptotic approximation for the distribution of ϕ -divergence multinomial goodness-of-fit statistic under local alternatives*. J. Japan Statist. Soc., **31**(2), 207–224.
- [11] Taneichi, N., Sekiya, Y. and Suzukawa, A. (2002) *Asymptotic approximations for the distributions of the multinomial goodness-of-fit statistics under local alternatives*. J. Multivariate Anal., **81**, 335–359.
- [12] Taneichi, N. and Sekiya, Y. (2007) *Improved transformed statistics for the test of independence in $r \times s$ contingency tables*. J. Multivariate Anal., **98**, 1630–1657.
- [13] Taneichi, N., Sekiya, Y. and Toyama, J. (2014) *Transformed goodness-of-fit statistics for a generalized linear model of binary data*. J. Multivariate Anal., **123**, 311–329.
- [14] Yarnold, J. K. (1972) *Asymptotic approximations for the probability that a sum of lattice random vectors lies in a convex set*. Ann. Math. Statist., **43**, 1566–1580.

Performance of acceleration of ALS algorithm in nonlinear PCA

Yuichi Mori, *Okayama University of Science*, mori@soci.ous.ac.jp

Masahiro Kuroda, *Okayama University of Science*, kuroda@soci.ous.ac.jp

Masaya Iizuka, *Okayama University*, iizuka@okayama-u.ac.jp

Michio Sakakihara, *Okayama University of Science*, sakaki@mis.ous.ac.jp

Abstract. Nonlinear principal components analysis with optimal scaling (NLPCA-OS) is useful for analyzing mixed measurement level data. The algorithm in NLPCA-OS is based on the alternating least squares (ALS) algorithm, where optimal transformation and low-rank matrix approximation are alternated until convergence. We have proposed an accelerated ALS algorithm using the vector ε algorithm ($v\varepsilon$ -ALS) which increases the speed of convergence, and have observed that computational costs by $v\varepsilon$ -ALS are less expensive than those by ordinary ALS in small examples in which all variables are categorical. In this paper, we try to evaluate the performance of proposed $v\varepsilon$ -ALS by simulation, in which NLPCA with $v\varepsilon$ -ALS is applied to several simulated datasets which have large numbers of variables with a variety of mixing rates of numerical and categorical variables. The simulation study indicates that the performance of approximation by $v\varepsilon$ -ALS is improved for all simulated datasets and that the larger the number of categorical variables is and the higher the mixing rate is, the more the $v\varepsilon$ -ALS reduces the computational costs.

Keywords. Vector ε algorithm, Acceleration of convergence, Alternating least squares, Mixed measurement level data, Simulation study.

1 Introduction

Nonlinear principal components analysis with optimal scaling (NLPCA-OS) is useful for analyzing mixed measurement level (nominal, ordinal and numerical) data. The algorithm in NLPCA-OS is based on the alternating least squares (ALS) algorithm, where optimal transformation and low-rank matrix approximation are alternated until convergence, that is, the algorithm alternates between optimal scaling for quantifying nominal and ordinal data and ordinary PCA for the optimally scaled data. PRINCIPALS [6] and PRINCALS [1] are the typical ALS algorithms for NLPCA.

Kuroda et al. [2] have proposed an accelerated ALS algorithm for NLPCA using the vector ε ($v\varepsilon$) algorithm of Wynn [7] which increases the speed of convergence. We have applied the method to some numerical examples (e.g., [2] and [3]) and have proposed some more accelerated methods (e.g., two-step algorithm in [5] and re-starting method in [4]), and observed that computational costs of NLPCA with the $v\varepsilon$ alternating least squares ($v\varepsilon$ -ALS) are less expensive than those of NLPCA with ordinary ALS.

In the previous studies, we applied the proposed methods to datasets with small number of variables (the number of variables is 20 at most) and all datasets we used consist of only categorical (nominal) variables but not a mixture of numerical and categorical ones. In this paper, we try to evaluate the performance of $v\varepsilon$ -ALS in further detail to clarify how well the algorithm performs for large data and mixed measurement level data. To do this, we conduct some simulations in which NLPCA with $v\varepsilon$ -ALS is applied to several artificial datasets which have large numbers of variables with a variety of mixing rates of numerical and categorical variables.

We give an overview of NLPCA-OS and its acceleration by $v\varepsilon$ -ALS in Section 2 and illustrate numerical experiments on sixteen different types of datasets generated artificially (four different sizes of datasets with four different mixing rates of categorical variables) in Section 3. We discuss the performance of NLPCA with $v\varepsilon$ -ALS in Section 4.

2 Nonlinear PCA and its acceleration by vector ε ALS

Let $\mathbf{X} = (\mathbf{X}_1 \mathbf{X}_2 \cdots \mathbf{X}_p)$ be an $n \times p$ standardized matrix of observations on n objects and p numerical variables. PCA postulates that \mathbf{X} is approximated by the bilinear form

$$\hat{\mathbf{X}} = \mathbf{Z}\mathbf{A}^\top, \quad (1)$$

where \mathbf{Z} is an $n \times r$ matrix of n component scores on r ($1 \leq r \leq p$) components and \mathbf{A} is a $p \times r$ matrix of p component loadings on r components.

In order to handle any categorical data or mixture of numerical and categorical data, NLPCA requires the optimal scaled data \mathbf{X}^* , in addition to estimating \mathbf{Z} and \mathbf{A} , in which categorical variables in \mathbf{X} are optimally scaled and satisfies restrictions

$$\mathbf{X}^{*\top} \mathbf{1}_n = \mathbf{0}_p \quad \text{and} \quad \text{diag} \left[\frac{\mathbf{X}^{*\top} \mathbf{X}^*}{n} \right] = \mathbf{I}_p, \quad (2)$$

where $\mathbf{1}_n$ and $\mathbf{0}_p$ are vectors of ones and zeros of length n and p , respectively. Thus NLPCA is a least square problem to estimate optimal scaling parameter \mathbf{X}^* and model parameters \mathbf{Z} and \mathbf{A} simultaneously, which minimize

$$\theta = \text{tr}(\mathbf{X}^* - \hat{\mathbf{X}})^\top (\mathbf{X}^* - \hat{\mathbf{X}}) = \text{tr}(\mathbf{X}^* - \mathbf{Z}\mathbf{A}^\top)^\top (\mathbf{X}^* - \mathbf{Z}\mathbf{A}^\top). \quad (3)$$

The ALS algorithm can be used in NLPCA-OS. It alternates between ordinary PCA and optimal scaling, and minimizes θ^* in (3) under restriction (2). For given initial data $\mathbf{X}^{*(0)}$, the procedure based on PRINCIPALS [6] is to iterate the following two steps until convergence:

Step 1 *Model parameter estimation step:* Obtain $\mathbf{A}^{(t)}$ by solving an eigenvalue problem

$$\left[\frac{\mathbf{X}^{*(t)\top} \mathbf{X}^{*(t)}}{n} \right] \mathbf{A} = \mathbf{A} \mathbf{D}_r, \quad (4)$$

where $\mathbf{A}^\top \mathbf{A} = \mathbf{I}_r$ and \mathbf{D}_r is an $r \times r$ diagonal matrix of eigenvalues. Compute $\mathbf{Z}^{(t)}$ from $\mathbf{Z}^{(t)} = \mathbf{X}^{*(t)} \mathbf{A}^{(t)}$.

Step 2 *Optimal scaling step*: Calculate $\hat{\mathbf{X}}^{(t+1)} = \mathbf{Z}^{(t)} \mathbf{A}^{(t)\top}$ from Equation (1). Find $\mathbf{X}^{*(t+1)}$ such that

$$\mathbf{X}^{*(t+1)} = \arg \min_{\mathbf{X}^*} \text{tr}(\mathbf{X}^* - \hat{\mathbf{X}}^{(t+1)})^\top (\mathbf{X}^* - \hat{\mathbf{X}}^{(t+1)})$$

for fixed $\hat{\mathbf{X}}^{(t+1)}$ under measurement restrictions on each of the variables. Since $\mathbf{X}^{*(t+1)}$ is obtained by separately estimating \mathbf{X}_j^* for each j ($j = 1, \dots, p$), scale $\mathbf{X}^{*(t+1)}$ by columnwise centering and normalizing. Re-compute $\mathbf{X}_j^{(t+1)}$ by an additional transformation to keep the monotonicity restriction for ordinal variables and skip this computation for numerical variables.

The superscript (t) indicates the t -th iteration. From the above iteration, we obtain a convergence sequence $\{\mathbf{X}^{*(t)}\}_{t \geq 0}$. Although the true limit points are theoretically obtained at $t = \infty$, the solutions by NLPCA-OS are parameters based on $\mathbf{X}^{*(t)}$ obtained when the iteration converges by the criterion θ .

Here we accelerate the above NLPCA with ALS using the $v\varepsilon$ algorithm of Wynn [7] which is very effective to accelerate the slow convergence of a linearly convergent vector sequence. Let $\{\dot{\mathbf{X}}^{(t)}\}_{t \geq 0} = \{\dot{\mathbf{X}}^{(0)}, \dot{\mathbf{X}}^{(1)}, \dot{\mathbf{X}}^{(2)}, \dots\}$ be the accelerated sequence of $\{\mathbf{X}^{(t)}\}_{t \geq 0}$. We define the inverse of vector \mathbf{X} by $[\mathbf{X}]^{-1} = \mathbf{X} / \langle \mathbf{X}, \mathbf{X} \rangle$, where $\langle \cdot, \cdot \rangle$ is the inner product of vectors. Then, the $v\varepsilon$ algorithm generates $\{\dot{\mathbf{X}}^{(t)}\}_{t \geq 0}$ by using

$$\text{vec} \dot{\mathbf{X}}^{*(t-1)} = \text{vec} \mathbf{X}^{*(t)} + \left[[\text{vec}(\mathbf{X}^{*(t-1)} - \mathbf{X}^{*(t)})]^{-1} + [\text{vec}(\mathbf{X}^{*(t+1)} - \mathbf{X}^{*(t)})]^{-1} \right]^{-1}, \quad (5)$$

where $\text{vec} \mathbf{X}^* = (\mathbf{X}_1^{*\top} \ \mathbf{X}_2^{*\top} \ \dots \ \mathbf{X}_p^{*\top})^\top$. It is expected that this new sequence $\{\dot{\mathbf{X}}^{(t)}\}_{t \geq 0}$ converges to a limit point $\mathbf{X}^{(\infty)}$ of $\{\mathbf{X}^{(t)}\}_{t \geq 0}$ faster than $\{\mathbf{X}^{(t)}\}_{t \geq 0}$. Our previous numerical experiments (e.g., [2], [3], [4] and [5]) demonstrated that its speed of convergence is significantly higher than that of the ordinary ALS algorithm.

The procedure to accelerate the ALS algorithm in PRINCIPALS described above iterates the following two steps:

Step 1 *PRINCIPALS step*: Compute model parameters $\mathbf{A}^{(t)}$ and $\mathbf{Z}^{(t)}$ and determine optimal scaling parameter $\mathbf{X}^{*(t+1)}$.

Step 2 *Acceleration step*: Calculate $\dot{\mathbf{X}}^{*(t-1)}$ using $\{\mathbf{X}^{*(t-1)}, \mathbf{X}^{*(t)}, \mathbf{X}^{*(t+1)}\}$ from Equation (5) and check the convergence by

$$\left\| \text{vec}(\dot{\mathbf{X}}^{*(t-1)} - \dot{\mathbf{X}}^{*(t-2)}) \right\|^2 < \delta, \quad (6)$$

where δ is a desired accuracy.

3 Numerical experiments

We examine the performance of the proposed acceleration for PRINCIPALS using $v\varepsilon$ -ALS by employing simulated data generated as below, and demonstrate the advantage of $v\varepsilon$ accelerated PRINCIPALS ($v\varepsilon$ -ALS in NLPCA) over ordinary PRINCIPALS (ordinary ALS in NLPCA) in terms of the number of iterations and CPU time (in second) required for convergence.

Data generation

Since we are interested in the performance of the proposed $v\varepsilon$ accelerated PRINCIPALS when it is performed for data which have large numbers of variables and include both numerical and categorical variables, we generate random data matrices with the following four types of number of observations (n) and variables (p): (A) $n=100$, $p=20$, (B) $n=100$, $p=50$, (C) $n=500$, $p=100$ and (D) $n=200$, $p=150$. All categorical variables have 10 levels (10 categories). To each of the datasets we further set four kinds of mixing rates of categorical variables: 0.25, 0.50, 0.75 and 1.00. The mixing rate 0.25 means that 25% of variables (rounded) are processed as categorical data and 75% as numerical data, and so on. The number of components (r) is two for all datasets.

We apply ordinary PRINCIPALS and $v\varepsilon$ accelerated PRINCIPALS to the above datasets. Consequently we execute thirty-two types of experiments ($\{\text{four data types}\} \times \{\text{four mixing rates of categorical variables}\} \times \{\text{ALS and } v\varepsilon\text{-ALS}\}$).

Results of experiments

For all experiments, δ for convergence is set to 10^{-12} , and PRINCIPALS terminates when $|\theta^{(t+1)} - \theta^{(t)}| < 10^{-12}$, where $\theta^{(t)}$ is the t -th update of θ calculated from Equation (3). Each algorithm also stops when the number of iterations exceeds 10,000. The procedure is replicated 100 times. All computations are performed with the statistical package R executing on Intel Core i5 3.3 GHz with 4 GB RAM. CPU times taken are measured by the function `proc.time`.

Table 1 is summary statistics of the numbers of iterations from thirty-two 100 simulations. Figure 1 shows the same thirty-two simulations in boxplots. The first graph from the left in Figure 1 displays boxplots of eight simulations ($\{\text{four mixing rates of categorical variables}\} \times \{\text{ALS and } v\varepsilon\text{-ALS}\}$) for data type (A), the second for (B), the third for (C) and the last for (D). The CPU times are similarly summarized in Table 2 and Figure 2.

From these tables and figures, as the data size is increasing, a greater number of iterations and more CPU time are required, but $v\varepsilon$ -ALS greatly reduces the number of iterations and CPU time. In case of 0.25 mixing rate, for example, ordinary ALS needs 178 iterations with 1.9 seconds for dataset (A) ($p=20$) but 639 iterations and 140 seconds for dataset (D) ($p=150$). On the other hand, $v\varepsilon$ -ALS needs 49 iterations and 0.7 seconds for (A) but 188 iterations and 45 seconds for (D). We can observe similar results as the mixing rate of categorical variables increases. The increase of the number of categorical variables requires computational cost and the acceleration by $v\varepsilon$ -ALS is therefore effective. In case of dataset (D) ($p=150$), for example, ordinary ALS needs 639 iterations with 140 seconds for 0.25 mixing rate but 1697 iterations and 805 seconds for 1.00 mixing rate. On the other hand, $v\varepsilon$ -ALS needs 188 iterations and 45 seconds for 0.25 mixing rate but 644 iterations and 314 seconds for 1.00 mixing rate.

It can be observed that the $v\varepsilon$ -ALS converges almost 3 times faster than ordinary ALS in all simulations. The tables also show the average speed-up rates in [SpeedUp] row of each data type, which is computed by dividing the number of iterations (CPU time) required for ordinary ALS divided by the number of iterations (CPU time) required for $v\varepsilon$ -ALS. In Figure 3, we illustrate the speed-up rates of 100 simulations only for dataset (D) in boxplot. The similar boxplots of the speed-up rate can be obtained for other datasets. Regardless of the data size and the mixing rate of categorical variables, $v\varepsilon$ -ALS is smaller 2.62 – 3.61 times of iterations and shorter 2.50 – 3.13 times of CPU time than those of ordinary ALS, although the speed-up rates slightly decrease according to the increase of mixing rate.

Data type	Stats	0.25		0.50		0.75		1.00	
		ALS	$v\epsilon$ -ALS	ALS	$v\epsilon$ -ALS	ALS	$v\epsilon$ -ALS	ALS	$v\epsilon$ -ALS
(A) $n=100$ $p=20$	Min.	51	17	89	29	125	40	180	65
	1st Qu.	88.5	27	150.8	47.75	225.8	72.75	332.5	103.8
	Median	123.5	35	210	60	323	96	474.5	154.5
	Mean	178.2	49.33	302	93.81	448	133.69	605.7	193.8
	3rd Qu.	182.2	50.5	308.8	86	532.5	147.5	737.5	241
	Max.	2687	623	2464	1344	2752	935	3578	820
	[SpeedUp]		[3.61]		[3.22]		[3.35]		[3.13]
(B) $n=100$ $p=50$	Min.	85	26	170	58	254	101	296	94
	1st Qu.	150	48.75	290.5	97.5	476.8	147	586.5	194
	Median	218	68.5	435	133	674	224.5	798	285
	Mean	294.4	90.22	505.7	170.4	780.5	266.7	1032	372.3
	3rd Qu.	334.5	95.25	580.8	208.5	922.5	337.5	1200.5	410
	Max.	1799	409	1777	1267	3717	783	4894	1959
	[SpeedUp]		[3.26]		[2.97]		[2.93]		[2.77]
(C) $n=500$ $p=100$	Min.	83	27	150	53	228	85	307	105
	1st Qu.	181.8	61	308.8	100.2	468.5	152	615.8	227
	Median	261	83	414.5	135	593	207.5	792.5	304.5
	Mean	345.4	103.8	548.2	179.5	697.3	245.8	1147.2	437.6
	3rd Qu.	357.5	111	672.8	216.8	875.2	297.5	1178.5	418.8
	Max.	2187	499	3474	793	1708	1051	10000	2752
	[SpeedUp]		[3.33]		[3.05]		[2.84]		[2.62]
(D) $n=200$ $p=150$	Min.	190	63	378	116	418	179	501	202
	1st Qu.	341	107.8	619.8	224	898	329.8	1058	397
	Median	468.5	143.5	867	329.5	1194	431	1436	569.5
	Mean	639	187.7	1090.3	388.1	1411	527.3	1697	644.2
	3rd Qu.	670.8	212	1297	448.2	1638	601.8	1987	785.5
	Max.	8329	1263	3667	1534	4406	2354	7416	2192
	[SpeedUp]		[3.40]		[2.81]		[2.68]		[2.63]

Table 1: Summary of statistics of the numbers of iterations of ordinary ALS and $v\epsilon$ accelerated ALS for four data types and four mixing rates of categorical variables.

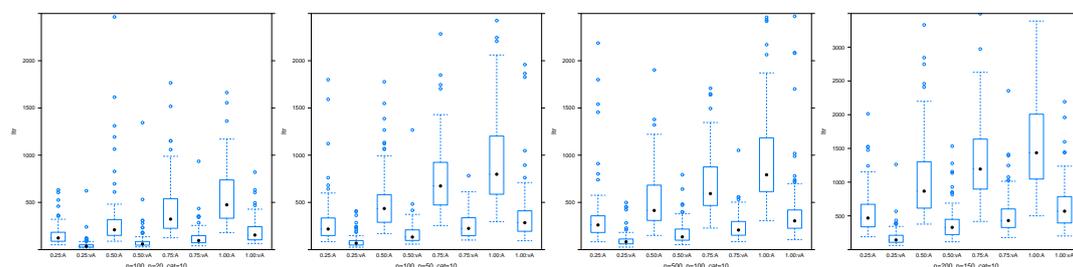


Figure 1: Boxplots of the number of iterations for data type (A) to (D) (from left to right in order).

4 Concluding remarks

In this paper, we examined the performance of the $v\epsilon$ -ALS algorithm which accelerates the convergence of the sequence generated from ordinary ALS. To do this, we applied ordinary ALS and $v\epsilon$ -ALS to several simulated datasets generated from four different data sizes and four different mixing rates of categorical variables. The numerical experiments for comparing the number of iterations and CPU time by ordinary ALS and $v\epsilon$ -ALS demonstrated that the larger the number of categorical variables is and the higher the mixing rate is, the more the $v\epsilon$ -ALS reduces the computational costs. They also indicated that the performance of approximation

Data type	Stats	0.25		0.50		0.75		1.00	
		ALS	$v\epsilon$ -ALS						
(A) $n=100$ $p=20$	Min.	0.66	0.34	1.21	0.55	1.96	0.77	3.07	1.26
	1st Qu.	1.018	0.4475	1.992	0.7775	3.37	1.258	5.593	1.948
	Median	1.35	0.53	2.695	0.92	4.745	1.6	7.855	2.795
	Mean	1.898	0.6744	3.775	1.3441	6.489	2.141	10.015	3.448
	3rd Qu.	1.925	0.685	3.86	1.2525	7.697	2.345	12.133	4.258
	Max.	26.15	6.43	29.44	16.8	38.88	13.83	58.15	13.93
	[SpeedUp]			[2.81]	[2.81]		[3.03]		[2.90]
(B) $n=100$ $p=50$	Min.	2.9	1.22	7.09	2.8	13	5.55	17.75	6.05
	1st Qu.	4.76	1.907	11.81	4.402	23.9	7.875	34.78	12.04
	Median	6.675	2.465	17.42	5.785	33.77	11.675	47.29	17.36
	Mean	8.887	3.104	20.22	7.246	39.02	13.812	60.95	22.53
	3rd Qu.	10.053	3.237	23.17	8.797	46.09	17.288	71.05	24.76
	Max.	52.24	12.5	70.13	50.84	184.89	39.74	287.42	116.24
	[SpeedUp]		[2.86]		[2.79]		[2.83]		[2.71]
(C) $n=500$ $p=100$	Min.	18.68	9.67	39.34	17.85	70.56	30.75	114.2	43.34
	1st Qu.	36.31	15.82	77.09	29.22	142.03	51.48	224	87.26
	Median	50.32	20.01	101.38	37.49	179.84	67.65	285.1	115.35
	Mean	65.14	23.84	132.91	48.49	209.99	79.31	410.4	164.23
	3rd Qu.	66.58	25.62	163.34	57.05	262.47	94.36	420.7	157.34
	Max.	392.5	96.43	825.62	197.59	507.83	323.42	3489.6	991.63
	[SpeedUp]		[2.73]		[2.74]		[2.65]		[2.50]
(D) $n=200$ $p=150$	Min.	43.41	16.78	115.4	38.93	163.9	74.04	238.9	101
	1st Qu.	75.57	26.95	188	72.22	348.6	133.48	501.9	195.4
	Median	103.07	35.03	261.6	104.67	463.4	173.78	681	278.1
	Mean	139.57	44.66	329	122.92	545.9	211.73	805.1	313.9
	3rd Qu.	146.21	49.99	392.7	141.39	630.3	241.1	942.5	382.1
	Max.	1788.1	284.06	1096.7	476.74	1714.3	932.91	3508.4	1062.2
	[SpeedUp]		[3.13]		[2.68]		[2.58]		[2.56]

Table 2: Summary of statistics of CPU times of ordinary ALS and $v\epsilon$ accelerated ALS for four data types and four mixing rates of categorical variables.

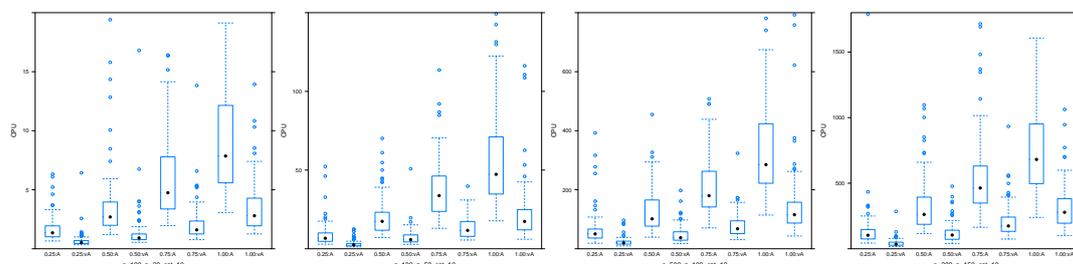


Figure 2: Boxplots of CPU time for data type (A) to (D) (from left to right in order).

by $v\epsilon$ -ALS is improved about 3 times of ordinary ALS for any number of categorical variables in data.

For future problems, we have to investigate how much the proposed acceleration improves computational efficiency when it is applied to more complex situations; such as variable selection problem. Since we are developing faster algorithms (e.g., re-starting ALS in [4]), we are trying to evaluate the performances of such algorithms in detail. Furthermore, there exist many other ALS types of algorithms, so we are attempting to speed up the convergence of their ALS algorithms by incorporating the proposed acceleration.

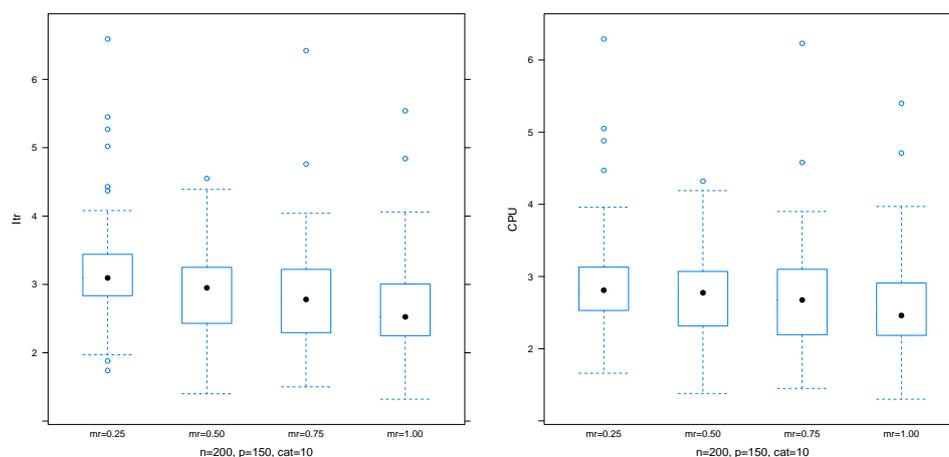


Figure 3: Boxplots of the speed-up rates of 100 simulations for data type (D) (Left: the number of iterations, Right: CPU time).

Acknowledgement

This work is supported by JSPS KAKENHI Grant Numbers 24500353, 26330052.

Bibliography

- [1] Gifi, A. (1990). *Nonlinear multivariate analysis*. John Wiley & Sons, Ltd.,
- [2] Kuroda, M., Mori, Y., Iizuka, M. and Sakakihara, M. (2011). *Accelerating the convergence of the EM algorithm using the vector epsilon algorithm*. *Computational Statistics and Data Analysis*, **55**, 143–153.
- [3] Kuroda, M., Mori, Y., Iizuka, M. and Sakakihara, M. (2012). *Acceleration of convergence of the alternating least squares algorithm for nonlinear principal components analysis*. In *Principal Component Analysis* (Sanguansat, P. (Ed.)), InTech Publications, 129-144.
- [4] Kuroda, M., Mori, Y., Izuka, M. and Sakakihara, M. (2013). *Accelerating and re-starting the alternating least squares algorithm for non-linear principal components analysis*. *Proceedings of the 59th World Statistics Congress*, 5426-5431.
- [5] Kuroda, M., Sakakihara, M., Mori, Y. and Izuka, M. (2012). *Two-stage acceleration for non-linear PCA*. *Proceedings of COMPSTAT 2012*, 461-471.
- [6] Young, F.W., Takane, Y., and de Leeuw, J. (1978). *Principal components of mixed measurement level multivariate data: An alternating least squares method with optimal scaling features*. *Psychometrika*, **43**, 279–281.
- [7] Wynn, P. (1962). *Acceleration techniques for iterated vector and matrix problems*. *Mathematics of Computation*, **16**, 301–322.

Finite-Sample Multivariate Tests for ARCH in Vector Autoregressive Models

Niklas Ahlgren, *Hanken School of Economics*, niklas.ahlgren@hanken.fi
Paul Catani, *Hanken School of Economics*, paul.catani@hanken.fi

Abstract. In this paper we propose finite-sample multivariate tests for ARCH effects in the errors of vector autoregressive (VAR) models using Monte Carlo testing techniques and the bootstrap. The tests under consideration are combined equation-by-equation LM tests, multivariate LM tests and LM tests of constant error covariance matrix. We use a parametric bootstrap to circumvent the problem that the test statistics in VAR models are not free of nuisance parameters under the null hypothesis. The tests are evaluated in simulation experiments and the bootstrap tests are found to have excellent size and power properties. The LM tests of constant error covariance matrix outperform the combined LM tests and multivariate LM tests in terms of power.

Keywords. Conditional heteroskedasticity, Vector autoregressive model, Monte Carlo test, Bootstrap

1 Introduction

The Lagrange multiplier (LM) test for autoregressive conditional heteroskedasticity (ARCH) by [5] is widely used as a diagnostic test in time series models. It is easy to compute from an auxiliary regression involving the squared least squares (LS) residuals. The LM statistic is asymptotically distributed as χ^2 under the null hypothesis. The multivariate generalisation of the test (see e.g. [6]) requires estimating a large number of parameters in the auxiliary regression. The test performs poorly for small and moderate sample sizes, particularly when the dimensions are large (see e.g. [6]). Other multivariate LM tests for ARCH have been proposed (see [3] and [4]), but these have not been much used.

Monte Carlo (MC) test techniques may be used to overcome some of the problems with multivariate tests for ARCH. MC testing techniques deliver exact finite-sample tests in regression models when the regressors are exogenous. In this paper we propose finite-sample multivariate

LM tests for ARCH in vector autoregressive (VAR) models by following a suggestion in [3] of replacing an exact test by a bootstrap test when the model includes lags. Our paper differs from [3] because we consider VAR models instead of regression models with exogenous regressors. We consider multivariate LM tests, whereas [3] consider combined equation-by-equation univariate LM tests and multivariate Portmanteau tests for ARCH. The paper is organised as follows. Multivariate LM tests for ARCH are described in Section 2. The bootstrap algorithm is outlined in Section 3. The results of MC experiments investigating the properties of the tests in finite samples are reported in Section 4. The tests are applied to credit default swap (CDS) prices in Section 5. Section 6 concludes.

2 Multivariate LM Test for ARCH

We consider multivariate LM tests for conditionally heteroskedastic (ARCH) errors in the n -variate vector autoregressive (VAR) model

$$\mathbf{y}_t = \mathbf{\Pi}_1 \mathbf{y}_{t-1} + \cdots + \mathbf{\Pi}_p \mathbf{y}_{t-p} + \mathbf{u}_t, \quad t = 1, \dots, T. \quad (1)$$

The null hypothesis is that the errors \mathbf{u}_t are IID($\mathbf{0}, \mathbf{\Omega}$) against the alternative hypothesis that they are conditionally heteroskedastic: $\mathbf{u}_t = \mathbf{H}_t^{1/2} \varepsilon_t$, where $\mathbf{H}_t = \mathbf{E}(\mathbf{u}_t \mathbf{u}_t' | \mathcal{F}_{t-1})$ is the conditional covariance matrix of the errors \mathbf{u}_t , \mathcal{F}_{t-1} is the σ -field generated by all available information until time $t-1$ and $\{\varepsilon_t\}$ is a sequence of IID($\mathbf{0}, \mathbf{I}_n$) random variables.

Combined Univariate Tests

The Lagrange multiplier (LM) test for ARCH [5] of order h in equation i is a test of $b_1 = \cdots = b_h = 0$ in the auxiliary regression $\hat{u}_{it}^2 = b_0 + b_1 \hat{u}_{i,t-1}^2 + \cdots + b_h \hat{u}_{i,t-h}^2 + e_{it}$. The test statistic has the form $LM_i = TR_i^2$, where R_i^2 is the coefficient of determination in the auxiliary regression for equation i . The LM statistic is asymptotically distributed as $\chi^2(h)$ under the null hypothesis. Following [3], standardised versions of the test statistics are obtained by replacing \hat{u}_{it} by the Cholesky-standardised residual \tilde{w}_{it} . The combined statistic is constructed as follows (see [3]):

$$\widetilde{LM} = 1 - \min_{1 \leq i \leq n} (p(\widetilde{LM}_i)), \quad (2)$$

where $p(\widetilde{LM}_i)$ are the individual p -values associated with the standardised LM statistics \widetilde{LM}_i . The p -values may be derived from the asymptotic distribution of \widetilde{LM}_i , which is a $\chi^2(h)$ distribution. The combined test is closely related to a Bonferroni-type testing procedure, but different from the Bonferroni bound the MC procedure delivers a simulated joint p -value [3].

Multivariate LM Tests

The multivariate LM test for ARCH is based on the auxiliary regression

$$\text{vech}(\hat{\mathbf{u}}_t \hat{\mathbf{u}}_t') = \mathbf{b}_0 + \mathbf{B}_1 \text{vech}(\hat{\mathbf{u}}_{t-1} \hat{\mathbf{u}}_{t-1}') + \cdots + \mathbf{B}_h \text{vech}(\hat{\mathbf{u}}_{t-h} \hat{\mathbf{u}}_{t-h}') + \mathbf{e}_t. \quad (3)$$

The operator vech stacks the elements on and below the main diagonal of an $n \times n$ matrix into a $\frac{1}{2}n(n+1)$ -dimensional vector. The null hypothesis is that $\mathbf{B}_1 = \cdots = \mathbf{B}_h = \mathbf{0}$. The multivariate LM statistic can be shown to be of the form

$$MLM = \frac{1}{2} T n(n+1) - T \text{tr}(\widehat{\mathbf{\Omega}}_{\text{vech}} \widehat{\mathbf{\Omega}}^{-1}), \quad (4)$$

where $\widehat{\boldsymbol{\Omega}}_{\text{vech}}$ is the estimator of the error covariance matrix from the auxiliary model (3) and $\widehat{\boldsymbol{\Omega}} = T^{-1} \sum_{t=1}^T \widehat{\mathbf{u}}_t \widehat{\mathbf{u}}_t'$ is the estimator of the error covariance matrix from the VAR model (1) [6]. Following [3], a standardised version of the test statistic is obtained by replacing $\widehat{\mathbf{u}}_t$ by $\widetilde{\mathbf{w}}_t$, the multivariate standardised residual. The MLM statistic is asymptotically distributed as $\chi^2(hn^2(n+1)^2/4)$ under the null hypothesis.

[4] propose a test for constant error covariance matrix. When testing for ARCH, a suitable alternative is the constant conditional correlation autoregressive conditional heteroskedasticity (CCC-ARCH) process of order h . Then $\mathbf{H}_t = \mathbf{D}_t \mathbf{P} \mathbf{D}_t$, where $\mathbf{D}_t = \text{diag}(h_{1t}^{1/2}, \dots, h_{nt}^{1/2})$ is a diagonal matrix of conditional standard deviations of the errors \mathbf{u}_t . Further, $\mathbf{D}_t^{-1} \mathbf{u}_t = \boldsymbol{\varepsilon}_t$, where $\boldsymbol{\varepsilon}_t \sim \text{IID}(\mathbf{0}, \mathbf{P})$ and \mathbf{P} is a positive definite matrix of conditional correlations. The conditional variance $\mathbf{h}_t = (h_{1t}, \dots, h_{nt})'$ follows a CCC-ARCH(h) process

$$\mathbf{h}_t = \mathbf{a}_0 + \sum_{k=1}^h \mathbf{A}_k \mathbf{u}_{t-k}^{(2)}, \tag{5}$$

where \mathbf{a}_0 is an n -dimensional vector of positive constants, $\mathbf{A}_1, \dots, \mathbf{A}_h$ are $n \times n$ diagonal matrices and $\mathbf{u}_t^{(2)} = (u_{1t}^2, \dots, u_{nt}^2)'$. The null hypothesis is $\mathbf{A}_1 = \dots = \mathbf{A}_h = \mathbf{0}$. The LM statistic is

$$LM_{CCC} = T \bar{\mathbf{s}}_T(\tilde{\boldsymbol{\theta}}) \tilde{\mathbf{I}}_T^{-1}(\tilde{\boldsymbol{\theta}}) \bar{\mathbf{s}}_T(\tilde{\boldsymbol{\theta}}), \tag{6}$$

where $\bar{\mathbf{s}}_T(\tilde{\boldsymbol{\theta}})$ and $\tilde{\mathbf{I}}_T(\tilde{\boldsymbol{\theta}})$ are the relevant blocks of the average score vector and information matrix, respectively, estimated under the null hypothesis (see [4]). The LM_{CCC} statistic is asymptotically distributed as $\chi^2(nh)$ under the null hypothesis.

3 Bootstrap Tests for ARCH

In this section we present the Monte Carlo (MC) testing technique and bootstrap algorithm. [3] develop a framework for MC tests which employs Cholesky-standardised multivariate residuals from the multivariate linear regression model

$$\mathbf{Y} = \mathbf{X} \mathbf{B} + \mathbf{U}, \tag{7}$$

where $\mathbf{Y} = (\mathbf{y}_1, \dots, \mathbf{y}_n)$ is a $T \times n$ matrix, \mathbf{X} is a $T \times k$ matrix of full column rank, \mathbf{B} is a $k \times n$ parameter matrix and $\mathbf{U} = (\mathbf{u}_1, \dots, \mathbf{u}_n)$ is a $T \times n$ matrix of errors. The VAR model can be written in the linear regression form (7) with $\mathbf{X}_t = (\mathbf{y}_{t-1}, \dots, \mathbf{y}_{t-p})'$ a typical row of \mathbf{X} and \mathbf{B} is an $np \times n$ parameter matrix. The distribution of test statistics in the VAR model based on Cholesky-standardised multivariate residuals are not free of nuisance parameters.

The bootstrap algorithm is a modification of the algorithm in [3] to autoregressions. Following a suggestion in [3], the LS estimator $\widehat{\mathbf{B}}$ of \mathbf{B} under the null hypothesis is used in the parametric bootstrap in step 3. The tests based on the parametric bootstrap are not exact in finite samples. They are only exact as the sample size tends to infinity.

Algorithm 3.1.

Bootstrap Monte Carlo tests for ARCH

Step 1 From the observed data, compute \widetilde{LM} in (2) and denote it $\widetilde{LM}^{(0)}$.

Step 2 Obtain N draws from $\mathbf{W}_1, \dots, \mathbf{W}_T \sim \text{NID}(\mathbf{0}, \mathbf{I}_n)$ and denote the drawn variates $\mathbf{W}^{(j)}$, $j = 1, \dots, N$.

Step 3 For each draw j , conditional on the observed regressor matrix \mathbf{X} , the Cholesky factor $\mathbf{S}_{\widehat{\mathbf{U}}}$ of the residuals $\widehat{\mathbf{U}}$ and the LS estimator $\widehat{\mathbf{B}}$ of \mathbf{B} , construct a bootstrap replication

$$\mathbf{Y}^{(j)} = \mathbf{X}\widehat{\mathbf{B}} + \mathbf{W}^{(j)}\mathbf{S}_{\widehat{\mathbf{U}}}, \quad j = 1, \dots, N.$$

Regress $\mathbf{Y}^{(j)}$ on \mathbf{X} and obtain the associated residual matrix $\widehat{\mathbf{U}}^{(j)}$, covariance matrix $\widehat{\mathbf{\Omega}}^{(j)} = T^{-1}\widehat{\mathbf{U}}^{(j)'}\widehat{\mathbf{U}}^{(j)}$ and its Cholesky factor $\mathbf{S}_{\widehat{\mathbf{U}}^{(j)}}$. Obtain the simulated standardised residuals

$$\widetilde{\mathbf{W}}^{(j)} = \widehat{\mathbf{U}}^{(j)}(\mathbf{S}_{\widehat{\mathbf{U}}^{(j)}}^{(j)})^{-1} = (\widetilde{\mathbf{w}}_1^{(j)}, \dots, \widetilde{\mathbf{w}}_n^{(j)}),$$

where $\widetilde{\mathbf{w}}_i^{(j)} = (\widetilde{w}_{i1}^{(j)}, \dots, \widetilde{w}_{iT}^{(j)})'$, $i = 1, \dots, n$.

Step 4 Compute the LM statistic for equation i and MC draw j , denoting it $\widetilde{LM}_i^{(j)}$. Compute $\widetilde{LM}^{(j)} = 1 - \min_{1 \leq i \leq n}(p(\widetilde{LM}_i^{(j)}))$ using (2) as in step 1.

Step 5 Given $\widetilde{LM}^{(j)}$, $j = 1, \dots, N$, compute the number of simulated values greater than or equal to $\widetilde{LM}^{(0)}$ (denoted $N\widehat{G}_N(\widetilde{LM}^{(0)})$). The MC p -value is

$$\widehat{p}_N(\widetilde{LM}) = [N\widehat{G}_N(\widetilde{LM}^{(0)}) + 1]/(N + 1).$$

The null hypothesis is rejected at the significance level α if $\widehat{p}_N(\widetilde{LM}) \leq \alpha$. The same algorithm is used with the multivariate LM tests MLM and LM_{CCC} . The asymptotic validity of the bootstrap tests follows from Theorem 1 of [3] and consistency of the LS estimator $\widehat{\mathbf{B}}$ of \mathbf{B} .

4 Simulations

We conduct Monte Carlo simulations for size and power of the multivariate tests for ARCH with $n = 2$. The samples sizes are $T = 100, 200$ and 400 . The number of Monte Carlo replications is 5000 for $T = 100$ and 200 , and 2000 for $T = 400$. In the bootstrap tests the number of replications is $N = 499$. The model for the conditional mean is a stationary VAR(2) model with $\mathbf{\Pi}_1 = \text{diag}(0.5)$ and $\mathbf{\Pi}_2 = \text{diag}(0.3)$. Five different data generating processes (DGPs) are considered for the errors \mathbf{u}_t . In DGP 1, $\mathbf{u}_t \sim \text{NID}(\mathbf{0}, \mathbf{I}_n)$. In DGPs 2 and 3, \mathbf{u}_t follows CCC-GARCH(1, 1) processes:

$$\mathbf{u}_t = \mathbf{H}_t^{1/2} \varepsilon_t, \quad \mathbf{H}_t = \mathbf{D}_t \mathbf{P} \mathbf{D}_t, \quad \mathbf{D}_t = \text{diag}(h_{1t}^{1/2}, \dots, h_{nt}^{1/2}), \quad \mathbf{h}_t = \mathbf{a}_0 + \mathbf{A}_1 \mathbf{u}_{t-1}^{(2)} + \mathbf{B}_1 \mathbf{h}_{t-1}.$$

Furthermore, $\varepsilon_t \sim \text{NID}(\mathbf{0}, \mathbf{P})$, where $\mathbf{P} = (\rho_{ij})$. The parameter values for \mathbf{A}_1 and \mathbf{B}_1 are $\mathbf{A}_1 = \text{diag}(0.08)$ and $\mathbf{B}_1 = \text{diag}(0.90)$ in DGP 2 and $\mathbf{A}_1 = \text{diag}(0.5)$ and $\mathbf{B}_1 = \mathbf{0}$ in DGP 3. The constant vector \mathbf{a}_0 has all its elements $a_{0i} = 0.02$, $i = 1, \dots, n$. The conditional correlation parameter is $\rho_{ij} = 0.5$ for $i \neq j$. DGP 4 is a diagonal BEKK-GARCH(1, 1) model given by

$$\mathbf{H}_t = \mathbf{C} + \mathbf{A}'_1 \mathbf{u}_{t-1} \mathbf{u}'_{t-1} \mathbf{A}_1 + \mathbf{B}'_1 \mathbf{H}_{t-1} \mathbf{B}_1,$$

where \mathbf{C} is a matrix with elements 0.1 on the main diagonal and off-diagonal elements 0.2, $\mathbf{A}_1 = \text{diag}(\sqrt{0.08})$ and $\mathbf{B}_1 = \text{diag}(\sqrt{0.9})$. DGP 5 is an Extended CCC (ECCC) GARCH model

DGP	1	2	3	4	5	1	2	3	4	5
Test	$h = 2$					$h = 5$				
	$T = 100$									
LM_1	0.035	0.120	0.596	0.120	0.119	0.035	0.145	0.466	0.144	0.148
LM_2	0.030	0.104	0.482	0.111	0.065	0.034	0.123	0.394	0.130	0.063
\widetilde{LM}	0.040	0.165	0.737	0.175	0.137	0.049	0.215	0.636	0.223	0.173
MLM	0.045	0.166	0.736	0.148	0.125	0.039	0.199	0.549	0.169	0.157
MLM^*	0.036	0.143	0.708	0.130	0.110	0.040	0.201	0.551	0.169	0.160
LM_{CCC}	0.024	0.151	0.780	0.144	0.107	0.026	0.183	0.696	0.186	0.137
LM_{CCC}^*	0.040	0.180	0.813	0.178	0.139	0.048	0.235	0.743	0.236	0.176
	$T = 200$									
LM_1	0.039	0.319	0.919	0.312	0.326	0.046	0.393	0.854	0.405	0.414
LM_2	0.039	0.275	0.820	0.275	0.115	0.040	0.361	0.731	0.366	0.144
\widetilde{LM}	0.046	0.432	0.976	0.429	0.329	0.049	0.537	0.936	0.555	0.422
MLM	0.052	0.417	0.977	0.419	0.324	0.049	0.532	0.927	0.520	0.409
MLM^*	0.046	0.389	0.974	0.389	0.301	0.045	0.519	0.921	0.503	0.393
LM_{CCC}	0.038	0.456	0.988	0.441	0.330	0.038	0.563	0.975	0.568	0.424
LM_{CCC}^*	0.046	0.479	0.989	0.463	0.352	0.048	0.586	0.977	0.594	0.449
	$T = 400$									
LM_1	0.046	0.392	0.994	0.652	0.634	0.042	0.519	0.987	0.752	0.767
LM_2	0.042	0.352	0.972	0.588	0.243	0.046	0.466	0.951	0.731	0.305
\widetilde{LM}	0.048	0.533	1.000	0.803	0.640	0.047	0.672	0.997	0.890	0.765
MLM	0.052	0.521	1.000	0.841	0.619	0.048	0.650	0.998	0.925	0.757
MLM^*	0.045	0.498	1.000	0.825	0.598	0.043	0.639	0.997	0.923	0.746
LM_{CCC}	0.042	0.574	1.000	0.837	0.668	0.046	0.702	1.000	0.918	0.776
LM_{CCC}^*	0.049	0.578	1.000	0.846	0.679	0.050	0.714	1.000	0.921	0.787

Table 1: Simulated size and power of LM tests for ARCH when $n = 2$. The nominal significance level is 5%.

similar to DGP 2, but with off-diagonal elements $a_{12} = 0.001$, $a_{21} = b_{12} = 0.004$ and $b_{12} = 0.02$. All estimations and numerical calculations are done using code written in R, version 2.15.2.

Table 1 presents the results for testing against ARCH of orders $h = 2$ and 5 in bivariate models. In addition to the multivariate tests, the table shows the results for the individual LM tests (denoted LM_1 and LM_2 , respectively). The multivariate tests for ARCH tend to be slightly undersized, with the exception of the multivariate LM test MLM . In particular LM_{CCC} is undersized, with size against $h = 2$ of 2.4% when $N = 100$, 3.8% when $N = 200$ and 4.2% when $N = 400$. Bootstrapping the test brings its size closer to the nominal level. Turning to power, we see that LM_{CCC} is the most powerful test when the DGP is a CCC-GARCH model. Despite being slightly undersized in small samples, the asymptotic LM_{CCC} test performs well in terms of power. If the errors are CCC-ARCH (DGP 3), then LM_{CCC} is outperformed only by its bootstrap version LM_{CCC}^* . The test also has the highest power when the correlations are not constant as in DGP 4 and when there is volatility interaction between the errors in DGP 5, although the differences between the tests are small in the latter case. The combined LM test \widetilde{LM} has lower power. The multivariate LM tests MLM and MLM^* have lower power than the

other multivariate tests with the exception DGP 4 and $T = 400$.

Outliers frequently occur in time series with conditional heteroskedasticity. We investigate the robustness of the tests to outliers in DGPs 1 and 2, when $T = 200$ and $h = 2$. Following [7] we consider additive and innovational outliers with outlier parameter $\omega = (3.5, 3.5)'$ and $\omega = (8, 8)'$ (see [7] for details). We concentrate on the case where there is a simultaneous outlier in both series at $t = 101$. We find that all tests are oversized in the presence of additive outliers, in particular the combined test \widetilde{LM} and the multivariate tests MLM and LM_{CCC} have size 17.9%, 29.6% and 22.6%, respectively, when the nominal significance level is 5%. Innovational outliers, on the other hand, have little impact on the size of the tests. The univariate tests and the combined test are slightly undersized, the size of LM_1 , LM_2 and \widetilde{LM} being 2.0%, 3.0% and 3.1% respectively, while the size of the multivariate tests increases by about 1 percentage point compared to the case with no outliers. In DGP 2 for power, \widetilde{LM} , MLM and LM_{CCC} have rejection probabilities 53.5%, 55.1% and 78.5% in the presence of additive outliers and 27.5%, 28.6% and 33.0% in the presence of innovational outliers. The full set of results are reported in the full length paper.

5 Application to Credit Default Swap Prices

We apply the multivariate LM tests for ARCH to VAR models estimated on credit spread and credit default swap (CDS) prices data. We take a subsample of the companies in Table 1 of [1]. The companies in our subsample are Bank of America, Citigroup, Goldman Sachs, Barclays Bank and Vodafone. We use 5-year maturity CDS prices and credit spreads from Datastream. The data are daily observations from 1 January 2009 to 31 January 2012. The number of daily observations for each company is $T = 804$. In addition to the whole sample period, we divide the data into 2 sub-periods of $T = 402$ observations and 4 sub-periods of $T = 201$ observations. The reason for considering sub-periods is that we want to detect differences between the tests for smaller values of T than the full sample size $T = 804$. The lag length of the VAR model is $p = 2$ for Bank of America, $p = 3$ for Citigroup, $p = 3$ for Goldman Sachs, $p = 4$ for Barclays Bank and $p = 3$ for Vodafone. The estimated VAR models contain dummy variables taking the value 1 for the date in question and 0 otherwise: 25 February 2009, 10 April 2009 and 8 June 2009 for Citigroup, 9 April 2009 for Goldman Sachs, 6 February 2009 and 4 June 2009 for Barclays Bank, and 8 June 2009 for Vodafone. There is evidence that the bond and CDS markets share periods of high volatility; for all companies large movements in one series is matched by large movements in the other series. This suggests that multivariate tests for ARCH effects will be more powerful than either univariate tests or combined tests. For the full sample period of $T = 804$ observations, all tests are significant at the 5% level and all tests are significant at the 1% level, except the univariate LM test for $h = 5$ in the equation for the credit spread for Bank of America. In fact, most p -values are either 0.000 or 0.001. For the sub-period of $T = 402$ observations, the multivariate tests MLM , MLM^* , LM_{CCC} and LM_{CCC}^* are almost all significant at the 5% and 1% levels.

The results for sub-periods of $T = 201$ observations (p -values reported in Table 2 for sub-periods 2 and 3) are more interesting from the point of view of being able to detect differences between the tests. We observe that the p -values of the asymptotic MLM tests are larger than the p -values of the bootstrap MLM^* tests, whereas the opposite holds for the asymptotic LM_{CCC} tests and bootstrap LM_{CCC}^* tests, which is in agreement with the findings in the simulations

Company	h	LM_{CDS}	LM_{CS}	MLM	LM_{CCC}	\widetilde{LM}	MLM^*	LM_{CCC}^*
Sub-period 2								
Bank of America	2	0.000	0.044	0.000	0.000	0.001	0.001	0.001
	5	0.000	0.007	0.000	0.000	0.001	0.001	0.001
	10	0.000	0.085	0.000	0.000	0.001	0.001	0.001
Citigroup	2	0.333	0.737	0.014	0.020	0.551	0.033	0.022
	5	0.103	0.803	0.004	0.000	0.167	0.009	0.001
	10	0.280	0.980	0.000	0.000	0.427	0.001	0.001
Goldman Sachs	2	0.053	0.003	0.000	0.000	0.010	0.001	0.001
	5	0.001	0.037	0.000	0.000	0.001	0.001	0.001
	10	0.000	0.221	0.000	0.000	0.002	0.001	0.001
Barclays Bank	2	0.000	0.000	0.000	0.000	0.001	0.001	0.001
	5	0.000	0.000	0.000	0.000	0.001	0.001	0.001
	10	0.000	0.001	0.000	0.000	0.001	0.001	0.001
Vodafone	2	0.000	0.000	0.000	0.000	0.001	0.000	0.001
	5	0.000	0.000	0.000	0.000	0.001	0.001	0.001
	10	0.000	0.008	0.000	0.000	0.001	0.001	0.001
Sub-period 3								
Bank of America	2	0.119	0.013	0.005	0.000	0.020	0.009	0.001
	5	0.148	0.065	0.020	0.000	0.100	0.032	0.001
	10	0.431	0.342	0.020	0.000	0.566	0.021	0.001
Citigroup	2	0.103	0.015	0.005	0.000	0.031	0.014	0.001
	5	0.340	0.036	0.049	0.000	0.064	0.044	0.001
	10	0.029	0.237	0.013	0.000	0.057	0.010	0.001
Goldman Sachs	2	0.136	0.567	0.025	0.063	0.241	0.035	0.050
	5	0.386	0.379	0.093	0.023	0.612	0.100	0.033
	10	0.131	0.015	0.156	0.000	0.026	0.121	0.001
Barclays Bank	2	0.746	0.008	0.039	0.000	0.015	0.052	0.001
	5	0.724	0.070	0.630	0.000	0.120	0.586	0.001
	10	0.959	0.369	0.522	0.008	0.575	0.467	0.007
Vodafone	2	0.093	0.292	0.563	0.003	0.157	0.507	0.004
	5	0.404	0.258	0.373	0.002	0.421	0.356	0.001
	10	0.471	0.708	0.591	0.013	0.689	0.565	0.015

Table 2: Tests for ARCH in the estimated VAR models for CDS prices in sub-periods 2 and 3.

that the MLM test is slightly oversized, whereas LM_{CCC} is conservative. On balance, the univariate tests (denoted LM_{CDS} and LM_{CS} for the CDS and credit spread series respectively) and the combined test only detect ARCH effects in about half of the cases. The multivariate tests find more evidence of ARCH. More rejections are recorded for LM_{CCC} and LM_{CCC}^* than for MLM and MLM^* , and the p -values of the former are smaller than the p -values of the latter. The bootstrap test LM_{CCC}^* finds the most evidence of ARCH.

6 Conclusions

In this paper we have introduced and evaluated multivariate bootstrap tests for ARCH in vector autoregressive models. The tests are based on standardised multivariate least squares residuals and are therefore easy to calculate. The results show that the bootstrap tests outperform the asymptotic tests in terms of both size and power. Our results also show that a less frequently used test against constant conditional correlation GARCH is more powerful than other multivariate LM tests such as combined univariate LM tests and multivariate LM tests which assume no particular alternative to the null hypothesis. The tests are applied to credit default swap (CDS) prices. The multivariate tests find significant ARCH effects in almost all series.

Acknowledgement

P. Catani acknowledges financial support from The Society of Swedish Literature in Finland.

Bibliography

- [1] Blanco, R., Brennan, S. and Marsh, I. W. (2005) *An empirical analysis of the dynamic relation between investment-grade bonds and credit default swaps*. *Journal of Finance* **60**, 2255–2281.
- [2] Dufour, J.-M. (2006) *Monte Carlo tests with nuisance parameters: A general approach to finite-sample inference and nonstandard asymptotics in econometrics*. *Journal of Econometrics* **133**, 443–478.
- [3] Dufour, J.-M., Khalaf, L. and Beaulieu, M.-C. (2010) *Multivariate residual-based finite-sample tests for serial dependence and ARCH effects with applications to asset pricing models*. *Journal of Applied Econometrics* **25**, 263–285.
- [4] Eklund, B. and Teräsvirta, T. (2007) *Testing constancy of the error covariance matrix in vector models*. *Journal of Econometrics* **140**, 753–780.
- [5] Engle, R. F. (1982), *Autoregressive conditional heteroskedasticity with estimates of the variance of United Kingdom inflation*. *Econometrica* **50**, 987–1007.
- [6] Lütkepohl, H. (2006) *New Introduction to Multiple Time Series Analysis*. Berlin: Springer-Verlag.
- [7] Tsay, R. S., Pena, D. and Pankratz, A. E. (2000), *Outliers in multivariate time series*. *Biometrika* **87**, 789–804.

Behaviour of the quality index in acceptance sampling by variables: computation and Monte Carlo simulation

Miguel Casquilho, *Department of Chemical Engineering, Instituto Superior Técnico, Universidade de Lisboa (University of Lisbon)*, mcasquilho@tecnico.ulisboa.pt

Fátima Rosa, *Department of Chemical Engineering, Instituto Superior Técnico, Universidade de Lisboa (University of Lisbon)*, fatimacoelho@tecnico.ulisboa.pt

Abstract. Quality is nowadays indispensable in every activity, but its control has been circumvented by many, because of the statistical technicality of the subject and the apparent uselessness of acceptance sampling (AS), dealt with in this study. With the current computing power and the access to the Internet, the control of Quality can be used where fit. For Gaussian variables and their acceptance sampling *by variables*, the usual standards are based on the quality index, the behaviour of which is addressed. Its computation is reviewed and, as our main objective, made available directly on our open website. As the acceptance criterion is based on the non-central t -distribution, its computation is commented and made available on the Internet, through a computer program prepared for this purpose. A Monte Carlo solution is also provided, which might be used if the computation of the distribution were not feasible.

Keywords. Quality Control, acceptance sampling, inspection by variables, Gaussian variable, international standards, “Form 1”, non-central t -distribution.

1 Fundamentals and scope

Quality has become a necessity in every activity, manufacturing or services, the customer being a driving force that promotes the need for improvement, responsibility, competitiveness. Achieving quality cannot dispense with measurement, hence statistic control. Although many in business circumvent the harder, statistical aspects of Quality and its control, as clearly remarked by Gunter ([8]), there is no way to substitute them. Numerous studies were done by researchers some decades ago, when the theory was being constructed (among many, *e.g.*, [9],

[13], [15]), and, in our own work, we have insisted (*e.g.*, [3], [4]) in the importance of acceptance sampling, which is a sharp example of a technique necessary to control the quality at the frontiers of a production system, unless there is a solid connection with the suppliers or the clients, which fortunately is becoming more frequent. Nowadays, the Internet can facilitate these control actions, as will be seen in the present study.

Acceptance sampling (AS), and statistical process control are the two parts usually considered in Quality Control. In AS *by variables*, the *quality index* is key to take decisions in the case of having to meet a single specification limit, such as it is dealt with in the standards for quality control, such as the American standard of the year 2008 ([2], the successor of [12]) or its equivalent international ISO standard of 2013 ([10]).

In this study, the computations underlying “Form 1” in the standard are addressed and made available to a user on the Internet. Indeed, with the current availability of computing power and access via the Internet, no simplifications are needed, such as the one ingeniously proposed by Hamaker ([9]) in the past. Based on a Gaussian variable that is the quality characteristic of interest, the procedure aims at controlling the quality of a lot by means of a random sample from it, leading to a comparison that dictates the decision, the comparison between the *quality index*, Q , and the *acceptability constant*, k , the lot being accepted if $Q \geq k$, and rejected otherwise. This is, of course, the test of a hypothesis, and the essence of the technique is the knowledge of the distribution of the statistic, Q , to find its critical value, k .

Whenever, contrary to the present case, neither an analytical nor a reasonably feasible numerical solution are available, a Monte Carlo approach is an alternative. Here, as an experimental “confirmation” of the numerical solution, the computing by a Monte Carlo simulation is also presented, both paths (numerical and simulated) being made solvable in our websites.

2 Acceptance criterion

The classical equations for Type I and Type II errors underlying inspection sampling lead to the sampling plan, *i.e.*, sample size, n , and acceptability constant, k (*e.g.*, [3]):

$$\begin{cases} P_{ac}(\varpi = \text{AQL}) = 1 - \alpha \\ P_{ac}(\varpi = \text{LTPD}) = \beta, \end{cases} \quad (1)$$

in which P_{ac} is the probability of acceptance, ϖ is the fraction defective (currently “fraction nonconforming”), AQL (“Acceptance Quality Limit”⁹) is the maximum percent defective (traditional for “fraction nonconforming”) with probability $1 - \alpha$, and LTPD (“Lot Tolerance Percent Defective”) is the maximum fraction defective for β . Values such as the following can be found: AQL = 1.5%, $\alpha = 5\%$, LTPD = 12%, $\beta = 10\%$. The acceptance criterion is given by the following condition, in which \bar{X} and S are, respectively, the sample average and standard deviation.

$$Q = \frac{\bar{X} - L}{S} \geq k \quad (2)$$

where Q is the quality index, and k the acceptability constant. (The *equals* sign is important because the practical comparison is made with Q rounded to the same number of significant figures as in k .) Note that a different definition of the quality index,

⁹Instead of the historical “Acceptable Quality Level”.

$$Q' = \frac{\bar{X} - L}{S/\sqrt{n}} = Q\sqrt{n} \geq k' \quad (3)$$

might be more natural from a statistician's standpoint. This, making $Q' \equiv t$ (t in Eq. 4, below), would lead directly to a known distribution. Then, *e.g.*, for $n = 10$ and AQL = 4%, instead of testing $Q \geq 1.23$ (from Table 1), it would be $Q' \geq 3.89$ ($= 1.23\sqrt{10}$), which would, however, make the calculations more cumbersome to the user. It has, nevertheless, since long been the laudable intention of the founders of Quality, themselves statisticians, to make its rules easily applicable by the general users.

When both specification limits, lower, L , and upper, U , are present, it is usual to write Q_L and Q_U , respectively, and in the case of the upper limit the numerator is changed to the "symmetric" $U - \bar{X}$, for obvious practical convenience. We arbitrarily chose here to work with the lower specification limit just because it is closer to $(\bar{X} - \mu)/S$. (The case of both limits is, however, out of our present scope, and constitutes the so-called "Form 2".)

Regarding the distribution of the statistic Q , the seminal work by Resnikoff and Lieberman ([11]) and the recent synthesis book by Schilling and Neubauer ([14]) are solid sources: Q is related to t , such that it is $t = Q\sqrt{n}$, and t follows a non-central t -distribution,

$$t \sim F(t; \nu, \delta) \quad (4)$$

where the parameters are, in this application, given by

$$\begin{aligned} \nu &= n - 1 \\ \delta &= \sqrt{n} \Phi^{-1}(1 - \text{AQL}) \end{aligned} \quad (5)$$

with Φ the standard Gaussian distribution (and Φ^{-1} its inverse function). From this will come the acceptance criterion, for a single specification limit (in this case, the lower one), as mentioned.

For the non-central t -distribution, the cumulative distribution function (cdf), F , can be given by (*e.g.*, [1])

$$F(t; \nu, \delta) = C_\nu \int_0^\infty \Phi(tu/\sqrt{\nu} - \delta) u^{\nu-1} e^{-u^2/2} du \quad (6)$$

with

$$C_\nu = [\Gamma(\nu/2) 2^{\nu/2-1}]^{-1} \quad (7)$$

where $\Gamma(\cdot)$ is the common Gamma function (extension of the factorial).

The criterion in Eq. 2 depends on the computation of the non-central t . This computation is described in the next section, and, as is the objective of this study, is made available on our website.

3 Computation

The computing of Eq. 6 was done numerically, as it appears that the integral in it is not amenable to analytical treatment, with the details as follows.

The computation of C_ν is easy, as the argument of the Gamma function in this application ($\nu/2$, with $\nu = n - 1$) is always a positive integer or half-integer, so the function becomes a factorial or a simple multiple of $\sqrt{\pi}$.

Inside of the integrand, for the Gaussian integral, $\Phi(\cdot)$, in computer languages such as the one used, Fortran 90, we relied on

$$\Phi(z) = \frac{1}{2} \left[1 + \operatorname{erf}\left(\frac{z}{\sqrt{2}}\right) \right] \quad (8)$$

where 'erf' is the *error function*, as given by the compiler.

The computation of Eq. 6 thus would boil down to a common numerical integration were it not for the upper integration limit (∞). For the purpose of this study, this difficulty was solved by determining experimentally which would be a "sufficiently" large value, say, M , for the upper integration limit (instead of ∞) for various possible, more or less favourable combinations of t , ν , and δ . We did not delve into a deeper search for accuracy in the values of k , provided full agreement is obtained, as the values available for comparison are just those in the standards, which of course are rounded.

The integral in Eq. 6 achieved subsequent agreement (all the figures in the numbers) with the values of k in Table B-1 reproduced in Table 1, for an upper integration limit of

$$M = \lfloor \sqrt{5n} \rceil \quad (9)$$

(meaning rounding or nearest integer), integer for simplicity, which proved sufficient for all the sample sizes tested. This heuristic expression that we propose for M is inspired in the fact that Resnikoff and Lieberman ([11]) present their tables of F (Eq. 4) as a function of (in their notation) x/\sqrt{f} , with f the degrees of freedom (ν here), and a still more prudent (greater) $n = \nu + 1$.

n	α	M	AQL = 1.50	AQL = 4.00
7	10 %	6	1.50	1.15
10	10 %	7	1.58	1.23
20	7.8 %	10	1.69	1.33
35	6 %	13	1.76	1.39
50	5 %	16	1.80	1.42

Table 1: Values of k from Table B-1 in [2] (AQL in %), with their underlying values of α , for verification.

In order to verify the values in Table 1, the computation of k as a function of n , AQL, and the adequate α , with the proposed M (or a user-supplied one), can be done at our dedicated webpage ([6]), through a computer program of ours that computes the non-central t -distribution. (The computing done on the website, *i.e.*, simply using a browser, is limited to about 30 s.) The limit M , with an integration step of 2×10^{-3} , led to computing times of about 10–70 s, in the Computing Center system of IST (CIIST) with Amd64 machines, at 2 GHz, running Debian Linux. Preliminary computing experiments were done in a parallel, MPI computing system (Milipeia), but the system was not considered indispensable, all the more because it is not anonymously accessible via the Internet. The verification can also be experimentally done, by Monte Carlo simulation, with numerical agreement, in another webpage of ours ([5]), a concomitant website, [7], being available for the above cited "Form 2".

TABLE B-1 **Standard Deviation Method**
Master Table for Normal and Tightened Inspection for Plans Based on Variability Unknown
(Single Specification Limit—Form 1)

Sample size code letter	Sample size	Acceptable Quality Levels (normal inspection)											
		T	.10	.15	.25	.40	.65	1.00	1.50	2.50	4.00	6.50	10.00
		k	k	k	k	k	k	k	k	k	k	k	k
B	3	↓	↓	↓	↓	↓	↓	↓	↓	1.12	.958	.765	.566
C	4	↓	↓	↓	↓	↓	↓	1.45	1.34	1.17	1.01	.814	.617
D	5	↓	↓	↓	↓	↓	1.65	1.53	1.40	1.24	1.07	.874	.675
E	7	↓	↓	↓	2.00	1.88	1.75	1.62	1.50	1.33	1.15	.955	.755
F	10	↓	↓	2.24	2.11	1.98	1.84	1.72	1.58	1.41	1.23	1.03	.828
G	15	2.53	2.42	2.32	2.20	2.06	1.91	1.79	1.65	1.47	1.30	1.09	.886
H	20	2.58	2.47	2.36	2.24	2.11	1.96	1.82	1.69	1.51	1.33	1.12	.917
I	25	2.61	2.50	2.40	2.26	2.14	1.98	1.85	1.72	1.53	1.35	1.14	.936
J	35	2.65	2.54	2.45	2.31	2.18	2.03	1.89	1.76	1.57	1.39	1.18	.969
K	50	2.71	2.60	2.50	2.35	2.22	2.08	1.93	1.80	1.61	1.42	1.21	1.00
L	75	2.77	2.66	2.55	2.41	2.27	2.12	1.98	1.84	1.65	1.46	1.24	1.03
M	100	2.80	2.69	2.58	2.43	2.29	2.14	2.00	1.86	1.67	1.48	1.26	1.05
N	150	2.84	2.73	2.61	2.47	2.33	2.18	2.03	1.89	1.70	1.51	1.29	1.07
P	200	2.85	2.73	2.62	2.47	2.33	2.18	2.04	1.89	1.70	1.51	1.29	1.07
		.10	.15	.25	.40	.65	1.00	1.50	2.50	4.00	6.50	10.00	
Acceptable Quality Levels (tightened inspection)													

All AQL values are in percent nonconforming. T denotes plan used exclusively on tightened inspection and provides symbol for identification of appropriate OC curve.

Figure 1: Table B-1 in [2] of k , given “sample size” (n) and “acceptable quality level” (AQL, currently *acceptance quality limit*).

Conclusions

We think that the rigorous tools of Statistics as applied to Quality Control (QC) must be brought to general attention, after an epoch in which they have been circumvented to facilitate the matters of Quality. The advent of ubiquitous computing power, namely through the Internet, makes QC accessible, even to non-specialists. Acceptance sampling (AS), one of the two branches of statistical, the other being Statistical Process Control, can and should nowadays be applied without restraints.

AS *by variables*, for the typical Gaussian random variable, was shown in its basic “Form 1”, as described in the generally adopted international standards, where a conveniently simple criterion is available to the decision maker. The underlying computations were explained, and an open website is available to anyone needing to assess a quality index, Q , in its comparison with its critical value, the acceptability constant, k .

Acknowledgement

We thank the Department of Chemical Engineering, Instituto Superior Técnico (IST), and CERENA, “Centro de Recursos Naturais e Ambiente” (*Centre for Natural Resources and the Environment*), which has included “Centro de Processos Químicos” (*Centre for Chemical Pro-*

cesses), at IST, Universidade de Lisboa (*University of Lisbon*), Lisbon, Portugal. We also thank CIIST (*Computing Centre of IST*), and “Milipeia” (Laboratory for Advanced Computing), University of Coimbra. We also thank the reviewers, whose comments contributed to improve this text.

Bibliography

- [1] Amos, D. E. (1964) *Representations of the central and non-central t-distributions*. *Biometrika*, **51**, 451–458.
- [2] ANSI/ASQC Z1.9-2008 (2008) *Sampling procedures and tables for inspection by variables for percent nonconforming*. ASQ, Milwaukee, WI (USA).
- [3] Carolino, E., Casquilho, M., and Barão, M. I. (2007) *Amostragem de aceitação para uma variável assimétrica, a exponencial*. (Acceptance sampling for an asymmetric variable, the Exponential.) Proceedings of the “XIV Congresso Anual da Sociedade Portuguesa de Estatística (XIV Annual Congress of the Portuguese Statistical Society), Covilhã (Portugal), 281–292.
- [4] Casquilho, M. (2010) *Numerical integration in tabular form*. ICEE-2010, Proceedings of the “International Conference on Engineering Education”, Gliwice (Poland).
- [5] Casquilho, M. (2012) *Convergence to non-central t curve.*, <http://web.tecnico.ulisboa.pt/~mcasquilho/compute/qc/F-tncConverg.php>, accessed 2014-Jan-15.
- [6] Casquilho, M. (2013) *Acceptance sampling by variables*, <http://web.tecnico.ulisboa.pt/mcasquilho/compute/qc/Fx-ASbyvariables.php>, accessed 2014-Jan-15.
- [7] Casquilho, M. (2014) *Transformation of Q into ϖ in acceptance sampling by variables*, <http://web.tecnico.ulisboa.pt/mcasquilho/compute/qc/Fx-ASvarQtoLotFD.php>, accessed 2014-Jan-15.
- [8] Gunter, B. (1998) *Farewell fusillade*. *Quality Progress*, **31**, No. 4, 111–114.
- [9] Hamaker, H. C. (1979) *Acceptance sampling for percent defective by variables and by attributes*. *Journal of Quality Technology*, **11**, 3, 139–148.
- [10] ISO 3951-2:2013 (2013) *Sampling procedures for inspection by variables – Part 2: General specification for single sampling plans indexed by acceptance quality limit (AQL) for lot-by-lot inspection of independent quality characteristics*. ISO, International Organization for Standardization, Geneva (Switzerland).
- [11] Resnikoff, G. J., and Lieberman, G. J. (1957) *Tables of the non-central t-distribution: density function, cumulative distribution function, and percentage points.*, Stanford University Press, Redwood City, CA (USA).
- [12] MIL-STD-414 (1957) *Sampling procedures and tables for inspection by variables for percent defective*. Office of the Assistant Secretary of Defense, Washington, D.C. (USA).

- [13] Owen, D. B. (1969) *Summary of recent work on variables acceptance sampling with emphasis on non-normality*. *Technometrics*, **11**, 4, 631–637.
- [14] Schilling, E. G., Neubauer, D. V. (2009) *Acceptance sampling in Quality Control*, 2.nd ed., Chapman and Hall / CRC, New York, NY (USA).
- [15] Wetherill, G. B., and Chiu, W. K. (1975) *A review of acceptance sampling schemes with emphasis on the economic aspect*. *International Statistical Review / Revue Internationale de Statistique*, **43**, 2, 191–210.

Sparse exploratory factor analysis

Sara Fontanella, *Open University*, Sara.Fontanella@open.ac.uk

Nickolay Trendafilov, *Open University*, Nickolay.Trendafilov@open.ac.uk

Kohei Adachi, *Osaka University*, adachi@hus.osaka-u.ac.jp

Abstract. Sparse principal component analysis is a very active research area in the last decade. In the same time, there are very few works on sparse factor analysis. We propose a new contribution to the area by exploring a procedure for sparse factor analysis where the unknown parameters are found simultaneously.

Keywords. ℓ_1 penalties, Matrix manifolds, Projected gradients.

1 Introduction

Exploratory factor analysis (EFA) is a model-based multivariate technique that aims to explain the relationships among p manifest random variables by r ($\ll p$) latent random variables called *common* factors. The EFA model assumes that some portion of the variation of each observed variable remains unaccounted for by the common factors. Thus, p additional latent variables called *unique* factors are introduced, each of which accounts for this portion of variance of the corresponding manifest variable [12]. In formal terms, the EFA model represents/approximates a given $n \times p$ data matrix Z of p observed (standardized) variables on n observations as a linear combination of r common and p unique factors

$$Z \approx F\Lambda^\top + U\Psi, \quad (1)$$

where Λ is a $p \times r$ parameter matrix of *factor loadings*. The choice of r is either subjective or based on preliminary validation. In both case its value is subject to some limitations [12]. The r -factor model (1) assumes that all involved random variables (Z , F and U) have zero means and unit variances, and that both common and unique factors are uncorrelated. Most importantly, they are also assumed *mutually* uncorrelated, and the $p \times p$ matrix Ψ is assumed diagonal with *non-zero* diagonal entries. Following the r -model defined above and the assumptions made, it can be found that the sample correlation matrix R is presented/approximated by EFA as:

$$R \approx R_{ZZ} = \Lambda\Lambda^\top + \Psi^2. \quad (2)$$

Thus, the main problem of EFA is to find the pair $\{\Lambda, \Psi\}$ which gives the best fit in some sense to the sample correlation matrix R (for certain r). If the data are assumed normally

distributed the maximum likelihood principle can be applied [12]. Then, finding $\{\Lambda, \Psi\}$ can be formulated as minimizing the following negative loglikelihood function [9, 12]:

$$\min_{\Lambda, \Psi} \log(\det(\Lambda\Lambda^T + \Psi^2)) + \text{trace}((\Lambda\Lambda^T + \Psi^2)^{-1}R), \quad (3)$$

which for short is called ML-EFA.

If nothing is assumed about the distribution of the data, the loglikelihood function (3) can still be used as a measure of the discrepancy between the model and the sample correlation matrices, R_{ZZ} and R . There are a number of other discrepancy measures [9] which are used in place of (3). A natural choice is the least squares approach for fitting the factor analysis model (2), which can be formulated as the following general class of weighted least squares problems:

$$\min_{\Lambda, \Psi} \|(R - \Lambda\Lambda^T - \Psi^2)V\|^2, \quad (4)$$

where V is a matrix of weights, and $\| \cdot \|$ denotes the Frobenius matrix norm $\|A\|^2 = \text{trace}A^T A$. The case of $V = I_p$ is known as the least squares factor analysis, LS-EFA. The second special case $V = R^{-1}$, is known as the generalized least squares problem, GLS-EFA.

The minimization problems ML, LS and GLS listed above are not *unconstrained*. The unknowns Λ and Ψ are sought subject to the following constraints [9]: for ML and GLS,

$$\Lambda^T \Psi^{-2} \Lambda \text{ to be diagonal}, \quad (5)$$

and for LS,

$$\Lambda^T \Lambda \text{ to be diagonal}. \quad (6)$$

The constraint (5) explains why Ψ is required by EFA to have non-zero diagonal entries. This assumption is equivalent to the assertion that no observable random variable can ever be explained entirely by a common factor. This assumption and several other features, e.g. factor scores indeterminacy [12], make the EFA model highly controversial, which probably explains why EFA is far less popular dimension reduction technique than principal components (PCA).

For any orthogonal $r \times r$ matrix Q we have:

$$R_{ZZ} = \Lambda\Lambda^T + \Psi^2 = \Lambda Q Q^T \Lambda^T + \Psi^2 = \Lambda Q (\Lambda Q)^T + \Psi^2, \quad (7)$$

which is known as the rotation indeterminacy in EFA. Indeed, the constraint (5) eliminates the indeterminacy (7), however such solutions are usually difficult for interpretation. Instead, the common practice is to make use of (7): rotate the initially found factor loadings Λ by some kind of “simple structure” rotation [12] to make them more interpretable. By “interpretable” it is meant that each factor has only few large loadings. The rule is to ignore, effectively make *zero*, the remaining rather small ones. In fact, the factor loadings interpretation relies on artificially constructed *sparse* loadings Λ , many of which are neglected, and thus considered zeros.

We propose to modify the EFA fitting problems (3) and (4) by introducing sparse-inducing constraints. Then, the resulting factor loadings Λ will be sparse in an optimal way. This strategy is not new. The same interpretation problem occurs in PCA. Its solution led in the last decade to developing a great number of new procedures directly producing sparse component loadings, which considerably simplifies their interpretation. In contrast, there are very few works on sparse EFA, e.g. [3, 13]. The proposed work will be a further contribution to this new research area.

2 New EFA parameters

It has been argued in [15], that, in fact, the constraints (5) and (6) facilitate the algorithms for numerical solution of the different EFA definitions (3) and (4), see for details e.g. [9, 12]. As we mentioned, occasionally (5) and (6) may facilitate the interpretation of Λ , but in general this is not the case. The alternative traditional approach to rotate the initial factor loadings Λ to “simple structure” gives, in turn, rotated factor loading violating (5) and (6).

In this work we adopt the new formulation of the EFA estimation problems (3) and (4) proposed in [15]. The constraints (5) and (6) will not be needed any more. The only *natural* constraints inferred from the r -factor analysis model (2) are that the $p \times r$ matrix Λ should have full column rank, and that the $p \times p$ diagonal matrix Ψ^2 should be positive definite. Additionally, we relax the second condition and assume positive *semi*-definite diagonal Ψ^2 . There are two reasons for this. From EFA model point of view this constraint seems too restrictive. From numerical point of view the algorithms developed in [15] do not rely on $\Psi^2 > 0$. Moreover, maintaining $\Psi^2 > 0$ may contradict to achieving high level of sparseness (Section 5).

Consider the eigenvalue decomposition of the positive semi definite $\Lambda\Lambda^T$ of rank at most r in (2), i.e. let $\Lambda\Lambda^T = QD^2Q^T$, where D^2 is an $r \times r$ diagonal matrix composed by the largest (nonnegative) r eigenvalues of $\Lambda\Lambda^T$ arranged in descending order and Q is a $p \times r$ orthonormal matrix containing the corresponding eigenvectors. Note that for this reparameterization $\Lambda^T\Lambda$ is diagonal, i.e. the condition (6) is fulfilled automatically. Then (2) can be rewritten as:

$$R_{ZZ} = QD^2Q^T + \Psi^2 . \quad (8)$$

Thus, instead of the pair $\{\Lambda, \Psi\}$, a triple $\{Q, D, \Psi\}$ is sought in [15]. Note, that the model (8) does not permit rotations, only permutations are possible. Thus, the new factor loadings Λ are given by QD . Clearly, when Q is sparse, Λ will have the same sparseness. In order to maintain the factor analysis constraints, the triple $\{Q, D, \Psi\}$ should be sought such that Q be an $p \times r$ orthonormal matrix, and D and Ψ – diagonal. Note, that we do not insist for non-singular Ψ , however the singularity of D implies failing of the r -factor analysis model.

The new formulation of the factor analysis estimation problems is straightforward. Indeed, for a given sample correlation matrix R , the ML-EFA is reformulated as follows:

$$\min_{Q, D, \Psi} \log(\det(QD^2Q^T + \Psi^2)) + \text{trace}((QD^2Q^T + \Psi^2)^{-1}R) , \quad (9)$$

and the LS- and the GLS-EFA estimation problems are rewritten as:

$$\min_{Q, D, \Psi} \|(R - QD^2Q^T - \Psi^2)V\|^2 . \quad (10)$$

3 Sparse factor loadings

Let q_i denote the i th column of Q , i.e. $Q = (q_1, q_2, \dots, q_r)$, and $\tau = (\tau_1, \tau_2, \dots, \tau_r)$ be a vector of tuning parameters, one for each column of Q . We consider a penalized version of EFA, where the ℓ_1 norm of each of the columns of Q is penalized, i.e. $\|q_i\|_1 \leq \tau_i$ for all $i = 1, 2, \dots, r$. Introduce the following discrepancy vector $q_\tau = (\|q_1\|_1, \|q_2\|_1, \dots, \|q_r\|_1) - \tau$, which can also be expressed as $q_\tau = 1_p^\top [Q \odot \text{sign}(Q)] - \tau$, where $\text{sign}(Q)$ is a matrix containing the signs of the elements of Q , and 1_p is a vector with p unit elements. We adapt the scalar penalty function $\max\{x, 0\}$

used by [16] to introduce the following vector penalty function $P_\tau(Q) = [q_\tau \odot (1_p + \text{sign}(q_\tau))]/2$. Then, the penalized versions of (9) and (10) can be defined, for the ML-EFA as:

$$\min_{Q,D,\Psi} \log(\det(QD^2Q^T + \Psi^2)) + \text{trace}((QD^2Q^T + \Psi^2)^{-1}R) + P_\tau(Q)^\top P_\tau(Q), \quad (11)$$

and for the LS- and the GLS-EFA as:

$$\min_{Q,D,\Psi} \|(R - QD^2Q^T - \Psi^2)V\|^2 + P_\tau(Q)^\top P_\tau(Q). \quad (12)$$

Note, that $P_\tau(Q)^\top P_\tau(Q)$ penalizes the sum of squares of $\|q_i\|_1 - \tau_i$ for all $i = 1, 2, \dots, r$, i.e. precise fit of $\|q_i\|_1$ to each tuning parameter τ_i cannot be achieved.

4 Gradients and Stiefel gradients

The gradients of the ML-, LS- and GLS-EFA objective functions with respect to the unknowns $\{Q, D, \Psi\}$ are given in [15] as the following block-matrix: $(-YQD^2, -Q^T YQ \odot D, -Y \odot \Psi)$. For ML-EFA, one has $Y = 2R_{ZZ}^{-1}(R - R_{ZZ})R_{ZZ}^{-1}$, and for LS- and GLS-EFA it changes to $Y = 4V(R - R_{ZZ})V$. Now we need to find the gradient ∇_Q of the penalty term $P_\tau(Q)^\top P_\tau(Q)$ with respect to Q , which should be added to $-YQD^2$.

Making use of the identity $\text{trace}(A \odot B)C = \text{trace}A(B^\top \odot C)$, we find that:

$$\nabla_Q = \frac{1}{2}W \odot [1_p(w \odot P_\tau)], \quad (13)$$

where 1_p is a $p \times 1$ vector and $1_{p \times r}$ is a $p \times r$ matrix with unit entries, and

$$w = 1_p + \text{th}(\gamma q_\tau) + (\gamma q_\tau) \odot [1_p - \text{th}^2(\gamma q_\tau)], \quad (14)$$

and

$$W = \text{th}(\gamma Q) + (\gamma Q) \odot [1_{p \times r} - \text{th}^2(\gamma Q)]. \quad (15)$$

The dynamical system approach employed in [15] can be readily applied for solving (11) and (12). It involves numerical integration of matrix ordinary differential equations (ODE) for $\{Q, D, \Psi\}$ defined by their projected gradients. Particularly, it involves projected gradient dynamical system for Q on the Stiefel manifold of all $p \times r$ orthonormal matrices. There exist a number of specialized numerical methods for solving such problem, e.g. [4] and others listed in [15]. In contrast to the standard EFA alternating approaches [9, 12], the dynamical system approach gives matrix algorithms which produce *simultaneous* solution for $\{Q, D, \Psi\}$ exploiting the geometry of their specific matrix structures. Moreover, such algorithms are *globally* convergent, i.e. the convergence is reached *independently* of the starting (initial) point.

The numerical ODE solvers currently available in **MATLAB** [11] are not suitable for solving large optimization problems. They track the whole trajectory defined by the ODE which is time-consuming and undesirable when the asymptotic state is of interest only. This limits the application of the proposed approach to solving (11) and (12) for rather small data sets.

An alternative way is to employ iterative algorithms directly working on matrix manifolds [1, 5, 17]. The listed above gradients can be readily used for solving (11) and (12) by employing **MANOPT**, a free **MATLAB**-based software for optimization on matrix manifolds [2].

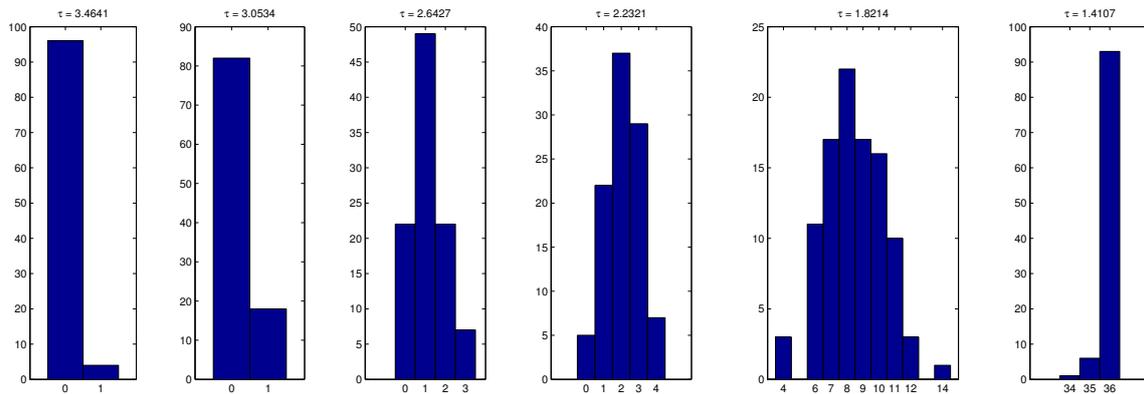


Figure 1: Number of zeros obtained in 100 runs of sparse ML-EFA (11) for different τ .

5 Numerical examples

In this Section we first explore the behavior of the proposed sparse EFA on simulated data considered in [3]. Then, in contrast to [3, 13], we consider two examples from the classic EFA.

Simulated data [3]

We examine the performance of the proposed approach by employing the simulated data constructed in [3]. They take a hypothetical 12×4 sparse loadings matrix Λ with the following non-zero entries: $\lambda_{11} = \lambda_{21} = \lambda_{31} = 1.8$, $\lambda_{42} = \lambda_{52} = \lambda_{62} = 1.7$, $\lambda_{73} = \lambda_{83} = \lambda_{93} = 1.6$ and $\lambda_{10,4} = \lambda_{11,4} = \lambda_{12,4} = 1.5$, and $\Psi^2 = \text{Diag}(1.27, .61, .74, .88, .65, .81, .74, 1.3, 1.35, .74, .92, 1.32)$. The "population" covariance matrix is created by (2), and then we normalize it to obtain a correlation matrix used to generate normally distributed zero mean independent samples.

We generate 100 data matrices each of which is analyzed by sparse ML-EFA. For this reason we solve (11) for six decreasing values of $\tau (= \sqrt{12}, 3.0534, 2.6427, 2.2321, 1.8214, 1.4107)$. The solution for any particular τ is used as a starting value for the next run with the consecutive τ . The starting values for the first $\tau (= \sqrt{12} = 3.4641)$ are chosen randomly. The number of the zero loadings among all $12 \times 4 = 48$ for each τ are depicted in Figure 1. For $\tau = \sqrt{12}$, nearly all factor loadings matrices are dense, only 4 of them contain a single zero entry. For $\tau = 2.6427$, there are 22 factor loadings matrices with no zero entry, 49 – with a single zero entry, 22 – with two zero entries, and the rest seven have three zero loadings. For $\tau = 1.4107$, there are 93 factor loadings matrices with 36 zero entries, 6 – with a 35 zeros, and only one – with 34 zero entries. In other words, with $\tau = 1.4107$ the sparse ML-EFA achieves 93% exact recovery of the underlying sparseness. The case $\tau = 1$ is not depicted, as it produces excessive sparseness. Clearly, the correct tuning parameter for this problem is around $\tau = 1.4107$. After the correct sparseness is localized, one can perform further runs to achieve the best corresponding fit.

Harman's Five Socio-Economic Variables [8, p.14]

First, we illustrate the proposed procedures for sparse EFA on a well known data set from classic EFA, namely the Harman's Five Socio-Economic Variables [8, p.14]. This small data set is interesting because the two- and the three-factor solutions from LS- and ML-EFA are 'Heywood cases' [8, 12], i.e. Ψ^2 contains zero diagonal entries, or $\Psi^2 \geq 0$. One-factor solution is not considered interesting as it explains only 57.47% of the total variance.

Table 1 contains several sparse LS-EFA solutions of (12) starting with $\tau = \sqrt{5} = 2.2361$, which is equivalent to the standard (non sparse) LS-EFA solution. For all of them we have $\Psi^2 \geq 0$. Clearly, POP, EMPLOY and HOUSE tend to be explained by the common factors only, which is already suggested by the non sparse solution ($\tau = \sqrt{5}$). Increasing the sparseness of the factor loadings results in variables entirely explained by either a common or unique factor. The presence of loadings with magnitudes over 1 demonstrates the well known weakness of LS-EFA in fitting the unit diagonal of a correlation matrix. It is well known that ML-EFA does not exhibit this problem which is illustrated by the next example.

VARS	$\tau = \sqrt{5}$		$\tau = 1.824$		$\tau = 1.412$		$\tau = 1$					
	<i>QD</i>	Ψ^2	<i>QD</i>	Ψ^2	<i>QD</i>	Ψ^2	<i>QD</i>	Ψ^2				
POP	-.62	-.78	.00	.07	1.0	.00	-.00	1.0	.00	.00	-.99	.00
SCHOOL	-.70	.52	.23	.94	-.20	.07	.85	-.00	.27	-.28	-.00	.92
EMPLOY	-.70	-.68	.04	.19	.87	.21	-.00	1.0	.00	-.00	-.99	.00
SERVICES	-.88	.15	.20	.78	.23	.34	.58	.13	.65	-.18	-.00	.97
HOUSE	-.78	.60	.03	1.0	-.22	.00	1.1	-.07	.00	-1.2	.00	.00

Table 1: LS-EFA solutions for Five Socio-Economic Variables, [8, p.14].

Holzinger-Harman's Twenty-Four Psychological Tests [8, p.123]

Finally, we illustrate the proposed procedures for sparse EFA on another well known data set from classic EFA, namely the Holzinger-Harman' Twenty-Four Psychological Tests [8, p.123]. It is widely used to illustrate different aspects of classic EFA [8, 12].

The correlation matrix [8, p.124] of these data is non-singular and we apply ML-EFA (11). The first five columns of Table 2 contain the solution (factor loadings *QD* and unique variances Ψ^2) of (11) with $\tau = \sqrt{24} = 4.899$, i.e. the standard ML-EFA solution, which is nearly identical to the ML solution obtained in [8, p.215]. Then, we rotate (with normalization) the factor loadings *QD* from the first four columns by VARIMAX from MATLAB [11], and the result is given in the next four columns of Table 2. The loadings in bold correspond to non-zero loadings of the sparse ML-EFA solution of (11) obtained with $\tau = 2.2997$ and depicted in the last columns of Table 2. Further decrease of τ results in sparser loadings, but regarded as too simplified. Note, that to interpret the VARIMAX solution, one must subjectively ignore the loadings with small absolute values. The sparse factor loadings are easily interpreted only by focusing on the nonzero loadings.

VARS	$\tau = \sqrt{24} = 4.899$					Varimax				$\tau = 2.2997$			
	QD			Ψ^2		Rotated QD				QD	Ψ^2		
1	.60	.39	-.22	.02	.44	.69	.16	.16	.19	.88		.32	
2	.37	.25	-.13	-.03	.78	.44	.12	.10	.08	.28		.85	
3	.41	.39	-.14	-.12	.64	.57	.14	.11	-.02	.54		.70	
4	.49	.25	-.19	-.10	.65	.53	.23	.08	.10	.55		.69	
5	.69	-.28	-.03	-.30	.35	.19	.74	.15	.21		.82	.35	
6	.69	-.20	.08	-.41	.31	.20	.77	.23	.07		.84	.32	
7	.68	-.29	-.08	-.41	.28	.20	.81	.07	.15		.86	.29	
8	.67	-.10	-.12	-.19	.49	.34	.57	.13	.24		.64	.54	
9	.70	-.21	.08	-.45	.26	.20	.81	.23	.04		.87	.27	
10	.48	-.49	-.09	.54	.24	-.12	.17	.17	.83	-.18		.91	.28
11	.56	-.14	.09	.33	.55	.12	.18	.37	.51			.63	.59
12	.47	-.14	-.26	.51	.44	.21	.02	.09	.72			.72	.50
13	.60	.03	-.30	.24	.49	.44	.19	.08	.53	.30		.47	.51
14	.42	.02	.41	.06	.65	.05	.20	.55	.08		-.47		.74
15	.39	.10	.36	.09	.70	.12	.12	.52	.07		-.53		.70
16	.51	.35	.25	.09	.55	.41	.07	.53	.06		-.57		.68
17	.47	-.00	.38	.20	.60	.06	.14	.57	.22		-.72		.54
18	.52	.15	.15	.31	.59	.29	.03	.46	.34		-.65		.61
19	.44	.11	.15	.09	.76	.24	.15	.37	.16		-.35		.82
20	.61	.12	.04	-.12	.59	.40	.38	.30	.12	.34			.76
21	.59	.06	-.12	.23	.58	.38	.17	.22	.44			.51	.69
22	.61	.13	.04	-.11	.60	.40	.37	.30	.12	.30			.79
23	.69	.14	-.10	-.04	.50	.50	.37	.24	.24	.58		.04	.63
24	.65	-.21	.02	.18	.50	.16	.37	.30	.50			.63	.58

Table 2: ML-EFA solutions for Twenty-Four Psychological Tests [8, p.123].

6 Conclusion

We propose a new method to construct sparse factor loadings for the classic EFA. This is, in fact, a new approach to EFA, which readily produces interpretable EFA results. Unfortunately, this can be achieved on the expense of losing some portion of the fit of the sparse EFA model (2) to the sample correlation matrix R . Further research is needed to quantify this loss, and possibly relate it to the sparseness of the factor loadings in new sparse EFA algorithms.

There are few methods for sparse PCA, e.g. [6, 14, 16], able to produce either orthonormal component loadings or uncorrelated components. In contrast to PCA, the factor loadings $\Lambda (= QD)$, both original and sparse, are not orthonormal. However, how the sparse factor loadings affect the correlations among the estimated factors remains to be studied.

Acknowledgement

This work is supported by a grant RPG-2013-211 from The Leverhulme Trust, UK.

Bibliography

- [1] Absil, P.-A., Mahony, R., and Sepulchre, R. (2008) *Optimization Algorithms on Matrix Manifolds*, Princeton: Princeton University Press.
- [2] Boumal, N., Mishra, B., Absil, P.-A. and Sepulchre, R., (2014) Manopt: a Matlab toolbox for optimization on manifolds, *The Journal of Machine Learning Research*, to appear.
- [3] Choi, J., Zou, H. and Oehlert, G. (2011) A penalized maximum likelihood approach to sparse factor analysis. *Statistics and Its Interface*, 3, 429–436.
- [4] Del Buono, N. and Lopez, L. (2001) Runge-Kutta type methods based on geodesics for systems of ODEs on the Stiefel manifold, *BIT Numerical Mathematics*, 41, 912–923.
- [5] Edelman, A., Arias, T., and Smith, S. T. (1998) The geometry of algorithms with orthogonality constraints, *SIAM Journal on Matrix Analysis and Applications*, 20, 303–353.
- [6] Farcomeni, A. (2009) An exact approach to sparse principal component analysis, *Computational Statistics*, 24, 583–604.
- [7] Hage, C., and Kleinsteuber, M. (2014) Robust PCA and subspace tracking from incomplete observations using ℓ_0 -surrogates, *Computational Statistics*, to appear.
- [8] Harman, H. H. (1976) *Modern Factor Analysis*, 3rd Ed., Chicago: University of Chicago Press.
- [9] Jöreskog, K. G. (1977) Factor analysis by least-squares and maximum likelihood methods, in *Mathematical Methods for Digital Computers*, (K. Enslein, A. Ralston and H.S. Wilf, Eds.) Vol. 3, pp. 125–153, New York: Wiley.
- [10] Luss, R., and Teboulle, M. (2012) Conditional gradient algorithms for rank-one matrix approximations with a sparsity constraint, <http://arxiv.org/pdf/1107.1163.pdf>.
- [11] MATLAB (2011) *MATLAB R2011a*, The MathWorks, Inc., New York.
- [12] Mulaik, S. A. (2010) *The Foundations of Factor Analysis*. 2nd ed., Chapman and Hall/CRC, Boca Raton, FL.
- [13] Ning, L., and Georgiou, T. (2011) Sparse factor analysis via likelihood and ℓ_1 -regularization, *50th IEEE Conference on Decision and Control and European Control Conference (CDC-ECC)*, Orlando, FL, USA, December 12-15, 2011, 5188–5192.
- [14] Qi, X., Luo, R., Zhao, H. (2013) Sparse principal component analysis by choice of norm, *Journal of Multivariate Analysis*, 114, 127–160.
- [15] Trendafilov, N. T. (2003) Dynamical system approach to factor analysis parameter estimation. *British Journal of Mathematical and Statistical Psychology* 56, 27–46.
- [16] Trendafilov, N. T., and Jolliffe, I. T. (2006) Projected gradient approach to the numerical solution of the SCoTLASS, *Computational Statistics and Data Analysis* 50, 242–253.
- [17] Wen, Z., and Yin, W. (2013) A feasible method for optimization with orthogonality constraints, *Mathematical Programming*, 142, 397–434.

Efficiency of partially reduced-bias mean-of-order- p versus minimum-variance reduced-bias extreme value index estimation

M. Ivette Gomes, *Universidade de Lisboa and CEAUL*, ivette.gomes@fc.ul.pt
Frederico Caeiro, *Universidade Nova de Lisboa, FCT and CMA*, fac@fct.unl.pt

Abstract. A recent class of estimators of a positive extreme value index (EVI), related to a mean-of-order- p (MOP) class of EVI-estimators is enlarged and studied for finite samples through a Monte-Carlo simulation study. A comparison of this class and a representative class of minimum-variance reduced-bias (MVRB) EVI-estimators is performed. The class of MVRB EVI-estimators is related to a direct removal of the dominant component of the bias of the most popular estimator of a positive EVI, the Hill estimator, performed in such a way that the minimal asymptotic variance is kept at the same level.

Keywords. Heavy right-tails, Monte-Carlo simulations, Semi-parametric estimation, Statistics of extremes

1 The estimators under study and scope of the paper

Let X_1, \dots, X_n be independent, identically distributed (i.i.d.) random variables (r.v.'s) with a common distribution function (d.f.) F . Let us denote the associated ascending order statistics (o.s.) by $X_{1:n} \leq \dots \leq X_{n:n}$ and let us assume that there exist sequences of real constants $\{a_n > 0\}$ and $\{b_n \in \mathbb{R}\}$ such that the maximum, $X_{n:n}$, linearly normalized, i.e., $(X_{n:n} - b_n)/a_n$, converges in distribution to a non-degenerate r.v. Then the limiting distribution is necessarily an *extreme value* (EV) distribution, with the functional form

$$EV_\xi(x) = \begin{cases} \exp(-(1 + \xi x)^{-1/\xi}), & 1 + \xi x > 0, & \text{if } \xi \neq 0, \\ \exp(-\exp(-x)), & x \in \mathbb{R}, & \text{if } \xi = 0. \end{cases} \quad (1)$$

The d.f. F is said to belong to the max-domain of attraction of EV_ξ , and we write $F \in \mathcal{D}_M(EV_\xi)$. The parameter ξ is the *extreme value index* (EVI), the primary parameter of extreme events.

The EVI measures the heaviness of the right tail function $\bar{F} := 1 - F$, and the heavier the tail, the larger the EVI is. In this paper we work with Pareto-type distributions, with a strict positive EVI, i.e. in $\mathcal{D}_M^+ := \mathcal{D}_M(EV_\xi)_{\xi > 0}$. Essentially due to the fact that asymptotic properties of second-order parameters' estimators are known when $\rho < 0$, we often assume a right tail function,

$$\bar{F}(x) = 1 - F(x) = Cx^{-1/\xi} (1 + D_1x^{\rho/\xi} + o(x^{\rho/\xi})), \text{ as } x \rightarrow \infty, \xi > 0, \tag{2}$$

for $C > 0, D_1 \neq 0, \rho < 0$ (see [13]). Then, with the possible parameterization, $A(t) = \xi\beta t^\rho, \rho < 0$, and denoting by $U(t)$ the tail quantile function, $U(t) := F^{\leftarrow}(1 - 1/t), t > 1$, with $F^{\leftarrow}(y) := \inf\{x : F(x) \geq y\}$, the generalized inverse function of F , we have

$$\lim_{t \rightarrow \infty} \frac{\ln U(tx) - \ln U(t) - \xi \ln x}{A(t)} = \frac{x^\rho - 1}{\rho}, \tag{3}$$

a result more generally proved in [6] for $\rho \leq 0$ and $F \in \mathcal{D}_M^+$, where $|A|$ is necessarily a regularly varying function with an index of regular variation equal to ρ . Further note that (2) is equivalent to (3) with $\rho < 0$.

The class of EVI-estimators under play

For Pareto-type models, the most common EVI-estimators are the *Hill* (H) estimators, introduced in [14], which are the averages of the log-excesses, $\ln X_{n-i+1:n} - \ln X_{n-k:n}, 1 \leq i \leq k < n$, and can thus be written as

$$H(k) := \frac{1}{k} \sum_{i=1}^k \ln \frac{X_{n-i+1:n}}{X_{n-k:n}} = \sum_{i=1}^k \ln \left(\frac{X_{n-i+1:n}}{X_{n-k:n}} \right)^{1/k} = \ln \left(\prod_{i=1}^k \frac{X_{n-i+1:n}}{X_{n-k:n}} \right)^{1/k}, \quad 1 \leq k < n. \tag{4}$$

The H EVI-estimator is thus the logarithm of the geometric mean (or mean-of-order-0) of $\underline{U} := \{U_{ik} := X_{n-i+1:n}/X_{n-k:n}, 1 \leq i \leq k < n\}$. More generally, the authors in [2] considered as basic statistics the *mean-of-order-p* (MOP) of \underline{U} , with $p \geq 0$, now written for any $p \in \mathbb{R}$:

$$A_p(k) = \begin{cases} \left(\frac{1}{k} \sum_{i=1}^k U_{ik}^p \right)^{1/p}, & \text{if } p \neq 0, \\ \left(\prod_{i=1}^k U_{ik} \right)^{1/k}, & \text{if } p = 0, \end{cases}$$

and an associated class of MOP EVI-estimators, that more generally than in [2], [3], [7] and [8], can be defined as

$$H_p(k) \equiv MOP_p(k) := \begin{cases} \frac{1 - A_p^{-p}(k)}{p} = \frac{1 - \left(\frac{1}{k} \sum_{i=1}^k U_{ik}^p \right)^{-1}}{p}, & \text{if } p < 1/\xi, p \neq 0, \\ \ln A_0(k) = H(k), & \text{if } p = 0, \end{cases} \tag{5}$$

with $H_0(k) \equiv H(k)$, given in (4). In [11], for $p = 0$, and in [2], for $0 < p < 1/\xi$, was proved that if $F \in \mathcal{D}_M^+$ and k is intermediate, i.e. $k = k_n, 1 \leq k < n$, is such that

$$k = k_n \rightarrow \infty \quad \text{and} \quad k_n = o(n), \text{ as } n \rightarrow \infty, \tag{6}$$

the estimators $H_p(k)$, in (5) are consistent for the estimation of ξ provided that $0 \leq p < 1/\xi$. If we further assume the validity of the second-order condition in (3), with ρ possibly null, we can write for $0 \leq p < 1/(2\xi)$ the asymptotic distributional representation,

$$H_p(k) \stackrel{d}{=} \xi + \frac{\sigma_{H_p} Z_k^{(p)}}{\sqrt{k}} + b_{H_p} A(n/k)(1 + o_p(1)),$$

$$b_{H_p} = b_{H_p}(\xi, \rho) = \frac{1 - p\xi}{1 - \rho - p\xi}, \quad \sigma_{H_p}^2 = \sigma_{H_p}^2(\xi) = \frac{\xi^2(1 - p\xi)^2}{1 - 2p\xi}, \quad (7)$$

where $Z_k^{(p)}$ is standard normal.

Remark 1.

We thus have an asymptotic normal behavior for $H_p(k)$, in (5), if $0 \leq p < 1/(2\xi)$ and k such that $\sqrt{k}A(n/k) \rightarrow \lambda$, finite. There is however a reasonably high asymptotic bias (a decreasing function of p) when $\lambda \neq 0$, i.e. when we slightly increase k up to values where the mean square error (MSE) of $H_p(k)$ is minimized.

Remark 2.

Further note that for $p = -1$, in (5), we get $H_{-1}(k) := \left(\frac{1}{k} \sum_{i=1}^k X_{n-k:n}/X_{n-i+1:n}\right)^{-1} - 1$, the so-called t -Hill EVI-estimator in [15]. Moreover, with the parameterization $p = 1 - \beta$, we get the functional studied in [1], for $\beta > 0$, or equivalently, $p < 1$.

Working just for technical simplicity in the class of models in (2), the representation in (7), for $p = 0$, with $b_{H_0} = 1/(1 - \rho)$, led the authors in [4] to directly remove the dominant component of the bias of the H EVI-estimator, given by $\xi\beta(n/k)^\rho/(1 - \rho)$, considering, for adequate second-order parameters' estimators, $(\hat{\beta}, \hat{\rho})$, provided in [10], among others, the *corrected-H* (CH) *minimum-variance reduced-bias* (MVRB) EVI-estimator,

$$CH(k) \equiv CH_{\hat{\beta}, \hat{\rho}}(k) := H(k) \left(1 - \frac{\hat{\beta}}{1 - \hat{\rho}} \left(\frac{n}{k}\right)^{\hat{\rho}}\right). \quad (8)$$

Similarly, and with values of p such that the asymptotic normality of the estimators in (5) was known to hold at the time, i.e. for $0 \leq p < 1/(2\xi)$, as proved in [2], the authors in [3] noticed that there is an optimal value

$$p \equiv p_M = \varphi_\rho/\xi, \quad \text{with} \quad \varphi_\rho = 1 - \rho/2 - \sqrt{\rho^2 - 4\rho + 2}/2, \quad (9)$$

which maximises the asymptotic efficiency of the class of estimators in (5). Then, they considered an *optimal* MOP (OMOP) r.v., defined by $OMOP(k) := H_{p_M}(k)$, with $H_p(k)$ given in (5), deriving its asymptotic behaviour. Such a behaviour and some extra developments in [7], led the author in [8] to introduce a class of *partially* RBMOP (PRBMOP) EVI-estimators based on $H_p(k)$, in (5), given by

$$RB_p(k; \hat{\beta}, \hat{\rho}) \equiv PRBMOP_p(k) := H_p(k) \left(1 - \frac{\hat{\beta}(1 - \varphi(\hat{\rho}))}{1 - \hat{\rho} - \varphi(\hat{\rho})} \left(\frac{n}{k}\right)^{\hat{\rho}}\right), \quad (10)$$

still dependent on a tuning parameter p .

We shall further estimate the optimal k -value for the H EVI-estimation, as given in [12], computing $\hat{k}_{0|H_0} = ((1 - \hat{\rho})n^{-\hat{\rho}}/(\hat{\beta} \sqrt{-2\hat{\rho}}))^{2/(1-2\hat{\rho})}$, $H_{00} := H(\hat{k}_{0|H_0})$, and considering next the RB EVI-estimator

$$RB^*(k) := H_{\hat{p}_M}(k), \quad \hat{p}_M = \varphi_{\hat{\rho}}/H_{00}, \quad (11)$$

with φ_{ρ} given in (9).

Scope of the paper

In Section 2, we state and prove a theorem related to the EVI-estimator in (5), that provides also the asymptotic behaviour of the class of EVI-estimators in (10), for any negative real p . In Section 3, and through the use of Monte Carlo simulation techniques, we derive finite sample distributional properties of the class of PRBMOP EVI-estimators, in (10), and the adaptive OMOP EVI-estimator in (11), comparatively to the class of MVRB EVI-estimators, in (8).

2 Asymptotic behaviour of the EVI-estimators for negative p

We next state and prove the main theoretical result in the article:

Theorem 2.1. *If $F \in \mathcal{D}_M^+$ and (6) holds, the estimators $H_p(k)$ in (5) are, for any real $p < 1/\xi$, consistent for the estimation of ξ . Moreover, for consistent estimators $(\hat{\beta}, \hat{\rho})$ such that*

$$\hat{\rho} - \rho = o_p(1/\ln n), \quad \text{as } n \rightarrow \infty, \quad (12)$$

a property so far known to be achievable for models in (2), the $RB_p(k)$ estimators in (10) are also consistent for the EVI-estimation.

Under the validity of the second-order condition in (3), with ρ possibly null, the asymptotic distributional representation in (7) follows for any real $p < 1/(2\xi)$. Moreover, also for any real $p < 1/(2\xi)$, and under condition (12),

$$RB_p(k; \hat{\beta}, \hat{\rho}) \stackrel{d}{=} \xi + \frac{\sigma_{H_p} Z_k^{RB_p}}{\sqrt{k}} + b_{RB_p} A(n/k)(1 + o_p(1)),$$

$$b_{RB_p} = b_{RB_p}(\xi, \rho) = \frac{\rho(p\xi - \varphi_{\rho})}{(1 - p\xi - \rho)(1 - \rho - \varphi_{\rho})}, \quad (13)$$

with a dominant bias component $o_p(A(n/k))$ if and only if $p = p_M = \varphi_{\rho}/\xi$, with φ_{ρ} given in (9).

Proof. Note that, with $U(\cdot)$ the tail quantile function, we can write the distributional identity $X = U(Y)$, with Y a standard Pareto r.v., i.e. a r.v. with d.f. $F_Y(y) = 1 - 1/y$, $y \geq 1$. For the o.s. associated with a strict Pareto sample (Y_1, \dots, Y_n) , we have $Y_{n-i+1:n}/Y_{n-k:n} \stackrel{d}{=} Y_{k-i+1:k}$, $1 \leq i \leq k$. Moreover, $kY_{n-k:n}/n \xrightarrow{p} 1$, as $n \rightarrow \infty$, i.e. $Y_{n-k:n} \stackrel{p}{\sim} n/k$. Consequently, and provided that $k \rightarrow \infty$, with $k/n \rightarrow 0$, as $n \rightarrow \infty$, $U_{ik} \stackrel{p}{\sim} Y_{k-i+1:k}^{\xi}$, $1 \leq i \leq k$.

Further note that

$$\mathbb{E}(Y^a) = \frac{1}{1-a} \quad \text{if } a < 1, \quad \text{Var}(Y^a) = \frac{a^2}{(1-a)^2(1-2a)}, \quad \text{if } a < 1/2. \quad (14)$$

Working next under the second-order framework in (3), and even more generally assuming that we can have $\rho = 0$, using the interpretation of the Box-Cox function as the logarithm when the power equals 0, we can write

$$\begin{aligned} T_p(k) &:= \frac{1}{k} \sum_{i=1}^k \left(\frac{X_{n-i+1:n}}{X_{n-k:n}} \right)^p = \frac{1}{k} \sum_{i=1}^k Y_i^{p\xi} \left(1 + A(n/k) (Y_i^\rho - 1)/\rho + o_p(A(n/k)) \right)^p \\ &= \frac{1}{k} \sum_{i=1}^k Y_i^{p\xi} + pA(n/k) \frac{1}{k} \sum_{i=1}^k Y_i^{p\xi} (Y_i^\rho - 1)/\rho + o_p(A(n/k)). \end{aligned}$$

On the basis of (14), a derivation similar to the one in [2], enables us to get the result in the theorem related to $H_p(k)$, in (5), for all $\xi > 0$ and $p < 1/(2\xi)$, even a negative real number, and for the EVI-estimator $H_p(k) = (1 - T_p^{-1}(k))/p$.

Noticing next that $RB_p(k; \beta, \rho) := H_p(k) \left(1 - \frac{\beta(1-\varphi_\rho)}{1-\rho-\varphi_\rho} \left(\frac{n}{k} \right)^\rho \right)$, we easily derive that the dominant component of the bias is given by

$$\frac{(1 - p\xi)A(n/k)}{1 - p\xi - \rho} - \frac{(1 - \varphi_\rho)A(n/k)}{1 - \rho - \varphi_\rho} = \frac{\rho(p\xi - \varphi_\rho)A(n/k)}{(1 - p\xi - \rho)(1 - \rho - \varphi_\rho)},$$

i.e. it is null only for $p = \varphi_\rho/\xi$. If we estimate consistently β and ρ through the estimators $\hat{\beta}$ and $\hat{\rho}$, and condition (12) holds, we can use Cramer’s delta-method, and in the lines of [8], we get the result in the theorem not only for $p \geq 0$, but for any negative real p . \square

In Figure 1, to visualize the reduction in bias achieved by the PRBMOP EVI-estimation, a representation of $b_{H_p} = b_{H_p}(\xi, \rho)$ and $b_{RB_p} = b_{RB_p}(\xi, \rho)$, respectively given in (7) and (13), as functions of ξ , for $p = 0.1, 0.5, 1$ and $\rho = -1$, can be found.

Quick and simple MVRB and PRBMOP EVI-estimators

Since the second-order reduced bias estimators in (8) and (10) depend on the estimation of the second order parameters β and ρ , and just as suggested in [10] for the MVRB EVI-estimator in (8), we could also have considered *quick and simple* estimators, a by-product of the estimators in (8) and (10), setting there $\hat{\beta} = 1$ and $\hat{\rho} = -1$. We then get

$$\begin{aligned} \overline{CH}(k) &:= H_{1,-1}(k) = H_0(k) \left(1 - \frac{a_{CH}k}{n} \right) \quad \text{and} \\ \overline{RB}_p(k) &= RB_p(k; 1, -1) = H_p(k) \left(1 - \frac{a_{RB}k}{n} \right) \quad (15) \end{aligned}$$

with $a_{CH} = 1/2$, $a_{RB} = (1 - \varphi(-1))/(2 - \varphi(-1))$, $\varphi(-1) = (3 - \sqrt{7})/2$, and where $H_p(k)$ stands for the MOP EVI-estimator in (5), with $H_0(k) \equiv H(k)$, the Hill EVI-estimator in (4).

If we consider the replacement of the estimators in (5) and (8) by the quick and simple estimators in (15), the dominant component of bias changes, but the asymptotic variance is still kept equal to $\sigma_{RB_p}^2$, in (7). The dominant component of the asymptotic bias of $\overline{CH}(k)$ is then of a smaller order than $A(n/k)$ if and only if $(\beta, \rho) = (1, -1)$. For both $\overline{CH}(k)$ and $\overline{RB}_p(k)$ the dominant component of bias, of the order of k/n , is of a larger order than the one of the $H_p(k)$ EVI-estimators for models with $\rho < -1$.

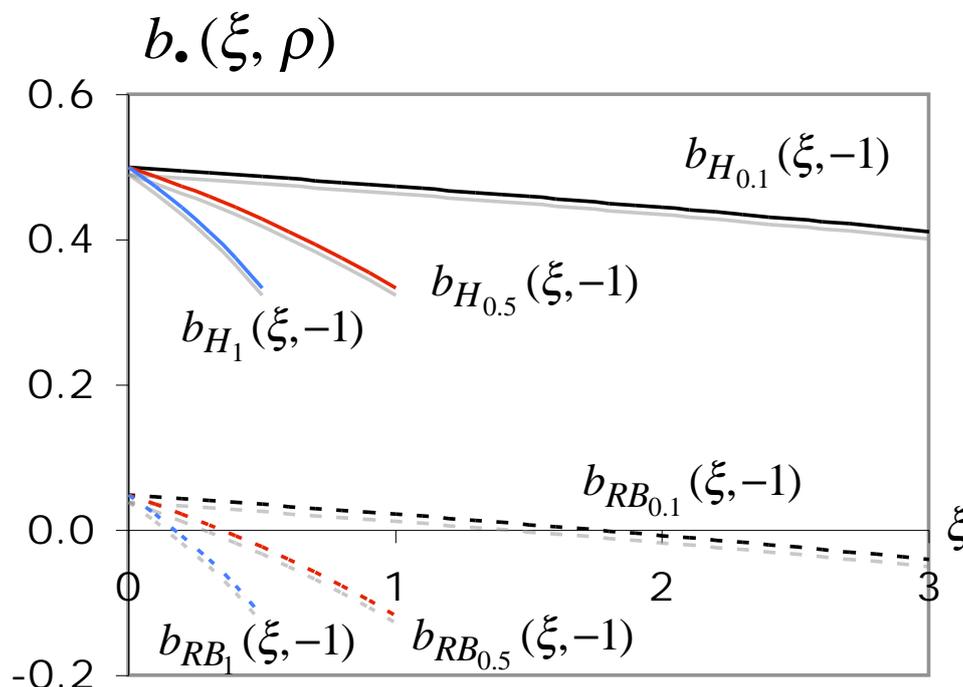


Figure 1: Values of $b_{H_p} = b_{H_p}(\xi, \rho)$ and $b_{RB_p} = b_{RB_p}(\xi, \rho)$, as functions of ξ , for $p = 0.1, 0.5, 1$ and $\rho = -1$

Remark 3.

Note that the MOP EVI-estimator is a special case of the so-called quick and simple EVI-estimators, since it uses $\hat{\beta} = 0$ in (5).

Remark 4.

The quick and simple estimators in (15) are adequate only if the guess $\beta = 1, \rho = -1$ captures properly the deviation of the underlying tail from a strict Pareto tail, but educated guesses may be much more precise than are the usually noisy estimates of higher order parameters. Note however that the second-order parameters' estimates proposed in this paper are quite reliable in the class of Hall-Welsh models.

3 Monte Carlo simulations

We have performed extensive simulations associated with the Generalized Pareto (GP) model, related to $EV_{\xi}(\cdot)$, in (1), through the relationship $F(x) = 1 + \ln EV_{\xi}(x) = 1 - (1 + \xi x)^{-1/\xi}$, $x \geq 0, \xi > 0$, for which $\rho = -\xi$, and the Burr $_{\xi, \rho}$ model, with d.f. $F(x) = 1 - (1 + x^{-\rho/\xi})^{1/\rho}$, $x \geq 0, \xi > 0, \rho < 0$. In all Monte-Carlo simulation experiments we have considered multi-sample simulations of size 5000×20 and sample sizes $n = 100, 200, 500, 1000, 2000$ and 5000 . For details on multi-sample simulation, we refer [9].

Mean values and MSE paths

For each value of n and for each of the aforementioned models, we have first simulated, as functions of k , the number of top o.s. involved in the estimation, and on the basis of first run of size 5000, the mean values (E) and root MSEs (RMSEs) of the EVI-estimators in (8) and (10), for values of $p = -1, -0.5, -0.25, -0.1$ and $p = \ell/(10\xi)$, $\ell = 1(1)9$, so that $p < 1/\xi$ and we have valid estimates. As an illustration, we present Figure 2 associated with $Burr_{1,-0.5}$ parents. We further notice that for all simulated parents with $\rho \in (-1, 0)$, RB^* is always in between the $PRBMOP$ for $\ell = 1$ (close to CH) and $\ell = 2$, both regarding mean values and RMSEs. In this same region of ρ -values, the best performance is always achieved in the region $5 \leq \ell \leq 9$, as can be seen in Figure 2.

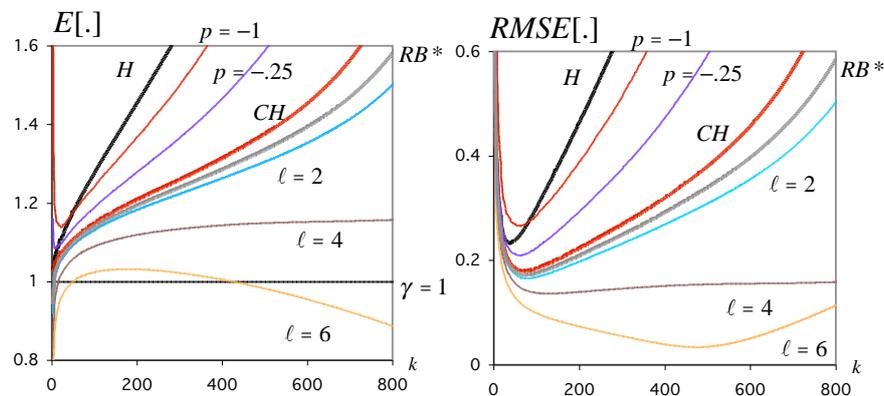


Figure 2: Mean values (left) and RMSEs (right) of $H(k)$, $CH(k)$, $RB^*(k)$ and $RB_p(k)$, $p = -1, -0.25$ and $p = \ell/10\xi$, $\ell = 2, 4, 6$, for a $Burr_{1,-0.5}$ underlying parent

Mean values and MSEs at optimal levels

We have computed the Hill EVI-estimator at the simulated value of $k_{0|H} := \arg \min_k RMSE(H(k))$, the simulated optimal k in the sense of minimum RMSE. We have also computed RB_{p0} , i.e. the PRBMOP EVI-estimator $RB_p(k)$ computed at the simulated value of $k_{0|RB_p} := \arg \min_k RMSE(RB_p(k))$. As an illustration of the bias reduction achieved with the PRBMOP EVI-estimators in (10) at optimal levels, see Table 1, related to the model $GP_{0.25}$. We there present, for $n = 100, 200, 500, 1000, 2000$ and 5000 , the simulated mean values at optimal levels of H_{00} , CH_0 and RB_{p0} for $p = -1, -0.5, -0.25, -0.1$ and $p = \ell/(10\xi)$, considering the two regions, $\ell = 1, 2, 3, 4$, where we can guarantee consistency and asymptotic normality, and $\ell = 5, 6, 7, 8, 9$, where only consistency is assured by Theorem 2.1. We further consider RB_0^* . Information on 95% confidence intervals, computed on the basis of the 20 replicates with 5000 runs each, is also provided. For each region, and among the estimators considered, the one providing a smaller squared bias than the best one in the previous region is written in *italic*, and underlined whenever it turns out to be the best in a region, beating the best estimator in the previous region.

We next present in Table 2 the simulated values of the indicators,

$$REFF_{RB_p|H} := RMSE(H_{00})/RMSE(RB_{p0}), \tag{16}$$

Table 1: Simulated mean values of H_{00} , CH_0 , RB_{p0} , $p = -1, -0.5, -0.25$, RB_0^* and RB_{p0} , $p = \ell/(10\xi)$, $\ell = 2(1)9$, for $GP_{0.25}$ underlying parents, together with 95% confidence intervals

GP_ξ parent, $\xi = 0.25$						
n	100	200	500	1000	2000	5000
H_{00}	0.419 ± 0.0024	0.390 ± 0.0028	0.365 ± 0.0018	0.347 ± 0.0016	0.335 ± 0.0012	0.320 ± 0.0011
CH_0	0.406 ± 0.0030	0.382 ± 0.0017	0.360 ± 0.0017	0.345 ± 0.0018	0.333 ± 0.0013	0.319 ± 0.0009
$p = -1$	0.452 ± 0.0027	0.416 ± 0.0026	0.381 ± 0.0016	0.361 ± 0.0019	0.345 ± 0.0013	0.327 ± 0.0008
$p = -0.5$	0.431 ± 0.0032	0.400 ± 0.0025	0.371 ± 0.0017	0.353 ± 0.0018	0.340 ± 0.0013	0.323 ± 0.0009
$p = -0.25$	0.419 ± 0.0031	0.391 ± 0.0017	0.366 ± 0.0018	0.350 ± 0.0019	0.337 ± 0.0014	0.321 ± 0.0009
RB_0^*	0.388 ± 0.0031	0.370 ± 0.0023	0.349 ± 0.0015	0.336 ± 0.0015	0.327 ± 0.0012	0.314 ± 0.0008
$\ell = 2$	0.361 ± 0.0027	0.351 ± 0.0020	0.338 ± 0.0016	0.328 ± 0.0015	0.321 ± 0.0012	0.310 ± 0.0007
$\ell = 3$	0.337 ± 0.0022	0.330 ± 0.0024	0.323 ± 0.0016	0.317 ± 0.0013	0.311 ± 0.0011	0.304 ± 0.0008
$\ell = 4$	0.295 ± 0.0002	0.293 ± 0.0002	0.293 ± 0.0001	0.293 ± 0.0001	0.293 ± 0.0001	0.292 ± 0.0001
$\ell = 5$	0.249 ± 0.0002	0.250 ± 0.0001				
$\ell = 6$	0.250 ± 0.0001					
$\ell = 7$	0.249 ± 0.0001	0.250 ± 0.0001				
$\ell = 8$	0.243 ± 0.0002	0.247 ± 0.0001	0.249 ± 0.0001	0.250 ± 0.0001	0.250 ± 0.0001	0.250 ± 0.0001
$\ell = 9$	0.226 ± 0.0002	0.232 ± 0.0001	0.238 ± 0.0001	0.242 ± 0.0001	0.245 ± 0.0001	0.247 ± 0.0001

again for a $GP_{0.25}$ model. Similar REFF-indicators have also been computed for the CH and RB^* EVI-estimators. In the first row of Table 2, we provide $RMSE_{00}$, the RMSE of H_{00} , so that we can easily recover the RMSE of all other estimators. The following rows provide the REFF-indicators of CH and RB_p , for the same values of p as in Table 1. We further present a similar REFF-indicator for RB^* . A similar mark (*italic* and/or underlined) is used. Confidence intervals are not provided for REFF-indicators larger than 10, but are available from the authors.

Remark 5.

An indicator higher than one means a better performance than the H estimator, i.e. the higher these indicators are, the better the associated EVI-estimators perform, comparatively to H_{00} .

Table 2: Simulated RMSE of H_{00} (first row) and REFF-indicators of CH , RB_p , $p = -1, -0.5, -0.25$, RB^* and RB_p , $p = \ell/(10\xi)$, $\ell = 2(1)9$, and for $GP_{0.25}$ underlying parents, together with 95% confidence intervals

GP_ξ parent, $\xi = 0.25$						
n	100	200	500	1000	2000	5000
$RMSE_{00}$	0.237 ± 0.1756	0.196 ± 0.1684	0.155 ± 0.1592	0.131 ± 0.1525	0.112 ± 0.1460	0.092 ± 0.1377
CH_0	1.148 ± 0.0049	1.118 ± 0.0027	1.088 ± 0.0025	1.069 ± 0.0018	1.057 ± 0.0018	1.042 ± 0.0012
$p = -1$	0.911 ± 0.0058	0.926 ± 0.0035	0.939 ± 0.0032	0.941 ± 0.0032	0.947 ± 0.0030	0.948 ± 0.0019
$p = -0.5$	1.006 ± 0.0055	1.005 ± 0.0030	1.001 ± 0.0029	0.995 ± 0.0025	0.994 ± 0.0023	0.989 ± 0.0015
$p = -0.25$	1.064 ± 0.0050	1.051 ± 0.0027	1.037 ± 0.0026	1.026 ± 0.0023	1.021 ± 0.0020	1.012 ± 0.0012
RB^*	1.231 ± 0.0034	1.194 ± 0.0021	1.157 ± 0.0020	1.133 ± 0.0021	1.115 ± 0.0016	1.094 ± 0.0014
$\ell = 2$	1.445 ± 0.0032	1.350 ± 0.0023	1.263 ± 0.0023	1.213 ± 0.0024	1.177 ± 0.0023	1.138 ± 0.0022
$\ell = 3$	1.717 ± 0.0038	1.563 ± 0.0033	1.420 ± 0.0036	1.340 ± 0.0032	1.279 ± 0.0032	1.215 ± 0.0035
$\ell = 4$	5.203 ± 0.0301	4.460 ± 0.0218	3.582 ± 0.0214	3.037 ± 0.0131	2.636 ± 0.0122	2.172 ± 0.0080
$\ell = 5$	34.900	40.247	53.187	65.706	76.072	96.346
$\ell = 6$	34.768	40.360	53.204	65.298	75.078	93.574
$\ell = 7$	30.255	36.245	47.763	57.696	65.384	77.688
$\ell = 8$	18.053	23.072	31.802	38.688	44.293	51.241
$\ell = 9$	8.791 ± 0.0708	9.467 ± 0.0735	10.932	12.456	14.347	17.536

4 Concluding remarks

1. Note that for $5 \leq \ell \leq 9$, we have consistency of the estimators either in (5) or in (10), but no guarantee of asymptotic normality, and even of bias reduction comparatively to the MOP EVI-estimators. Despite of this comment the EVI-estimators in this region can be the ones that exhibit the best performance regarding both mean values and RMSEs.
2. The estimators for negative values of p can beat the Hill at optimal levels but never the CH EVI-estimators regarding both bias and RMSE. They are not indeed efficient, despite of the fact that the similar MOP EVI-estimator for $p = -1$ (see [1]) has revealed to be the most robust EVI-estimator, but with a low efficiency. This comment could thus lead to a discussion of robustness versus efficiency and to the need of an indicator that takes both concepts into account (see e.g. [5]).
3. For both mean values and RMSEs at optimal levels, and again if we restrict ourselves to the region of p -values where we can so far guarantee asymptotic normality, the best results were obtained at $p = 4/(10\xi)$ for most of the simulated models.
4. Regarding RMSE, the consistent and asymptotically normal PRBMOP EVI-estimators at optimal levels, can always beat the MVRB EVI estimators also at optimal levels for all $0 < p < 1/(2\xi)$. They can however be beaten by the only consistent PRBMOP EVI-estimators ($1/(2\xi) \leq p < 1/\xi$), at optimal levels, for most of the simulated parents.

Acknowledgement

Research partially supported by National Funds through **FCT**—Fundação para a Ciência e a Tecnologia, projects PEst-OE/MAT/UI0006/2014 (CEA/UL) and PEst-OE/MAT/UI0297/2014 (CMA/UNL).

Bibliography

- [1] Beran, J., Schell, D. and Stehlik, M. (2014) *The harmonic moment tail index estimator: asymptotic distribution and robustness*. Ann. Inst. Statist. Math., 66, 193–220.
- [2] Brilhante, M.F., Gomes, M.I. and Pestana, D. (2013) *A simple generalization of the Hill estimator*. Computational Statistics and Data Analysis, 57:1, 518–535.
- [3] Brilhante, M.F., Gomes, M.I. and Pestana, D. (2013) *The MOP EVI-Estimator Revisited*. In Pacheco, A., Santos, R., Oliveira, M.R., and Paulino, C.D. (eds.), *New Advances in Statistical Modeling and Application*. Studies in Theoretical and Applied Statistics, 163–175, Springer International Publishing, Switzerland.
- [4] Caeiro, F., Gomes, M.I. and Pestana, D. (2005) *Direct reduction of bias of the classical Hill estimator*. Revstat, 3:2, 113–136.
- [5] Dell’Aquila, R. and Embrechts, P. (2006) *Extremes and robustness: a contradiction?* Fin. Mkts. Portfolio Mgmt., 20, 103–118.
- [6] Geluk, J. and de Haan, L. (1987) *Regular Variation, Extensions and Tauberian Theorems*. CWI Tract 40, Center for Mathematics and Computer Science, Amsterdam, Netherlands.

- [7] Gomes, M.I., Brilhante, M.F. and Pestana, D. (2013) *New reduced-bias estimators of a positive extreme value index*. Comm. Statist. Simulation Comput., in press.
- [8] Gomes, M.I., Brilhante, M.F., Caeiro, F. and Pestana, D. (2014) *A new partially reduced-bias mean-of-order p class of extreme value index estimators*. Notas e Comunicações CEAUL 04/2014.
- [9] Gomes, M.I. and Oliveira, O. (2001) *The bootstrap methodology in Statistics of Extremes: choice of the optimal sample fraction*. Extremes, 4:4, 331–358.
- [10] Gomes, M.I. and Pestana, P. (2007) *A sturdy reduced-bias extreme quantile (VaR) estimator*. J. American Statistical Association, 102:477, 280–292.
- [11] de Haan, L. and Peng, L. (1998) *Comparison of extreme value index estimators*. Statistica Neerlandica, 52, 60–70.
- [12] Hall, P. (1982) *On some simple estimates of an exponent of regular variation*. J. Royal Statistical Society B, 44, 37–42.
- [13] Hall, P. and Welsh, A.W. (1985) *Adaptive estimates of parameters of regular variation*. Ann. Statist., 13, 331–341.
- [14] Hill, B.M. (1975) *A simple general approach to inference about the tail of a distribution*. Ann. Statist., 3, 1163–1174.
- [15] Stehlík M., Potocký R., Waldl H. and Fabián Z. (2010) *On the favourable estimation of fitting heavy tailed data*. Computational Statistics, 25, 485–503.

Data-driven wavelet resolution choice in multichannel box-car deconvolution with long memory

J.R. Wishart, *University of New South Wales*, j.wishart@unsw.edu.au

Abstract. In wavelet deconvolution, the finest resolution level is a key parameter which needs to be chosen carefully. In this paper a data-driven method is presented that selects the finest resolution level using a blockwise thresholding method in the Fourier domain. In particular, we present a method that applies to the general multichannel model whereby a practitioner observes many box-car convolutions of a signal of interest (with possible different levels of box-car ‘blur’) with additive long memory noise. The box-car functions governing the blur are assumed to have Badly Approximable (BA) width. To the best of the author’s knowledge, no automatic fine resolution selection method exists for the box-car wavelet deconvolution paradigm. We present a method that selects the optimal level that is adaptive to box-car width and noise levels and conduct a short numerical study to supplement the findings.

Keywords. Box-car, Badly Approximable, Data-driven, fractional Brownian motion, Fourier analysis, Meyer Wavelet, Multichannel deconvolution, Wavelet Analysis

1 Introduction

Wavelet deconvolution methods have been a popular area of research for inverse problems in recent history. One of the popular algorithms is the `WaveD` method first proposed in [7]. This method is a hard-thresholding wavelet deconvolution estimator that is attractive since it is a non-iterative technique that, utilising the Fast Fourier Transform, is fast and easy to compute and also has desirable theoretical properties. The methodology has been extended to the multichannel and long memory cases in recent years in [5, 13, 15, 9, 1], among others. Most recently, [1] and [9] consider a framework with both a multichannel signal and additive long memory errors. The particular context given in [9] will be the focus here.

consider an estimator and measure its performance from a minimax perspective. They also consider a more general in the sense that it encompasses the case when the additive errors are Gaussian or sub-Gaussian (a type of moment condition) but considers the context where both

the number of channels in the multichannel model and the number of observations per channel diverge to infinity. This paper will focus on the context of [9] which considers the number of channels to be fixed and constant, the error is driven by a fractional Brownian motion process and optimality is measured from a maxiset type perspective.

The WaveD method, does however, require calibration or tuning of its key parameters. These key parameters are the resolution levels to include in the wavelet expansion and the threshold levels at each resolution. The theoretical optimal choice of parameters for the thresholds is well understood and addressed in detail in [9] for a maxiset type perspective and in [1] for a minimax type perspective. In both cases an asymptotic condition is required on the finest resolution level. However, specific calibration of this parameter in finite cases, when box-car convolution is apparent, has not been addressed to the authors knowledge. The asymptotic theory of [9] requires that the finest resolution level for box-car blur, j_1^B , satisfies,

$$2^{j_1^B} \asymp \left(\frac{n^{\alpha_*}}{\log n} \right)^{1/(2\tilde{\nu}_*+1)} \quad (1)$$

where $\tilde{\nu}_*$ is a parameter that summarises the overall degree of ill-posedness of the multichannel model and $a \asymp b$ means there exists constants $c, C > 0$ such that $cb < a < Cb$. The condition on j_1^B in (1) is an asymptotic condition to assure bounds on L^p -risk for estimation (see [9]). As such, it cannot be used to determine the appropriate level of truncation in the wavelet expansion from a finite sample. Data driven methods do exist for fine resolution level calibration. However, they seem to consider only the case of regular smooth convolution whose spectrum decays with a power law. This method for smooth blur is not directly suitable and we present an alternative that is better for the box-car convolution case.

The paper is organised in the following way. In Section 2, a review of the the wavelet deconvolution expansion and the multichannel model with long memory is given. In Section 3, the fine resolution estimator is presented. In Section 4 a numerical study is conducted with concluding remarks given in Section 5 along with some comparison to other frameworks.

2 Preliminaries

Consider the problem of recovering $f \in L^2(T)$, periodic on $T = [0, 1]$. Let K_ℓ , $\ell = 1, 2, \dots, M$; be a set of box-car blurring kernels, also defined on T that generate the set of convolutions,

$$K_\ell * f(t) = \int_{-c_\ell}^{c_\ell} f(t-x) dx, \quad t \in T, \quad \ell = 1, 2, \dots, M; \quad (2)$$

where $c_\ell > 0$ for each $\ell = 1, 2, \dots, M$; and c_ℓ are referred to as the box-car half widths. Suppose (2) is only observable with additive long memory noise with,

$$dY_\ell(t) = K_\ell * f(t) dt + \sigma_\ell n^{-\alpha_\ell/2} dB_{H_\ell}(t), \quad t \in T, \quad \ell = 1, 2, \dots, M. \quad (3)$$

Here, B_{H_ℓ} are independent standard fractional Brownian motions with known Hurst parameters $H_\ell = 1 - \alpha_\ell/2 \in [1/2, 1)$. The noise level is governed by the usual parametrisation, $\sigma_\ell n^{-\alpha_\ell/2}$. See [14, 15, 9] for a link between the asymptotic theoretical model, (3), and the discrete model faced by practitioners. An example of model (3) is given in the left plot of Figure 1.

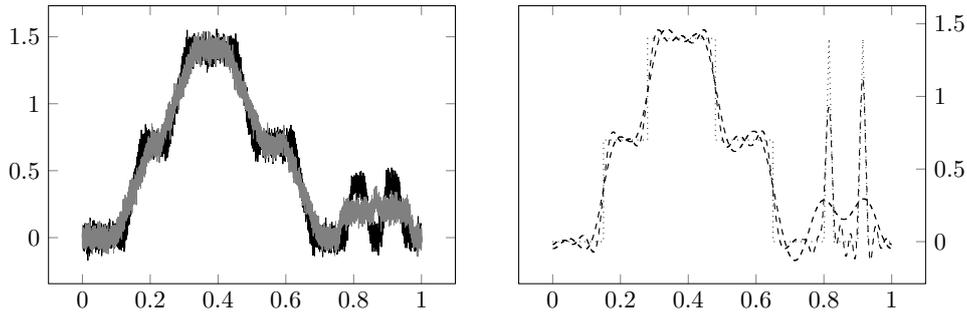


Figure 1: Left plot: signal with two channels; black line = first channel with $c_1 = 1/\sqrt{353}$, grey line = second channel with $c_2 = 1/\sqrt{89}$. Right plot: dotted line = target signal, solid line = MWaveD estimate using block method, $\hat{j}_1^B = 6$; dashed line = MWaveD estimate using smooth method, $\hat{j}_1^S = 4$.

In the Fourier domain, model (3) takes the form,

$$y_{m,\ell} = k_{m,\ell} \cdot f_m + \frac{\sigma_\ell}{n^{\alpha_\ell/2}} z_{m,\ell}, \quad \ell = 1, 2, \dots, M; \tag{4}$$

where $y_{m,\ell} = \int e^{-2\pi i m x} dY(x)$, $z_{m,\ell} = \int e^{-2\pi i m x} dB_{H_\ell}(x)$. The Fourier coefficients of K_ℓ and f are denoted $k_{m,\ell}$ and f_m respectively.

Meyer Wavelet Basis. Denote (Ψ, Φ) to be the periodised Meyer wavelet and scaling functions (cf. [11, 10]). These functions induce a set of orthonormal basis functions,

$$\Phi_{j,k}(t) = 2^{j/2} \Phi(2^j t - k), \quad \Psi_{j,k}(t) = 2^{j/2} \Psi(2^j t - k), \quad j \geq 0, \quad k \in \mathbb{Z}, \quad t \in \mathbb{R},$$

known as the dilated and translated wavelet functions at resolution level j and scale position $k/2^j$. With these wavelet functions, any periodic $f \in L^2$ has expansion,

$$f(x) = \sum_k a_{j,k} \Phi_{j_0,k}(x) + \sum_{j \geq j_0} \sum_k \Psi_{j,k}(x), \quad \text{where } a_{j,k} = \int f \Phi_{j,k} \text{ and } b_{j,k} = \int f \Psi_{j,k}.$$

The values $a_{j,k}$ and $b_{j,k}$ are known as the scaling and wavelet coefficients. The Meyer wavelet basis is particularly useful here both mathematically and computationally since it is band-limited and thus has a simple behaviour in the Fourier domain.

MWaveD estimator. The WaveD estimator of [7] was extended recently by [9] to handle the multichannel context in the presence of long memory noise. This extended estimator, referred to here as MWaveD, is defined,

$$\hat{f}(x) = \sum_{k=0}^{2^{j_0}-1} \hat{a}_{j_0,k} \Phi_{j_0,k}(x) + \sum_{j=j_0}^{j_1} \sum_{k=0}^{2^j-1} \hat{b}_{j,k} \mathbb{1}_{\{|\hat{b}_{j,k}| > \lambda_j\}} \Psi_{j,k}(x) \tag{5}$$

where indices j_0 and j_1 correspond to the coarsest and finest resolution levels in the expansion and k indexes the detail locations inside each resolution.

The `MWaveD` coefficients, $\widehat{b}_{j,k}$, are computed with a weighted average of the observations in model (4),

$$\widehat{b}_{j,k} = \sum_{m \in C_j} \frac{\overline{\Psi_m^{j,k}} \sum_{\ell=1}^M \sigma_\ell^{-2} n^{\alpha_\ell} |m|^{1-\alpha_\ell} \overline{k_{m,\ell}} y_{m,\ell}}{\sum_{\ell=1}^M \sigma_\ell^{-2} n^{\alpha_\ell} |m|^{1-\alpha_\ell} |k_{m,\ell}|^2} \quad (6)$$

where $\Psi_m^{j,k}$ denotes the Fourier transform of $\Psi_{j,k}$ (and \overline{f} denotes the conjugate of f) and $C_j = \{m \in \mathbb{Z} : \Psi_m^{j,k} \neq 0\}$. De-noising is applied through hard-thresholding with the scale level thresholds $\lambda_j = \zeta \tau_j \sqrt{\log n}$ where ζ is a smoothing parameter and τ_j is the standard error of (6) with

$$\tau_j^2 = \sum_{m \in C_j} |\Psi_m^{j,k}|^2 \left(\sum_{\ell=1}^M \sigma_\ell^{-2} n^{\alpha_\ell} |m|^{1-\alpha_\ell} |k_{m,\ell}|^2 \right)^{-1}. \quad (7)$$

If the noise levels, σ_ℓ , are unknown to the practitioner, they are estimated using $\widehat{\sigma}_\ell$ using a standard approach. A typical estimator of $\widehat{\sigma}_\ell$ is the median absolute deviation of the estimated wavelet coefficients at the highest possible resolution level, $J = \lfloor \log_2 n \rfloor - 1$. Then both $\widehat{b}_{j,k}$ and τ_j are computed by replacing σ_ℓ with $\widehat{\sigma}_\ell$.

These `MWaveD` estimates are computed in Figure 1 using the suggested blockwise selection method presented below and comparing with the existing stopping method. The suggested blockwise selection method allows higher resolution levels and consequently a better estimator that is closer to the true signal.

3 Fine resolution estimation

The goal is to find the highest possible resolution level in the wavelet expansion, while preserving the optimal properties of `MWaveD` in [9]. The deconvolution process in the Fourier domain involves division by the Fourier coefficients of the convolution kernel. This has the impact of inflating the noise process when the kernel decays at higher frequencies. Data-driven methods have been considered to alleviate this in [3] and [4] where the blurring function K_ℓ is unknown and observed in noise. It was extended to the long memory and multichannel context in [15, 9]. In this scenario, the convolution is assumed to be a regular smooth type where the convolution kernel in the Fourier domain obeys a power law decay,

$$|k_m| \asymp |m|^{-\nu}, \quad m \in \mathbb{R} \quad (8)$$

for some $\nu > 0$. They then employ a stopping rule in the Fourier domain to select the finest resolution level when the Fourier convolution coefficients drop below the maximum of the threshold for the noise level. This is the point where the price paid by inflating the noise outweighs any benefit of higher scale information in the signal and the wavelet expansion is unstable. Specifically, the stopping rule suggests (cf. [9]),

$$\widehat{j}_1^S = \max_{\ell=1, \dots, M} \log_2 \left[\min \left\{ m \in \mathbb{Z}^+ : |k_{m,\ell}| \leq m^{\alpha_\ell/2} n^{\alpha_\ell} / \sigma_\ell \log(n_\ell^\alpha / \sigma_\ell^2) \right\} \right] - 1. \quad (9)$$

Method (9), will be referred to as the `smooth` rule. Unfortunately, the box-car convolution coefficients do not exhibit the same simple power law decay as (8) in the Fourier domain. The Fourier coefficients of the box-car function take the form,

$$k_{m,\ell} = \frac{\sin(2\pi m c_\ell)}{2\pi m c_\ell}, \quad m \in \mathbb{Z}, \quad \ell = 1, 2, \dots, M; \quad \Rightarrow |k_{m,\ell}| \asymp |m c_\ell| / |m c_\ell| \quad (10)$$

where $|x| = \inf \{|x - r| : r \in \mathbb{Z}\}$ is the distance from x and its closest integer (cf. [13, p.63]).

Information is lost at some frequencies if the box-car half width is a rational, i.e. for rational $c_\ell = a/b$, then $k_{m,\ell}$ vanishes for all frequencies, m , that are a multiple of b . The problem is less severe if one considers the half-widths to be Badly Approximable (BA) numbers (the reader is referred to [6, 7, 5, 13] for more detail). The BA numbers contain the quadratic irrationals and we will focus on these for simulation examples in this paper.

As observed by [8], one of the main numerical difficulties with box-car deconvolution is that the Fourier coefficients vary wildly due to approximations in $|\cdot|$ given in (10).

However, this instability is avoided when taking sums over dyadic blocks, e.g. over the sets C_j which have cardinality $|C_j| = 2^{j+1}$. In particular, an equivalence between sums over dyadic blocks of the box-car spectrum with BA half widths and regular smooth decay when $\nu = 3/2$ is shown in [7, Proposition 2]. Figure 2 shows the instability in the box-car spectrum compared with regular smooth decay. The `smooth` method picks a lower final resolution than is necessary since the log spectrum of the box-car drops below the log bound earlier than the ‘equivalent’ smooth spectrum with $\nu = 3/2$ for the `smooth` method due to the instabilities of the box-car. This is shown in the right plot of Figure 3.

Blockwise selection. With this in mind, we adapt the methods of [2] who consider a blockwise estimation of fine scale levels in the image-deblurring case. The deconvolution process at resolution j involves averaging over dyadic blocks, C_j (see (6)). The cost of this deconvolution process inflates the variance of the noise coefficients to τ_j^2 given by (7). The blockwise selection rule keeps all resolution levels such that the cost of deconvolution is no more than the maximum allowed noise level. The maximum noise level we can allow whilst preserving the properties of [9] is of the order,

$$\left(2^j \sum_{\ell=1}^M \log(n^{\alpha_\ell} / \sigma_\ell^2)\right)^{-1}. \tag{11}$$

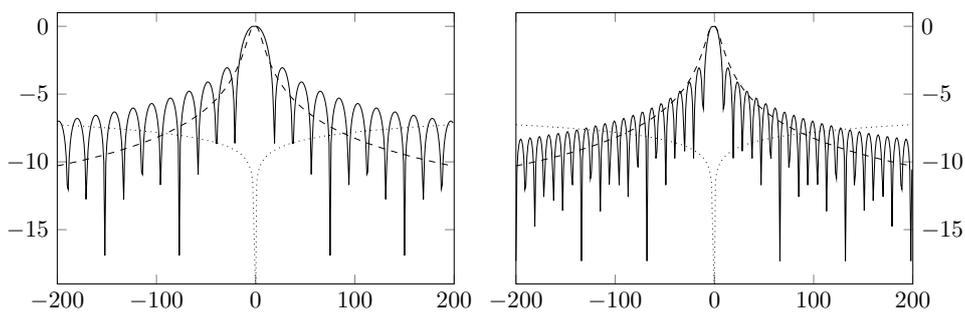


Figure 2: Comparison of the log decay of box-car Fourier coefficients and the noise bounds. Solid lines are $y = 2 \log |k_{m,\ell}|$ for width levels $c_\ell = 1/\sqrt{353}$ on the left plot and $c_\ell = 1/\sqrt{89}$ on the right plot. Dashed line = log spectrum of $\Gamma(3/2, 0.025)$ density in Fourier domain (example of (8) with $\nu = 3/2$). Dotted line = bound on log scale for the smooth method (see right hand side of inequality in (9)).

Thus, the blockwise selection method involves choosing \hat{j}_1^B to be the largest resolution such that, τ_j^2 does not exceed (11):

$$\hat{j}_1^B = \min \left\{ j \geq 0 : \tau_j^2 \geq \left(2^j \sum_{\ell=1}^M \log(n^{\alpha_\ell} / \sigma_\ell) \right)^{-1} \right\} - 1. \tag{12}$$

This blockwise selection method in (12) will be referred to as the **block** method.

This maximum noise level is of the appropriate order due to the following argument. From the results of [9], the degree of ill-posedness of box-car blur in the multichannel model, $\tilde{\nu}_*$ defined in (1), satisfies the equation $2\tilde{\nu}_* + 1 = 2 + \alpha^* + 1/(2M)$ where $\alpha^* = \max_{\ell=1, \dots, M} \alpha_\ell$. Similarly, define $\alpha_* = \min_{\ell=1, \dots, M} \alpha_\ell$, then using the results in [9, Section 7] with (12) implies,

$$2^{-\hat{j}_1^B} \left(\sum_{\ell=1}^M \log(n^{\alpha_\ell} / \sigma_\ell^2) \right)^{-1} \asymp \tau_{\hat{j}_1^B}^2 \leq C n^{-\alpha_* \hat{j}_1^B} 2^{\hat{j}_1^B(1+\alpha^*+1/M)}. \tag{13}$$

Then, rearranging (13) shows that (1) agrees with the estimator in (12),

$$C \left(\frac{n^{\alpha_*}}{\log n} \right) \leq 2^{\hat{j}_1^B(2+\alpha^*+1/M)} = 2^{\hat{j}_1^B(2\tilde{\nu}_*+1)}$$

for some constant $C > 0$. If the maximum noise level was increased, then this optimal level could be violated and the optimal properties of [9] would not hold. Thus, the **block** method is a more appropriate choice to make for this context of multichannel box-car blur.

A visual comparison of the two resolution selection methods is given in Figure 3, which applies the **block** and **smooth** methods to the multichannel data displayed in the left plot of Figure 1. The results of Figure 3 were then used to generate the right plot of Figure 1 using the **MWaved** approach.

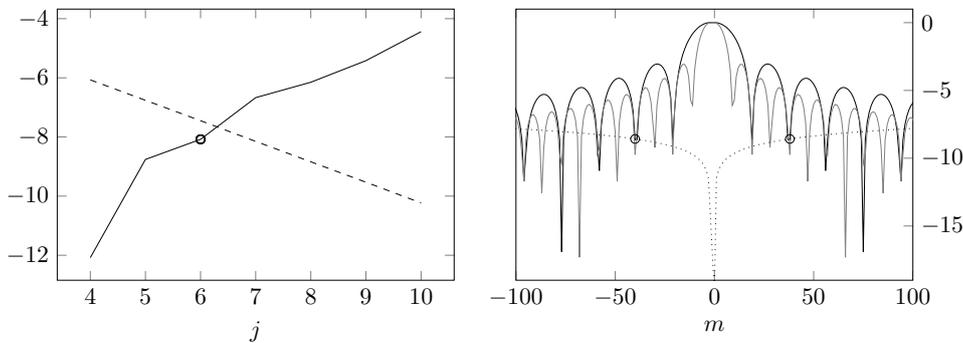


Figure 3: Block vs smooth method. Left plot = block method using log scale; solid line = $\log \tau_j^2$; dashed line = $-j \log 2 - \log \sum_{\ell} \log(n^{\alpha_\ell} / \sigma_\ell^2)$; $\hat{j}_1^B = 6$. Right plot = smooth method; black and grey lines = log decay of box-car spectrum in first and second channels; dotted = log bound; $\hat{j}_1^S = \lfloor \log_2 38 \rfloor - 1 = 4$

4 Numerical study

To justify our small theoretical result, we supplement our findings with a simulation study to compare the performance of the two methods discussed in the paper. Simulations were conducted

on the four test signals of the LIDAR, Bumps, Blocks and Doppler which are common in the literature (see e.g. [3, 9]). However, due to brevity and space constraints only the LIDAR case is reported here for the forthcoming calibrations of (3). Performance was measured by computing the Root Mean Integrated Square Error (RMISE) for the MWaveD estimates, \hat{f} , from model (3) with $M = 1, 2, 3$ or 4. The MWaveD tuning parameters used the theoretical thresholds, λ_j , from [9] with the smoothing parameter set to be $\zeta = 2\sqrt{\alpha_*}$ and the highest resolution was selected using either the `block` method or the `smooth` method. The RMISE was approximated using 1024 replications of each scenario. Each simulation was generated using $Mn = 4096$ values which were split amongst the available channels giving $n = \lfloor 4096/M \rfloor$ observations for each channel. This ensures a fair comparison of equal information across the multichannel model. The box-car half width was fixed across all channels, using $c_\ell = 1/\sqrt{353}$ or $1/\sqrt{89}$. Further, the dependence level in the noise was set to be $\alpha_\ell = 0.8$ ($H_\ell = 0.6$) in all channels (noise simulated using the `fracdiff` package from CRAN). Three noise level were studied using the signal to noise ratio, measured in dB, the low noise case (30dB), medium noise (20dB) and high noise case (10dB).

Table 1 shows the results for the simulation study where each cell entry denotes the RMISE of the estimator at each noise level and box-car width, with the average of the estimated fine resolution level (to the nearest integer) shown in parenthesis. The `block` method always produced a higher fine resolution level compared to the `smooth` method. Also, the RMISE was globally smaller for the `block` method across all channels, noise levels and box-car widths, indicating that the higher resolutions included important information resulting in a better estimate. Thus, it is clear that the `smooth` method consistently underestimates the appropriate fine resolution level. This is not surprising since the `smooth` method was designed for regular smooth blur and the `block` method has been tuned to correspond to the optimal properties in [9].

	One Channel			Two Channels		
	10dB	20dB	30dB	10dB	20dB	30dB
$1/\sqrt{353}$						
<code>block</code>	0.1647(4)	0.1297(5)	0.0818(5)	0.1647(4)	0.1297(5)	0.0818(5)
<code>smooth</code>	0.1950(3)	0.1450(4)	0.1441(4)	0.1950(3)	0.1450(4)	0.1441(4)
$1/\sqrt{89}$						
<code>block</code>	0.2039(4)	0.1606(5)	0.1061(5)	0.2039(4)	0.1606(5)	0.1061(5)
<code>smooth</code>	0.2054(3)	0.1945(3)	0.1449(4)	0.2054(3)	0.1945(3)	0.1449(4)
	Three Channels			Four Channels		
	10dB	20dB	30dB	10dB	20dB	30dB
$1/\sqrt{353}$						
<code>block</code>	0.1754(4)	0.1254(5)	0.0804(5)	0.1625(4)	0.1167(5)	0.0770(5)
<code>smooth</code>	0.1921(3)	0.1914(3)	0.1423(4)	0.1949(3)	0.1941(3)	0.1441(4)
$1/\sqrt{89}$						
<code>block</code>	0.2215(4)	0.1608(4)	0.1023(5)	0.2034(3)	0.1591(4)	0.1066(5)
<code>smooth</code>	0.2229(3)	0.1926(3)	0.1557(4)	0.2346(2)	0.1945(3)	0.1940(3)

Table 1: Monte-Carlo approximations to RMISE and resolution levels for MWaveD estimates for $M = 1, 2, 3$ or 4 in the low, medium and high noise scenarios with box-car widths = $1/\sqrt{353}$ or $1/\sqrt{89}$

5 Conclusion

We have presented a data-driven method for choosing the finest possible wavelet resolution level for wavelet deconvolution of box-car blur. This `block` method is compared with the `smooth` stopping method for resolution level estimation and shown to be superior for the box-car case. The presented method automatically adapts to the level of box-car blur (c_ℓ half widths), and to the level of noise, σ_ℓ , for all $\ell = 1, 2, \dots, M$.

It may be possible to consider the prospect of resolution estimation with noisy deconvolution ($k_{m,\ell}$ being only observable in noise). However, the `MWaveD` estimation would perhaps need to be re-considered since the weighting regime in (6) depends on $k_{m,\ell}$ directly.

The `block` methodology might also be relevant for the recent work in [1]. The authors in [1] consider a similar asymptotic condition on j_1 (denoted J in their paper). Their framework is more general in some ways than the context of [9] considered here with [1] having more relaxed conditions on the error variables being either a Gaussian variable or sub-Gaussian variable (type of moment condition). They are also some extra challenges since they consider a more general functional deconvolution type model with long memory instead of the ‘standard’ deconvolution model in (3) and [1] assume the number of channels diverges to infinity with n in their context. They also consider the block-thresholding wavelet estimator where the wavelet coefficients at each resolution level are grouped into sub-blocks of size $\log n$ and thresholded separately compared to [9] which implement the simple hard thresholded wavelet estimator where the coefficients are given a single threshold for each resolution. It seems possible that the methodology could be adapted to the context of [1] but it is outside the scope of the current work presented here. It would likely require a more delicate `block` resolution estimation method to assess the variability within each of the sub-blocks of length $\log n$ created at each resolution level implicit in the construction of the general blockwise thresholding wavelet estimator. In addition, the method might need to adapt to the diverging number of channels, M , to ensure consistency with their results and as postulated by [12] might require extra nontrivial results on number theory to allow this divergence.

Bibliography

- [1] Benhaddou, R., Kulik, R., Pensky, M., and Sapatinas, T. (2014) *Multichannel deconvolution with long-range dependence: A minimax study*, J. Statist. Plann. Inference, **148**, 1–19 (invited paper).
- [2] Cavalier, L., and Raimondo, M. (2006) *On choosing wavelet resolution in image deblurring*. Proceedings of the International Conference on Computer Graphics, Imaging and Visualization., 177–181.
- [3] Cavalier, L., and Raimondo, M. (2007) *Wavelet deconvolution with noisy eigenvalues*. IEEE Trans. Signal Process., **55**, 2414–2424.
- [4] Cavalier, L., and Hengartner, N.W. (2005) *Adaptive estimation for inverse problems with noisy operators*. Inverse Problems., **21**, 1345–1361.
- [5] De Canditiis, D., and Pensky, M. (2006) *Simultaneous wavelet deconvolution in periodic setting*. Scand. J. Statist., **33**, 293–306.

- [6] Johnstone, I. and Raimondo, M. (2004) *Periodic boxcar deconvolution and diophantine approximation*. Ann. Statist., **32**, 1781–1804.
- [7] Johnstone, I., Kerkyacharian, G., Picard, D., and Raimondo, M. (2004) *Wavelet deconvolution in a periodic setting*. J. R. Stat. Soc. Ser. B Stat. Methodol., **66**, 547–573.
- [8] Kerkyacharian, G., Picard, D. and Raimondo, M. (2007) *Adaptive boxcar deconvolution on full Lebesgue measure sets*, Statist. Sinica **17**, 317–340.
- [9] Kulik, R., Sapatinas, T. and Wishart, J.R. (2014) *Multichannel deconvolution with long range dependence: upper bounds on the L^p -risk ($1 \leq p < \infty$)*, Appl. Comput. Harmon. Anal. (to appear in).
- [10] Mallat, S. (1999) *A wavelet tour of signal processing*, Academic Press Inc.
- [11] Meyer, Y. (1992) *Wavelets and operators*, Cambridge University Press **37**.
- [12] Pensky, M. and Sapatinas, T. (2010) *On convergence rates equivalency and sampling strategies in functional deconvolution models*, Ann. Statist., **38**, 1793–1844.
- [13] Pensky, M. and Sapatinas, T. (2011) *Multichannel boxcar deconvolution with growing number of channels*, Electron. J. Stat., **5**, 53–82.
- [14] Wang, Y. (1996) *Function estimation via wavelet shrinkage for long-memory data*. Ann. Statist., **24**, 466–484.
- [15] Wishart, J.R. (2013) *Wavelet deconvolution in a periodic setting with long-range dependent errors*. J. Statist. Plann. Inference, **143**, 867–881.

Comparison of block bootstrap testing methods of mean difference for paired longitudinal data

Hirohito Sakurai, *National Center for University Entrance Examinations*, sakurai@rd.dnc.ac.jp

Masaaki Taguri, *Chuo University*, tagurimm@yahoo.co.jp

Abstract. This paper compares three block bootstrap testing methods for detecting the difference of two means in longitudinal data when the data of two groups are paired. The block resampling techniques used in this paper include moving block bootstrap, circular block bootstrap and stationary bootstrap. These are used to approximate the null distributions of test statistics. In each test we here consider the following four types of test statistics: (i) sum of absolute values of difference between two mean sequences, (ii) sum of squares of difference between two mean sequences, (iii) estimator of area-difference between two mean curves, and (iv) difference of kernel estimators based on two mean sequences. Monte Carlo simulations are carried out in order to examine the sizes and powers of the testing methods.

Keywords. Moving block bootstrap, Circular block bootstrap, Stationary bootstrap, Test of mean difference, Longitudinal data

1 Introduction

Comparison of two means or regression curves of two populations is one of important problems in statistics and related fields. Suppose now that there are paired two samples given by $\{(Y_i(t), X_i(t))\}_{i=1}^q$ for $t = 1, \dots, n$, where q and n are the numbers of subjects and observed points, respectively. And assume that, for fixed t , $Y_1(t), \dots, Y_q(t)$ are independent over q subjects, and that $X_1(t), \dots, X_q(t)$ are independent over q subjects, where $Y_i(t)$ and $X_i(t)$ are continuous in t . Then we consider the model

$$Y_i(t) = f(t) + \varepsilon_i(t), \quad X_i(t) = g(t) + \eta_i(t), \quad i = 1, \dots, q, \quad t = 1, \dots, n, \quad (1)$$

where $f(t)$ and $g(t)$ are unknown regression functions, and $\varepsilon_i(t)$ and $\eta_i(t)$ are the error terms having means 0 and finite variances. Our problem is then to test

$$H_0 : f(t) = g(t) \text{ for all } t \quad \text{vs.} \quad H_1 : f(t) \neq g(t) \text{ for some } t, \quad (2)$$

where H_0 and H_1 are the null and alternative hypotheses, respectively.

For example, Figure 1 shows a real dataset of wind velocity measured by an artificial satellite and a radar on the earth, where $q = 11$ and $n = 13$. From this dataset, we want to know whether the mean behavior of the two devices in measuring wind velocity is equal or not. Then the problem is formulated as (2) and the significant difference between them is investigated by some methods, which is briefly explained in Section 4.

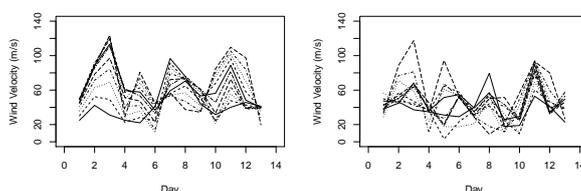


Figure 1: Wind velocity data (left: satellite, right: radar)

For (2), there are several methods assuming that the error terms are independent and identically distributed (i.i.d.) and are normal. However, it may be unrealistic to put such assumptions when we analyze a real dataset. If we cannot assume the normality, the nonparametric approach, for example by [1], is available. Another possible choice would be an application of resampling such as nonparametric bootstrap. In i.i.d. setting, the bootstrap [2] is a quite useful tool, however, in time series analysis, the naive application fails to capture the dependent structure of data because it ignores the order of observations. In order to overcome this problem, model-based and block resampling approaches are proposed; see, for example, [5] and references therein.

In this paper, we focus on a paired two-sample problem which can be reduced to a one-sample problem, and compare three block bootstrap testing methods, [8], [9] and [10], for detecting the difference of two means in paired longitudinal data, which can be viewed as tests for functional data (e.g. [11]). The contribution of this paper is to clarify the mutual relationships of the level and power properties among [8], [9] and [10], since such investigation has not been done yet. The block bootstraps considered in this paper include moving block bootstrap [4], circular block bootstrap [6] and stationary bootstrap [7]. In Section 2, we review the testing methods, [8], [9] and [10], which calculates p -values (achieved significance levels) using [4], [6] and [7], respectively. In order to investigate the properties of sizes and powers of the above testing methods, Monte Carlo simulations are carried out in Section 3, and an example of real data analysis and some concluding remarks are given in Section 4.

2 Testing methods using block bootstraps

In this section, we review the testing methods using block bootstrap proposed by [8], [9] and [10]. Note that the area-difference given by $A = \int |f(t) - g(t)| dt$ is 0 under H_0 and positive under H_1 . Then, the hypothesis of our interest reduces to testing

$$H_0 : A = 0 \quad \text{vs.} \quad H_1 : A > 0. \quad (3)$$

We first introduce several test statistics to detect the difference between $f(t)$ and $g(t)$ in (1). The following statistic is proposed by [3]:

$$S_n = S_n(D_1, \dots, D_n) = \left[\sum_{j=0}^{n-1} \left(\sum_{t=j+1}^{j+h} D_t \right)^2 \right] \left[n \sum_{t=1}^{n-1} \frac{(D_{t+1} - D_t)^2}{2} \right]^{-1}, \quad (4)$$

where $D_t = Y_t - X_t$ for $t = 1, \dots, n$, or $D_t = Y_{t-n} - X_{t-n}$ for $t = n+1, \dots, n+h$, $Y_t = \sum_{i=1}^q Y_i(t)/q$, $X_t = \sum_{i=1}^q X_i(t)/q$, $h = [np]$ is the integer part of np , and p is a tuning constant satisfying $0 < p < 1$ which is determined by the fully data-driven approach; the second approach described in [3, pp.1043–1044]. The statistic (4) is essentially based on kernel estimators of $f(t)$ and $g(t)$. As another type of test statistics, we can use

$$T_{1n} = T_{1n}(D_1, \dots, D_n) = \sum_{t=1}^n |D_t|, \quad T_{2n} = T_{2n}(D_1, \dots, D_n) = \sum_{t=1}^n D_t^2. \quad (5)$$

Further, taking account of the area-difference A ,

$$T_{3n} = T_{3n}(D_1, \dots, D_n) = \frac{1}{2} \sum_{t=1}^{n-1} (|D_t| + |D_{t+1}|) I_+ + \frac{1}{2} \sum_{t=1}^{n-1} \frac{|D_t|^2 + |D_{t+1}|^2}{|D_t| + |D_{t+1}|} I_- \quad (6)$$

is also available as a test statistic, where $I_+ = I\{D_t D_{t+1} \geq 0\}$, $I_- = I\{D_t D_{t+1} < 0\}$ and $I\{\cdot\}$ is the indicator function, respectively. The test statistic (6) seems to have a complicate form, however it is an estimator of A in (3) constructed by the trapezoidal rule with linear interpolations of adjacent observation values. Intuitively the values (4), (5) and (6) are small when H_0 is true, while they are large when H_0 is false. Therefore, T_{1n}, T_{2n}, T_{3n} and S_n enable us to measure the discrepancy between $f(t)$ and $g(t)$.

Next we explain three testing algorithms referring to [4], [6] and [7]. We call them Moving Block Bootstrap (MBB), Circular Block Bootstrap (CBB) and Stationary Bootstrap (SB) tests, respectively. The main idea of these methods is that we apply MBB, CBB and SB techniques in order to approximate the null distribution of (4), (5) and (6). Let $D_{0,t} = D_t - \bar{D} = D_t - \sum_{t=1}^n D_t/n$ for $t = 1, \dots, n$, and d_t and $d_{0,t}$ denote realizations of D_t and $D_{0,t}$. For simplicity, let T be a generic notation for T_{1n}, T_{2n}, T_{3n} or S_n . For a given significance level α , MBB, CBB and SB tests are described in Algorithms 2.1, 2.2 and 2.3, respectively.

Algorithm 2.1 (MBB test).

Step 1 Divide the centered observations $d_{0,1}, \dots, d_{0,n}$ into $k (= n - \ell + 1)$ successive overlapping blocks $\{\xi_1, \dots, \xi_k\}$ with each length $\ell (\leq n)$, where $\xi_t = \{d_{0,t}, \dots, d_{0,t+\ell-1}\}$ for $t = 1, \dots, k$.

Step 2 Draw m blocks $\{\xi_1^{*b}, \dots, \xi_m^{*b}\}$ randomly with replacement from $\{\xi_1, \dots, \xi_k\}$, and put first n elements of $\{\xi_1^{*b}, \dots, \xi_m^{*b}\}$ as a resample $\{d_1^{*b}, \dots, d_n^{*b}\}$, where $m = [n/\ell]$ (if n/ℓ is an interger) or $m = [n/\ell] + 1$ (otherwise).

Step 3 Calculate $t^{*b} = T(d_1^{*b}, \dots, d_n^{*b})$ based on the resample in Step 2.

Step 4 Calculate the achieved significance level, $\widehat{ASL} = \sum_{b=1}^B I\{t^{*b} \geq t_{obs}\}/B$, by repeating Steps 2 and 3 an appropriate number of times B , and reject H_0 if $\widehat{ASL} \leq \alpha$, where $t_{obs} = T(d_1, \dots, d_n)$.

Algorithm 2.2 (CBB test).

CBB test is different from MBB test only in the construction of blocks, so Steps 3 and 4 are same as in MBB test. Steps 1 and 2 in CBB test are given as follows:

Step 1 Divide $\{d_{0,1}, \dots, d_{0,n}\}$ into n blocks with each length ℓ in the manner of circular block bootstrap [6], and put

$$\xi_j = \begin{cases} \{d_{0,j}, \dots, d_{0,j+\ell-1}\}, & j = 1, \dots, n - \ell + 1, \\ \{d_{0,j}, \dots, d_{0,n}, d_{0,1}, \dots, d_{0,j+\ell-n-1}\}, & j = n - \ell + 2, \dots, n. \end{cases}$$

Step 2 Draw m blocks $\{\xi_1^{*b}, \dots, \xi_m^{*b}\}$ randomly with replacement from $\{\xi_1, \dots, \xi_n\}$, and take first n elements of $\{\xi_1^{*b}, \dots, \xi_m^{*b}\}$ as a resample $\{d_1^{*b}, \dots, d_n^{*b}\}$ ($b = 1, \dots, B$), where m is defined in Step 2 of Algorithm 2.1.

Algorithm 2.3 (SB test).

Step 1 Define $\xi^{*b}(t, \ell)$ as the block starting from $d_{0,t}$ with length $\ell (\geq 1)$, and

$$\xi^{*b}(t, \ell) = \begin{cases} \{d_{0,t}, \dots, d_{0,t+\ell-1}\}, & t = 1, \dots, n - \ell + 1, \\ \{d_{0,t}, \dots, d_{0,n}, d_{0,1}, \dots, d_{0,t+\ell-n-1}\}, & t = n - \ell + 2, \dots, n. \end{cases}$$

Step 2 Generate $L_1^{*b}, \dots, L_K^{*b} \stackrel{i.i.d.}{\sim} \text{Geo}(s)$ and $I_1^{*b}, I_2^{*b}, \dots, I_K^{*b} \stackrel{i.i.d.}{\sim} \text{DU}(n)$ for $K = \min\{k : \sum_{i=1}^k L_i^{*b} \geq n\}$ in the manner of SB [7], where $\text{Geo}(s)$ and $\text{DU}(n)$ denote geometric distribution with parameter $s = 1/\ell$, and discrete uniform distribution on $\{1, 2, \dots, n\}$, respectively.

Step 3 Combine K blocks: $\xi^{*b} = \{\xi^{*b}(I_1^{*b}, L_1^{*b}), \dots, \xi^{*b}(I_K^{*b}, L_K^{*b})\}$.

Step 4 Construct a resample $\{d_1^{*b}, \dots, d_n^{*b}\}$ by putting the first n elements of ξ^{*b} .

Step 5 Calculate $t^{*b} = T(d_1^{*b}, \dots, d_n^{*b})$ for $b = 1, \dots, B$, and reject H_0 if $\widehat{ASL} \leq \alpha$ similar to Step 4 in MBB and CBB tests.

3 Numerical examination

We carry out Monte Carlo simulations to investigate the size and power properties of the testing methods considered in Section 2. For comparison, we also conduct Bowman and Young's test for paired data [1, p.85] (hereafter termed "BY" for short). In our level and power studies, the nominal level is $\alpha = 0.05$ and 0.10 . All our results are based on independent 2000 simulation replications of paired two samples, $\{(Y_i(t), X_i(t))\}$, where $B = 2000$ replications of resampling are applied to every two samples in Algorithms 2.1, 2.2 and 2.3. Naturally, the same initial samples are used for the comparisons.

We generate initial samples according to (1) whose means are specified by $f(t) = c$ and $g(t) = 0$, where $c = 0, 0.2, 0.4, 0.6, 0.8, 1.0$. The case of $c = 0$ or $c \neq 0$ corresponds to the null hypothesis or the alternative hypothesis being true. The values, q and n , are $q = 10, 20, 30$ and $n = 10$. As for the error terms $\varepsilon_i(t)$ and $\eta_i(t)$, if $f(t)$ and $g(t)$ explain most of the correlation structure contained in the data, then it may be realistic to consider that the errors are nearly

i.i.d., or very weak dependency exists in $\varepsilon_i(t)$ and $\eta_i(t)$. Thus, we choose the following Gaussian AR(1) errors: $\varepsilon_i(t) = \phi\varepsilon_i(t-1) + z_{1i}(t)$ and $\eta_i(t) = \phi\eta_i(t-1) + z_{2i}(t)$, where $z_{1i}(t) \stackrel{i.i.d.}{\sim} N(0, \tau^2)$, $z_{2i}(t) \stackrel{i.i.d.}{\sim} N(0, \tau^2)$, $\phi = 0, \pm 0.1, \pm 0.2$, $\tau^2 = (1 - \phi^2)V(\varepsilon_i(t)) = (1 - \phi^2)V(\eta_i(t))$, and $V(\varepsilon_i(t)) = 1, 3, 5$. The computation has been carried out for all combinations of these parameters, however, to save space, the results for the case of $\alpha = 0.05$, $q = 10, 30$ and $V(\varepsilon_i(t)) = 3$ are given in Tables 1 and 2.

Since it is preferable that the empirical level is nearly equal to the nominal level α , our choice of ℓ in MBB, CBB and SB tests is done so that the empirical level is close to α . If there are some candidates which have the same level errors, we make the conservative choice, viz., we choose ℓ such that the empirical level is less than the nominal level. Further if there are some candidates whose empirical levels are equal, we select ℓ to maximize the empirical power among them.

The resulting choices of ℓ are given in Table 1. This table shows that CBB and SB tests need longer ℓ for $\phi \leq 0$ and shorter ℓ for $\phi > 0$, and that MBB test does not need longer ℓ for both cases. In general, since the block bootstrap methods take account of dependency structure of observations, it is expected that we need longer ℓ in all MBB, CBB and SB tests, however the results of Table 1 show that MBB test with T_{rn} ($r = 1, 2, 3$) may reduce to the case of i.i.d. resampling.

q	ϕ	MBB				CBB				SB			
		T_{1n}	T_{2n}	T_{3n}	S_n	T_{1n}	T_{2n}	T_{3n}	S_n	T_{1n}	T_{2n}	T_{3n}	S_n
10	-0.2	1	1	1	3	5	6	4	2	9	5	3	2
	-0.1	1	1	1	2	7	4	2	1	5	4	2	1
	0	1	1	1	2	4	1	1	1	3	1	1	1
	0.1	1	1	1	2	2	1	1	1	1	1	1	1
	0.2	1	1	1	2	1	1	1	1	1	1	1	1
30	-0.2	1	1	1	3	6	5	4	2	9	8	4	2
	-0.1	1	1	1	3	7	3	2	2	8	3	2	2
	0	1	1	1	3	4	1	1	2	3	1	1	1
	0.1	1	1	1	2	2	1	1	1	1	1	1	1
	0.2	1	1	1	2	1	1	2	2	1	1	1	2

Table 1: Optimum ℓ in MBB, CBB and SB tests for $\alpha = 0.05$ and $V(\varepsilon_i(t)) = 3$

Now, we first summarize the results of the level studies. The empirical levels of MBB, CBB, SB and BY tests are given in Table 2. From this table, we can observe that BY test has a large level error, while MBB, CBB and SB tests have a tendency to keep the nominal level α as a whole. In particular, the level error of CBB and SB tests is small for $\phi \leq 0$. For $\phi > 0$, the level error of MBB, CBB and SB tests with T_{1n} and S_n is small, however those with T_{2n} and T_{3n} seems to be slightly large.

Next, we discuss the power studies. Since we found similar tendencies among the nine cases of $(q, V(\varepsilon_i(t)))$, we show the results for $(q, V(\varepsilon_i(t))) = (20, 3)$ with $\phi = 0, \pm 0.2$. The empirical powers were affected by the number of subjects q as well as the variance of noise. The increase of noise variance causes the decrease of empirical power as a whole, however the power properties among the three variances given above were nearly equal to each other. Thus, we choose and discuss the case where $V(\varepsilon_i(t)) = 3$.

Since the bad behavior of BY test was observed in our level study, we exclude the results of BY test and our discussion below concentrates on the properties of MBB, CBB, and SB tests. Figure 2 compares the empirical powers corresponding to T_{1n}, T_{2n}, T_{3n} , and S_n when the block

q	ϕ	MBB				CBB				SB				BY
		T_{1n}	T_{2n}	T_{3n}	S_n	T_{1n}	T_{2n}	T_{3n}	S_n	T_{1n}	T_{2n}	T_{3n}	S_n	
10	-0.2	0.012	0.026	0.026	0.056	0.045	0.052	0.050	0.053	0.048	0.048	0.047	0.051	0.531
	-0.1	0.021	0.036	0.042	0.047	0.056	0.048	0.051	0.035	0.051	0.052	0.052	0.035	0.532
	0	0.029	0.051	0.054	0.048	0.049	0.049	0.055	0.042	0.049	0.050	0.054	0.043	0.535
	0.1	0.041	0.068	0.081	0.043	0.057	0.067	0.078	0.059	0.040	0.068	0.077	0.054	0.546
	0.2	0.057	0.092	0.111	0.050	0.059	0.095	0.111	0.068	0.059	0.092	0.112	0.069	0.537
30	-0.2	0.009	0.021	0.024	0.051	0.043	0.054	0.052	0.043	0.045	0.051	0.053	0.043	0.414
	-0.1	0.018	0.032	0.037	0.051	0.052	0.045	0.047	0.054	0.050	0.051	0.051	0.051	0.400
	0	0.029	0.046	0.051	0.053	0.054	0.047	0.053	0.052	0.047	0.049	0.052	0.057	0.415
	0.1	0.044	0.070	0.078	0.045	0.053	0.070	0.080	0.061	0.045	0.070	0.077	0.066	0.435
	0.2	0.064	0.100	0.120	0.043	0.064	0.099	0.110	0.068	0.064	0.100	0.120	0.070	0.434

Table 2: Empirical levels of MBB, CBB, SB and BY tests for $\alpha = 0.05$ and $V(\varepsilon_i(t)) = 3$

bootstrap method is fixed in each test. This figure shows that the empirical power of T_{3n} is most powerful among them, and that the relationship among powers corresponding to the four test statistics is given by $T_{3n} \geq T_{2n} \geq T_{1n} \geq S_n$ in most cases. This indicates the numerical superiority of these tests using T_{3n} in power. On the other hand, Figure 3 shows the comparison of empirical powers corresponding to MBB, CBB and SB tests when the test statistic is fixed. From this figure, we can observe that CBB and SB tests are more powerful than MBB test, and that the empirical powers of CBB and SB tests are quite similar to each other.

4 Wind velocity data analysis and some concluding remarks

Applying MBB, CBB and SB tests with every possible ℓ to the wind velocity data in Figure 1, we obtain the results that ASL's of MBB, CBB and SB tests with T_{1n} , T_{2n} and T_{3n} are all 0.000 for $\ell = 1, \dots, 12$; those with S_n for $\ell = 1, \dots, 12$ are 0.087, 0.293, 0.820, 0.875, 0.960, 0.639, 0.496, 0.250, 0.000, 0.000, 0.000, 0.252 in MBB test; 0.083, 0.093, 0.128, 0.162, 0.172, 0.154, 0.178, 0.154, 0.122, 0.093, 0.130, 0.133 in CBB test; and 0.086, 0.135, 0.133, 0.160, 0.148, 0.145, 0.151, 0.151, 0.158, 0.148, 0.150, 0.153 in SB test. BY test rejects the null hypothesis in (3). Therefore, there is a possibility of the significant difference between the satellite and radar in measuring wind velocity.

In this paper we have compared three block bootstrap testing methods MBB, CBB and SB for two means in paired longitudinal data. Our numerical studies indicate the applicability of MBB, CBB and SB tests for weakly dependent data even when the sample size is very small. In some cases, we have confirmed the effectiveness of application of CBB and SB tests using T_{2n} and T_{3n} as test statistics. The problem on block length selection in the block resampling is very important, and the development of a fully data-driven approach to selecting (mean) block length in the above tests will be needed for practical data analyses. Further, extensions of the testing methods to several curves and/or grouped data cases would be required in the future.

Bibliography

- [1] Bowman, A. and Young, S. (1996) *Graphical comparison of nonparametric curves*. Applied Statistics, **45**, 83–98.

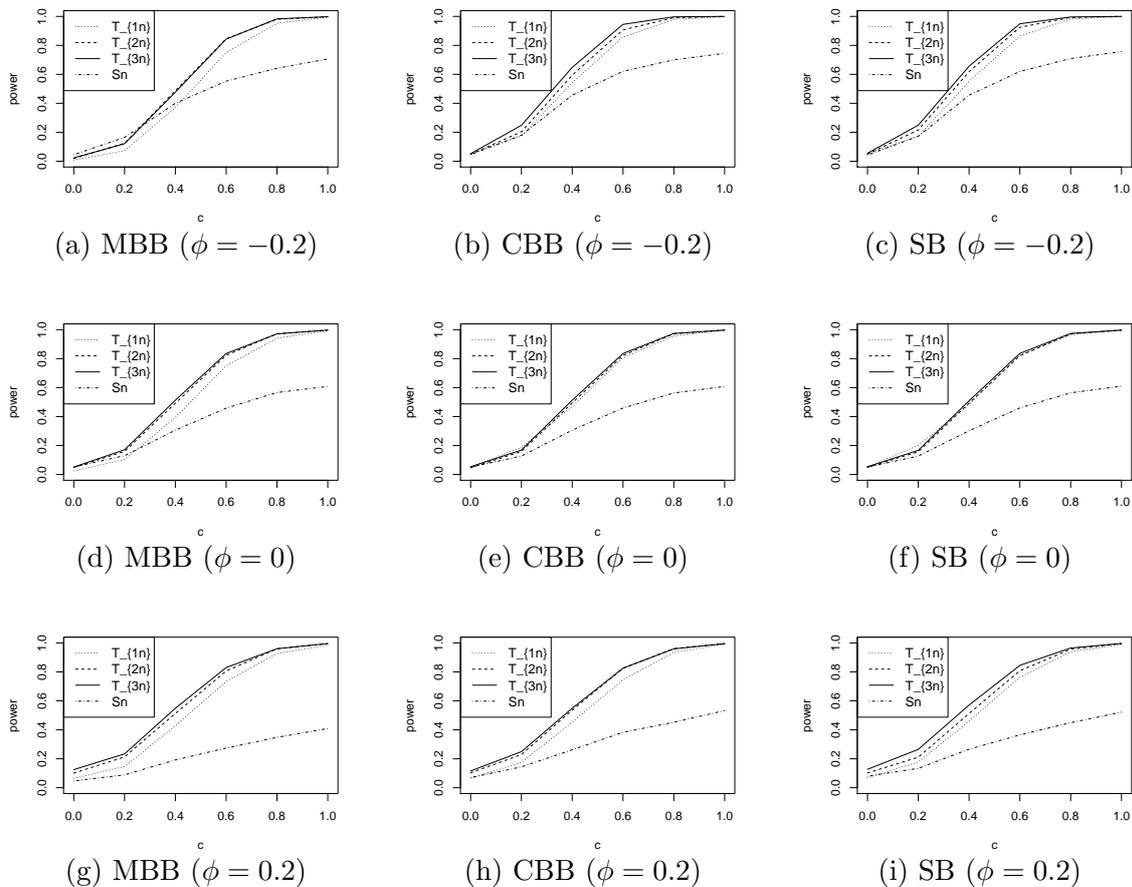


Figure 2: Comparison of four statistics for $\alpha = 0.05$, $q = 20$ and $V(\varepsilon_i(t)) = 3$

- [2] Efron, B. (1979) *Bootstrap methods: another look at the jackknife*. Annals of Statistics, **7**, 1–26.
- [3] Hall, P. and Hart, J. D. (1990) *Bootstrap test for difference between means in nonparametric regression*. Journal of the American Statistical Association, **85**, 1039–1049.
- [4] Künsch, H. R. (1989) *The jackknife and the bootstrap for general stationary observations*. Annals of Statistics, **17**, 1217–1241.
- [5] Lahiri, S. N. (2003) *Resampling Methods for Dependent Data*. Springer.
- [6] Politis, D. N. and Romano, J. P. (1992) *A circular block-resampling procedure for stationary data*. Exploring the Limit of Bootstrap (eds. LePage, R. and Billard, L.), Wiley, 263–270.
- [7] Politis, D. N. and Romano, J. P. (1994) *The stationary bootstrap*. Journal of the American Statistical Association, **89**, 1303–1313.
- [8] Sakurai, H. and Taguri, M. (2004) *Moving block bootstrap test for mean difference in paired longitudinal data*. Proceedings of the Eighth China-Japan Symposium on Statistics, 240–243.

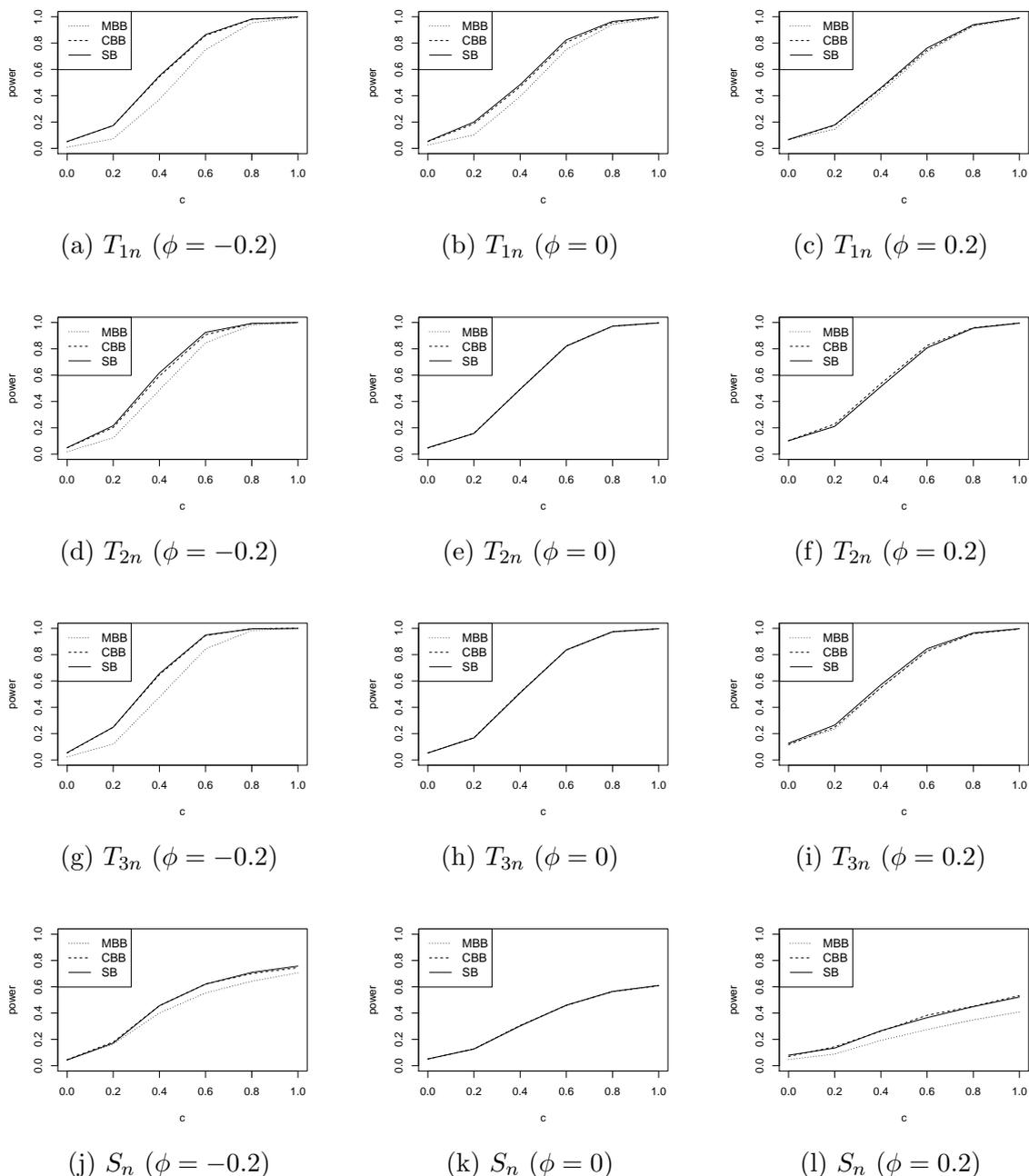


Figure 3: Comparison of MBB, CBB and SB tests for $\alpha = 0.05$, $q = 20$ and $V(\varepsilon_i(t)) = 3$

- [9] Sakurai, H. and Taguri, M. (2008) *Test of mean difference for paired longitudinal data based on circular block bootstrap*. COMPSTAT2008 Proceedings in Computational Statistics (ed. Brito, P.), Physica-Verlag, 679–687.
- [10] Sakurai, H. and Taguri, M. (2011) *Test of mean difference for paired longitudinal data using stationary bootstrap*. Program Book of Joint Meeting of the 2011 Taipei International

Statistical Symposium and 7th Conference of the Asian Regional Section of the IASC, 184.

- [11] Zhang, J.-T., Liang, X. and Xiao, S. (2010) *On the two-sample Behrens-Fisher problem for functional data*. *Journal of Statistical Theory and Practice*, **4**, 571–587.

Functional data modeling to measure exposure to ozone

M. Arisido, *University of Padova, Department of Statistical Science, Italy*, Arisido@stat.unipd.it

Abstract. One of the many challenges involved in environmental studies of pollutants on human health is how to measure the daily exposure to ozone. Despite hourly measures of ozone concentrations are available, studies on short-term effects of ozone and human health reduce the hourly measures to a single daily summary measure, such as daily average, daily maximum etc. This reduction leads to disregard the non-uniform temporal distribution of the pollutant, and can be an issue in modelling the association between short-term effects of ozone and human health outcome. We present alternative approach by treating all hourly measures of a day as one function. The functional form of ozone incorporates all hourly measures and aids to uncover important features of the daily patterns of ozone. To investigate the effect of the hourly records on health, we consider a functional generalized linear model (FGLM) in which the predictor is functional ozone and the response is daily hospital admissions. The model allows to estimate the effect of ozone as a function of daily hour that allows to examine the influence of the pollutant throughout the day. Thus, the portion of daily ozone function potentially linked to health can be recognized. We demonstrate the superiority of our approach over the classical models that use daily summary measures using out-of-sample predictive performance.

Keywords. functional data; hospital admission; lag; Ozone.

1 Introduction

Despite the increasing number of statistical studies on pollutants and health, there have been rare methodology that take into account the daily patterns of the concentrations. Often, although hourly measures of concentrations are available, studies on short-term effects of ozone reduce the hourly measures to a single daily summary statistics such as, daily maximum, daily average etc [4, 13]. Those summary statistics lead to the use of classical statistical methods such as the generalized additive models (GAMs) [5] to investigate the effect of ozone on health [2]. However, those summary statistics are rough syntheses of the daily patterns of ozone concentrations. Using them, totally disregard the non-uniform temporal distribution of the pollutant.

We propose the use of functional regression models, that allow to incorporate all hourly measurements of ozone. The interest is to investigate the association between exposure to ozone and daily hospital admissions (Morbidity) in a functional data analysis (FDA) setting. We adopt the functional generalized linear model (FGLM) to explain the Poisson daily hospital admissions. We show that aligning common features of the ozone functions can help to identify the portion of ozone curve potentially linked to health.

The study used data limited to the summer periods (June-July-August) of the years 1996-2002 for city of Milan, Italy. Daily hospital admission data for the specified period were obtained from the regional health informative system for all hospitals located in the city of Milan. Meteorological and environmental data for the same periods were obtained from the regional agency for environmental protection (ARPA) of Lombardia.

2 Methods: Functional data Modelling and Application

The term 'Functional data' was introduced by [11] to denote samples that consists of curves. The idea behind analyzing data using FDA technique is to change the discrete timely observed data into functions, thereby one function is considered as a single observation. In our application, let $X_i(t)$, $i = 1, \dots, N$ denote the day i ozone concentrations recorded over the daily hours t . In practice, ozone concentrations are measured at discrete grid of points t_j , $j = 1, \dots, J$ for each day. For day i hourly ozone concentrations were measured at $J = 24$ discrete time points. To estimate the continuous curve $X_i(t)$, we used a set of B-spline basis functions ϕ_1, \dots, ϕ_K , so that $X_i(t)$ can be specified in the form

$$X_i(t) = \sum_{k=1}^K c_k \phi_k(t) \quad (1)$$

where K is the number of basis functions, c_k are the coefficients of the basis to be estimated using smoothing technique. We apply the roughness penalty approach criterion of [10] to smooth the ozone data as:

$$SSE(c) = \sum_j^N (y_{ij} - X_i(t_j))^2 + \lambda \int_0^T (X_i''(t_j))^2 dt \quad (2)$$

where y_{ij} is the the recorded ozone data in the day i at daily hour j , and the integrated, squared, second derivative measures the roughness of the function. The parameter λ continuously controls the smoothness of the fit, and can be selected by the generalized cross-validation criterion (GCV) [12]. Sample of 20 functional data obtained using (1) are shown in Figure 1 (left). As can be observed from the plotted functions, the curves exhibit identifiable features (landmarks) such as daily maxima or minima. However, the location of these features varies from day to day; each function reaches its maximum concentrations level at different hour; which make difficult to compare day to day variability of the pollutant at their respective maxima. The curves could be aligned to have the maximal ozone concentrations at the same time point [11, 3]. This may be an important idea in the context of air pollutant and health to capture the portion of ozone curve near maximum that can be potentially harmful. Formally, the i^{th} function, $x_i(t)$ is transformed to $X_i^*(t) = X_i(h_i(t))$, where $h_i(t)$ is called time-warping function which defines the transformation [11]. Figure 1 (right) displays the same sample functional data after alignment.

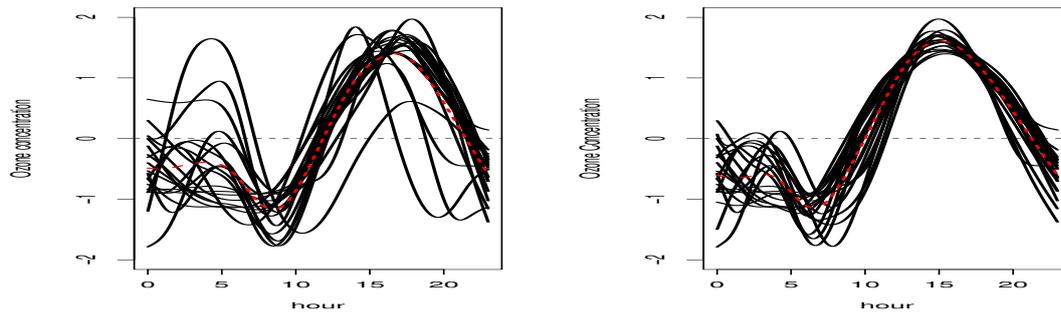


Figure 1: Sample of 20 ozone functional data (left) and the same ozone functional data after alignment using hour at which the ozone concentrations of a function is maximum (right).

Functional regression models

The daily number of hospital admissions y_i in day i is assumed to be distributed as a Poisson variable with mean denoted by $\mu_i = E(y_i)$. To model the data, the link function $\log(\mu_i)$ is connected to the predictor. This was done using functional generalized linear models (FGLM), the extension of generalized linear models [8] to the case in which the predictor is functional and the response is scalar. The model is given by

$$\log \mu_i = \alpha + \int_0^T X_i(t)\beta(t)dt \quad (3)$$

where X_i is the day i ozone concentrations curve as obtained in Section 2, $\beta(t)$ is the functional coefficient which describes the influence of the pollutant at time t on daily number of admissions. For example, the value of $\beta(t)$ evaluated at the daily hour 1 pm describes the influence of this hourly ozone on one day total number of admissions. The intercept is represented by α . The functional coefficient $\beta(t)$ can be specified in the form $\beta(t) = \sum_{l=1}^L b_l \theta_l(t)$, where b_l are coefficients of the B-spline basis θ_l with dimension L . To estimate b_l , we used P-spline type penalty given by [7], which requires using a simple difference penalty on the coefficients b_l . Formally, let $\Delta b_j = b_j - b_{j-1}$, the second difference penalty is $\lambda \sum_j (\Delta^2 b_j)^2$, where λ still controls the weight of the penalty. We included other confounding predictors including other pollutant in 3 either as linear components or non-linear smooth functions. Particularly, particulate matter (PM), a complex mixture of small particles and liquid droplets including acids and other organic chemicals, may account the observed relationship between ozone and health outcome [6]. For this reason, we considered the daily average of of particulate matter smaller than 10 micrometers in diameter (PM_{10}) and its one day lag. As it is commonly done, we included the daily maximum temperature as smooth non-linear function ($f(\text{Temp}_i)$) instead of restricted parametric form. To control seasonal confounding, we also included day of the week (DOW) and calendar year (Year) as factor predictors. For day of the week, (Monday = 1, ..., Saturday = 6) and Sunday is the baseline. Similarly, for calendar year, (1997 = 1, ..., 2002 = 6) and year 1996 is the baseline. The full flexible model is

$$\log \mu_i = \alpha + \int_0^T X_i(t)\beta(t)dt + PM_{10} + f(\text{Temp}_i) + \sum_{j=1}^6 \delta_j \text{DOW}_j + \sum_{k=1}^6 \gamma_k \text{Year}_k \quad (4)$$

Often, the impact of the pollutants on health persists for some days from the date of exposure. For this reason measures of pollutants are often lagged by a number of days to explain the current day health outcome. To examine the best time lag to use, we consider four different lagged values (0-3 lags) of functional ozone.

Model selection

For model selection, two methods can be considered: the Unbiased Risk Estimate (UBRE) and the generalized cross-validation (GCV) which are suggested by [14] in the context of generalized additive model (GAM). The UBRE is suggested when the scale parameter of the distribution of the data is known as the Poisson case in our data. For which the UBRE score is given as $\frac{D}{n} + 2\frac{P}{n}$, where D is the deviance (twice the difference between the log-likelihood for the saturated model and the log-likelihood for the present model), P is the total degrees of freedom and n is number of observations (functions). The model with small UBRE score is the better model in terms of goodness of fit. If the scale parameter is unknown, model selection can be done using the well known generalized cross-validation(GCV) criterion [5, 12], computed as $\frac{nD}{(n-p)^2}$.

3 Results

Four different models were fitted according to number of lagged days for functional ozone. The estimates of $\beta(t)$ together with 95% point-wise intervals for each model were shown in Figure 2. The estimates were achieved using 8 B-spline basis of order 3 and second order difference penalty. The results were insensitive to the change in other values of B-spline. The confidence interval of the estimated coefficient curve of the current day ozone exposure (lag 0) involves zero almost throughout the day. The functional estimate of one day lagged values of ozone (lag 1) is significant mainly in the afternoon of the daily hours, which indicates that exposure to ozone in the previous day is associated to the current day hospital admissions. The 95% point-wise confidence intervals for the estimate of two days lagged values of ozone (lag 2) and three days lagged values (lag 3) involve zero in the majority of the regions, but they show a significant effect in the night hours. Note that the same overall patterns for lag 2 and lag 3 estimates, which may indicate the persistence may last up to 3 days. To confirm this, we fitted a model using lag 7 to investigate if the estimate would be different from lag 2 and lag 3; the resulting estimate was very close to that of the estimate from lag 3.

The results of the confounding predictors included as linear components were summarized in Table 1. The estimated value (associated standard deviations in the bracket) are given for the current day PM₁₀ and its one day lag (PM₁₀^{lag}), which are significantly associated with daily number of hospital admissions. The factor predictors days of the week and calendar year are significant each at 6 degrees of freedom.

The Unbiased Risk Estimate (UBRE) (see model selection in Sect.2) is used to select the best time lag for ozone exposure. The UBRE scores given in Table 1 show that lag 1 ozone exposure produced lower UBRE score compared to the other lags; the dependence of hospital admissions on one day lag functional ozone is selected as best.

A second model estimates were considered for the aligned ozone curves. The main objective is to improve the estimated coefficient curve by aligning the observed curves at the daily maximum

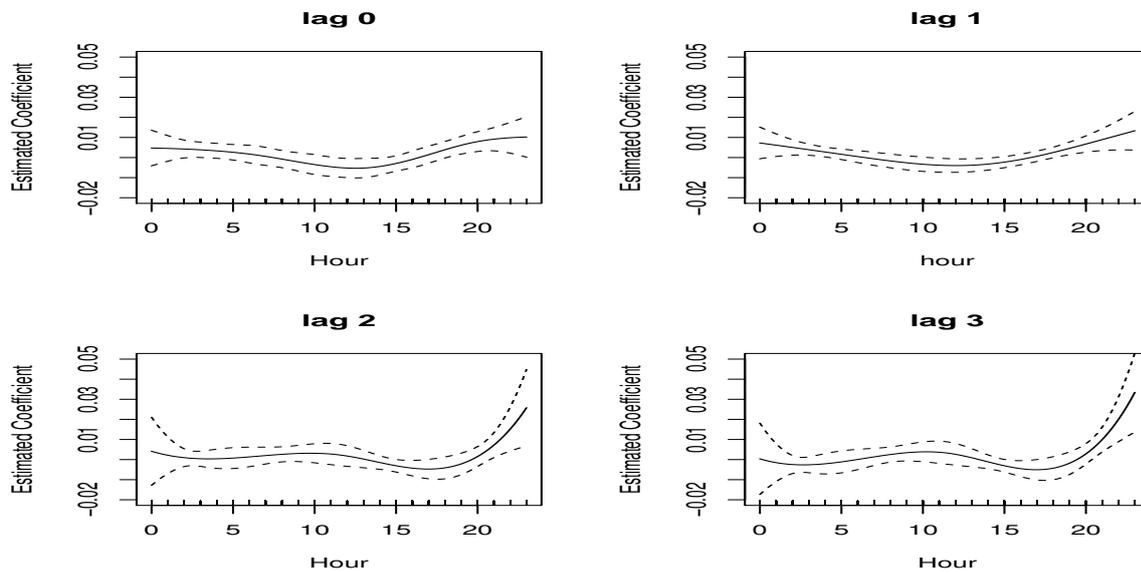


Figure 2: Estimate of ozone functional coefficient, $\beta(t)$, with different lag controlling other confounding predictors

The functional ozone with the following lags:				
	lag 0	lag 1	lag 2	lag 3
PM ₁₀	0.025*	0.024*	0.024*	0.026*
	(0.008)	(0.007)	(0.008)	(0.008)
PM ₁₀ ^{lag}	0.019*	0.020*	0.020*	0.021*
	(0.008)	(0.075)	(0.007)	(0.008)
Observations	599	598	597	596
UBRE	0.511	0.504	0.516	0.523

Table 1: Results for predictors with linear components under the functional linear regression models using different lag structure for the functional ozone. The estimates (associated standard deviation in the bracket) are given for particulate matter (PM₁₀) and the one day lag particulate matter (PM₁₀^{lag})

* indicates the significance of the predictor
 UBRE: Unbiased Risk Estimate.

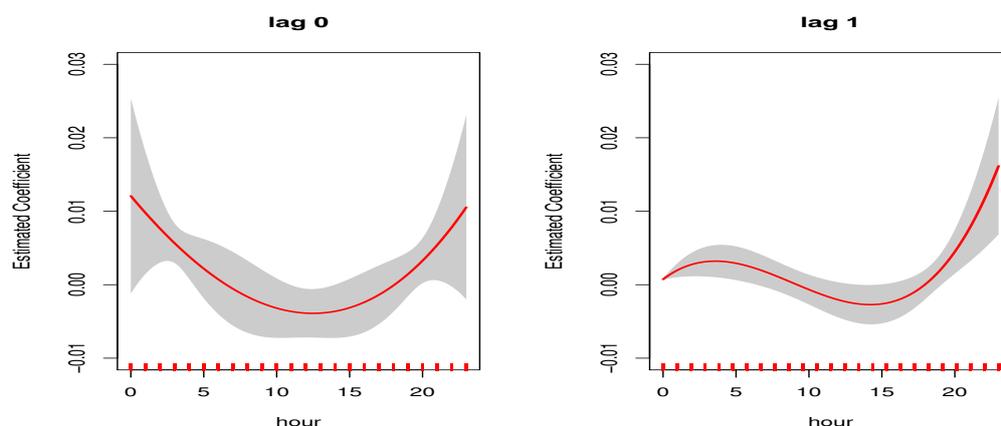


Figure 3: Estimated coefficient curves of ozone under linear functional model using aligned ozone curves at daily maximum for the same day ozone exposure (lag 0) and the previous day ozone exposure (lag 1)

(3 pm). The estimates of the confounding predictors and their significance levels were the same as the analysis under the original (unaligned) curves. Using the UBRE criterion, still the lag 1 ozone exposure model came out the best as compared to the other lag structure under aligned ozone curve analysis. In fact, the estimates for lag 2 and lag 3 were not significantly different from their estimates in Figure 2. As such, we report the estimates for the the same day ozone (lag 0) and the previous day ozone exposure (lag 1) in Figure 3. These estimates are fairly improved under the alignment. Particularly, it is interesting to note that the lag 1 estimate (Figure 3, right) does not involve zero starting from the point at which the original curves were aligned (3 pm). Thus, the region of the estimated curve starting from the daily maximum is significantly associated to hospital admissions.

We notice that the FGLM results agree with preliminary results where models were fitted at each hourly ozone concentrations using classical generalized additive models (GAM). The model with concentrations level in the evening came out as the preferred one according to UBRE (see model selection in Sec. 2). Figure 4 reports the estimated effected with associated 95% confidence intervals and the UBRE score of the models. Although the estimated ozone effect is significant for each model, the estimated effect is maximum at the daily hour of 9 pm and the UBRE is minimum at the same time point. The Figure includes two more models fitted to the daily average and the daily maximum (the vertical lines, in the right side) to compare with the results of the hourly estimates. We note that the hourly measure and the daily summaries (average and maximum) data are scaled appropriately prior to model fitting. The daily maximum produced higher estimate than the hourly measures, but once again, the UBRE indicates the model with the daily maximum may not be the better model.

4 Predicting hospital admission

We compare the current approach predictive accuracy with four other approaches of ozone exposure models. Two of these models involve a generalized additive model using daily average

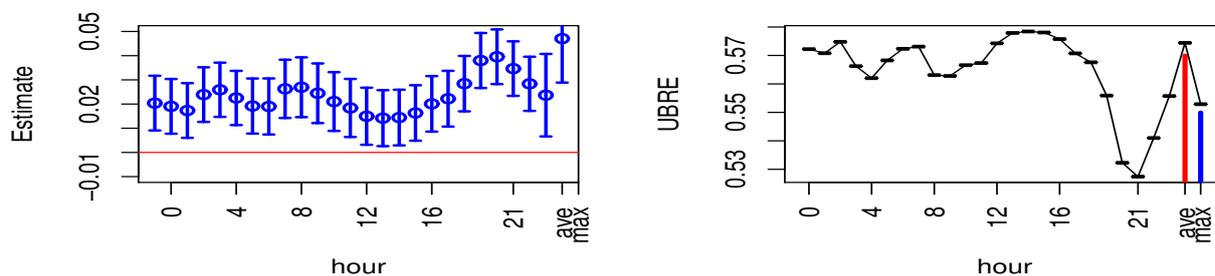


Figure 4: The estimated hourly ozone effect with associated 95% confidence interval (left) and the UBRE scores (right) under the standard GAM. Results for the daily average (ave) and maximum (max) are also displayed at the far right-hand side using vertical lines. Data for the hourly measures and the daily summaries are scaled appropriately prior to model fitting.

Ozone exposure model:						
	GAM-Average	GAM-Maximum	GAM-CP1	GAM-CP2	FGLM-O	FGLM-A
RMSE	33.391	33.392	28.415	28.419	19.034	19.011

Table 2: Out-of-sample RMSE for different ozone exposure models including the functional regressions approach.

and daily maximum ozone as main predictors by adjusting the same confounders used in 4. For comparison, we also included two alternative ozone exposure measure in [1]. These are (1) the difference between day time maximum hourly concentrations and threshold level and (2) night time average concentrations. They showed that their method capture the daily variability of ozone better than classical summary measures. We refer to these as GAM-CP1 and GAM-CP2 respectively. We consider FGLM based on the observed ozone curves (FGLM-O) and based on aligned curves of ozone (FGLM-A). The time period of the study in years 1996-2002 is divided in two groups. Data for years 1996, 1997, 1998 and 1999 were used as training set. The data for years 2000, 2001 and 2002 were used as validation or test set. The idea is to use the first four years data to predict the daily hospital admission for the latter three years and then to compute the out-of-sample residual mean squared error (RMSE). We report the prediction results in Table 2. In this context, the RMSE of a model indicates the performance of the model in predicting daily admission in the validation group using the training group. When the RMSE score is smaller, the model is better in predicting the response. Consequently, the functional regressions approach in Table 2 have higher predicting ability compared to the other approaches. Here, functional regression based on aligned curve (FGLM-A) has improved the forecasting accuracy slightly in terms of RMSE measure. Clearly, exposure model based on daily summary measures perform worse than other approaches.

Acknowledgement

The author would like to thank University of Padova for financial support, Francesco Pauli and Monica Chiogna for crucial advice, critical reading of the paper and useful suggestions.

Bibliography

- [1] Chiogna, M. and Pauli, F. (2011). *Modelling short-term effects of ozone on morbidity: an application to the city of Milano, Italy, 1995 - 2003*. Environmental and Ecological Statistics **18**, 169-184.
- [2] Dominici, F., McDermott, A., Zeger, S. L. and Samet, J. M. (2002). *On the use of generalized additive models in time-series studies of air pollution and health*. American journal of epidemiology **156**, 193-203.
- [3] Gasser, T. and Kneip, A. (1995). *Searching for structure in curve samples*. Journal of the American Statistical Association **90**, 1179-1188.
- [4] Goldberg, M. S., Burnett, R. T., Brook, J., Bailar, J. C., Valois, M. F., and Vincent, R. (2001). *Associations between daily cause-specific mortality and concentrations of ground-level ozone in Montreal, Quebec*. American journal of epidemiology **154**, 817-826.
- [5] Hastie, T. and Tibshirani, R. (1990). *Generalized additive models*. Statistical Science **1**, 297-318.
- [6] Levy, J. I., Chemerynski, S. M. and Sarnat, J. A. (2005). *Ozone exposure and mortality: an empiric bayes metaregression analysis*. Epidemiology **16**, 458-468.
- [7] Marx, B.D and Eilers,P.H. (1996). *Flexible smoothing with B-splines and penalties*. Statistical Science **11**, 89-121.
- [8] McCullagh, P. and Nelder, J.A (1989). *Generalized Linear Models*. New York: Chapman and Hall.
- [9] McLean, M. W., Hooker, G., Staicu, A. M., Scheipl, F. & Ruppert, D. (2012). *Functional generalized additive models*. Journal of Computational and Graphical Statistics, (just-accepted).
- [10] O’Sullivan, F. (1986). *A statistical perspective on ill-posed inverse problems*. Statistical science, 502-518.
- [11] Ramsay, J.O. and Silverman, B.W. (2005). *Functional Data Analysis*. New York: Springer.
- [12] Wahba, G., Golub, G. H. and Heath, M. (1979). *Generalized cross-validation as a method for choosing a good ridge parameter*. Technometrics **21**, 215-223.
- [13] Wong, C. M., Ma, S., Hedley, A. J., and Lam, T. H. (2001). *Effect of air pollution on daily mortality in Hong Kong*. Environmental health perspectives **109**, 335.
- [14] Wood,S.N. (2006). *Generalized additive Models An Introduction with R*. USA: Chapman & Hall/CRC.

Estimation based on covariances from multiple one-step randomly delayed measurements with noise correlation

Raquel Caballero-Águila, *Universidad de Jaén*, raguila@ujaen.es
Aurora Hermoso-Carazo, *Universidad de Granada*, ahermoso@ugr.es
Josefa Linares-Pérez, *Universidad de Granada*, jlinares@ugr.es

Abstract. This paper is concerned with the recursive optimal least-squares linear estimation problem for a class of discrete-time linear stochastic systems with measured outputs perturbed by autocorrelated and cross-correlated noises. It is assumed that the multiple measurements are subject to one-step delays with different delay rates, and the measurement delay phenomenon occurs randomly. Under these assumptions and by an innovation approach, recursive algorithms with a simple structure and easily implementable are obtained for the prediction, filtering and fixed-point smoothing problems. It is assumed that the signal evolution model is unknown and the recursive estimation algorithms are derived requiring only information about the mean and covariance functions of the processes involved in the observation model, as well as the knowledge of the delay probabilities. A simulation example is shown to illustrate the effectiveness of the proposed algorithms.

Keywords. Least-squares estimation, innovation approach, correlated noises, delayed measurements.

1 Introduction

In the past decades, the development of network technologies has promoted the study of the estimation problem in multi-sensor systems, where the imperfection of the communication channels usually causes random sensor delays and/or multiple packet dropouts during the transmission process. Standard observation models are not appropriate to describe these random uncertainties, and classical estimation algorithms cannot be applied directly. Several modifications of the standard estimation algorithms have been proposed to incorporate the effects of randomly

delayed measurements (see [2] and [5], among others). Although in these papers all sensors are assumed to have the same delay characteristics, this assumption has been generalized considering multiple delayed sensors with different delay characteristics (see, for example, [1] and [3]).

On the other hand, the assumption of independent white noise processes can be restrictive in many real-world problems where correlation and cross-correlation of the noises may be present. For this reason, the estimation problem in systems with correlated and cross-correlated noises, is becoming an active research topic (see [4] and references therein).

Motivated by the above analysis, this paper addresses the signal estimation problem from multiple one-step randomly delayed measurements with different delay characteristics, under the assumption that the measured outputs are perturbed by one-step autocorrelated and cross-correlated observation noises.

The main contributions of this paper are: (a) unlike [1], where the measured outputs are perturbed by white observation noises, the current model includes the possibility of simultaneous one-step auto-correlated and cross-correlated observation noises in the measured outputs and random delays after transmission; (b) the proposed estimators are globally optimal in the linear least-squares sense and the algorithms structure is recursive and suitable for online applications. As in [1], the proposed algorithms are obtained without requiring full knowledge of the state-space model generating the signal process and by applying an innovation approach. However, the uncorrelation assumption of the observation noises in [1] guarantees that the noise estimators are zero, while this is not true for the problem at hand, where the noise estimators must be taken into account in the innovation process derivation, due to the correlation assumptions of the noise processes.

The rest of the paper is organized as follows. In Section 2, we present the delayed measurement model to be considered and the assumptions and properties under which the LS linear estimation problem is addressed. Recursive filtering and fixed-point smoothing algorithms are derived in Section 3 and the performance of the proposed algorithms is illustrated by a numerical simulation example in Section 4.

Notation: The notation used throughout the paper is standard. A^T and A^{-1} denote the transpose and inverse of a matrix A , respectively. The shorthand $Diag(a_1, \dots, a_m)$ denotes a diagonal matrix whose diagonal entries are a_1, \dots, a_m . $\mathbf{1} = (1, \dots, 1)^T$ denotes the all-ones vector and I the identity matrix. If the dimensions of matrices are not explicitly stated, they are assumed to be compatible for algebraic operations. The notation \circ denotes the Hadamard product ($[C \circ D]_{ij} = C_{ij}D_{ij}$) and δ_{k-s} represents the Kronecker delta function.

2 Problem formulation

The optimal least-squares (LS) linear estimation problem of a discrete-time signal z_k using multiple one-step randomly delayed measurements, with correlated noises, is addressed under the assumption that the evolution model of the signal is unknown and only information about its mean and covariance functions is available; specifically:

Assumption 1: The n -dimensional signal process $\{z_k; k \geq 1\}$ has zero mean and its autocovariance function is expressed in a separable form, $E[z_k z_j^T] = A_k B_j^T$, $j \leq k$, where A and B are known $n \times M$ matrix functions.

Consider the following observation model with one-step random delays:

$$\begin{aligned} \tilde{y}_k &= H_k z_k + \tilde{v}_k, \quad k \geq 1, \\ y_k &= (I - \Gamma_k) \tilde{y}_k + \Gamma_k \tilde{y}_{k-1}, \quad k > 1; \quad y_1 = \tilde{y}_1, \end{aligned} \tag{1}$$

where $\{\tilde{y}_k; k \geq 1\}$ is the m -dimensional measured output process; H_k for $k \geq 1$ are known matrices, $\{\tilde{v}_k; k \geq 1\}$ is the measurement noise, and $\Gamma_k = \text{Diag}(\gamma_k^1, \dots, \gamma_k^m)$, where $\gamma_k^i, i = 1, \dots, m$, are Bernoulli random variables. The following assumptions are established:

Assumption 2: $\{\tilde{v}_k; k \geq 1\}$ is a zero-mean process with $\text{Cov}[\tilde{v}_k, \tilde{v}_s] = \tilde{R}_{k,k} \delta_{k-s} + \tilde{R}_{k,s} \delta_{k-1-s} + \tilde{R}_{k,s} \delta_{k+1-s}$.

Assumption 3: For $i = 1, 2, \dots, m$, the process $\{\gamma_k^i; k > 1\}$ is a sequence of independent Bernoulli random variables with known probabilities $P[\gamma_k^i = 1] = p_k^i, \forall k > 1$. For $i, j = 1, 2, \dots, m$, the variables γ_k^i and γ_s^j are independent for $k \neq s$, and $\text{Cov}[\gamma_k^i, \gamma_k^j]$ are known.

Assumption 4: The signal process, $\{z_k; k \geq 1\}$, and the processes $\{\tilde{v}_k; k \geq 1\}$ and $\{\Gamma_k; k > 1\}$ are mutually independent.

The above assumptions lead to the following properties:

- The random matrices $\{\Gamma_k; k > 1\}$ are independent or, equivalently, the m -dimensional process $\{\gamma_k; k > 1\}$, where $\gamma_k = (\gamma_k^1, \dots, \gamma_k^m)^T$, is a white sequence. The first and second-order moments of these processes are known, and the following notation will be used:

$$\begin{aligned} \bar{\Gamma}_k &\equiv E[\Gamma_k] = \text{Diag}(p_k^1, \dots, p_k^m) \\ K_k^\gamma &\equiv E[\gamma_k \gamma_k^T], \quad K_k^{1-\gamma} \equiv E[(\mathbf{1} - \gamma_k)(\mathbf{1} - \gamma_k)^T], \quad K_k^{\gamma, 1-\gamma} \equiv E[\gamma_k(\mathbf{1} - \gamma_k)^T]. \end{aligned}$$

- If G is a $m \times m$ random matrix independent of $\{\Gamma_k, k > 1\}$, the Hadamard product properties lead to

$$E[\Gamma_k G \Gamma_k] = K_k^\gamma \circ E[G], \quad E[\Gamma_k G (I - \Gamma_k)] = K_k^{\gamma, 1-\gamma} \circ E[G], \quad E[(I - \Gamma_k) G (I - \Gamma_k)] = K_k^{1-\gamma} \circ E[G].$$

To simplify future expressions, the observation model (1) will be written equivalently as

$$\begin{aligned} y_k &= x_k + v_k, \quad k \geq 1, \\ x_k &= (I - \Gamma_k) H_k z_k + \Gamma_k H_{k-1} z_{k-1}, \quad k \geq 2; \quad x_1 = H_1 z_1, \\ v_k &= (I - \Gamma_k) \tilde{v}_k + \Gamma_k \tilde{v}_{k-1}, \quad k \geq 2; \quad v_1 = \tilde{v}_1. \end{aligned} \tag{2}$$

From the assumptions and properties of the observation model, the following first and second order properties of the processes $\{x_k; k \geq 1\}$ and $\{v_k; k \geq 1\}$ and, consequently, those of the observation process, $\{y_k; k \geq 1\}$, are easily obtained.

(I) The process $\{x_k; k \geq 1\}$ has zero mean and $K_k^x \equiv \text{Cov}[x_k, x_k]$ is given by

$$\begin{aligned} K_k^x &= K_k^{1-\gamma} \circ (H_k A_k B_k^T H_k^T) + K_k^{1-\gamma, \gamma} \circ (H_k A_k B_{k-1}^T H_{k-1}^T) \\ &\quad + K_k^{\gamma, 1-\gamma} \circ (H_{k-1} B_{k-1} A_k^T H_k^T) + K_k^\gamma \circ (H_{k-1} A_{k-1} B_{k-1}^T H_{k-1}^T), \quad k \geq 2; \\ K_1^x &= H_1 A_1 B_1^T H_1^T. \end{aligned} \tag{3}$$

(II) The process $\{v_k; k \geq 1\}$ has zero mean and

$$\text{Cov}[v_k, v_s] = R_{k,k}\delta_{k-s} + R_{k,s}\delta_{k-1-s} + R_{k,s}\delta_{k-2-s}, \quad s \leq k,$$

where

$$\begin{aligned} R_{k,k} &= K_k^{1-\gamma} \circ \widetilde{R}_{k,k} + K_k^{1-\gamma,\gamma} \circ \widetilde{R}_{k,k-1} + K_k^{\gamma,1-\gamma} \circ \widetilde{R}_{k-1,k} + K_k^\gamma \circ \widetilde{R}_{k-1,k-1}, \quad k > 1; \\ R_{k,k-1} &= (I - \overline{\Gamma}_k) \widetilde{R}_{k,k-1} (I - \overline{\Gamma}_{k-1}) + \overline{\Gamma}_k \widetilde{R}_{k-1,k-1} (I - \overline{\Gamma}_{k-1}) + \overline{\Gamma}_k \widetilde{R}_{k-1,k-2} \overline{\Gamma}_{k-1}, \quad k > 2; \\ R_{k,k-2} &= \overline{\Gamma}_k \widetilde{R}_{k-1,k-2} (I - \overline{\Gamma}_{k-2}), \quad k > 3; \\ R_{1,1} &= \widetilde{R}_{1,1}, \quad R_{2,1} = (I - \overline{\Gamma}_2) \widetilde{R}_{2,1} + \overline{\Gamma}_2 \widetilde{R}_{1,1}, \quad R_{3,1} = \overline{\Gamma}_3 \widetilde{R}_{2,1}. \end{aligned} \quad (4)$$

(III) The processes $\{x_k; k \geq 1\}$ and $\{v_k; k \geq 1\}$ are uncorrelated and, consequently,

$$K_k^y \equiv \text{Cov}[y_k, y_k] = K_k^x + R_{k,k}, \quad k \geq 1, \quad (5)$$

with K_k^x and $R_{k,k}$ given in (3) and (4), respectively.

3 LS linear estimation problem

To obtain a recursive algorithm for the LS linear estimator, $\widehat{z}_{k/L}$, of the signal, z_k , based on the randomly delayed observations $\{y_1, \dots, y_L\}$, an innovation approach has been used. In this section, recursive algorithms for the signal LS linear predictor ($L < k$), filter ($L = k$) and fixed-point smoother ($L > k$) are presented.

Prediction and filtering recursive algorithm *The signal predictors $\widehat{z}_{k/L}$, $L < k$, and the signal filter, $\widehat{z}_{k/k}$, are obtained as*

$$\widehat{z}_{k/L} = A_k O_L, \quad L < k; \quad \widehat{z}_{k/k} = A_k O_k, \quad (6)$$

where the vectors O_k are recursively calculated from

$$O_k = O_{k-1} + J_k \Pi_k^{-1} \mu_k, \quad k \geq 1; \quad O_0 = 0.$$

The matrix function J is given by

$$\begin{aligned} J_k &= G_{B_k}^T - r_{k-2} G_{A_k}^T - J_{k-2} \Pi_{k-2}^{-1} R_{k,k-2}^T - J_{k-1} \Pi_{k-1}^{-1} T_{k,k-1}^T, \quad k > 2; \\ J_2 &= G_{B_2}^T - J_1 \Pi_1^{-1} T_{2,1}^T, \quad J_1 = B_1^T H_1^T \end{aligned}$$

with $r_k = E[O_k O_k^T]$ recursively obtained from

$$r_k = r_{k-1} + J_k \Pi_k^{-1} J_k^T, \quad k \geq 1; \quad r_0 = 0.$$

The innovation, μ_k , satisfies

$$\begin{aligned} \mu_k &= y_k - G_{A_k} O_{k-2} - R_{k,k-2} \Pi_{k-2}^{-1} \mu_{k-2} - T_{k,k-1} \Pi_{k-1}^{-1} \mu_{k-1}, \quad k > 2; \\ \mu_2 &= y_2 - T_{2,1} \Pi_1^{-1} \mu_1, \quad \mu_1 = y_1, \end{aligned}$$

where $T_{k,k-1} = E[y_k \mu_{k-1}^T]$ is recursively obtained from

$$\begin{aligned} T_{k,k-1} &= G_{A_k} J_{k-1} + R_{k,k-1} - R_{k,k-2} \Pi_{k-2}^{-1} T_{k-1,k-2}^T, \quad k > 2; \\ T_{2,1} &= G_{A_2} B_1^T H_1^T + R_{2,1}. \end{aligned}$$

The innovation covariance matrix, Π_k , is given by

$$\begin{aligned} \Pi_k &= K_k^y - G_{A_k} r_{k-2} G_{A_k}^T - R_{k,k-2} \Pi_{k-2}^{-1} R_{k,k-2}^T - T_{k,k-1} \Pi_{k-1}^{-1} T_{k,k-1}^T \\ &\quad - G_{A_k} J_{k-2} \Pi_{k-2}^{-1} R_{k,k-2}^T - R_{k,k-2} \Pi_{k-2}^{-1} J_{k-2}^T G_{A_k}^T, \quad k > 2; \\ \Pi_2 &= K_2^y - T_{2,1} \Pi_1^{-1} T_{2,1}^T, \quad \Pi_1 = K_1^y. \end{aligned}$$

The matrices K_k^y and $R_{k,s}$, for $s = k, k-1, k-2$, are given in (5) and (4), respectively. Finally, the matrices G_{A_k} and G_{B_k} are defined by

$$G_{\Psi_k} = (I - \bar{\Gamma}_k) H_k \Psi_k + \bar{\Gamma}_k H_{k-1} \Psi_{k-1}, \quad \Psi = A, B.$$

The performance of the LS estimators $\hat{z}_{k/L}$, $L \leq k$, is measured by the covariance matrices of the estimation errors, $\Sigma_{k/L} = E[(z_k - \hat{z}_{k/L})(z_k - \hat{z}_{k/L})^T]$. Using expression (6), and taking into account that $r_L = E[O_L O_L^T]$, we obtain the following expressions

$$\Sigma_{k/L} = A_k [B_k^T - r_L A_k^T], \quad L \leq k.$$

Note that the computation of the prediction and filtering error covariance matrices does not depend on the current set of observations, as it only needs the matrices A_k and B_k , which are known, and the matrices r_L , which are recursively calculated from (3); hence, the error covariance matrices provide a measure of the estimator performance even before we get any observed data.

Fixed-point smoothing algorithm. The fixed-point smoother $\hat{z}_{k/L}$, $L > k$, is calculated as

$$\hat{z}_{k/L} = \hat{z}_{k/L-1} + S_{k,L} \Pi_L^{-1} \mu_L, \quad L > k,$$

with initial condition given by the filter, $\hat{z}_{k/k}$, and

$$\begin{aligned} S_{k,L} &= [B_k - E_{k,L-2}] G_{A_L}^T - S_{k,L-2} \Pi_{L-2}^{-1} R_{L,L-2}^T - S_{k,L-1} \Pi_{L-1}^{-1} T_{L,L-1}^T, \quad L > k, \quad (L > 2) \\ S_{1,2} &= B_1 G_{A_2}^T - S_{1,1} \Pi_1^{-1} T_{2,1}^T \end{aligned}$$

with $S_{k,k-1} = A_k J_{k-1}$ and $S_{k,k} = A_k J_k$. The matrices $E_{k,L}$ satisfy the recursive formula

$$E_{k,L} = E_{k,L-1} + S_{k,L} \Pi_L^{-1} J_L^T, \quad L > k; \quad E_{k,k-1} = A_k r_{k-1}, \quad E_{k,k} = A_k r_k.$$

The filter $\hat{z}_{k/k}$, the matrices G_{A_L} , $T_{L,L-1}$ and J_L , the innovations ν_L and their covariance matrices Π_L are obtained from the linear filtering algorithm.

Using the recursive formula of the fixed-point smoother, the following expression for the fixed-point smoothing error covariance matrices, $\Sigma_{k/L} = E[(z_k - \hat{z}_{k/L})(z_k - \hat{z}_{k/L})^T]$, $L > k$, is immediately deduced

$$\Sigma_{k/L} = \Sigma_{k/L-1} - S_{k,L} \Pi_L^{-1} S_{k/L}^T, \quad L > k,$$

with the filtering error covariance matrix, $\Sigma_{k/k}$, as initial condition.

Remark. Further research topics include the extension of our results to more general observation models. For example, a similar study to that performed in this paper would allow us to extend the current results to the case of randomly delayed measurements correlated at consecutive sampling times. This extension would cover some real-world problems, such as transmission models with stand-by sensors, where two successive observations cannot be delayed.

4 Numerical simulation example

Let $\{z_k; k \geq 1\}$ be a zero-mean scalar signal with autocovariance function $E[z_k z_j] = 1.025641 \times 0.95^{k-j}$, $j \leq k$, which is factorizable according to Assumption 1 just taking, for example, $A_k = 1.025641 \times 0.95^k$ and $B_k = 0.95^{-k}$. For the simulations, the signal is assumed to be generated by an autoregressive model, $z_{k+1} = 0.95z_k + w_k$, where $\{w_k; k \geq 1\}$ is a zero-mean white Gaussian noise with variance 0.1, for all k .

Consider measurements coming from two sensors,

$$\tilde{y}_k = \begin{pmatrix} \tilde{y}_k^1 \\ \tilde{y}_k^2 \end{pmatrix} = \begin{pmatrix} 1.5 \\ 0.5 \end{pmatrix} z_k + \begin{pmatrix} \tilde{v}_k^1 \\ \tilde{v}_k^2 \end{pmatrix}, \quad k \geq 1$$

where $\{\tilde{v}_k^i; k \geq 1\}$, $i = 1, 2$, are defined by $\tilde{v}_k^i = c_i(\eta_k + \eta_{k+1})$, $i = 1, 2$, with $c_1 = 2$, $c_2 = 1.5$ and $\{\eta_k; k \geq 1\}$ a zero-mean Gaussian white process with variance 0.5.

Now, according to the proposed observation model, it is assumed that, at any sampling time $k > 1$, the measured output from the i -th sensor, \tilde{y}_k^i , can be randomly delayed by one sampling period during network transmission; that is,

$$y_k^i = (1 - \gamma_k^i)\tilde{y}_k^i + \gamma_k^i\tilde{y}_{k-1}^i, \quad k > 1; \quad y_1^i = \tilde{y}_1^i, \quad i = 1, 2,$$

where $\{\gamma_k^i; k > 1\}$, $i = 1, 2$, are independent sequences of independent Bernoulli random variables with $P[\gamma_k^1 = 1] = p^1$, $k > 1$.

Firstly, to compare the performance of the predictor, $\hat{z}_{k/k-1}$, filter, $\hat{z}_{k/k}$, and fixed-point smoothers, $\hat{z}_{k/L}$, with $L = k + 1$, $k + 2$, $k + 3$, the corresponding error variances are calculated considering constant delay probabilities, $p^1 = 0.3$ and $p^2 = 0.4$. The results are displayed in Figure 1 which shows that the error variances corresponding to the fixed-point smoother are less than those of the filter and the filtering error variances are smaller than the prediction ones, thus confirming that the smoother has the best performance while the predictor has the worst performance. This figure also shows that the performance of the fixed-point smoothers improves as the number of available observations increases. Analogous results are obtained for other values of the probabilities p^i , $i = 1, 2$.

Next, we study the filtering error variances, $\Sigma_{k/k}$, when the delay probabilities p^1 and p^2 are varied from 0.1 to 0.9. In all the cases we have examined, the filtering error variances present insignificant variation from the 10th iteration onwards and, consequently, only the error variances at a specific iteration are shown here. Figure 2(a) displays the filtering error variances at $k = 50$ versus p^1 (for constant values of $p^2 = 0.1$ to 0.5) and Figure 2(b) shows these variances versus p^2 (for constant values of $p^1 = 0.1$ to 0.5).

From these figures it is concluded that the performance of the filter improves as the delay probabilities, p^i , $i = 1, 2$, decrease. Consequently, more accurate estimations are obtained as p^i comes nearer to 0, case in which all the observations arrive on time.

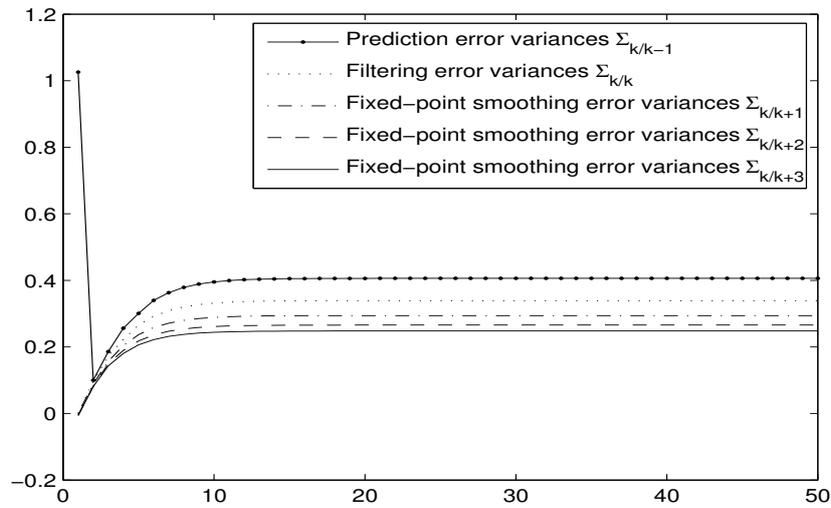


Figure 1: Prediction, filtering and smoothing error variances, when $p^1 = 0.3$ and $p^2 = 0.4$.

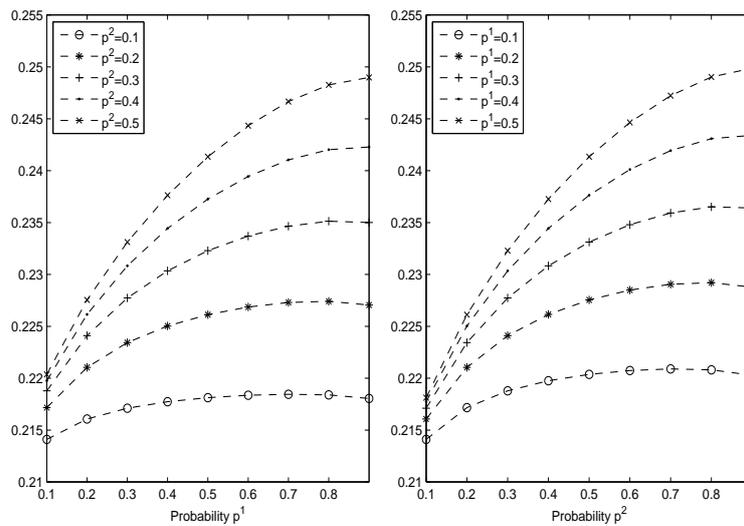


Figure 2: (a) Filtering error variances, $\Sigma_{50/50}$, versus p^1 (for constant values of $p^2 = 0.1$ to 0.5). (b) Filtering error variances $\Sigma_{50/50}$ versus p^2 (for constant values of $p^1 = 0.1$ to 0.5).

Funding

This research is funded by Ministerio de Educación y Ciencia (grant No. MTM2011-24718).

Bibliography

- [1] Caballero-Águila, R. Hermoso-Carazo, A. and Linares-Pérez, J. (2013) *Linear estimation based on covariances for networked systems featuring sensor correlated random delays*. International Journal of Systems Science, **44**, 1233–1244.
- [2] Hermoso-Carazo, A. and Linares-Pérez, J. (2008) *Linear and quadratic least-squares estimation using measurements with correlated one-step random delay*. Digital Signal Processing, **18**, 450–464.
- [3] Hounkpevi, F. O. and Yaz, E. E. (2007) *Minimum variance generalized state estimators for multiple sensors with different delay rates*. Signal Processing, **87**, 602–613.
- [4] Hu, J., Wang, Z. and Gao, H. (2013) *Recursive filtering with random parameter matrices, multiple fading measurements and correlated noises*. Automatica, **49**, 3440–3448.
- [5] Yang, Y. H., Fu, M. Y. and Zhang, H. S. (2013) *State estimation subject to random communication delays*. Acta Automatica Sinica, **39**, 237–243.

Density and Distribution Function estimation through iterates of fractional Bernstein Operators

Claude Manté, *Aix-Marseille Université, Université du Sud Toulon-Var, CNRS/INSU, IRD, MIO, UM 110, Campus de Luminy Marseille, France, claude.mante@mio.osupytheas.fr*

Abstract. We describe a method for distribution function and density estimation with Bernstein polynomials. We take advantage of results about the eigenstructure of the Bernstein operator to refine the Sevy's convergence acceleration method, based on iterates of this operator; the original Sevy's algorithm is improved by introducing fractional operators. The proposed algorithm has better convergence properties than the classical one; the price to pay is a controllable loss of the shape-preserving properties of the Bernstein approximation (monotonicity and positivity in the Density Estimation setting). The method is tested on simulated data.

Keywords. Density Estimation, Bernstein operator, root of operators, Bernstein polynomials, Lagrange polynomials

1 Introduction

Bernstein simultaneously introduced in 1912 the polynomials and the operator that bear his name in a famous paper [2]. But, as Farouki [8] noticed, this approximation has been seldom used, due to its slow convergence. For instance, to approach $f(t) = t^2$ on the unit interval with a maximal error of 10^{-4} , we need a polynomial of degree 2500 [8] ! Nevertheless, this operator (denoted B_n) has attractive shape-preserving properties: if f is positive (or monotone, or convex), its image $B_n[f]$ is so (see [5] for further properties). Consequently, the structure of a distribution function (**d.f.**) is preserved by B_n ; this point strongly motivated the use of the Bernstein approximation in Density Estimation [19, 1, 3, 12, 13, 14].

Notations

We will work in the Banach space $C[0, 1]$ of continuous functions on $[0, 1]$, equipped with the norm $\|f\| := \max_{x \in [0, 1]} |f(x)|$. The subspace of polynomials of degree $\leq n$ will be denoted \mathfrak{P}_n , and

$\overline{\mathfrak{P}}_n$ will be the supplementary of \mathfrak{P}_1 in \mathfrak{P}_n . We will denote \mathcal{F}^+ the closed convex cone of positive functions of $C[0, 1]$, and \mathcal{F}^1 the closed convex set of functions of $C[0, 1]$ integrating to 1.

Consider an operator $U : C[0, 1] \rightarrow C[0, 1]$; for $n \geq 2$ (fixed), its restriction to \mathfrak{P}_n will be denoted $\overset{\circ}{U}$, and its restriction to $\overline{\mathfrak{P}}_n$ will be denoted \overline{U} . For the sake of simplicity, the restrictions of the identity operator to these subspaces will be denoted 1 instead of $\overset{\circ}{1}$ or $\overline{1}$; $Mat(U; B_1, B_2)$ will denote the matrix representation of U with respect to the bases B_1 and B_2 of $C[0, 1]$. We will use the matrix p -norms $\|U\|_p := \sup_{v \neq 0} \frac{\|U(v)\|_p}{\|v\|_p}$ where $\|v\|_p$ is the usual vector ℓ^p -norm. Notice that $\|U\|_1 := \max_{1 \leq k \leq dim(U)} \sum_{j=1}^{dim(U)} |U_{j,k}|$, $\|U\|_2$ is the spectral norm, and $\|U\|_\infty := \max_{1 \leq j \leq dim(U)} \sum_{k=1}^{dim(U)} |U_{j,k}|$ (see [7]).

2 Expression of powers of the Bernstein operator into different bases

The Bernstein operator $B_n : C[0, 1] \rightarrow C[0, 1]$ is defined by [4, 15, 18]:

$$B_n[f](x) := \sum_{j=0}^n w_{n,j}(x) f\left(\frac{j}{n}\right)$$

with $w_{n,j}(x) := \binom{n}{j} x^j (1-x)^{n-j}$; its range $\mathcal{R}(B_n) \subseteq \mathfrak{P}_n$. Cooper and Waldron [4] gave its spectral decomposition, which can be also written into the form hereunder [16].

Theorem 2.1. *The Bernstein operator can be represented in the diagonal form*

$$B_n[f] = \sum_{j=0}^n \lambda_j^{[n]} \pi_j^{[n]} \otimes \pi_j^{*[n]} (\mathcal{L}_n[f])$$

where $f \in C[0, 1]$, $\lambda_j^{[n]} = \frac{n!}{(n-j)!n^j} \in]0, 1]$ and $\pi_j^{[n]} \in \mathfrak{P}_n$ are its eigenvalues and eigenvectors, $\pi_j^{*[n]}$ is the dual vector of $\pi_j^{[n]}$, and $u \otimes v^*(w) := u \langle v^*, w \rangle$.

We will need the Lagrange interpolation operator (**equispaced case**) $\mathcal{L}_n : C[0, 1] \rightarrow C[0, 1]$ defined by: $\mathcal{L}_n[f](x) := \sum_{j=0}^n \ell_{n,j}(x) f\left(\frac{j}{n}\right)$, with

$$\ell_{n,j}(x) := \prod_{\substack{k=0 \\ k \neq j}}^n \frac{nx - k}{j - k}.$$

Three bases of \mathfrak{P}_n will be needed:

1. the Bernstein's basis $W_n := \{w_{n,j}(x), 0 \leq j \leq n\}$
2. the Lagrange's basis $L_n := \{\ell_{n,j}(x), 0 \leq j \leq n\}$
3. the eigenvectors of B_n , $\Pi_{[n]} := \{\pi_j^{[n]}(x), 0 \leq j \leq n\}$.

Let us denote $LW_{[n]}$ the transformation matrix associated with the bases L_n and W_n , whose j^{th} column consists in the coordinates of $w_{n,j}$ in the basis L_n . The following results can be easily demonstrated [16]:

Lemma 2.2. $Mat\left(\overset{\circ}{B}_n; L_n, W_n\right) = I_n$ and $Mat\left(\overset{\circ}{B}_n; W_n, W_n\right) = LW_{[n]}$.

Thank to this lemma, we obtain for any $k \geq 2$ a first matrix representation of B_n^k from the diagram:

$$B_n^k : C[0, 1] \xrightarrow{\mathcal{L}_n} (\mathfrak{P}_n, L_n) \xrightarrow{I_n} (\mathfrak{P}_n, W_n) \xrightarrow{LW_{[n]}^{k-1}} (\mathfrak{P}_n, W_n). \tag{1}$$

Besides, Theorem 2.1 gives an alternative representation of this operator:

$$B_n^k : C[0, 1] \xrightarrow{\mathcal{L}_n} (\mathfrak{P}_n, L_n) \xrightarrow{L\Pi_{[n]}} (\mathfrak{P}_n, \Pi_{[n]}) \xrightarrow{\Lambda_{[n]}^k} (\mathfrak{P}_n, \Pi_{[n]}) \xrightarrow{\Pi W_{[n]}} (\mathfrak{P}_n, W_n) \tag{2}$$

where $\Lambda_{[n]}$ is the diagonal matrix associated with the vector

$$(1, 1, 1 - 1/n, (3n - 2)/n^2), \dots, n!/n^n)$$

of eigenvalues of B_n , and $L\Pi_{[n]}$ and $\Pi W_{[n]}$ are transformation matrices associated with the three bases.

3 Sevy’s sequences for d.f. and density approximation

We saw that in the elementary case $f(t) = t^2$, the speed of convergence of $B_n[f]$ towards f is only $O\left[\frac{1}{n}\right]$ [8]; the situation is worse in the special case of d.f.s, since it can be proven [15] that one should rather expect $O\left[\frac{1}{\sqrt{n}}\right]$. To get a sequence of approximations converging faster than B_n , Sevy [17] proposed to supersede B_n by the iterated operator

$$\mathfrak{J}_n^I := (1 - (1 - B_n)^I) \tag{3}$$

and proved the following result.

Theorem 3.1. ([18], see also [4]) For $n \geq 1$ fixed, and any function F defined on $[0,1]$, we have:

$$\left\| \mathfrak{J}_n^I[F] - \mathcal{L}_n[F] \right\| \xrightarrow{I \rightarrow \infty} 0.$$

Such a sequence build a bridge between $\mathfrak{J}_n^1[F] = B_n[F]$ and $\mathcal{L}_n[F]$. It is worth noting that $\mathcal{L}_n[F]$ interpolates the data but can be very bumpy and that in the equispaced case, the interpolation errors are maximal ([6, Ch. 2]; [11, Ch. 5]). Suppose now F is a d.f.; $B_n[F]$ is also a d.f., but in general $\mathcal{L}_n[F]$ will not share the same characteristics. Thus, it is natural to try to determine some optimal number of iterations $I^* \geq 1$ in order that $\mathfrak{J}_n^{I^*}[F]$ has the structure of a d.f., while $\mathfrak{J}_n^{I^*+1}[F]$ has not. In other words, the density approximation $\widehat{f}_n^{(I^*)}(x) := \frac{d}{dx} \mathfrak{J}_n^{I^*}[F](x)$ should be *bona fide*, i.e. should belong to $\mathcal{F}^+ \cap \mathcal{F}^1$, while $\widehat{f}_n^{(I^*+1)} \notin \mathcal{F}^+ \cap \mathcal{F}^1$ (see [15]).

4 Interpolating Sevy’s sequences (see [16])

To refine Sevy’s sequences, we build for $K \geq 2$ the K^{th} “root” of the operator $G_n := (1 - B_n)$ involved in Formula 3. Because B_n only preserves \mathfrak{P}_1 , the eigenvalues of $\overline{G_n}$ belong to $]0, 1[$. Thus, thanks to classical results about convergent series of operators (see [10] for instance), one may consistently define the fractional operator

$$\overline{G_n}^{(1/K)} := \exp\left(\frac{1}{K} \log(\overline{G_n})\right). \tag{4}$$

One can easily verify the following lemma.

Lemma 4.1. $\forall I \geq 1,$

$$\mathfrak{J}_n^I = (1 - (1 - B_n)^I) = \left(1 - \left(1 - \overset{\circ}{B}_n\right)^I\right) \circ \mathcal{L}_n.$$

Consequently, we can proceed as if $f \in \mathcal{R}(\mathcal{L}_n)$ and don’t have to worry about the “Lagrange residual” $f - \mathcal{L}_n[f]$. Since \mathfrak{P}_1 is preserved by B_n and because of Lemma 4.1, $\mathfrak{J}_n^k(f) = \mathcal{L}_1[f] + \mathfrak{J}_n^k(\mathcal{L}_n[f] - \mathcal{L}_1[f])$, and we can set the definition of K-fractional Sevy’s sequences.

Definition 4.2. *Let $K \geq 2$, and $f \in C[0, 1]$. The K -fractional Sevy’s sequence of approximations of f is:*

$$\mathfrak{J}_{n;K}^j[f] := \mathcal{L}_1[f] + \left(1 - \overline{G_n}^{(j/K)}\right) (\mathcal{L}_n[f] - \mathcal{L}_1[f]), \quad j \geq 1.$$

Such a sequence interpolates the original one, since $\mathfrak{J}_{n;K}^{jK}[f] = \mathfrak{J}_n^j(f)$. Its matrix representation stems from diagram 2.

Lemma 4.3. *Mat $\left(\overset{\circ j}{\mathfrak{J}}_{n;K}; L_n, W_n\right) = \Pi W_{[n]} \circ \Lambda_{[n]}^{(j/K)} \circ L \Pi_{[n]}$, where $\Lambda_{[n]}^{(j/K)}$ is the diagonal matrix associated with the vector*

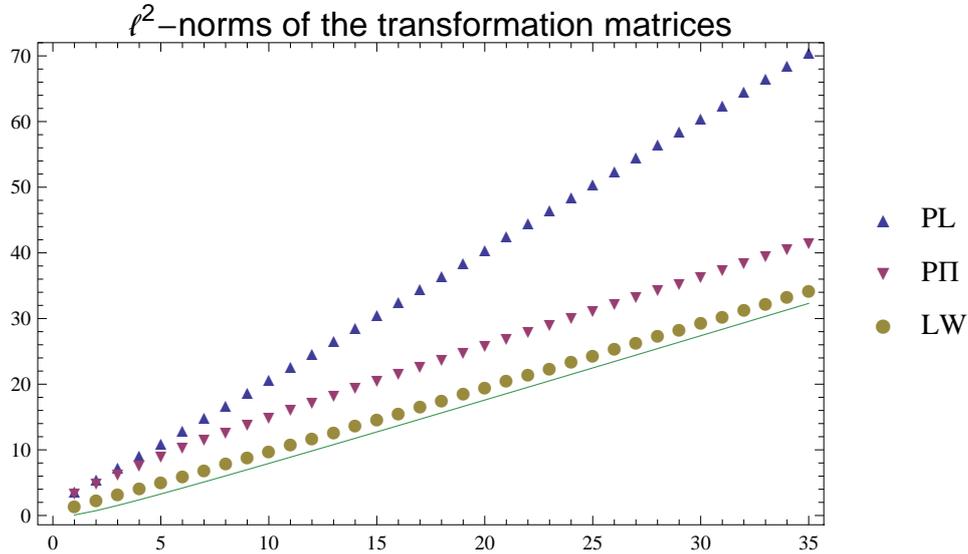
$$\left(1, 1, 1 - \left(\frac{1}{n}\right)^{(j/K)}, 1 - \left(\frac{3n - 2}{n^2}\right)^{(j/K)}, \dots, 1 - \left(1 - \frac{n!}{n^n}\right)^{(j/K)}\right).$$

5 Numerical issues

Because of Lemmas 2.2 and 4.1, building a classical Sevy’s sequence amounts to compute powers of the transformation matrix $LW_{[n]}$ (see diagram 1). The condition number of this matrix in the ℓ^2 -norm is [7]: $\frac{\|LW_{[n]}\|_2}{\|LW_{[n]}^{-1}\|_2} = \frac{\lambda_0^{[n]}}{\lambda_n^{[n]}} = \frac{n^n}{n!} \approx \frac{e^n}{\sqrt{2\pi n}}$ (asymptotically - see [9]). Thus, $LW_{[n]}$ is ill-conditioned, and one must expect to meet numerical problems when n is big. The situation is potentially worse for fractional sequences, since Lemma 4.3 shows that the matrix of the restricted operator depends on both the ill-conditioned transformation matrices $L \Pi_{[n]}$ and $\Pi W_{[n]}$ (see Figure 1).

But the point for us is merely to control numerical errors in computing $\mathfrak{J}_{n;K}^j[f]$! Notice that on the one hand $Mat\left(\overset{\circ}{B}_n; L_n, W_n\right) = I_n$ (Lemma 2.2), while on the other hand $Mat\left(\overset{\circ}{B}_n; L_n, W_n\right) = \Pi W_{[n]} \circ \Lambda_{[n]} \circ L \Pi_{[n]}$ (diagram 2). Consequently, the matrix norms

Figure 1: Logarithm of the condition numbers of the transformation matrices $PL_{[n]}$, $P\Pi_{[n]}$ and $LW_{[n]}$; the continuous line corresponds to the asymptotic value $n - \frac{1}{2}\text{Log}(2\pi n)$.



$$\left\| \begin{array}{l} \Pi W_{[n]} \circ \Lambda_{[n]} \circ L\Pi_{[n]} - I_n \\ \Pi W_{[n]} \circ \Lambda_{[n]} \circ L\Pi_{[n]} - I_n \end{array} \right\|_1 \quad (5)$$

are convenient indicators of loss of numerical accuracy imputable to the ill-conditioning of the transformation matrices. Since the only easy-to-handle basis is the power basis, the transformation matrices $PL_{[n]}$, $P\Pi_{[n]}$ and $PW_{[n]}$ are straightforwardly computed, and we can write:

$$\begin{aligned} \Pi W_{[n]} &= P\Pi_{[n]}^{-1} \circ PW_{[n]} \\ L\Pi_{[n]} &= PL_{[n]}^{-1} \circ P\Pi_{[n]} \end{aligned} \quad (6)$$

(formally). But we can derive from Figure 1 that these inverse matrices cannot be computed with sufficient accuracy in general. Thus, it's necessary to supersede in (6) the inverse matrices by the Moore-Penrose generalized inverses $P\Pi_{[n]}^+$ and $PL_{[n]}^+$. This gives rise to the regularized operators:

$$\begin{aligned} \widetilde{\Pi W}_{[n]} &:= P\Pi_{[n]}^+ \circ PW_{[n]} \\ \widetilde{L\Pi}_{[n]} &:= PL_{[n]}^+ \circ P\Pi_{[n]}. \end{aligned} \quad (7)$$

On Figure 2, we plotted the logarithm of the second indicator of Formula (5), for n ranging from 1 to 35 (a similar graph can be obtained for the first indicator). Two cases must be distinguished on this plot: the “symbolic” one, where polynomial eigenfunctions were computed from the recurrence formula given by [4], and the “numerical” one, where they were computed by polynomial interpolation of the eigenvectors of $LW_{[n]}$, giving rise to the alternative basis $\widehat{\Pi}_{[n]} := \{ \widehat{\pi}_j^{[n]}(x), 0 \leq j \leq n \}$. Of course, we should have $\widehat{\pi}_j^{[n]} = \pm \pi_j^{[n]} \forall 0 \leq j \leq n$ if there were not different roundoff errors on both sides, imputable to different algorithms! That is why we

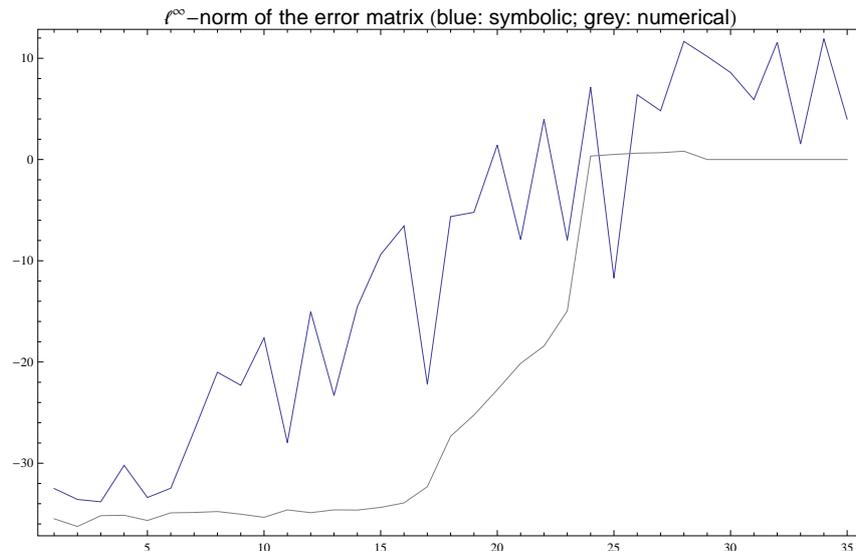


Figure 2: The sequences $\left\{ \text{Log} \left(\left\| \widetilde{\Pi} \widetilde{W}_{[n]} \circ \Lambda_{[n]}^{Tr} \circ \widetilde{L} \widetilde{\Pi}_{[n]} - I_n \right\|_{\infty} \right), 1 \leq n \leq 35 \right\}$ and $\left\{ \text{Log} \left(\left\| \widetilde{\mathbf{W}}_{[n]} \circ \Lambda_{[n]}^{Tr} \circ \widetilde{L} \widetilde{\Pi}_{[n]} - I_n \right\|_{\infty} \right), 1 \leq n \leq 35 \right\}$; $\Lambda_{[n]}^{Tr}$ is the diagonal matrix obtained by setting to zero each eigenvalue $< 10^{-12}$.

took into account the numerical rank of $\overset{\circ}{B}_n$, discarding from the computation of Formula (5) eigenvectors associated with eigenvalues smaller than 10^{-12} (see Figure 2 and its legend).

It is worth noting that the computational cost in the symbolic case is considerable: it took about 6600 seconds to produce the symbolic part of Figure 2, while the numeric part was obtained in 80 seconds.

6 Application to density and d.f. estimation

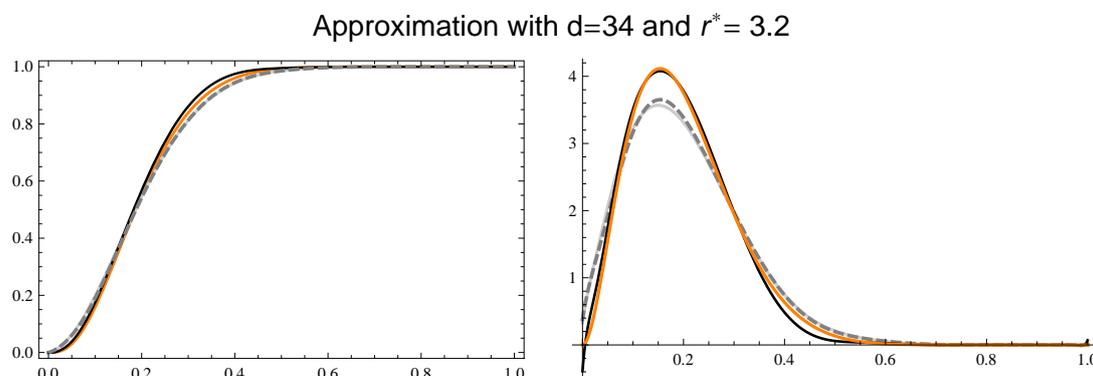
Suppose F is some differentiable d.f. associated with a random variable X defined on $[0, 1]$, and that $S_N := \{X_1, \dots, X_N\}$ is a N -sample of X , giving rise to the empirical d.f. $F_N(x)$. Babu *et al.* [1] proposed to estimate F by a Bernstein polynomial $\widetilde{F}_{N,m}$ of degree m :

$$\widetilde{F}_{N,m}(x) := \sum_{k=0}^m F_N\left(\frac{k}{m}\right) w_{m,k}(x) = B_m[F_N] \tag{8}$$

with $m \leq m_0 := \lceil N/\text{Log}(N) \rceil$. The proposed method consists in superseding $B_{m_0}[F_N]$ by some $\mathcal{J}_{m^*,K}^I[F_N]$, where $m^* \leq m_0$ and $I^* \geq K$ (fixed) are convenient values of the degree of the estimator and of the number of iterations in Definition 4.2.

As an illustration, we displayed first on Figure 3 the results obtained with a sample of size 200 of $\beta(3, 12)$, with $K = 10$. We found that $I^* = 32$ iterations of the fractional operator (4) simultaneously corresponded to a satisfactory fit of the e.d.f. and an approximately *bona fide* density estimation. Thus, in this case, the fractional number of iterations was $r^* = 1 + \frac{22}{10}$. On this plot, we superimposed to the true d.f. three estimators: the Babu's one, of degree $m_0 = 38$,

Figure 3: Estimation of the $\beta(3, 12)$ d.f. and density from a sample. Left panel: the true d.f. (orange), the Babu's one (gray and dashed, of degree $m_0 = \lceil 200/\text{Log}(200) \rceil = 38$), the classical Bernstein estimator of degree $m = 34$ (gray), and the proposed one (black), of degree 34 too. Right panel: density estimators obtained by deriving the d.f.s estimated.



the Bernstein estimator of degree $m = 34$, and the iterated estimator (black), of degree 34 too. The density estimators are derivatives of these d.f.s

In addition, we collected in Table 1 results from simulations carried on with 30 samples of size $N = 150$ ($\Rightarrow m_0 = 30$) of four Beta distributions. For sake of simplicity, we fixed $I^* = 20$ (see [16] for a theoretical justification). For each one of these samples and for each estimator (4 estimators of the d.f. and 3 estimators of the density, since the e.d.f. is not differentiable), the Integrated Squared Error (ISE) $\int (\hat{F}(x) - F(x))^2 dx$ and the L^1 error norm $\int |\hat{f}(x) - f(x)| dx$ were computed. Clearly, even in this suboptimal situation ($I^* = 20$), the proposed estimators outperformed classical ones, excepted in the very simple case $\beta(1, 2)$ (uniform distribution). Notice the honorable performances of the good old e.d.f.!

Table 1: Simulations results. First group of columns: the distribution simulated, and optimal value of m (for further details, see [16]); second group: median of $10^3 \cdot ISE$ of estimated distribution functions; third group: median of the L^1 error norms for estimated densities. Best result are in bold characters.

<i>Probability</i>	m^*	<i>e.d.f.</i>	B_{30}	B_{m^*}	$\mathfrak{J}_{m^*}^{20}$	B'_{30}	B'_{m^*}	$\mathfrak{Y}_{m^*}^{20}$
$\beta(1, 2)$	16	0.497	0.415	0.38	0.569	0.1	0.09	0.108
$\beta(2, 4)$	18	0.6	0.51	0.56	0.368	0.108	0.12	0.099
$\beta(3, 12)$	25	0.32	0.783	0.908	0.258	0.197	0.207	0.118
$\beta(10, 10)$	25	0.318	1.16	1.37	0.289	0.248	0.263	0.153

Bibliography

- [1] Babu, G. J., Canty, A. J. and Chaubey, Y. P. (2002) *Application of Bernstein polynomials for smooth estimation of a distribution and density function*. Journal of Statistical Planning and Inference, **105**, 377-392.

- [2] Bernstein, S. N. (1912) *Démonstration du théoreme de Weierstrass fondée sur le calcul des probabilités*. Commun. Soc. Math. Kharkov, **13**, 1-2.
- [3] Bouezmarni, T. and Rolin, J.M. (2007) *Bernstein estimator for unbounded density function*. Journal of Nonparametric Statistics, **19**, **3**, 145-161.
- [4] Cooper, S. and Waldron, S. (2000) *The eigenstructure of the Bernstein operator*. Journal of Approximation Theory, **105**, 133-165.
- [5] Davis, P. J. (1963) *Interpolation and approximation*. Blaisdell, New York
- [6] de Boor, C. (1978) *A practical guide to splines*. Applied Mathematical Sciences, 27, Springer-Verlag, New York.
- [7] Farouki, R. T. (1991) *On the stability of transformations between power and Bernstein polynomials forms*. Computer Aided Geometric Design, **8**, 29-36.
- [8] Farouki, R.T. (2012) *The Bernstein polynomial basis: a centennial retrospective*. Computer Aided Geometric Design, **29**, 379-419.
- [9] Impens, C. (2003) *Stirling's series made easy*. Amer. Math'l Monthly, **110**, 730-735.
- [10] Kato, T. (1995) *Perturbation theory for linear operators*. Springer-Verlag, Berlin, Heidelberg.
- [11] Laurent, P.-J. (1972) *Approximation et optimisation*. Enseignement des sciences, **13**, Hermann, Paris.
- [12] Leblanc, A. (2010) *A Bias-reduced approach to density estimation using Bernstein polynomials*. Journal of Nonparametric Statistics, **22**, **4**, 459-475.
- [13] Leblanc, A. (2012) *On estimating distribution functions using Bernstein polynomials*. Ann. Inst. Stat. Math., **64**, 919-943.
- [14] Leblanc, A. (2012) *On the boundary properties of Bernstein polynomial estimators of density and distribution functions*. Journal of Statistical Planning and Inference, **142**, 2762-2778.
- [15] Manté, C. (2012) *Application of iterated Bernstein operators to distribution function and density approximation*. Applied Mathematics and Computation, **218**, 9156-9168.
- [16] Manté, C. (in revision) *Iterated Bernstein operators for bona fide distribution function and density estimation. Balancing between the iterations number and the polynomial degree*.
- [17] Sevy, J. C. (1993) *Convergence of iterated boolean sums of simultaneous approximants*. Calcolo, **30**, 41-68.
- [18] Sevy, J. C. (1995) *Lagrange and least-squares polynomials as limits of linear combinations of iterates of Bernstein and Durrmeyer polynomials*. Journal of Approximation Theory, **80**, 267-271.
- [19] Vitale, R.A. (1975) *A Bernstein polynomial approach to density function estimation*. Statistical Inference and Related Topics, **2**, 87-99.

Bayesian cluster detection via adjacency modelling

Craig Anderson, *University of Glasgow*, c.anderson.3@research.gla.ac.uk

Duncan Lee, *University of Glasgow*, Duncan.Lee@glasgow.ac.uk

Nema Dean, *University of Glasgow*, nema.dean@glasgow.ac.uk

Abstract. The aim of disease mapping is to estimate the spatial pattern in disease risk across a set of areal units, in order to identify units which have elevated disease risk. Existing methods use Bayesian hierarchical models with spatially smooth conditional autoregressive priors to estimate disease risk, but these methods cannot identify the geographical extent of spatially contiguous high-risk clusters of areal units. We propose a two stage approach, which first produces a set of potential cluster structures for the data and then chooses the optimal structure by fitting an extended Bayesian hierarchical model. The first stage uses a hierarchical agglomerative clustering algorithm, spatially adjusted to account for the neighbourhood structure of the data. This algorithm is applied to data prior to the study period, and produces a set of n potential cluster structures. The second stage fits a Poisson log-linear model to the data, in which the optimal cluster structure and the spatial pattern in disease risk is estimated via a Markov Chain Monte Carlo (MCMC) algorithm. After assessing the methodology with a simulation study, it was applied to a study of respiratory disease risk in Glasgow, Scotland, where a number of high risk clusters were identified.

Keywords. Clustering, Conditional autoregressive model, Disease mapping

1 Introduction

Disease risk varies geographically as a result of many factors, including differences in environmental exposures, and cultural and behavioural differences between the inhabitants of different areas. Even within a city such as Glasgow, there are substantial inequalities in terms of health and disease risk, with poverty being one of the most important reasons for these differences. It is of interest to health agencies to be able to identify clusters of areas which have similar disease risks in order to make informed policy decisions. Many different approaches have been proposed for the identification of the spatial extent of high-risk clusters in spatial disease maps, including Bayesian hierarchical modelling [3], scan statistics [6] and point process methodology [4]. The first of these is typically based on a Poisson log-linear model, where covariates and/or a

set of random effects are used to represent the spatial disease risk pattern. The random effects are included to account for spatial correlation in the response that was not captured by the covariates; and are typically modelled by a conditional autoregressive (CAR) prior [2]. CAR priors make the naive assumption of global correlation between all pairs of random effects in geographically adjacent areal units, and therefore produce a spatially smooth risk surface. However, such smoothing is detrimental to our main aim, which is to identify groups of areas which have much higher (or lower) risks compared with surrounding areas, so an alternative approach is required.

Therefore, this paper outlines new methodology which allows for the estimation of the spatial pattern in disease risk, whilst simultaneously detecting the spatial extent of high or low risk clusters. In doing so, the cluster structure is accounted for when estimating disease risk, so that high risk clusters are not smoothed towards their geographical neighbours that do not exhibit elevated risks. The methodology brings together hierarchical agglomerative clustering techniques and conditional autoregressive models in a two-stage approach. The first stage is a spatially-adjusted hierarchical agglomerative clustering algorithm first proposed in [1], which respects the spatial contiguity of the study region. This algorithm is applied to disease data preceding the study period to elicit candidate cluster configurations. The second stage fits an extension of the Poisson log-linear model originally developed by [8] to the study data, where Markov Chain Monte Carlo (MCMC) simulation methods are used to estimate both the optimal cluster structure and disease risk.

2 Bayesian disease mapping

Disease maps allow us to graphically illustrate the differences within a geographical area. Such maps are produced by partitioning the study region into non-overlapping areal units such as electoral wards or census tracts, and then calculating the overall risk of disease for the population living in each areal unit.

Study Design and Modelling

The study region \mathcal{A} is partitioned into n non-overlapping areal units $\mathcal{A} = \{\mathcal{A}_1, \dots, \mathcal{A}_n\}$, and $\mathbf{Y} = (Y_1, \dots, Y_n)$ and $\mathbf{E} = (E_1, \dots, E_n)$ represent the observed and expected numbers of disease cases in each unit during the study period. The latter are constructed by external standardisation, based on the age and sex demographics of the population living in each areal unit. A Poisson log-linear model is commonly used to estimate disease risk, and a general form is given by

$$\begin{aligned} Y_i | E_i, R_i &\sim \text{Poisson}(E_i R_i) & i = 1, \dots, n, \\ \ln(R_i) &= \mathbf{x}_i^T \boldsymbol{\beta} + \phi_i. \end{aligned} \quad (1)$$

Here R_i represents disease risk in areal unit \mathcal{A}_i , and is modelled by a vector of covariates $\mathbf{x}_i^T = (1, x_{i1}, \dots, x_{ip})$, with coefficients $\boldsymbol{\beta} = (\beta_0, \dots, \beta_p)$, and a random effect ϕ_i . The random effects $\boldsymbol{\phi} = (\phi_1, \dots, \phi_n)$ account for the spatial autocorrelation induced into the disease data by factors such as unmeasured confounding, neighbourhood effects and grouping effects. They are modelled

by a conditional autoregressive (CAR) prior, which induces spatial correlation via a binary neighbourhood matrix W , where $w_{ij} = 1$ if areal units $(\mathcal{A}_i, \mathcal{A}_j)$ share a common border (denoted $i \sim j$) and $w_{ij} = 0$ otherwise. Note that $w_{ii} = 0$ for all i . CAR priors can be specified as a set of n univariate conditional distributions $f(\phi_i | \phi_{-i})$, where $\phi_{-i} = (\phi_1, \dots, \phi_{i-1}, \phi_{i+1}, \dots, \phi_n)$. The simplest of these CAR priors was the intrinsic prior proposed by [2], and this model is given by

$$\phi_i | \phi_{-i} \sim N \left(\frac{\sum_{j=1}^n w_{ij} \phi_j}{\sum_{j=1}^n w_{ij}}, \frac{1}{\tau^2 (\sum_{j=1}^n w_{ij})} \right) \quad i = 1, \dots, n, \quad (2)$$

where τ^2 is a conditional precision parameter. The conditional expectation of ϕ_i is the mean of the random effects in neighbouring areal units, while the variance is inversely proportional to the number of neighbouring units. This set of conditional distributions correspond to a multivariate Gaussian distribution, with mean zero but an improper precision matrix given by $Q = (\text{diag}(W\mathbf{1}) - W)$, where $W\mathbf{1}$ is a vector containing the number of neighbours for each areal unit.

3 Method

We propose a two-stage approach for estimating the spatial pattern in disease risk and identifying spatially contiguous clusters that exhibit either elevated or reduced disease risks. In the first stage we utilise the spatially adjusted hierarchical agglomerative clustering algorithm proposed by [1], and use it to elicit a set of candidate cluster configurations for the data. In the second stage we propose a hierarchical Bayesian model for the disease data which can simultaneously select the optimal cluster configuration from the candidates elicited in Stage 1 and also estimate disease risk.

Stage 1 - Eliciting cluster configurations using hierarchical agglomerative clustering

The method of clustering (for details see [5]) involves grouping together objects that are similar whilst separating those that are different, which is appropriate here because we wish to identify groups of areal units with similar disease risks. The clustering algorithm is taken from [1] and is applied to disease data preceding the study period, because it is likely to exhibit a similar spatial risk pattern to the study data unless substantial urban regeneration has taken place. Let $(\mathbf{Y}^{(1)}, \mathbf{E}^{(1)}), \dots, (\mathbf{Y}^{(q)}, \mathbf{E}^{(q)})$ denote the observed and expected disease counts for the q time intervals (usually years) preceding the study period. These earlier data are used to elicit a set of n potential cluster configurations for the study data, which are denoted here by $\{\mathcal{C}_1, \dots, \mathcal{C}_n\}$. Here $\mathcal{C}_k = \{\mathcal{C}_k(1), \dots, \mathcal{C}_k(k)\}$ partitions the n areal units $\mathcal{A} = \{\mathcal{A}_1, \dots, \mathcal{A}_n\}$ into k spatially contiguous groups, where $\mathcal{C}_k(j)$ is the j th cluster. The set of all possible spatially contiguous cluster configurations for the study region \mathcal{A} is very large, so we use this clustering step to vastly reduce the number of potential cluster structures to be considered in stage 2.

The data are clustered on the log standardised incidence ratio scale, that is $\ln(\mathbf{Y}^{(j)} / \mathbf{E}^{(j)})$, because this corresponds to the linear predictor scale in (1). Let $\boldsymbol{\psi} = [\ln(\mathbf{Y}^{(1)} / \mathbf{E}^{(1)}), \dots, \ln(\mathbf{Y}^{(q)} / \mathbf{E}^{(q)})]$ be the $n \times q$ matrix whose columns comprise $\ln(\mathbf{Y}^{(j)} / \mathbf{E}^{(j)})$ for $j = 1, \dots, q$, and denote the i th row by $\boldsymbol{\psi}_i = [\ln(Y_i^{(1)} / E_i^{(1)}), \dots, \ln(Y_i^{(q)} / E_i^{(q)})]$, the vector of q values for areal unit \mathcal{A}_i . The data

are clustered using a modified hierarchical agglomerative clustering algorithm, which initially considers each data point as its own singleton cluster, and then joins together the two least dissimilar clusters at each stage to form a larger cluster. This process is repeated until only one cluster containing all data points remains. For a configuration with k clusters the dissimilarity, d_{ij} , between clusters i ($\mathcal{C}_k(i)$) and j ($\mathcal{C}_k(j)$) can be measured by a number of metrics called linkage methods, but based on the results of [1] we will proceed with centroid linkage in this paper.

Centroid linkage measures the dissimilarity as the Euclidean distance between the average of the two clusters, that is $d_{ij} = \|\bar{\mathcal{C}}_k(i) - \bar{\mathcal{C}}_k(j)\|$, where $\bar{\mathcal{C}}_k(i) = (1/n_i) \sum_{f: \mathcal{A}_f \in \mathcal{C}_k(i)} \psi_f$, and n_i is the number of areal units in cluster $\mathcal{C}_k(i)$. The hierarchical agglomerative clustering algorithm described above is extended so that it produces spatially contiguous clusters, which is achieved by only allowing clusters containing two areal units which share a common border to be merged at each step. The algorithm produces a set of candidate cluster structures $\{\mathcal{C}_1, \dots, \mathcal{C}_n\}$.

Stage 2 - A model for estimating the cluster structure and disease risk

The study data are denoted by (\mathbf{Y}, \mathbf{E}) , and the best cluster structure for these data from the set of n candidates $\{\mathcal{C}_1, \dots, \mathcal{C}_n\}$ elicited from stage 1 is estimated together with disease risk by extending the Poisson log-linear CAR model given in Section 2 in two main ways. This approach takes advantage of the natural ordering of the cluster structures to allow the number of clusters to be considered as a univariate parameter within the model. The mechanism for implementing a given cluster structure is the neighbourhood matrix W , which is altered so that w_{ij} only equals one if areal units $(\mathcal{A}_i, \mathcal{A}_j)$ share a border and are in the same cluster. Thus if two adjacent areal units are in the same cluster their random effects are partially correlated and are smoothed over in the modeling, while if they are in different clusters they are conditionally independent and are not smoothed over. Thus there is a one-to-one relationship between the number of clusters and the value of W , and the n candidate values of W are denoted by (W_1, \dots, W_n) . Here W_1 corresponds to a single cluster and thus equals W , the original adjacency structure of the region. This value thus enforces strong spatial smoothing across the region, as no high or low risk clusters have been identified. In contrast, W_n corresponds to all n areal units being assigned to their own cluster of size one, and thus W_n is the zero matrix. This value thus corresponds to independent random effects with no spatial smoothing constraints.

However, the intrinsic CAR prior is not appropriate here, since our model could produce a neighbourhood matrix W in which an areal unit has no neighbours due to it being a singleton cluster. If this was areal unit i , this would cause $\sum_{j=1}^n w_{ij} = 0$, yielding an infinite mean and variance in (2). Instead, we use the localised CAR model outlined in [8], where an extended random effects vector $\tilde{\phi} = (\phi, \phi^*)$ is used, with ϕ^* being the global random effect which is potentially common to all areas and prevents the infinite mean and variance problem outlined above. An extended $(n+1) \times (n+1)$ neighbourhood matrix \tilde{W} is specified for this vector, which takes the form

$$\tilde{W} = \begin{pmatrix} W & \mathbf{w}_* \\ \mathbf{w}_*^T & 0 \end{pmatrix}$$

where $\mathbf{w}_* = (w_{1*}, \dots, w_{n*})$ and $w_{i*} = I[\sum_{i \sim j} (1 - w_{ij}) > 0]$. Here, $I[\cdot]$ denotes an indicator function, which sets $w_{i*} = 1$ if any entry in row i of the neighbourhood matrix W is changed from a 1 to a 0 due to a neighbouring area being in a different cluster. Otherwise, $w_{i*} = 0$. Based on this extended neighbourhood matrix, $\tilde{\phi}$ is modelled as $\tilde{\phi} = N(\mathbf{0}, \tau^2 Q(\tilde{W}, \epsilon)^{-1})$ with the precision matrix

$$Q(\tilde{W}, \epsilon) = \text{diag}(\tilde{W}\mathbf{1}) - \tilde{W} + \epsilon\mathbf{1}. \tag{3}$$

This corresponds to the intrinsic CAR model for the extended random effects vector $\tilde{\phi}$, with a small positive constant added to the diagonal of the precision matrix to ensure that it is invertible. The invertibility of $Q(\tilde{W}, \epsilon)$ is required as its determinant is computed when updating W , and [8] suggest that the results are insensitive to ϵ and set $\epsilon = 0.001$. The full conditionals of this extended CAR model are given by

$$\begin{aligned} \phi_i | \tilde{\phi}_{-i} &\sim N\left(\frac{\sum_{j=1}^n w_{ij} \phi_j + w_{i*} \phi_*}{\sum_{j=1}^n w_{ij} + w_{i*} + \epsilon}, \frac{\tau^2}{\sum_{j=1}^n w_{ij} + w_{i*} + \epsilon}\right), \\ \phi_* | \tilde{\phi}_{-*} &\sim N\left(\frac{\sum_{j=1}^n w_{j*} \phi_j}{\sum_{j=1}^n w_{j*} + \epsilon}, \frac{\tau^2}{\sum_{j=1}^n w_{j*} + \epsilon}\right). \end{aligned} \tag{4}$$

This means that the conditional expectation is a weighted average of the random effects in neighbouring areas and the global random effect ϕ^* , with binary weights based on the current choice of W matrix. Here, $\tilde{\mathbf{W}} = (\tilde{W}_1, \dots, \tilde{W}_n)$ is the set of extended neighbourhood matrices related to the set of cluster structures elicited in stage 1, where \tilde{W}_j is the matrix corresponding to the cluster structure with j clusters. Given this extended CAR prior, the overall Bayesian hierarchical model we propose is given by

$$\begin{aligned} Y_i | E_i, R_i &\sim \text{Poisson}(E_i R_i) \text{ for } i = 1, \dots, n, \\ \ln(R_i) &= \beta_0 + \phi_i, \\ \tilde{\phi} &\sim N(\mathbf{0}, \tau^2 Q(\tilde{W}, \epsilon)^{-1}), \\ \tilde{W} &\sim \text{Discrete}(\tilde{W}_1, \dots, \tilde{W}_n; \pi_1, \dots, \pi_n), \\ \pi_j &= \frac{\exp(-j\theta)}{\sum_{i=1}^n \exp(-i\theta)}, \\ \beta_0 &\sim N(0, 1000) \text{ for } j = 1, \dots, p, \\ \theta &\sim \text{Uniform}(0, 1), \\ \tau^2 &\sim \text{Uniform}(0, 1000). \end{aligned} \tag{5}$$

Initially, a discrete uniform prior was considered for \tilde{W} , but it may not be appropriate to give equal weighting to structures with extremely large numbers of clusters, as the spatial autocorrelation present in the data suggests the number of clusters will be relatively small. Therefore our prior probabilities for $(\tilde{W}_1, \dots, \tilde{W}_n)$ are given by (π_1, \dots, π_n) , with an additional parameter θ being introduced to control the strength of the weights. When $\theta = 0$ a discrete uniform prior is assumed for \tilde{W} , while $\theta = 1$ corresponds to a scaled exponential weighting which gives larger prior weight to values of W corresponding to fewer clusters.

4 Motivating application

The study region is the Greater Glasgow and Clyde health board, which contains the city of Glasgow in the east and the river Clyde estuary in the west. Glasgow is the largest city in Scotland, with a population of around 600,000 people. The health board is split into $n = 271$ administrative units known as intermediate geographies (IGs), containing populations of between 2,244 and 10,877 people with a median value of 4,239. The disease data are the numbers of hospital admissions with a primary diagnosis of respiratory disease in each IG in 2011, which corresponds to the International Classification of Disease tenth revision codes J00-J99 and R09.1. The expected hospital admission numbers were calculated using external standardisation, based on age and sex adjusted rates for the whole of Scotland. The top panel of Figure 1 displays the standardised incidence ratio (SIR) for respiratory hospital admission, which is the ratio of the observed to the expected numbers of cases. The figure identifies regions of high risk in the east of the city and directly to the south of the river, which include heavily deprived areas such as Easterhouse and Govan. In contrast, areas in the centre (just north of the river) and far south of the study region exhibit much lower risks, which are affluent areas such as the West End and Giffnock.

Results

The two-stage clustering model proposed in Section 3 was applied to these data, where the clustering step used respiratory disease data from 2008 to 2010. The fitted risk surfaces for these data sets exhibit similar spatial patterns to the 2011 study data, with Pearson's correlation coefficients of 0.86 (2010 data), 0.84 (2009 data) and 0.82 (2008 data) respectively. Markov-Chain Monte Carlo inference was used to obtain these results, with 5000 samples used for burn-in and a further 5000 used for the inference. The optimal cluster structure was chosen to be that corresponding to the mode cluster number, which in this case was 18. Our method has the advantage of being able to quantify the uncertainty in the number of clusters identified, and a 95% credible interval for this ranges between 17 and 28. In addition, the median cluster number was 21.

The estimated risk surface (greyscale) and cluster structure (white dots) for the configuration with 18 clusters are displayed in the bottom panel of Figure 1, which has the same scale as the SIR plot in the top panel of that figure. In the majority of cases, there do appear to be differences in risks between neighbouring clusters, and two of the more prominent clusters have been labelled A and B on the map. The low risk Cluster A is the affluent West End of the city which is surrounded on all sides by more deprived areas. Cluster B includes a number of prosperous areas to the north of the city, including Milngavie, Milton of Campsie and Lennoxton, which have much lower risks than neighbouring areas such as Kirkintilloch and the East End of the city, which are less affluent. The clusters appear to be based around grounds of socio-economic deprivation, which is well known to be linked with disease risk. The high risk areas in Figure 1 are generally areas with high levels of deprivation, while the lower risk areas are more affluent.

5 Conclusion

The main aim of this paper was to develop statistical methodology to simultaneously estimate the spatial pattern in disease risk and identify clusters of areas exhibiting high (and low) risk. To achieve this aim a new methodology has been developed which fuses together spatial agglomerative hierarchical clustering techniques with an extended conditional autoregressive model, with inference based on Markov-Chain Monte Carlo simulation. This approach allows us to identify an optimal cluster structure which best describes the data, and it extends [1] by quantifying the uncertainty in the cluster structure. The clustering techniques are applied to disease risk data prior to our study period, allowing us to elicit candidate cluster structures for the study data. These candidate structures have a natural ordering in terms of the number of clusters, which allows them to be considered as a univariate parameter in our Bayesian hierarchical model. This model estimates disease risk directly via the random effects, allowing for correlation between neighbouring areal units which are in the same cluster, but not enforcing it for areas in different clusters. Our approach differs from that used in [1], where the cluster structure was fixed when estimating the remaining model parameters. Here we are able to produce a credible interval for the number of clusters and can identify other potential alternative cluster structures which are supported by the data.

Acknowledgements

The work of the first author was funded by the Carnegie Trust. An extended version of this paper has been submitted to the Computational Statistics & Data Analysis.

Bibliography

- [1] Anderson, C., Lee, D. and Dean, N. (2014) *Identifying Clusters in Bayesian Disease Mapping*. Biostatistics to appear.
- [2] Besag, J., York, J. and Mollie, A. (1991) *Bayesian image restoration, with two applications in spatial statistics*. Annals of the Institute of Statistical Mathematics, **43**, 1–20.
- [3] Charras-Garrido, M., Abrial, D. and de Goer, J. (2012) *Classification method for disease risk mapping based on discrete hidden Markov random fields*. Biostatistics, **13**, 241–255.
- [4] Diggle, P., Rowlingson, B. and Su, T. (2005) *Point process methodology for on-line spatio-temporal disease surveillance*. Environmetrics, **16**, 423–434.
- [5] Hastie, T., Tibshirani, R. and Friedman, J. (2001) *The Elements of Statistical Learning*. **Chapter 14.3**.
- [6] Kulldorff, M. (1997) *A Spatial Scan Statistic*. Communications in Statistics, **26**, 1481–1496.
- [7] Lee, D., and Mitchell, R. (2013) *Locally adaptive spatial smoothing using conditional autoregressive models*. Journal of the Royal Statistical Society Series C, **62**, 593–608.

- [8] Lee, D., Rushworth, A. and Sahu, S. (2013) *A Bayesian localised conditional auto-regressive model for estimating the health effects of air pollution*. Biometrics to appear.



Figure 1: The top panel displays the standardised incidence ratio (grey-scale) for respiratory disease hospitalisation in 2011 in Greater Glasgow. The bottom plot displays the estimated risk surface (grey-scale) from the model with 18 clusters (white dots).

Robust test of restricted model

Jan Amos Visek, *Charles University in Prague*, visek@fsv.cuni.cz

Abstract. The paper proposes a test of restricted versus unrestricted model in the framework of linear regression model estimated by the least weighted squares - robust version of the ordinary least squares. The patterns of simulations of the quantiles of test statistic are included.

Keywords. Test of restricted model, robustness, the least weighted squares.

1 Introduction of basic framework

Test of restricted versus unrestricted model - based on comparison of residual sums of squares - belongs to the basic tool of classical regression analysis. A justification of the idea of utilizing the sums of squares has its roots in the L_2 -metric of Euclidean space and in a nearly exclusive employment of the least squares (LS) and/or the maximum likelihood (ML). Because the latter was employed mostly under assumption of normality, when LS and ML coincide, it implicitly accepted L_2 -metric, too. But this justification is not sufficient. We need the fact that the restricted model is nested in the unrestricted one (see Conditions C3). For the non-nested models the situation is much more complicated and the (heuristic) ideas supporting an alternative approach seem to be less convincing than those for the nested models, see e.g. [19].

In the robust approach the problem was firstly addressed on ICORS 2011 and ERCIM 2013 (see [18]) and in a bit different framework by Hannay and Stahel, see [5], [11]. There were also attempts to establish significance of the individual explanatory variable by the bootstrap, see [9] or [10], which can lead - when generalized into the framework with submodels - to the similar results as we present here.

The low attention devoted to the topics is quite understandable from the technical point of view. Robust estimators - usually defined as solutions of extremal problems - either distort the L_2 -metric or turn originally nested models into non-nested ones - we will see it below¹⁰. The paper is an attempt to solve the problem as a technical one, the philosophical side of it would require much more space. Let us fix notations.

¹⁰There are also other technical reasons - e.g. in the most cases we cannot apply Fisher-Cochran theorem because different weights given to different residual in fact changes the homoscedastic situation into heteroscedastic one - details again below.

Denote by N the set of all positive integers and for any $p \in N$ let R^p be the p -dimensional Euclidean space¹¹. Further, let (Ω, \mathcal{A}, P) be the probability space and write $Z(\omega)$ for the random variable (r.v.) Z when we need to emphasize that we consider its value of at a given point $\omega \in \Omega$. Finally, \mathbf{I} stays for the diagonal unit matrix and $\mathcal{N}(\mu, \Sigma)$ denotes the normal distribution with the mean vector μ and covariance matrix Σ .

For any $n \in N$ and a fixed $\beta^0 \in R^p$ consider the linear regression model

$$Y_i = X_i' \beta^0 + e_i, \quad i = 1, 2, \dots, n. \quad (1)$$

where X_i 's are p -dimensional explanatory variables, see Conditions $\mathcal{C}1$ below. In what follows we will need also matrix notations. So, let us write the model in the form

$$Y = X\beta^0 + e \quad (2)$$

where $Y = (Y_1, Y_2, \dots, Y_n)'$, $X = (X_1, X_2, \dots, X_n)'$ and $e = (e_1, e_2, \dots, e_n)'$. We will keep the traditional framework, the significance of explanatory variables was studied in, i. e. assuming the normality of error terms. So, we consider the framework implied by the assumptions:

Conditions $\mathcal{C}1$ *The sequence $\{(X_i', e_i)'\}_{i=1}^\infty$ is sequence of independent and identically distributed $(p+1)$ -dimensional random variables (i.i.d. r.v.'s) with distribution function $F_{X,e}(v, u) = F^{(1)}(v^{(1)}) \cdot F_{X,e}^{(2)}(v^{(2)}, u)$ where $F^{(1)}(v^{(1)}) : R^1 \rightarrow [0, 1]$ is d. f. degenerated at 1 and $F_{X,e}^{(2)}(v^{(2)}, u) = F_X(v^{(2)}) \cdot F_e(u)$ where $F_X(v^{(2)}) : R^{p-1} \rightarrow [0, 1]$ is absolutely continuous and $F_e(u) = \mathcal{N}(0, \sigma^2)$. Finally, $\text{cov}(X_1)$ is regular.*

Remark 1 *Conditions $\mathcal{C}1$ allow for intercept. Hence the first column of the design matrix is the n -dimensional vector of ones. If the model without intercept is considered a straightforward modification of the present conditions has to be adopted. All the considerations, made in what follows, then keep to hold.*

For any $\beta \in R^p$ and $i \in N$ let us denote the i -th residual as

$$r_i(\beta) = Y_i - X_i' \cdot \beta. \quad (3)$$

Definition 1 *For $n \in N$ put $W = \text{diag}\{w_1, w_2, \dots, w_n\}$ with $w_i \in [0, 1]$. Then the estimator*

$$\hat{\beta}^{(WLS, n, w)} = \arg \min_{\beta \in R^p} \sum_{i=1}^n w_i r_i^2(\beta) = (X'WX)^{-1} X'WY \quad (4)$$

is called the weighted least squares estimator (WLS)¹².

Denoting $r_{(i)}^2(\beta)$ the i -th order statistic among the squared residuals, i. e.

$$r_{(1)}^2(\beta) \leq r_{(2)}^2(\beta) \leq \dots \leq r_{(n)}^2(\beta),$$

we will study a robust version of $\hat{\beta}^{(OLS, n)}$, namely:

Definition 2 *The estimator*

$$\hat{\beta}^{(LWS, n, w)} = \arg \min_{\beta \in R^p} \sum_{i=1}^n w_i r_{(i)}^2(\beta) \quad (5)$$

¹¹To avoid possible confusions, hereafter all the vectors will be assumed to be the column ones.

¹²For the simplicity we have assumed that $X'WX$ is regular; otherwise we have to employ a generalized inverse.

is called the *least weighted squares estimator (LWS)*, see [14].

Remark 2 Notice that in the definition of $\hat{\beta}^{(LWS,n,w)}$ the weights are assigned to the order statistics of squared residuals rather than to the squared residuals directly. It is immediately clear that $\hat{\beta}^{(LWS,n,w)}$ is in fact $\hat{\beta}^{(WLS,n,w)}$ applied on some permutation, say $\pi = \pi(w)$, of the order of original data, see [17, 18] and discussion below.

Remark 3 Let $h \in N, \frac{n}{2} < h \leq n, w_h^* = 1$ and $w_i^{**} = 1, i = 1, 2, \dots, h$ with $w_i^* = w_i^{**} = 0$ otherwise. Then

$$\hat{\beta}^{(LWS,n,w^*)} = \arg \min_{\beta \in R^p} r_{(h)}^2(\beta) = \hat{\beta}^{(LMS,n,h)} \text{ and } \hat{\beta}^{(LWS,n,w^{**})} = \arg \min_{\beta \in R^p} \sum_{i=1}^h r_{(i)}^2(\beta) = \hat{\beta}^{(LTS,n,h)}$$

for $\hat{\beta}^{(LMS,n,h)}$ see [8] and for $\hat{\beta}^{(LTS,n,h)}$ [4].

In what follows we shall restrict ourselves on monotone weights and we will assume:

Conditions C2 The weight function $w(u)$ is continuous, strictly decreasing, $w : [0, 1] \rightarrow [0, 1]$ with $w(0) = 1$ and put $w_i = w\left(\frac{i-1}{n}\right)$.

The strict monotonicity of weights is only a technical task, in simulation below we will use $w(u)$ decreasing for $0 \leq u \leq h/100$ from 1 to 0.99, then linearly decreasing to 0.01 at $u = g/100$ and finally decreasing from this value to $w(1) = 0$.

Both, $\hat{\beta}^{(LMS,n,h)}$ as well as $\hat{\beta}^{(LTS,n,h)}$, can suffer by instability with respect to an (arbitrary) small shift or deletion of one observation, see [6, 12, 13, 15]. Moreover, in contrast to M - or MM -estimators - when we want to reach the scale- and regression-equivariance of the estimator - $\hat{\beta}^{(LWS,n,w^*)}$ does not require any studentization of residuals by the scale-equivariant and regression-invariant estimator of standard deviation of error term, see [2]. In other words, we can compute it “directly” by a simple and quick algorithm, without computing a preliminary scale estimate, compare it with MM -estimators in [20]. Finally, the speed of algorithm allows to tailor the weight function for unknown level of contamination in a way of the Forward Search, [1]. That was the reasons for proposing $\hat{\beta}^{(LWS,n,w)}$ which has improved some features of previous estimators and inherited the plausible properties. On the other hand, we have to pay for it. The price, to be paid, is highly involved proof of consistency (as we estimate implicitly the scale of error terms), [17] (see also [16]).

2 Preliminary findings on the least weighted squares

Let $\Pi^{(n)}$ be the set of all permutations of integers $\{1, 2, \dots, n\}$. Fix $\pi \in \Pi^{(n)}, \pi = \{\pi_1, \pi_2, \dots, \pi_n\}$ and put

$$Y_\pi = (Y_{\pi_1}, Y_{\pi_2}, \dots, Y_{\pi_n})', \quad X_\pi = (X_{\pi_1}, X_{\pi_2}, \dots, X_{\pi_n})' \quad \text{and} \quad e_\pi = (e_{\pi_1}, e_{\pi_2}, \dots, e_{\pi_n})'$$

and consider the model

$$Y_\pi = X_\pi \beta^0 + e_\pi.$$

Then denote

$$\hat{\beta}_\pi^{(WLS,n,w)} = \arg \min_{\beta \in R^p} \sum_{i=1}^n w_i (Y_{\pi_i} - X'_{\pi_i} \beta)^2 = (X'_\pi W X_\pi)^{-1} X'_\pi W Y_\pi \tag{6}$$

and (see Definition 1)

$$S_{\pi}^2 = \sum_{i=1}^n w_i (Y_{\pi_i} - X'_{\pi_i} \hat{\beta}_{\pi}^{(WLS,n,w)})^2 = \min_{\beta \in R^p} \sum_{i=1}^n w_i (Y_{\pi_i} - X'_{\pi_i} \beta)^2. \quad (7)$$

Obviously we have $S_{\pi}^2 = S_{\pi}^2(\omega)$. Finally, fix $\omega_0 \in \Omega$ and put

$$\hat{\pi}(\omega_0) = \arg \min_{\pi \in \Pi^{(n)}} S_{\pi}^2(\omega_0). \quad (8)$$

Notice that $\hat{\pi}$ depends on ω but naturally also on Y and X (which we did not include into notation - for a while - as they are fixed and moreover when fixing ω_0 , we have $Y = Y(\omega_0)$ and $X = X(\omega_0)$). The definition of $\hat{\beta}^{(WLS,n,w)}$, i. e. (6), and (8) yield

$$S_{\hat{\pi}(\omega_0)}^2(\omega_0) = \min_{\pi \in \Pi^{(n)}} \min_{\beta \in R^p} \sum_{i=1}^n w_i (Y_{\pi_i} - X'_{\pi_i} \beta)^2 \quad (9)$$

and hence

$$\hat{\beta}_{\hat{\pi}(\omega_0)}^{(WLS,n,w)}(\omega_0) = \hat{\beta}^{(LWS,n,w)}(\omega_0), \quad (10)$$

for more details see [17, 18]. Fixing some $\pi \in \Pi^{(n)}$, let us put

$$B(\pi) = \{\omega \in \Omega : \pi = \hat{\pi}(\omega)\},$$

remember (8). We have

$$\cup_{\pi \in \Pi^{(n)}} B(\pi) = \Omega$$

and the strict monotonicity of weights implies that from $\pi^{(1)} \neq \pi^{(2)}$ we have

$$P(B(\pi^{(1)}) \cap B(\pi^{(2)}) = \emptyset) = 1.$$

Let us consider a fix $\pi^0 \in \Pi^{(n)}$. Taking into account (10), we find for any $\omega \in B(\pi^0)$

$$\hat{\beta}^{(LWS,n,w)}(\omega) = \hat{\beta}_{\pi^0}^{(WLS,n,w)}(\omega) = \arg \min_{\beta \in R^p} \sum_{i=1}^n w_i (Y_{\pi_i^0}(\omega) - X'_{\pi_i^0}(\omega) \beta)^2. \quad (11)$$

Recalling that $\{(X'_i, e_i)'\}_{i=1}^{\infty}$ is a sequence of i.i.d. r.v.'s, we conclude that for all $\pi \in \Pi^{(n)}$ the set $B(\pi)$ has the same probabilistic features. In other words, $B(\pi)$'s are undistinguishable from the probabilistic point of view each from other. It means also that

$$P(B(\pi)) = (n!)^{-1}. \quad (12)$$

It opens a way for studying the tests statistics (proposed below) conditionally for selected $\pi \in \Pi^{(n)}$ and then "to take the mean value" employing (12).

We will fix $\pi \in \Pi^{(n)}$ once again. Then for any $\omega \in B(\pi)$ (6) and (11) imply

$$\hat{\beta}^{(LWS,n,w)}(Y, X) = \hat{\beta}_{\pi}^{(WLS,n,w)}(Y, X) = \arg \min_{\beta \in R^p} \{(Y_{\pi} - X_{\pi} \beta)' W (Y_{\pi} - X_{\pi} \beta)\}$$

$$= \arg \min_{\beta \in R^p} \left\{ (\tilde{Y}_\pi - \tilde{X}_\pi \beta)' (\tilde{Y}_\pi - \tilde{X}_\pi \beta) \right\} = \hat{\beta}^{(OLS,n)}(\tilde{Y}_\pi, \tilde{X}_\pi) \tag{13}$$

where for W see Definition 1 and

$$\tilde{Y}_\pi = W^{\frac{1}{2}} Y_\pi \quad \text{and} \quad \tilde{X}_\pi = W^{\frac{1}{2}} X_\pi. \tag{14}$$

So, although the *least weighted squares* - similarly as other robust estimators distorted the L_2 metric, but by the considerations, we have just concluded, we returned back to this metric, of course for the data transformed as given in (14), i. e. for the heteroscedastic data. Please remember in the rest of paper the relation (14). The equality

$$\hat{\beta}^{(LWS,n,w)}(Y, X) = \hat{\beta}_\pi^{(WLS,n,w)}(Y, X) = \hat{\beta}^{(OLS,n)}(\tilde{Y}_\pi, \tilde{X}_\pi) \tag{15}$$

is the fundamental for all further considerations, it open the way for all further results. Moreover, it hints that the idea to compare the residual sum of squares for restricted and unrestricted model can attain again the legitimacy. Unfortunately, it holds only partially due to the heteroscedasticity of data, see (14) and the next paragraph.

3 Establishing the test of restricted against unrestricted model

Let us start with fixing the framework. We want to test the models

$$Y = X\beta^0 + e \quad \text{against} \quad Y = Z\gamma^0 + v \tag{16}$$

where $\gamma^0 \in R^q$ and we will assume that the Conditions C1 hold with the appropriate modifications also for the latter model. Moreover, we assume - as in the classical framework - that models are nested:

Conditions C3 *The matrices X and Z are such that $\mathcal{M}(Z) \subset \mathcal{M}(X)$ where $\mathcal{M}(A)$ is the set of all linear combinations of columns of matrix A .*

Remark 5 *Prior to continuing let us recall that the test of restricted versus unrestricted model in the classical framework is based on the idea that the respective residual sums of squares are not significantly different. We apply the test to avoid an unpleasant situation of employing restricted model although it is worse than the unrestricted one. Hence the test is to reject the hypothesis if there is a suspicion that the restricted model need not be sufficient for explanation.*

At this moment we need to recall one thing from the previous paragraph. As we saw in (15), we have defined at any point $\omega \in \Omega$ $\pi = \pi(\omega)$ and $\tilde{Y}_\pi, \tilde{X}_\pi$ (see (14) and (8), respectively) and then $\hat{\beta}^{(LWS,n,w)}(Y, X) = \hat{\beta}^{(OLS,n)}(\tilde{Y}_\pi, \tilde{X}_\pi)$ We have also mentioned at the previous paragraph that $\pi = \pi(\omega, Y, X)$. Finally, we can study the model

$$\tilde{Y}_{\pi(\omega, Y, X)} = \tilde{X}_{\pi(\omega, Y, X)} \beta^0 + \tilde{e}_{\pi(\omega, Y, X)} \tag{17}$$

$$\text{with } \tilde{Y}_{\pi(\omega, Y, X)} = W^{\frac{1}{2}} Y_{\pi(\omega, Y, X)}, \tilde{X}_{\pi(\omega, Y, X)} = W^{\frac{1}{2}} X_{\pi(\omega, Y, X)} \quad \text{and} \quad \tilde{e}_{\pi(\omega, Y, X)} = W^{\frac{1}{2}} e_{\pi(\omega, Y, X)} \tag{18}$$

and employ the *ordinary least squares*. The residual sum of squares for this model is then (see (7) and (15))

$$S_{\pi(\omega, Y, X)}^2 = \sum_{i=1}^n w_i \left(Y_{\pi_i(\omega, Y, X)} - X'_{\pi_i(\omega, Y, X)} \hat{\beta}_{\pi(\omega, Y, X)}^{(LWS,n,w)} \right)^2$$

$$= \sum_{i=1}^n \left(\tilde{Y}_{\pi_i(\omega, Y, X)} - \tilde{X}'_{\pi_i(\omega, Y, X)} \hat{\beta}_{\pi(\omega, Y, X)}^{(OLS, n)} \left(\tilde{Y}_{\pi(\omega, Y, X)}, \tilde{X}_{\pi(\omega, Y, X)} \right) \right)^2. \tag{19}$$

We can do the same for the latter model in (16). But then generally the inclusion

$$\mathcal{M}(Z_{\pi(\omega, Y, Z)}) \subset \mathcal{M}(X_{\pi(\omega, Y, X)}) \tag{20}$$

does not hold because generally we have $\pi(\omega, Y, Z) \neq \pi(\omega, Y, X)$. In other words, the models $\tilde{Y}_{\pi(\omega, Y, X)} = \tilde{X}_{\pi(\omega, Y, X)}\beta^0 + \tilde{e}_{\pi(\omega, Y, X)}$ and $\tilde{Y}_{\pi(\omega, Y, Z)} = \tilde{Z}_{\pi(\omega, Y, Z)}\beta^0 + \tilde{e}_{\pi(\omega, Y, Z)}$ are not nested although the models in (16) were nested. So, we cannot utilize the residual sums of squares $S_{\pi(\omega, Y, X)}^2$ and $S_{\pi(\omega, Y, Z)}^2$, for comparing the “quality” of models in (16), at least not in the philosophy of classical regression analysis.

Now, we have basically two possibilities. Either to test the models in (16) by means of residual sum of squares of two models - which are however nested - or to test them without the “restriction of being nested” as it was done in the classical statistics for the non-nested models. We are going to study in the rest of paper the former possibility. We do not claim in any case that the latter one is a deadlock but it requires much more space and presentations of results of various attempts (for discussion see e. g. [7, 19]). So, we will consider the heteroscedastic models¹³ (its clear that we can drop ω from the subindex $\pi(\omega, Y, X)$ due to (12) and the considerations above it)

$$\tilde{Y}_{\pi(Y, X)} = \tilde{X}_{\pi(Y, X)}\beta^0 + \tilde{e}_{\pi(Y, X)} \quad \text{against} \quad \tilde{Y}_{\pi(Y, X)} = \tilde{Z}_{\pi(Y, X)}\gamma^0 + \tilde{v}_{\pi(Y, X)} \tag{21}$$

with

$$\tilde{Y}_{\pi(Y, X)} = W^{\frac{1}{2}}Y_{\pi(Y, X)}, \quad \tilde{X}_{\pi(Y, X)} = W^{\frac{1}{2}}X_{\pi(Y, X)}, \quad \tilde{e}_{\pi(Y, X)} = W^{\frac{1}{2}}e_{\pi(Y, X)}, \quad \mathcal{L}(\tilde{e}_{\pi(Y, X)}) = \mathcal{N}(0, \sigma^2W) \tag{22}$$

and

$$\tilde{Z}_{\pi(Y, X)} = W^{\frac{1}{2}}Z_{\pi(Y, X)}, \quad \tilde{v}_{\pi(Y, X)} = W^{\frac{1}{2}}v_{\pi(Y, X)} \quad \text{and} \quad \mathcal{L}(\tilde{v}_{\pi(Y, X)}) = \mathcal{N}(0, \sigma^2W). \tag{23}$$

Moreover, notice please that under Conditions $\mathcal{C}1$ - due to the fact that W is the diagonal matrix - \tilde{e} as well as \tilde{v} have independent coordinates - we will need it later.

So, we have transformed the task (16) to the form (21), the regression coefficients will be estimated by the *ordinary least squares* and the test statistic will be based on the residual sum of squares. At the first glance it seems simple, because in the classical framework such task was solved - but not for heteroscedastic error terms because the crucial step in the study of such a problem was the employment of Fisher-Cochran theorem, requiring the homoscedastic response variable. At this moment, an idea to transform the heteroscedastic data back to the homoscedastic ones and then carry out the test can appear. But then the all probabilistic considerations would be done in a different framework¹⁴. Due to the fact that the reordering of rows in the design matrices X and Z (as given in (21)) was the same, (20) holds. Then we can proceed in the same way as in the classical statistics employing the decomposition

$$\tilde{e}'\tilde{e} = \tilde{e}' \left[\mathbf{I} - \tilde{X} (\tilde{X}'\tilde{X})^{-1} \tilde{X}' \right] \tilde{e} + \tilde{e}' \left[\tilde{X} (\tilde{X}'\tilde{X})^{-1} \tilde{X}' - \tilde{Z} (\tilde{Z}'\tilde{Z})^{-1} \tilde{Z}' \right] \tilde{e} + \tilde{e}' \tilde{Z} (\tilde{Z}'\tilde{Z})^{-1} \tilde{Z}' \tilde{e} \tag{24}$$

¹³Realize please that the heteroscedasticity is somewhat “artificial”, i. e. the character of heteroscedasticity is given - except of “nuisance” parameter σ^2 - by the matrix W .

¹⁴An extended discussion with details was carried out in [18] where also some other references to the topic were given.

where (see (3))

$$\begin{aligned} \tilde{e}' \left[\mathbf{I} - \tilde{X} (\tilde{X}'\tilde{X})^{-1} \tilde{X}' \right] \tilde{e} &= e' W^{\frac{1}{2}} \left[\mathbf{I} - \tilde{X} (\tilde{X}'\tilde{X})^{-1} \tilde{X}' \right] W^{\frac{1}{2}} e \\ &\underbrace{=}_{\text{denote}} e' M_1 e = r' \left(\hat{\beta}^{OLS,n}(\tilde{Y}, \tilde{X}) \right) r \left(\hat{\beta}^{OLS,n}(\tilde{Y}, \tilde{X}) \right) \end{aligned} \tag{25}$$

and in the similar way we can write the second and the third term of (24) as

$$e' M_2 e \quad \text{and} \quad e' M_3 e.$$

Now we can employ the same steps as in the proof of Fisher-Cochran theorem, see e. g. [3]. Let the orthonormal vector base of R^n , say $\tilde{q}_1, \tilde{q}_2, \dots, \tilde{q}_n$, be selected so that the vectors $(\tilde{q}_1, \tilde{q}_2, \dots, \tilde{q}_{n-p})$, $(\tilde{q}_{n-p+1}, \dots, \tilde{q}_{n-q})$ and $(\tilde{q}_{n-q+1}, \tilde{q}_{n-q+2}, \dots, \tilde{q}_n)$ are the eigenvectors of the matrices M_1 , M_2 and M_3 corresponding to their positive eigenvalues, respectively. Moreover, denote $(\psi_1, \psi_2, \dots, \psi_{n-p}, 0, \dots, 0)$, $(0, 0, \dots, 0, \phi_{n-p+1}, \phi_{n-p+2}, \dots, \phi_{n-q}, 0, \dots, 0)$ and $(0, 0, \dots, 0, \theta_{n-q+1}, \theta_{n-q+2}, \dots, \theta_n)$ the eigenvalues corresponding to these eigenvectors of the three matrices, respectively. Finally put $Q = (\tilde{q}_1, \tilde{q}_2, \dots, \tilde{q}_n)$ and recall that then $Q'Q = QQ' = \mathbf{I}$ and also recall that $\mathcal{L}(e) = \mathcal{N}(0, \sigma^2 \mathbf{I})$.

Having put $\Psi = \text{diag}(\psi_1, \psi_2, \dots, \psi_{n-p}, 0, \dots, 0)$, $\Phi = \text{diag}(0, 0, \dots, 0, \phi_{n-p+1}, \dots, \phi_{n-q}, 0, \dots, 0)$ and $\Theta = \text{diag}(0, 0, \dots, 0, \theta_{n-q+1}, \theta_{n-q+2}, \dots, \theta_n)$, let us define random vectors $\varepsilon = \Psi^{\frac{1}{2}} Q' e$, $\tau = \Phi^{\frac{1}{2}} Q' e$ and $\kappa = \Theta^{\frac{1}{2}} Q' e$ so that $\mathbb{E}\varepsilon = 0$, $\text{cov}(\varepsilon) = \mathbb{E} \left[\Psi^{\frac{1}{2}} Q' e e' Q \Psi^{\frac{1}{2}} \right] = \sigma^2 \Psi^{\frac{1}{2}} Q' Q \Psi^{\frac{1}{2}} = \sigma^2 \Psi^{\frac{1}{2}} \Psi^{\frac{1}{2}} = \sigma^2 \Psi$ and hence $\mathcal{L}(\varepsilon) = \mathcal{N}(0, \sigma^2 \Psi)$. Similarly, $\mathbb{E}\tau = 0$, $\text{cov}(\tau) = \sigma^2 \Phi$ and $\mathcal{L}(\tau) = \mathcal{N}(0, \sigma^2 \Phi)$ and $\mathbb{E}\kappa = 0$, $\text{cov}(\kappa) = \sigma^2 \Theta$ and $\mathcal{L}(\kappa) = \mathcal{N}(0, \sigma^2 \Theta)$. Moreover, $\varepsilon_i \equiv 0$ for $i = n - p + 1, \dots, n$, $\tau_i \equiv 0$ for $i = 1, 2, \dots, n - p, n - q + 1, \dots, n$, $\kappa_i \equiv 0$ for $i = 1, 2, \dots, n - q$ and covariance matrix of the vector $(\varepsilon', \tau', \kappa)'$ is the diagonal one. Then their normality yields their independence. Now,

$$\varepsilon' \varepsilon = \sum_{i=1}^{n-p} \varepsilon_i^2 = e' Q \Psi Q' e = e' M_1 e = r' \left(\hat{\beta}^{OLS,n}(\tilde{Y}, \tilde{X}) \right) r \left(\hat{\beta}^{OLS,n}(\tilde{Y}, \tilde{X}) \right). \tag{26}$$

So, $r' \left(\hat{\beta}^{OLS,n}(\tilde{Y}, \tilde{X}) \right) r \left(\hat{\beta}^{OLS,n}(\tilde{Y}, \tilde{X}) \right)$ is the sum of $n - p$ independent r. v.'s which are squares of normally distributed r. v.'s. Then utilizing (25) (remember that $\tilde{e} = W^{\frac{1}{2}} e$)

$$\begin{aligned} \text{var} \left(\left\| r \left(\hat{\beta}^{OLS,n}(\tilde{Y}, \tilde{X}) \right) \right\| \right) &= \mathbb{E} \left(r' \left(\hat{\beta}^{OLS,n}(\tilde{Y}, \tilde{X}) \right) r \left(\hat{\beta}^{OLS,n}(\tilde{Y}, \tilde{X}) \right) \right) \\ &= \mathbb{E} \left(e' W^{\frac{1}{2}} \left[\mathbf{I} - \tilde{X} (\tilde{X}'\tilde{X})^{-1} \tilde{X}' \right] W^{\frac{1}{2}} e \right) = \mathbb{E} \left(\text{tr} \left(e' W^{\frac{1}{2}} \left[\mathbf{I} - \tilde{X} (\tilde{X}'\tilde{X})^{-1} \tilde{X}' \right] W^{\frac{1}{2}} e \right) \right) \\ &= \mathbb{E} \left(\text{tr} \left(W^{\frac{1}{2}} e e' W^{\frac{1}{2}} \left[\mathbf{I} - \tilde{X} (\tilde{X}'\tilde{X})^{-1} \tilde{X}' \right] \right) \right) = \sigma^2 W \left[\mathbf{I} - \tilde{X} (\tilde{X}'\tilde{X})^{-1} \tilde{X}' \right] = \sum_{i=1}^n w_i (1 - d_{ii}) \end{aligned}$$

where $d_{ii} = \left[\tilde{X} (\tilde{X}'\tilde{X})^{-1} \tilde{X}' \right]_{ii}$. Taking into account that $\varepsilon = \Psi^{\frac{1}{2}} Q' e$ and employing the (22), we find analogously $\text{var}(\|\varepsilon\|) = \sigma^2 \sum_{i=1}^{n-p} \psi_i$. Due to (26), we have $\sum_{i=1}^{n-p} \psi_i = \sum_{i=1}^n w_i (1 - d_{ii})$. It means that the residual sum of squares $\varepsilon' \varepsilon = r' \left(\hat{\beta}^{OLS,n}(\tilde{Y}, \tilde{X}) \right) r \left(\hat{\beta}^{OLS,n}(\tilde{Y}, \tilde{X}) \right)$ has the *generalized χ^2 -distribution*:

Definition 3 Let for $k \in N$ and $\sigma_i^2 > 0, i = 1, 2, \dots, k$ the sequence of random variables $\{\xi_i\}_{i=1}^k$ be normally distributed with zero means and variances σ_i^2 . Then the distribution of the random variable $\tau = \sum_{i=1}^k \xi_i^2$ will be called *generalized χ^2 -distribution* with $s =$

$\sum_{i=1}^k \sigma_i^2$ degrees of freedom and k -dimensional scale-parameter $\sigma^2 = (\sigma_1^2, \sigma_2^2, \dots, \sigma_k^2)$. Denote it by $\chi_{generalized}^2(s, \sigma^2)$. Further, let the random variable η be independent from τ and have the $\chi_{generalized}^2(t, \lambda^2)$ distribution. Then the distribution of random variable $\frac{\eta}{t} / \frac{\tau}{s}$ will be called the *generalized F-distribution with s and k degrees of freedom and scale-parameters λ^2 and σ^2* . Denote it by $F_{generalized}(s, t, \lambda^2, \sigma^2)$.

Along the similar lines we find that $\tau' \tau = (\hat{\beta}^{OLS,n}) (\hat{\beta}^{OLS,n}) - r' (\hat{\gamma}^{OLS,n}) r (\hat{\gamma}^{OLS,n})$ has the *generalized χ^2 -distribution with $\sigma^2 \sum_{i=n-p+1}^{n-q} \phi_i = \sigma^2 \sum_{i=1}^n w_i (d_{ii} - c_{ii})$ degrees of freedom ($c_{ii} = [\tilde{Z}' (\tilde{Z}' \tilde{Z})^{-1} \tilde{Z}']_{ii}$) and the $(p-q)$ -dimensional scale parameter $\sigma_\tau^2 = (\sigma^2 \phi_{n-p+1}, \sigma^2 \phi_{n-p+2}, \dots, \sigma^2 \phi_{n-q})$. Due to the independence of ε and τ , we have:*

Theorem 3.1.

$$F = \frac{r' (\hat{\beta}^{OLS,n}(\tilde{Y}, \tilde{X})) r (\hat{\beta}^{OLS,n}(\tilde{Y}, \tilde{X})) - r' (\hat{\gamma}^{OLS,n}(\tilde{Y}, \tilde{Z})) r (\hat{\gamma}^{OLS,n}(\tilde{Y}, \tilde{Z}))}{\sum_{i=1}^n w_i (d_{ii} - c_{ii})} \times \frac{\sum_{i=1}^n w_i (1 - d_{ii})}{r' (\hat{\beta}^{OLS,n}(\tilde{Y}, \tilde{X})) r (\hat{\beta}^{OLS,n}(\tilde{Y}, \tilde{X}))} \tag{27}$$

has the *generalized F-distribution with $\sigma^2 \sum_{i=1}^n w_i (d_{ii} - c_{ii})$ and $\sigma^2 \sum_{i=1}^n w_i (1 - d_{ii})$ degrees of freedom and the scale parameters $\sigma^{**} = (\sigma^2 \phi_{n-p+1}, \sigma^2 \phi_{n-p+2}, \dots, \sigma^2 \phi_{n-q})$ and $\sigma^* = (\sigma^2 \psi_1, \sigma^2 \psi_2, \dots, \sigma^2 \psi_{n-p})$.*

4 The simulation study

It is clear that employing the second terms of the right hand side of (24) and (25), we can easily simulate F from (27) (generating e.g. X and Z by p and q dimensional normal d.f. and normally distributed error terms). In simulations (patterns of results are given below) we used $w(u)$ decreasing for $0 \leq u \leq h/100$ from 1 to 0.99, then linearly decreasing to 0.01 at $u = g/100$ and finally decreasing from this value to $w(1) = 0$ (h and g are in the head of tables). For various sample sizes n we have generated $k = 1000$ F 's and then we found $round(n \cdot 0.975)$ and $round(n \cdot 0.995)$ order statistics. This was repeated $\ell = 1000$ times (k and ℓ also in the head of tables) and mean value and mean absolute deviations were computed (the latter are given in the below tables in round parentheses as subindices and multiplied by 100).

Tables

$k = 1000, \ell = 1000, h = 75, g = 95$ and $\alpha = 0.975$.

n	30	40	50	60	70	80
$\hat{F}_{0.995}(n)$	6.355 _(2.614)	6.032 _(2.409)	5.801 _(2.018)	5.679 _(1.953)	5.626 _(1.870)	5.545 _(1.723)
$F_{0.995}(n)$	6.598	6.188	5.974	5.843	5.755	5.691

$k = 1000, \ell = 1000, h = 75, g = 95$ and $\alpha = 0.995$.

n	30	40	50	60	70	80
$\hat{F}_{0.975}(n)$	4.222 _(1.432)	4.073 _(1.238)	3.962 _(1.085)	3.905 _(1.043)	3.879 _(0.964)	3.856 _(0.910)
$F_{0.975}(n)$	4.291	4.106	4.009	3.948	3.906	3.876

$$k = 1000, \ell = 1000, h = 55, g = 85 \text{ and } \alpha = 0.975.$$

n	30	40	50	60	70	80
$\hat{F}_{0.975}(n)$	4.226 _(1.978)	4.058 _(1.698)	3.943 _(1.434)	3.920 _(1.344)	3.861 _(1.197)	3.840 _(1.139)
$F_{0.975}(n)$	4.291	4.106	4.009	3.948	3.906	3.876

5 Conclusions

It is clear that much more extended study (for various levels of contamination, for various degrees of freedom, etc.) will be necessary to give a reliable picture about behaviour of the test. Nevertheless, already these 3 tables indicate that the quantiles of the corresponding distribution are not very far from the Fisher-Snedecor F which can help us for a rough idea about possibility to accept or reject the restricted model. Moreover, time which is necessary for given data (i. e. for given Z and X) to simulate the quantile is (due to the speed of present computational means) rather short, say maximally minutes (the software - written in MATLAB - is available on request).

Acknowledgement

The paper was written with the support of the Czech Science Foundation project 13-01930S *Robust methods for nonstandard situations, their diagnostics and implementations*.

Bibliography

- [1] Atkinson, A. C., M. Riani (2000): *Robust Diagnostic Regression Analysis*. New York: Springer Verlag.
- [2] Bickel, P. J. (1975) *One-step Huber estimates in the linear model*. JASA 70, 428–433.
- [3] Craig, A. T. (1943) *Note on the Independence of Certain Quadratic Forms*. Ann. Math. Statist.14, 107-204.
- [4] Hampel, F. R., E. M. Ronchetti, P. J. Rousseeuw, W. A. Stahel (1986) *Robust Statistics – The Approach Based on Influence Functions*. New York: J.Wiley & Son.
- [5] Hannay, M., W. A. Stahel (2013) *A robust score test for regression*. Book of abstracts of ICORS 2013, 23.
- [6] Hettmansperger, T.P., S. J. Sheather (1992) *A Cautionary Note on the Method of Least Median Squares*. The American Statistician 46, 79–83.
- [7] Judge, G. G., W.E. Griffiths, R. C. Hill, H. LHuÂtkepohl T. C. Lee (1985) *The Theory and Practice of Econometrics*. New York: J.Wiley and Sons.
- [8] Rousseeuw, P.J. (1984) *Least median of square regression*. JASA, 79, 871-880.
- [9] Salibian-Barrera, M., S. Van Aelst, G. Willems (2006) *Principal components analysis based on multivariate MM-estimators with fast and robust bootstrap*. JASA 101, 1198-1211.

- [10] Salibian-Barrera, M., R. H. Zamar (2002) *Bootstrapping robust estimates of regression*. *AS* 30, 556 - 582.
- [11] Stahel, W. A., M. Hannay (2013) *Exploring ideas for testing in robust regression*. Book of abstracts of ICORS 2013, 47.
- [12] Víšek, J. Á. (2000) *A cautionary note on the method of Least Median of Squares reconsidered*. Trans. of the Twelfth Prague Conference, Prague, 254 - 259.
- [13] Víšek, J. Á. (2000) *On the diversity of estimates*. *CSDA* 34, (2000) 67 - 89.
- [14] Víšek, J. Á. (2000) *Regression with high breakdown point*. *ROBUST 2000*, 324 - 356, eds. J. Antoch & G. Dohnal, The Czech Statistical Society 2001.
- [15] Víšek, J. Á. (2005) *Selection of robust method. Numerical examples and results*. Bulletin of the Czech Econometric Society, (2005) 11, 1 - 58.
- [16] Víšek, J. Á. (2006) *The least trimmed squares. Consistency, \sqrt{n} -consistency, asymptotic normality and Bahadur representation*. *Kybernetika* (2006), 42, 1 - 36, 181 - 202, 203 - 224.
- [17] Víšek, J. Á. (2011) *Consistency of the least weighted squares under heteroscedasticity*. *Kybernetika* 47 , 179-206.
- [18] Víšek, J. Á. (2014) *Diagnostics of robust identification of model*. Submitted¹⁵ to Advances in Data Mining and Robust Statistics, special volume of CSDA.
- [19] Wooldridge, J. M. (2006) *Introductory Econometrics. A Modern Approach*. MIT Press, Cambridge, Massachusetts, second edition 2009.
- [20] Yohai, V. J. (1987): High breakdown-point and high efficiency robust estimates for regression. *AS* 15, 642 - 656.

¹⁵A very first version was presented on ICORS 2011 (<http://samba.fsv.cuni.cz/~visekjan/ICORS/icors2011/>), an improved one on ERCIM 2013.

Conditional quantile estimation using optimal quantization: a numerical study

Isabelle Charlier, *Université Libre de Bruxelles and Université de Bordeaux*, ischarli@ulb.ac.be

Davy Paindaveine, *Université Libre de Bruxelles*, dpaindav@ulb.ac.be

Jérôme Saracco, *Université de Bordeaux*, jerome.saracco@math.u-bordeaux1.fr

Abstract. We construct a nonparametric estimator of conditional quantiles of Y given $X = x$ using optimal quantization. Conditional quantiles are particularly of interest when the conditional mean is not representative of the impact of the covariable X on the dependent variable Y . L_p -norm optimal quantization is a discretizing method used since the 1950's in engineering. It allows to construct the best approximation of a continuous law with a discrete law with support of size N . The aim of this work is then to use optimal quantization to construct conditional quantile estimators. We study the convergence of the approximation ($N \rightarrow \infty$) and the consistency of the resulting estimator for this fixed- N approximation. This estimator was implemented in R in order to evaluate the numerical behavior and to compare it with existing methods.

Keywords. Nonparametric estimation, Conditional quantile, Optimal quantization.

1 Introduction

Quantile regression allows to assess the impact of a covariable X on a (scalar) response variable Y and is an alternative to standard regression. It is particularly of interest when the mean does not provide an enough satisfactory picture of the distribution. We then get a more complete picture of the conditional distribution if we consider the conditional quantile functions

$$x \mapsto q_\alpha(x) = \inf\{y \in \mathbb{R} : F(y|x) \geq \alpha\}, \quad (1)$$

for various $\alpha \in (0, 1)$, where $F(\cdot|x)$ stands for the conditional distribution of Y given $X = x$. They are equivalently defined by solving the following optimization problem:

$$q_\alpha(x) = \arg \min_{a \in \mathbb{R}} E[\rho_\alpha(Y - a)|X = x], \quad (2)$$

where $\rho_\alpha(z) = \alpha z \mathbb{1}_{[z \geq 0]} - (1 - \alpha)z \mathbb{1}_{[z < 0]}$ is called the *check function*.

An important application of conditional quantiles is that they provide reference hypersurfaces (curves when $d = 1$) if we consider the quantile functions $x \mapsto q_\alpha(x)$ when x varies, and confident intervals of the form $I_\alpha = [q_\alpha(x), q_{1-\alpha}(x)]$ when x is fixed, which are widely used in many domains, as medicine, economics or lifetime analysis.

There exist many approaches to define conditional quantile estimators since the literature on quantile regression became really large in recent years. For example, [1] focuses on nearest-neighbor estimators of a conditional quantile while local linear estimator is investigated in [6].

We define in [2] a new estimator of conditional quantiles based on optimal quantization and we perform a numerical study of this estimator in [3]. Optimal quantization is a tool allowing to discretize any continuous distribution of a random vector X . It then provides an approximation of X by a discrete random vector with support of size N . This approximation is obtained by projecting X on a set of N points, called a grid. This grid is chosen in such a way that the L_p -norm difference between X and its discretized version is minimal. The reader can refer to [4, 5] for more details on optimal quantization.

We will first briefly recall in Section 2 the general idea of our method and explain the different steps in the construction of our estimator. Then, in Section 3, we provide a numerical comparison of our estimator with three alternative quantiles estimators.

2 Conditional quantile estimation through optimal quantization

In this section, we first explain the general idea of the construction of our estimator introduced in [2]. We then detail point by point this construction that is implemented in a R package called `QuantifQuantile` (available on the CRAN).

In the sequel, we denote by Y a real random variable and X a d -dimensional random vector. We define an estimator of conditional quantiles thanks to L_p -norm quantization. The idea is to replace X in (2) by a discrete version, obtained by projecting X on an optimal quantization grid. We then take an empirical version of this approximation. Let us specify this construction.

Assume that X belongs to L_p , *i.e.* $\|X\|_p := E[|X|^p]^{1/p} < \infty$. Let $\gamma^N \in (\mathbb{R}^d)^N$ a set of N points of \mathbb{R}^d , called a *grid*. We approximate X by the projection of X onto this grid, that we denote $\tilde{X}^{\gamma^N} := \text{Proj}_{\gamma^N}(X)$. Obviously, the quality of this approximation depends hugely on the choice of the grid. We then choose γ^N as the grid minimizing the *quantization error* $\|X - \tilde{X}^{\gamma^N}\|_p$. Classic result in quantization ensures the existence (but not the unicity) of such grid under the assumption that the law of X does not charge any hyperplanes. We will denote in the sequel \tilde{X}^N the projection of X onto an optimal grid. In practice, an optimal grid is constructed using a *stochastic gradient algorithm*. This algorithm is detailed further. The reader may refer to [4] for more details on the concept of quantization. We then define

$$\tilde{q}_\alpha^N(x) := \arg \min_{a \in \mathbb{R}} E[\rho_\alpha(Y - a) | \tilde{X}^N = \tilde{x}], \quad (3)$$

where \tilde{x} is the projection of x onto γ^N .

Let us now assume that we have n independent copies $(X'_1, Y_1)', \dots, (X'_n, Y_n)'$. We define an estimator of conditional quantiles by taking an empirical version of this approximation, denoted $\hat{q}_\alpha^{N,n}(x)$. Its construction is provided in the sequel.

We derived the following theorems for the convergence of $\tilde{q}_\alpha^N(x)$ when $N \rightarrow \infty$ and of $\hat{q}_\alpha^{N,n}(x)$ when $n \rightarrow \infty$ and N fixed. We need the following assumptions.

ASSUMPTION (A) (i) The random vector (X, Y) is generated through $Y = m(X, \varepsilon)$, where the d -dimensional covariate vector X and the error ε are mutually independent; (ii) the link function $(x, z) \mapsto m(x, z)$ is of the form $m_1(x) + m_2(x)z$, where the functions $m_1(\cdot) : \mathbb{R}^d \rightarrow \mathbb{R}$ and $m_2(\cdot) : \mathbb{R}^d \rightarrow \mathbb{R}_0^+$ are Lipschitz functions; (iii) $\|X\|_p < \infty$ and $\|\varepsilon\|_p < \infty$; (iv) the distribution of X does not charge any hyperplanes.

ASSUMPTION (B) (i) The support S_X of P_X is compact; (ii) ε admits a continuous density $f^\varepsilon : \mathbb{R} \rightarrow \mathbb{R}_0^+$ (with respect to the Lebesgue measure on \mathbb{R}).

To obtain rates of convergence, we will need the following reinforcement of Assumption (A).

ASSUMPTION (A') Same as Assumption (A), but with (iii) replaced by (iii)' there exists $\delta > 0$ such that $\|X\|_{p+\delta} < \infty$, and $\|\varepsilon\|_p < \infty$.

ASSUMPTION (C) P_X is continuous and has a compact support.

Under these assumptions, the underlying curve m is quite smooth, which avoids possible peaks or jumps.

Theorem 2.1. Fix $\alpha \in (0, 1)$. Then (i) under Assumptions (A)-(B),

$$\|\tilde{q}_\alpha^N(X) - q_\alpha(X)\|_p \leq 2 \sqrt{\max\left(\frac{\alpha}{1-\alpha}, \frac{1-\alpha}{\alpha}\right)} [m]_{\text{Lip}}^{1/2} \|L^N(X)\|_p^{1/2} \|X - \tilde{X}^N\|_p^{1/2},$$

for N sufficiently large, where $(L^N(X))$ is a sequence of X -measurable random variables that is bounded in L_p ; (ii) under Assumptions (A')-(B),

$$\|\tilde{q}_\alpha^N(X) - q_\alpha(X)\|_p = O(N^{-1/2d}), \quad \text{as } N \rightarrow \infty.$$

Theorem 2.2. Fix $\alpha \in (0, 1)$. Then, under Assumptions (A)-(B),

$$\sup_{x \in S_X} |\tilde{q}_\alpha^N(x) - q_\alpha(x)| \rightarrow 0, \quad \text{as } N \rightarrow \infty.$$

Theorem 2.3. Fix $\alpha \in (0, 1)$, $x \in S_X$ and $N \in \mathbb{N}_0$. Then, under Assumptions (A), (B)(i), and (C), we have that, as $n \rightarrow \infty$,

$$|\hat{q}_\alpha^{N,n}(x) - \tilde{q}_\alpha^N(x)| \rightarrow 0,$$

in probability, provided that quantization is based on $p = 2$.

More details on these theorems and their proofs can be found in [2].

We will now explain step by step the construction of $\hat{q}_\alpha^{N,n}(x)$. We will then complete this section with an illustration on some dataset.

Determining an optimal N -grid

Since the starting idea of our method consists in replacing X with a discrete version with support of size N , the first step is naturally dedicated to the choice of an optimal N -grid for X , with N fixed. Since no theoretical quantization result provides such a grid, the only way at our disposal

to get it is to use a stochastic gradient algorithm. Starting from an initial grid denoted $\hat{\gamma}^{N,0}$, we update the grid at step $t - 1$ thanks to the observation X_t , playing the role of stimuli. We then obtain the grid at step t , for $t = 1, \dots, n$. After n steps, we thus get a grid $\hat{\gamma}^{N,n}$ considered as optimal. Let us make it more precise.

Let $(\delta_t), t \in \mathbb{N}_0$, be a deterministic sequence in $(0, 1)$ such that

$$\sum_t \delta_t = \infty \quad \text{and} \quad \sum_t \delta_t^2 < \infty.$$

For N fixed, the algorithm works as follows.

Algorithm 2.1.

For $t = 1 \dots, n$,

Step 0 The initial grid $\hat{\gamma}^{N,0}$ in $(\mathbb{R}^d)^N$ is chosen by sampling randomly among the X_i 's without replacement.

Step t The grid at step t is defined recursively as

$$\hat{\gamma}_i^{N,t} = \begin{cases} \hat{\gamma}_i^{N,t-1} - \delta_t |\hat{\gamma}_i^{N,t-1} - X_t|^{p-1} \frac{\hat{\gamma}_i^{N,t-1} - X_t}{|\hat{\gamma}_i^{N,t-1} - X_t|} & \text{if } \text{Proj}_{\hat{\gamma}^{N,t-1}}(X_t) = \hat{\gamma}_i^{N,t-1} \\ \hat{\gamma}_i^{N,t-1} & \text{otherwise} \end{cases},$$

where $\hat{\gamma}_i^{N,t} \in \mathbb{R}^d$ denotes the i th component of $\hat{\gamma}^{N,t}$, $i = 1, \dots, N$.

We observe that only one point of the grid at step $t - 1$ moves at each step t : the one on which the stimuli X_t is projected.

The resulting grid $\hat{\gamma}^{N,n}$ allows thus to quantize X : we define $\widehat{X}^{N,n} = \text{Proj}_{\hat{\gamma}^{N,n}} X$. This is important to point out that this quantization step provides a grid that is chosen independently of Y . Thus, the link function m does not play any role in this step.

Estimating conditional quantiles

As above-mentioned, an approximation of conditional quantiles is defined by replacing X by its projection on the optimal N -grid in the definition. An estimator is then constructed by taking an empirical version of this approximation, as follows :

Algorithm 2.2.

Let $(X'_1, Y'_1), \dots, (X'_n, Y'_n)'$ be n independent copies of (X, Y) .

Step 1 We project each X_i on the grid $\hat{\gamma}^{N,n}$ and we write $\widehat{X}_i^{N,n} = \text{Proj}_{\hat{\gamma}^{N,n}}(X_i)$. We then work with the projected sample $\{(\widehat{X}_i^{N,n}, Y_i)'\}_{i=1, \dots, n}$.

Step 2 The conditional quantiles are then estimated by

$$\widehat{q}_\alpha^{N,n}(x) = \arg \min_{a \in \mathbb{R}} \sum_{i=1}^n \rho_\alpha(Y_i - a) \mathbb{1}_{[\widehat{X}_i^{N,n} = \hat{x}^N]},$$

where $\hat{x}^N = \text{Proj}_{\hat{\gamma}^{N,n}}(x)$. In practice, $\widehat{q}_\alpha^{N,n}(x)$ is simply evaluated as the sample α -quantile of the Y_i 's whose corresponding X_i admits \hat{x}^N as projection onto $\hat{\gamma}^{N,n}$.

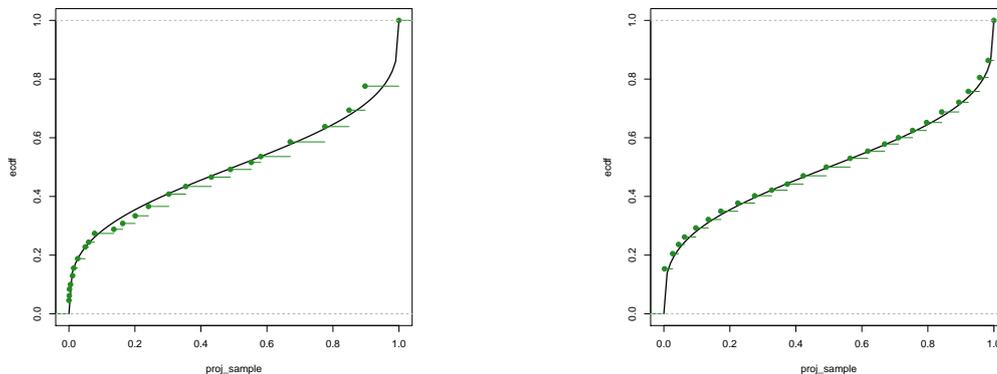


Figure 1: Comparison of population (black) and grid-projected sample (green) cdf for samples of size 500 (left) and 5000 (right) generated from a beta distribution (for the green one, the sample was projected onto an optimal quantization grid of size $N = 25$).

The first step of this algorithm is illustrated in Figure 1. We generate observations with law $\text{Beta}(0.3,0.3)$ and we consider $N = 25$. Using Algorithm 2.1, an optimal grid is constructed and we project the sample onto this grid. The left graph represents the grid-projected sample cumulative distribution function (cdf) in green and the population one in black for a sample size $n = 500$. The right one is similar with $n = 5000$. We observe that the grid-projected sample versions fit very well the population ones (better and better when n increases).

Nevertheless, the grid provided by the stochastic gradient algorithm may be a poor approximation of the optimal one when the sample size is small (when n is equal to 300 or less). Indeed, $\hat{\gamma}^{N,n}$ is constructed after n iterations. As the choice of the grid is the basis in the construction of our estimator, it has a major impact on the resulting reference curves that are not smooth. For this reason, we use bootstrap to introduce a more appropriate conditional quantile estimator.

For some integer B , we generate B samples of size n from our original sample $\{(X_i, Y_i)\}_{i=1,\dots,n}$ with replacement. Each bootstrap sample is then used as stimuli to construct a grid by performing the stochastic gradient algorithm. Thanks to these B grids, we get B estimations of $q_\alpha(x)$ thanks to Algorithm 2.2, that we denote $\hat{q}_\alpha^{(1)}(x), \dots, \hat{q}_\alpha^{(B)}(x)$. The bootstrap version of our estimator is then defined as:

$$\bar{q}_{\alpha,B}^{N,n}(x) = \frac{1}{B} \sum_{b=1}^B \hat{q}_\alpha^{(b)}(x). \tag{4}$$

We usually take $B = 50$ when X is univariate.

Figure 2 represents the curves of estimated conditional quantiles for a sample of size $n = 500$. The left panel of Figure 2 is obtained without bootstrap and the right one with bootstrap. This bootstrap version provides clearly smoother curves.

Selecting the number N of quantizers

The choice of the number N of quantizers is crucial: for too small N , the curves show a large bias and for too large N , the variability is important but the bias smaller.

We propose a data driven selection criterion for N . As explained in [3], we first investigate the MSE (Mean Squared Error) as a function of N (by taking some suitable family of possible

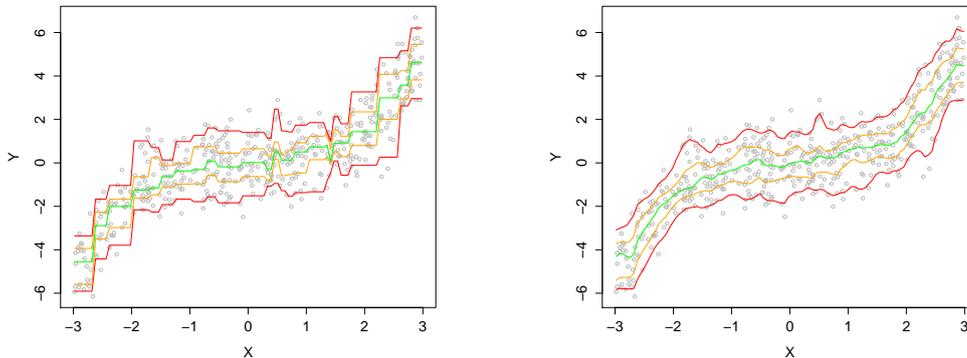


Figure 2: For $n = 500$, $X \sim U(-3, 3)$, $Y = X^3/5 + \epsilon$, with $\epsilon \sim \mathcal{N}(0, 1)$ independent of X , the curves of estimated conditional quantiles, without and with bootstrap respectively. They are obtained with $\alpha = 0.05, 0.25, 0.5, 0.75$ and 0.95 respectively (upwards). Left: Without bootstrap, Right: $B = 50$

values for N). These curves are actually convex and we choose an optimal value for N as the “arg min” of $\text{MSE}(N)$. Of course, the MSE is calculated using the true conditional quantiles. We then propose a bootstrap estimate of the MSE that only uses the observations. We observe that the corresponding curves are convex and minimized for a N close the optimal one for the true MSE (see [3] for more details). Let us specify this criterion.

Let $\{x_1, \dots, x_{\mathcal{N}_x}\}$ be a set of \mathcal{N}_x deterministic points for which we want to estimate $q_\alpha(x)$ (generally equispaced on the support of X). We actually generate $B + \tilde{B}$ bootstrap samples of size n from the initial sample. The first B bootstrap samples allows to construct $\tilde{q}_{\alpha, B}^{N, n}(x_j)$ as above explained. The last \tilde{B} are used to calculate \tilde{B} estimations of $q_\alpha(x_j)$, that we denote $\hat{q}_\alpha^{(\tilde{b})}(x_j)$, for $\tilde{b} = 1 \dots, \tilde{B}$. The true conditional quantiles are replaced by $\hat{q}_\alpha^{(\tilde{b})}(x_j)$ in the expression of the MSE, and we take the mean of these \tilde{B} versions. More precisely, we define

$$\widehat{\text{MSE}}(N) = \frac{1}{\mathcal{N}_x} \sum_{j=1}^{\mathcal{N}_x} \left(\frac{1}{\tilde{B}} \sum_{\tilde{b}=1}^{\tilde{B}} (\hat{q}_\alpha^{(\tilde{b})}(x_j) - \tilde{q}_{\alpha, B}^{N, n}(x_j))^2 \right). \quad (5)$$

We then select the optimal number N of quantizers as

$$\hat{N}^* = \arg \min_{N \in \mathfrak{N}} \widehat{\text{MSE}}(N), \quad (6)$$

where \mathfrak{N} denotes a grid of values for N chosen according to the sample size of the considered dataset.

3 Comparison with alternative conditional quantile estimators

We explained in the previous section the construction of our estimator and we proposed a selection criterion for the tuning parameter N . We now recall three well-known conditional quantile estimators and the selection criteria for their own tuning parameters. We then summarize the boxplot comparison realized in [3] and specify which estimators seem preferable in each situation.

The k nearest-neighbor is introduced in [1]. This estimator of $q_\alpha(x)$ is defined as follows. Let $X_i^* = |X_i - x|$ for $i = 1, \dots, n$ and let $X_{n1}^* < \dots < X_{nn}^*$ denote the order statistics of X_1^*, \dots, X_n^* and Y_{n1}, \dots, Y_{nn} the induced order statistics of $(X_1^*, Y_1), \dots, (X_n^*, Y_n)$, i.e. $Y_{ni} = Y_j$ if $X_{ni}^* = X_j^*$. For any positive integer $k \leq n$, the k nearest-neighbor estimator $\hat{q}_\alpha^k(x) = \hat{q}_\alpha^{k,n}(x)$ is the $[k\alpha]$ th order statistics of Y_{n1}, \dots, Y_{nn} . The idea is to select the k points of the data such that their X 's are the nearest of x , whence the name, and to calculate the quantile of order α of their Y 's. Of course, k plays the role of tuning parameter and must be specified. Since we did not find in the literature an efficient method to select k only based on the data, we choose k by taking it minimizing the mean squared error among an set of values for k , that we will denote k^* .

The kernel weighted local linear estimator introduced by [6] is the second competitor. This estimator is defined as $\hat{q}_\alpha^{YJ}(x) = \hat{a}$, with

$$(\hat{a}, \hat{b}) = \arg \min_{(a,b) \in \mathbb{R} \times \mathbb{R}} \sum_{i=1}^n \rho_\alpha(Y_i - a - b(X_i - x)) K\left(\frac{x - X_i}{h}\right),$$

where K is a kernel function, chosen as the standard normal density, and where h is the bandwidth. We choose h according to α as

$$h_\alpha = h_{\text{mean}} \left(\frac{\alpha(1 - \alpha)}{\varphi(\Phi^{-1}(\alpha))^2} \right),$$

where φ and Φ are respectively the standard normal density and distribution functions, and where h_{mean} is the optimal choice of h for regression mean estimation, selected thanks to a cross-validation criteria. We also consider the local constant version of this estimator. More precisely, it is defined as $\hat{q}_\alpha^{YJc}(x) = \hat{a}$, with

$$\hat{a} = \arg \min_{a \in \mathbb{R}} \sum_{i=1}^n \rho_\alpha(Y_i - a) K\left(\frac{x - X_i}{h}\right),$$

and where the kernel function and the bandwidth are chosen as in the local linear case.

Notice that an important point in conditional quantile estimation is the choice of the observations X_i that will be taken into account when estimating $q_\alpha(x)$. We see that $\hat{q}_\alpha^{YJ}(x)$ and $\hat{q}_\alpha^{YJc}(x)$ choose it thanks to some bandwidth while $\hat{q}_\alpha^k(x)$ is constructed using the k observations whose X -part is the closest to x . Our method is based on the observations whose X -part is projected on the same point of the grid as x (we call the set of such points a quantization cell). The main advantage of our method is then that the number of observations used to estimate $q_\alpha(x)$ is adaptive with x . The choice of a bandwidth is felt to be interesting when the observations X are quite uniformly distributed on the support of X but it may be disadvantageous when the density of the points is smaller in some regions of the support.

We then compare our estimator with these competitors. We consider different models (homoscedastic and heteroscedastic) and sample sizes ($n = 300, 500$ and 1000). For each of them, we generate 500 samples and we calculate the estimates $\bar{q}_{\alpha,B}^{N,n}(x)$, $\hat{q}_\alpha^{YJ}(x)$, $\hat{q}_\alpha^{YJc}(x)$ and $\hat{q}_\alpha^k(x)$. We then realize the boxplots of the mean squared error (MSE) according to each estimator. We generally observe that $\bar{q}_{\alpha,B}^{N,n}(x)$ generally outperforms its competitors when the covariate is not uniformly distributed. In case of uniformly distributed X , $\hat{q}_\alpha^{YJ}(x)$ is often better. We illustrate it in Figure 3 where we generate 500 samples of size $n = 300$ with $X = 6Z - 3$, with $Z \sim \text{Beta}(0.3, 0.3)$ and $Y = X^3/5 + \varepsilon$, where ε is a normal error term independent of X . We observe that $\bar{q}_{\alpha,B}^{N,n}(x)$ provides the smallest MSE, followed by $\hat{q}_\alpha^k(x)$. More details on this comparison study can be found in [3].

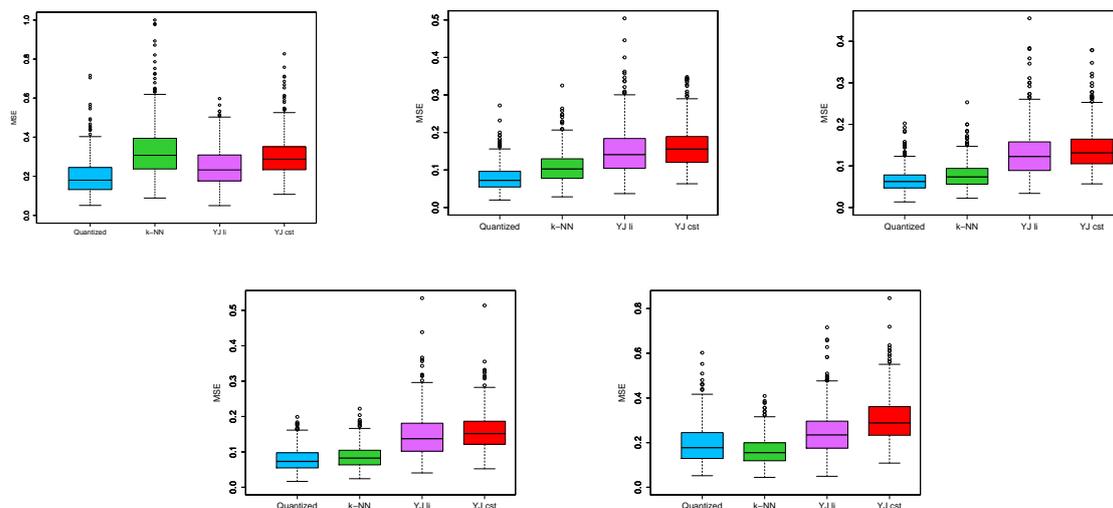


Figure 3: For 500 replications of sample of size $n = 300$ from model $Y = \frac{1}{5}X^3 + \varepsilon$, the boxplots of the MSE in the estimation of the conditional quantile curves: in blue, with $\hat{q}_{\alpha, B}^{N, n}(x)$, in green, with $\hat{q}_{\alpha}^k(x)$, in purple, with $\hat{q}_{\alpha}^{YJ}(x)$ and in red, with $\hat{q}_{\alpha}^{YJc}(x)$. Left to right: $\alpha = 0.05, 0.25, 0.5, 0.75, 0.95$

Bibliography

- [1] Bhattacharya, P.K. and Gangopadhyay, A.K. (1990) *Kernel and nearest-neighbor estimation of a conditional quantile*. *Annals of Statistics*, **7**(3), 1400–1414.
- [2] Charlier, I., Paindaveine, D. and Saracco, J. (2014) *Conditional quantile estimation through optimal quantization*. Submitted.
- [3] Charlier, I., Paindaveine, D. and Saracco, J. (2014) *Numerical study of a conditional quantile estimator based on optimal quantization*. Manuscript in preparation.
- [4] Pagès, G. (1998) *A space quantization method for numerical integration*. *Journal of Computational and Applied Mathematics*, **89**(1), 1–38.
- [5] Pagès, G. and Printems, J. (2003) *Optimal quadratic quantization for numerics: the Gaussian case*. *Monte Carlo Methods and Applications*, **9**(2), 135–165.
- [6] Yu, K. and Jones, M.C. (1998) *Local linear quantile regression*. *Journal of the American Statistical Association*, **93**(441), 228–237.

A combined nonparametric test for seasonal unit roots

Robert M. Kunst, *Institute for Advanced Studies Vienna and University of Vienna*, kunst@ihs.ac.at

Abstract. Nonparametric unit-root tests are a useful addendum to the toolbox of time-series analysis. They tend to trade off power for enhanced robustness features. We consider combinations of variants of the RURS (seasonal range unit roots) test statistic and of the level-crossings count. This combination exploits two main characteristics of seasonal unit-root models, the range expansion typical of integrated processes and the low frequency of changes among main seasonal shapes. The combination succeeds in achieving power gains over the component tests. Simulations explore the finite-sample behavior relative to traditional parametric tests.

Keywords. Seasonality, nonparametric tests, visualization, time series.

1 Introduction

The current literature on seasonal time series (see [9] or [6], for example) ascribes the origin of the basic discrimination problem among conflicting paradigms for the generation of seasonal features to [10].

The three main model worlds of concern are: (a) deterministic seasonal variation, as customarily expressed via seasonal dummy variables; (b) seasonal unit roots and seasonal integration; (c) stochastic stationary cyclical variation. A distinctive feature among these concepts is their implication for long-run forecasts of seasonal patterns. Deterministic seasonal patterns in an otherwise stationary environment entail that the sample average of seasonal shapes is the appropriate long-run predictor for future shapes. Seasonal integration emphasizes the importance of the most recent pattern as a shape predictor, even though this prediction will face increasing uncertainty at increasing horizon and persistent shape changes are to be expected. Stationary cyclical variation implies trivial predictions at longer horizons.

The most important statistical tools for discriminating among these main concepts of seasonal time-series generators evolved in the 1990s: the HEGY test by [11], the CH test by [5], and some further contributions that are conveniently summarized by [9]. These tests are parametric, build on Gaussian likelihoods, and optimize power properties for specific designs. By contrast, nonparametric tests aim at increased robustness at the price of reduced power.

Variance-ratio tests for seasonal unit roots were considered by [15] who generalized the testing concept by [3] to the seasonal case. Whereas these tests achieve additional robustness, they cannot be viewed as narrow-sense nonparametric. [12] and [13] based their RURS (RUR seasonal) test on the nonparametric RUR (range unit-root) test by [2] which uses a count of new records in time series as a unit-root criterion. The test suggested here combines a variant of this RURS test with another nonparametric unit-root testing idea that was investigated by [4] and [8] and relies on counting zero crossings.

We provide an additional motivation for our selection of component tests in our new test by first presenting the idea of jittered seasonal phase plots. These plots are constructed as follows. First, the information on seasonal shapes is condensed into classes, and then the transition patterns between these classes are recorded.

The paper is organized as follows. Section 2 introduces jittered seasonal phase plots as a visualization tool. Section 3 considers the nonparametric combination test. Section 4 applies the methods to exemplary time series. Section 5 concludes.

2 Jittered seasonal phase plots

Generally, we restrict attention to the quarterly case. In principle, the monthly case (or any other seasonal aspect) can be handled analogously. Due to the large number of possible seasonal shapes, however, the visualization tools outlined below are less attractive for monthly data.

Initially, consider the quarterly Gaussian seasonal random walk (SRW)

$$x_t = x_{t-4} + \varepsilon_t, t \geq 0, \quad x_t = 0, -3 \leq t \leq 0, \quad (1)$$

with Gaussian iid increments ε_t . A characteristic feature of such *seasonal unit-root processes* are the infrequent but persistent changes in the rank position of quarters. If the series is plotted by quarters—the spring series, the summer series etc.—the four seasonal curves cross each other rarely. If they do, however, they do not tend to cross back into their previous ranking order.

The idea of counting intersections of quarterly plots in order to obtain a non-parametric statistic for discriminating among the main seasonal generating models was mentioned by [12]. Each crossing represents a change of the qualitative seasonal pattern, in the sense that, for example, stronger sales of a product during the spring season give way to higher sales in summer.

In the following, we classify the shapes of four consecutive quarterly observations within a year into eight possible qualitative seasonal patterns $0 \leq m \leq 7$, according to whether the value rises or falls between two adjacent quarters. We code the cases as three-digit binary numbers, using ‘1’ for an increase between quarters and ‘0’ for a decrease. Note that the direction from the last quarter of a year to the first quarter of the next year is ignored in the discretization. This follows the idea that seasonality should be viewed independently from the year-to-year trend. Thus, the 111 pattern ($m = 7$) is not subdivided into cases with a strong slump at the beginning of the year and cases with a persistent upward movement.

Discretization of the sample space into eight classes entails a considerable loss in information and can only be justified if it assists in the discrimination problem of concern. Unfortunately, direct connected phase plots do not turn out to be very useful for the classification of the generating seasonal process, even with large samples.

A main problem in the simple phase plot is that it does not show the population within the classes. One suggestion would be to randomize the observations in an interval such as

$[m - 0.4, m + 0.4]$, which would correspond to the technique of *jittering*. A more sophisticated visualization may distribute the observations that belong to class m according to their ‘depth’ within the class. Given that an observation is classified to m according to increases or decreases in specific quarters, it is said to be ‘deep’ in the class when these increases or decreases are large, while almost constant patterns can be seen as ‘shallow’. There are three quarter-to-quarter movements, and an observation may be deep regarding quarter 1-2 and shallow in other quarters.

We decided to define the positions by maximum relative depth, calculated from maximum absolute inter-quarter increases or decreases. We opted to position the shallow points in the center of the interval and the deep points at the boundaries. Thus, classes are usually entered in the center as fresh and shallow members, and after an extended period in the class the boundaries are approached. For later reference, we denote the position within the bin as \tilde{x} , such that $\tilde{x} = 0$ corresponds to its center.

There is no coercive convention for the left and right parts of the interval, $[m - 0.4, m)$ and $(m, m + 0.4]$. We decided for uniform randomization, i.e. jittering in the literal sense. For seasonal unit-root processes, deep observations tend to be followed by deep ones, and this randomization tends to generate a St. Andrew’s cross or saltire, a pattern that sends a clear message to the observer. For an SRW, the result of this procedure is shown in the left part of Figure 1.

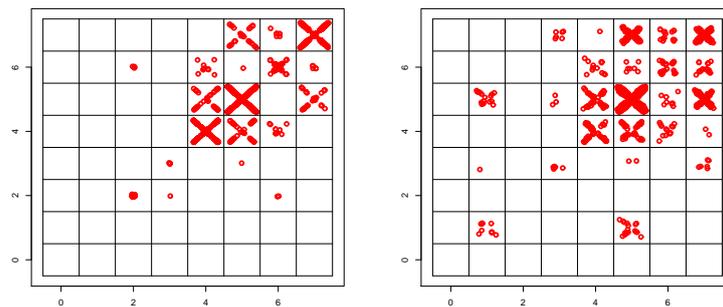


Figure 1: Jittered phase diagrams for pattern classes. Generating model for the left plot is a quarterly Gaussian seasonal random walk, for the right plot a process with deterministic seasonal variation. In both cases, 40,000 observations have been generated.

Whereas unit-root seasonality generates the hitherto highlighted features, i.e. rare transitions between bins and persistence within some bins, many real-world seasonal cycles, for example temperature series, support *deterministic models* (see the example in Section 4). A simple generating model for such deterministic seasonal variation is a stable ARMA model superseded with a repetitive cycle expressed via seasonal dummy variables. For exposition, we consider the model

$$x_t = 0.4x_{t-4} + \sum_{j=1}^4 d_j \delta_{j,t} + \varepsilon_t, \quad (2)$$

with standard Gaussian errors and $\delta_{j,t}$ denoting seasonal dummy constants. For the deterministic pattern, we impose $(d_1, \dots, d_4) = (0, 8, 3, 10)$. Such deterministic seasonal processes

are characterized by an inherent tendency to switch back to the old pattern regime after any crossings.

The impression is confirmed in a jittered phase plot shown in the right portion of Figure 1. There is a strong preference for one class, and occasional visits to other classes remain episodic.

3 A nonparametric test for seasonal unit roots

We first note that the classification decision based on the jittered plots mainly depends on two features: (i) the *transition frequency* across shape classes and (ii) the *precision of the saltire* shapes. Our presentation continues to be restricted to the quarterly case, although it is straight forward to generalize to other periodicities, unlike the phase-plot visualization.

Concerning the transition frequency, it is convenient to start from the case that the data-generating process is a seasonal random walk. The event of a shape transition, for example

$$x_{1,t} > x_{2,t}, x_{1,t-1} < x_{2,t-1}, \quad (3)$$

with $x_{i,j}$ denoting the quarter i in year j , clearly is equivalent to

$$x_{1,t} - x_{2,t} > 0, x_{1,t-1} - x_{2,t-1} < 0. \quad (4)$$

In an SRW, the quarters represent independent random walks. The difference between quarters is then a random walk itself, so the above event is a zero crossing for the random walk $x_{1,t} - x_{2,t}$.

Some facts on the distribution of zero crossings in random walks are known from the literature ([4], [8]). In particular, [4] showed that the modified zero crossings count

$$K_T^*(0) = \frac{\hat{\sigma}}{\widehat{\text{MAD}}} T^{-0.5} \sum_{t=1}^T I(X_{t-1} \leq 0, X_t > 0) + I(X_{t-1} > 0, X_t \leq 0) \quad (5)$$

is asymptotically distributed as $|N(0, 1)|$. Here, $\hat{\sigma}$ denotes an estimate of the standard error of the increments ΔX_t , whereas $\widehat{\text{MAD}}$ is an estimate of their absolute first moments. These two correction factors are suggested to be formed empirically as

$$\hat{\sigma} = \sqrt{T^{-1} \sum_{t=1}^T (\Delta X_t)^2}, \quad \widehat{\text{MAD}} = T^{-1} \sum_{t=1}^T |\Delta X_t|. \quad (6)$$

[8] (GS) then generalized this result to trend-corrected random walks, for which the modified crossings statistic converges to a Rayleigh distribution. In particular, however, GS showed that replacing the sample variance in $K_T^*(0)$ by a long-range variance in the vein of variance-ratio tests or the popular unit-root test by [14] succeeds in making the test robust to autocorrelation in the increments under the null of a unit root. The main argument in their proof relies on a result by [1]. Thus, the test that was strictly valid only for random walks in the version of [4] becomes a test for the null of a first-order integrated process, in symbols $I(1)$. Substituting the quarter-to-quarter differences $x_{i+1,t} - x_{i,t}$, $i = 1, 2, 3$, for the random walk X_t yields a seasonal variant of the crossings count

$$\zeta_1 = \frac{\hat{\sigma}}{\widehat{\text{MAD}}} T^{-0.5} \sum_{t=2}^T I(\Xi(x_{1,t}, \dots, x_{4,t}) \neq \Xi(x_{1,t-1}, \dots, x_{4,t-1})), \quad (7)$$

with $t = 1, \dots, T$ denoting the years in the sample and $\Xi(\cdot)$ the function that classifies four observations within a year into one of the eight seasonal shapes. In the modified statistic ζ_1 , the estimate $\hat{\sigma}$ is replaced by a long-run standard deviation estimate.

Concerning the precision of the saltire, we form a second nonparametric test statistic ζ_2 from the median distance between observations in the (i, j) bin and the saltire form. We use the Euclidean distance of the absolute point $(|\tilde{x}_{t-1}|, |\tilde{x}_t|)$ and the 45 degree line, if \tilde{x} denotes a properly normalized observation relative to the center of the corresponding bin, i.e. the maximum difference among adjacent quarters within a year t . This distance, in turn, equals $2^{-0.5}\Delta|\tilde{x}_t|$ in the diagonal (i, i) bins, and its average should converge to an absolute moment of a distribution of increments.

We note, however, that the bins have been scaled to minima and maxima. Such maxima are known to expand at the rate of $T^{0.5}$ for random walks and at a much slower rate of $\log T$ for stationary Gaussian variables. Thus, the average of the increments in a scaled world approaches zero at a rate of $T^{-0.5}$ for random walks, while its properties depend on characteristics of the data-generation process including error distributions for stationary variables. Accordingly, the nonparametric test statistic ζ_2 is defined as $T^{0.5}$ multiplied by the median distance of observations and the saltire, i.e.

$$\zeta_2 = T^{0.5} \text{med} \{ |\tilde{x}_t| - |\tilde{x}_{t-1}| \} / \sqrt{2} \quad (8)$$

We opt for the median rather than the mean for the sake of distributional robustness. A closely related nonparametric unit-root test statistic was suggested by [2] (AES) who used the count of new records in the time series. We note that the range is proportional to the number of records.

Unfortunately, a robustification step comparable to the correction by GS for the original suggestion by [4] is not available for the AES test. This property is rooted deeply in the construction principle of the tests. The level-crossings count is sensitive to moments of the generating distribution, which are then adjusted for by a correction factor, and this factor is in turn robustified in the GS version. In the AES test, by contrast, the adjustment term cancels out, and the distribution of the test statistic is independent of the moment properties of the generating law for pure random walks.

It pays to reconsider the shapes that we encountered by simulation in Section 2 in the light of this discussion. For an SRW, the saltire shape is approximated as the sample size increases, with the average distance from the saltire decreasing. For a non-seasonal random walk, both the horizontal and vertical directions of the bins expand at the rate $T^{0.5}$, and the bins are densely filled. For white noise, both directions expand slowly, and the bins are filled in a circular fashion, with points in the corners remaining rare.

In various simulation experiments using generating processes with and without seasonal unit roots, we found the discriminatory power of the two test statistics to be quite satisfactory. We also found that dependence between ζ_1 and ζ_2 is not too strong, particularly under the non-unit root alternative, so it pays to use both of them jointly.

Briefly consider a time-series plot of a trajectory of, for example, a random walk. The statistic ζ_1 increases at every zero crossing, and ζ_2 changes at every new extremum. This may insinuate that random-walk realizations with many crossings and thus a large value of ζ_1 have a slower expansion rate and thus larger ζ_2 . For very small samples, the two test statistics are indeed correlated by construction. This strong dependence soon disappears as the sample size increases.

The statistics ζ_1 and ζ_2 process different information. Whereas good or even optimal combinations of individual tests can be constructed by the Bonferroni principle, we here insist on the convenient simplicity of a combination that is evaluated quickly. A linear combination such as $\zeta_1 + c\zeta_2$ may have higher power than individual tests. Whereas relative scales would suggest $c = 7$, we opted for the larger value of $c = 17$ that was suggested by various comparative power simulations. From these simulations at varying sample sizes, we show two exemplary plots in Figure 2. Test power is seen to increase rapidly as c approaches values around $c = 17$ from below and to decrease slowly as c increases further. The power maxima appear as a nearly straight line and recommend a constant value of c . Lower optimal c , however, are sometimes found close to the null as well as at a large distance from the null, where the ζ_2 test is often unable to achieve power close to one.

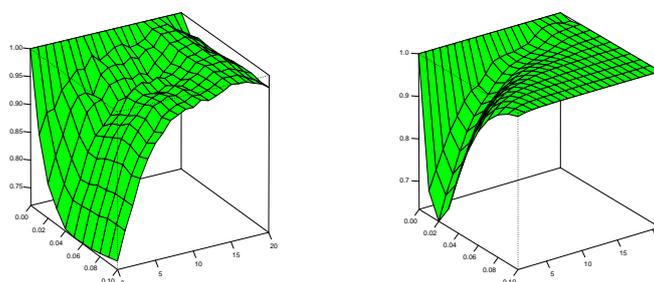


Figure 2: Ratios of test power along generated models $x_t = \phi x_{t-1} + \varepsilon_t$, for $\phi = 1 - \tilde{\phi}$ and $\tilde{\phi}$ on the y -axis. Tests are based on linear combinations $\zeta_1 + c\zeta_2$, with c on the x -axis. For given $\tilde{\phi}$, power is divided by the maximum, such that a value of 1 on the z -axis indicates maximum power. Left graph for $T = 100$, right graph for $T = 400$.

In the following, the statistics

$$\zeta = \frac{\zeta_1 + 17\zeta_2}{18}, \quad \tilde{\zeta} = \frac{\tilde{\zeta}_1 + 17\tilde{\zeta}_2}{18} \quad (9)$$

will be in focus. Table 1 provides some corresponding quantiles.

Based on the simulated quantiles, Table 2 adds some simple power simulations. Test power is investigated along the ray through the alternative

$$x_t = \phi x_{t-4} + \varepsilon_t, \quad (10)$$

with ϕ varying over $0.9 + j * 0.01$ with $j = 0, \dots, 10$, such that $j = 10$ represents the null and $j = 0$ implies $\phi = 0.9$. This rather crude simulation design corresponds to the very basic seasonal unit-root test that was suggested originally by [7], which is rarely used nowadays. Nonetheless, power turns out to be surprisingly good, although lower than for the parametric HEGY test or the test by [7], which was constructed particularly along the same ray and has the best power here. The combined version $\zeta = \frac{\zeta_1 + 17\zeta_2}{18}$ dominates the single tests convincingly, sometimes excepting an area close to the null, where ζ_2 alone shows a slight advantage. On the other hand,

Table 1: Significance points for the nonparametric tests.

	$T = 100$			$T = 400$			$T = 4000$		
	1%	5%	10%	1%	5%	10%	1%	5%	10%
ζ_1	2.10	1.53	1.27	2.12	1.55	1.29	2.15	1.58	1.29
ζ_2	0.25	0.20	0.17	0.23	0.19	0.17	0.22	0.18	0.16
$\frac{\zeta_1+17\zeta_2}{18}$	0.32	0.25	0.22	0.31	0.25	0.22	0.31	0.24	0.21
$\tilde{\zeta}_1$	1.66	1.18	0.94	1.93	1.37	1.11	2.06	1.49	1.20
$\frac{\tilde{\zeta}_1+17\tilde{\zeta}_2}{18}$	0.29	0.22	0.19	0.30	0.22	0.19	0.29	0.23	0.19

Table 2: Power properties of the nonparametric tests.

ϕ	$T = 100$			$T = 400$		
	ζ_1	ζ_2	ζ	ζ_1	ζ_2	ζ
1.00	0.050	0.050	0.050	0.050	0.050	0.050
0.99	0.074	0.083	0.086	0.184	0.268	0.267
0.98	0.102	0.121	0.131	0.367	0.575	0.577
0.97	0.140	0.171	0.188	0.548	0.798	0.806
0.96	0.183	0.229	0.253	0.712	0.914	0.927
0.95	0.229	0.283	0.319	0.825	0.965	0.978
0.94	0.279	0.339	0.388	0.895	0.986	0.993
0.93	0.326	0.394	0.452	0.940	0.995	0.998
0.92	0.377	0.446	0.520	0.967	0.997	0.999
0.91	0.427	0.497	0.585	0.984	0.999	1.000
0.90	0.473	0.544	0.643	0.992	1.000	1.000

Generating model is $x_t = \phi x_{t-4} + \varepsilon_t$. Significance level is 5%. $\phi = 1$ is the null.

ζ_2 shows unsatisfactory performance in small samples at a comparatively large distance from the null. Also [12] and [13] report low power in many directions for a non-parametric test that is similar to ζ_2 .

Table 2 evaluates the power for the original statistics ζ_1 and ζ and not for the adjusted statistics $\tilde{\zeta}_1$ and $\tilde{\zeta}$. Not much is lost, however, in this simple design if the adjustment is applied. The situation changes when autocorrelation is present under the null.

A known problem with both original nonparametric unit-root tests ζ_1 and ζ_2 and thus also with the combined ζ is that the statistical properties are sensitive to deviations from the pure SRW walk model under the null. The test becomes non-similar. Our experiments (unreported due to lack of space) confirm this problem. If the SRW generating mechanism is replaced by $x_t = x_{t-4} + u_t$ with u_t stable autoregressive, the test becomes undersized for negative and positive autocorrelation, with the size bias persisting at slightly larger samples. Under the alternative, the size bias entails lower power.

Implementing the long-run variance correction reduces the distortion problems. The component test based on ζ_2 cannot be adjusted, thus the combined test inherits problems from its

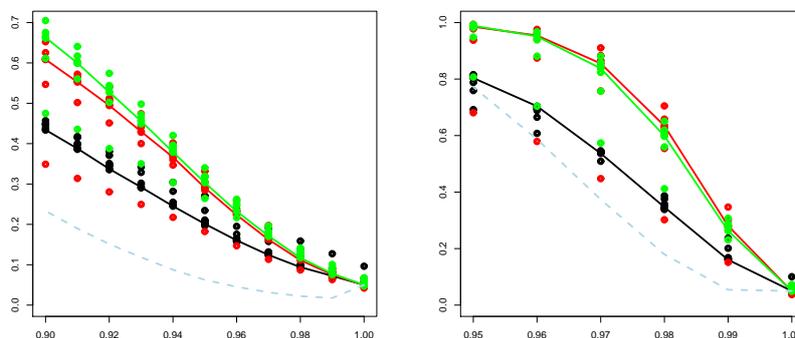


Figure 3: Power function for $T = 100$ and $T = 400$. Black: $\tilde{\zeta}_1$; red: ζ_2 ; green: $\tilde{\zeta}$. Light blue dashes: a HEGY variant. Generating model is $x_t = \phi x_{t-4} + u_t$, $u_t = \theta u_{t-1} + \varepsilon_t$, $\theta = 0.25j$, $j = -3, \dots, 3$.

typically strongly weighted component. The adjusted portion $\tilde{\zeta}_1$, however, succeeds in removing a good part of the distortion. Another feature, however, makes Figure 3 even more interesting. In these simulations, we ran control simulations based on the traditional parametric HEGY test. Actually, we considered two versions of the HEGY test: first, an F test for the seasonal unit roots at -1 and at $\pm i$; second, an F test for the same seasonal unit roots, although under the incorrect assumption of a unit root at $+1$. The power performance of both HEGY versions is very similar. The HEGY test is clearly beaten by the nonparametric tests under investigation.

The power performance of the HEGY test agrees well with literature sources, thus the difference between the tests is not due to specific problems of the parametric tests but rather to the effect of additional power of range expansion tests that was also emphasized by AES. We note that the dominance of the test $\tilde{\zeta}$ shrinks as T increases, and the HEGY test actually dominates at a larger distance from the null. The effect appears to be at odds with the usual statistical test construction that is based on asymptotic optimization, and it is difficult to interpret. A tentative explanation may be that the nonparametric tests concentrate on the true classification features of interest, such as zero crossings and range expansion, while the parametric tests are limited by the precision of the regression estimates.

4 Empirical applications

Whereas the simulated charts are mostly based on relatively large samples, empirical data sets typically are much shorter, whether they are taken from economics or from other disciplines.

Austrian industrial production is a quarterly variable that is available from 1957. The left panel of Figure 4 shows the seasonal jitter plot for the years 1957–2011. The seasonal pattern shows some variation, but it usually returns to its basic shape quickly. A blurred saltire forms in bin # 5, which represents the activity troughs during summer vacation and in the cold start of the year, with rising activity during the second and fourth quarters. The values of the test statistics are $(\tilde{\zeta}_1, \zeta_2, \tilde{\zeta}) = (0.27, 0.17, 0.17)$. These values are in the non-rejection region, maybe excepting ζ_2 which provides weak evidence for rejection, as it is close to the 10% quantile. Similar

results are obtained for other aggregate economic variables with strong seasonality, such as for example unemployment rates.

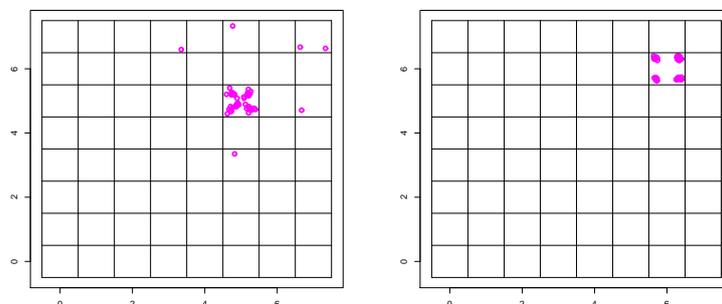


Figure 4: Jittered phase plot for Austrian industrial production 1957–2011, and for temperature at Heathrow airport 1984–2011, quarterly observations.

The right panel of Figure 4 shows the jitter plot for air temperature at Heathrow. This variable provides an example for purely deterministic seasonality with almost negligible dependence among seasonal patterns in adjacent years. Four symmetric spots in the bin # 6 represent the repetitive cycle of rise-rise-fall that is observed annually with no exception. Here, the values of the test statistics are $(\tilde{\zeta}_1, \zeta_2, \tilde{\zeta}) = (0, 0.21, 0.20)$, with ζ_2 playing a key role in rejecting seasonal unit roots.

It is of some interest to compare these test results to the outcome of traditional parametric tests for seasonal unit roots. For example, the most straightforward test of that type is the HEGY test due to [11], which rejects the unit root null for both considered variables. HEGY results vary slightly across lag-order specifications for augmenting terms.

5 Summary and conclusion

We demonstrate that the suggested nonparametric combination test tends to dominate its constituent component tests, and the suggested weight appears to be well chosen. Surprisingly, in some designs the test even dominates parametric tests for seasonal unit roots. Our results are in line with AES, whose simulations are quite supportive for their idea in traditional unit-root situations, as their RUR test often outperforms standard tests, such as the Dickey-Fuller test. Our results are less in line with the simulations of [4] who delineate a gloomy picture for the power of their nonparametric tests, even if combined with traditional tests.

We are also able to demonstrate that the visualization by jittered phase plots is an appealing and potentially helpful tool in the investigation of the nature of seasonality in time series. The slanted cross or saltire appears to be a well recognizable shape, and deviations from the pure form are easily spotted by the human eye.

A main problem with seasonality remains: good discrimination requires samples that are slightly larger than those that are typically available. Really long time series are needed in order to discriminate safely the case of pattern reversion from the case of episodic pattern change.

Bibliography

- [1] Akonom, J. (1993) *Comportement asymptotique du temps d'occupation du processus des sommes partielles*. Annales de l'Institut Henri Poincaré (B) Probabilités et Statistiques, **29**, 57–81.
- [2] Aparicio, F., Escribano, A. and Sipols, A.E. (2006) *Range unit-root (RUR) tests: robust against nonlinearities, error distributions, structural breaks and outliers*. Journal of Time Series Analysis, **27**, 545–576.
- [3] Breitung, J. (2002) *Nonparametric tests for unit roots and cointegration*. Journal of Econometrics, **108**, 343–363.
- [4] Burridge, P. and Guerre, E. (1996) *The Limit Distribution of Level Crossings of a Random Walk, and a Simple Unit Root Test*. Econometric Theory, **12**, 705–723.
- [5] Canova, F. and Hansen, B.E. (1995) *Are seasonal patterns constant over time? A test for seasonal stability*. Journal of Business & Economic Statistics, **13**, 237–252.
- [6] Caporale, M.C., Cunado, J. and Gil-Alana, L.A. (2012) *Deterministic versus stochastic seasonal fractional integration and structural breaks*. Statistics and Computing, **22**, 349–358.
- [7] Dickey, D.A., Hasza, D.P. and Fuller, W.A. (1984) *Testing for Unit Roots in Seasonal Time Series*. Journal of the American Statistical Association, **79**, 355–367.
- [8] García, A. and Sansó, A. (2006) *A generalization of the Burridge-Guerre Nonparametric Unit Root Test*. Econometric Theory, **22**, 756–761.
- [9] Ghysels, E. and Osborn, D. (2001) *The Econometric Analysis of Seasonal Time Series*. Cambridge University Press.
- [10] Hylleberg, S. (1986) *Seasonality in Regression*. Academic Press.
- [11] Hylleberg, S., Engle, R.F., Granger, C.W.J. and Yoo, B.S. (1990): *Seasonal integration and cointegration*. Journal of Econometrics, **44**, 215–238.
- [12] Kunst, R.M. (2009) *A Nonparametric Test for Seasonal Unit Roots*. Economics Series, No. 233, Institute for Advanced Studies, Vienna.
- [13] Kunst, R.M. and Franses, P.H. (2011) *Testing for seasonal unit roots in monthly panels of time series*. Oxford Bulletin of Economics and Statistics, **73**, 469–488.
- [14] Phillips, P.C.B. and Perron, P. (1988) *Testing for a unit root in time series regressions*. Biometrika, **75**, 335–346.
- [15] Taylor, A.M.R. (2005) *Variance ratio tests of the seasonal unit root hypothesis*. Journal of Econometrics, **124**, 33–54.

To Split or to Mix? Tree vs. Mixture Models for Detecting Subgroups

Hannah Frick, *Universität Innsbruck*, Hannah.Frick@uibk.ac.at

Carolin Strobl, *Universität Zürich*, Carolin.Strobl@psychologie.uzh.ch

Achim Zeileis, *Universität Innsbruck*, Achim.Zeileis@uibk.ac.at

Abstract. A basic assumption of many statistical models is that the same set of model parameters holds for the entire sample. However, different parameters may hold in subgroups (or clusters) which may or may not be explained by additional covariates. Finite mixture models are a common technique for detecting such clusters and additional covariates (if available) can be included as concomitant variables. Another approach that relies on covariates for detecting the clusters are model-based trees. These recursively partition the data by splits along the covariates and fit one model for each of the resulting subgroups. Both approaches are presented in a unifying framework and their relative (dis)advantages for (a) detecting the presence of clusters and (b) recovering the grouping structure are assessed in a simulation study, varying both the parameter differences between the clusters and their association with the covariates.

Keywords. finite mixture model, model-based clustering, model-based recursive partitioning

1 Introduction

A basic assumption of many statistical models is that its set of parameters applies to all observations. However, subgroups may exist for which different sets of parameters hold, e.g., the relationship between some response and regressors might be different for younger and older individuals. If the breakpoint which separates “younger” and “older” were known, parameter stability can simply be assessed by checking for parameter differences between these two specific subgroups. However, if the breakpoint is unknown or if there is a smooth transition between “young” and “old”, the subgroups can be still be detected in a data-driven way. Either a finite mixture model [5] can be employed, possibly using age as a concomitant variable [2] to model smooth transitions between clusters. Alternatively, model-based recursive partitioning [9] can capture the difference by one or more splits in the partitioning variable age, yielding a tree structure (similar to classification and regression trees, [1]) where each leaf is associated with a parametric model.

Given their shared goal of establishing subgroups to capture parameter instabilities across subgroups, how do these mixture models and model-based trees compare? A unifying framework for both methods is presented and their relative (dis)advantages for (a) detecting the presence of clusters with different parameters and (b) recovering the underlying grouping structure are assessed in a simulation study.

2 Theory

Although both mixture models and model-based trees can be applied to general parametric models estimated by means of the maximum likelihood (ML) principle, we focus on linear regression here because it is the most simple and commonly applied model:

$$y_i = x_i^\top \beta + \varepsilon_i \quad (1)$$

with response y_i , regressor vector x_i , and errors ε_i for observations $i = 1, \dots, n$. The unknown vector of regression coefficients β can be estimated by least squares, which yields the same estimates as ML estimation under the assumption of independent normal errors with variance σ^2 . In the latter case the log-likelihood is given by $\sum_{i=1}^n \log \phi(y_i; x_i^\top \beta, \sigma^2)$ where $\phi(\cdot)$ denotes the probability density function of the normal distribution.

Within this framework, both trees and mixture models can assess whether the same coefficient vector β holds for all n observations and, if parameter stability is violated, simultaneously find clusters/subgroups and estimate the associated cluster-specific coefficients. Further covariates z_i can be used as concomitant or partitioning variables, respectively, to establish these clusters.

Finite mixture models

Finite mixture models assume that the data stem from K different subgroups with unknown subgroup membership and subgroup-specific parameters $\beta_{(k)}$ and $\sigma_{(k)}$ ($k = 1, \dots, K$). The full mixture model is a weighted sum over these separate models (or components):

$$f(y_i; x_i, z_i, \beta_{(1)}, \sigma_{(1)}, \dots, \beta_{(K)}, \sigma_{(K)}) = \sum_{k=1}^K \pi_k(z_i) \cdot \phi(y_i; x_i^\top \beta_{(k)}, \sigma_{(k)}^2). \quad (2)$$

The component weights may depend on the additional covariates z_i through a concomitant variable model [2], typically a multinomial logit model

$$\pi_k(z_i) = \frac{\exp(z_i^\top \alpha_{(k)})}{\sum_{g=1}^K \exp(z_i^\top \alpha_{(g)})} \quad (3)$$

with component-specific coefficients $\alpha_{(k)}$. For identifiability, one group (typically the first) is used as a reference group and the coefficients of this group are set to zero: $\alpha_{(1)} = 0$. This also includes the special case without concomitant variables, where $z_i = 1$ is just an intercept yielding component-specific weights $\pi_k(z_i) = \pi_{(k)}$.

Given the number of subgroups K , all parameters in the mixture model are typically estimated simultaneously by ML using the expectation-maximization (EM) algorithm. To choose the number of subgroups K , the mixture model is typically fitted for $K = 1, 2, \dots$ and then the best-fitting model is selected by some information criterion. Here, we employ the Bayesian Information Criterion (BIC).

Model-based recursive partitioning

Model-based recursive partitioning [9] can also detect subgroups for which different model parameters hold. These subgroups are separated by sample splits in the covariates z_i used for partitioning. The algorithm performs the following steps:

1. Estimate the model parameters in the current subgroup.
2. Test parameter stability along each partitioning variable z_{ij} .
3. If any instability is found, split the sample along the variable z_{ij^*} with the highest instability. Choose the breakpoint with the highest improvement in model fit.
4. Repeat 2–4 on the resulting subsamples until no further instability is found.

Here, we only briefly outline how these parameter instability tests work and refer to [8] for the theoretical details. The basic idea is that the scores, i.e., the derivative of $\log \phi(\cdot)$ with respect to the parameters, evaluated at the estimated coefficients behave similar to least squares residuals: They sum to zero and if the model fits well they should fluctuate randomly around zero. However, if the parameters change along one of the partitioning variables z_{ij} , there should be systematic departures from zero. Such departures *along* a covariate can be captured by a cumulative sum of the scores (ordered by the covariate) and aggregated to a test statistic, e.g., summing the absolute or squared cumulative deviations. Summing the scores along a categorical partitioning variable then leads to a statistic that has an asymptotic χ^2 distribution while aggregation along numeric partitioning variables can be done in a way that yields a maximally-selected score (or Lagrange multiplier) test. In either case, the p -value p_j can be obtained for each ordering along z_{ij} without having to reestimate the model. The p -values are then Bonferroni-adjusted to account for testing along multiple orderings and partitioning continues until there is no further significant instability (here at the 5% level).

Since each split can be expressed through an indicator function $I(\cdot)$ (for going left or right), each branch of the tree can be represented as a product of such indicator functions. Therefore, the model-based tree induced by recursive partitioning is in fact also a model of type (2), albeit with rather different weights:

$$\pi_k(z_i) = \prod_{j=1}^{J_k} I(s_{(j|k)} \cdot z_{i(j|k)} > b_{(j|k)}) \quad (4)$$

where $z_{(j|k)}$ denotes the j -th partitioning variable for terminal node k , $b_{(j|k)}$ is the associated breakpoint, $s_{(j|k)} \in \{-1, 1\}$ the sign (signaling splitting to the left or right), and J_k the number of splits leading up to node k .

Differences and similarities

While both methods are based on the same linear regression model and aim at detecting subgroups with stable parameters, certain differences arise:

- Because K is fixed for each mixture model, it is based on model selection via an information criterion whereas the selection of K through a tree is based on significance tests.

- Covariates z are optional for mixture models and latent subgroups can be estimated. For a tree, those covariates are required. Furthermore, if no covariates associated with the subgroups are available, the groups cannot be detected.
- The concomitant model (3) assumes a smooth, monotonic transition between subgroups. The sample splits of a tree (4) represent abrupt shifts, multiple splits in a covariate are able to represent a non-monotonic transition. While variable selection is inherent to trees, it requires an additional step for mixtures models.
- Trees yield a hard clustering and mixtures a probabilistic clustering of the observations.

To investigate how the aforementioned differences between the two methods affect their ability to detect parameter instability, a simulation study is conducted, which is described in the next section.

3 Simulation study

To determine how well mixture models and model-based trees detect parameter instability, two basic questions are asked. First, is any instability found at all? Second, if so, are the correct subgroups recovered? These two aspects are potentially influenced by several factors: How does the relationship between the response y and the regressors x differ between the subgroups and how strongly does it differ? If there are any additional covariates z available, how and how strongly are those covariates connected to the subgroups? In general, we expect the following:

- Given the covariates z are associated strongly enough with the subgroups, trees are able to detect smaller differences in $\beta_{(k)}$ than mixtures because they employ a significance test for each parameter rather than an information criterion for full sets of parameters. In contrast, mixtures are more suitable to detect subgroups if they are only loosely associated with the covariates z , as long as the differences in $\beta_{(k)}$ are strong enough.
- If the association between covariates and subgroups is smooth and monotonic, mixtures are more suitable to detect the subgroups whereas trees are more suitable if the association is characterized by abrupt shifts and possibly non-monotonic.
- If several covariates determine the subgroups simultaneously, the mixture is more suitable, whereas trees are more suitable if z includes several noise variables unconnected to the subgroups.

Motivated by these considerations, the simulation design is explained in the next section.

Simulation design

A single regressor x and four additional covariates z_1, \dots, z_4 are drawn from a uniform distribution on $[-1, 1]$. The response y is computed with errors drawn from a standard normal distribution. Two subgroups of equal size are simulated. How they differ is governed by the form of β – either in their *intercept*, *slope*, or *both* – and the magnitude of their differences is governed by the simulation parameter κ (Table 1). How the covariates are connected to the subgroups is governed by the form of $\pi_k(z)$ – either via a logistic or a step function – and the

	Label	Details
Coefficients	intercept	$\beta_{(1)} = (\kappa, 0)^\top$ $\beta_{(2)} = (-\kappa, 0)^\top$
	slope	$\beta_{(1)} = (0, \kappa)^\top$ $\beta_{(2)} = (0, -\kappa)^\top$
	both	$\beta_{(1)} = (\kappa, -\kappa)^\top$ $\beta_{(2)} = (-\kappa, \kappa)^\top$
Covariates	axis1	logistic with $\alpha_{(2)} = (\exp(\nu), 0, 0, 0)^\top$
	diagonal	logistic with $\alpha_{(2)} = (\exp(\nu), -\exp(\nu), 0, 0)^\top$
	double step	tree with $\pi_2(z) = I(z_1 > -0.5)I(z_1 < 0.5)$

Table 1: Simulation scenarios for regression coefficients and covariates.

strength of this association is governed by simulation parameter ν . Here, three scenarios for $\pi_k(z)$ are considered: a smooth logistic transition along z_1 , a smooth logistic transition along z_1 and z_2 simultaneously, and a sharp transition along z_1 with two breakpoints (labeled *axis1*, *diagonal*, and *double step*, respectively). The corresponding parameter vector of the logistic function and the breakpoints can be found in Table 1. The simulation parameters cover the following ranges: $\kappa \in \{0, 0.05, \dots, 1\}$ and $\nu \in \{-1, -0.5, \dots, 2\}$. Note that $\beta_{(1)}$ and $\beta_{(2)}$ are identical if $\kappa = 0$ and thus only one subgroup is simulated. Each coefficient scenario is combined with every covariate scenario and the sample size $n \in \{200, 500, 1000\}$ is varied. The covariates z_3 and z_4 are always noise variables and thus either included or excluded in z . For each of these conditions, 500 datasets are drawn and three methods applied: a model-based tree, a plain mixture, and a mixture with concomitant variables. Both mixtures are fitted with $K = \{1, \dots, 4\}$ and \hat{K} selected via BIC. For all computations, the R system for statistical computing [7] is used along with the add-on packages **partykit** [4] and **flexmix** [3].

Outcome assessment

To address the first question of whether or not any instability is found, the *hit rate* is computed: This is the rate of selecting more than one subgroup – this corresponds to splitting at least once for a tree and to selecting $\hat{K} > 1$ for a mixture. To address the second question if the right subgroups are found, the estimated clustering is compared to the true clustering. Many external cluster indices such as the Rand index favor a “perfect match”, i.e., splitting one true subgroup into several estimated subgroups (as might be unavoidable in a tree) is penalized by the index. Cramér’s coefficient is invariant against such departures from a perfect match [6] and thus employed here.

Simulation results

Exemplary results are shown for the scenario with differences in *both* coefficients with $n = 200$ observations, without the noise variables z_3 and z_4 , and the *double step* scenario as well as the logistic scenarios *axis1* and *diagonal* with three levels $\nu = \{-1, 0, 1\}$ of separation between subgroups. For the *double step* scenario, the hit rate for detecting instability is depicted in the left panel of Figure 1. The tree clearly outperforms both mixtures. For the logistic scenarios *axis1* and *diagonal*, the hit rates are depicted in Figure 2. If the covariates are only weakly associated with the subgroups ($\nu = -1$, left column), the tree is unable to detect the subgroups,

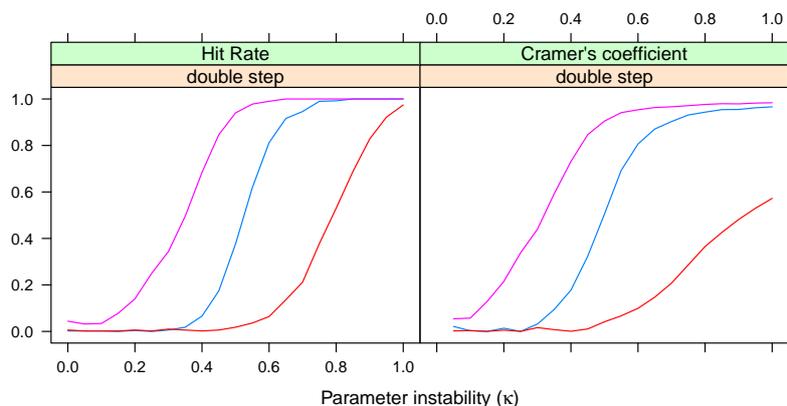


Figure 1: Hit Rate (left panel) and Cramér's coefficient (right panel) for the *double step* covariate scenario. Line type: dashed for tree, solid for mixture, dot-dashed for plain mixture.

regardless of how strongly they differ in their regression coefficients. Both mixtures are able to detect instability beyond a threshold of $\kappa = 0.7$ and reach hit rates of almost 1. For a medium association ($\nu = 0$), the tree is able to detect smaller differences in the regression coefficients than the mixtures but for larger differences both mixtures equally outperform the tree. If the association is strong ($\nu = 1$), the tree outperforms both mixtures. The mixture with concomitants in turn outperforms the plain mixture which is per definition invariant to (changes in) the association between covariates and subgroups. Interestingly, the tree performs rather similarly for the scenarios *axis1* and *diagonal*, indicating that approximation through sequential splits works rather well. For $\kappa = 0$ only one true subgroup exists and mixtures nearly always select $\hat{K} = 1$ while trees incorrectly select $\hat{K} > 1$ subgroups only in less than 5% of the cases (which is the significance level employed in the parameter stability tests).

The corresponding recovery of the subgroups as measured by Cramér's coefficient is depicted in the right panel of Figure 1 for the *double step* scenario. Similar to the detection of instability, the tree outperforms both mixtures. For the logistic scenarios, the Cramér's coefficients are shown in Figure 3. For low and medium levels of association between covariates and subgroups, both mixtures outperform the tree for stronger instabilities. For a medium level of association ($\nu = 0$) and small instabilities, the tree's advantage in detecting instabilities translates into an advantage of also uncovering the correct subgroups. However for a stronger association ($\nu = 1$), the mixture with concomitants recovers the true subgroups better than the other two methods once the hit rates are similar across methods. Despite its good hit rates, the tree never exceeds a Cramér's coefficient of about 0.6. This is the case regardless of how strong the regression coefficients differ, indicating that the tree's ability to uncover the correct subgroups is limited by the (relative) weakness of association between covariates and subgroups. For an even stronger association ($\nu > 1$, not depicted here), the tree recovers the subgroups as well as the concomitant mixture in the *axis1* scenario but fails to do so in the *diagonal* scenario.

For larger numbers of observations ($n = 500$ or 1000) and the other two coefficients scenarios (*intercept* and *slope*), results are similar to those shown here, just being generally more pronounced. Including two additional noise variables z_3 and z_4 affected both the tree and the concomitant mixture, with hit rates dropping slightly stronger for the mixture.

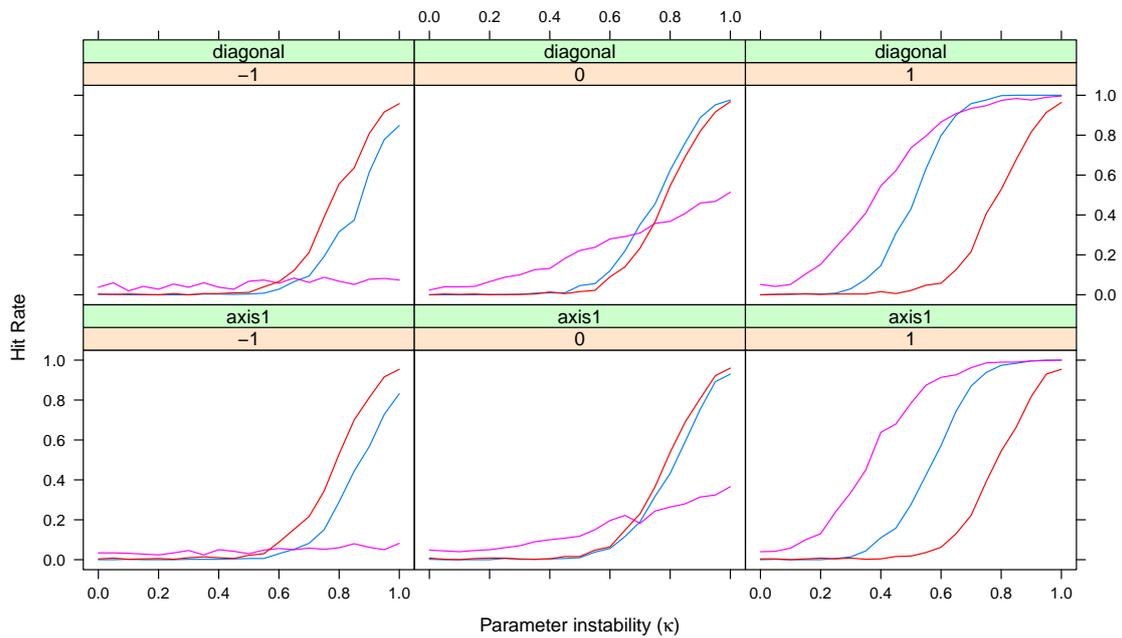


Figure 2: Hit Rate for the logistic covariate scenarios for three levels of $\nu \in \{-1, 0, 1\}$. Line type: dashed for tree, solid for mixture, dot-dashed for plain mixture.

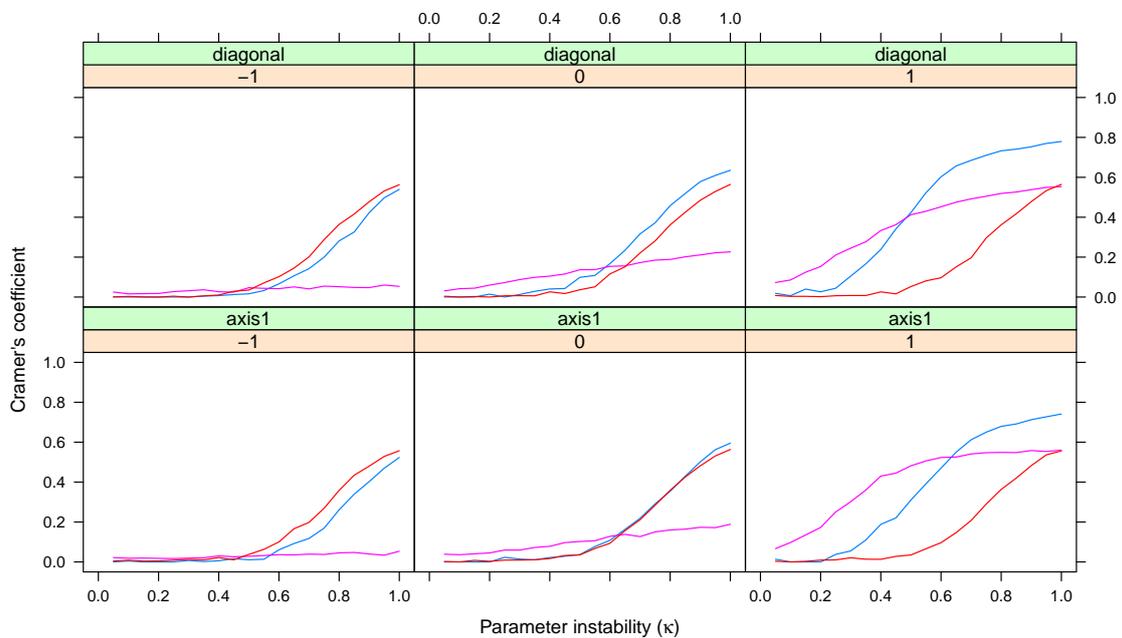


Figure 3: Cramér's coefficient for the logistic covariate scenarios for three levels of $\nu \in \{-1, 0, 1\}$. Line type: dashed for tree, solid for mixture, dot-dashed for plain mixture.

4 Discussion

Both methods are suitable to detect parameter instability (or lack thereof) and recover the subgroups (if any). Which method is more suitable depends largely on the association between the subgroups and covariates as well as how strongly the subgroups differ in their respective parameter vectors. If the association between subgroups and covariates is strong, the tree is able to detect smaller differences in the parameters than the mixtures. The approximation of a smooth transition between classes through sample splits works rather well. In addition, the tree can represent a non-monotonic association which the mixtures cannot. If the association between subgroups and covariates is weak but the difference in the parameters reasonably strong, mixture models are more suitable than the tree. Mixture models are also capable of detecting latent subgroups without any association to covariates. It would be interesting to investigate whether these relationships also apply to situations with more subgroups and mixtures with higher numbers of components. Further questions for further research include the assessment of subgroup recovery on a test set rather than in-sample and variable selection for the concomitant models of mixtures, which could also be accomplished with an information criterion.

In summary, both methods have their relative advantages and thus are more suitable to detect parameter instability and uncover subgroups in different situations. As the exact structure is unknown in practice, we suggest using both methods to gain better insight into the data.

Bibliography

- [1] Breiman, L., Friedman, J.H., Olshen, R.A. and Stone, C.J. (1984) *Classification and Regression Trees*. Wadsworth, California.
- [2] Dayton, C.M. and Macready, G. (1988) *Concomitant-variable latent-class models*. Journal of the American Statistical Association, **83**(401):173–178.
- [3] Grün, B. and Leisch, F. (2008) *FlexMix version 2: Finite mixtures with concomitant variables and varying and constant parameters*. Journal of Statistical Software, **28**(4):1–35.
- [4] Hothorn, T. and Zeileis, A. (2014) *partykit: A modular toolkit for recursive partytioning in R*. Working Paper 2014-10, Research Platform eeecon, Universität Innsbruck.
- [5] McLachlan, G. and Peel. D. (2000) *Finite Mixture Models*. John Wiley & Sons, New York.
- [6] Mirkin, B. (2001) *Eleven ways to look at chi-squared coefficients for contingency tables*. The American Statistician, **55**(2):111–120.
- [7] R Core Team (2013) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.
- [8] Zeileis, A. and Hornik, K. (2007) *Generalized M-fluctuation tests for parameter instability*. Statistica Neerlandica, **61**(4):488–508.
- [9] Zeileis, A., Hothorn, T. and Hornik, K. (2008) *Model-based recursive partitioning*. Journal of Computational and Graphical Statistics, **17**(2):492–514.

Propensity score matching with clustered data: an application to birth register data

Massimo Cannas, *University of Cagliari*, `massimo.cannas@unica.it`
Bruno Arpino, *Universitat Pompeu Fabra*, `bruno.arpino@upf.edu`

Abstract. In this paper we consider the implementation of propensity score matching for clustered data. Different approaches to reduce bias due to cluster level confounders are considered: matching within clusters and random or fixed effects models for the estimation of the propensity score. All the methods are illustrated with an application to the estimation of the effect of caesarean section on the Apgar score using birth register data from Sardinia hospitals.

Keywords. Causal inference, Propensity score, Matching, Multilevel data, Caesarean section, Apgar score.

1 Introduction

Methods based on the propensity score are widely used in many fields to estimate causal effects with observational data. When treatment assignment is not randomized but it is reasonable to assume that selection is on observables, matching (as well as weighting and stratification) methods are used to adjust for different distributions of the observed characteristics in the treated and the control groups [7]. Apart from few exceptions [2, 8, 11] these methods have been considered only for unstructured data. However, in many applications data show a hierarchical structure (e.g., students nested into schools, patients nested into hospitals, individuals nested into geographical areas). We consider situations where both individual and cluster-level (e.g., hospital) characteristics can influence both treatment intake and the outcome. In these contexts ignoring cluster-level confounding factors would introduce a bias.

In this paper, we consider different approaches to take into account the hierarchical structure of the data with the aim of reducing the bias due to group-level characteristics. These methods are particularly useful when it is not possible to measure all cluster-level confounders. To illustrate the methods, we consider estimating the effect of caesarian section on the Apgar score. In our application, the relevant structure is represented by a hierarchy of 2 levels (individuals

nested into hospitals) and we will consider this type of data structure in the following. However, the approaches we consider can be easily adapted to more complex structures.

Propensity score matching with clustered data

Suppose we have a two-level data structure where N micro units at the first level, indexed by i ($i = 1, 2, \dots, n_j$), are nested in J macro units at the second level (clusters), indexed by j ($j = 1, 2, \dots, J$). We consider a binary treatment administered at the individual level, T , and an outcome variable, Y also measured at the individual level. Pre-treatment variables can be first (X) or second level (Z) variables.

Under the potential outcome framework, let $Y_{ij}(t)$ be the potential outcome if unit ij was assigned to treatment t , $t \in \{0, 1\}$. An individual causal effect is a comparison of $Y_{ij}(1)$ with $Y_{ij}(0)$, yet only one of the two potential outcomes is observed depending on the value of T_{ij} . Usually, the Average Treatment effect on the Treated (ATT) is considered as an interesting summary of individual causal effects: $ATT = E(Y_{ij}(1) - Y_{ij}(0) | T_{ij} = 1)$.

To identify the ATT with observational data, the following assumptions are often invoked:

- SUTVA: If $T = T'$ then $Y(T) = Y(T')$ for all T, T' in $\{0, 1\}^N$
- Unconfoundedness: $Y(1), Y(0) \perp T | (X, Z)$;
- Overlap: $0 < P(T = 1 | (X, Z)) < 1$.

The Stable Unit Treatment Value Assumption (SUTVA, [9]) requires that potential outcomes for a unit are not affected by the treatment received by other units, and there are no hidden versions of the treatment. Unconfoundedness asserts that the probability of assignment to a treatment does not depend on the potential outcomes conditional on observed covariates [9]. Unconfoundedness essentially assumes that within subpopulations defined by values of the covariates, we have random assignment of the treatment; it rules out the role of unobserved variables and therefore is often referred to also as selection on observables [7].

Rosenbaum and Rubin [9] showed that under the previous assumptions, adjustment on the propensity score eliminates bias due to observed confounders. The propensity score, e , is defined for each unit as the probability to receive the treatment conditional given its covariate values. In our setting, assuming that all covariates are observed we have $e_{ij} = Pr(T_{ij} = 1 | (X_{ij}, Z_j))$. The propensity score is a one-dimensional summary of the multidimensional set of covariates, such that when the propensity score is balanced across the treatment and control groups, the distribution of all covariates are balanced in expectation across the two groups. In this way the problem of adjusting for a multivariate set of observed characteristics reduces to adjusting for the one-dimensional propensity score and this can be done using several Propensity Score Matching (PSM) algorithms that, for each given unit, determine a set of units in the opposite treatment condition with similar value for the propensity score.

In observational studies the propensity score is not known and must be estimated from the data, usually using logit or probit models. Obviously, an incorrectly estimated propensity score may lose its balancing property. More importantly, if one or more variables affecting the selection into treatment and potential outcomes are not observed, then unconfoundedness is violated and

ATT estimators based on PSM will be biased. In fact, PSM can only balance variables used in the propensity score model. In the following we shall assume that we have good measurement on all individual level confounders, X , but we may have no information on all or some of the second-level confounders, Z . We consider different approaches to implement PSM with a 2-level data structure. Two groups of strategies can be adopted in order to take into account the hierarchical structure of the data: implementing the matching within clusters; using a model for the estimation of the propensity score that takes the hierarchical structure explicitly into account. Therefore, the approaches we compare are as follows:

- A Single-level propensity score; matching on the pooled dataset;
- B Single-level propensity score; matching only within-clusters;
- C Single-level propensity score; preferential within-cluster matching;
- D Random-effect propensity score; matching on the pooled dataset;
- E Fixed-effect propensity score; matching on the pooled dataset.

Approach A ignores completely the hierarchical structure. In this case, if we do not include all relevant confounders at the second level in the propensity score and obtain a good balance on all of them, our ATT estimator based on the PSM will be biased. Approach B deals with this problem by matching units within clusters only. This automatically guarantees that all cluster-level variables (measured and unmeasured) are perfectly balanced. This can come to a cost. Control units to be matched with treated units are only searched within the same cluster. In this way it could be that we lose some good match and so the balancing of individual level variables could be worse. Moreover, if we impose a caliper it could be that we do not find a control matched unit that we would find in other clusters. So, an additional problem could be losing some treated units.

To avoid these problems and combine the benefits of approaches A and B, approach C starts by searching control units within cluster. If none is found, control units are searched in other clusters. This approach improves the balancing of cluster level variables with respect to approach A and avoids the lost of units of approach B.

In alternative to exploiting the hierarchical structure in the implementation of the matching, approaches D and E take it into account when modelling the propensity score. In particular, approach D and E use a random or fixed effect, respectively, to represent unmeasured cluster level variables. Arpino and Mealli [2] and Thoemmes and West [11] showed that PSM using random or fixed effects models are able to reduce the bias of ATT due to unmeasured cluster level variables. However, our simulation exercise is more realistic because it is inspired by a real case studies, it involves a larger number of individual level variables and strongly unbalanced dataset.

Estimating the effect of caesarian section on Apgar score

Apart from individual level variables, the literature suggested the relevance of hospital level factors both on the decision of taking a medical treatment and on the medical outcomes for several procedures. In other words, these cluster level variables may act as confounders and so the researcher should adjust the analysis accordingly. For example, Caceras et al. [4] and Bragg

et al. [3] indirectly measured the impact of hospital variables on the likelihood of a caesarean delivery. Similarly, since the work of Hughes et al. [6] it is clear that these variables may also affect the quality of the outcome. When we refer to unobserved variables at the hospital level we are referring to variables whose role has been proved or conjectured by previous studies; for example variables which do not vary at the hospital level for a reasonably long period of time, like obstetrician practice, physician's preferences and guidelines promoting or restricting the liberal use of caesarean sections. Clearly, it is not always possible to observe all hospital level factors that contribute to the decision of operating a caesarean section and may also impact on the infant's health as measured by the Apgar score. To this end we adopt the strategies detailed in the previous section.

2 Data

The data set we consider contains information on deliveries occurred in the 22 hospitals of the Italian region of Sardinia in 2010 and 2011. The source is the official form on the birth event (known as CedAP) filled by physicians after the birth and accounting for all hospitalized births in the specified period. The form is divided in three parts containing sociodemographic information on the mother, the pregnancy and the infant. From the initial population of 23,925 observations we extracted the subset of non-complicated pregnancies in order to better isolate the effect of the caesarian section on the target variable. In particular, we selected nulliparous women at 32 or more weeks of gestational age with a singleton and living infant in vertex (head-down) position, without birth anomalies. We further restrict the sample to mothers aged between 15 and 44. The subset of non-complicated pregnancies is widely used in observational studies related to cesarean section, for example [3, 4] make analogous variable selections, but the former study also limits the sample to hospitals with almost 500 deliveries per year. The selected subset contains 14,757 cases clustered in 20 hospitals (the observations of two hospitals were removed since after the selection they contained only treated or untreated women). Proportions of caesarean sections across hospitals vary from a minimum of 0.11 to a maximum of 0.64 with an average of 0.35 (see Table 1). We focus on the 5-minute Apgar score as the outcome variable. This score is a simple and widely established indicator of the infant's health. It is well known that low Apgar scores are strongly associated with high mortality rates [1]. In our sample the proportion of low (< 7) scores is 0.0064. The score distribution is highly skewed with an average score of 9.54.

We built the propensity score model for the probability of caesarean section relying on a set of clinical (X) and social (Z) variables that proved significant in previous studies. In the first group of predictor we have infant weight, mother's gestational age, induction of labour and pregnancy related pathologies. In the second group we have socio-demographic information like maternal age and maternal education

3 Empirical Results

We start by reporting in Table 2 the mean differences of covariates across treated and untreated women for each balancing strategies. The last row of the table averages the (absolute) differences over all covariates and it known as the standardized bias (ASAM), an overall measure of covariate balance. We report the balance before matching and compare it with the balance we obtain with approaches A, B, C, D and E. Several variables showed a standardized bias higher than

Hospital	N. births	N. caesarean sections	% caesarean sections	‰ low apgar infants
1	2,532	1,166	46.0	16.5
2	1,788	623	34.8	2.7
3	1,687	540	32.0	5.3
4	1,473	632	42.9	14.2
5	1,253	410	32.7	0.7
6	1,197	428	35.7	3.3
7	980	240	24.4	2.0
8	875	238	27.2	5.7
9	529	190	35.9	3.7
10	434	135	31.1	6.9
11	403	164	40.6	0
12	396	117	29.5	7.5
13	351	134	38.1	8.5
14	266	74	27.8	7.5
15	208	99	47.5	9.6
16	191	122	63.8	10.4
17	103	40	38.8	9.7
18	50	9	18.0	20
19	32	13	40.6	0
20	9	1	11.1	0
Total	14,757	5,375		
Mean	737.8	268.7	35.0	6.75

Table 1: Number of cesarean sections and low Apgar infants by hospital.

commonly accepted threshold (5% or 10%) representing substantive unbalance before matching. All considered approaches were effective in reducing imbalance even if approaches B and C show a slightly worse balance. However, these methods compared to method A take into account possible hospital level confounding effects and give anyway acceptable balance of all individual covariates. In particular, method B should be the preferred one given that it automatically balances all hospital level factors but still guarantees good balance of individual observed confounders compared to the other approaches. Finally, approaches D and E give slightly better ASAM than B and C for individual level covariates even if the balance of unobserved covariates at the hospital level is not guaranteed as is in within cluster matching.

In Table 3 the total number of treated units dropped due to the caliper option is shown. Here the caliper is 0.25 in standard deviation units so all treated units with a propensity score (e) outside the range $(e - 2\sigma_e, e + 2\sigma_e)$, where σ_e indicates the standard deviation of the propensity score, will be discarded. When matching within hospitals we keep the same criterion by using the standard deviation of the clusters as the reference value. The matched dataset were obtained using macros based on the Matching package [10]. It is interesting noting that the number of drops is not a constant proportion of the cluster size (not shown), as the covariate distribution may vary across clusters.

In Table 4 we show the ATT estimate for unmatched (i.e. the raw effect prior to any

Variable	Before	A	B	C	D	E
<i>Maternal Age (years)</i>						
< 20	-14.942	-0.632	-0.552	-0.551	-0.936	-1.685
20-24	-12.461	1.254	2.278	2.269	1.593	1.223
25-29	-15.048	0.151	1.778	1.780	0.915	1.257
30-35	-6.119	-0.708	-0.854	-0.818	-2.297	0.288
> 35	26.672	0.128	-1.435	-1.461	1.035	-1.383
<i>Maternal Education</i>						
Less than High School	-2.534	0.239	-4.264	-4.360	-2.495	-3.746
High School	0.575	-1.359	1.794	1.1784	0.452	2.172
Graduate or more	2.802	-0.997	3.063	2.998	0.581	-0.235
Missing	-0.056	0.828	0.418	0.688	2.849	2.910
<i>Infant Weight (grams)</i>						
< 2500	21.498	0.524	0.413	0.402	0.620	-0.291
2500-4000	-23.880	-1.700	-2.542	-2.544	-0.120	0.193
>4000	9.138	2.187	3.782	3.856	1.160	0.104
<i>Labor Induction</i>	-5.038	-1.547	0.393	0.437	-2.562	-2.813
<i>Gestational Age</i>						
Preterm (< 37 weeks)	23.273	-1.789	-1.584	-1.622	-1.937	0.193
Early norm (37 – 38 weeks)	26.950	0.400	-1.583	-1.486	-0.099	-1.933
Late norm (\geq 39 weeks)	-40.737	0.798	2.522	2.495	1.367	1.697
<i>Pathology during pregnancy*</i>	20.756	0.353	4.225	4.088	2.616	1.447
ASAM	14.863	0.917	1.970	1.981	1.390	1.386

* This is a dichotomous variable set to 1 if one (or more) of the following diseases occurred during pregnancy: Diabetes mellitus, Eclampsia, Hypertension, Placenta Previa.

Table 2: Mean differences of mothers characteristics before and after matching.

Hospital	N. births	N. caesarean sections	N. drops A	N. drops B	N. drops C	N.drops D	N.drops E
	14,757	5,375	0	38	0	0	0

Table 3: Number of dropped treated units.

adjustment) and matched datasets. The effect of caesarean is consistently estimated to be positive: it increases the risk of low Apgar score. It is worth noting that approaches B and C that control for hospital factors show considerably lower estimates than approach A. This may signal a possible overestimation of the effect of caesarean section when hospital confounding effects associated to higher prevalence of this section mode are not taken into account. Similarly, also multilevel and fixed effect propensity score models (approaches D and E) yield a pooled

estimate lower than that of approach A. Clearly, approaches B-C and D-E have a higher mean ASAM than approach A (1.197-1.198 and 1.390-1.386 versus 0.94) and this should be considered the cost of balancing the potential confounders at the hospital level. is not surprising; indeed these two matching strategies are expected to diverge when there is strong imbalance at the hospital level but not globally.

METRICS	STRATEGY	Without match	A	B	C	D	E
<i>Balance</i>							
Drops		0	0	38	0	0	0
ASAM		14.8	0.91	1.97	1.98	1.39	1.38
<i># of outcomes (every 1000 individuals)</i>							
in treated		10.9	10.9	11.0	10.9	10.9	10.9
in untreated		5.2	9.1	9.6	9.7	9.9	9.9
ATT		5.75	1.80	1.40	1.23	1.02	1.07

Table 4: Empirical results for unmatched and matched subsets (strategies A-E). For each strategy: Drops is the number of dropped treated units; ASAM is the average standardized mean difference in covariates values across treated and untreated units; ATT is the mean difference between the number of outcomes in treated and untreated groups.

Simulation study

Motivated by previous empirical analysis we made a simulation experiment which illustrates the implications of different matching strategies when there is unobserved confounding at the cluster level. We followed a semi-empiric simulation strategy (see for example Huber et al. [5]) in the sense that we kept the original set of covariates and introduced an additional hospital level variable (H) to analyze the confounding effect. The variable H is set up constant for all observations in the same hospital. We then simulated the effect of a null, mild and strong confounding effect of H on the balance and the ATT by increasing its coefficient (β_H) in the outcome and treatment equations.

Simulation results show that when there is no unobserved confounding ($\beta_H = 0$) approaches B-E yield a similar average balance, which is only slightly higher than the balance attained in approach A, which is the best approach in this situation. However, when the size of the confounding effect increases, approaches B-E yield considerably lower average balance and bias than approach A and so should be preferred when unobserved confounding at the cluster level is suspected.

4 Concluding remarks

In this paper we discuss the advantages and drawbacks of different techniques to implement propensity score matching with clustered data. We apply these techniques to a population dataset containing information on the birth event in a two year period, clustered in twenty

hospitals. When clusters size are big as in our application and there is potential confounding due to unobserved hospital level variables, an effective approach consists in implementing the matching within clusters or starting with a within matching approach and then use the pooled sample for remaining unmatched cases.

Acknowledgement

We would like to thank the Autonomous Region of Sardinia for providing the anonymized data used in the empirical application.

Bibliography

- [1] Annibale D.J., Hulsey T.C., Wagner C.L. et al. (1995) *Comparative neonatal morbidity of abdominal and vaginal deliveries after uncomplicated pregnancies*, Arch Pediatr Adolesc Med Aug, **149**(8),862-7.
- [2] Arpino, B. and Mealli, F. (2011) *The specification of the propensity score in multilevel observational studies*, Computational Statistics and Data Analysis, **55**, 1770 -1780.
- [3] Bragg, F., Cromwell, D.A., Edozien,L. et al. (2010) *Variation in rates of caesarean section among English NHS trusts after accounting for maternal and clinical risk: cross sectional study*. British Medical Journal; doi:10.1136/bmj.c506.
- [4] Caceres, I.A., Arcaya, M., Declercq, E. et al. (2013) *Hospital Differences in Cesarean Deliveries in Massachusetts (US) 2004-2006: The Case against Case-Mix Artifact*, PLOS ONE **8**(3), doi:10.1371/journal.pone.0057817.
- [5] Huber, M., Lechner, M. and Wunsch, C. (2013) *The performance of estimators based on the propensity score*, Journal of Econometrics **175**, 1-21.
- [6] Hughes, R.G., Hunt., S.S. and Luft, H.S. (1987) *Effects of surgeon volume and hospital volume on quality of care in hospitals*, Med Care **25**, 489-503.
- [7] Imbens, G.W. (2004) *Nonparametric estimation of average treatment effects under exogeneity: a review*, Review of Economics and Statistics, **86**, 4-30.
- [8] Li, F., Zaslavsky, A. M., and Landrum, M. B.(2013) *Propensity score weighting with multilevel data*, Statistics in Medicine, **32**(19), 3373-3387.
- [9] Rosenbaum, P.R., Rubin, D.B (1983) *The central role of the propensity score in observational studies for causal effects*, Biometrika, **70**, 41-55.
- [10] Sekhon, J.S. (2011) *Multivariate and Propensity Score Matching Software with Automated Balance Optimization*, Journal of Statistical Software, **42**(7), 1-52.
- [11] Thoemmes, F.J. and West, S.G. (2011) *The use of propensity scores for nonrandomized designs with clustered data*, Multivariate Behavioral Research, **46**(3), 514-543.

Monitoring the shape parameter of a Weibull distribution

Fernanda Figueiredo, *Faculdade de Economia da Universidade do Porto and Centro de Estatística e Aplicações, Universidade de Lisboa*, otilia@fep.up.pt

M. Ivette Gomes, *Centro de Estatística e Aplicações, Faculdade de Ciências, Universidade de Lisboa*, ivette.gomes@fc.ul.pt

Adelaide Figueiredo, *Faculdade de Economia da Universidade do Porto and LIAAD-INESC Porto*, adelaide@fep.up.pt

Abstract. A control chart based on the quantile function to monitor the shape parameter of a Weibull distribution is proposed and its performance is analyzed by Monte Carlo simulation. The importance of monitoring the shape parameter even when the other parameters of the Weibull distribution are assumed known is further enhanced, together with motivating examples.

Keywords. Control charts, Monte Carlo simulations, Quantile function, Statistical Process Control

1 Introduction and Motivation

Even when the mean value and the standard deviation conform to the process specifications, we can have quite different distributions in terms of skewness and kurtosis. As an example we refer the Weibull distribution, with cumulative distribution function (cdf) given by

$$F(x) = 1 - \exp\left(-\left(\frac{x-\lambda}{\delta}\right)^\alpha\right), \quad x \geq \lambda, \quad (1)$$

with $\lambda \in \mathbb{R}$, $\delta > 0$ and $\alpha > 0$, commonly used to model asymmetric positive data. The versatility induced by the shape parameter α enables us to obtain different distributional shape, as can be seen from its probability density function (pdf), $f(x) = dF(x)/dx$, represented in Figure 1 (left), for $\lambda = 0$ and $\delta = 1$. This versatility has contributed for its prominent role in many areas of research. In particular, the Weibull distribution is commonly used in many life testing and reliability studies, allowing us to obtain a failure rate function $h(x) := f(x)/(1 - F(x)) = (\alpha/\delta) ((x - \lambda)/\delta)^{\alpha-1}$, $x \geq \lambda$, with different monotonic behavior, according to the value of α : if $\alpha = 1$ we have a constant failure rate (CFR) function, if $\alpha < 1$ we obtain a decreasing failure rate (DFR) function, and if $\alpha > 1$ we have an increasing failure rate (IFR) function. In

Economics and Social Sciences, the Weibull distribution has also been frequently used to describe income distributions, to define inequality indexes of wealth, and as a survival model of firms, for instance. But we can also find applications of the Weibull distribution in other areas, such as in Health Sciences, Hydrology and Meteorology. It should also be noted that the coefficients of skewness and kurtosis of the Weibull distribution depend only on the shape parameter, α . The illustrative situations previously mentioned undoubtedly reveal the importance in detecting possible changes in the shape parameter of the Weibull distribution.

Apart from the importance of the three-parameter Weibull distribution in several applications, even with the expense of higher difficulties in which concerns the parameter's estimation when the location parameter is unknown (see, for instance, [6], [25], [10]), we are going to focus on the most commonly used two-parameter Weibull distribution, with location at zero, and cdf obtained from (1) by replacing λ by 0. In this case we have several possible and efficient estimators for the scale and the shape parameters (see [11], [1], [2], [27], [13] and [14], among others). In the literature several control charts associated with Weibull processes are also provided: among others, we refer control charts for monitoring the scale or the shape parameters ([24], [23], [22], [21]), and the percentiles ([16], [15], [3], [7]). Here we propose a new control chart for the shape parameter based on the quantile function, along the lines of [4], although the methodology used to define the chart can be extended to detect shifts in other parameters, even associated with the three-parametric family. The paper is organized as follows. Section 2 provides some information about the quantile and the sample quantile functions here considered. Section 3 presents a control chart based on the quantile function and Section 4 concludes with the analysis of the control chart performance and some comments.

2 Quantile Function and Sample Quantile Function

Definition 2.1. *Given a general cdf $F(x)$ of a random variable (rv) X , the quantile function $Q(u)$, is defined by*

$$Q(u) = F^{-1}(u) = \inf \{x : F(x) \geq u\}, \quad 0 \leq u \leq 1.$$

Remark 6.

$Q(u)$ can be used to define some (non)parametric location, scale and position measures of a rv X , such as the median $Q(0.5)$, the interquartile range $Q(0.75) - Q(0.25)$, the percentiles $Q(p/100)$, the mean value $E(X) = \int_0^1 Q(u) du$ and the variance $V(X) = \int_0^1 (Q(u) - E(X))^2 du$, among others. For other details see [17] and [18].

For the standard Weibull($\lambda = 0, \delta = 1, \alpha$) \equiv Weibull(α) distribution, the quantile function, $Q_\alpha(u)$, is given by

$$Q_\alpha(u) = (-\ln(1-u))^{\frac{1}{\alpha}}, \quad 0 \leq u \leq 1. \quad (2)$$

In Figure 1 (right) we picture the quantile function for the Weibull(α) distribution. From this figure we observe that small changes in α result in large differences in the distribution in terms of skew and tail-weight, less evident for the Weibull distributions with $\alpha < 1$. Even for small values of u we can observe differences between the quantile functions. Finally, the distributional shape of the Weibull models with $\alpha \leq 1$ is very different from the one obtained for $\alpha > 1$. Let (X_1, \dots, X_n) be a sample of size n of a rv X and $(X_{1:n} \leq \dots \leq X_{n:n})$ the sample of associated

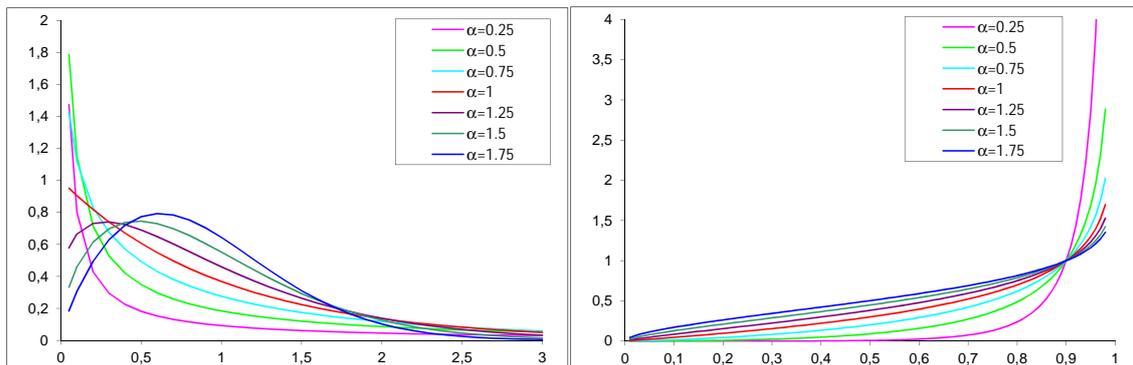


Figure 1: Probability density function (left) and quantile function (right) for the Weibull distribution with $\alpha = 0.25, 0.5, 0.75, 1, 1.25, 1.5, 1.75$.

ascending order statistics. For the sample (empirical) quantile function, $\tilde{Q}(u)$, there are several possible definitions. Here we consider the following one:

Definition 2.2. *The sample (empirical) quantile function is the piecewise linear function:*

$$\tilde{Q}(u) = n \left(\frac{i}{n} - u \right) X_{i-1:n} + n \left(u - \frac{i-1}{n} \right) X_{i:n} \quad \text{for} \quad \frac{i-1}{n} \leq u \leq \frac{i}{n}, \quad i = 1, \dots, n. \quad (3)$$

3 Control chart based on the quantile function

To implement any control chart the nominal process parameters must be either assumed known (given from past experience with similar processes or fixed from engineering specifications) or estimated from a Phase I reference sample, taken when the process is assumed stable and in-control. The estimation of such parameters will have effect on the performance of the control chart (see, for instance, [8]). Here we only consider the known parameter's case, and without loss of generality, a standard Weibull distribution associated with the data process X , with shape parameter $\alpha > 0$, that may change from the in-control value α_0 , to the out-of-control value α_1 .

The control chart we propose to detect changes in the shape parameter α is based on the test of hypothesis

$$H_0 : X \sim F(x, \alpha_0) \text{ versus } H_1 : X \sim F(x, \alpha_1), \quad \alpha_1 \neq \alpha_0,$$

where $F(x, \alpha)$ denotes the cdf of X , and the associated control statistic is defined by

$$\chi^2 = (\tilde{\mathbf{Q}} - \mathbf{Q}_0)' \Sigma_0^{-1} (\tilde{\mathbf{Q}} - \mathbf{Q}_0), \quad (4)$$

where \mathbf{Q}_0 denotes a vector of quantile function values associated with the in-control distribution evaluated in r points $u_i, i = 1, \dots, r, 0 < u_1 < \dots < u_r < 1$, where we must monitor the quantile function and $\tilde{\mathbf{Q}}$ denotes a vector of the sample quantile function evaluated in the same values u_i .

If there exist large differences between the observed values of the sample quantile function and the expected values, we conclude that the underlying data distribution has shifted. Thus,

we have the following decision rule: Reject H_0 if $\chi_{obs}^2 \geq \chi_{1-p}$, i.e. when the observed value of the control statistic is greater or equal to the $(1-p)$ -probability quantile of the control statistic distribution, χ_{1-p} , with $p = 1/ARL_0$ and ARL_0 denoting the desired in-control Average Run Length (ARL) of the chart.

Following Mosteller [12] and Kubat and Epstein [9] the distribution of the control statistic in (4) can be approximated with either a central or a non-central chi-square distribution. They state that under H_0 , the vector $\tilde{\mathbf{Q}} = (\tilde{Q}(u_1), \dots, \tilde{Q}(u_r))$ has approximately, for a large sample of size n , a r -dimensional normal distribution with a mean vector $\mathbf{Q}_0 = [\mu_i]_{r \times 1}$, with $\mu_i = Q_0(u_i)$ and a variance-covariance matrix $\Sigma_0 = [\sigma_{ij}]_{r \times r}$, with

$$\sigma_{ij} = \frac{u_i(1-u_j)}{nf(Q_0(u_i))f(Q_0(u_j))} \quad \text{for } i \leq j \quad \text{and} \quad \sigma_{ij} = \sigma_{ji}.$$

Then, the control statistic has approximately a chi-square distribution with r degrees of freedom (d.f.). Under H_1 , the control statistic has approximately a noncentral chi-square distribution with r d.f. and noncentrality parameter $\eta(\alpha) = (\mathbf{Q}_\alpha - \mathbf{Q}_0)' \Sigma_0^{-1} (\mathbf{Q}_\alpha - \mathbf{Q}_0)$, with \mathbf{Q}_α denoting the vector of quantile function values defined in equation (2). Consequently, the ARL of the chart is approximately given by

$$ARL(\alpha) = \frac{1}{1 - F_{\chi_{r,\eta(\alpha)}^2}(\chi_{r,1-p}^2)}, \quad \alpha > 0,$$

where $F_{\chi_{r,\eta(\alpha)}^2}$ denotes the cdf of the noncentral chi-square rv with r d.f. and noncentrality parameter $\eta(\alpha) = (\mathbf{Q}_\alpha - \mathbf{Q}_0)' \Sigma_0^{-1} (\mathbf{Q}_\alpha - \mathbf{Q}_0)$, and $\chi_{r,1-p}^2$ denotes the $(1-p)$ -probability quantile of the chi-square distribution with r d.f.

In our case, i.e., for the Weibull(α_0), the vector $\mathbf{Q}_0 = [\mu_i]$ and the matrix $\Sigma_0 = [\sigma_{ij}]$ have the following elements, for $1 \leq i \leq j \leq r$:

$$\mu_i = (-\ln(1-u_i))^{\frac{1}{\alpha_0}} \quad \text{and} \quad \sigma_{ij} = \frac{u_i(1-u_i)^{-1}}{n\alpha_0^2 (\ln(1-u_i) \ln(1-u_j))^{1-\frac{1}{\alpha_0}}}.$$

Remark 7.

The choice of the tuning parameters u_i , $1 \leq i \leq r$, where to monitor the quantile function must take into consideration the process parameters of most interest to be monitored, or if we have candidates for the alternative distribution, in our case the magnitudes of the shift in the shape parameter we must detect, we must evaluate the corresponding quantile functions in values where they most differ. The distribution of the control statistic in (4), and consequently, the performance of the chart, also depends on the number r of the tuning parameters. To avoid taking decisions on the basis of a very conservative test, r should be small.

4 Control chart performance and some comments

To illustrate the performance of the previous control chart we consider data from the following in-control distributions: Weibull(α_0), $\alpha_0 = 0.5, 0.75, 1, 1.25, 1.5, 2$, corresponding to distributions with different failure rate behavior and different skew and kurtosis (see Table 1).

α_0	0.5	0.75	1	1.25	1.5	2
Failure rate	DFR	DFR	CFR	IFR	IFR	IFR
Skewness	6.619	3.121	2.000	1.430	1.072	0.631
Kurtosis	87.720	18.987	9.000	5.802	4.390	3.245

Table 1: Type of failure rate, skewness and kurtosis of the in-control distributions, Weibull(α_0).

We have taken samples of size $n = 25$ (this value is a bit larger than it is usual in Statistical Quality Control, but not large in other applications, for instance, in Economics) and we have chosen $r = 3$ tuning parameters, more precisely, $u_1 = 0.915$, $u_2 = 0.955$ and $u_3 = 0.995$. The piecewise linear quantile function in (3) lead us to the values $\tilde{Q}(u_1) = 0.125X_{22:25} + 0.875X_{23:25}$, $\tilde{Q}(u_2) = 0.125X_{23:25} + 0.875X_{24:25}$ and $\tilde{Q}(u_3) = 0.125X_{24:25} + 0.875X_{25:25}$.

The sample size $n = 25$ was chosen taking into consideration that quantile-based estimation is adequate for large samples and attempting to validate the approximation of the distribution of the statistic in (4) to a chi-square. The selection of r and the vector \mathbf{u} of the values u_i was done on the basis of the quantile function pictured in Figure 1 (left), choosing u_i values in the region where the quantile functions under H_0 and H_1 most differ, but also taking into consideration the works of Dubey [2] and Hassanein [6], that suggest estimators based on a few very high and very low sample percentiles (although the situation under study is not exactly the same). In [2], efficient percentile-based estimators for the Weibull parameters are proposed: an estimator of the shape parameter based on the 17th and the 97th sample percentiles, which does not depend upon any knowledge about the scale; an estimator for the scale based on the 40th and the 82nd sample percentiles, in case of unknown shape parameter; the 24th and 93rd sample percentiles that jointly are efficient estimators for both the scale and the shape parameters. In [6], the optimum spacings of the few (2, 4 or 6) sample quantiles used in estimators of the location and the scale parameters of a Weibull distribution when the shape parameter is known, also led to the choice of very high and low sample percentiles, between 86th and 98th, and 1th and 14th.

As we are working with very asymmetric models, we have decided to simulate the in-control distribution of the control statistic in order to validate the previously advanced approximated chi-square distribution. We have considered a sample of 1000000 values for each scenario ($F(\alpha_0)$, n , $Q_0(u_i)$, $\tilde{Q}(u_i)$ and u_i , $i = 1, \dots, 3$). The analysis of the simulated distribution lead us to conclude that the suggested approximation to the χ_3^2 distribution is not very accurate in this case. Table 2 presents the $(1 - p)$ -probability quantiles of the simulated control statistic distribution when we consider samples of size $n = 25, 100$, and of the χ_3^2 distribution for comparison. As we can observe the high quantiles of the simulated distribution are, in general, very different from the ones obtained for the χ_3^2 distribution. Although these differences get smaller as n increases, if the shape parameter is very small the differences persist large, even for $n = 100$. Indeed we have considered other combinations of n , r and \mathbf{u} , such as, $n = 20$, $r = 5$, $\mathbf{u} = (0.475, 0.525, 0.625, 0.725, 0.975)$ and $\mathbf{u} = (0.625, 0.675, 0.725, 0.775, 0.975)$, but we have obtained similar conclusions.

Therefore, for each scenario we have considered the 99.5% probability quantile of the simulated distribution, in order to have an $ARL_0 \simeq 200$, and then, we have computed by simulation the in-control ARL of the chart. Similarly, we have obtained the out-of-control ARL by simulation. A bit surprisingly, we get a control chart very efficient only to detect decreases in

n	25					100				
ARL ₀	100	200	370.4	500	1000	100	200	370.4	500	1000
$1 - p$	99%	99.5%	99.73%	99.8%	99.9%	99%	99.5%	99.73%	99.8%	99.9%
Weibull(0.5)	13.697	19.348	25.704	29.078	38.177	13.057	16.754	20.519	26.685	27.602
Weibull(0.75)	9.227	12.243	15.514	17.187	21.440	10.252	12.752	15.254	16.566	19.757
Weibull(1)	7.796	10.064	12.402	13.644	16.697	9.417	11.471	13.489	14.563	17.186
Weibull(1.25)	7.365	9.198	11.083	12.120	14.624	9.162	10.942	12.738	13.689	16.085
Weibull(1.5)	7.368	8.979	10.485	11.395	13.642	9.094	10.742	12.417	13.270	15.448
Weibull(2)	7.267	8.872	10.230	11.012	12.898	9.102	10.665	12.181	12.979	14.942
χ_3^2	11.345	12.838	14.156	14.796	16.266	11.345	12.838	14.156	14.796	16.266

Table 2: $(1 - p)$ -probability quantiles of the simulated control statistic distribution and of the χ_3^2 distribution.

the shape parameter, i.e., from α_0 to $\alpha < \alpha_0$. Looking again to the values presented in Table 2 we observe that the value of a given quantile increases (decreases) substantially with the skew and the kurtosis (with the value of the shape parameter α) of the underlying distribution. Thus, when the value of the shape parameter decreases, the values of the control statistic easily overpass the $(1 - p)$ -probability quantile of the in-control distribution of the control statistic, and hardly overpass it when the shape parameter increases. As we can observe from the ARL values presented in Table 3, the control chart is very efficient to detect decreases in the shape parameter of a Weibull distribution, or in other words, to detect increases in the asymmetry and kurtosis of the underlying data distribution in comparison with the reference distribution, providing simulated ARL values very small.

α	<u>0.5</u>	0.45	0.4	0.35	0.3	0.25	0.2	0.15			
ARL	<u>202.9</u>	38.4	10.0	3.7	1.9	1.3	1.1	1.0			
α	<u>0.75</u>	0.7	0.65	0.6	0.55	0.5	0.4	0.3	0.25		
ARL	<u>202.3</u>	66.8	24.8	10.5	5.2	3.0	1.5	1.1	1.0		
α	<u>1</u>	0.9	0.8	0.75	0.7	0.6	0.4	0.3	0.25		
ARL	<u>202.4</u>	42.2	11.1	6.4	4.0	2.0	1.3	1.1	1.0		
α	<u>1.25</u>	1.2	1.15	1.1	1.05	1	0.75	0.5	0.25		
ARL	<u>205.4</u>	110.6	60.6	34.0	19.7	11.9	2.0	1.1	1.0		
α	<u>1.5</u>	1.45	1.4	1.35	1.3	1.25	1.2	1.1	1	0.75	0.5
ARL	<u>203.9</u>	127.8	79.2	49.2	30.9	19.8	13.0	6.1	3.4	1.4	1.0
α	<u>2</u>	1.9	1.8	1.75	1.7	1.6	1.5	1.25	1	0.75	0.5
ARL	<u>202.0</u>	116.4	60.9	43.4	31.0	16.1	8.8	2.7	1.4	1.1	1.0

Table 3: ARL of the chart implemented to detect decreases in the shape parameter of the Weibull distribution, from $\alpha_0 = 0.5, 0.75, 1, 1.25, 1.5, 2$ to α , based on the simulated distribution of the control statistic. The values α_0 and $ARL_0 \simeq 200$ are underlined.

Acknowledgement

Research partially supported by National Funds through **FCT**—Fundação para a Ciência e a Tecnologia, within projects PEst-OE/MAT/UI0006/2014 (CEA/UL) and FCOMP-01-0124-

FEDER-037281, and by the ERDF European Regional Development Fund through the COMPETE Programme (operational programme for competitiveness).

Bibliography

- [1] Cohen, A. C. (1965) *Maximum likelihood estimation in the Weibull distribution based on complete and on censored samples*. *Technometrics*, **7**(4), 579–588.
- [2] Dubey, S. D. (1967) *Some percentile Estimators for Weibull Parameters*. *Technometrics*, **9**, 119–129.
- [3] Erto, P., Pallotta, G. and Park, S. H. (2008) *An example of data technology product: a control chart for Weibull processes*. *International Statistical Review*, **76**(2), 157–166.
- [4] Grimshaw, S. and Alt, F. (1997) *Control Charts for Quantile Function Values*. *J. Qual. Technol.*, **29**, 1–7.
- [5] Guo, B. and Wang, B. X. (2014) *Control Charts For Monitoring the Weibull Shape Parameter Based On Type-II Censored Sample*. *Qual. Reliab. Engng. Int.*, **30**, 13–24.
- [6] Hassanein, K. M. (1971) *Percentile estimators for the parameters of the Weibull distribution*. *Biometrika*, **58**(3), 673–676.
- [7] Huang, X. and Pascual, F. (2011) *ARL-unbiased control charts with alarm and warning lines for monitoring Weibull percentiles using first-order statistic*. *J. of Statistical Computation and Simulation*, **81**, 1677–1696.
- [8] Jensen, W. A., Jones-Farmer, L. A., Champ, C. W. and Woodall, W. H. (2006) *Effects of parameter estimation on control chart properties: a literature review*. *J. Qual. Technol.*, **38**, 349–364.
- [9] Kubat, P. and Epstein, B. (1980) *Estimation of Quantiles of Location-Scale Distributions Based on Two or Three Order Statistics*. *Technometrics*, **22**, 575–581.
- [10] Lockhart, R. A. and Stephens, M. A. (1994) *Estimation and Tests of fit for the three-parameter Weibull distribution*. *J. R. Stat. Soc.*, **56**(3), 1491–500.
- [11] Menon, M. V. (1963) *Estimation of the Shape and Scale parameters of the Weibull distribution*. *Technometrics*, **5**(2), 175–182.
- [12] Mosteller, F. (1946) *On some useful inefficient statistics*. *Ann. Math. Statist.*, **17**, 377–408.
- [13] Murthy, V. K. and Swartz, G. B. (1975) *Estimation of Weibull parameters from two-order statistics*. *J. R. Stat. Soc., series B (methodological)*, **37**(1), 96–102.
- [14] Newby, M. J. (1980) *The properties of moment estimators for the Weibull distribution based on the sample coefficient of variation*. *Technometrics*, **22**(2), 187–194.
- [15] Nichols, M. D. and Padgett, W. J. (2006) *A Bootstrap Control Chart for Weibull Percentiles*, *Qual. Reliab. Engng. Int.*, **22**, 141–151.
- [16] Padgett, W. J. and Spurrier, J. D. (1990) *Shewhart-type charts for percentiles of strength distributions*. *J. Qual. Technol.*, **22**, 283–288.

- [17] Parzen, E. (2004) *Quantile Probability and Statistical Data Modeling*. Statistical Science, **19**, 652–662.
- [18] Parzen, E. (1979) *Nonparametric Statistical Data Modeling*. J. Am. Stat. Assoc., **74**, 105–121.
- [19] Pascual, F. (2010) *EWMA Charts for the Weibull Shape Parameter*. J. Qual. Technol., **42**(4), 400–416.
- [20] Pascual, F. (2013) *Individual and Moving Ratio Charts for Weibull Processes*. In Stochastic Orders in Reliability and Risk: In honor of Professor Moshe Shaked, Lecture Notes in Statistics, Springer.
- [21] Pascual, F. and Li, S. (2012) *Monitoring the Weibull Shape Parameter by Control Charts for the Sample Range of Type II Censored Data*. Qual. Reliab. Engng. Int., **28**(2), 233–246.
- [22] Pascual, F. and Nguyen, D. (2011) *Moving Range Charts for Monitoring the Weibull Shape Parameter with Single-Observation Samples*. Qual. Reliab. Engng. Int., **27**, 905–919.
- [23] Pascual, F. and Zhang, H. (2011) *Monitoring the Weibull Shape Parameter by Control Charts for the Sample Range*. Qual. Reliab. Engng. Int., **27**, 15–25.
- [24] Ramalhoto, M. F. and Morais, M. (1999) *Shewhart control charts for the scale parameter of a Weibull control variable with fixed and variable sampling intervals*. J. of Applied Statistics, **26**, 129–160.
- [25] Smith, R. L. and Naylor, J.C. (1987) *A Comparison of Maximum Likelihood and Bayesian Estimators for the Three-Parameter Weibull Distribution*. J. R. Stat. Soc., series C (Applied Statistics), **36**(3), 358–369.
- [26] Sze, C. and Pascual, F. (2013) *Control Charts for Monitoring Weibull Distribution*. In Technical Report Series, Department of Mathematics, Washington State University.
- [27] Thoman, D. R, Bain, L. J. and Antle, C. E. (1969) *Inferences on the Parameters of the Weibull Distribution*. Technometrics, **11**, 445–460.

How far is the Corpus Callosum of an Average Individual from Albert Einstein's?

Mingfei Qiu, *Florida State University*, mingfeiqiu@stat.fsu.edu

Vic Patrangenaru, *Florida State University*, vic@stat.fsu.edu

Leif Ellingson, *Texas Tech University*, leif.ellingson@ttu.edu

Abstract. The optic nerves meet at the Optic Chiasma (OC) in a midsagittal plane at the base of the brain. There, half of the axons from each nerve cross over into the other nerve, so that some visual information from the left eye travels in parallel with information from the right eye within each of the two nerves. The blending of the two eye images allows one to perceive the projective shape of the scene. The Corpus Callosum (CC) connects the two cerebral hemispheres and facilitates interhemispheric communication. It is the largest white matter structure in the brain. Albert Einstein's brain was removed shortly after his death, weighted, dissected and photographed by a pathologist. High resolution versions of those pictures were quantitatively studied in two recent papers listed in the references. Contours of CC midsagittal sections are extracted from MRI images. Given that Einstein passed at 76, we extracted a small subsample of CC brain contour, in the age group 64-83, and tested how far is the average CC contour from Einstein's. The analysis was performed on the Hilbert manifold of planar contours, following the methodology recently developed by the authors.

1 Albert Einstein's Brain Image Data and Shape of Euclidean Contours

Einstein's brain was removed shortly after his death (most likely without prior family consent), weighted, dissected and photographed by a pathologist. Among other pictures, a digital scan of a picture of the General Relativity creator's half brain taken at the autopsy is displayed below; we extracted the contour of the CC from this Einstein's brain image, the shape of which would be set as a null hypothesis in our testing problem (see Figure 1). Fletcher (2013)[8] extracted contours of CC midsagittal sections from MRI images, to study possible age related changes in this part of the human brain. His study points out to certain age related shape changes, in the corpus callosum. Given that Einstein passed at 76, we consider a subsample of corpus callosum

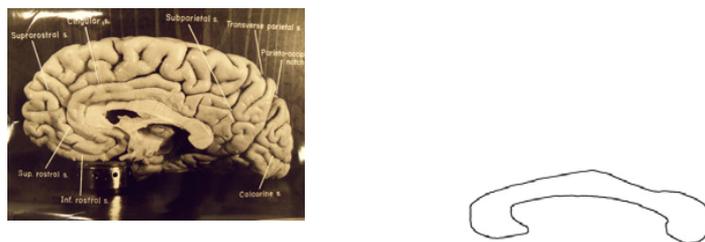


Figure 1: Right hemisphere of Einstein's brain including CC mid-sagittal section (left) and its contour (right).

brain contours from Fletcher(2013)[8], in the age group 64-83, to test how far is the average CC contour from Einstein's. The data is displayed in Figure 2.

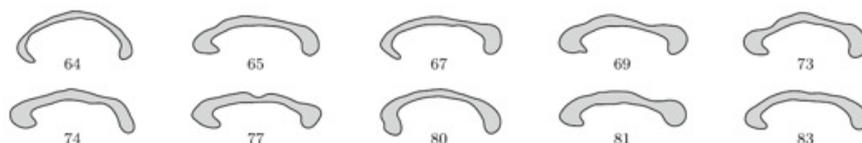


Figure 2: Corpus callosum mid-sagittal sections shape data, in subjects ages - 65 to 83

We consider contours, boundaries of 2D topological disks in the plane. To keep the data analysis stable, and to assign a *unique* labeling, we make the *generic* assumption that across the population there is a unique anatomical or geometrical landmark starting point p_0 on such a contour of perimeter one, so that the label of any other point p on the contour is the “counterclockwise” travel time at constant speed from p_0 to p . A *regular contour* $\tilde{\gamma}$ is regarded as the range of a piecewise differentiable *regular* arclength parameterized function $\gamma : [0, L] \rightarrow \mathbb{C}, \gamma(0) = \gamma(L)$, that is one-to-one on $[0, L)$. Two contours $\tilde{\gamma}_1, \tilde{\gamma}_2$ have the same *direct similarity shape* if there is a direct similarity $S : \mathbb{C} \rightarrow \mathbb{C}$, such that $S(\tilde{\gamma}_1) = \tilde{\gamma}_2$. Two regular contours $\tilde{\gamma}_1, \tilde{\gamma}_2$ have the same similarity shape if their centered counterparts satisfy to $\tilde{\gamma}_{2,0} = \lambda \tilde{\gamma}_{1,0}$, for some $\lambda \in \mathbb{C} \setminus \{0\}$. Therefore Σ_2^{reg} , set of all direct similarity shapes of regular contours, is a dense and open subset of $P(\mathbf{H})$, the projective space corresponding to the Hilbert space \mathbf{H} of all square integrable centered functions from S^1 to \mathbb{C} . (see Ellingson et al (3013)[5]).

All CC contours were matched by selecting an initial anatomic landmark on the lower posterior CC and traveling along the contour of the midsection in arclength time. The algorithm 1.1 below was used for the matching.

Algorithm 1.1.

This algorithm randomly selects k matched sampling points from the uniform distribution over $[0, L_j)$ for a sample of n contours, where L_j is the perimeter of contour j and $j = 1, 2, \dots, n$.

Step 1 *Select a common starting point for all n contours such that this represents the contour at time s_1 .*

Step 2 Generate $s_2, s_3, \dots, s_k \sim \text{Uniform}(0, 1)$. Sort these in increasing order and relabel them as t_1, t_2, \dots, t_k .

Step 3 Obtain matched sampling points for each contour.

For $j=1:n$

Evaluate contour j at times $t_1 * L_j, t_2 * L_j, \dots, t_k * L_j$ to obtain $z(t_1 * L_j), z(t_2 * L_j), \dots, z(t_k * L_j)$

End

For the matching algorithm applied to the CC data, see Figure 3.

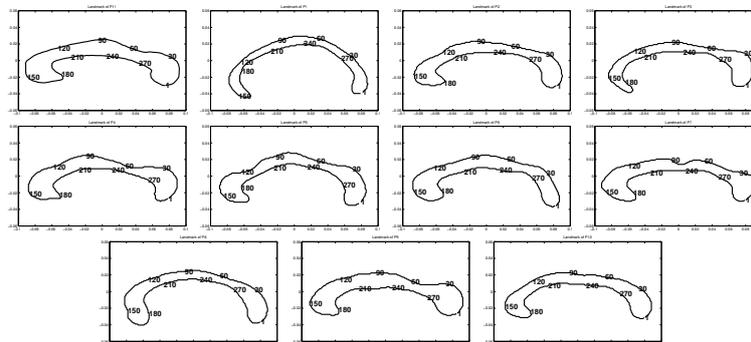


Figure 3: Matched sampling points on midsagittal sections in for CC data (Einstein's is the upper left CC).

2 Asymptotic distributions for means on Hilbert manifolds

Definition 2.1. Assume \mathbf{H} is a separable, infinite dimensional Hilbert space over the reals. A chart on a separable metric space (M, ρ) is a one to one homeomorphism $\varphi : U \rightarrow \varphi(U)$ defined on an open subset U of M to a Hilbert space \mathbf{H} . A Hilbert manifold is a separable metric space M , that admits an open covering by domain of charts, such that the transition maps $\varphi_V \circ \varphi_U^{-1} : \varphi_U(U \cap V) \rightarrow \varphi_V(U \cap V)$ are differentiable.

The projective space $P(\mathbf{H})$ of a Hilbert space \mathbf{H} , set of all one dimensional linear subspaces of \mathbf{H} , has a natural structure of Hilbert manifold modeled over \mathbf{H} . Define the distance between two vector lines as their angle, and, given a line $\mathbb{L} \subset \mathbf{H}$, a neighborhood $U_{\mathbb{L}}$ of \mathbb{L} can be mapped via a homeomorphism $\varphi_{\mathbb{L}}$ onto an open neighborhood of the orthocomplement \mathbb{L}^{\perp} by using the decomposition $\mathbf{H} = \mathbb{L} \oplus \mathbb{L}^{\perp}$. Then if $\mathbb{L}_1 \perp \mathbb{L}_2$, the map is a $\varphi_{\mathbb{L}_1} \circ \varphi_{\mathbb{L}_2}^{-1}$ is differentiable map between open subsets in \mathbb{L}_1^{\perp} , respectively in \mathbb{L}_2^{\perp} .

Definition 2.2. An embedding of a Hilbert manifold M in a Hilbert space \mathbb{H} is a one-to-one differentiable function $j : M \rightarrow \mathbb{H}$, such that for each $x \in M$, the differential $d_x j$ is one to one, and the range $j(M)$ is a closed subset of \mathbb{H} and the topology of M is induced via j by the topology of \mathbb{H} .

As an example, we consider the Veronese-Whitney (VW) embedding $j : P(\mathbf{H})$ in $\mathcal{L}_{HS} = \mathbf{H} \otimes \mathbf{H}$, introduced in the finite dimensional case by Kent (1992)[9], given by

$$j([\gamma]) = \frac{1}{\|\gamma\|^2} \gamma \otimes \gamma^*, [\gamma] \in P(\mathbf{H}). \tag{1}$$

Definition 2.3. *If $j : \mathcal{M} \rightarrow \mathbb{H}$ is an embedding and given a random object X on \mathcal{M} , the associated Fréchet function is $\mathcal{F}_j(x) = E(\|j(X) - j(x)\|^2)$. The set of all minimizers of \mathcal{F}_j is the extrinsic mean set of X . If the extrinsic mean set has one element only, that element is called the extrinsic mean and is labeled μ_j .*

Lemma 2.4. *(Patrangenaru(1998)[12]) Consider a random object X on \mathcal{M} and assume $j(X)$ has the mean vector μ . Then the extrinsic mean set is the set of all points $x \in \mathcal{M}$, such that $j(x)$ is at minimum distance from μ . (iii) In particular, μ_j exists if there is a unique point on $j(\mathbf{M})$ at minimum distance from μ , the projection $P_j(\mu)$ of μ on $j(\mathbf{M})$, and in this case $\mu_j = j^{-1}(P_j(\mu))$.*

The Veronese-Whitney mean (VW mean) is the extrinsic mean for a random object $X = [\Gamma]$ on $P(\mathbf{H})$ with respect to the VW embedding, and it exists if and only if $E(\frac{1}{\|\Gamma\|^2} \Gamma \otimes \Gamma^*)$ has a simple largest eigenvalue. In this case, the VW mean is $\mu_j = [\gamma]$, where γ is an eigenvector for this eigenvalue.

3 Test statistic for the one sample neighborhood hypothesis

Assume Σ_j is the extrinsic covariance operator of a random object X on the Hilbert manifold \mathcal{M} , with respect to the embedding $j : \mathcal{M} \rightarrow \mathbb{H}$ (see Ellingson et al.(2013)[5]). Let \mathbf{M}_0 be a compact submanifold of \mathcal{M} . Let $\varphi_0 : \mathcal{M} \rightarrow \mathbb{R}$ be the function

$$\varphi_0(p) = \min_{p_0 \in \mathbf{M}_0} \|j(p) - j(p_0)\|^2, \tag{2}$$

and let $\mathbf{M}_0^\delta, \mathbb{B}_0^\delta$ be given respectively by

$$\mathbf{M}_0^\delta = \{p \in \mathcal{M}, \varphi_0(p) \leq \delta^2\}, \mathbf{B}_0^\delta = \{p \in \mathcal{M}, \varphi_0(p) = \delta^2\}. \tag{3}$$

Since φ_0 is Fréchet differentiable and all small enough $\delta > 0$ are regular values of φ_0 , it follows that \mathbf{B}_0^δ is a Hilbert submanifold of codimension one in \mathcal{M} . Let ν_p be the normal space at a points $p \in \mathbf{B}_0^\delta$, orthocomplement of the tangent space to \mathbf{B}_0^δ at p . We define $\mathbb{B}_0^{\delta,X}$

$$\mathbb{B}_0^{\delta,X} = \{p \in \mathbf{B}_0^\delta, \Sigma_j|_{\nu_p} \text{ is positive definite}\}. \tag{4}$$

Definition 3.1. *The neighborhood hypothesis consists in the following two alternatives:*

$$H_0 : \mu_j \in M_0^\delta \cup \mathbb{B}_0^{\delta,X} \text{ vs. } H_1 : \mu_j \in (M_0^\delta)^c \cap (\mathbb{B}_0^{\delta,X})^c. \tag{5}$$

Munk et al. (2008)[11] show that, in the case of random objects on Hilbert spaces, the test statistic for these types of hypotheses has an asymptotically standard normal distribution for large sample sizes. Here, we consider neighborhood hypothesis testing for the particular situation in which the submanifold \mathbf{M}_0 consists of a point m_0 on \mathcal{M} . We set $\varphi_0 = \varphi_{m_0}$, and since $T_{m_0}\{m_0\} = 0$ we will prove the following result (see Ellingson et. al. (2013) [5]).

Theorem 3.2. *If $M_0 = \{m_0\}$, the test statistic for the hypotheses specified in (3) has an asymptotically standard normal distribution and is given by:*

$$T_n = \sqrt{n}\{\varphi_{m_0}(\hat{\mu}_j) - \delta^2\}/s_n, \quad s_n^2 = 4\langle \hat{\nu}, S_{j,n}\hat{\nu} \rangle \text{ where} \quad (6)$$

$$S_{j,n} = \frac{1}{n} \sum_{i=1}^n (\tan_{\hat{\mu}} d_{\overline{j(X)_n}} P_j(j(X_i) - \overline{j(X)_n})) \otimes (\tan_{\hat{\mu}} d_{\overline{j(X)_n}} P_j(j(X_i) - \overline{j(X)_n})) \quad (7)$$

is the extrinsic sample covariance operator for $\{X_i\}_{i=1}^n$, and

$$\hat{\nu} = (d_{\hat{\mu}_{j,n}} j)^{-1} \widehat{\tan}_{j(\hat{\mu}_{j,n})}(j(m_0) - j(\hat{\mu}_{j,n})). \quad (8)$$

4 Neighborhood hypothesis for the mean shape of an Euclidean contour

Given any VW-nonfocal probability measure Q on $P(\mathbf{H})$, from Section 3 we see that if $\gamma_1, \dots, \gamma_n$ is a sample from Γ , then $\hat{\mu}_{j,n}$ is the projective point of the eigenvector corresponding to the largest eigenvalue of $\frac{1}{n} \sum_{i=1}^n \frac{1}{\|\gamma_i\|^2} \gamma_i \otimes \gamma_i^*$. Given n i.i.d.r. objects (i.i.d.r.o.'s) from a VW-nonfocal distribution on $P(\mathbf{H})$, the asymptotic distribution of $\overline{j(X)_n}$ converges as follows

$$\sqrt{n}(\overline{j(X)_n} - \mu) \rightarrow_d \mathcal{G} \text{ as } n \rightarrow \infty, \quad (9)$$

where \mathcal{G} has a Gaussian distribution $N_{\mathcal{L}_{HS}}(0, \Sigma)$ on \mathcal{L}_{HS} a zero mean and covariance operator Σ . It follows that the projection $P_j : \mathcal{L}_{HS} \rightarrow j(P(\mathbf{H})) \subset \mathcal{L}_{HS}$ is given by

$$P_j(A) = \nu_A \otimes \nu_A^*, \quad (10)$$

where ν_A is the eigenvector of norm 1 corresponding to the largest eigenvalue δ_1^2 of A , $P_j(\mu) = j(\mu_j)$, and $P_j(\overline{j(X)_n}) = j(\hat{\mu}_{j,n})$

Applying the delta method to (9), Ellingson et al.(2013)[5] arrived at a CLT for the VW extrinsic sample mean $\hat{\mu}_{j,n}$. Because of the infinite dimensionality, in practice, a sample estimate for the covariance operator is always degenerate, so studentization does not work. We then reduce the dimensionality via the neighborhood hypothesis methodology. Suppose that $j : P(\mathbf{H}) \rightarrow \mathcal{L}_{HS}$ is the VW embedding in (1) and $\delta > 0$ is a given positive number. Using the notation in Section 3, we now can apply the results above to random shapes of regular contours. Assume $x_r = [\gamma_r]$, $\|\gamma_r\| = 1$, $r = 1, \dots, n$ is a random sample from a VW-nonfocal probability measure Q . Asymptotically the tangential component of the VW-sample mean around the VW-population mean has a complex multivariate normal distribution. In particular, if we extend the CLT for VW-extrinsic sample mean Kendall shapes in Bhattacharya and Patrangenaru (2005)[2], to the infinite dimensional case, the j -extrinsic sample covariance operator $S_{j,n}$, when regarded as an infinite Hermitian complex matrix has the following entries

$$S_{j,n,ab} = n^{-1}(\hat{\delta}_1^2 - \hat{\delta}_a^2)^{-1}(\hat{\delta}_1^2 - \hat{\delta}_b^2)^{-1} \sum_{r=1}^n \langle e_a, \gamma_r \rangle \langle e_b, \gamma_r \rangle^* \quad | \langle e_1, \gamma_r \rangle |^2, a, b = 2, 3, \dots \quad (11)$$

with respect to the complex orthobasis e_2, e_3, e_4, \dots of unit eigenvectors in the tangent space $T_{\hat{\mu}_{j,n}}P(\mathbf{H})$. Recall that this orthobasis corresponds via the differential $d_{\hat{\mu}_{j,n}}$ with an orthobasis

(over \mathbb{C}) in the tangent space $T_{j(\hat{\mu}_{j,n})}j(P(\mathbf{H}))$, therefore one can compute the components $\hat{\nu}^a$ of $\hat{\nu}$ from equation (8) with respect to e_2, e_3, e_4, \dots , and derive for s_n^2 in (6) the following expression

$$s_n^2 = 4 \sum_{a,b=2}^{\infty} S_{E,n,ab} \hat{\nu}^a \overline{\hat{\nu}^b}, \quad (12)$$

where $S_{E,n,ab}$ given in equation (11) are regarded as entries of a Hermitian matrix.

5 Bootstrap confidence regions for means of contours and Corpus Callosum one-sample test

Similarly to the standard arithmetic mean, we see that the extrinsic mean provides a summary of the shapes by reducing the variability (see Ellingson et al. (2013a)[6]). Another plus of the extrinsic mean is that the computation is fast (see Bhattacharya et. al.(2012)[3]).

One method for performing inference, is through nonparametric nonpivotal bootstrap (Efron (1979)[4]). The nonparametric bootstrap algorithm for constructing a confidence region for extrinsic mean contour is given below.

Algorithm 5.1.

INPUT x : (x is $k \times n$ complex matrix with norm one columns); k : number of matched points on contours ; n : number of contours(columns); N : number of bootstraps
OUTPUT CR: confidence region

Step 1 Compute extrinsic mean of x

For $i=1:n$

$$X_i = x_i * x_i^*$$

End

$$\bar{X} = \text{sum}(X)/n$$

μ_{VW} = eigenvector corresponds to the largest eigenvalue of \bar{X}

Step 2 Bootstrap

For $j=1:N$

$$u_1, \dots, u_n = \text{random integer} \sim \text{uniform}(1, n)$$

$$y_{1:n} = x_{u_1}, \dots, x_{u_n}$$

For $i=1:n$

$$Y_i = y_i * y_i^*$$

End

$$\bar{Y} = \text{sum}(Y)/n$$

μ_{BVWj} = eigenvector corresponds to the largest eigenvalue of \bar{Y}

ϕ_{BVWj} = real part of trace($(\bar{Y} - \bar{X})(\bar{Y} - \bar{X})^T$)

End

$cutoff=95\% \text{ quantile of } \phi_{BVMj}$

$$CR=\{\mu_{BVWj} \mid ((\phi_{BVWj} < cutoff) * \mu_{BVWj}) \cap (\mu_{BVWj} \neq 0), j = 1, \dots, N\}$$

We will use the neighborhood hypothesis test on the manifold of planar contours to test if the average shape of the CC in a population of sixty five to eighty three years old people is close to the shape of Einstein's CC. Data in Figure 2 was used to test the hypothesis that the mean CC shape is in a small neighborhood around the shape of Einstein's CC (Figure 5). The closest representatives of the VW sample mean of the shapes of contours of the CC midsections vs the shape of Einstein's CC midsection are displayed in Figure 4. The overlaps of the two contours are rare, which visually shows that the average CC contour shape is significantly different from Einstein's. The maximum value for δ where the test is significant was found to be 0.1367, which

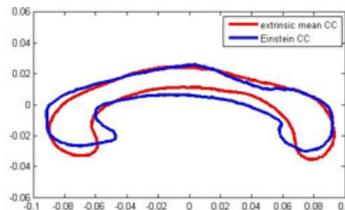


Figure 4: Superimposed icons for 2D direct similarity shapes of CC midsections : sample mean (red) vs Albert Einstein's (blue)

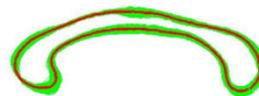


Figure 5: 95% bootstrap confidence region for the extrinsic mean CC contour by 1000 resamples.

is quite large taking into account the fact that the diameter of any finite dimensional complex projective space with the Fubini-Study metric, is $\frac{\pi}{2}$. The corresponding neighborhood for this value of δ is pictured in Figure 5.

Although other recent studies focussed more on size, rather than shape, they tell the same story: an average brain is not close to A. Einstein's brain, although it weighted less than the average (see Falk et al. (2013)[7] and Man et al. (2014)[10]). Our results, while from a different perspective than in Man et. al.(2014)[10], point as well to the fact that A.Einstein's corpus callosum, being relatively thicker, allowed for a better connectivity between the left and right hemispheres of the brain, than the connectivity in the average subject in his group age.

Acknowledgement

We acknowledge support from awards NSF-DMS-1106935 and NSA-MSP-H98230-14-1-0135 and the 2013/2014 LDHD program at SAMSI, RTP, NC. We would like to thank the referees for their suggestions that helped us improve the manuscript.

Bibliography

- [1] Amaral, G. J. A., Dryden, I. L., Patrangenaru, V. and Wood, A.T.A. (2010). Bootstrap confidence regions for the planar mean shape. *J. Statist. Plann. Inference.* **140**, 3026-3034.
- [2] Bhattacharya, R.N. and Patrangenaru, V. (2005). Large sample theory of intrinsic and extrinsic sample means on manifolds- Part II, *Ann. Statist.*, **33**, No. 3, 1211- 1245.
- [3] Bhattacharya, R.N; Ellingson, L; Liu, X; Patrangenaru, V; and Crane, M. (2012) Extrinsic Analysis on Manifolds is Computationally Faster than Intrinsic Analysis, with Application to Quality Control by Machine Vision. *Applied Stochastic Models in Business and Industry.* **28**, 222-235.
- [4] Efron, B. (1979) Bootstrap methods: another look at the jackknife. *Ann. Statist.* **7**, 1–26.
- [5] Ellingson, L., Patrangenaru, V. and Ruymgaart, F. (2013). Nonparametric Estimation of Means on Hilbert Manifolds and Extrinsic Analysis of Mean Shapes of Contours. *Journal of Multivariate Analysis.* **122**, 317–333.
- [6] Ellingson, L., Ruymgaart, F. H., and Patrangenaru, V. (2013a). Data analysis on Hilbert manifolds and shapes of planar contours. *Statistical Models and Methods for non-Euclidean Data with Current Scientific Applications, The 32nd Leeds Annual Statistical Research Workshop 2nd-4th July 2013* (Edited K. V. Mardia and Jochen Voss). 23–27.
- [7] Falk, D., Lepore, F. E. and Noe, A. (2013). The cerebral cortex of Albert Einstein: a description and preliminary analysis of unpublished photographs. *Brain*, **136**, 1304 – 1327
- [8] Fletcher, T. P. (2013) Geodesic Regression and the Theory of Least Squares on Riemannian Manifolds. *Int. J. Comput. Vis.* **105**, 171 - 185.
- [9] Kent, J.T. (1992), New directions in shape analysis. *The Art of Statistical Science, A Tribute to G.S. Watson*, 115–127. Wiley Ser. Probab. Math. Statist. Probab. Math. Statist., Wiley, Chichester, 1992.
- [10] Man, W., Falk, D., Sun, T., Chen, W., Li, J., Yin, J., Zang, D. and Fan M. (2014). The corpus callosum of Albert Einsteins brain: another clue to his high intelligence? *Brain*, **137**, 1 - 8.
- [11] Munk, A., Paige, R., Pang, J., Patrangenaru, V. and Ruymgaart, F. H.(2008). The One and Multisample Problem for Functional Data with Applications to Projective Shape Analysis. *J. of Multivariate Anal.* **99**, 815-833.
- [12] Patrangenaru, V. (1998). Asymptotic Statistics on Manifolds, *Ph.D. Dissertation* , *Indiana University*.

Statistical Registration of Frontal View Gait Silhouette with Application to Gait Analysis

Kosuke Okusa, *Kyushu University*, okusa@design.kyushu-u.ac.jp

Toshinari Kamakura, *Chuo University*, kamakura@indsys.chuo-u.ac.jp

Abstract. We study the problem of analyzing and classifying frontal view gait video data. In this study, we focus on the shape scale changing in the frontal view human gait, we estimate scale parameters using the statistical registration and modeling on a video data. To demonstrate the effectiveness of our method, we apply our model to the frontal view gait authentication. As a result, our model shows good performance for the scale estimation and human gait authentication.

Keywords. Shape analysis, Gait analysis, Scale estimation

1 Introduction

We study the problem of analyzing and classifying frontal view gait video data. A study on the human gait analysis is very important in the fields of the health/sports management, medical research, and the biometrics.

Gait analysis is mainly based on motion capture system and video data. The motion capture system can give the precise measurements of trajectories of moving objects, but it requires the laboratory environments and this system cannot be used in the field study. On the other hand, the video camera is handy to observe the gait motion in the field study.

From the standpoint of health/medical research area. Gage [1] proposed brain paralysis gait analysis using gait video data. Kadaba *et al.* [2] discussed importance of lower limb in the human gait using gait video data too. Many gait analysis have recently analyzing using video analysis software (e.g. Dartfish, Contemplas, Silicon Coach). For example, Borel *et al.* [3] and Grunt *et al.* [4] proposed infantile paralysis gait analysis using lateral view gait video data.

From the standpoint of statistics, Olshen *et al.* [5] proposed the bootstrap estimation for confidence intervals of the functional data with application to the gait cycle data observed by the motion capture system.

However, most studies have not focused on frontal view gait analysis, because such data has many restrictions on analysis based on the filming conditions.

The video data filmed from the frontal view is difficult to analyze, because the subject getting close in to the camera, and data includes the scale-changing parameters [6, 7]. To cope with this, Okusa *et al.* [8] and Okusa & Kamakura [9] proposed a registration for scales of moving object using the method of nonlinear least squares, but Okusa *et al.* [8] and Okusa & Kamakura [9] did not focus on the human leg swing. Okusa & Kamakura [10] focus on the gait analysis using arm and leg swing model with estimated parameters and application to the normal/abnormal gait analysis. However, their models have many of parameters, and it raise calculation cost and instability of parameter estimation.

On the other hand, from the stand point of biometrics, many of this area's researchers mainly using human silhouette shape for the gait authentication. However, they did not focus on the scale registration in the frontal view gait analysis case. In this area, just normalize the human silhouette and it apply to the gait authentication. It is reasonable to suppose that the normalize of the human silhouette lost a lot of gait information.

In this study, we focus on the scale changing of human shape in the frontal view gait analysis, we estimate scale parameters using the statistical registration and modeling on a video data. To demonstrate the effectiveness of our method, we apply our model to the frontal view gait authentication. As a result, our model shows good performance for the scale estimation and gait authentication.

The organization of the rest of the paper is as follows. In section 2, we discuss the advantage and problem of frontal view gait analysis. In section 3, primarily, we describe archetype model proposed in Okusa & Kamakura [11]. Next, we discuss the modified model for the shape scale registration. In section 4, we validate our modified model using the frontal view gait authentication. We conclude with a summary in section 5.

2 Frontal View Gait Data

In this section, we describe an overview of frontal view gait data. Many of gait analysis using lateral view gait data (e.g. Borel *et al.* [3], Grunt *et al.* [4], Barnich & Droogenbroeck [6], Lee *et al.* [7]), because observed data not includes the scale-changing parameters, it is easy to detect the human gait features. In a corridor like structure, the subject is approaching a camera. Such case is difficult observe lateral view gait.



Figure 1: Frontal view gait data

In a lateral view gait, at least two cycles or four steps are needed. For more robust estimation of the period of walking, about 8m is recommended. To capture this movement, the camera

distance required is about 9m. Practically, having such a wide space is difficult. On the other hand, frontal view gait video is easy to observe 8m (or more) gait steps [7].

Figure 1 is an example of frontal view gait data recorded by Figure 2 situation. Figure 1 illustrates difficulty of frontal view gait analysis. This figure indicates the subjects getting close in to the camera and the subjects scaling is changing. The frontal view gait analysis requires registration of scale-changing component. Figure 3 shows subject's width time-series behavior of frontal view gait data. This figure illustrates frontal view gait data contains many of time-series components.

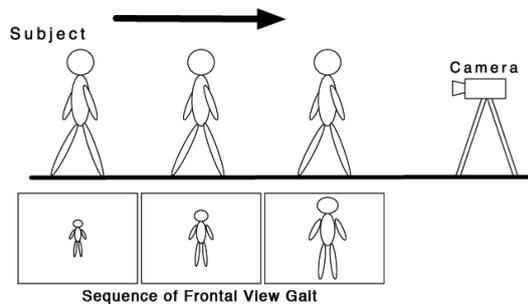


Figure 2: Filming situation of frontal view gait data

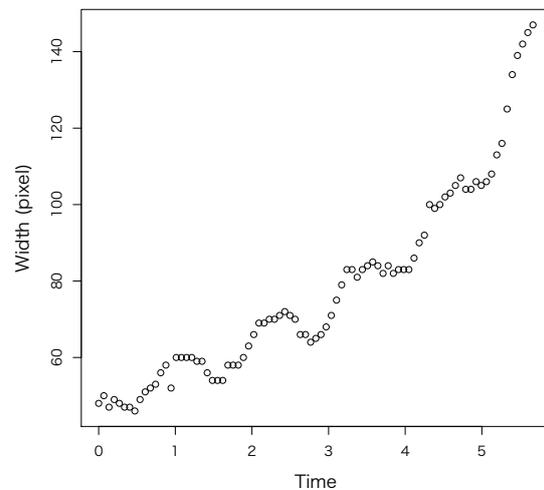


Figure 3: Time-series behavior of frontal view subject width

3 Modeling of frontal view gait data

Preprocessing

The raw video data is difficult to observe subject width and height time-series behavior, because data contains background. We separate subject from background using inter-frame subtraction method (Eq. 1).

$$\Delta^{(T)} = |I^{(T+1)} - I^{(T)}|, \quad T = 1, \dots, (n - 1),$$

$$\Delta^{(T)}(p, q) = \begin{cases} 1 & (\Delta^{(T)}(p, q) > 0) \\ 0 & (\text{Otherwise}). \end{cases} \quad (1)$$

Here, $\Delta^{(T)}$ is an inter-frame subtraction image, $I^{(T)}$ is grey scaled video data image at frame T , (p, q) is the pixel coordinate.

Generally, this method is difficult to apply to the field study data (e.g. security camera), because the all background pixels are not static. However, in the experimental environment case,

background pixels are tunable. We can assume that inter-frame subtraction method is reasonable. In the field study data, many of researchers are using “dynamic background subtraction” method (see Tamersoy [12]).

Subject Width/Height Calculation Inter-frame subtraction method can separate the subject and background. However, it is difficult to measure the time-series behavior of the subject width and height. In this section, we describe the subject width and height calculation method using inter-frame subtraction data.

Let us suppose that inter-frame subtraction image is binary matrix. We can measure the subject height and width by integration calculation of row and column at each frame. In this study, we focus on the human gait arm and leg swing of the frontal view gait. We assume that subject width and height time-series behavior consist of the arm and leg swing behavior.

Relationship between camera and subject

In this section, we describe archetype model proposed in Okusa & Kamakura [11]. Figure 4 shows a relationship between camera and subject. From figure 4, width and height model has same structure. In this section, we describe the subject’s width modeling. We can assume simple camera structure. We consider the virtual screen exists between observation point and subject, and we define x_i as subject width on the virtual screen at i -th frame ($i = 1, \dots, n$).

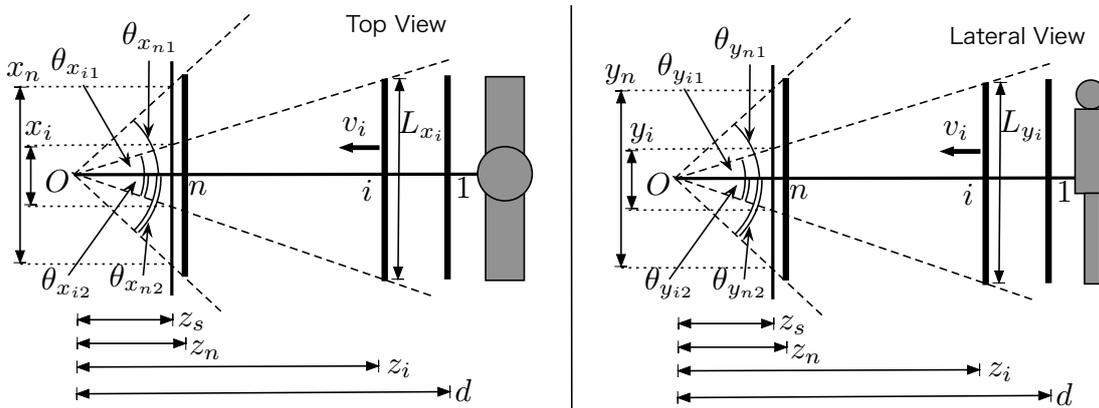


Figure 4: Relationship between camera and subject

Here we define z_i, z_j as distance between observation point and subject at i -th, j -th frame, z_s as distance between observation point and virtual screen, $\theta_{x_{i1}}, \theta_{x_{i2}}$ as subject angle of view from observation point at i -th frame, d as distance between observation point and 1st frame, v_i as subject speed at i -th frame. Okusa *et al.* [8] defined the subject length L was constant. We assume that L has the time-series behavior and we define L_i is the subject length at i -th frame.

x_i at i -th frame depends on $\theta_{x_{i1}}, \theta_{x_{i2}}$ as shown in Figure 4.

$$x_i = z_s(\tan \theta_{x_{i1}} + \tan \theta_{x_{i2}}). \tag{2}$$

Similarly, the subject length at i -th frame is

$$L_{x_i} = z_i(\tan \theta_{x_{i1}} + \tan \theta_{x_{i2}}). \tag{3}$$

From Eq.(2), Eq.(3), ratio between x_n and x_i is

$$\frac{x_n}{x_i} = \frac{L_{x_n} z_i}{L_{x_i} z_n} \quad (4)$$

Frame interval is equally-spaced (15 fps). Okusa *et al.* [8] assumed the average speed is constant. We can assume that average speed from i -th frame is $(n - i) = (z_i - z_n)/\bar{v}$, therefore z_i is $z_i = z_n + \bar{v}(n - i)$. We substitute z_i to Eq.(4)

$$x_i = \frac{M_{x_i} \gamma}{\gamma + (n - i)} x_n + \epsilon_i, \quad (5)$$

where γ is z_n/\bar{v} , M_{x_i} is L_{x_i}/L_{x_n} , ϵ_i is noise. From Eq.(5), predicted value $\hat{x}_i^{(n)}$ is registration from i -th frame's scale to n -th frame's scale

$$\hat{x}_i^{(n)} = \frac{\gamma + (n - i)}{M_{x_i} \gamma} x_i. \quad (6)$$

Next, we discuss the modified model for the shape scale registration.

Modified model for the shape scale registration

Let us consider the scale of shape, it seems that subject's width and height's has same relationships.

From Eq.(5), we can define subject height as

$$y_i = \frac{M_{y_i} \gamma}{\gamma + (n - i)} y_n + \epsilon_i, \quad (7)$$

where M_{y_i} is L_{y_i}/L_{y_n} .

We can assume that subject's width and height are same scale changing components γ , we can estimate common scale parameter γ from the following nonlinear least squares equation.

$$S(\gamma, M_x, M_y) = \sum_{i=1}^n \left\{ x_i - \frac{M_x \gamma}{\gamma + (n - 1)} x_n \right\}^2 + \sum_{i=1}^n \left\{ y_i - \frac{M_y \gamma}{\gamma + (n - 1)} y_n \right\}^2 \rightarrow \min. \quad (8)$$

We set the initial value γ as $\frac{1}{2} \left\{ \frac{1}{n} \sum_{i=1}^N \frac{x_i(n-i)}{x_i - M_x x_n} + \frac{1}{n} \sum_{i=1}^N \frac{y_i(n-i)}{y_i - M_y y_n} \right\}$ where mean value of solve Eq.(5), Eq.(7) for γ , and M_x, M_y as 1 (Okusa *et al.* [8]).

In next session, we validate the effectiveness of our model.

4 Experiments and Results

In this section, we validates our modified model using the frontal view gait authentication. To validate the effectiveness of our model, we observes frontal view walking video data (10 steps, Male, average height: 176.4cm, sd: 3.07cm) and apply to our proposed model.

Figure 5 is plot of the one of the 10 subjects width(pixel) time-series behavior. Here, continuous line represent fitted value of Eq.5. Similarly, Figure 6 is plot of subject height(pixel) time-series behavior. Here, continuous line represent fitted value of Eq.7. From Figure 5, Figure 6, proposed model is good fitting for frontal view gait data.

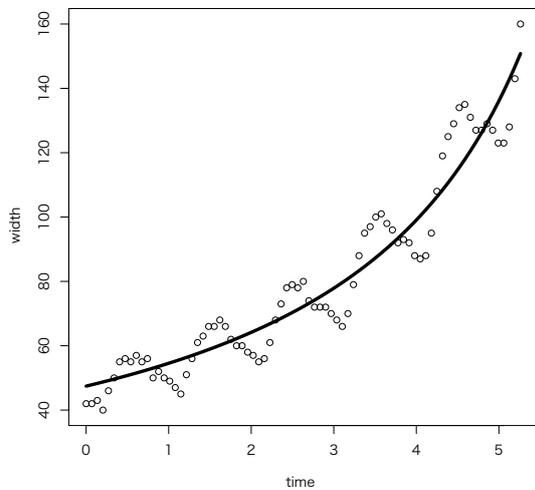


Figure 5: Fitted value of subject's width

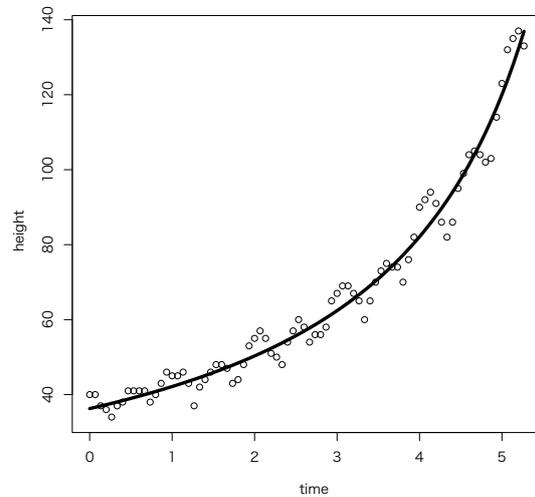


Figure 6: Fitted value of subject's height

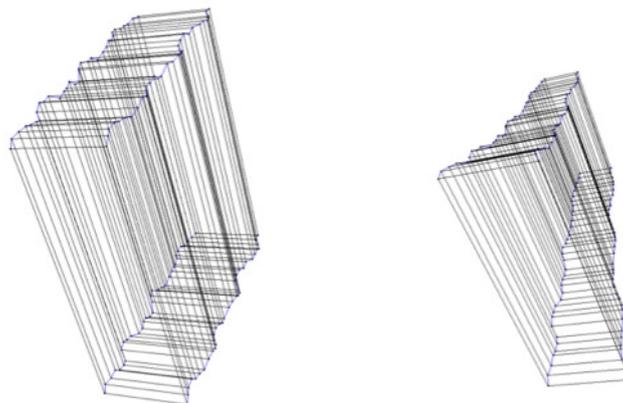


Figure 7: Scale-corrected result of circumscribed quadrangle of human gait silhouette (left side), non-corrected result (right side)

Figure 7 is plot of the scale-corrected result of circumscribed quadrangle of human gait silhouette based on the proposed method. Left side picture and right side picture are scale-corrected and non-corrected results, respectively. From Figure 7, proposed model is able to correct the scale changing components. Proposed method keeps arm and leg swing components after the registration.

Frontal View Gait Authentication

In this section, we validates our modified model using the frontal view gait authentication using CASIA gait dataset [13]. We compared our method with Barnich & Droogenbroeck [6] method

and “compendium method”.

Barnich & Droogenbroeck [6] is focus on the frontal view gait authentication, but they did not focus on the scale registration. This method is focus on the temporal evolution of the walking human silhouette. First, they normalize the human gait silhouette and calculate the integrated value of row and column at each frame. Secondly, they make a authentication classifier using random forests and apply to the gait authentication. We modifies Barnich & Droogenbroeck [6] method’s gait silhouette normalization using our proposed method. Our method is not required the normalization, it is able to correct the scale-changing components.

Additionally, we compared another method and our method. We named this method as “compendium method”. This method is simply scaling gait silhouettes to have the same fixed height, the silhouette will be re-scaled to have the fixed height = Y and the width = $x \times Y/y$. We modifies Barnich & Droogenbroeck [6] method’s gait silhouette normalization using this compendium method. It is seems to be low calculation cost.

To evaluate these methods, we apply these parameters to leave-one-out cross-validation test.

Table 1: Average authentication rate (%) and calculation time

	Authentication Rate (%)	Calc. Time (sec)
Proposed method	88.46	3.24
Compendium method	83.84	< 1
Barnich & Droogenbroeck [6]	81.5	< 1

Table 1 is average authentication rate (%) and calculation time of proposed method, compendium method and Barnich & Droogenbroeck [6] method. This result indicates our model shows good performance for the scale estimation and human gait authentication. Compendium method has higher authentication rate than Barnich & Droogenbroeck method, but proposed method shows more good performance for gait authentication. It is probable that low authentication rate of compendium method is caused by invalidation of the time-series behavior of subject’s height.

On the other hand, Barnich & Droogenbroeck [6] method and compendium method calculation cost is faster than proposed method. It is reasonable to think that it caused by parameter estimation step. We need to speed up the calculation cost in next step.

5 Conclusion

In this article, we focus on the shape scale changing in the frontal view human gait, we estimate scale parameters using the statistical registration and modeling on a video data. To demonstrate the effectiveness of our method, we apply our model for the frontal view human gait authentication. As a result, we also show that our method may be used for the frontal view human gait authentication.

In next phase, we need to speed up the calculation cost of our method. Additionally, we need to implement the gait authentication system based on the proposed method and demonstrate it.

Bibliography

- [1] J. R. Gage, “Gait analysis for decision-making in cerebral palsy.” *Bull. Hosp. Jt. Dis. Orthop. Inst.*, vol. 43, no. 2, pp. 147–163, 1982.
- [2] M. P. Kadaba, H. K. Ramakrishnan, and M. E. Wootten, “Measurement of lower extremity kinematics during level walking.” *J. Orthop. Res.*, vol. 8, no. 3, pp. 383–392, 1990.
- [3] S. Borel, P. Schneider, and C. J. Newman, “Video analysis software increases the interrater reliability of video gait assessments in children with cerebral palsy,” *Gait & posture*, vol. 33, no. 4, pp. 727–729, 2011.
- [4] S. Grunt, P. J. van Kampen, M. M. Krogt, M. A. Brehm, C. A. M. Doorenbosch, and J. G. Becher, “Reproducibility and validity of video screen measurements of gait in children with spastic cerebral palsy.” *Gait & posture*, vol. 31, no. 4, pp. 489–494, 2010.
- [5] R. A. Olshen, E. N. Biden, M. P. Wyatt, and D. H. Sutherland, “Gait analysis and the bootstrap,” *Ann. Statist.*, pp. 1419–1440, 1989.
- [6] O. Barnich and M. V. Droogenbroeck, “Frontal-view gait recognition by intra-and inter-frame rectangle size distribution,” *Pattern Recognit. Lett.*, vol. 30, pp. 893–901, 2009.
- [7] T. K. M. Lee, M. Belkhatir, and P. A. Lee, “Fronto-normal gait incorporating accurate practical looming compensation,” *Pattern Recognit.*, 2008.
- [8] K. Okusa, T. Kamakura, and H. Murakami, “A statistical registration of scales of moving objects with application to walking data. (in Japanese),” *Bull. Jpn. Soc. Comput. Statist.*, vol. 23, no. 2, pp. 94–111, 2011.
- [9] K. Okusa and T. Kamakura, “A statistical registration of scale changing and moving objects with application to the human gait analysis. (in Japanese).” *Bull. Jpn. Soc. Comput. Statist.*, vol. 24, no. 2, 2012.
- [10] K. Okusa and T. Kamakura, “Gait parameter and speed estimation from the frontal view gait video data based on the gait motion and spatial modeling,” *Int’l J. Appl. Math.*, vol. 43, no. 1, pp. 37–44, 2013.
- [11] K. Okusa and T. Kamakura, “Fast Gait Parameter Estimation for Frontal View Gait Video Data Based on the Model Selection and Parameter Optimization Approach,” *IAENG International Journal of Applied Mathematics*, vol. 43, no. 4, pp. 220–225, 2013.
- [12] B. Tamersoy, “Background subtraction – lecture notes,” 2009.
- [13] S. Zheng, J. Zhang, K. Huang, R. He, and T. Tan, “Robust view transformation model for gait recognition,” *International Conference on Image Processing(ICIP)*, pp. 2073–2076, 2011.

Application of Kalman Filter with alpha-stable distribution

Pavel Mozgunov, *National Research University Higher School of Economics*, pmozgunov@gmail.com¹⁶

Abstract. In this paper we consider the behavior of Kalman Filter state estimates in the case of distribution with heavy tails. The simulated linear state space models with Gaussian measurement noises were used. Gaussian noises in state equation are replaced by components with alpha-stable distribution with different parameters alpha and beta. We consider the case when "all parameters are known" and two methods of parameters estimation are compared: the maximum likelihood estimator (MLE) and the expectation-maximization algorithm (EM). It was shown that in cases of large deviation from Gaussian distribution the total error of states estimation rises dramatically. We conjecture that it can be explained by underestimation of the state equation noises covariance matrix that can be taken into account through the EM parameters estimation and ignored in the case of ML estimation.

Keywords. Kalman Filter, alpha-stable distribution, MLE, EM-algorithm

1 Introduction

State-space model (SSM) is convenient way for expressing dynamic systems that involve unobserved variables. If the model is linear and noises are Gaussian, the technique of Kalman Filter (KF) [1] can be applied. However, the condition of Gaussian noises in SSM is a strong enough and can significantly restrict the area of application of Kalman Filter. Some modification of Kalman Filter in cases of non-Gaussian noises was proposed in [7], but the case of non-Gaussian disturbances in measurement equation is considered, and the question of parameters estimation is still open. In this paper we used the KF technique in case of non-Gaussian state noises, and studied the behaviour of KF and methods of estimation and find some useful properties of EM-estimation that allows in process of parameters estimation (with out any additional computations) get acceptable results.

¹⁶This study (research grant No 14-05-0007) was supported by The National Research University Higher School of Economicsâ Academic Fund Program in 2014.

In Section 2 we consider linear SSM in general, and suppose a case of alpha-stable disturbances in state-equation. In Section 3, we overview two methods of SSM parameters estimation: MLE and EM. In section 4, we present the results of Kalman Filter estimation for different parameters of alpha-stable distribution and demonstrate useful properties of EM-algorithm that were found out in this research. We make conclusion in Section 5.

2 State-Space Model and Kalman Filter.

State-space model

Consider linear state-space model

$$X_{k+1} = A_k X_k + V_{k+1} \in R^m, \quad k = 0, 1, \dots \quad (1)$$

$$Y_k = C_k X_k + W_k \in R^d, \quad k = 0, 1, \dots \quad (2)$$

Where (1) is called *state equation* and (2) is *measurement equation*. A_k - matrix $m \times m$, V_k - state noise, C_k - matrix $d \times m$, W_k - measurement noise. To be the Kalman Filter optimal estimator in the least squares sense [1], it is necessary that noises and initial state vector should be Gaussian. It means that, $V_k \sim N(0, \delta_{kl} Q_k)$; $W_k \sim N(0, \delta_{kl} R_k)$ and $X_0 \sim N(\mu, \Sigma)$

Alpha -stable distribution.

We replace the Gaussian noises in equation (1) by disturbances with alpha-stable distribution and consider one-dimensional case. α - stable distribution is fully determined by its characteristic function [6]:

$$\log \phi(t) = -\sigma^\alpha |t|^\alpha \left\{ 1 - i\beta \operatorname{sign}(t) \tan \frac{\pi\alpha}{2} \right\} + i\mu t; \quad \alpha \neq 1 \quad (3)$$

$$\log \phi(t) = -\sigma |t| \left\{ 1 + i\beta \operatorname{sign}(t) \frac{2}{\pi} \log |t| \right\} + i\mu t; \quad \alpha = 1 \quad (4)$$

where $\alpha \in (0; 2]$, $\beta \in [-1, 1]$, $\sigma > 0$, $\mu \in R$, and α - characteristic exponent (refer to heavy tails of distribution); μ - location parameter; σ - scale parameter; β - skewness parameter. We denote the random variable with α - stable distribution in the following way: $X \sim S_\alpha(\sigma, \beta, \mu)$

Class of α -stable distributions was chosen because the case of $\alpha = 2$ and $\beta = 0$ ($S_2(\sigma, 0, \mu)$) corresponds to the Gaussian random variable $N(\mu, 2\sigma^2)$. It means that it can be studied, how deviations in α from 2, affects the sum of Kalman Filter prediction error.

Moreover, we put the distribution of initial state vector to be α -stable, with the same parameter α as in the state noise. Due to the property that the sum of two alpha-stable variables with the same parameter α is a alpha-stable variable again with parameter α , we get that the state vector has α stable distribution with the same α in all moments of time.

To simulate one-dimension alpha-stable distribution we used the same method as in Weron(1996)[5].

For $\alpha \neq 1$

$$X = S_{\alpha, \beta} \left(\frac{\sin \alpha (V + B_{\alpha, \beta})}{(\cos V)^{1/\alpha}} \right) \left(\frac{\cos(V - \alpha(V + B_{\alpha, \beta}))}{W} \right)^{(1-\alpha)/\alpha} \quad (5)$$

$$S_{\alpha,\beta} = [1 + \beta^2 \tan^2 \frac{\pi\alpha}{2}]^{1/(2\alpha)}$$

$$B_{\alpha,\beta} = \frac{\arctan(\beta \tan \frac{\pi\alpha}{2})}{\alpha}$$

And for $\alpha = 1$

$$X = \frac{2}{\pi} [(\pi/2 + \beta V) \tan V - \beta \log(\frac{\pi/2 W \cos V}{(\pi/2) + \beta V})] \tag{6}$$

Where V is uniformly distributed in interval $[-\pi/2; \pi/2]$ ($V \sim U[-\pi/2; \pi/2]$) and W exponentially distributed with parameter 1 ($W \sim Exp(1)$).

So the model that will be considered and simulated in this paper can be expressed in the following way:

$$x_{k+1} = Ax_k + \varepsilon_{k+1}; \quad \varepsilon_k \sim S_{\alpha}(\sigma, \beta, \mu); \quad x_0 \sim S_{\alpha}(\sigma_2, \beta, \mu)$$

$$y_k = Cx_k + \mu_k; \quad \mu_k \sim \mathcal{N}(0, R)$$

Kalman Filter

To get the Kalman Filter equations first, assuming that vector of parameters $\theta_k = [\mu, \Sigma, A_k, C_k, Q_k, R_k]$ is known for all k . The aim is to estimate state variable at time k , based on available information at time k and the error of this estimation. It can be formulated the following way:

$$\hat{X}_{k|k} = E[X_k | \mathcal{Y}_k];$$

$$\Sigma_{k|k} = E[(X_k - \hat{X}_{k|k})(X_k - \hat{X}_{k|k})^* | \mathcal{Y}_k], \quad k=0,1,2,\dots,N, \text{ where}$$

$$\mathcal{Y}_k = \sigma\{Y_0, \dots, Y_k\} \text{-sigma-algebra generated by } Y_0, \dots, Y_k, k=0,1,2,\dots,N.$$

Suppose that on the iteration k one has $\hat{X}_{k|k}$ and $\Sigma_{k|k}$, and it is necessary to find $\hat{X}_{k+1|k+1}$ and $\Sigma_{k+1|k+1}$. Before receiving a new observation Y_{k+1} one makes a prediction based on (1). Then when new observation received, the correction is started. G_k (Kalman Gain) is coefficient that shows the measure of uncertainty in new observation. All equations of Kalman Filter are received in assumption of Gaussian noises.

Prediction:

$$\hat{X}_{k+1|k} = A_k \hat{X}_{k|k}$$

$$\Sigma_{k+1|k} = A_k \Sigma_{k|k} A_k^* + Q_{k+1}$$

Innovations:

$$\nu_{k+1} = Y_{k+1} - C_{k+1} \hat{X}_{k+1|k}$$

$$H_{k+1|k} = C_{k+1} \Sigma_{k+1|k} C_{k+1}^* + R_{k+1}$$

Kalman Gain:

$$G_{k+1} = \Sigma_{k+1|k} C_{k+1}^* H_{k+1|k}^{-1}$$

Correction:

$$\hat{X}_{k+1|k+1} = \hat{X}_{k+1|k} + G_{k+1} \nu_{k+1}$$

$$\Sigma_{k+1|k+1} = (I - G_{k+1} C_{k+1}) \Sigma_{k+1|k}$$

Given $\hat{X}_{k|k}$ and $\Sigma_{k|k}$ for all k , results can be improved by and find smoothed estimates of states: $\hat{X}_{k|N}$ and $\Sigma_{k|N}$ for $k=0,1,2,\dots,N$ (all formulas can be found in [4])

3 Methods of parameters estimation

All equations above make sense only when all parameters are known. Assume that parameters are independent on time: $\theta = [\mu, \Sigma, A, C, Q, R]$, and the aim is to estimate this vector based only on realisation y_1, \dots, y_k .

To obtain the MLE [2], it is necessary to maximize the likelihood function of innovations by numerical technique (e.g. quasi-Newton-Raphson). The likelihood function of innovations:

$$L_Y(\theta) = \prod_{k=1}^N \frac{1}{(2\pi)^{n/2}} |H_{k|k-1}(\theta)|^{-1/2} \exp(-\frac{1}{2} \nu_k(\theta)^* H_{k|k-1}^{-1} \nu_k(\theta)) \tag{7}$$

The idea to use EM-algorithm was proposed in [3]. At E-step it is necessary to find the conditional expectation of the following function:

$$Q(\theta, \theta') = E_{\theta'}[\log \frac{dP_{\theta}}{dP_{\theta'}} | \mathcal{Y}_N], \log \frac{dP_{\theta}}{dP_{\theta'}} = \sum_{l=0}^N \log \bar{\lambda}_l + \text{constant}$$

where $\bar{\lambda}_k =$

$$\frac{|Q|^{-1/2} \phi(Q^{-1/2}(X_k - AX_{k-1}))}{\phi(X_k)} \cdot \frac{|R|^{-1/2} \psi(R^{-1/2}(Y_k - CX_k))}{\psi(Y_k)}$$

and $X_k \sim \mathcal{N}(0, I_n)$; $Y_k \sim \mathcal{N}(0, I_m)$; $\phi(x)$ and $\psi(y)$ is probability density function of standard Gaussian random variable.

At the M-step, (at $j + 1$ step), one maximises $Q(\theta, \theta')$: $\theta_{j+1} = \text{argmax}_{(\theta)} Q(\theta, \theta')$. The update recursive equations for all parameters can be found in [4] except matrix C, so we give it here.

$$C = \sum_{k=0}^N Y_k \hat{X}_{k|N}^* (\sum_{k=0}^N [\Sigma_{k|N} + \hat{X}_{k|N} \hat{X}_{k|N}^*])^{-1}$$

4 Simulation results and findings

The following parameters were chosen for simulation:

$$\begin{aligned} x_{k+1} &= x_k + \varepsilon_{k+1}; & \varepsilon_k &\sim S_{\alpha}(20, \beta, 0) \\ y_k &= 1.2x_k + \mu_k; & \mu_k &\sim \mathcal{N}(0, 150) \\ x_0 &\sim S_{\alpha}(50, \beta, 100) \end{aligned}$$

Each sample contents 1000 observations. For all parameters α and β under consideration the mean values of Z simulations are used. To apply the Kalman Filter, we used that α -stable distribution with $\alpha = 2$ and $\beta = 0$ ($S_2(\sigma, 0, \mu)$) corresponds to Gaussian random variable $N(\mu, 2\sigma^2)$, and we simply replaced each α to $\alpha = 2$. It means that we ignore true heavy tails.

All parameters are known. Consider how α and β influence total error of prediction of unobserved states. Firstly, we studied the case of all parameters are known.

Figure 1 demonstrates the mean error of estimation that was calculated as

$$\text{Error} = \frac{1}{Z} \sum_{m=0}^Z \sum_{k=0}^N (X^{(m)} - \hat{X}^{(m)}_{k|k})^2 \quad (8)$$

It can be seen from the Figure 1, that in cases of large deviation from $\alpha = 2$ (refer to Gaussian distribution) the total error is rising dramatically, and by decreasing α error continues to rise, e.g. for $\alpha = 1.1$ the total error was 4000 times higher then in Gaussian case (error displayed as a constant on the first picture because of great mistake for small α , so "zoomed" graphs are provided). Moreover, the Kalman Smoother gives more accurate state estimates only for α in near [1.85;2]. Note that in the interval [1.85;2] the KF and KS are not so sensitive to deviation in α (for $\alpha=1.85$, the total error increases by 20%) - then the slopes of both curves only rise.

The only parameter that was not set to true values is **distribution of state equation noise** and initial state vector. So it is logical to assume that such large total error of predictions is strongly connected with the non-observance of Gaussian assumption of state equation noise. To

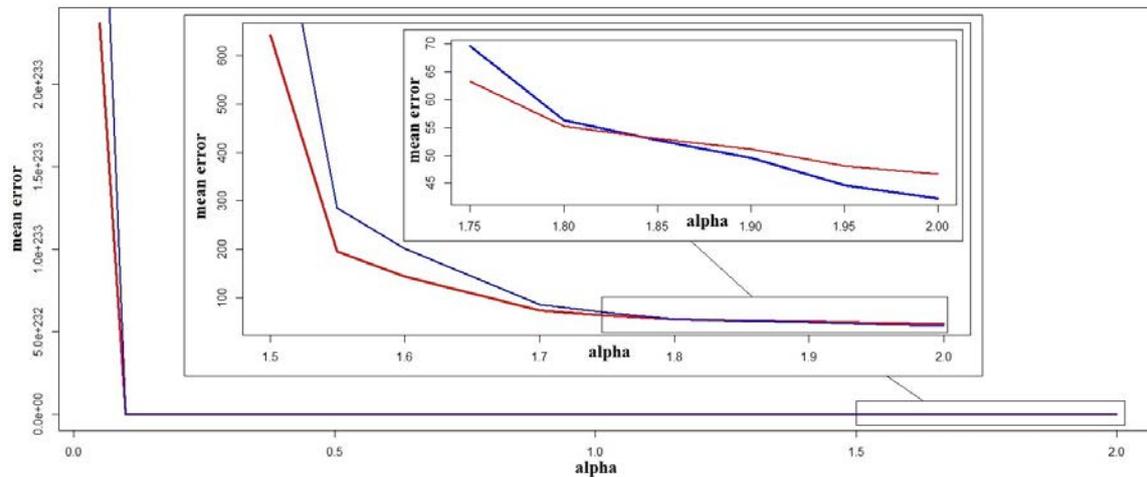


Figure 1: Mean error (10000 simulation) of KF (red) and KS (blue) estimation. $\alpha \in [0.05; 2]$, step=0.05, $\beta = 0$

be more precise, we conjecture that it can be explained by **underestimation of the state equation noises covariance matrix** (matrix $Q = 2\sigma^2$ in our model set up).

It is enough to compare α - stable distribution and Gaussian distribution with the same covariance matrix, to understand it¹⁷:

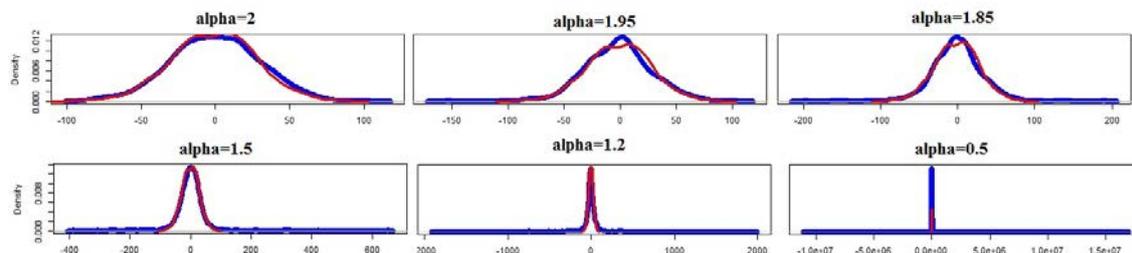


Figure 2: Kernel densities of α - stable distribution(blue)($\sigma = 20; \beta = 0$) and Gaussian(red) distribution with same covariance matrix.

The consequences of the assumption of Gaussian noise when it truly alpha-stable is shown on Figure 2. To obtain KF estimates we substitute true distribution (that corresponds to blue line) by Gaussian distribution with the same covariance matrix (that corresponds to red lines). However, it is seen that long tails of blue graphs are not covered by red ones, so as a result **algorithm can not cannot "detect" jumps** in a process that are appearing because of heavy tail distribution of disturbances. It is obvious that with α closer to zero, the tails become longer and the underestimation becomes larger and we met larger total error (when α is close to 2, the difference is not so dramatic). In terms of Kalman filter, algorithm consider state equation as not so noisy how it is truly is, and it assign small weight (Kalman Gain) to received observation.

The same mean total error as before for different β and same α is plotted in Figure 3.

¹⁷Gaussian variables were simulated by Box-Muller, and α -stable were as in [5]

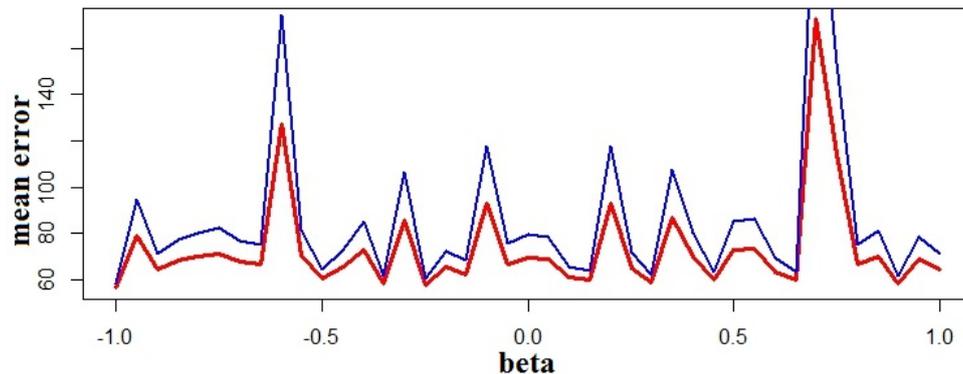


Figure 3: Mean error (10000 simulations) of KF (red) and KS (blue) estimation. $\beta \in [-1; 1]$ step=0.05 $\alpha = 1.75$

There is probably no dependence between total error and parameter β . The case of symmetric tails ($\beta = 0$) do not refer to the least total error. Moreover, the fluctuations of the total error are not so large. The deviation in heavy tails (α) is more significant for state estimates than in β , so we switch to different α only in estimation methods.

Parameters estimation. We consider the behaviour of two estimation procedures. Initial values were set to true ones, and the following results were found.¹⁸

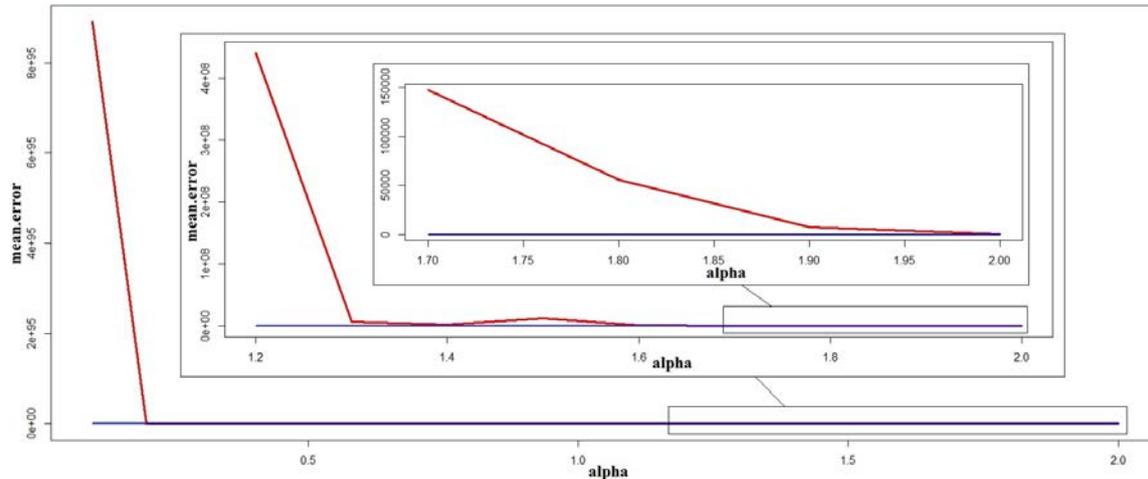


Figure 4: Mean error (1000 simulations). Parameters estimation by MLE (red) and by EM (blue). $\alpha \in [0.1; 2]$, step=0.1 $\beta = 0$

Figure 4 shows that the total error of estimation by MLE is increasing sharp, and as expected MLE is *extremely sensitive to deviation* from Gaussian distribution. Because of different ranges of scale, we plot the error of EM estimation separately (Figure 5).

¹⁸Because of numerical optimization of likelihood function in ML estimation, the number of simulations was reduced.

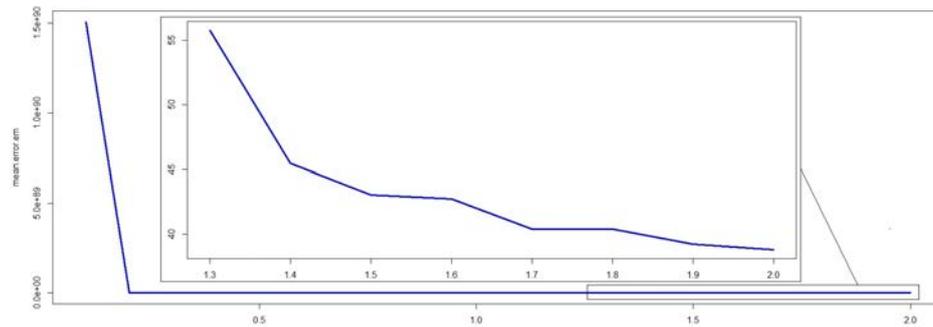


Figure 5: Mean error (1000 simulations). EM (blue) parameters estimation. $\alpha \in [0.1; 2]$, step=0.1 $\beta = 0$

Figure 5 shows that total error of KF prediction when parameters are estimated by EM increases slowly in the interval of [1.3;2], e.g. for $\alpha = 1.4$ error increases only by 14% (in average), against near 622% when "all parameters are known" and it is more satisfying result than before. But out of this interval the total error is large but less than in case of MLE estimation. To understand the nature of such good behaviour of EM we turn to figures of parameters estimation (Figure 6).

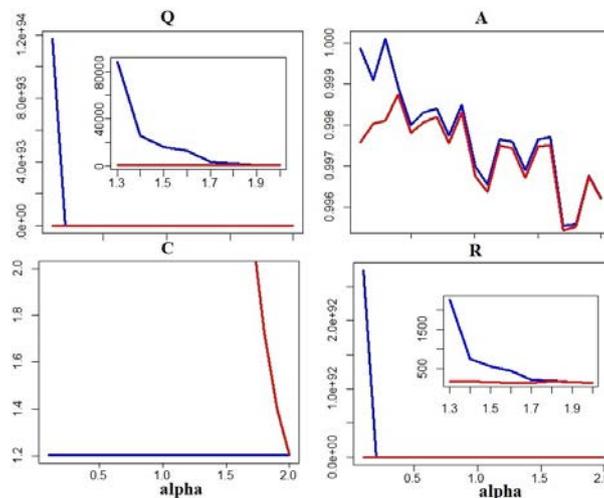


Figure 6: Mean (1000 simulation) MLE (red) and EM (blue) parameters estimation. $\alpha \in [0.1; 2]$, step=0.1 $\beta = 0$

It is obvious that results of *EM-estimation* are more accurate than MLE ones. The parameters of A and C are estimated correctly by EM-algorithm. But *EM-algorithm overestimate* the covariance matrix Q (compare to real value). **However this over estimation allows to take into account heavy tails.** The *overestimation of matrix Q* tends to larger Kalman Gain. It means that in a moment of correction we **assign larger weight to observation** we received than to our prediction (because of large covariance matrix). It is important to mention, that received results were confirmed when EM estimation was used only for Q, and all other parameters were true, so the cause of little increase in prediction error is overestimation of Q. Of course,

we increase the confidence interval of our state prediction, but it seems to be quite acceptable cost for considerable error decreasing. In our simulation example we decrease (e.g. for $\alpha = 1.3$ the total error 140 times while $\text{Sigma}_{k|k}$ increase only by 1.28). Unfortunately, overestimation of Q is not enough in cases of larger deviation in α . But in cases when it is necessary to estimate parameters and $\alpha \in [1.3; 2]$ it is acceptable not to complicate KF, and use only EM.

5 Conclusion

By simulation it was shown that in spite of the heavy tails of state equation noise, EM and ML estimate parameter A properly. EM-algorithm can **overestimate** covariance matrix of state noise in such way that the total error of prediction increases a little (compared to the Gaussian case) in the interval $[1.3; 2]$. It was confirmed that exactly due to the overestimation of covariance matrix Q , it is possible to prevent large prediction error. Without any additional computations, only in process of parameters estimation by EM and if $\alpha \in [1.3; 2]$, we can get approximately true values of unobserved states. ML estimation does not demonstrate such good properties, and more likely gives wrong estimates of states. It is a possible evidence of unacceptable application MLE for KF estimation, because it can lead to incorrect results. Although it is evidence of useful properties of EM that can allow to apply standard KF procedure in cases of α -stable distribution. The detecting how proposed method can reduce the prediction error comparing to modification of KF in [7], and how to estimate all parameters in this modification are questions for our further research.

Bibliography

- [1] Bucy, R.S., Kalman, R.E. (1961). *New Results in Linear Filtering and Prediction Theory*. Trans. ASME J. of Basic Engineering. 83, 95-108
- [2] Gupta, N.K., R.K.Mehra. (1974). *Computational Aspects of Maximum Likelihood Estimation and Reduction in Sensitivity Function Calculation*. IEEE Trans. Aut.Cont. AC-19 774-783
- [3] Shumway , R.H., Stoffer, D.S. (1982). *An approach to time series smoothing and forecasting using the EM algorithm*. Journal of Time Series Analysis 3(4): 253-264.
- [4] Shumway, R.H. and Stoffer, D.S. (2006). *Time Series Analysis and Its Applications (with R Examples)*. Springer.
- [5] Weron, R.(1996) . *On the Chambers-Mallows-Stuck method for simulating skewed stable random variables.*, Statist. Probab. Lett. 28, 165-171
- [6] Zolotarev, V. (1986) *One dimensional Stable Distributions*. American Mathematical Society, Providence, RI. Russian original, 1983
- [7] Xu Sun, Jinqiao Duan, Xiaofan Li, Xiangjun Wang (2013). *State estimation under non-Gaussian Lévy noise: A modified Kalman filtering method*. arXiv:1303.2395.v1

Acknowledgement. I would like to thank Prof. Valentin Konakov for the statement of problem and for stimulating discussions.

Combining sub(up)-approximations of different type to improve a solution

Bernard Fichet, *Aix-Marseille University*, bernard.fichet@lif.univ-mrs.fr

Abstract. This paper deals with the ultrametric approximation of a given dissimilarity according to the supremum norm. It has been established that the solution set is the finite union of some ultrametric intervals. In order to improve the homogeneity of such a solution interval, we want the bounds of it to share some constraints, such as to have same tree structure or common compatible order. We solve here those new problems. Besides, many results are presented in a general framework, where the concepts of subdominant (updominated) or submaximal (upminimal) approximations play a key role.

Keywords. subdominant, updominated, submaximal, upminimal, supremum norm, ultrametric, compatible order.

1 Introduction

Hierarchical classification is a very important part of clustering. The famous one-to-one correspondence between indexed hierarchies and ultrametric spaces, see [7], highlights the role played by ultrametricity. Indeed, given a dissimilarity d on some finite set I , producing a dendrogram as visual display of I , turns out to realise an approximation of d by an ultrametric \hat{d} .

L_1 -norm and L_2 -norm approximations lead to NP-hard problems, see [1] and [8]. A contrario, polynomial algorithms have been developed for the L_∞ -norm approximation, see [4]. In [2], the authors emphasise the link between the L_∞ -norm approximation and the subdominant approximation, so obtaining a simple algorithm and nice mathematical properties on these concepts. Later on, the previous link has been extended through upminimal approximations, leading to a characterisation of the L_∞ -norm solutions, as a finite number of ultrametric intervals.

All these concepts are recalled, developed and sometimes extended in the next two sections. This is done in a general framework where a particular structure is simply a reference set in a vector space, the set of ultrametries being nothing but a subset of the space of dissimilarities. In

particular, the characterisation of the solution set in terms of intervals is here extended, while there is no subdominant and no updominated approximation.

The last section is devoted to the improvement of an ultrametric interval solution, by adding some constraint to ultrametricity, while preserving the ultrametric optimal error. De facto, that turns out to intersect the ultrametric structure to another one, leading in a general case to new challenging problems about intersecting two structures. How does the intersection structure inherit from the two original structures? This is first treated in our general framework, where we propose an algorithm about subdominants, then secondly for ultrametricity under a constraint of tree structure or compatibility with a fixed order.

2 The framework

Here we use the general setup as developed in [2]. The basic set is a real vector space \mathcal{E} with finite dimension p . For a fixed basis, a vector u of \mathcal{E} has coordinates u_1, \dots, u_p . The space is endowed with the supremum norm, simply denoted by $\|\cdot\|$, so that $\|u\| = \max_j |u_j|$. We also note $\mathbf{1}$ the vector with unit coordinates, generating the main diagonal \mathcal{L} of \mathcal{E} , i.e. the set of vectors $c\mathbf{1}$, $c \in \mathbb{R}$.

With respect to the fixed basis of \mathcal{E} , a partial order between vectors is defined by a pairwise comparison of the coordinates: $u \preceq v$ if and only if $u_j \leq v_j$, for all $j = 1, \dots, p$. Then, every nonempty bounded subset \mathcal{A} of \mathcal{E} has a least upper bound and a greatest lowerbound with j -coordinates $\sup\{u_j : u \in \mathcal{A}\}$ and $\inf\{u_j : u \in \mathcal{A}\}$, respectively. In other words, (\mathcal{E}, \preceq) is a (conditionally) complete lattice. Again, for every nonempty subset \mathcal{A} of \mathcal{E} and every $u \in \mathcal{E}$, we note $\mathcal{A}_{\preceq}(u) := \{x \in \mathcal{A} : x \preceq u\}$, and $\mathcal{A}_{\succeq}(u) := \{x \in \mathcal{A} : x \succeq u\}$. Of course, such sets may be empty.

Now, we go further by considering a fixed nonempty subset \mathcal{K} in \mathcal{E} . It corresponds to a fixed classification structure under consideration and will be the approximating reference set. So, given u in \mathcal{E} , we define the following problem (P) :

(P) : $\inf\{\|u - x\| : x \in \mathcal{K}\}$. The optimal error which always exists will be denoted by \hat{e} .

Of course, our ability to treat the problem (P) , in particular to prove the existence of a solution and to compute it, will strongly depend on some geometrical, algebraical and topological properties of \mathcal{K} . Here are a few some of them. We say that \mathcal{K} obeys the *cylinder condition* if it is invariant under translation along the line \mathcal{L} , i.e. $x \in \mathcal{K}$ implies $(x + c\mathbf{1} \in \mathcal{K})$ for every $c \in \mathbb{R}$. \mathcal{K} obeys the *half positive or the half negative cylinder condition* if it is invariant under translation along the half line \mathcal{L}_+ or \mathcal{L}_- , i.e. $x \in \mathcal{K}$ implies $(x + c\mathbf{1} \in \mathcal{K})$ for every $c \in \mathbb{R}_+$, or $c \in \mathbb{R}_-$, respectively. Topological closeness of \mathcal{K} makes sure the existence of a solution. If \mathcal{K} is convex and z, z' are solutions of (P) , any vector of the segment $[z, z']$ is a solution, a property which is valid for any norm defining (P) . Again, due to the choice of the L_∞ -norm, any vector of the interval (z, z') of \mathcal{K} is a solution, whenever z, z' are solutions satisfying $z \preceq z'$.

Examples We give here relevant examples of data analysis and classification, where approximation problems can be treated within this scheme.

- Isotonic regressions. (X, \preceq) is a finite poset or a subset of the partially ordered set \mathbb{R}^p . \mathcal{E} is the space \mathbb{R}^X of numerical functions on X , endowed with its canonical basis and \mathcal{K} is the set of all *isotonic functions* on X : $f \in \mathcal{K}$ iff $x, y \in X$ and $x \preceq y$ implies $f(x) \leq f(y)$. The set \mathcal{K} obeys the cylinder condition and is a closed convex polyhedral cone.

- Dissimilarities. I is a finite set. \mathcal{E} is the set \mathcal{D} of pre-dissimilarities, *i.e.* the subset of functions d of $\mathbb{R}^{I \times I}$ such that: $\forall i, j \in I, d(i, j) = d(j, i), d(i, i) = 0$. According to the canonical basis of \mathcal{D} , a *dissimilarity* is an element of \mathcal{D}_+ . We still note $\mathbf{1}$, the unit dissimilarity.
 - semi-metric spaces. \mathcal{K} is the set \mathcal{D}_m of *semi-distances* of \mathcal{D} , *i.e.* dissimilarities obeying $\forall i, j, k \in I, d(i, j) \leq d(i, k) + d(j, k)$. The set \mathcal{D}_m is a closed convex polyhedral cone of \mathcal{D} . It obeys the half positive cylinder condition.
 - semi-ultrametric spaces. \mathcal{K} is the set \mathcal{D}_u of *semi-ultrametrics* of \mathcal{D} , *i.e.* dissimilarities obeying $\forall i, j, k \in I, d(i, j) \leq \max[d(i, k), d(j, k)]$. The set \mathcal{D}_u is a closed cone of \mathcal{D} . Extended to pre-ultrametrics with negative values, it obeys the cylinder condition.
 - θ -compatible dissimilarities. Given a fixed total order θ on I , those are dissimilarities which are θ -compatible, *i.e.* obeying: $(i\theta j\theta k)$ implies $d(i, k) \geq \max[d(i, j), d(j, k)]$, see [3], for example. The corresponding set \mathcal{D}_θ forms a closed convex polyhedral cone. Extended to pre-dissimilarities, it obeys the cylinder condition.

3 Lower and greater approximations

This section is devoted to L_∞ -norm approximations, as introduced in the general setup of Section 2, see the problem (P) . We mainly study the link between approximating an element u of \mathcal{E} by a vector of \mathcal{K} and approximating u by a vector of \mathcal{K} less (or greater) than u , if any. Thus, with the problem (P) , we associate two new problems:

$$(\mathbf{P}') : \inf\{\|u - x\| : x \in \mathcal{K}_{\preceq}(u)\}; \quad (\mathbf{P}'') : \inf\{\|u - x\| : x \in \mathcal{K}_{\succeq}(u)\}.$$

The problems are well-defined whenever $\mathcal{K}_{\preceq}(u)$ and $\mathcal{K}_{\succeq}(u)$ are nonempty. In that case, the optimal errors which always exist will be denoted by ϵ_* and ϵ^* , respectively.

Subdominant and updominated approximations We recall here some basic notions.

Definition 1.

Given the structure \mathcal{K} in the vector space \mathcal{E} ,

i) \mathcal{K} is said to admit a subdominant (updominated), if for every $u \in \mathcal{E}$ such that $\mathcal{K}_{\preceq}(u)$ ($\mathcal{K}_{\succeq}(u)$) is non empty, the latter set has a greatest (lowest) element, the \mathcal{K} -subdominant (\mathcal{K} -updominated) u_* (u^* of u).

ii) \mathcal{K} is said to be sup-closed (inf-closed) if for every subset \mathcal{M} of \mathcal{K} , bounded from above (below), $\sup\{x \in \mathcal{M}\} \in \mathcal{K}$ ($\inf\{x \in \mathcal{M}\} \in \mathcal{K}$).

Actually, the two definitions are equivalent. The proof is almost obvious. This is:

Proposition 1.

The two definitions coincide.

Clearly, the concepts of sup-closeness and inf-closeness are dual. Thus, any statement, property or any formula admits a dual formulation in this paragraph. We leave to the reader the transposition of any of them. For instance, if \mathcal{K} is sup-closed, u_* is the greatest solution of problem (P') .

The following proposition highlights two main hypothesis on \mathcal{K} , see proposition 1 of [2].

Proposition 2.

Let \mathcal{K} be sup-closed and obey the cylinder condition. Then:

- i) $\epsilon_* = \|u - u_*\| = 2\hat{\epsilon}$, and $u_* + \hat{\epsilon}\cdot\mathbf{1}$ is the greatest solution of problem (P) .
 ii) $\epsilon_* = \epsilon^*$, and $u_* + \epsilon_*\cdot\mathbf{1}$ is the greatest solution of problem (P'') .

Let us observe that if \mathcal{K} is both sup-closed and inf-closed, then the three problems (P) , (P') and (P'') have an interval as set of solutions, the extreme values of them deriving from the subdominant and the updominated approximations of u , after translation along with the main diagonal.

Thus, under the hypothesis of proposition 2 the computation of the greatest solutions of problems (P) , (P') or (P'') strongly depends on our ability to compute a subdominant. We here give some examples.

- isotonic regressions. This is a very simple example, where there are both a subdominant and an updominated approximation. Besides, there are analytic solutions for those approximations, hence an easy interval solution with the cylinder condition clearly fulfilled.

Let us note that θ -compatible dissimilarities can be treated by this way. Indeed, the conditions of θ -compatibility define a partial order between pairs of I , so that a θ -compatible dissimilarity turns out to be isotonic on the poset of pairs.

- semi-ultrametrics. The set \mathcal{D}_u of semi-ultrametrics on I is clearly sup-closed and obeys the cylinder condition, provided that we accept negative values. So, any dissimilarity d admits a subdominant d_* . Moreover, many efficient procedures have been developed to compute d_* . The single linkage algorithm builds an indexed hierarchy which is nothing but the one associated with d_* . Again, the subdominant is the ultrametric associated with a spanning tree derived from the complete graph defined by (I, d) , see [6]. A greedy method by shrinking the triangles works well too [7]. Thus we get the greatest L_∞ -approximation in a simple way [2], compared with the algorithmical way previously obtained by [4].
- semi-metrics. The set \mathcal{D}_m of semi-metrics on I also is sup-closed, but it only obeys the half positive cylinder condition. So, proposition 2 does not apply. Yes, there is still a metric subdominant, with some efficient procedures to compute it, but it infers only results along the positive direction of the main diagonal. We only deduce that $\hat{\epsilon}$ is less than $\epsilon_*/2$.

Submaximal and upminimal approximations. The previous paragraph highlights the key role played by the subdominant when it is associated with the cylinder condition. When the subdominant does not exist, it may be replaced by a weaker concept, intensively developed for ultrametricity, through upminimal approximations. Before going back to ultrametricity, we introduce this concept in our general framework, in order to establish conditions for general results. Of course, there is still a perfect duality, and we leave the reader free to transpose concepts and properties.

Definition 3.1. *Given u in \mathcal{E} and the reference set \mathcal{K} of \mathcal{E} , a submaximal or upminimal element of u , is any maximal element u_* or minimal element u^* of $\mathcal{K}_{\preceq}(u)$ or $\mathcal{K}_{\succeq}(u)$, if any, respectively.*

In the sequel, we will note $\mathcal{S}(u)$ and $\mathcal{U}(u)$ the sets of submaximal elements and upminimal elements of u respectively.

Proposition 3.

If \mathcal{K} is (topologically) closed and $\mathcal{K}_{\preceq}(u)$ is nonempty, then $\mathcal{S}(u)$ is nonempty. Moreover, each element of $\mathcal{K}_{\preceq}(u)$ is bounded by an element of $\mathcal{S}(u)$.

Proof. Let x be any element of $\mathcal{K}_{\preceq}(u)$. Since \mathcal{K} is closed, so is the non-empty set $\mathcal{A} := \mathcal{K}_{\preceq}(u) \cap \mathcal{K}_{\succeq}(x)$. Whence, \mathcal{A} is compact. Endowing here \mathcal{E} with the L_1 -norm, the problem: $\sup\{\|y - x\|_1 : y \in \mathcal{A}\}$ has a solution in \mathcal{A} , say u_* . We claim that u_* is maximal in $\mathcal{K}_{\preceq}(u)$. Let z be in the latter set and obey : $u_* \preceq z$. Then: $x \preceq u_* \preceq z$. Thus, $z \in \mathcal{A}$ and $\|u_* - x\|_1 \leq \|z - x\|_1$, so that $z = u_*$. □

Corollary 3.2. *Under the assumptions of proposition 3,*

$$\epsilon_* = \inf\{\|u - u_*\| : u_* \in \mathcal{S}(u)\}.$$

Proof. Immediate. □

Let us note that one may have infinitely many submaximal or upminimal elements. As an example, consider the reference set \mathcal{D}_m of semi-metrics on I . Let d be the dissimilarity on $I = \{i, j, k\}$ with $d(i, j) = d(j, k) = 1; d(i, k) = 3$. It exists a metric subdominant d_* . One has: $d_*(i, j) = d_*(j, k) = 1; d_*(i, k) = 2$. But it is quite easy to see that the upminimal elements of d are of the type: $d^*(i, j) = 1 + a, d^*(j, k) = 2 - a, d^*(i, k) = 3, 0 \leq a \leq 1$. Observe that this example remains valid for star-metrics or tree-metrics which are equivalent to metrics on three points.

However, we may sometimes guarantee finiteness of those optimal approximations. This is the case, somewhat usual, where there are bounding elements of $\mathcal{K}_{\preceq}(u)$ having coordinates in a fixed finite pool of numerical values, noted $V(u)$, the set of distinct coordinates of u for example.

Proposition 4.

Let u be in \mathcal{E} , and suppose $\mathcal{K}_{\preceq}(u) \neq \emptyset$. Assume that for every x in $\mathcal{K}_{\preceq}(u)$, there is y of $\mathcal{K}_{\preceq}(u)$ and greater than x , with coordinates in a fixed finite pool set $V(u)$. Then, $\mathcal{S}(u)$ is finite and nonempty, and each element of $\mathcal{K}_{\preceq}(u)$ is bounded by an element of $\mathcal{S}(u)$.

Proof. Clearly, sub-maximal elements of $\mathcal{K}_{\preceq}(u)$ are those of the finite subset of vectors of $\mathcal{K}_{\preceq}(u)$ with coordinates in $V(u)$. □

Observe that, unlike proposition 3, the latter proposition does not require closeness of \mathcal{K} .

The following corollaries connect sub-maximal or up-minimal elements with the L_∞ -norm approximations of u in \mathcal{K} . Their proofs are somewhat easy. They depict the set of solutions in terms of union of intervals, sensu the partial order on \mathcal{E} , as evoked in Section 2.

Corollary 1.

Assume that \mathcal{K} obeys the cylinder condition, $\mathcal{S}(u)$ is finite, each vector of $\mathcal{K}_{\preceq}(u)$ being covered by an element of $\mathcal{S}(u)$.

Then, the set of L_∞ -norm approximations has finitely many maximal elements, all of them being written as $\hat{u} = u_ + \hat{\epsilon} \cdot \mathbf{1}$, where u_* is in $\mathcal{S}_*(u) := \{u_* \in \mathcal{S}(u) : \|u - u_*\| = \epsilon_*\}$. Any L_∞ -norm solution is bounded by a maximal solution.*

Corollary 2.

Assume that \mathcal{K} obeys hypotheses of corollary 1 and in addition the dual condition: $\mathcal{U}(u)$ is finite, each vector of $\mathcal{K}_{\succeq}(u)$ covering an element of $\mathcal{U}(u)$.

Then, the set of L_∞ -norm approximations is the finite union of some intervals, each of them being of the type (\hat{u}', \hat{u}) , where \hat{u}', \hat{u} are the minimal and maximal elements of the set of L_∞ -norm solutions, respectively, satisfying $\hat{u}' \preceq \hat{u}$.

Corollary 3.

Assume that \mathcal{K} obeys hypotheses of corollary 1 and in addition: \mathcal{K} is inf-closed.

Then, the set of L_∞ -norm approximations is the finite union of the intervals of the type : $(u^* - \hat{\epsilon}.\mathbf{1}, u_* + \hat{\epsilon}.\mathbf{1})$, where u^* is the up-dominated of u , and u_* is in $\mathcal{S}_*(u)$.

4 Intersecting structures

In this section, we mainly focus on ultrametric approximations. Corollary 3 of Section 3 applies, so that the set of solutions in approximating a dissimilarity d , is the finite union of the intervals of the type : $(d^* - \hat{\epsilon}.\mathbf{1}, d_* + \hat{\epsilon}.\mathbf{1})$, where d^* are the up-minimal ultrametries at an optimal norm ϵ^* of d and d_* is the subdominant of d . Of course, we need here an algorithm to compute an optimal up-minimal ultrametric, and we may use the one mentioned in [5], which extends a procedure developed by [9] to compute any up-minimal ultrametric.

Now, we are going to introduce some new constraints, as an help in choosing an interval solution. The aim is to get an interval with bounds sharing some common features, such as a similar tree hierarchical structure or a common compatible order. Although ultrametricity remains here the key concept through optimal ultrametric approximations, adding new constraints to select a solution may somewhere appears as intersecting two structures. If there is no privileged one, a new challenge occurs: in terms of mathematical properties or algorithms, what is the impact of the intersection. Before going back to ultrametricity, we now give a few results in this way. Still, there are established in our general framework of Section 2.

Assume we are given two reference sets \mathcal{K}' and \mathcal{K}'' in \mathcal{E} . Define $\mathcal{K} := \mathcal{K}' \cap \mathcal{K}''$ and suppose $\mathcal{K} \neq \emptyset$. The set \mathcal{K} clearly inherits some simple properties, provided that \mathcal{K}' and \mathcal{K}'' fulfill these properties, such as the cylinder condition. But the most basic and obvious property is the following:

Proposition 5.

If \mathcal{K}' and \mathcal{K}'' admit a subdominant (updominated), then \mathcal{K} does.

If u in \mathcal{E} has subdominants u_* , u'_* and u''_* , leading to optimal errors ϵ_* , ϵ'_* and ϵ''_* , with respect to \mathcal{K} , \mathcal{K}' and \mathcal{K}'' , respectively, then $\epsilon_* \geq \max[\epsilon'_*, \epsilon''_*]$. Suppose now we have efficient algorithms to compute u'_* and u''_* . One may think of the following algorithm for u_* :

Algorithm 4.1.

Initialization: $v^0 := u$.

Step r Having v^r , compute $v^{r+1} := \min[v_*^r, v_*^{r'}]$. If $v^r = v^{r+1}$, then stop.

The following proposition justifies the algorithm.

Proposition 6.

The sequence $\{v^r\}$ is convergent. Its limit is u_* , the subdominant in \mathcal{K} , provided that one of the following two conditions is satisfied:

- i) $\{v^r\}$ converges in a finite number of steps.
- ii) \mathcal{K}' and \mathcal{K}'' are closed.

Proof. Since u_* is in both \mathcal{K}' and \mathcal{K}'' , u_* is inductively less than v^r . Besides, $\{v^r\}$ is clearly non-increasing, whence the convergence to a limit say w , with $w \succcurlyeq u_*$. Moreover, if $v^{r+1} = \min[v_*^r, v_*^{r'}] = v^r$, then $v_*^r = v_*^{r'} = v^r$, so that v^r is in \mathcal{K} . Condition i) is proven. When \mathcal{K}' and \mathcal{K}'' are closed, the sub-sequences $\{v_*^r\}$ and $\{v_*^{r'}\}$ which lie between v^r and v^{r+1} , converge to the

same limit w . Therefore, w is in both \mathcal{K}' and \mathcal{K}'' , hence in \mathcal{K} . Thus, $w \preceq u_*$ and condition ii) is proven. □

For the computational aspect, let us note that we could use an alternative and more tractable procedure, by alternatingly taking the subdominant in \mathcal{K}' , then in \mathcal{K}'' , and so on.

Corollary 4.

If the subdominants of u in \mathcal{K}' and \mathcal{K}'' have coordinates with values in a finite set $V(u)$, the sequence $\{v^n\}$ converges to u_ in a finite number of steps.*

Now, let us go back to ultrametricity, with the aim to select an ultrametric interval solution, by adding some new constraints, as written in the beginning of this section. We wish the bounds to satisfy a fixed constraint, such as to produce a fixed tree hierarchical structure τ or to have a fixed compatible order θ . Recall that a hierarchy on I is a class \mathcal{H} of nonempty subsets, the clusters, satisfying the following axioms: i) $I \in \mathcal{H}$, ii) $H, H' \in \mathcal{H}$ implies $H \cap H' \in \{H, H', \emptyset\}$ (nestedness axiom), iii) the minimal elements of \mathcal{H} cover I . By a tree hierarchical structure τ , we mean the poset (\mathcal{H}, \subseteq) or its covering graph which is a tree due to the nestedness condition. A so-called level index (in the broad sense) f on \mathcal{H} may be viewed as an isotonic non-negative function vanishing on the minimal clusters. In the 1-1 correspondence with an ultrametric d , for a pair (i, j) of I , one has $d(i, j) = f(H_{ij})$, where H_{ij} is the smallest cluster containing i and j . For a fixed tree structure τ , i.e. a fixed poset (\mathcal{H}, \subseteq) , we note \mathcal{D}_τ the set of ultrametries defined by all level indices (in the broad sense) f on \mathcal{H} . Adding also the set \mathcal{D}_θ of θ -compatible dissimilarities, we confront two intersecting structures: $\mathcal{D}_\tau \cap \mathcal{D}_u = \mathcal{D}_\tau$ and $\mathcal{D}_\theta \cap \mathcal{D}_u$. The sets \mathcal{D}_τ , \mathcal{D}_θ have nice properties: they obey the cylinder condition and they admit a subdominant and an updominated. However, to preserve the ultrametric optimality, that is to preserve the optimal error $\hat{\epsilon}$, we choose the tree structure and the compatible order to be associated with a bound of an optimal ultrametric interval solution. Then, whatever the intersecting structure is involved, this bound will be preserved, producing the same optimal error. It remains to compute the new second bound according to the intersecting set in consideration. The problem is quite easy for \mathcal{D}_τ . We treat now this problem, while preserving the upper bound of the optimal solution with respect to \mathcal{D}_u .

Proposition 7.

Let $(d^ - \hat{\epsilon}.\mathbf{1}, d_* + \hat{\epsilon}.\mathbf{1})$ be an optimal ultrametric interval solution with respect to \mathcal{D}_u . Let τ be the hierarchical structure associated with the subdominant d_* . Let d^+ be the updominated in \mathcal{D}_τ .*

Then, $(d^+ - \hat{\epsilon}.\mathbf{1}, d_ + \hat{\epsilon}.\mathbf{1})$ is an optimal interval solution in \mathcal{D}_τ .*

Proof. By definition, d_* is in \mathcal{D}_τ which is included in \mathcal{D}_u , so that it is both less and greater than the subdominant in \mathcal{D}_τ , hence equal to it. Whence the interval solution in \mathcal{D}_τ , as given in the proposition, with $\hat{\epsilon}$ as optimal error in \mathcal{D}_τ . □

Observe that d^+ dominates some optimal upminimal d^* , showing that the interval solution in \mathcal{D}_τ , is included in an interval solution in \mathcal{D}_u .

The problem is much more complicated with compatible orders, if we try to preserve the upper bound of the interval solutions derived from the subdominant d_* . Indeed, we will have then to compute some upminimal element in $\mathcal{D}_u \cap \mathcal{D}_\theta$, and we have no algorithm for this task,

less again to compute an optimal one! Fortunately, the opposite way is possible. It is based on an easy computation of the subdominant in $\mathcal{D}_u \cap \mathcal{D}_\theta$.

Lemma 4.1. *Let a dissimilarity be θ -compatible. Then its ultrametric subdominant is too.*

Proof. The chain defined the linear order θ , is clearly a minimum spanning tree of the graph associated with the dissimilarity: apply Prim's algorithm to see this. Then, the ultrametric defined by this chain fulfills θ -compatibility. □

Corollary 5.

Let d be in \mathcal{D} , d_- be its subdominant in \mathcal{D}_θ . Then, its subdominant in $\mathcal{D}_u \cap \mathcal{D}_\theta$ is $(d_-)_$.*

Proposition 8.

Let $(d^ - \hat{\epsilon}.\mathbf{1}, d_* + \hat{\epsilon}.\mathbf{1})$ be an optimal ultrametric interval solution with respect to \mathcal{D}_u . Let θ be a linear order compatible with the updominated ultrametric d^* . Let $(d_-)_*$ be the subdominant in $\mathcal{D}_u \cap \mathcal{D}_\theta$.*

Then, $(d^ - \hat{\epsilon}.\mathbf{1}, (d_-)_* + \hat{\epsilon}.\mathbf{1})$ is an optimal interval solution in $\mathcal{D}_u \cap \mathcal{D}_\theta$.*

The proof is similar to the one of the proposition 7. Observe that $(d_-)_*$ is less than d_* , so that the interval solution in $\mathcal{D}_u \cap \mathcal{D}_\theta$ is included in the original interval solution in \mathcal{D}_u .

Bibliography

- [1] Day, W.H.E. (1987) *Computational complexity of inferring phylogenies from dissimilarities matrices*. Bulletin of Mathematical Biology, **49**, 461–467.
- [2] Chepoi, V. and Fichet, B. (2000) *l_∞ -Approximation via Subdominants*. Journal of Mathematical Psychology, **44**, 600–616.
- [3] Critchley, F. and Fichet, B. (1994) *The partial order by inclusion of the principal classes of dissimilarity on a finite set, and some of their basic properties*. In: B. van Cutsem (ed.). Classification and dissimilarity analysis. Lecture Notes in Statistics. New York: Springer-Verlag, 5–65.
- [4] Farach, M., Kannan, S. and Warnow, T. (1995) *A robust model for finding optimal evolutionary trees*. Algorithmica, **13**, 155–179.
- [5] Fichet, B. (2012) *Intervals as ultrametric approximations according to the supremum norm*. In: M. Deza, M. Petitjean and K. Markov (eds.). Mathematics of Distances and Applications. ITHEA Sofia, 147.
- [6] Gower, J. and Ross, G. (1969) *Minimum spanning tree and single linkage cluster analysis*. Applied Statistics, **18**, 54–64.
- [7] Jardine, N., and Sibson, R. (1971) *Mathematical Taxonomy*. New York, Wiley.
- [8] Křivánek, M. and Morávek J. (1986) *NP-Hard problems in hierarchical-tree clustering*. Acta informatica **23**, 311–323.
- [9] Leclerc, B. (1986) *Caractérisation, construction et dénombrement des ultramétriques supérieures minimales*. Statistique et Analyse des données, **11**, 2, 26–50.

Log-linear multidimensional Rasch model for capture-recapture

Elvira Pelle, *University of Milano-Bicocca*, e.pelle@campus.unimib.it

David J. Hessen, *Utrecht University*, D.J.Hessen@uu.nl

Peter G.M. Van der Heijden, *Utrecht University, University of Southampton*,

P.G.M.vanderheijden@uu.nl

Abstract. The traditional capture-recapture method assumes homogeneity of the capture probabilities. However, differences of character or behaviour between individuals may occur and models that allow for varying susceptibility to capture over individuals and unequal catchability have been proposed and psychometric models, such as the Rasch model, were successfully applied. In the present work, we propose the use of the multidimensional Rasch model in the capture-recapture context. We assume that lists may be divided into two or more subgroups, such that they can be viewed as indicators of the latent variables which account for correlations among lists. We show how to express the probability of a generic capture profile in terms of log-linear multidimensional Rasch model and apply the methodology to a real data set.

Keywords. Rasch model, capture-recapture, heterogeneity, log-linear model, EM algorithm

1 Introduction

The capture-recapture method is a statistical method originally used to estimate the size of wildlife populations [1] based on a sequence of trapping experiments where individual trapping histories are used to estimate the population size.

The capture-recapture method has been successfully applied to other contexts, like human populations, where the common labels are *multiple-recapture*, *multiple-records systems* and *multiple-records systems method* [2]. In general, it can be applied to any situation in which two or more lists are available. Here, the estimation of the population size uses two or more incomplete but overlapping lists. Each list is regarded as a capture sample and data are usually arranged in an incomplete 2^s contingency table where the missing cell corresponds to absence in all s lists; then log-linear models are used to analyse the data [3].

The traditional capture-recapture method assumes that the S registrations are independent. If we allow possible dependencies between registrations, this results in interaction terms in log-linear models [4].

Another central assumption in the traditional capture-recapture approach is the homogeneity of the capture probability. However, differences of character or behaviour between individuals may cause indirect dependence between lists. Models that allow for varying susceptibility to capture through individuals and unequal catchability have been proposed either in the case of human populations [5] or in animal population studies [6] and psychometric models, such as the Rasch model, were successfully applied.

In applying the dichotomous Rasch model to the capture-recapture context, correct or incorrect answers to an item are replaced by "being observed" or "not being observed" in a list and, if all lists are supposed to be of the same kind, it is possible to treat heterogeneity in terms of constant apparent dependence between lists (Darroch [5], Agresti [6], International Working Group for Disease Monitoring and Forecasting [2]). Bartolucci and Forcina [7], shown how to relax the basic assumptions of the Rasch model (conditional independence and unidimensionality) by adding some suitable columns to the design matrix of the model. Bartolucci and Pennoni [8] proposed an extension of the latent class model for behaviour effects in which the latent class of a subject follows a Markov chain with transition probabilities depending on the previous capture history.

In the present work, we propose the use of the multidimensional Rasch model in the capture-recapture context. In particular, we assume that lists may be divided into two or more subgroups, such that they can be viewed as indicators of the latent variables which account for correlations among lists. To do so, the extension of the Dutch Identity for the multidimensional partial credit model (Hessen [9]) can be utilized. The Dutch Identity is a tool proposed by Holland [10], useful in the study of the structure of item response models, used by psychometricians to explain the characteristics and performance of a test. We use the results of Hessen [9], typically used in psychometric context, in the capture-recapture framework to express the probability of a generic capture profile in terms of log-linear multidimensional Rasch model.

We proceed as follows: in Section 2 we discuss the situation with three lists and two latent variables. In Section 3 we present the model that allows for the presence of a stratifying variable. In Section 4 we describe the method for a more general situation with S lists and J strata. In Section 5 we apply the methodology proposed to a real data set on children born with a neural tube defect (NTD's) in the Netherlands.

2 Three lists

Consider a situation in which three lists R1, R2 and R3 are available. Let $n_{i_1 i_2 i_3}$ denote the observed frequencies of the data, where $i_s = (0, 1)$, $i = 1, 2, 3$ and $i_s = 0$ denotes "not observed" and $i_s = 1$ denotes "observed". n_{000} is the number of individuals "not observed" in any list that has to be estimated in order to estimate the total unknown population size N . Data can be arranged in a 2^3 contingency table with a missing cell corresponding to absence in all the three lists (see Table 1).

Let I_s , $s = 1, 2, 3$ be the random variables denoting the presence or absence of an individual in the corresponding list. Assume that there are two latent variables which explain the correlation among lists, and suppose that I_1, I_2 and I_3 are conditionally independent given the two latent variables. Let $\Theta = (\Theta_1, \Theta_2)$ denotes the vector of latent variables and $\theta = (\theta_1, \theta_2)$ denotes a realization.

		R3			
		Observed		Not Observed	
		R2		R2	
		Observed	Not Observed	Observed	Not Observed
R1	Observed	n_{111}	n_{101}	n_{110}	n_{100}
	Not Observed	n_{011}	n_{001}	n_{010}	0^*

* Missing cell is treated as structurally zero cell

Table 1: Contingency table for three lists

We are interested in analysing a log-linear model that allows the presence of the two latent variables.

Let $\pi_{0_s}, s = 1, 2, 3$ be the probability of not being observed in the s -th list and let $\pi_{1_s} = 1 - \pi_{0_s}$ be the probability of being observed in the s -th list. The probability of inclusion in list s given the vector of latent variables may be expressed in a logistic form in the following way:

$$\pi_{1_s|\boldsymbol{\theta}} = \frac{e^{\mathbf{u}'_s \boldsymbol{\theta} - \delta_s}}{1 + e^{\mathbf{u}'_s \boldsymbol{\theta} - \delta_s}} \tag{1}$$

where δ_s is the parameter for the list s , θ_r is the parameter for the r -th latent variable and \mathbf{u}'_s is the row vector of the (3×2) full column rank matrix $\mathbf{U} = [u_{sr}]$ of weights for the latent variables, where

$$u_{sr} = \begin{cases} 1 & \text{if the list } R_s \text{ is indicator of the } r\text{-th latent variable} \\ 0 & \text{otherwise} \end{cases}$$

Let $\mathbf{t} = (t_1, t_2)$ be the vector of total scores, where the total scores are given by $t_1 = u_{11}i_1 + u_{21}i_2 + u_{31}i_3$ and $t_2 = u_{12}i_1 + u_{22}i_2 + u_{32}i_3$.

Let $\mathbf{i} = (i_1, i_2, i_3)$ denotes a generic capture profile for an individual. According to standard probability theory, the probability of a generic capture profile ($\pi_{i_1 i_2 i_3}$) may be written as

$$\pi_{i_1 i_2 i_3} = \int \dots \int \pi_{i_1 i_2 i_3|\boldsymbol{\theta}} f(\boldsymbol{\theta}) d\boldsymbol{\theta} \tag{2}$$

where $\pi_{i_1 i_2 i_3|\boldsymbol{\theta}}$ is the probability of a generic capture profile conditional to $\boldsymbol{\theta}$ and $f(\boldsymbol{\theta})$ is the multivariate density of $\boldsymbol{\theta}$. Under the assumption that the posterior distribution of the vector of latent variables conditional to the capture profile $i_1 i_2 i_3 = 000$ follows a multivariate normal distribution, we have

$$\begin{aligned} \pi_{i_1 i_2 i_3} &= \pi_{000} \exp \left\{ \sum_{s=1}^3 i_s \delta_s + t_1 \mu_1 + t_2 \mu_2 + \frac{1}{2} t_1^2 \gamma_{11} + \frac{1}{2} t_2^2 \gamma_{22} + t_1 t_2 \gamma_{12} \right\} \\ &= \pi_{000} \exp \left\{ \sum_{s=1}^3 i_s \delta_s + \mathbf{t}' \boldsymbol{\mu} + \frac{1}{2} \mathbf{t}' \boldsymbol{\Gamma} \mathbf{t} \right\} \end{aligned} \tag{3}$$

where $\mathbf{t} = (t_1, t_2) = \mathbf{i}' \mathbf{U}$, $\boldsymbol{\mu}$ is the mean vector and $\boldsymbol{\Gamma} = [\gamma_{ir}]$ is symmetric.

Let n the number of individuals observed in all lists. Since the probability of a generic capture profile $i_1 i_2 i_3$ has multinomial distribution, we can express the expected frequencies of

$n_{i_1 i_2 i_3}$ as

$$m_{i_1 i_2 i_3} = n\pi_{i_1 i_2 i_3} \tag{4}$$

Substituting (4) in (3) and taking the logarithm we obtain:

$$\ln m_{i_1 i_2 i_3} = \delta + \sum_{s=1}^3 i_s \delta_s + \mathbf{t}'\boldsymbol{\mu} + \frac{1}{2}\mathbf{t}'\boldsymbol{\Gamma}\mathbf{t} \tag{5}$$

where $\delta = \ln(n\pi_{000})$.

The model in equation (5) is not identified. Setting $\boldsymbol{\mu} = \mathbf{0}$ for identification, the log-linear multidimensional Rasch model can be rewritten as:

$$\begin{aligned} \ln m_{i_1 i_2 i_3} &= \delta + \sum_{s=1}^3 i_s \delta_s + \frac{1}{2}\mathbf{t}'\boldsymbol{\Gamma}\mathbf{t} \\ &= \delta + i_1 \delta_1 + i_2 \delta_2 + i_3 \delta_3 + \frac{1}{2}t_1^2 \gamma_{11} + \frac{1}{2}t_2^2 \gamma_{22} + t_1 t_2 \gamma_{12} \end{aligned} \tag{6}$$

Note that there are $2(2 + 1)/2 = 3$ parameters to account for the two latent variables θ_1 and θ_2 . In particular, γ_{11} and γ_{22} represent, respectively, the variance of the first latent variable and the variance of the second latent variable, given the total scores t_1 and t_2 , while γ_{12} represents the covariance between the two latent variables, given the total scores t_1 and t_2 . In general, to account for q latent variables, we need $q(q + 1)/2$ parameters.

3 Model with a stratifying variable

Suppose now that a stratifying variable is available. Let $n_{i_1 i_2 i_3 j}$ and $\pi_{i_1 i_2 i_3 j}$ denote the observed frequencies and the probabilities for strata j , respectively. In this case, n_{000j} indicates the frequency of individuals not observed in any lists in the j -th strata. With two strata the contingency table has two missing cells, as shown in Table 2.

		R3			
		Observed		Not Observed	
		R2		R2	
Year	R1	Observed	Not Observed	Observed	Not Observed
1	Observed	n_{1111}	n_{1011}	n_{1101}	n_{1001}
	Not Observed	n_{0111}	n_{0011}	n_{0101}	0*
2	Observed	n_{1112}	n_{1012}	n_{1102}	n_{1002}
	Not Observed	n_{0112}	n_{0012}	n_{0102}	0*

* Missing cell are treated as structurally zero cells

Table 2: Contingency table for three lists and two strata

We assume that lists are indicators of the latent variables which explain correlations among lists and the posterior distribution of the latent variables (given the capture profile of not observed) follows a multivariate normal distribution; similarly to the previous case, we have

$$\ln m_{i_1 i_2 i_3 j} = \delta_j + \sum_{s=1}^3 i_s \delta_{sj} + \frac{1}{2}\mathbf{t}'\boldsymbol{\Gamma}_j\mathbf{t} \tag{7}$$

where $\delta_j = \ln(n\pi_{000j})$, $\mathbf{\Gamma}_j$ is a symmetric matrix, \mathbf{t} is the vector of total scores and the mean vector for the j -th strata $\boldsymbol{\mu}_j$ is set equal to zero for identification.

If parameters are equal across the strata $\delta_{sj} = \delta_s, \forall j$, that is the assumption of measurement invariance holds, we can test whether $\boldsymbol{\mu}_j = \boldsymbol{\mu} = \mathbf{0}$ and $\mathbf{\Gamma}_j = \mathbf{\Gamma}$ for all j .

If the simultaneous hypothesis holds, then the model in (7) becomes

$$\ln m_{i_1 i_2 i_3 j} = \delta_j + \sum_{s=1}^3 i_s \delta_s + \frac{1}{2} \mathbf{t}' \mathbf{\Gamma} \mathbf{t}. \quad (8)$$

4 Generalization

The extension to a more general situation with S registrations and J strata is straightforward.

Let $n_{i_1 \dots i_s j}$ and $\pi_{i_1 \dots i_s j}$ be the observed frequencies and the probabilities, respectively, where the index $i_s, s = 1, 2, \dots, S$ denotes the cross-classification of S lists and $j = (1, 2, \dots, J)$ is the index denoting the strata. The resulting contingency table has J structural zeros (one for each stratum).

Suppose now that the covariances between the random variables I_1, \dots, I_S can be explained by q latent variables. Let \mathbf{u}'_s denotes the s -th row of the $SJ \times q$ full column rank matrix $\mathbf{U} = [u_{sr}]$, where $u_{sr} = 1$ if list RS is indicator of the r -th latent variable and 0 otherwise, and let $\mathbf{t} = (t_1, \dots, t_q)$ be the vector of the total scores of the latent variables, that is $t_r = \sum_{s=1}^S u_{sr} i_s$.

Under the assumption of a multivariate normal distribution of the posterior distribution of the latent variables (conditional to the capture profile of individuals not observed in any list) the log-linear multidimensional Rasch model takes the form:

$$\ln m_{i_1 \dots i_s j} = \delta_j + \sum_{s=1}^S i_s \delta_{sj} + \mathbf{t}' \boldsymbol{\mu}_j + \frac{1}{2} \mathbf{t}' \mathbf{\Gamma}_j \mathbf{t} \quad (9)$$

where $\boldsymbol{\mu}_j$ is the mean vector for the j -th strata, \mathbf{t} is the vector of total scores and $\mathbf{\Gamma}_j$ is a symmetric matrix.

Without any additional constraints the model is not identified. If we set $\boldsymbol{\mu}_j$ equal to $\mathbf{0}$ for identification we have:

$$\ln m_{i_1 \dots i_s j} = \delta_j + \sum_{s=1}^S i_s \delta_{sj} + \frac{1}{2} \mathbf{t}' \mathbf{\Gamma}_j \mathbf{t} \quad (10)$$

The model in (10) can be treated as a traditional log-linear model. Thus, to estimate the parameters of the model it is possible to follow the approach proposed by Sanathanan [11] which consists of maximizing the conditional likelihood given the distribution of the observed frequencies (for more details see Sanathanan [11] and Bishop et al. [4]). Once the parameters have been estimated they can be used to obtain the estimate of the portion of population missed by all lists and thus the total unknown population size N .

5 Application

We apply the methodology described in the preceding Sections to the data set of five lists described by Zwane et al. [12] on children born with a NTD's in the Netherlands. Data cover a period of 11 years (from 1988 through 1998) and Year (denoted by Y_{cat}) is treated as a

stratifying variable. Since the five lists cover different but overlapping periods of time, we use the EM algorithm proposed by Zwane et al. [12] to estimate the missing cells resulting from the fact that lists partially overlap. None of these lists record all cases of NTD's in the Netherlands, and the scope of this application is to estimate the total unknown number of children affected by NTD's.

We assume that the five lists R1, R2, R3, R4 and R5 may be divided into two subgroups such that they can be view as indicators of the latent variables which account for correlations among lists. In particular, we consider two multidimensional Rasch models obtained using the methodology described above and compare them with other log-linear models. In total, we take into account five models.

Table 3 summarizes the results of these models fitted to the data; for each model we report the number of parameters, the degrees of freedom, the deviance, the value of AIC, the value of BIC and the estimated total population size \hat{N} . Table 4 presents the yearly estimates $\hat{N}_j, j = 1988, \dots, 1998$ for each model.

Model	Design matrix	Par	df*	Dev	AIC	BIC	\hat{N}
1	R1+R2+R3+R4+R5+ Y_{cat}	16	213	400	432	487	2229
2	1+(R1R2+...+R4R5)	26	203	298	350	439	3077
3	1+H1	17	212	349	383	441	3009
4	1+ $\theta_1 + \theta_2$	19	210	324	362	427	2793
5	1+ $\theta_3 + \theta_4$	19	210	311	349	414	3041

Table 3: Selected models with deviance, AIC and BIC

Model	\hat{N}_{88}	\hat{N}_{89}	\hat{N}_{90}	\hat{N}_{91}	\hat{N}_{92}	\hat{N}_{93}	\hat{N}_{94}	\hat{N}_{95}	\hat{N}_{96}	\hat{N}_{97}	\hat{N}_{98}
1	199	224	234	206	222	186	189	202	178	210	179
2	275	309	323	285	302	258	261	280	246	290	248
3	272	305	319	281	303	249	252	271	238	280	239
4	251	282	295	260	280	232	235	252	222	261	223
5	271	305	318	281	300	255	258	277	244	287	245

* There are 229 observed cells

^SH1 is the first-order heterogeneity term

[†] $\theta_1 = R1 + R2$ and $\theta_2 = R3 + R4 + R5$

[‡] $\theta_3 = R1 + R2 + R4$ and $\theta_4 = R3 + R4 + R5$

Table 4: Selected models with yearly estimates

Note that the model with only the main-effect parameters does not fit the data well, as it has a high deviance. The model with the first-order heterogeneity parameter improves the fit, while adding all two-factor interaction parameters to Model 1 results in a smaller deviance.

Both of the multidimensional Rasch models fit the data well and Model 5, in which lists R1, R2 and R4 are assumed to be indicators of the first latent variable (θ_3) and lists R3, R4 and R5 are supposed to be indicators of the second latent variable (θ_4), is the best model, since it has the smallest value of AIC and BIC; thus, it is the selected model.

6 Conclusion

In this manuscript we proposed the use of the multidimensional Rasch model in the capture-recapture framework.

We assumed that lists may be divided into two or more subgroups (not necessarily disjoint) which constitute the latent variables accounting for correlations among lists. As consequence, the random variables denoting the presence or absence of an individual into a list are assumed to be conditionally independent, given the latent variable.

Under the assumption that the posterior distribution of the latent variables follows a multivariate normal distribution, we used the extension of the Dutch Identity proposed by Hessen [9] in a psychometric context to the capture-recapture framework and we showed that it is possible to re-express the probability of a generic capture-profile in terms of the log-linear multidimensional Rasch model.

Finally, we applied the methodology we proposed to a dataset on NTD's in the Netherlands from 1988 through 1998. Since lists did not cover the same time periods, we used the EM algorithm proposed by Zwane *et. al* [12] to estimate the missing entry in the data set. The results showed that the selected model for inference is one of the log-linear multidimensional Rasch models obtained by applying the methodology proposed. In fact, it was preferable among the other log-linear model, as it presented the smallest value of both AIC and BIC.

Bibliography

- [1] Seber, G. A. F. (1982) *The Estimation of Animal abundance and Related Parameters*. London: Griffin
- [2] International Working Group for Disease Monitoring and Forecasting, (IWGDMF). (1995) *Capture-Recapture and Multiple-Record Systems Estimation I: History and Theoretical Development*. American Journal of Epidemiology, **142**, 1059–1068.
- [3] Fienberg, S. E. (1972) *The multiple-recapture census for closed populations and the 2^k incomplete contingency tables*. Biometrika, **59**, 591–603.
- [4] Bishop, Y. M. M., Fienberg, S. E. and Holland, P. W. (1975) *Discrete Multivariate Analysis: Theory and Practice*. MIT Press: Cambridge.
- [5] Darroch, J. N., Fienberg, S. E., Glonek, G. F. V. and Junker, B. W. (1993) *A three-sample multiple-capture approach to census population estimation with heterogeneous catchability*. Journal of the American Statistical Association, **88**, 1137–1148.
- [6] Agresti, A. (1994) *Simple capture-recapture models permitting unequal catchability and variable sampling effort*. Biometrics, **50**, 494–500.
- [7] Bartolucci, F. and Forcina, A. (2001) *Analysis of Capture-Recapture data with a Rasch-Type Model allowing for Conditional Dependence and Multidimensionality*. Biometrics, **57**, 714–719.
- [8] Bartolucci, F. and Pennoni, F. (2007) *A Class of Latent Markov Models for Capture-Recapture data Allowing for Time, Heterogeneity, and Behavior Effects*. Biometrics, **63**, 568–578.

- [9] Hessen, D. J. (2012) *Fitting and testing conditional multinormal partial credit models*. Psychometrika, **77**, 693–709.
- [10] Holland, P. W. (1990) *The Dutch Identity: a new tool for the study of item response model*. Psychometrika, **55**, 5–18.
- [11] Sanathanan, L. (1972) *Models and estimation methods in visual scanning experiments*. Technometrics, **14**, 813–829.
- [12] Zwane, E. N., van der Pal, K. and van der Heijden, P. G. M. (2004) *The multiple-record systems estimator when registrations refer to different but overlapping populations*. Statistics in Medicine, **23**, 2267–2281.

Monitoring the process variability using STATIS

Adelaide Figueiredo, *Faculdade de Economia da Universidade do Porto and LIAAD-INESC Porto*, adelaide@fep.up.pt

Fernanda Figueiredo, *Faculdade de Economia da Universidade do Porto and Centro de Estatística e Aplicações, Universidade de Lisboa*, otilia@fep.up.pt

Abstract. In real situations the evaluation of the global quality of either a product or a service depends on more than one quality characteristic. In order to monitor the variability of multivariate processes and identify the variables responsible for changes in the process, we will use the STATIS (Structuration des Tableaux A Trois Indices de la Statistique) methodology, a three-way data analysis method. For this purpose we consider a control chart based on a similarity measure between two positive semi-definite matrices, the RV coefficient, and we evaluate the performance of this control chart for monitoring multivariate normal data.

Keywords. Control chart, Monte Carlo simulation, Process monitoring, RV coefficient, STATIS, Statistical Quality Control.

1 Introduction

The evaluation of the global quality of either a product or a service in real situations depends on several quality characteristics. Often the quality characteristics of interest are correlated and so multivariate techniques of process control are more appropriate than univariate methods for monitoring the individual characteristics. In a multivariate quality control procedure it is crucial to identify the out-of-control variables when the control chart gives an out-of-control signal. The control charts are the tools commonly used for process monitoring in Statistical Quality Control (SQC). The control charts were introduced by Shewhart at Bell Laboratories in 1924 and they initially emerged for monitoring industrial processes, but nowadays they are currently applied in several areas, including Health, Medicine, Genetics, Environment and Finance. The purpose of these graphical representations (control charts) is to help make decisions about the state of the process that is being monitored, whether it is in-control or out-of-control. Whenever a control chart triggers an out-of-control signal, which may possibly be a false alarm, it is necessary to investigate the causes responsible for the emission of such signal so that appropriate corrective actions are taken. Most of the multivariate control schemes are based on the Hotelling T^2 statistic

and are implemented under the assumption of multivariate normal data. These schemes consist of simultaneously monitoring the mean vector and the covariance matrix of the process or, separately monitor the mean vector and the covariance matrix. Several control charts have been proposed for monitoring the mean vector of a multivariate process such as a control chart based on the Hotelling statistic (1947) and refinements of this control chart, among others. Control charts for controlling the variability (covariance matrix) of a process have also been proposed based on the generalized variance and its refinements (Alt, 1985), and control charts based on the maximum of the sample variances or on the maximum of the ranges of the p characteristics under study (Costa and Machado, 2008a, 2008b), among other charts. Additionally several control schemes have appeared in the literature to monitor both the mean vector and covariance matrix of the process (Chen et al., 2005, Zhang and Chang, 2008, etc).

The STATIS (Structuration des Tableaux a Trois Indices de la Statistique) methodology was introduced by L'Hermier des Plantes (1976) and later developed by Lavit (1988) and Lavit et al. (1994). STATIS enables us to analyse simultaneously several data tables measured on the same individuals or variables for different circumstances or time instants. If the individuals are the same in all tables, we compare the structure of individuals in all tables and this analysis is called STATIS method. If the variables are the same in all tables, we compare the relations between variables in all tables and this analysis is called Dual STATIS method. If the individuals and variables are the same in all tables, we can apply both methods.

The STATIS methodology involves three steps. In the first step, termed *interstructure*, we globally compare the data tables. In STATIS we compare the structure of individuals through the scalar product matrices and, in Dual STATIS, we compare the relations between variables through the covariance matrices. In the second step, termed *intrastructure*, we describe the common structure in all data tables through the determination of the compromise and the respective euclidean image. In the last step, we identify which individuals or variables contribute the most to the observed differences among the data tables. We represent the individuals' or variables' trajectories on the compromise Euclidean image. These trajectories enable us to identify the individuals (STATIS) or variables (Dual STATIS) which most contribute for the differences among the data tables. In this step, we also do the decomposition of the squared distances between any two pairs of tables into contributions of individuals or variables to identify which individuals (variables) more contribute to the differences among the tables.

STATIS methodology has been applied in various areas, and in particular in statistical quality control to monitor batch processes by several authors. For instance, Scepi (2002) focus on the non parametric control schemes both for simple data as well as for batch and time dependent data, Gourvéneec et al. (2005) applied STATIS to batch process data to monitor the evolution in time of batches, Niang et al. (2009) proposed a non parametric quality control strategy based on STATIS and convex hull peeling for monitoring batch processes with constant or variable duration.

Our aim in this study is to apply the STATIS methodology for monitoring the covariance matrix of p quality characteristics, and beyond that, through this methodology, identify which variables are responsible for the emission of an out-of-control signal. In Section 2 we propose a control chart based on RV coefficient (Escoufier, 1973) between the compromise covariance matrix obtained from a set of reference samples and the covariance matrix of a new sample. In Section 3 we evaluate the performance of the RV -chart for monitoring a multivariate normal process. In Section 4 we present an illustrative example.

2 Control chart based on STATIS

We consider K reference samples of size n measured on p variables taken in K different time instants, when the process is in the in-control state, and we represent these matrices by their covariance matrices V_k 's.

The RV coefficient (Escoufier, 1973) between V_k and $V_{k'}$ is defined by

$$RV(V_k, V_{k'}) = \frac{Tr(V_k Q V_{k'} Q)}{\sqrt{Tr(V_k Q)^2 Tr(V_{k'} Q)^2}},$$

where Tr denotes the trace operator of a matrix and Q is the metric in the individuals space, defined by the identity matrix or by a diagonal matrix whose main elements are equal to the reciprocal of the variances of the variables. The RV coefficient varies between 0 and 1. The closer the RV coefficient is to 1, the more similar the two covariance matrices V_k and $V_{k'}$ are. We determine the compromise covariance matrix, V , which is defined as a weighted mean of the K covariance matrices V_k 's:

$$V = \sum_{k=1}^K \alpha_k V_k,$$

where α_k are the weights representing the agreement between the K tables and the compromise. These weights are the elements of the eigenvector associated with the largest eigenvalue of the following matrix Z containing the RV coefficients between the V_k 's:

$$Z = \begin{pmatrix} 1 & RV(V_1, V_2) & \cdots & RV(V_1, V_K) \\ RV(V_2, V_1) & 1 & \cdots & RV(V_2, V_K) \\ \vdots & \vdots & \ddots & \vdots \\ RV(V_K, V_1) & RV(V_K, V_2) & \cdots & 1 \end{pmatrix}$$

The control chart here proposed, denoted RV -chart, is implemented as follows. For a new time instant $k + 1$, we compare its covariance matrix V_{k+1} with the compromise covariance matrix V through the RV coefficient. Denoting CL the control limit of the chart, we consider the following decision criterion:

- If $RV(V, V_{k+1}) \geq CL$ we consider that the process is in-control.
- Otherwise, we decide that the process is out-of-control. In this case we identify which variables are responsible for this situation, through the decomposition of the distance between V and V_{k+1} into percentages of variables' contributions.

The exact distribution of the RV coefficient is unknown, and thus we fix CL at an empirical percentile of the sampling distribution of the RV coefficient.

When the process is declared to be in the out-of-control state, for identifying the variables which contribute the most for the differences between V and V_{k+1} , we decompose the squared Hilbert Schmidt distance between V and V_{k+1} , defined by $d_{HS}^2(V, V_{k+1}) = Tr[(V - V_{k+1})Q]^2$ into percentages of variables' contributions. The percentage of contribution of variable i for d_{HS}^2 is given by:

$$c_{var\ i, d_{HS}^2} = \frac{q_{ii} \sum_{j=1}^p q_{ij} (V^{ij} - V_{k+1}^{ij})^2}{d_{HS}^2(V, V_{k+1})},$$

where q_{ii} is the i th diagonal element of Q , V^{ij} is the ij -element of V and V_{k+1}^{ij} is the ij -element of V_{k+1} .

3 Performance of the control chart

For evaluating the efficiency of the proposed RV -chart, we computed by simulation the Average Run Length (ARL), the most commonly used measure of performance of control charts. We generated multivariate normal processes $N_p(\boldsymbol{\mu}, \Sigma)$, for $p=2,3$ assuming different structures for the covariance matrices when the process is in-control and out-of-control. In each case, we obtained the compromise covariance matrix based on 4 reference samples generated when the process is in-control. For $\alpha=0.005$, we determined the control limit of the chart, i.e., the percentile 0.5% of the distribution of the RV coefficient, obtained through a Monte Carlo simulation experiment of size 100000 and we calculated the in-control and out-of-control ARL values through 10000 replicates for the different structures of the covariance matrix.

More precisely, we generated samples from a bivariate normal distribution $N_2(\boldsymbol{\mu}, \Sigma)$, with mean vector $\boldsymbol{\mu} = (0, 0)'$ and covariance matrix $\Sigma = \begin{pmatrix} 1 & \sigma_{12} \\ \sigma_{12} & 1 \end{pmatrix}$. Note that we could consider another mean vector because we will work with centered data. The unit variances in Σ imply that the covariance is equal to the correlation coefficient. Some obtained results are presented in Tables 1 and 2. We also generated samples from a multivariate normal distribution $N_3(\boldsymbol{\mu}, \Sigma)$ with

mean vector $\boldsymbol{\mu} = (0, 0, 0)'$ and covariance matrix $\Sigma = \begin{pmatrix} 1 & \sigma_{12} & \sigma_{13} \\ \sigma_{12} & 1 & \sigma_{23} \\ \sigma_{13} & \sigma_{23} & 1 \end{pmatrix}$. As previously we

could use another mean vector and the unit variances imply covariances equal to the correlation coefficients. Some obtained results are indicated in Tables 3 and 4.

$\sigma_{12}=0$ in-control				$\sigma_{12}=0.75$ in-control			
n	5	10	15	n	5	10	15
CL	0.360	0.593	0.698	CL	0.390	0.747	0.863
σ_{12}	ARL			σ_{12}	ARL		
0	<u>198.55</u>	<u>188.31</u>	<u>193.35</u>	0.75	<u>178.42</u>	<u>204.85</u>	<u>180.29</u>
0.3	161.88	105.72	66.85	0.5	39.69	16.64	9.43
0.5	121.79	48.75	22.59	0.3	18.10	5.57	3.09
0.75	67.95	15.84	5.96	0.1	9.68	2.78	1.63
0.95	31.87	6.26	2.50	0	7.24	2.12	1.37
-0.3	168.65	105.71	66.88	-0.3	3.53	1.27	1.05
-0.5	122.25	48.17	22.40	-0.5	2.28	1.08	1.01
-0.95	31.24	6.26	2.46	-0.75	1.42	1.00	1.00

Table 1: Control limit and ARL for $n=5,10,15$ being $\sigma_{12}=0$ or $\sigma_{12}=0.75$ when the process is in-control. The in-control ARL values are underlined.

From the Tables 1-4, the control limit of the chart and the ARL depend on the sample size n and on the in-control correlation matrix. When the process is in-control the ARL is large and approximately equal to 200, and when the process is out-of-control, the ARL is smaller and decreases as the sample size increases.

$\sigma_{12}= 0.5$ in-control				$\sigma_{12}= -0.5$ in-control			
n	5	10	15	n	5	10	15
CL	0.338	0.581	0.721	CL	0.338	0.607	0.740
σ_{12}	ARL			σ_{12}	ARL		
0.5	<u>191.55</u>	<u>202.53</u>	<u>185.43</u>	-0.5	<u>191.55</u>	<u>206.25</u>	<u>188.17</u>
0.3	85.65	47.43	29.62	0	31.94	9.95	4.99
0	32.94	10.14	4.96	0.5	8.06	1.83	1.20
-0.3	14.03	3.25	1.76	0.75	3.96	1.16	1.01
-0.75	3.95	1.16	1.01	0.95	1.99	1.01	1.00

Table 2: Control limit and ARL for $n=5,10,15$ being $\sigma_{12}= 0.5$ or $\sigma_{12}= -0.5$ when the process is in-control. The in-control ARL values are underlined.

$\sigma_{ij} = 0, i \neq j$, in-control				$\sigma_{ij} = 0.5, i \neq j$, in-control			
n	5	10	15	n	5	10	15
CL	0.325	0.561	0.671	CL	0.300	0.607	0.740
$\sigma_{12}, \sigma_{13}, \sigma_{23}$	ARL			$\sigma_{12}, \sigma_{13}, \sigma_{23}$	ARL		
0,0,0	<u>202.34</u>	<u>194.99</u>	<u>201.12</u>	0.5,0.5,0.5	<u>189.35</u>	<u>197.09</u>	<u>190.17</u>
0.5,0.5,0.5	77.78	18.30	6.85	0.3,0.3,0.3	71.61	32.84	20.27
0.7,0.7,0.7	40.00	6.43	2.39	0.1,0.1,0.1	30.24	8.33	4.16
0.5,0.2,0.9	43.42	9.19	3.52	0,0,0	19.80	4.57	2.25
0.9,0.75,0.9	21.07	3.19	1.42	0,1,0.3,0.9	70.41	27.47	13.15

Table 3: Control limit and ARL for $n=5,10,15$ being $\sigma_{ij} = 0, i \neq j$ and $\sigma_{ij} = 0.5, i \neq j$ when the process is in-control. The in-control ARL values are underlined.

n	5	10	15
CL	0.340	0.561	0.671
$\sigma_{12}, \sigma_{13}, \sigma_{23}$	ARL		
0.75,0.75,0.75	<u>199.36</u>	<u>202.22</u>	<u>193.50</u>
0.5,0.5,0.5	36.85	11.92	6.59
0.3,0.3,0.3	14.58	3.47	1.95
0.1,0.1,0.1	6.75	2.49	1.14
0,0,0	4.52	1.75	1.05
0.9,0.5,0.1	24.53	7.62	2.74
0.1,0.9,0.3	18.94	4.06	2.11

Table 4: Control limit and ARL for $n=5,10,15$ being $\sigma_{ij} = 0.75, i \neq j$ when the process is in-control. The in-control ARL values are underlined.

If the correlations are very small when the process is in-control, the chart detects the existence of positive or negative correlations, being the large correlations (positive or negative) easily detected.

If the correlations are large and positive (negative) when the process is in-control, the chart detects decreases in the correlations, negative (positive) correlations and also absence of corre-

lation. In this case the chart does not detect increases in positive (negative) correlations. The chart is more sensible to detect large negative (positive) correlations.

Thus, the control chart enables us to detect changes in the correlations between variables and it detects such a change as fast as we move away from the in-control covariance matrix.

4 Illustrative example of a multivariate normal process

The following example is based on the example given in Hawkins and Maboudou-Tchao (2008). Suppose a process with 4 quality characteristics X_1, X_2, X_3, X_4 with normal multivariate distribution $N_4(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Assume that when the process is in-control, the mean vector and the covariance matrix are given by

$$\boldsymbol{\mu}_0 = \begin{pmatrix} 126.61 \\ 77.48 \\ 80.95 \\ 97.97 \end{pmatrix} \text{ and } \boldsymbol{\Sigma}_0 = \begin{pmatrix} 15.04 & & & \\ 8.66 & 5.83 & & \\ 10.51 & 5.56 & 15.17 & \\ 12.04 & 7.5 & 8.79 & 10.57 \end{pmatrix},$$

respectively. Consider that the compromise matrix is obtained from 10 reference samples of size $n=15$ generated when the process is in-control, and given by

$$V = \begin{pmatrix} 17.85 & & & \\ 9.94 & 6.56 & & \\ 12.10 & 6.14 & 16.16 & \\ 14.14 & 8.55 & 9.81 & 12.14 \end{pmatrix}.$$

To illustrate the performance of the RV -chart, we consider the following three out-of-control situations and shifts of magnitude $\theta=5\%, 10\%, 15\%, 20\%, 25\%, 50\%, 75\%, 100\%$.

1. Increase of θ in the variances, decrease of θ in the covariances and no change in the mean vector.
2. Decrease of θ in the covariances and no changes in the variances nor in the mean vector.
3. Increase of θ in the variances and no changes in the covariances nor in the mean vector.

The control limit of the chart for a false alarm rate $\alpha=0.005$, computed by simulation based on 100000 samples of size 15 generated when the process is in-control is 0.852. The corresponding in-control ARL, also obtained by simulation using 10000 samples generated when the process is in-control is equal 184.65 (a bit smaller than the nominal ARL value equal to 200).

The out-of-control ARL values obtained with 10000 samples generated in the three situations and shifts of magnitude θ is indicated in Table 5. From this table we observe that the out-of-control ARL decreases as fast as θ increases.

Next we exemplify how to determine the variables responsible for the out-of-control situation. Suppose the following two samples taken from the process, that have lead us to conclude for an out-of-control state. The covariance matrices associated to these samples are given by

$$V_1 = \begin{pmatrix} 17.80 & & & \\ 1.53 & 15.2 & & \\ 3.98 & 3.43 & 18 & \\ -0.29 & 9.5 & 1.56 & 14.3 \end{pmatrix} \text{ and } V_2 = \begin{pmatrix} 7.61 & & & \\ 0.39 & 3.32 & & \\ 2.55 & -0.31 & 19.63 & \\ -0.33 & 2.38 & -5.1 & 6.89 \end{pmatrix},$$

Shift θ	5%	10%	15%	20%	25%	50%	75%	100%
Situation 1	65.3	27.7	13.5	7.7	4.8	1.4	1.1	1.0
Situation 2	105.2	62.1	37.9	23.4	13.8	2.6	1.2	1.0
Situation 3	110.0	70.2	47.0	32.8	23.7	7.7	4.1	2.7

Table 5: ARL for the three out-of-control situations for each θ

respectively. Computing the coefficients $RV(V, V_1)$ and $RV(V, V_2)$, we obtained the values 0.745 and 0.506, respectively, and consequently, we conclude in both cases that the process is out-of-control. For identifying the variables that contributed the most for the out-of-control state, we decomposed $d^2(V, V_1)$ and $d^2(V, V_2)$ into percentage of variables' contributions (see Table 6). We usually retain the variables' contributions higher than the mean (in our case, $> 100\%/4 = 25\%$) and then we select the corresponding variables as the ones responsible for the out-of-control signal.

Variables	Contribution (%) $d^2(V, V_1)$	Contribution (%) $d^2(V, V_2)$
X_1	<u>32.9</u>	17.9
X_2	22.2	7.2
X_3	14.8	<u>44.8</u>
X_4	<u>30.1</u>	<u>30.1</u>

Table 6: Percentages of variables' contributions

From Table 6 we conclude that the variables X_1 and X_4 contributed the most for the out-of-control signal for the first given sample and X_3 and X_4 contributed the most for the out-of-control signal associated with the second sample.

Acknowledgement

This work is financed by the ERDF – European Regional Development Fund through the COMPETE Programme (Operational Programme for Competitiveness) and by National Funds through the FCT – Fundação para a Ciência e Tecnologia (Portuguese Foundation for Science and Technology) within the project FCOMP-01-0124-FEDER-037281 and the project PEst-OE/MAT/UI0006/2014 (CEA/UL).

Bibliography

- [1] Alt, F. B. (1985). *Multivariate quality control*. In *Encyclopedia of Statistical Sciences*, S. Kotz and N. L. Johnson, eds, Wiley, New York.
- [2] Chen, G., Cheng, S. W. and Xie, H. (2005). *A new multivariate control chart for monitoring both location and dispersion*. *Communications in Statistics: Simulation and Computation*, **34**, 203–217.

- [3] Costa, A. F. B. and Machado, M. A. G. (2008a). *A new chart based on sample variances for monitoring the covariance matrix of multivariate processes*. The International Journal of Advanced Manufacturing Technology, **41**, 770–779.
- [4] Costa, A. F. B. and Machado, M. A. G. (2008b). *A new multivariate control chart for monitoring the covariance matrix of bivariate processes*. Communications in Statistics: Simulation and Computation, **37**, 1453–1465.
- [5] Escoufier, Y. (1973) *Le traitement des variables vectorielles*. Biometrics, **29**, 751–760.
- [6] Gourvéneç, S., Stanimirova, I. and Saby, O. A. (2005) *Monitoring batch process with the STATIS approach*. Journal of Chemometrics, **19**, 288–300.
- [7] Hawkins, D. M. and Maboudou-Tchao, E. M. (2008) *Multivariate exponentially weighted moving covariance matrix*. Technometrics, **50**, 155–166.
- [8] Hotelling, H. (1947) *Multivariate quality control, illustrated by the air testing of sample bombsights*. Techniques of Statistical Analysis, McGraw Hill, New York, pp. 111-184.
- [9] Lavit, C. (1988) *Analyse Conjointe de Tableaux Quantitatifs*. Collection Méthodes+Programmes, Masson.
- [10] Lavit, C., Escoufier, Y., Sabatier, R. and Traissac, P. (1994) *The ACT (Statis method)*. Computational Statistics and Data Analysis, **18**, 97–119.
- [11] L'Hermier Des Plantes, H. (1976) *Structuration des Tableaux a Trois Indices de la Statistique*. Thèse de 3^{ème} cycle. Université de Montpellier II.
- [12] Niang, N., Fogliatto, F. S. and Saporta, G. (2009) *Batch Process Monitoring by three-way data analysis approach*. The XIII International Conference Applied Stochastic Models and Data Analysis (ASMDA 2009), June 30-July 3, Vilnius, Lithuania.
- [13] Scepi, G. (2002) *Parametric and non parametric multivariate quality control charts*. IN Lauro, C. et al. (Eds) Multivariate Total Quality Control, 163–189. Heidelberg: Physica-Verlag.
- [14] Zhang, G. and Chang, S. I. (2008) *Multivariate EWMA control charts using individual observations for process mean and variance monitoring and diagnosis*. International Journal of Production Research, **46**, pp. 6855-6881.

Estimation of Lévy CARMA models in the `yuima` package: application on the financial time series

Stefano M. Iacus, *University of Milan*, stefano.iacus@unimi.it
Lorenzo Mercuri, *University of Milan*, lorenzo.mercuri@unimi.it

Abstract. In this work we show how to use the R package `yuima` available on CRAN for the estimation of a Continuous Autoregressive Moving Average (CARMA) model on the real data. When dealing with the CARMA model, one of the advantages of the `yuima` package is the possibility of recovering the increments of the underlying noise and choosing the appropriate Lévy model. The estimation of the parameters for the underlying Lévy process makes `yuima` package appealing for modeling financial time series. Indeed, identifying the appropriate noise for a CARMA model allows to capture asymmetry and heavy tails observed in the real data.

Keywords. `yuima` package, CARMA model, financial time series

1 Introduction

The aim of the developers of the `yuima` package is to build an environment based on S4 classes and methods for the R language. The package allows the user to deal with a broad class of stochastic differential equations containing unidimensional or multidimensional diffusion processes, fractional Brownian motions, jump and jump diffusion processes. The main class is called `yuima-class` and it is composed by slots.

In particular, the slot `data` contains an object of class `yuima-data` where it is possible to store the empirical or simulated data. Slot `model` is an object of class `yuima.model` and gives a mathematical description of the chosen stochastic differential equation while the slot `sampling` gives information on how the data have been collected or generated. The slot `characteristic` gives additional information about the statistical model. The slot `functional` is used to specify functionals of the chosen model and the expected values can be approximated through asymptotic expansion formulas. This last slot is useful for option pricing purpose.

In this note we show how to apply the R package `yuima` for the identification and the estimation of a CARMA model driven by a general Lévy process. The Continuous Autoregressive

Moving Average (CARMA) model driven by a standard Brownian Motion was first introduced in literature by [7] as a continuous counterpart of the discrete-time ARMA process. [2] considers a CARMA model driven by a Lévy process with finite second order moments where the gaussianity assumption is relaxed. In this way the marginal distribution of a CARMA process is allowed to be asymmetric and heavy-tailed making it appealing for applications in financial econometrics.

We use the results described in [9] to fit a real market dataset in order to recover the underlying Lévy process by means of the `yuima` package once the estimation of the coefficients is done. The existing R packages available on CRAN deal only with CARMA(p,q) models driven by a standard Brownian Motion [12] or Gaussian CAR(p) models [16] while we are able to consider Lévy distributions of a general form.

2 CARMA model in the `yuima` package

In this Section we review the functions available in the `yuima` package for the estimation of a CARMA model driven by a Lévy process¹⁹. Let L_t be a Lévy process with $E(L_1^2) < +\infty$. The CARMA(p,q) model Y_t as in [2] is defined as a stationary solution of the following linear differential equation:

$$a(D)Y_t = b(D)DL_t \quad (1)$$

where D denotes the differentiation with respect to time and the polynomials $a(z)$ and $b(z)$ are defined as:

$$\begin{aligned} a(z) &= z^p + a_1z^{p-1} + \dots + a_p \\ b(z) &= b_0 + b_1z^1 + \dots + b_{p-1}z^{p-1} \end{aligned}$$

where a_1, \dots, a_p are the autoregressive parameters and b_0, \dots, b_{p-1} are the moving average parameters such that $b_q \neq 0$ and $b_j = 0, \forall j > q$.

We remark that the representation in (1) is not useful since higher order derivatives for the Lévy are not well defined so we use the state space representation is:

$$\begin{aligned} Y_t &= b'X_t \\ dX_t &= AX_tdt + edL_t \end{aligned} \quad (2)$$

where the state variable $X_t = [X_{0,t}, \dots, X_{p-1,t}]'$ is a vector process of dimension p and

$$e = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix}, \quad b = \begin{bmatrix} b_0 \\ \vdots \\ 0 \\ b_{p-1} \end{bmatrix}, \quad A = \begin{bmatrix} 0 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1 \\ -a_p & -a_{p-1} & \dots & -a_1 \end{bmatrix}$$

The `yuima` package contains several `S4` classes and methods for simulation, estimation and inference of stochastic differential equations that have the following general form:

$$dX_t = b(t, X_t)dt + \sigma(t, X_t)dW_t^H + c(t, X_t)dZ_t.$$

where $b(t, X_t)$, $\sigma(t, X_t)$ and $c(t, X_t)$ are coefficients defined by the user. W_t^H is a fractional Brownian motion (by default the Hurst index is fixed to $\frac{1}{2}$) and Z_t is a pure jump Lévy process.

¹⁹A detailed discussion of the methodologies and routines implemented in the `yuima` package for CARMA models is reported in [9]

In such a context, the estimation of a Lévy CARMA model can be done using two methods implemented in `yuima`: `setCarma` and `qmle`.

The `setCarma` function returns an object of class `yuima.carma`²⁰ that contains a mathematical description of the CARMA model based on the state space representation. For this function necessary arguments are the orders of autoregressive p and moving average q parameters. By default the underlying noise is a standard Brownian motion and other Lévy measures can be specified using the arguments `measure` and `measure.type` (See [3] for more details).

The `qmle` function requires as arguments a list with the initial values for the parameters and an object of class `yuima` where the slot `data` contains the observed time series, the slot `model` is filled with an object of class `yuima.carma` and the slot `sampling` indicates the frequency of observed data. The algorithm behind the function performs a three-step estimation procedure described below:

- The autoregressive a_1, \dots, a_p and the moving average b_0, \dots, b_q parameters are obtained using the quasi-maximum likelihood method where the unobservable state variable X_t is estimated using the Kalman Filter algorithm.
- Given the CARMA parameters, the increments of the underlying noise are obtained following the approach developed in [4]
- The parameters of the chosen Lévy measure are estimated using the increments obtained in the previous step. In this case, if the density has not an analytical form, the likelihood function is computed by Fourier Transform.

3 Application on a real dataset

In this Section we use the `yuima` package for the estimation of a CARMA model on the VIX data. The VIX, introduced by the Chicago Board Options Exchange [5] in 1993 and modified in 2003, measures the 30-day expected volatility of the S&P 500 index. As a first step, we need to load the `quantmod` package [11] for downloading financial data and `TSA` package [6] for a standard time series analysis. Our dataset is composed by the closing daily VIX values from October 24-th 2008 to April 30-th 2014. Table 1 reports the main statistics of the dataset.

VIX	DATA	VIX.Close
1	1st Qu.: 01-03-2010	1st Qu.: 15.79%
2	Median: 19-07-2011	Median: 19.08%
3	3rd Qu.: 06-12-2012	3rd Qu.: 25.95%

Table 1: Statistics for the VIX data from October 24-th 2008 to April 30-th 2014.

The VIX time series is obtained from the `yahoo-finance` page using the code listed below.

```
> getSymbols("^VIX",src="yahoo")
> daysInYear<-252
```

²⁰The class `yuima.carma` extends the `yuima.model` since contains additional informations related to the CARMA model

```

> Years<-6
> end.day<-dim(VIX)[1]
> start.day<-end.day-daysInYear*Years
> VIX.df<-data.frame(index(VIX),coredata(VIX),stringsAsFactors=FALSE)
> colnames(VIX.df)<-c("date","Open","High","Low","Close","Vol","Adj")

```

In our empirical analysis we decide to estimate a CARMA model on the logarithm of the squared VIX values and we use the following command lines in order to store the new time series.

```

> VIX.returns1<-(VIX.df$Close/100)^2
> dataForAnalysis<-ts((log(VIX.returns1[start.day:end.day])),frequency=daysInYear)

```

We conduct a preliminary qualitative analysis for identifying a possible pair of autoregressive and moving average parameters. This result is achieved using the Autocorrelation function (ACF) for detecting the linear dependence on time series and the Extended Autocorrelation function (EACF) developed in [14] for recognizing the (p,q) orders of an ARMA model.

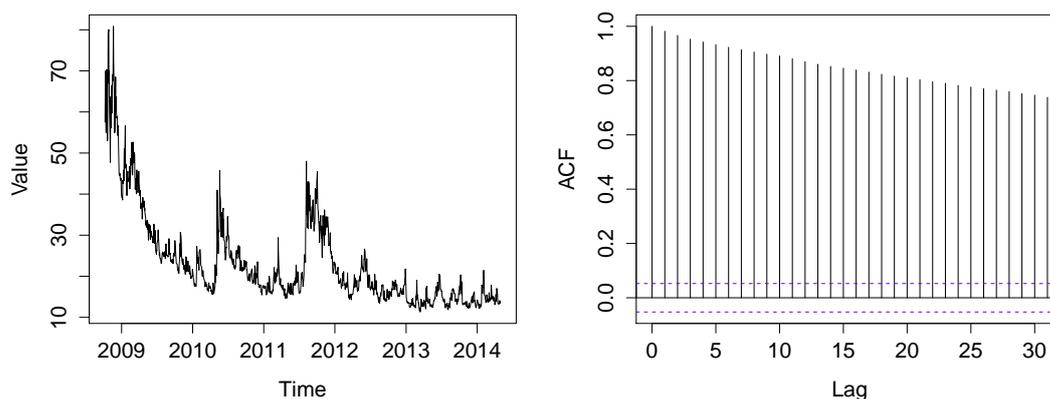


Figure 1: Time series from October 24-th 2008 to April 30-th 2014 (left side) and Autocorrelation function (right side) of the close VIX Index

AR - MA	0	1	2	3	4
0	X	X	X	X	X
1	X	X	X	O	O
2	X	O	O	O	O
3	X	X	O	O	O
4	X	X	X	O	O

Table 2: Shape of the Extended Autocorrelation Function obtained using VIX data. We have X for absolute values of the corresponding EACF greater than or equal to $\frac{2}{\sqrt{T}}$ and O for lower values. The bound $\frac{2}{\sqrt{T}}$ is twice the asymptotic standard error of the EACF with T being the number of observations.

The shape of the EACF in Table 2 suggests that the most appropriate model is the CARMA(2,1) since the upper vertex of the O's triangle coincides with the pair (p=2,q=1) (see [13] for a complete treatment).

In order to estimate the parameters of a CARMA(2,1) with location parameter²¹ μ , we need to build an object of class `yuima` that contains the observed time series stored in an object of class `yuima.data` and a mathematical description of the model in an object of class `yuima.carma`.

```
> mydata<-setData(dataForAnalysis)
> mymodel<-setCarma(p=2,q=1,loc.par="mu")
```

We remark that, by default the `setCarma` builds a model driven by a standard Brownian motion. Using the following code we fill the slots `data` and `model` in an object of class `yuima` that is used as an input in the estimation algorithm implemented in the `yuima` package.

```
> samp <- setSampling(Terminal=Years,n=Years*daysInYear)
> myCARMA<-setYuima(data=mydata,model=mymodel,sampling=samp)
```

We run the `qmlc` function that returns an object of class `yuima.carma.qmlc` containing the estimated parameters and the increments of the underlying noise (see [9] and [3] for further details).

```
> myParm0<-list(a1=36,a2=56,b0=21,b1=1,mu=0)
> myRes<-qmlc(yuima=myCARMA,start=myParm0)
```

The results we get with the `qmlc` function are reported in Table 3.

	a_1	a_2	b_0	b_1	μ
Est.	80.37	26.06	94.55	2.37	-2.75
Std.	21.35	30.79	33.98	0.06	1.09
$-2 \times \log L$			-1575.13		

Table 3: Estimates of the CARMA(2,1) model driven by a standard Brownian motion .

We check whether the residuals are independent and normally distributed. The independence assumption is studied by the means of the ACF and quantitative instruments such as Ljung-Box and Box-Pierce tests (we refer to [13] for more details). The shape of the ACF reported in Figure 2 seems to confirm the absence of correlation in the increments. Moreover, in Table 4 the p-values for both tests are greater than 0.05.

	Test	D.F.	p-value
Ljung-Box	14.89	10	13.59%
Box-Pierce	14.80	10	13.96%

Table 4: Results of the Ljung-Box and Box-Pierce tests on the estimated increments

²¹The `setCarma` is able to define a more general CARMA model with location and scale parameters, see [9] for more details.

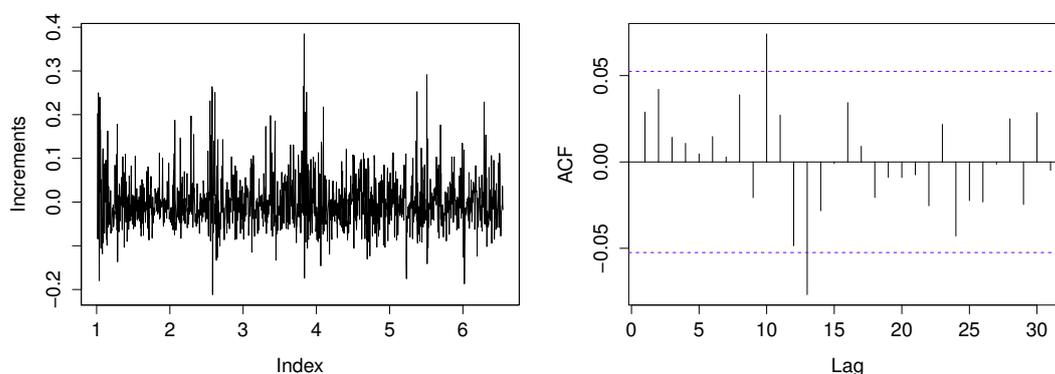


Figure 2: Estimated Increments (left side) and corresponding Autocorrelation (right side).

Looking at Figure 3 we observe a departure from the normality assumption for the distribution of the increments. In particular, the shape of the qq-plot indicates the presence of heavy tails. We use two different Lévy processes as underlying noise for the CARMA model:

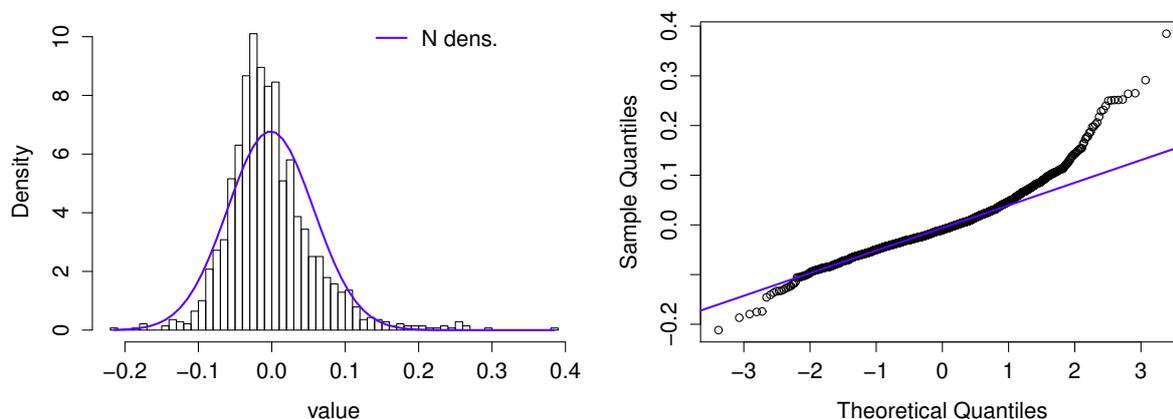


Figure 3: Empirical density (left side) and qq-plot (right side) of the estimated increments

the Variance Gamma [10] and the Normal Inverse Gaussian [1]. The advantage of the `yuima` package is the possibility of dealing with different processes for the underlying noise term. We specify the Lévy measure through the `setCarma` function:

```
> mymodel.VG<-setCarma(p=2, q=1,loc.par="mu",
  measure=list("rngamma(z,lambda,alpha,beta,mu0)"),
  measure.type="code")
> mymodel.NIG<-setCarma(p=2, q=1,loc.par="mu",
  measure=list(df=list("rNIG(z, alpha, beta, delta1, mu0)")),
  measure.type="code")
```

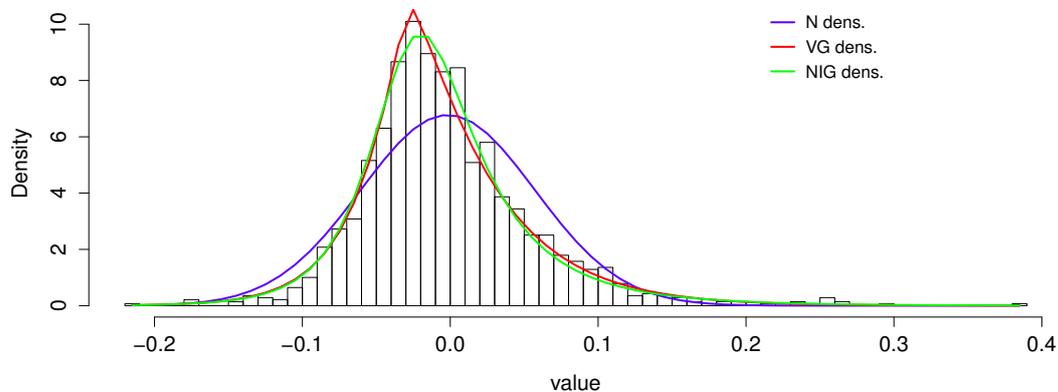


Figure 4: Empirical, Variance Gamma, Normal Inverse Gaussian and Normal densities fitted to the increments obtained applying the `qmle` function to the VIX time series.

We build two `yuima` objects that contain the VIX time series and the mathematical description of the considered Lévy CARMA models:

```
> myCARMA.VG<-setYuima(data=mydata,model=mymodel.VG, sampling=samp)
> myCARMA.NIG<-setYuima(data=mydata,model=mymodel.NIG sampling=samp)
```

The `qmle` function returns the autoregressive, the moving average and the underlying Lévy parameters²².

```
> myParm0.VG<-list(a1=36,a2=56,b0=21,b1=1,mu=0,lambda=1,alpha=1,beta=0,mu0=0)
> myParm0.NIG<-list(a1=36,a2=56,b0=21,b1=1,mu=0,alpha=2,beta=1,delta1=1,mu0=0)
> myRes.VG<-qmle(yuima=myCARMA.VG,start=myParm0.VG,aggregation=FALSE)
> myRes.NIG<-qmle(yuima=myCARMA.NIG,start=myParm0.NIG,aggregation=FALSE)
```

The estimated parameters are reported in Table 5. Figure 4 shows a comparison of the empirical and estimated densities on the Lévy increments. It is worth noting that, since we set `aggregation=FALSE`, the Lévy parameters are estimated using increments on intervals with time length $\Delta t = \frac{1}{252}$ and the `qmle` function computes the parameters defined on intervals of unitary length (annual basis in our case).

Distr.	α	β	δ_1	λ	μ_0	$-2 \times \log L$
NIG	17.67 (1.90)	6.87 (1.20)	12.43 (0.86)	... (...)	-6.71 (0.64)	-4215.21
VG	29.66 (1.94)	7.65 (1.01)	... (...)	307.63 (30.71)	-7.19 (0.50)	-4212.70

Table 5: Estimated Parameters of Lévy measure. The standard errors are reported in the brackets.

²²See [8] for a complete discussion on the meaning of the Variance Gamma and Normal Inverse Gaussian parameters

Bibliography

- [1] Barndorff, N. O. (1977). *Exponentially decreasing distributions for the logarithm of particle size*. Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences (The Royal Society), **353** 401–409.
- [2] Brockwell, P.J. (2001). *Lévy-Driven Carma Processes*. Annals of the Institute of Statistical Mathematics, **53** 113–124.
- [3] Brouste, A. Fukasawa, M. Hino, H. Iacus, S. M., Kamatani, K. Koike, Y. Masuda, H. Nomura, R. Ogihara, T. Shimuzu, Y. Uchida, M. and Yoshida N. (2014). *The YUIMA Project: A Computational Framework for Simulation and Inference of Stochastic Differential Equations*. Journal of Statistical Software, **57** 1–51. URL <http://www.jstatsoft.org/v57/i04/>
- [4] Brockwell, P. J. Davis, A. and Yang, Y. (2011) *Estimation for Non-Negative Lévy-Driven CARMA Processes*. Journal of Business, **29** 250–259.
- [5] Chicago Board Options Exchange (20030) *Vix-cboe volatility index*. URL <http://www.cboe.com/micro/vix/vixwhite.pdf>.
- [6] Chan, K. S. and Ripley B. (2012). *TSA: Time Series Analysis*. R package version 1.01. URL <http://CRAN.R-project.org/package=TSA>.
- [7] Doob, J.I. (1944) *The elementary gaussian process*. Ann. Math. Stat., **15** 229–282.
- [8] Iacus, M. S. (2011) *Option Pricing and Estimation of Financial Models with R*. Wiley Series.
- [9] Iacus, M. S. and Mercuri L. (2014) *Implementation of Lévy CARMA model in yuima package*. Working paper series.
- [10] Madan, D.B. and Seneta, E. (1990). *The variance gamma (V.G.) model for share market returns*. Journal of Business, **63** 511–524.
- [11] Ryan, J. A. (2013). *quantmod: Quantitative Financial Modelling Framework* R package version 0.4-0. URL <http://CRAN.R-project.org/package=quantmod>.
- [12] Tómasson, H. (2013) *Some computational aspects of gaussian CARMA modelling*. Statistics and Computing (in press), 1–13.
- [13] Tsay, R. S. (2005) *Analysis of financial time series*. John Wiley & Sons.
- [14] Tsay, R. and Tiao, G. (1984) *Consistent Estimates of Autoregressive Parameters and Extended Sample Autocorrelation Function for Stationary and Nonstationary ARMA Models*. Journal of the American Statistical Association, **79** 84–96.
- [15] Todorov, V. (2008) *Econometric Analysis of Jump-Driven Stochastic Volatility Models*. Journal of Econometrics, **160** 12–21.
- [16] Tunnicliffe, W. G. and Wang, Z. (2013). *cts: Continuous Time Autoregressive Models*. R package version 1.0-15, URL <http://CRAN.R-project.org/package=cts>

Modelling multivariate time series by structural equations modelling and segmentation approach

Christian Derquenne, *Electricité de France - R&D*, christian.derquenne@edf.fr

Abstract. We propose a method to build a complex model for irregular and multivariate time series. First, to take into account non-linearity, break, volatility, we chose to segment these series as linear trends to eliminate non-stationarity, standardizing the raw series with equation regression and standard error of each segment. Then, we used an exploratory approach using free structural equation modelling to establish links between these standardized variables. The latter allows building blocks (groups) of homogeneous time series, and then remove ones with latent variables, to seek significant links between them and finally apply the Partial Least Squares Approach on to the proposed free model. In the application, we compared the results of two models, the first on differentiated series, the second on the standardized series using segmentation. This has shown a greater consistency of results in terms of coefficients stability, significant relationships and business aspects for the model in standardized series using segmentation. Our future work will involve comparison with other methods to exhibit drawbacks and advantages of our method, including state-space models and forecasting time series contained in the target or with the free model approach or with the model set by the expert approach.

Keywords. Time series, cointegration, structural equations modelling, segmentation

1 Context and issues

In many applications in finance, environment, energy management, reliability, etc., the data can be observed as a univariate or multivariate time series. Usually, the first ones former (univariate) are stationarized by differentiation, then an ARMA model or a cointegration model on these pre-processed data can be applied. In the multivariate case, a state-space model or a multivariate cointegration model can be used. When there are predictors for modelling univariate or multivariate response, a structural model using link variables can be assessed by the expert. This structure is represented as a graph corresponding to the equations of the chosen model. These graphical models are also used for non-temporal data and are named: structural equations

modelling (SEM) (cf. fig. 1(a) & (b)). They consist of an outer model (or measurement model) that links the observable variables to unobservable latent variables and secondly, an inner model (or structural model) linking some latent variables between them.

Two statistical approaches are generally used to estimate these models. The LISREL method (LInear Structural RELationships) [7] which is based on the covariance matrix between endogenous and exogenous variables, depends on the structure imposed by the domain's expert. The maximum likelihood estimator is usually used. The second method is the PLS approach (Partial Least Squares) introduced in [9]. This approach is based on the observed variables starting from the structure of the individual, using the least squares estimator. Whatever the type of data: static, dynamic, univariate, multivariate, these theoretical models are proposed by an expert and the two previous methods aim to confirm or disprove their hypotheses. In contrast, when the model is not known a priori, it is necessary to discover the groups of observed variables (outer model), the relationships between latent variables and their orientations (inner model). In this case, the model is said to "free" [1] because it is built by a non-supervised approach. Furthermore, the relationships between the observed variables that can be static or dynamic are not always linear. Therefore, the previous approaches (state space model, cointegration, LISREL, PLS) are no longer valid in a reasonable manner.

Indeed, in the case of time series, the relationships between data may be non-linear, linear but with a non-constant variability, with breakpoints, etc. To overcome this problem, it is possible to use segmentation methods [2, 3, 4, 6, 8]. These ones allow to capture linear local trends of the time series. In the other words, the time-domain is partitioned into a few time-intervals on which the series exhibits a linear trend. Cutting can be very fruitful in the search for links between time series. The time series can then be stationarized by piecewise regression in normalizing using information provided by the segmentation [5]. Then, the relationships between these transformed variables can be analyzed and used to build of the state space models or structural equations models. Indeed, this process allows to obtain stationarized time series, but also it provides either a linear relationship or either a lack of relationship.

In this paper, we present briefly the segmentation method, then we formalize the structural equations model (fixed by expert and free model) based on segmentation, finally we apply this approach to a real example of application, but with anonymized data. We conclude on inputs, limits and perspectives of this work.

Times series modelling: SEM and segmentation

We have introduced a segmentation method of time series [2] completely unsupervised which allows to reduce the complexity compared with respect to the other methods, but above all proposes solutions segmentation of the series containing constant segments, increasing or decreasing slopes and with different levels of dispersion. Our method is original because it offers a decision support for time series, step by step. It contains two main phases: data preparation to obtain a first segmentation of data and modelling of segments based on a Gaussian heteroskedastic linear model by successive adaptations. Each of the two phases is repeated a few times depending on the degree of smoothing applied to the data. The degree of smoothing can vary from 2 to T theory. It corresponds to the number of observations included in moving median used in the phase of data preparation. The empirical complexity is $O(T\sqrt{T})$ and the theoretical complexity is $O(T^2)$. To improve this method, we introduced a preliminary phase, better consideration of variance components of the series by means of an appropriate transformation of the data [3] and

an approach of meta-segmentation [4] which consists of aggregation the better segmentations. This method has been tested on many series and has provided encouraging results on both simulated data to assess the quality of reconstruction of the series: detection of break points and modeling segments, but mainly on real data, especially in the area of price formation energy market [5].

Let (X_1, \dots, X_p) be, p time series of size T , where $X_j = (X_{j(t)})_{j=1,p;t=1,T}$ and $(\tilde{X}_1, \dots, \tilde{X}_p)$ their transformations in segments. Each \tilde{X}_j has m_j segments of respective sizes T_{jk} , with $\sum_{k=1}^{m_j} T_{jk} = T$. Then with aid of the segmentation, each estimation of $X_{j(t)}$ is as follows: $\tilde{X}_{j(t)} = \sum_{k=1}^{m_j} (\alpha_{jk}^{(0)} + \alpha_{jk}^{(1)} t) \cdot 1_{[t \in \tau_{jk}]}$ where τ_{jk} corresponds to a segment number k of the segmentation of X_j . Lastly, we denote to (X_1^*, \dots, X_p^*) , the stationarized time series with aid of associated segmentations, such as $X_{j(t)}^* = ((X_{j(t)} - \tilde{X}_{j(t)})/s_{jk}) \cdot 1_{[t \in \tau_{jk}]}$, where $s_{jk} = \sqrt{\sum_{t \in \tau_{jk}} (X_{j(t)} - \tilde{X}_{j(t)})^2 / T_{jk}}$.

Structural equations modelling for times series

Let $(Y_1, \dots, Y_k, \dots, Y_q)$ be, q time series responses and let $(X_1, \dots, X_j, \dots, X_p)$ be, p time series predictors. ΔX_i denotes, the first differentiated times series of X_i . For instance, a multivariate cointegration model is: $\Delta Y_{1,t} = \alpha_{11} \Delta X_{1,t-1} + \alpha_{12} \Delta X_{1,t} + \epsilon_{1,t}$ et $\Delta Y_{2,t} = \beta_1 \Delta Y_{1,t} + \alpha_{21} \Delta X_{2,t-1} + \alpha_{22} \Delta X_{2,t} + \epsilon_{2,t}$, where α 's are linking coefficients between predictors and responses, β_1 links the both responses Y_1 and Y_2 , and the ϵ 's are independent white noises. Its graphical model is given in figure 1(a). Another approach is to use structural equations modelling. The measurement model links the observable variables (X, Y) to the unobservable variables (latent variables) (Z, U) ; the inner model links them some latent variables (see fig. 1(b)). In the following example, the outer model is as follows: $\Delta X_{1,t} = \lambda_{11} Z_1 + \delta_{11,t}$; $\Delta X_{1,t-1} = \lambda_{12} Z_1 + \delta_{12,t-1}$; $\Delta X_{2,t} = \lambda_{21} Z_2 + \delta_{21,t}$; $\Delta X_{2,t-1} = \lambda_{22} Z_2 + \delta_{22,t-1}$; $\Delta Y_{1,t} = \lambda_{31} U_1 + \epsilon_{11,t}$; $\Delta Y_{2,t} = \lambda_{32} U_1 + \epsilon_{12,t}$, while the structural model is: $U_1 = \gamma_1 Z_1 + \gamma_2 Z_2 + \zeta_1$. In this case, each time series (manifest variable) is a reflection of its latent variable Z_1 , Z_2 or U_1 . The unidimensionality of each block of manifest variables assumption is therefore required. In other words, these are dependent variables and complementary to each other. Loadings λ 's are used to model the manifest variables with their associated latent variable, the ϵ 's and δ s correspond to the measurement errors, the coefficients γ 's link the endogenous latent variable U_1 to its exogenous latent variables Z_1 and Z_2 . Finally, ζ_1 is the prediction error of U_1 . The adjustment is made by LISREL method or by PLS approach.

One of the major problems in modeling time series, with predictors is that they are strongly linked together to explain one or more event(s). It appears then multicollinearity between explanatory variables and this can make artificially high variance coefficients, causing a decrease in statistical t -test, and therefore the statistical are not significant then the predictor variable does not explained the response. In this case, the cointegration models become unusable if the expert wants to keep the all predictors selected in order to understand the formation of the response.

The use of standardized series then allows the use of models for cointegration but simpler than those possible in structural equations modelling configurations. Two approaches are reasonable for a finer statistical analysis, expert establishes his theoretical model, then the estimation SEM method aims to confirm or disprove their hypotheses. In contrast, it happens in many cases that then the relationships between variables are not completely known then an exploratory approach is preferred. It will then construct a "free" model using information available in the data.

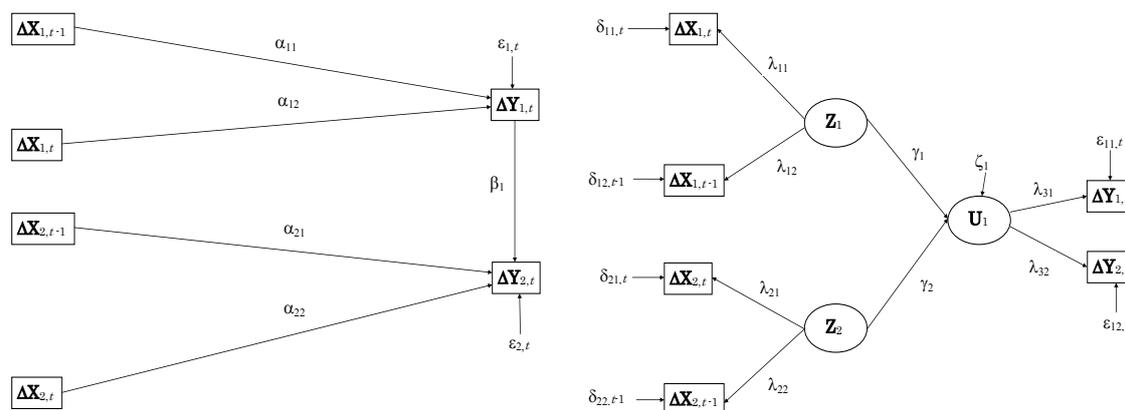


Figure 1: (a) Cointegration model (b) Structural equations model

The building principle of this type of model involves five steps: (i) building blocks of variables, (ii) estimation of latent variables, (iii) establishing statistical relationships between these latent variables, (iv) orientation links between latent variables and (v) application of the LISREL method or PLS approach to estimate the proposed free model. The first step returns to cluster time series standardized or differentiated using exploratory factor analysis (Principal Components Analysis with oblique rotation, for example). This provides several one-dimensional groups of variables (usually all the eigenvalues are smaller than one, except the largest). The estimation of latent variables is to select for each block of variables, the first principal component. The relationships between the set of first principal components are constructed using partial linear correlation coefficients. A link between two latent variables will be considered significant if the p -value of the statistical test is less than a threshold (eg, 0.01). Then, the orientation of links is generally left to the expert, although it is possible to optimize a structural model by maximizing, for example, a global R^2 . Finally, the proposed free model is estimated by LISREL method or PLS approach.

2 Application

We have 7 predictors (X_1, \dots, X_7) and 7 responses (Y_1, \dots, Y_7), time series (blue curve on fig. 2(a)). Each them has its segmented version (red curve on fig. 2 (a)). For instance, X_1 appears on row 4 and column 1, whereas Y_1 is given on row 1 and column 2. Many treatments have been performed on the actual anonymized data such as cointegration models, structural equations models fixed by an expert and free models. We only present the results of two free models. The first ones is built on differentiated series ; the second ones is applied on the standardized curves by segmentation. The figure 2(b) provides the scatter plots of X 's (in column) and Y 's (in row) deflated on other regressors. For instance, the link between X_7 and Y_7 (row 1 - col. 7) is disturbed by two outliers corresponding to two common breakpoints ($t = 140$ and $t = 231$), whereas there is a quite linear trend between X_4 and Y_3 (row 5 - col. 4). The scatter plots on differentiated data and on standardized data with segmentation are given on figures 3(a) and 3(b). Obviously we can see that the problem of disturbance due to the outliers stays for the differentiated data

with respect to the standardized data by segmentation. The latter transformation has the great advantage to capture better linear trend because it takes into account the breakpoint in the time series. For instance, there is a partial linear trend between X_7 and Y_7 .

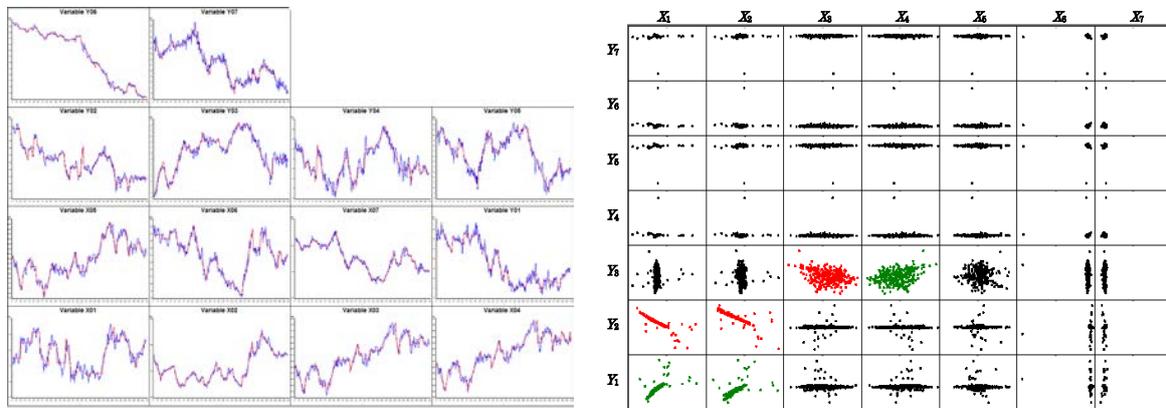


Figure 2: (a) Original and segmented data (b) Scatter plots on original data

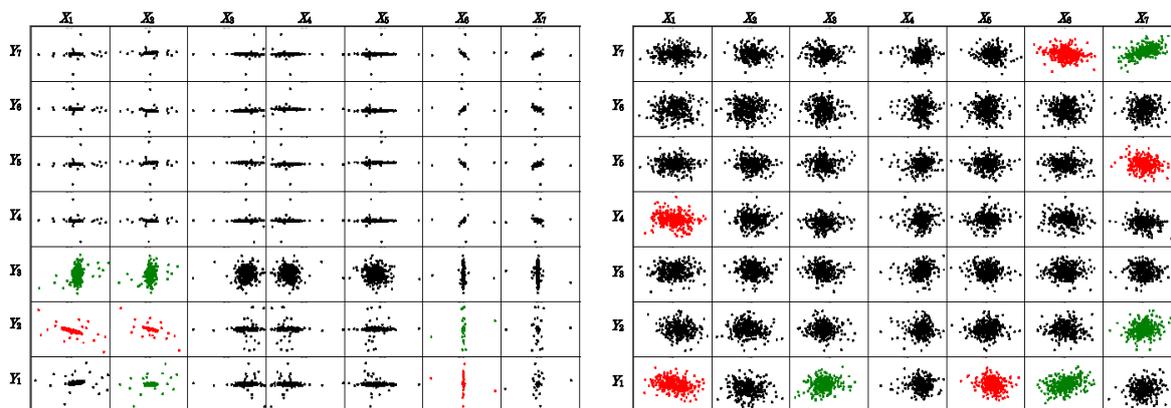


Figure 3: (a) Scatter plots on differentiated data (b) Scatter plots on standardized data

In addition, we have added to these, 14 time series, their delayed $t - 1$ series. However, this work is not limited to 1-lag series. It can be straightforwardly extended to p -lag series without altering anything in the method. The advantage is that $AR(p)$ processes can thus be modeled. Besides, it is not really necessary to know the autoregression order of each series involved (and indeed, original series may have different autoregression orders): one of the interests of the SEM-based methodology is to be able to deal with too many variables (current series, lagged series) and yet draw a regularized and robust model. For each model, the clustering of variables was applied separately on all of $X_j^{(t)}, X_j^{(t-1)}$ and $Y_k^{(t)}, Y_k^{(t-1)}$, for $j, k = 1, 7$ to keep an explanatory structure between predictors and responses. For both models, the same number of groups for

the X 's and Y 's were obtained, 6 and 8, respectively (see table 1). We note that the contents of the groups of the two models are quite similar, both for the X 's and for the Y 's. The only difference is the series grouping with their delayed for segmentation model series: $G_5^{(X)}$ and $G_1^{(Y)}$. Indeed $G_5^{(X)} = \{X_2^{(t-1)}; X_2^{(t)}\}$ and $G_1^{(Y)} = \{Y_4^{(t-1)}; Y_4^{(t)}\}$ for normalized data by segmentation, whereas $G_3^{(X)} = \{X_2^{(t-1)}, X_6^{(t-1)}, X_7^{(t-1)}\}$ and $G_4^{(X)} = \{X_2^{(t)}, X_6^{(t)}, X_7^{(t)}\}$ on the one hand and $G_3^{(Y)} = \{Y_2^{(t-1)}, Y_4^{(t-1)}\}$ and $G_4^{(Y)} = \{Y_2^{(t)}, Y_4^{(t)}\}$, on the other hand for differentiated data. Each group is framed in red on the figures 4(a), 4(b), 5(a) and 5(b).

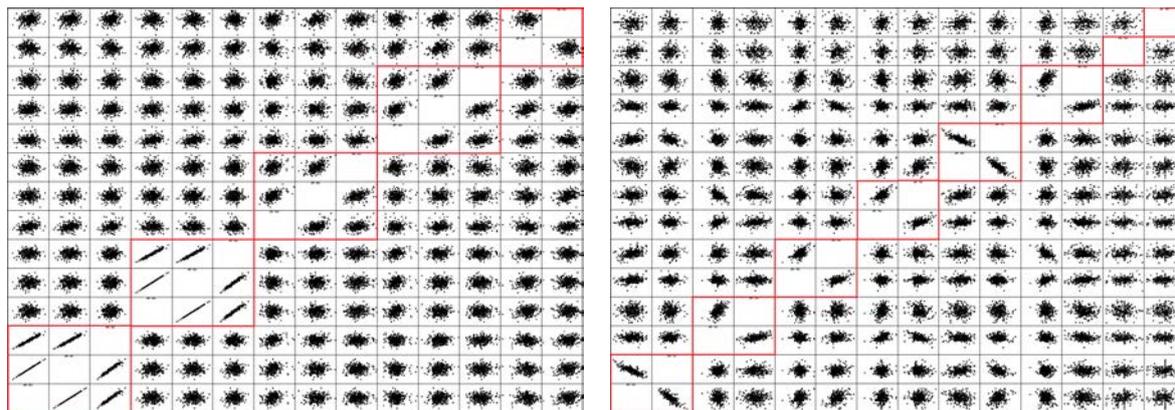


Figure 4: (a) Clustering of differentiated data on X (b) Clustering differentiated data on Y

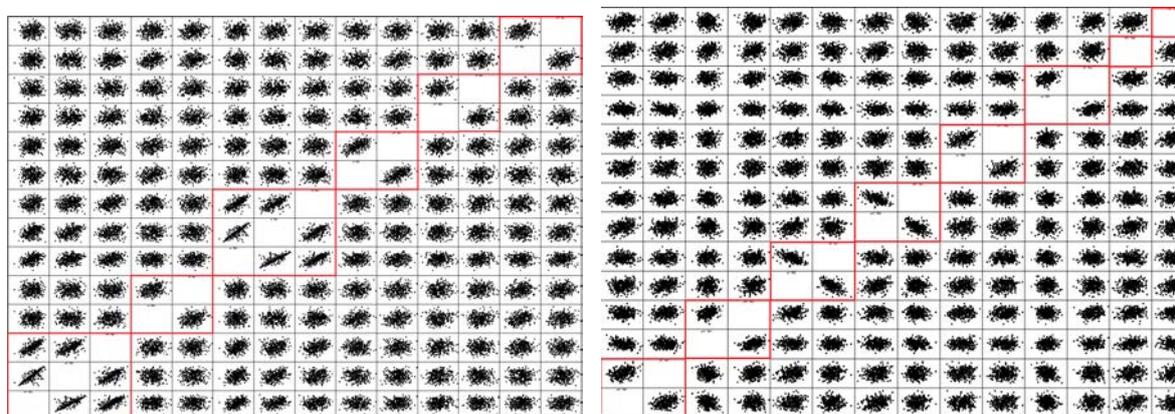


Figure 5: (a) Clustering on standardized data on X (b) clustering on standardized data on Y

Figures 6(a) and 6(b) show two free structural equations models constructed using the PLS approach. For differentiated data model, the latent variables associated with groups of predictors and responses are respectively named \mathbf{V}_g and \mathbf{W}_g , where g is the number of the associated group ; for the second model, we have \mathbf{Z}_g and \mathbf{U}_g . The links between latent variables were

established only if p -value of the test of partial correlation coefficient was less than 0.01 in order to avoid overly complex models. 27 and 23 edges, respectively, were obtained for differentiated and standardized series. In addition, the latent variables \mathbf{W}_7 , \mathbf{W}_8 and \mathbf{U}_5 have no significant relationship with others. This corresponds to the series Y_6 delayed. In the first model $\mathbf{W}_2, \mathbf{W}_3, \mathbf{W}_4, \mathbf{W}_5$ are targets (they point to any other latent variable), whereas in the standard model on segmentation series, only two targets were found: \mathbf{U}_3 and \mathbf{U}_8 (the R^2 are framed).

Two targets are common to both models: \mathbf{W}_5 and \mathbf{U}_3 summarizing $Y_5^{(t)}$ and $Y_7^{(t)}$. These correspond to very important business variables. However, the latent variables that explain (Student's t on the arrows) in the two models do not contain the same variable: $X_1^{(t-1)}, X_1^{(t)}$ and $X_2^{(t)}$ for the differentiated series and $X_3^{(t)}, X_4^{(t)}$ and $X_5^{(t)}$ for the standardized series. These last three appeared more consistent than the first in terms of business. Target \mathbf{W}_4 and \mathbf{U}_8 are also carriers of information with their variables $Y_2^{(t)}, Y_4^{(t)}$ and $Y_2^{(t)}$, respectively. This is not the case for the target \mathbf{W}_2 . The structural model of differentiated data tends to over-estimate Student's t and R^2 higher links between latent variables from the model using standard segmentation series, which can be symptomatic example of poor consideration of nonlinear relationships or a non taken into account of breakdown in multivariate time series for the first free model based on the differentiated data while this is not the case for the segmentation approach.

Grp. X	SEM diff. (V)	SEM segm. (Z)	Grp. Y	SEM diff. (W)	SEM segm. (U)
$G_1^{(X)}$	$X_3^{(t-1)}, X_4^{(t-1)}, X_5^{(t-1)}$	$X_3^{(t-1)}, X_4^{(t-1)}, X_5^{(t-1)}$	$G_1^{(Y)}$	$Y_5^{(t-1)}, Y_7^{(t-1)}$	$Y_4^{(t-1)}, Y_4^{(t)}$
$G_2^{(X)}$	$X_3^{(t)}, X_4^{(t)}, X_5^{(t)}$	$X_6^{(t)}, X_7^{(t)}$	$G_2^{(Y)}$	$Y_1^{(t)}, Y_3^{(t)}$	$Y_1^{(t-1)}, Y_3^{(t-1)}$
$G_3^{(X)}$	$X_2^{(t-1)}, X_6^{(t-1)}, X_7^{(t-1)}$	$X_3^{(t)}, X_4^{(t)}, X_5^{(t)}$	$G_3^{(Y)}$	$Y_2^{(t-1)}, Y_4^{(t-1)}$	$Y_5^{(t)}, Y_7^{(t)}$
$G_4^{(X)}$	$X_2^{(t)}, X_6^{(t)}, X_7^{(t)}$	$X_1^{(t-1)}, X_1^{(t)}$	$G_4^{(Y)}$	$Y_2^{(t)}, Y_4^{(t)}$	$Y_5^{(t-1)}, Y_7^{(t-1)}$
$G_5^{(X)}$	$X_1^{(t)}$	$X_2^{(t-1)}, X_2^{(t)}$	$G_5^{(Y)}$	$Y_5^{(t)}, Y_7^{(t)}$	$Y_6^{(t-1)}, Y_6^{(t)}$
$G_6^{(X)}$	$X_1^{(t-1)}$	$X_6^{(t-1)}, X_7^{(t-1)}$	$G_6^{(Y)}$	$Y_1^{(t-1)}, Y_3^{(t-1)}$	$Y_1^{(t)}, Y_3^{(t)}$
n.a.	n.a.	n.a.	$G_7^{(Y)}$	$Y_6^{(t-1)}$	$Y_2^{(t-1)}$
n.a.	n.a.	n.a.	$G_8^{(Y)}$	$Y_6^{(t)}$	$Y_2^{(t)}$

Table 1: Clusters of the both models

3 Concluding remarks, application and further research

In this paper, we built a model to understand how the connections formed between predictors and multivariate responses for irregular time series. First, to take into account non-linearity, break point, volatility between variables, we chose to segment these series as linear segments to eliminate the non-stationarity, standardizing the raw series. Then, we used an exploratory approach to build a free structural equation modelling to establish links between these standardized variables. The latter allows building blocks (groups) of homogeneous variables, then to summarize with latent variables, to research significant links between them and finally to apply the PLS approach on the proposed free model. In the application, the comparison of the results of the two models (differentiated *vs* standardized series) has shown a greater consistency of results in terms of stability coefficients, significant relationships and business aspects for the model with normalized series by segmentation. Our future work will involve to extend our method on the forecasting of multivariate time series. We will use the two approaches: non parametric

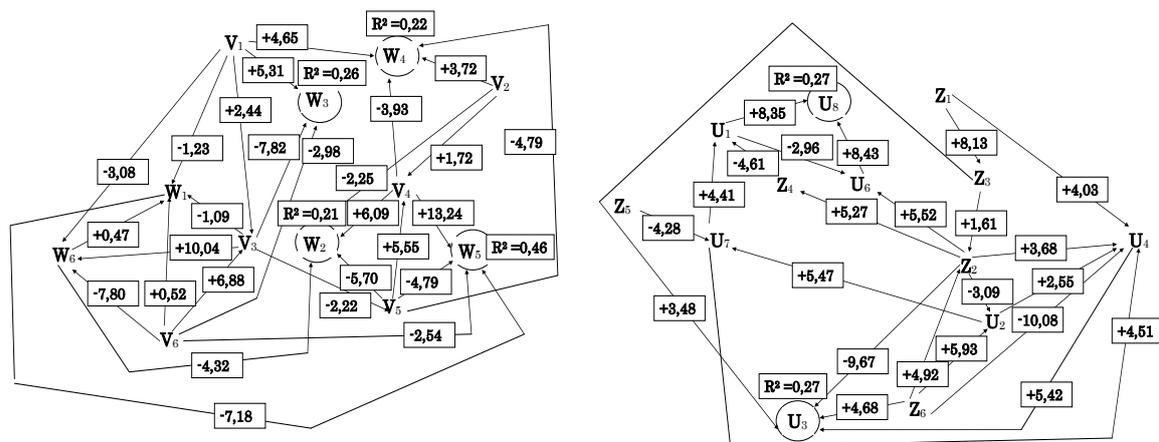


Figure 6: (a) Free model on differentiated data (b) Free model on standardized data

forecasting with detection of similar linear trends (segments) and Markov chains to estimate the probability of transition from one segment to another.

Bibliography

- [1] Derquenne, Ch. et Hallais, C. (2003): Une méthode alternative à l'approche PLS : comparaison et application aux modèles conceptuels marketing, RSA, LII, 37-72.
- [2] Derquenne, C. (2010): An Explanatory Segmentation Method for Time Series, in *Proceedings of Compstat 2010*, Y. Lechevallier & G. Saporta (eds.), 1st Edition, pp. 935-942.
- [3] Derquenne, C. (2011): Segmentation of Time Series with Heteroskedastic Components, 58th *World Statistical Congress of ISI*, Dublin, Ireland.
- [4] Derquenne, C. (2012): Meta-segmentation of time series for searching a better segmentation, in *Proc. of Compstat 2012*, Limassol, Cyprus, pp. 191-204.
- [5] Derquenne, C. (2013): Clustering of time series via a segmentation approach, IFCS 2013, *Program and Book of Abstracts*, Tilburg, The Netherlands p. 134.
- [6] Hébrail G., Hugueney B., Lechevallier Y. and Rossi F. (2010): Exploratory analysis of functional data via clustering and optimal segmentation. *Neurocomputing* 73(7-9): 1125–1141.
- [7] Joreskog, K.G. & Sorbun, D. (1979): *Advance in Factor Analysis and Structural Equation Models*, Abt Books, Cambridge.
- [8] Lavielle, M. (2009): Detection of Changes using a Penalized Contrast (the DCPC algorithm), http://www.math.u-psud.fr/~lavielleprogrammes_lavielle.html.
- [9] Wold, H. (1982): Soft Modelling : The Basic Design and Some Extensions, in Joreskog K.G. and Sorbun D., Editors, *Systems under Indirect Observation*, 2, 1-54, N-H, Amsterdam.

Study on the choice of regression quantile threshold in a POT model

Martin Schindler, *Technical University of Liberec*, martin.schindler@tul.cz

Jan Picek, *Technical University of Liberec*, jan.picek@tul.cz

Jan Kyselý, *Technical University of Liberec*, kysely@ufa.cas.cz

Abstract. We adopt the Peak Over Threshold (POT) method with a non-stationary threshold to estimate high quantiles. We use a linear regression quantile as a time-dependent threshold, assuming that a linear trend is present in the data. Using Monte Carlo simulations we try to find the threshold (regression quantile) which would be optimal with respect to the reliability of the estimates of high quantiles. The reliability is measured by the coverage probability of the confidence interval. We investigate how the choice of the optimal threshold changes if we change the sample size, estimated quantile or the estimate itself. We give particular recommendation in case of underlying Gumbel distribution specifying how the threshold should be decreased with decreasing sample size or increasing confidence of the confidence interval. Besides we conclude that the heavier the tails of the distribution, the lower the threshold should be.

Keywords. Monte Carlo simulation, Peak Over Threshold model, regression quantiles, return level.

1 Introduction

The Peak Over Threshold (POT) model, see e.g. Coles (2001), is frequently used and represents a very important branch of extreme value modelling. It is based on the Pickands-Balkema-de Haan theorem, see Balkema and de Haan (1974) and Pickands (1975), which says that for a large class of underlying distribution functions excesses over a given (sufficiently high) threshold are approximately distributed according to a generalized Pareto distribution (GPD), given that the threshold is exceeded.

We wish to estimate the m -year return level (quantile of a distribution) from a data sample using the POT method. The purpose of this work is to find the optimal threshold. We assume that the data comes from the following model (i.i.d. with linear trend):

- data y_1, \dots, y_n follow a model $y_i = e_i + \beta_0 + \beta_1 i$, $i = 1, \dots, n$, where e_i is a random sample with a distribution function $F(x)$ (density function $f(x)$).

- m -year return level (line) is

$$F^{-1}\left(1 - \frac{1}{m}\right) + \beta_1 i, \quad 1 \leq i \leq n$$

which is the population $(1 - \frac{1}{m})$ -th regression quantile line.

In a usual model a certain ordinary quantile is taken as the threshold. Here the natural choice for the threshold is thus an α -th regression quantile $(\hat{\beta}_0(\alpha), \hat{\beta}_1(\alpha))^T$ which is a solution of the following minimization problem:

$$\min_{(t_0, t_1) \in \mathbf{R}^2} \sum_{i=1}^n \rho_\alpha(y_i - t_0 - t_1 i), \text{ where}$$

$\rho_\alpha(u) = |u| \{ (1 - \alpha)I[u < 0] + \alpha I[u > 0] \}$, $u \in \mathbf{R}$, $\alpha \in (0, 1)$. The first component of the solution $\hat{\beta}_0(\alpha)$ estimates $\beta_0 + F^{-1}(\alpha)$ and the second component $\hat{\beta}_1(\alpha)$ estimates β_1 . The symbol I denotes the indicator function.

This minimization is in fact a linear programming problem, so it can be solved by some modification of the simplex algorithm, which is described in Koenker and Bassett (1978). If the sample size n is very large, it can be advantageous to use alternative computation methods described in Koenker and Portnoy (1997). As stated by Koenker and Bassett (1978), the set of the α -th regression quantiles has at least one element $(\hat{\beta}_0(\alpha), \hat{\beta}_1(\alpha))^T$ defining such a regression hyperplane so that at least two elements of $\{(i, y_i); i = 1, \dots, n\}$ lie on it. These two elements identify the regression quantile uniquely. Portnoy (1991) showed that, if we consider this element of the solution $(\hat{\beta}_0(\alpha), \hat{\beta}_1(\alpha))^T$ as a function of α ($0 < \alpha < 1$), the number of such distinct elements is proportional to $n \log n$ as n increases to infinity.

The regression quantile procedures is (compared to e.g. least square line) more robust and resistant against deviations in the response variable and also more flexible tool for estimation of threshold compared to taking e.g. an ordinary quantile of residuals from the least square line.

The aim is to find α such that the α -th regression quantile line $\hat{\beta}_0(\alpha) + \hat{\beta}_1(\alpha)i$ is the “optimal” threshold for the POT method, where the criterion for the threshold to be “optimal” is taken to be the maximization of the probability that a confidence band covers the real return level.

Maximum likelihood estimation is used to compute the GPD parameters, followed by the delta method to construct the confidence intervals (bands) for return levels. Although superior methods for constructing confidence intervals (e.g. bootstrap) exist, we have chosen the delta method because of its computational feasibility. The bootstrap method is used for similar study in Kyselý (2010). Moreover, Arcones (2003), Cheung and Lee (2005) and Lee (1999) describe the ways how to construct a bootstrap confidence interval in such a situation. Nevertheless, bootstrap is not used in this study due to the computational hardness.

2 Simulation setting

We have used Monte Carlo simulations to assess the optimal threshold for a POT model. As applications of the POT method in climatology are of our interest, we want the data we simulate to mimic typical datasets of summer maximum daily temperature or rainfall but at the same time to have a relatively simple structure. Since this kind of data has mostly distribution with

exponential tails and maxima from such distribution converge asymptotically to the Gumbel distribution, we have chosen the underlying distribution to be the standard Gumbel. So as the first step, we have used the following setting of parameters in this study:

- Choice of $F(x)$: standard Gumbel distribution
- Sample size $n = (20, 40, 80, 160)$ years $\times 90$ (90 corresponds to the length of season)
- We set trend $\beta_1 = 0.05/90$ (this parameter is not important due to equivariancy of regression quantiles)
- We wish to estimate (20, 50, 100, 200)-year return levels
- We use (75, 80, 85, 87, 89, 90, 91, 92, 93, 94, 95, 96, 97, 98)% regression quantile as threshold
- We compute (80, 85, 90, 95, 99, 99.9)% confidence bands to estimate the return levels
- For every setting of the parameters we generate 6600 sets of data.

The results of the simulations are described by plots in figure 1 that show the estimates of the probability that the estimated confidence interval (band) covers the return level (line) with respect to the regression quantile threshold used.

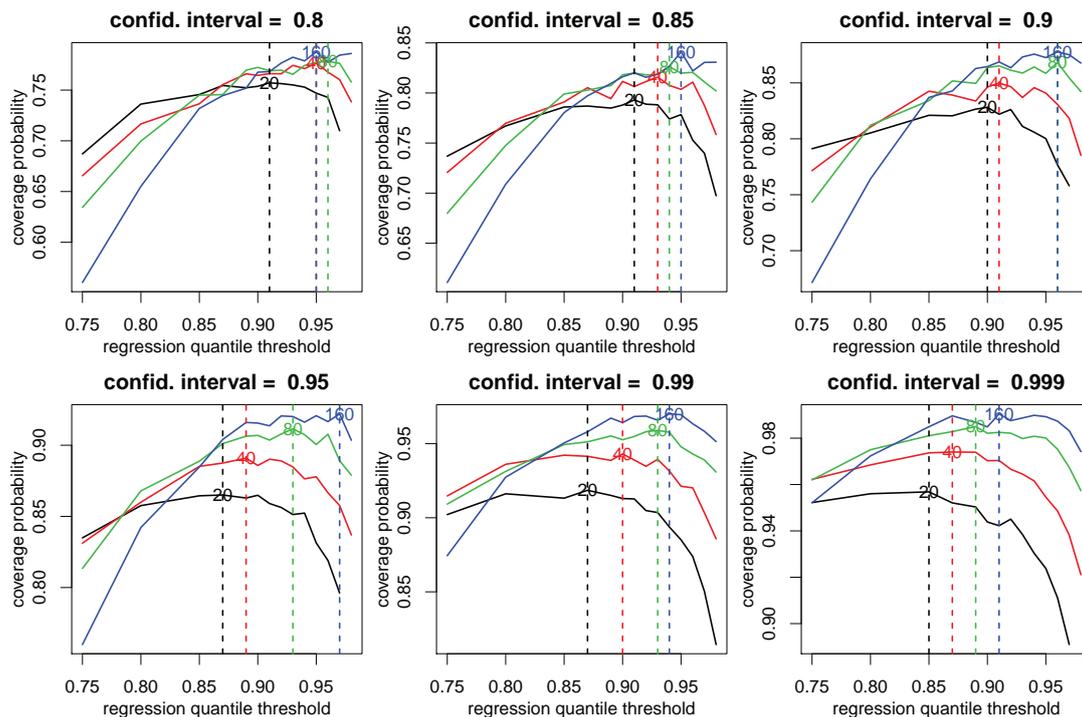


Figure 1: Average coverage probability for different sample sizes. Return level: 100 years

In figure 1 the different solid curves represent different sample sizes (in years). The vertical dashed lines always depicts the optimal regression quantile threshold, i.e. the threshold for which the maximum coverage probability is achieved.

3 Results

On the plots in figure 2 the optimal regression quantile with respect to sample size (in years) can be seen. The plots in figure 3 show the coverage probability for the optimal regression quantile threshold with respect to sample size (in years). In all the figures the different plots stand for different confidence intervals (bands) or return periods (levels). Moreover, the coverage probabilities for the optimal regression quantiles are summarized in table 1.

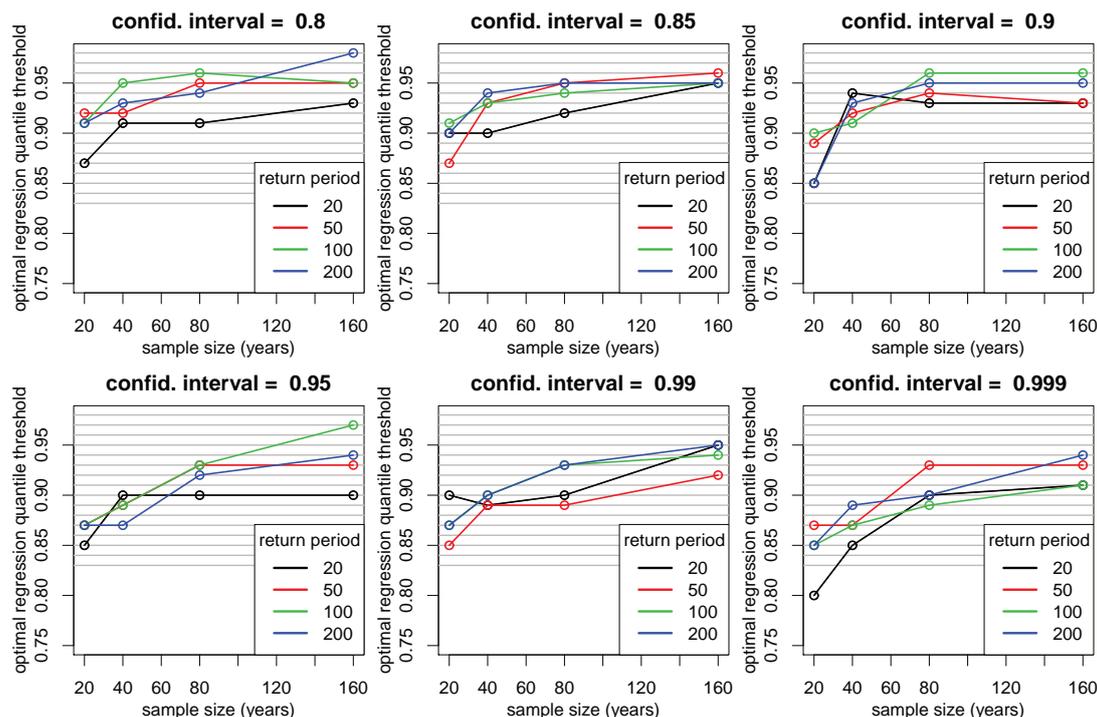


Figure 2: Optimal regression quantile wrt. sample size

It is quite clear from the figures that the optimal threshold is usually lower than the 95% regression quantile (see e.g. figure 2). The optimal threshold is approximately equal to the 90% regression quantile, unless low confidence for interval is used and very large sample size is available. If sample size decreases, the optimal regression quantile threshold decreases as well. If confidence of the confidence interval increases, the optimal regression quantile threshold decreases.

Figure 1 also shows that the real coverage probability is lower than expected and decreases (at least for the optimal threshold) with decreasing sample size. It can be seen that lowering threshold to the 75% regression quantile causes that the increasing bias of the estimate decreases the coverage probability (and this effect of bias is stronger for larger sample sizes). On the other hand, with increasing threshold, the increasing variance of the estimate causes decreasing coverage probability (and the smaller sample size, the more evident this effect of variance is). This is probably the reason for the fact that the optimal regression quantile threshold rises with increasing sample size.

The negative effect of bias is also the more striking the lower the confidence of interval is (see

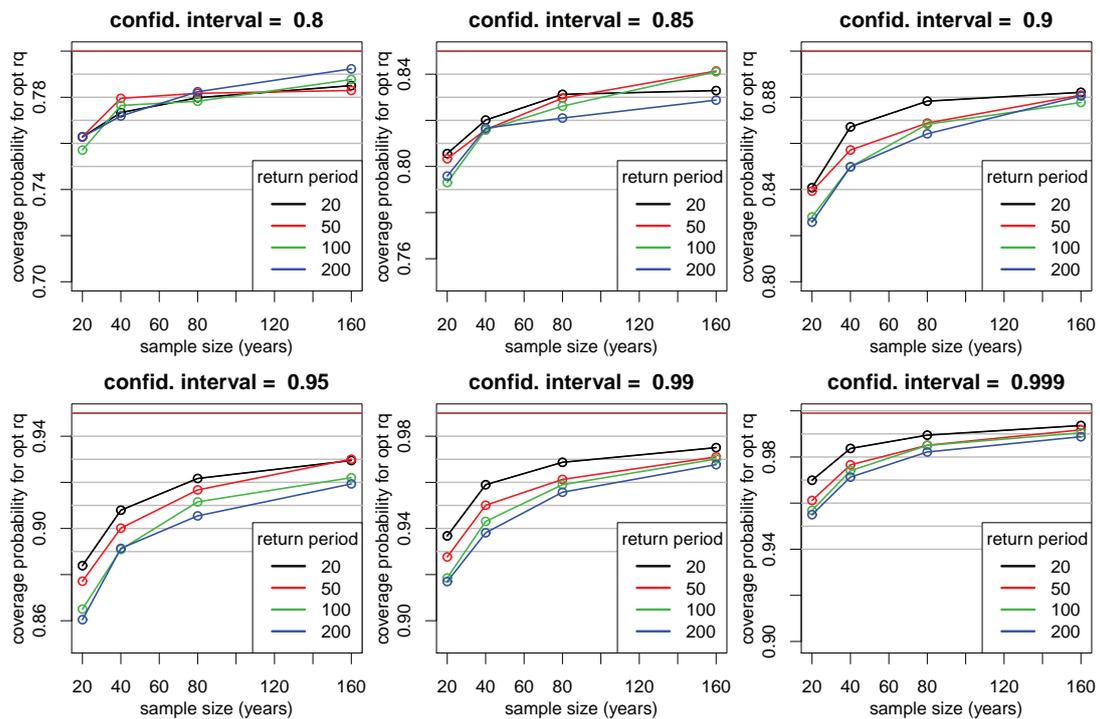


Figure 3: Average coverage probability for optimal regression quantile wrt. sample size.

return period = 20							return period = 50					
confid. int.	0.8	0.85	0.9	0.95	0.99	0.999	0.8	0.85	0.9	0.95	0.99	0.999
sample size												
20	0.763	0.806	0.841	0.884	0.937	0.970	0.763	0.803	0.839	0.877	0.928	0.961
50	0.773	0.820	0.867	0.908	0.959	0.984	0.780	0.816	0.857	0.900	0.950	0.977
100	0.780	0.831	0.878	0.922	0.969	0.989	0.782	0.830	0.869	0.917	0.961	0.985
200	0.785	0.833	0.882	0.929	0.975	0.994	0.783	0.841	0.881	0.930	0.971	0.992

return period= 100							return period= 200					
confid. int.	0.8	0.85	0.9	0.95	0.99	0.999	0.8	0.85	0.9	0.95	0.99	0.999
sample size												
20	0.757	0.793	0.828	0.865	0.919	0.957	0.763	0.796	0.826	0.860	0.917	0.955
50	0.776	0.816	0.850	0.891	0.943	0.974	0.772	0.817	0.850	0.891	0.938	0.971
100	0.778	0.826	0.868	0.912	0.959	0.985	0.782	0.821	0.864	0.905	0.956	0.982
200	0.788	0.841	0.878	0.922	0.970	0.990	0.792	0.829	0.881	0.919	0.968	0.989

Table 1: Average coverage probability for optimal regression quantile

figure 1). This is probably due to the fact that the lower confidence means narrower interval (band) and so higher vulnerability to bias. This seems to explain why the optimal threshold declines with increasing confidence.

So figures 1 and 2 suggest that for i.i.d. observations from Gumbel distribution the 95% regression quantile could be the optimal threshold if sample size were 160 years (7200 observations) and if 80% confidence interval (band) were used. Then for twice smaller sample size

the threshold should be decreased by 2% points and when confidence of the interval rises from $(1 - \alpha)100\%$ to $(1 - \frac{\alpha}{3})100\%$ the threshold should be decreased by 1% point. These recommendations were based on the assumption that the optimal choice of the threshold more or less does not depend on the return period (level) estimated. This assumption can be made upon inspection of figure 2.

The next notable fact is that the real coverage probability of confidence intervals (bands) is much lower than expected. This is visible particularly in figure 3 and table 1. For real 95% confidence (estimated by the coverage probability) we would need to construct approximately 99% confidence interval. Moreover, with decreasing sample size or increasing return period the real confidence goes down. Interesting feature is that the highest deviation from expected confidence occurs for 95% confidence interval.

Simulations also imply (not presented) that confidence interval is biased downwards, i.e. return level is too often higher than the confidence interval indicates. This is a striking fact that can be very dangerous in practice, and illustrates an important limitation of the delta method.

4 Effect of the underlying distribution

Let \mathbf{X} be the design matrix of our linear model and denote the sample regression quantile by $\hat{\beta}_n(\alpha) = (\hat{\beta}_0(\alpha), \hat{\beta}_1(\alpha))^T$. If $\lim_{n \rightarrow \infty} n^{-1}\mathbf{X}'\mathbf{X} = \mathbf{Q}$ is a positive definite matrix, then it can be shown, see Koenker and Bassett (1978), that $\sqrt{n}(\hat{\beta}_n(\alpha) - \beta(\alpha))$ converges in distribution to two-variate Gaussian with mean zero and covariance matrix $\frac{\alpha(1-\alpha)}{f^2(F^{-1}(\alpha))} \mathbf{Q}^{-1}$.

Thus the variance of the intercept as well as of the slope of the estimated regression quantile depends on the underlying distribution, more precisely is proportional to $\frac{\alpha(1-\alpha)}{f^2(F^{-1}(\alpha))}$. It is easy to show that for F standard Gumbel distribution $\frac{\alpha(1-\alpha)}{f^2(F^{-1}(\alpha))} \doteq \frac{1}{1-\alpha}$ for α in the vicinity of 1. For F standard normal distribution (light-tailed distribution) $\frac{\alpha(1-\alpha)}{f^2(F^{-1}(\alpha))} \doteq \frac{1}{\sqrt{1-\alpha}}$. On the other hand, if F is GEV distribution with shape parameter 0.8 (heavy-tailed distribution) this term can be approximated by $\frac{1}{(1-\alpha)^{2.8}}$. So the variance of the regression quantile increases faster as α increases to 1 which results in faster increase of non-coverage of the confidence belt, for heavy-tailed distribution. This is the main reason why the optimal regression threshold is lower for heavy-tailed distribution, see figure 4.

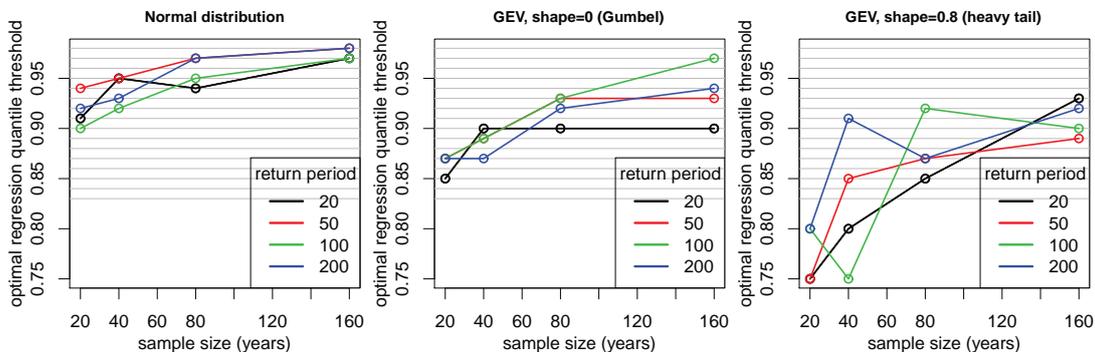


Figure 4: Optimal regression quantile wrt. sample size. Confidence interval: 95%

5 Conclusions

We have shown that the usual 95% regression quantile does not always have to be the best choice for a threshold in a POT model. In our model a 90% regression quantile seems to be a better choice for moderate sample sizes and 95% confidence interval. The simulation study also suggested that the threshold should be decreased with decreasing sample size or increasing confidence of the confidence interval (band) that is used. Very unfavourable fact is that the actual coverage of the confidence intervals (bands) is much lower than expected. Solution, although only partial, can be using e.g. bootstrap confidence intervals.

It was also shown that the underlying distribution also influence the optimal choice of the threshold. For light-tailed distribution (e.g. normal) the optimal regression quantile as threshold is higher (usually around 95%) than for the Gumbel distribution. Contrary, for heavy-tailed distribution (e.g. GEV with shape parameter 0.8) the optimal threshold is lower (usually around 85%) than for the Gumbel distribution.

All the results presented were derived under simplified assumptions (e.g. the i.i.d. assumption). To be of more practical use we should probably look at data with a different structure, e.g. autocorrelated data. But that would be already a subject of a further research.

Acknowledgements

Martin Schindler and Jan Picek were supported by ESF operational programme “Education for Competitiveness” in the Czech Republic in the framework of project Support of engineering of excellent research and development teams at the Technical University of Liberec No. CZ.1.07/2.3.00/30.0065. Jan Kyselý was supported by ESF project CZ.1.07/2.3.00/20.0086 (KLIMATEXT).

Bibliography

- [1] Arcones M. A. (2003) *On the asymptotic accuracy of the bootstrap under arbitrary resampling size*. Annals of the Institute of Statistical Mathematics **55**, 563–583.
- [2] Balkema A. and de Haan L. (1974) *Residual life time at great age*. Annals of Probability **2**, 792–804.
- [3] Cheung K. Y. and Lee S. M. S. (2005) *Variance estimation for sample quantiles using the m out of n bootstrap*. Annals of The Institute of Statistical Mathematics **57**, 279–290.
- [4] Coles S. (2001) *An Introduction to Statistical Modeling of Extreme Values*. Springer, London.
- [5] Davison, A. C. and Hinkley, D. V. (1997). *Bootstrap Methods and their Application*. Cambridge University Press, Cambridge.
- [6] Koenker R. and Bassett G. (1978): *Regression quantiles*. Econometrica **46**, 33-50.
- [7] Koenker R. and Portnoy S. (1997): *The Gaussian Hare and the Laplacean Tortoise: Computability of Squared-error vs Absolute Error Estimators, (with discussion)*. Statistical Science **12**, 279-300.

- [8] Kyselý J. (2010) *Coverage probability of bootstrap confidence intervals in heavy-tailed frequency models, with application to precipitation data*. Theoretical and Applied Climatology **101**, 345–361.
- [9] Kyselý J., Pícek J. and Beranová R. (2010) *Estimating extremes in climate change simulations using the peaks-over-threshold method with a non-stationary threshold*. Global and Planetary Change **72**, 55–68.
- [10] Lee S. M. S. (1999) *On a class of m out of n bootstrap confidence intervals*. Journal Of The Royal Statistical Society. Series B: Statistical Methodology **61**, 901–911.
- [11] Pickands J. (1975) *Statistical inference using extreme order statistics*. Annals of Statistics **3**, 119–131.
- [12] Portnoy S. (1975) *Asymptotic behavior of the number of regression quantile breakpoints*. SIAM Journal of Scientific and Statistical Computing **12**, 867–883.

Reduced K-means with sparse loadings

Ryo Takahashi, *Osaka University*, takahashi@bm.hus.osaka-u.ac.jp

Abstract. For a data matrix of objects by variables, De Soete and Carrollâs (1994) reduced K-means (RKM) clustering is formulated as simultaneously clustering the objects into a smaller number of clusters and finding the principal components summarizing the variables. In this paper, we propose a modified RKM procedure which produces a sparse loading matrix including a number of zero elements. Such a sparse matrix facilitates interpreting the relationships of variables to components, as they can be captured only by focusing on nonzero loadings. In our proposed method, the RKM loss function is minimized over membership, cluster center and loading matrices subject to the following two constraints; [1] the one constraining the cardinality of loadings to be a specified integer, and [2] an orthonormality condition for components. A key property of the procedure is that its loss function is decomposed as the sum of a term irrelevant to loadings and their function being easily minimized under the cardinality constraint. Using this property, we present an efficient alternating least-squares algorithm. The proposed new RKM is illustrated with a real data set.

Keywords. Cluster analysis, Sparse PCA, Dimension reduction, Perfect simple structure

1 Introduction

Let \mathbf{X} be an n -objects \times p -variables data matrix. De Soete & Carroll (1994) proposed a method simultaneously partitioning the n objects in \mathbf{X} into K clusters and finding the q components that summarizes p variables, with $K < n$ and $q < p$. This method, called *Reduced K-means* (RKM), is formulated by combining the loss function of the K-means clustering and that for principal component analysis (PCA) into a single criterion to be optimized. Therefore, RKM could identify the low-dimensional space that keeps the information about the cluster structure underlying a data set \mathbf{X} . The model of RKM is written as

$$f_{RKM}(\mathbf{U}, \mathbf{C}, \mathbf{A}) = \|\mathbf{X} - \mathbf{UCA}'\|^2, \quad (1)$$

where \mathbf{U} is an $n \times K$ binary indicator matrix, \mathbf{C} is a $K \times q$ matrix of cluster centroids, and \mathbf{A}

is a $p \times q$ loading matrix. The constraint $\mathbf{A}'\mathbf{A} = \mathbf{I}$ is imposed on the loading matrix.

The purpose of this paper is to modify RKM so that loadings matrix \mathbf{A} is easier to interpret. If the matrix is sparse, i.e., includes a number of zero elements, its interpretation is facilitated. That is, we propose a new RKM procedure that constrains the loading matrix to be sparse. In the proposed method, the cardinality constraint on the loading matrix is introduced that controls the number of zero elements in the loading matrix. Furthermore, an effective alternating least-squares algorithm is also presented. A key property of the algorithm is that the loss function is decomposed as the sum of a term irrelevant to loadings and their function being easily minimized under the cardinality constraint.

2 Notation

For the convenience for readers, the notation common to all sections is listed here.

- $n, p \dots$ number of objects and variables respectively;
- $K \dots$ number of clusters;
- $q \dots$ number of principle components ($q \leq p$);
- $\mathbf{X} \dots (n \times p)$ data matrix; If the variables are expressed by different units of measurement they are standardized to have mean zero and unit variance;
- $\mathbf{U} \dots (n \times K)$ binary object membership matrix specifying for each objects the membership to each clusters, i.e., $u_{ij} = 1$ if the i th object belongs to the j th cluster, $u_{ij} = 0$ otherwise;
- $\mathbf{C} \dots (K \times q)$ cluster centroid matrix; The k th row of \mathbf{C} corresponds to the coordinate of the k th cluster centroid in a reduced space;
- $\mathbf{A} \dots (p \times q)$ loading matrix, where a_{jl} are the coefficients of the linear combinations of the observed variables;
- $m \dots$ number of cardinality on the loading matrix \mathbf{A} ($m < p \times q$);
- $\mathbf{F} \dots (n \times q)$ component score matrix. The f_{il} is the score of the i th object for the l th component;
- $\mathbf{I}_q \dots (q \times q)$ identity matrix;

3 Proposed Model

The proposed model is mathematically specified as

$$\mathbf{X} = \mathbf{UCA}' + \mathbf{E}, \quad (2)$$

where \mathbf{E} contains error. The model (2) is of the same form as the RKM model, but we introduce the following constraints:

$$\mathbf{C}'\mathbf{U}'\mathbf{U}\mathbf{C} = n\mathbf{I}_q, \quad (3)$$

$$\text{Card}(\mathbf{A}) = m. \quad (4)$$

In (3), we use the multiplier n for obtaining standardized score. The constraint (4) is the cardinality constraint on the loading matrix \mathbf{A} , which controls the number of non-zero elements of \mathbf{A} . The cardinality parameter m must be prespecified by users. The squared norm of \mathbf{E} in (2) is minimized with the constraints (3) and (4). That is, the proposed procedure is formulated as

$$\text{Min}_{\mathbf{U}, \mathbf{C}, \mathbf{A}} \|\mathbf{X} - \mathbf{U}\mathbf{C}\mathbf{A}'\|^2 \quad \text{s.t. } \mathbf{C}'\mathbf{U}'\mathbf{U}\mathbf{C} = n\mathbf{I}, \text{ Card}(\mathbf{A}) = m \quad (5)$$

over \mathbf{U}, \mathbf{C} and \mathbf{A} .

4 Algorithm

The constrained minimization problem of (5) can be solved by the following alternating least-squares algorithm.

Update \mathbf{U}

First, we present how to update the membership matrix \mathbf{U} with \mathbf{C} and \mathbf{A} being kept fixed.

For each $i = 1, \dots, n$ and each $k = 1, \dots, K$, we update $\mathbf{U} = (u_{ik})$ by

$$\mathbf{U} = (u_{ik}) = \begin{cases} 1 & \text{iff } \|\mathbf{x}_i - \mathbf{c}_k\mathbf{A}'\|^2 = \min\{\|\mathbf{x}_i - \mathbf{c}_s\mathbf{A}'\|^2 : s = 1, \dots, K; s \neq k\} \\ 0 & \text{otherwise} \end{cases}, \quad (6)$$

where \mathbf{x}_i ($i = 1, \dots, n$) is the i th row vector of \mathbf{X} , and \mathbf{c}_k ($k = 1, \dots, K$) is the k th row vector of \mathbf{C} . The equation (6) implies that each objects should be assigned to the single cluster such that the Euclidean distances between the objects and the cluster centroids to which they belong is minimal.

Update \mathbf{C}

We next propose the way to update \mathbf{C} with \mathbf{U} and \mathbf{A} being fixed. Due to the constraint (3), we should not take the way using by the conventional K-means algorithms. Instead, we adopt the following procedure.

Firstly, we define the matrix \mathbf{C}^\dagger by

$$\mathbf{C}^\dagger = \mathbf{M}^{\frac{1}{2}}\mathbf{C}, \quad (7)$$

where the matrix \mathbf{M} is constructed by $\mathbf{M} = n^{-1}\mathbf{U}'\mathbf{U}$. Hence, \mathbf{C}^\dagger is a columnwise orthonormal

matrix. By using \mathbf{C}^\dagger , the loss function (5) can be rewritten as follow:

$$\begin{aligned} f(\mathbf{C}^\dagger|\mathbf{U}, \mathbf{A}) &= \left\| \mathbf{X} - \mathbf{U}\mathbf{M}^{-\frac{1}{2}}\mathbf{C}^\dagger\mathbf{A}' \right\|^2 \\ &= \text{tr}\mathbf{X}'\mathbf{X} - 2\text{tr}\mathbf{A}'\mathbf{X}'\mathbf{U}\mathbf{M}^{-\frac{1}{2}}\mathbf{C}^\dagger + \text{tr}\mathbf{A}'\mathbf{A} \end{aligned} \quad (8)$$

In the expansion of (8), the first and third term is constant for \mathbf{C}^\dagger . So we can solve this optimization problem by finding \mathbf{C}^\dagger which maximize the second term: $\text{tr}\mathbf{A}'\mathbf{X}'\mathbf{U}\mathbf{M}^{-\frac{1}{2}}\mathbf{C}^\dagger$. Since \mathbf{C}^\dagger is columnwise orthonormal, this can be solved by using the following singular value decomposition (SVD).

$$\mathbf{A}'\mathbf{X}'\mathbf{U}\mathbf{M}^{-\frac{1}{2}} = \mathbf{K}\mathbf{\Lambda}\mathbf{L}' \quad (9)$$

with $\mathbf{K}'\mathbf{K} = \mathbf{I}_q$, $\mathbf{L}'\mathbf{L} = \mathbf{I}_K$, and $\mathbf{\Lambda}$ is a $q \times q$ diagonal matrix.. Then, the optimal \mathbf{C}^\dagger can be obtained by

$$\mathbf{C}^\dagger = \mathbf{L}\mathbf{K}' \quad (10)$$

Thus, the optimal \mathbf{C} is defined as follows.

$$\mathbf{C} = \mathbf{M}^{-\frac{1}{2}}\mathbf{L}\mathbf{K}' \quad (11)$$

Update A

The loss function (5) can be decomposed as follows:

$$\begin{aligned} \left\| \mathbf{X} - \mathbf{U}\mathbf{C}\mathbf{A}' \right\|^2 &= \left\| (\mathbf{X} - \mathbf{U}\mathbf{C}\mathbf{B}') + (\mathbf{U}\mathbf{C}\mathbf{B}' - \mathbf{U}\mathbf{C}\mathbf{A}') \right\|^2 \\ &= \left\| \mathbf{X} - \mathbf{U}\mathbf{C}\mathbf{B}' \right\|^2 + n \left\| \mathbf{B} - \mathbf{A} \right\|^2 \end{aligned} \quad (12)$$

where $\mathbf{B} = n^{-1}\mathbf{X}'\mathbf{U}\mathbf{C}$ (see Adachi & Trendafilov, 2014). The last identity follows from the orthogonality of $(\mathbf{X} - \mathbf{U}\mathbf{C}\mathbf{B}')$ and $(\mathbf{U}\mathbf{C}\mathbf{B}' - \mathbf{U}\mathbf{C}\mathbf{A}')$, or equivalently

$$\begin{aligned} (\mathbf{X} - \mathbf{U}\mathbf{C}\mathbf{B}')'(\mathbf{U}\mathbf{C}\mathbf{B}' - \mathbf{U}\mathbf{C}\mathbf{A}') &= \mathbf{X}'\mathbf{U}\mathbf{C}\mathbf{B}' - \mathbf{X}'\mathbf{U}\mathbf{C}\mathbf{A}' - \mathbf{B}\mathbf{C}'\mathbf{U}'\mathbf{U}\mathbf{C}\mathbf{B}' + \mathbf{B}\mathbf{C}'\mathbf{U}'\mathbf{U}\mathbf{C}\mathbf{A}' \\ &= n\mathbf{B}\mathbf{B}' - n\mathbf{B}\mathbf{A}' - n\mathbf{B}\mathbf{B}' + n\mathbf{B}\mathbf{A}' = {}_p\mathbf{O}_q \end{aligned} \quad (13)$$

with ${}_p\mathbf{O}_q$ the $p \times q$ matrix of zeros. For updating the loading matrix \mathbf{A} , we use the decomposed loss function (12), instead of (5). In (12), \mathbf{A} is only related to the second term of the right-hand side. Thus, the optimal \mathbf{A} can be obtained by minimizing the following function with the constraint (18).

$$\text{Min}_{\mathbf{A}} \left\| \mathbf{B} - \mathbf{A} \right\|^2 \quad \text{s.t. Card}(\mathbf{A}) = m \quad (14)$$

The matrix \mathbf{B} is given by $\mathbf{B} = n^{-1}\mathbf{X}'\mathbf{U}\mathbf{C}$. Since \mathbf{U} and \mathbf{C} are fixed, \mathbf{B} is also fixed. Then, we can solve the minimization problem (14) in the following manner:

$$a_{jl} = \begin{cases} b_{jl} & \text{iff } |b_{jl}| \text{ is the } m\text{th largest absolute value in } \mathbf{B} \\ 0 & \text{otherwise} \end{cases}, \tag{15}$$

where a_{jl} and b_{jl} is the (j, l) th element of \mathbf{A} and \mathbf{B} , respectively.

We described the artificial example of this process below.

Example.1: For given \mathbf{B} , the optimal loading matrix \mathbf{A} with $\text{Card}(\mathbf{A}) = 5$ is follow ($p = 4, q = 2$).

$$\mathbf{B} = \begin{bmatrix} -0.8 & 0.2 \\ 0.6 & -0.1 \\ 0.3 & -0.7 \\ -0.4 & 0.5 \end{bmatrix} \implies \begin{bmatrix} -0.8 & 0 \\ 0.6 & 0 \\ 0 & -0.7 \\ -0.4 & 0.5 \end{bmatrix} = \mathbf{A}$$

In this example, the $m(= 5)$ th largest absolute value in \mathbf{B} is $0.4 (= |b_{14}|)$. Thus, if $|b_{jl}| < 0.4$ ($j = 1, \dots, p; l = 1, \dots, q$), the corresponding elements of \mathbf{A} are set to be zero.

Initialization

In this algorithm, we need to set the initial values of \mathbf{U} and \mathbf{C} , preliminarily.

The initial value of \mathbf{U} can be chosen randomly or in a rational way. If \mathbf{U} has an empty column, we should restart this step.

Due to the constraint (3), the initial value of \mathbf{C} can not be given straightforwardly. Thus, we choose the initial \mathbf{C} by the following steps.

Step1 : Draw the element of a $K \times q$ matrix \mathbf{C}_{raw} from the uniform distribution $U(-1, 1)$.

Step2 : Calculate $\mathbf{C}_{cent} = \mathbf{C}_{raw} - \mathbf{C}_{mean}$. \mathbf{C}_{mean} is defined as follows:

$$\mathbf{C}_{mean} = n^{-1} \left[\mathbf{1}'_n \mathbf{U} \mathbf{C}_{raw}, \dots, \mathbf{1}'_n \mathbf{U} \mathbf{C}_{raw} \right]' \tag{16}$$

where $\mathbf{1}_n$ is $n \times 1$ unit vector.

Step3 : Perform the eigenvalue decomposition defined as $\mathbf{C}'_{cent} \mathbf{U}' \mathbf{U} \mathbf{C}_{cent} = \Gamma \Upsilon^2 \Gamma'$.

Then, the initial value of \mathbf{C} satisfying the constraint (3) is

$$\mathbf{C} = n^{\frac{1}{2}} \mathbf{C}_{cent} \Gamma \Upsilon^{-1} . \tag{17}$$

Iteration

For a given data matrix \mathbf{X} , loss function (5) can be minimized simply by iterating the updates of \mathbf{U} , \mathbf{C} , and \mathbf{A} . The whole of this iterative algorithm follows such steps :

Step1 : Initialize \mathbf{U} and \mathbf{C} by (17).

Step2 : Compute $\mathbf{B} = n^{-1}\mathbf{X}'\mathbf{U}\mathbf{C}$ and update \mathbf{A} with (15).

Step3 : Update \mathbf{U} with (6).

Step4 : Perform SVD in (9) and update \mathbf{C} with (11).

Step5 : Compute the function value for the present \mathbf{U} , \mathbf{C} , and \mathbf{A} . If the updates of \mathbf{U} , \mathbf{C} , and \mathbf{A} have decreased the function value, back to Step2. Otherwise, the process has converged.

This algorithm monotonically decreases the loss function value. Because of the binary constraint on \mathbf{U} , this procedure can be expected to be rather sensitive to local optima. Thus, we recommend to use many randomly started runs to decrease the chance of missing the global optima.

5 Perfect Simple Structure Loadings

The cardinality constraint (4) can be easily extended to the following row-wise constraint.

$$\text{Card}(\mathbf{a}_j) = m_j \quad (j = 1, \dots, p) \quad (18)$$

where \mathbf{a}_j ($j = 1, \dots, p$) represents the j th row vector of \mathbf{A} .

In this case, the equation (14) can be rewritten as follows;

$$\sum_{j=1}^p \|\mathbf{b}_j - \mathbf{a}_j\|^2 \quad \text{s.t.} \quad \text{Card}(\mathbf{a}_j) = m_j \quad (j = 1, \dots, p) \quad (19)$$

Then, we should replace the updating manner (15) by following process;

$$a_{jl} = \begin{cases} b_{jl} & \text{iff } |b_{jl}| \text{ is the } m_j\text{th largest absolute value in the } j\text{th row of } \mathbf{B} \\ 0 & \text{otherwise} \end{cases}, \quad (20)$$

Particularly, if we set $m_j = 1$ for all $j = 1, \dots, p$, the loading matrix \mathbf{A} is said to have perfect simple structure.

Example.2: The following loading matrix (6 variables \times 3 components) is one of the perfect simple structured matrix (where # means non-zero values).

$$\begin{bmatrix} \# & 0 & 0 \\ 0 & 0 & \# \\ 0 & \# & 0 \\ 0 & \# & 0 \\ \# & 0 & 0 \\ 0 & 0 & \# \end{bmatrix}$$

The perfect simple structured loadings is the simplest and easiest to interpret the relationships of the observed variables to principal components. Browne (2001) refers to this structure as "perfect cluster solution". From this point of view, the proposed method achieves the partitioning of the objects into K classes and the variables into q classes, simultaneously. From this point forward, we refer this case ($m_j = 1 : j = 1, \dots, p$) as $m = \text{PS}$ (Perfect Simple structure).

Table 1: The loading matrix \mathbf{A} obtained by the proposed method ($m = 5$, PS) and RKM.

	Proposed method ($m = 5$)		Proposed method (PS)		RKM	
	Component 1	Component 2	Component 1	Component 2	Component 1	Component 2
GDP	0	0.406	0	0.408	-0.116	0.787
LI	0	0	0.235	0	0.230	0.087
UR	0	0.530	0	0.530	0.163	0.410
IR	-0.718	0	-0.718	0	-0.653	0.102
TB	0.699	0	0.699	0	0.652	-0.049
NNS	0	-0.825	0	-0.824	-0.235	-0.438

6 Real Data Example

The short-term scenario (September 1999) on microeconomic performance of national economies of twenty countries, member of OECD, has been analyzed in Vichi & Kiers (2001) and Vichi & Saporta (2009). Vichi & Kiers (2001) have used this data to evaluate the ability of their proposed method in identifying groups of similar economic conditions and help to interpret the relationships among the observed variables. This data set has six variables: Gross Domestic Product (GDP), Leading Indicator (LI), Unemployment Rate (UR), Interest Rate (IR), Trade Balance (TB), and Net National Savings (NNS). These variables were standardized such that they have mean 0.00 and variance 1.00. For convenience, we used the first two principal components ($q = 2$). For comparison, we carried out the proposed method ($m = 5$, PS) and the RKM.

We should denote that each method is sensitive to local optima, therefore we run each algorithm from 500 different initial values. Then, the solution that most minimizes the objective function was defined as the optimal solution.

The loading matrices obtained by the proposed method ($m = 5$, PS) and the RKM are shown in Table.1. Due to some exactly zero loadings, our proposed model enjoyed quite interpretable loadings, whereas the many intermediate values occurred in the loadings of the RKM. Especially, in case of $m = 5$, the variable "LI" was eliminated from the extracted space. It implies that the proposed method has the role as the variable selection in some cases.

The plot of the countries in the first two components extracted by the proposed method ($m = 5$) is shown in Fig .1. Each class of variables is highlighted by a dotted axis. Obviously, the cardinality of \mathbf{A} and the information for clustering are related to the transactions. However, the proposed method shows homogeneous clusters (between cluster deviance of the proposed method is 74.67% of total deviance, while between cluster deviance of the tandem clustering is about 40% of the total deviance).

7 Conclusion

In this paper, we propose a new RKM procedure that aims at simultaneously grouping the objects and finding the interpretable low-dimensional space of variables. The proposed method in which the cardinality constraint is used provides the loading matrix including a number of exactly zero elements. Thus, it becomes easier to interpret the relationships between the variables and the principal components. Furthermore, the cardinality of a loading matrix can also be constrained

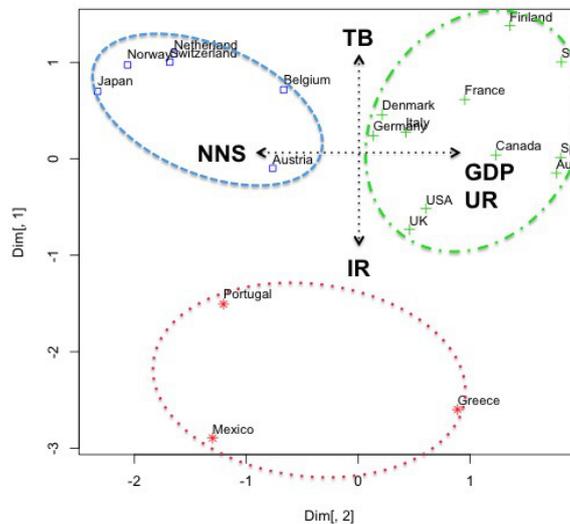


Figure 1: The clustering result of the proposed method ($m = 5$): Clusters of countries are highlighted by ellipse.

row-wise. In this case, the resulting loading matrix has perfect simple structure with only one non-zero element in each rows.

In the proposed method, how many loadings should be zero must be specified by users. A method for specifying it remains for future studies. One candidate may be to use information criteria.

Bibliography

- [1] Adachi, K., & Trendafilov, N. (2014). Penalty-free sparse PCA. *to appear in COMPSTAT 2014 proceedings*.
- [2] Browne, M.W. (2001). An overview of analytic rotation in exploratory factor analysis. *Multivariate Behavioral Research*, **36**, pp. 111-150.
- [3] De Soete, G., & Carroll, J.D. (1994). K-means clustering in a low-dimensional Euclidean space. In Diday, E., Lechevallier, Y., Schader, M., Bertrand, P., & Burtschy, B. (Eds.) *New Approaches in Classification and Data Analysis*, pp. 212-219. Heidelberg: Springer.
- [4] Vichi, M., & Kiers, H.A.L. (2001). Factorial k-means analysis for two-way data. *Computational Statistics and Data Analysis*, **37**, pp. 49-64.
- [5] Vichi, M., & Saporta, G. (2009). Clustering and disjoint principal component analysis. *Computational Statistics and Data Analysis*, **53**(8), pp. 3194-3208.

Spatial dependence monitoring over distributed data streams

Antonio Irpino, *Second University of Naples*, antonio.irpino@unina2.it

Antonio Balzanella, *Second University of Naples*, antonio.balzanella@unina2.it

Rosanna Verde, *Second University of Naples*, rosanna.verde@unina2.it

Abstract. This paper proposes a strategy for monitoring spatial dependence in multiple, spatially located, data streams. The interest on this topic is motivated by the number of real world applications in which data collected by sensor network depends on the geographic location of each sensing device. For instance, surface air temperatures streams are more likely to be similar when measured at nearby locations rather than if they are detected in distant places. The strategy we propose for addressing this challenge is based on distributed processing. At each sensor, it is performed a summarization of the data by means of a micro-clustering strategy for histogram data. At the central processing node, it is measured the spatial dependence and it is evaluated its evolution over time introducing a new tool: the variogram for histogram data.

Keywords. Data Stream Mining, Histogram Data, Variogram

1 Introduction

Massive datasets having the form of continuous streams with no fixed length are becoming very common due to the availability of sensor networks which can perform, at a very high frequency, repeated measurements of some variable.

The knowledge extraction from such data must consider the technological characteristics of the tools for data acquisition as well as the nature of the monitored phenomenon. Often, data acquisition is performed by sensors having some limited storage and processing resources. Moreover, the communication among sensors is constrained by the physical distribution or by limited bandwidths. Finally, the recorded data often concerns highly evolving phenomena requiring algorithms able to adapt the knowledge as new observations arrive.

The prevailing paradigm for the analysis of data in this context is the centralized data stream analysis. Observations, recorded by sensors, are organized and processed by a single unit which provides the results of queries. In this case, the single processing unit should guarantee space and time efficiency so that data has to be processed on the fly, at the speed in which it is recorded,

and algorithms need to adapt their behavior over time, consistently with the dynamic nature of data.

More recently, the literature is moving toward distributed stream processing in which each sensor can perform a local processing of data, within its resource constraints, and communicate efficiently with other processing units. In this case, in addition to space and time efficiency, algorithms should account for the communication load imposed by the network infrastructure. Usually, this is carried out limiting the communication between processing units to the simple transmission and reception of synthetic synopsis of data.

In the framework of distributed stream processing, this paper deals with the task of monitoring the evolution of spatial dependence among data streams.

The interest in this topic is motivated by the number of real world applications in which data, collected by a sensor network, depends on the geographic location of each sensing device. The First Law of Geography, also frequently known as simply Toblers Law[8], states: “*Everything is related to everything else, but near things are more related than distant things*”. This law, which finds its major developments in Geostatistics, is still valid in the framework of data stream mining, when data is collected by spatially located sensors. For instance, surface air temperatures streams, are more likely to be similar when measured at nearby locations rather than if they are detected in distant places.

Due to the high evolving nature of streaming data and to their potentially unbounded size, the spatial dependence can evolve itself over time, thus, the main challenges are to measure, on-line, the spatial dependence and to keep track of its evolution.

To our knowledge, these challenges have not been dealt in the data stream mining literature. Methods for analyzing spatial data streams focus, mainly, on traditional data mining tasks (clustering, classification, summarization) on streams recording the position of moving objects [5][7]. Our challenge is, instead, the analysis of data produced by sensors having a fixed, known, spatial location.

The strategy we propose has the following features: 1) It is based on a new tool for measuring the spatial dependence, named Variogram for Histogram data, which extends the classical Variogram to histogram data; 2) It allows to update, on-line, the Variogram for Histogram data; 3) The statistics needed for the computation and updating of the Variogram for Histogram data can be computed on distributed processing units, using a low network load.

The paper is organized as follows: Section 2 introduces the data definition and the processing scheme; Section 3 describes our strategy for summarizing the parallel arriving data streams through histograms; Section 4 provides the details of our proposal for spatial dependence monitoring; Section 5 provides some experimental results.

2 Data definition and processing setup

Let $Y_i = \{(y_i^1, t_1), \dots, (y_i^j, t_j), \dots\}$ be a data stream made by real valued observations y_i^j on a discrete time grid $T = \{t_1, \dots, t_j, \dots\}$, with $t_j \in \mathcal{R}$ and $t_j > t_{j-1}$. The data stream Y_i is made by observations recorded by a sensor located at $s_i \in S$, with $S \subseteq \mathcal{R}^2$ be the geographic space.

Our aim is to measure and monitor the spatial dependence in a set of n data streams $Y = \{Y_1, \dots, Y_i, \dots, Y_n\}$ assuming that the time grid T is common to all the data streams.

To reach this aim, the data recorded by each sensor over time is split into non overlapping windows whose identifier is $w = 1, \dots, \infty$. A window, which is an ordered subset of T having

size b , frames a subsequence $Y_i^w = \{(y_i^j, t_j), \dots, (y_i^{j+b}, t_{j+b})\}$ for each Y_i .

A subsequence Y_i^w is summarized by a histogram $H_i^w = \{(I_{i,l}^w, \pi_{i,l}^w), \dots, (I_{i,L}^w, \pi_{i,L}^w)\}$ made by L weighted intervals (bins). Each histogram partitions the support $D_i^w = [\underline{y}_i; \bar{y}_i]$ of a subsequence Y_i^w into a set of non overlapping intervals, or bins, so that:

$$D_i^w = \{I_{i,1}^w, \dots, I_{i,l}^w, \dots, I_{i,L}^w\}, \text{ where } I_{i,l}^w = [\underline{y}_{i,l}, \bar{y}_{i,l}) \text{ (for } l = 1, \dots, L)$$

The use of histograms as tool for describing the empirical distribution of the data in a window is motivated by their capability of collecting information such as the location of data, the variability, the symmetry, the curtosis, etc. Histograms are fast to compute and there are efficient methods for computing the distance between them.

In particular, in [4] the authors develop a ℓ_2 version of the Wasserstein distance for histogram data. The ℓ_2 Wasserstein distance introduced in [6] can be interpreted as the Euclidean distance between quantile functions:

Let F and G the distribution functions of two random variables Y_1 and Y_2 and F^{-1} and G^{-1} the corresponding quantile functions. The ℓ_2 Wasserstein can be expressed as follows:

$$d_W(Y_1, Y_2) := \sqrt{\int_0^1 (F^{-1}(t) - G^{-1}(t))^2 dt}$$

Its main issue is related to the inverse of the distribution functions which is impossible to do analytically for most distributions. The authors in [4][9], address this problem, by introducing an exact and efficient way to compute this distance when data are histograms.

Given a histogram description H_i^w , the quantities $\tau_{i,l}^w$ can be defined in order to represent the cumulative weights associated with the elementary intervals of H_i^w :

$$\tau_{i,l}^w = \begin{cases} 0 & l = 0 \\ \sum_{h=1, \dots, l} \pi_{i,h}^w & l = 1, \dots, L \end{cases} \quad (1)$$

Using (1), and assuming a uniform density for each $I_{i,l}^w$, the inverse distribution function is a piecewise function defined as follows:

$$F_i^{w-1}(t) = \underline{y}_{i,l} + \frac{t - \tau_{i,l-1}^w}{\tau_{i,l}^w - \tau_{i,l-1}^w} (\bar{y}_{i,l} - \underline{y}_{i,l}) \quad \tau_{i,l-1}^w \leq t < \tau_{i,l}^w.$$

To compute the distance between two histogram descriptions H_i^w and H_j^w it is needed to identify a set of common uniformly dense intervals. Let τ^w be the set of the cumulated weights of the two distributions $\tau^w = [\tau_0^w, \dots, \tau_l^w, \dots, \tau_q^w]$, where: $\tau_0^w = 0$ $\tau_q^w = 1$ and $\pi_l^w = \tau_l^w - \tau_{l-1}^w$. To solve the problem of finding a common set τ^w of cumulated weights associated with the quantiles of the two distributions, we consider equi-depth histograms. In this case, histograms H_i^w and H_j^w , involved in the distance computation, are characterized by the same set of weights $\pi_{i,l}^w = \pi_{j,l}^w = \frac{1}{L}$ and $q = L$.

For each interval (bin) of the histogram, it is possible to compute the centers and radii, as follows:

$$c_{i,l} = (F_i^{w-1}(\tau_l^w) + F_i^{w-1}(\tau_{l-1}^w))/2 \quad r_{i,l} = (F_i^{w-1}(\tau_l^w) - F_i^{w-1}(\tau_{l-1}^w))/2.$$

Because intervals are uniformly distributed, it is possible to express them as a function of their centers and radii, and to rewrite the distance as follows:

$$d_W^2(H_i^w, H_j^w) = \sum_{l=1}^q \pi_l^w \left[(c_{i,l} - c_{j,l})^2 + \frac{1}{3} (r_{i,l} - r_{j,l})^2 \right]. \quad (2)$$

The analysis of the incoming data is performed keeping a part of the processing at the single sensors and a part at a centralized computation node. The former, performs an analysis, window by window, of a single data stream in order to provide a summarization of the data. A set of outcomes of the processing at the sensor, is sent to the centralized node for the computation of the spatial dependence.

In the next section we provide the details of the processing at a sensor.

3 On-line summarization of a data stream through CluStream for Histogram data

The analysis of the data recorded by a sensor is performed through the CluStream algorithm for Histogram data [1]. Every time a new window of data becomes available and the corresponding histogram is built, the CluStream updates a summarization of the stream consisting in a set of histograms.

CluStream for Histogram data provides the summarization of the stream Y_i by keeping updated a set μC_i of synopsis data structures named micro-clusters. Basically, a micro-cluster μC_i^k , with $k = 1, \dots, K$, records a Histogram centroid \overline{H}_i^k and the number of allocated histograms n_i^k .

The incoming histogram H_i^w is allocated to the micro-cluster μC_i^k such that $d_W^2(H_i^w, \overline{H}_i^k) < d_W^2(H_i^w, \overline{H}_i^{k'})$ (with $k \neq k'$ and $k = 1, \dots, K$), if $d_W^2(H_i^w, \overline{H}_i^k) < u$. The threshold value u controls the size of each micro-cluster ensuring the representativity of the centroid.

The allocation of a histogram to a micro-cluster causes the need to update n_i^k by $n_i^k = n_i^k + 1$ and the micro-cluster centroid \overline{H}_i^k . The latter can be performed keeping into account a further implication of the Wasserstein distance for histogram data: the average histogram is the histogram having as center of each bin, the average of the centers and as radii, the average of the radii. Thus, the updating can be performed by using the information stored in the micro-cluster and the new allocated histogram.

If the condition $d_W^2(H_i^w, \overline{H}_i^k) < u$ is not satisfied for any micro-cluster, a new one is started setting the allocated histogram as centroid and $n_i^k = 1$. This procedure, can bring the number of micro-clusters K to grow so much to exceed the available memory resources at the computation node. We propose, in this case, to merge the two nearest micro-clusters into one.

The proposed procedure, performed in a parallel way on all the streams, permits to keep, at each time instant, a snapshot of the data behavior. This is due to availability of the set of histograms used as representatives.

As we will see in the next section, the evaluation of the spatial dependence is performed keeping into account the behavior of the data of each stream. Thus, we need to send a set of statistics to the central computation node which contemporary guarantee a low network load and a high quality of the result. To reach this aim, the communication between the processing node at the sensor and the centralized processing unit is made by two tasks. The first task, which

is performed at predefined time stamps (for example, every 20 windows), consists in sending a snapshot of the micro-cluster centroids to the central computation node. The second task, which is performed at every time window, consists in sending the identifier of the micro-cluster to which the histogram of the window has been allocated.

The idea is to summarize the data in a window with the centroid to which it has been allocated so that, if an updated set of centroids is kept by the central node (task 1), it is sufficient to send the identifier of the micro-cluster rather than its data (task 2).

4 Spatial dependence monitoring

In this section, we introduce the analysis performed at the central node for measuring and monitoring the spatial dependence. With this aim we introduce a new tool named variogram for histogram data, which extends the classical variogram to histogram data.

The variogram function

In the traditional geostatistics literature, a widely used tool for evaluating the spatial dependence is the variogram. Given a random process Y , it is defined as the variance of the difference between process values at two locations, across realizations of the process [3]. If the process is stationary and isotropic, the variogram γ can be represented as a function of the distance $h = \|s_i - s_j\|$ between spatial locations. An unbiased estimator of the variogram, for a set of observations y_i , $i = 1, \dots, k$ located at s_i is the empirical variogram whose formal expression is the following:

$$\hat{\gamma}(h) = \frac{1}{|N(h)|} \sum_{i,j \in N(h)} (y_i - y_j)^2 \quad (3)$$

where $N(h)$ is the set of observations such that $\|s_i - s_j\| = h$ and $|N(h)|$ is the number of pairs in the set.

In spatially dependent data, the value of the variogram function increases with the lag distance h until a limit is reached. This limit permits to identify the level beyond which the variability is no longer dependent on the spatial distance.

The variogram estimator cannot be computed at every lag distance h , due to the reduced availability of observations. This involves that the empirical variogram is not ensured to be valid so that in applied geostatistics, the empirical variograms are often approximated by model function ensuring validity [2].

The empirical variogram introduced above, can be still used as a pure exploratory tool for investigating spatial dependence in the data. In this sense, there is a distinction between the variogram model, estimated using the empirical variogram and than some fitting model function, and the experimental variogram which corresponds to the empirical variogram computed on the data without making formal assumptions on the process generating the data.

We are interested in this second use of the variogram function, in order to monitor the evolution of the spatial dependence among the data streams.

Since the input data streams are represented as sequences of histograms, the next section will introduce the variogram for histogram data.

Experimental variogram for histogram data

According to the processing setup introduced above, at the central computation node it is kept a snapshot of micro-cluster centroids of each stream. Every time a new window becomes available, it is possible to measure the spatial dependence by receiving, for each data stream, the identifier of the micro-cluster to which the histogram of the window has been allocated. This approach, allows to measure the spatial dependence of the data in a window using the micro-cluster centroids rather than the raw sensor data.

Starting from this assumption, we define the experimental variogram for histogram data consistently with the ℓ_2 Wasserstein metric introduced above and with its expression for histogram data in 2:

$$\gamma_H(h) = \frac{1}{|N(h)|} \sum_{i,j \in N(h)} (d_W^2(\overline{H}_i^k, \overline{H}_j^{k'})) \quad (4)$$

where:

$N(h)$ is the set of observations such that $\|s_i - s_j\| = h$

$|N(h)|$ is the number of pairs in the set $N(h)$.

\overline{H}_i^k and $\overline{H}_j^{k'}$ are the micro-cluster centroids to which, respectively, the histograms H_i^w and H_j^w of the stream Y_i^w , Y_j^w have been allocated.

The variogram $\gamma_H(h)$ is nonnegative since it is the average of squared distances and shares the characteristics of the experimental variogram for traditional scalar data.

For irregularly spaced data where there are not enough observations exactly separated by h , $N(h)$ is modified to $(s_i, s_j) : \|s_i - s_j\| \in (h - \epsilon, h + \epsilon)$, with $\epsilon > 0$. This involves that $\gamma_H(0) \geq 0$ (nugget effect).

According to the provided definition of variogram for histogram data, we can measure the spatial dependence between distributions (histograms) into a time window. If there is spatial dependence, we expect that near sensors tend to have lower values of average distances while far sensors tend to be more different so that $\gamma_H(h)$ is an increasing function.

Our strategy is to update the variogram every time a new set of identifiers is received at the central computation node. We still record a snapshot of the computed and updated variogram, at predefined time stamps, in order to keep track of the data behavior over a time.

The variogram is obtained by computing the average of pairwise distances at each lag distance h , thus, to obtain a variogram $\gamma_H^w(h)$ for the window w starting from the variogram $\gamma_H^{w-1}(h)$ at the window $w - 1$, we have to update such average values:

$$\gamma_H^w(h) = \frac{1}{2|N^w(h)|} \left(\gamma_H^{w-1}(h) |N(h)| + \sum_{i,j \in N(h)} (d_W^2(\overline{H}_i^k, \overline{H}_j^{k'})) \right) \quad (5)$$

where $N^w(h) = N^{w-1}(h)$ is the set of observations such that $\|s_i - s_j\| = h$ and $|N^w(h)|$ is the number of pairs in the set.

It should be noted that the set of pairs i, j involved in the average computation depends only on the spatial location of the sensors which record the data. Such locations are fixed and known a priori so that at each window, always the same pairs participate to the computation of the variogram value. Moreover, due to our processing setup, the histogram centroids are available at the central processing node. This involves that the computation of the variogram for histogram data can be carried out faster if the distances between micro-clusters are stored

into a lookup table. Such lookup table is only updated when the sensor send new histogram centroids. This occurs only at predefined time instants, as stated above.

In order to evaluate the evolutions in spatial dependence, we still propose a measure which is computed on the variograms of two time periods:

$$EVO = \sqrt{\sum_h (\gamma_H^w(h) - \gamma_H^{w-k}(h))^2} \quad (6)$$

The computation of the measure EVO is feasible since the data streams have a fixed spatial location so that the compared experimental variograms are computed for the same values of the lag h

5 Experimental results on real data

We have made some preliminary test for evaluating the performance of the proposed strategy in keeping track of the spatial dependence among sensor data using a public dataset of real data, available at <http://db.csail.mit.edu/labdata/labdata.html>.

The dataset collects the records of 54 sensors placed at the Intel Berkeley Research lab between February 28th and April 5th, 2004. Mica2Dot sensors with weather boards collected timestamped topology information, along with humidity, temperature, light and voltage values once every 31 seconds. Data was collected using the TinyDB in-network query processing system, built on the TinyOS platform. The dataset includes the x and y coordinates of sensors (in meters relative to the upper right corner of the lab).

We have analyzed the temperature records of each sensor so that we have a set of 54 time series each one made by 65000 observations.

In order to run the test, we have set the size of each window to $s = 200$ and the number of bins for each histogram to $L = 10$. In fig.1a and in fig.1b we have plotted the variogram for histogram data obtained by processing, respectively, the first 32500 time stamps and the latest 32500 time stamps. Looking at the plots, we can note a change in the spatial dependence between the first batch of data and the second one, however both highlight the presence of a spatial dependence since the two plots tend to increase with the lag distance h .

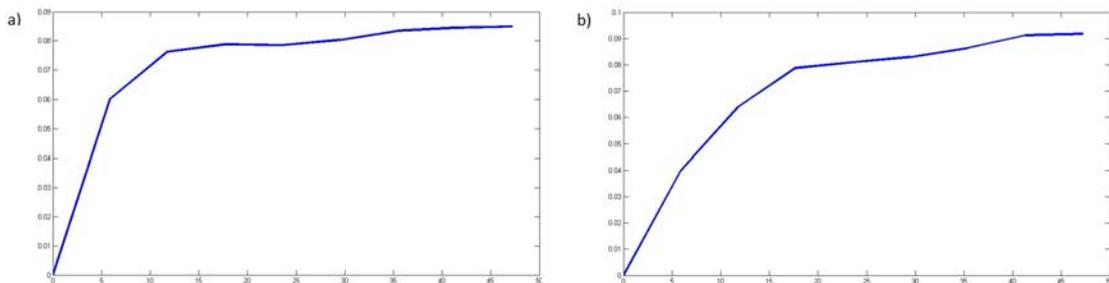


Figure 1: Variogram function for histogram data computed on the first 32500 time stamps (a) and on the latest 32500 time stamps (b)

6 Conclusions

In this paper we have introduced a strategy for monitoring the spatial dependence of data recorded by spatially located sensors. To reach this aim, we have proposed to summarize the incoming data streams through sets of histograms, then, we have introduced the variogram for histogram data for measuring the spatial dependence. The proposed processing setup allows to make a part of the computation at distributed nodes using a low communication load. Preliminary results confirm the effectiveness of the method, however, in future works, we will explore the possibility to reduce the computational effort required for keeping an updated variogram over time.

Bibliography

- [1] Balzanella A., Rivoli L., Verde R. (2013) *Data stream summarization by histograms clustering*. In: Statistical Models for Data Analysis. Eds: Giudici; Ingrassia; Vichi, Springer. ISBN:978-3-319-00031-2 (2013)
- [2] Chiles, J. P., Delfiner P. (1999) *Geostatistics, Modelling Spatial Uncertainty*. Wiley-Interscience
- [3] Cressie, N. (1993) *Statistics for spatial data*. Wiley Interscience
- [4] Irpino, A., Verde, R., Lechevallier, Y. (2006) *Dynamic clustering of histograms using Wasserstein metric* In: COMPSTAT, pp 869-876
- [5] Ling-Yin Wei, Wen-Chih Peng. (2013) *An incremental algorithm for clustering spatial data streams: exploring temporal locality*. In: Knowledge and Information Systems Volume 37, Issue 2 , pp 453–483
- [6] Mallows, C. L. (1972) *A Note on Asymptotic Joint Normality* In: The Annals of Mathematical Statistics, **43** (2), pp 508-515.
- [7] Qiang Ding, Qin Ding, Perrizo W.(2002) *Decision tree classification of spatial data streams using Peano Count Trees*. In: Proceedings of the 2002 ACM symposium on Applied computing, March 11-14, Madrid, Spain
- [8] Tobler W., (1970) *A computer movie simulating urban growth in the Detroit region*. Economic Geography, 46(2): 234-240
- [9] Verde, R., Irpino, A. (2007) *Dynamic Clustering of Histogram Data: Using the Right Metric*. In: P. Brito, G. Cucumel, P. Bertrand, F. Carvalho (Eds.), Selected Contributions in Data Analysis and Classification, Studies in Classification, Data Analysis, and Knowledge Organization, chap. 12, Springer Berlin Heidelberg, Berlin, Heidelberg, 123-134
- [10] Wasserstein, L. (1969) *Markov processes over denumerable products of spaces describing large systems of automata*. In: Prob. Inf. Transmission, **5**, 47-52

Imputation of complex dependent data by conditional copulas: analytic versus semiparametric approach

F. Marta L. Di Lascio, *Free University of Bozen-Bolzano, Italy*, marta.dilascio@unibz.it
Simone Giannerini, *University of Bologna, Italy*, simone.giannerini@unibo.it
Alessandra Reale, *ISTAT, Italian Statistical Institute, Italy*, reale@istat.it

Abstract. Missing data occur in almost all the surveys and may create serious problems because restricting the analysis to complete cases leads to loss of precision and invalid inferences. Hence, missing data are commonly treated by imputation, that is, they are filled in with plausible values. In a previous work we proposed a copula-based method that allows to impute by accounting for both the (complex) dependence structure underlying the data and the shape of the margins. The method employs the conditional density functions of the missing variables given the observed ones. These functions are derived analytically once parametric models for the margins and the copula are specified. In this paper, we extend our method in a semiparametric fashion in that the margins are estimated non-parametrically through local likelihood methods. We compare the performance of the two versions of the imputation method in terms of the preservation of both the dependence structure and the microdata in different simulated scenarios by varying copula, marginal distributions and the level of the dependence parameter. The method has a wide range of applicability and has been implemented in the R software package `CoImp`.

Keywords. Conditional copula function, Dependence structure, Imputation method, Missing data.

1 Introduction

The problem of missing data arises in many applied fields. For example, in large surveys, a unit that does not respond either to a particular question or to the entire survey creates a missing item or a missing record, respectively. From the point of view of statistical analysis this is a serious issue because incomplete data sets are generally challenging. If there are only few missing values, it is possible to restrict the analysis to the complete cases but, in many applications deleting incomplete cases leads to a reduction of the sample size so that proper

inferences are precluded. Hence, it is customary to resort to imputation methods that fill in the missing as to create a complete data set.

Imputation of missing data is one of the major tasks of data analysis in many areas and different techniques have been developed to tackle the problem (see [9]). Here, we focus on stochastic single imputation, that is, the imputation is performed by filling in a random value for each missing data until a complete data set is obtained. In brief, an appropriate stochastic model is fitted to the available data and imputed values are simulated from it. In this context, one of the main goals is to perform imputation by preserving the dependence structure of the data.

Di Lascio et al. [3], [4] proposed an imputation method based on copula function [10]. The theory of copula [2] provides a flexible and powerful approach to construct multivariate distributions by separating marginal distributions from the joint dependence structure of the data generating process, which is described by the copula. Hence, the margins can have any kind of distribution and the copula represents the multivariate dependence between variables and is able to account for many different dependence structures, such as asymmetry, heavy-tail and so on. The method proposed in [4] employs the conditional density functions of the missing variables given the observed ones to impute complex dependent data. The conditional density functions are defined through the joint copula model so that the imputation takes into account the multivariate structure of the data generating process. An alternative approach available in literature is the so-called fully conditional specification [11], [12] that builds the multivariate model by a series of conditional models, one for each incomplete variable. Such a method is flexible but its statistical properties are difficult to establish and the implied joint distribution may not exist theoretically due to the incompatibility of conditionals [1]. This problem is overcome by Di Lascio et al. [4] who derive the conditional densities starting from a multivariate model defined via copula. Their approach is fully-parametric since it specifies a (possibly different) parametric model for each univariate margins and for the copula; then, the conditional densities are derived analytically and estimation is carried out through the two-step maximum likelihood due to Joe and Xu [6]. In this paper we extend the method in Di Lascio et al. [4] so that the margins are modelled non-parametrically by means of local likelihood estimators. The extension allows us to avoid the analytical problems that might arise in the derivation of the conditional densities for specific combinations of copula and margins. The resulting method is powerful, flexible and easy to use also when the missing data are high-dimensional and the dependence is complex. We compare the performance of the novel semiparametric copula-based imputation method with the fully parametric version in a simulation study.

The paper is organized as follows. In Section 2 we present the rationale of the semiparametric imputation method based on copula function. In Section 3 we describe the simulation study and discuss the results, whereas Section 4 contains a brief discussion.

2 Copula-based Imputation method

The basic idea of the method proposed by Di Lascio et al. [4] is to derive the conditional density functions of the missing variables given the observed ones through the corresponding conditional copulas. Once the conditional densities are available the missing values are imputed by drawing observations from them. This idea is motivated from practical problems that arise frequently: *i*) modeling multivariate distributions with different margins and complex dependence structures

and *ii*) imputing missing values by preserving the dependence underlying the data. Our proposal allows to accomplish both tasks; moreover, it can be easily used independently from the dimension and the kind (monotone or non monotone) of the missing patterns. We describe the proposal by focusing on bivariate distributions defined via the three copula models belonging to the Archimedean family: the Clayton, the Gumbel and the Frank copula. For an extended review of copula models see [8].

Suppose we have two continuous random variables X_1 and X_2 with distribution functions F_1, F_2 and densities f_1 and f_2 , such that their probability integral transforms are $U_1 \sim F_1(X_1)$ and $U_2 \sim F_2(X_2)$, respectively. Further, assume that for some records X_1 is missing whereas X_2 is always observed. We derive the conditional density function $f(x_1|x_2)$ through *i*) the canonical representation of the joint density via the density copula $c(\cdot)$ (see also [2]):

$$f(x_1, x_2) = c(F_1(x_1), F_2(x_2)) \prod_{j=1}^2 f_j(x_j) \quad (1)$$

and *ii*) the conditional copula density $c(u_1|u_2)$ defined by using Bayes' rule (see [13], p.89) so that:

$$f(x_1|x_2) = c(u_1|u_2)f_1(x_1). \quad (2)$$

In order to impute missing observations, we perform the following steps:

1. estimate the margins on the available data through the local log-likelihood function (see [7])

$$l(f_j, x) = \sum_{i=1}^n W\left(\frac{X_i - x}{h}\right) \log f_j(X_i) - n \int_{\mathcal{X}_j} W\left(\frac{z - x}{h}\right) \exp(f_j(z)) dz \quad (3)$$

where W is a suitable nonnegative weight function, h the bandwidth parameter, \mathcal{X}_j is the domain of the variable X_j with $j = 1, 2$ and $\log f_j(z)$ with $j = 1, 2$ is modeled by a local polynomial (for technical details see [5] and [7]);

2. estimate the dependence parameter θ of the copula model c on the available data by the pseudo-maximum log-likelihood method

$$\hat{\theta} = \operatorname{argmax}_{\theta} \sum_{i=1}^n \log c(\hat{U}_{i1}, \hat{U}_{i2}; \theta) \quad (4)$$

where $\hat{U}_{ij} = R_{ij}/(n+1)$ with $j = 1, 2$ and R_{ij} being the rank of X_{ij} among X_{1j}, \dots, X_{nj} are the pseudo-observations of the two margins;

3. derive the conditional density functions in eq. (2);
4. impute missing observations by drawing observations from the conditional densities computed at the previous step by means of the Hit or Miss Monte Carlo method as described in [4].

Note that the method does not depend on the kind of missing pattern. Moreover, the scheme in its multivariate version has been implemented in the R package `CoImp`.

3 A simulation study

In this section we compare the performance of the semiparametric copula-based imputation method with the fully parametric approach introduced in Di Lascio et al. [4]. In order to perform the comparison we reproduce their scenarios as summarized in Table 1.

Table 1: The 3 scenarios of the simulation study. Notes: (δ, γ) are the shape and rate parameters of a Gamma distribution; ν is the degrees of freedom parameter of a Chi-square distribution; λ is the rate parameter of an Exponential distribution.

Scenario	Copula	Margin X_1	Margin X_2
1	Clayton	Gamma: $X_1 \sim G(\delta = 3, \gamma = 3)$	Chi-Square: $X_2 \sim \chi^2(\nu = 4)$
2	Gumbel	Uniform: $X_1 \sim U(0, 1)$	Exponential: $X_2 \sim \text{Exp}(\lambda = 1/3)$
3	Frank	Chi-Square: $X_1 \sim \chi^2(\nu = 2)$	Chi-Square: $X_2 \sim \chi^2(\nu = 4)$

Figure 1 shows the contour plots of the three bivariate distributions described in Table 1 for the two levels of the dependence parameter θ corresponding to the Kendall's correlation coefficient $\tau = 0.35$ (upper panel) and $\tau = 0.7$ (lower panel). Note the different scale of the central upper plot due to the asymmetry. The Monte Carlo study consists of the following steps:

- i) generate a sample of $n = 200$ observations from bivariate random vectors (X_1, X_2) as in Table 1 by varying the dependence parameter θ such that Kendall's correlation results $\tau = 0.35$ and $\tau = 0.70$;
- ii) introduce 40% of (artificial) missing values completely at random into the original data set;
- iii) apply the semiparametric version of the copula-based imputation method to the raw data set as to obtain the imputed observations x_m^{im} with $m = 1, \dots, M$, where M is the total number of missing values;
- iv) repeat steps ii) and iii) $K = 50$ times in order to take into account the source of variability deriving from the randomness of the mechanism generating the missing data;
- v) assess the goodness of the imputation methods in terms of the preservation of both the microdata and the strength of the dependence by using
 1. the *mean absolute relative error* (MARE) between imputed (x_m^{im}) and original (x_m^{ob}) values:

$$\text{MARE} = \frac{1}{K} \sum_{k=1}^K \left[\frac{1}{M} \sum_{m=1}^M \left| \frac{x_m^{\text{im}} - x_m^{\text{ob}}}{x_m^{\text{ob}}} \right| \right] \quad (5)$$

2. the *relative bias* (RB) and the *relative root mean squared error* (RRMSE) of the dependence parameter θ of the copula:

$$\text{RB} = \frac{1}{K} \sum_{k=1}^K \left(\frac{\hat{\theta}_k - \theta_0}{\theta_0} \right); \quad \text{RRMSE} = \sqrt{\frac{1}{K} \sum_{k=1}^K \left(\frac{\hat{\theta}_k - \theta_0}{\theta_0} \right)^2} \quad (6)$$

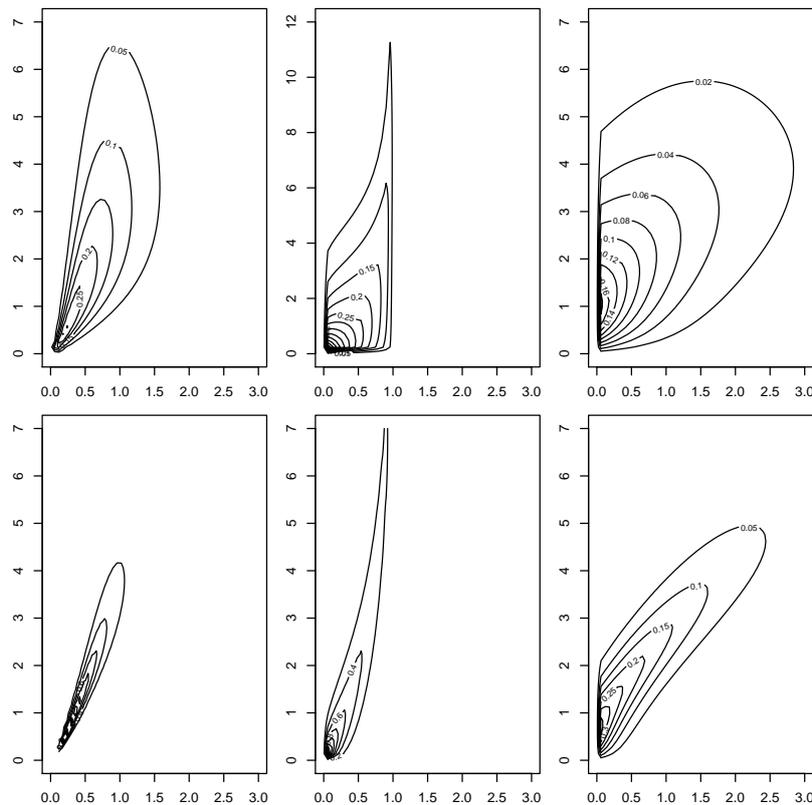


Figure 1: Contour plots of the three bivariate distributions described in Table 1 and defined via Clayton copula (left panel), Gumbel copula (middle panel) and Frank copula (right panel). The dependence parameter θ is such that Kendall's correlation coefficient is $\tau = 0.35$ (mild dependence, upper panel) and $\tau = 0.7$ (high dependence, lower panel).

where θ_0 is the true value of the dependence parameter and $\hat{\theta}_k$ is the estimated θ for the k -th simulated sample.

The results are shown in Table 2. The copula-based imputation method in its semiparametric version shows a performance very similar to the analytical version. Slight differences between the two versions of the method emerge for the MARE. The biggest discrepancy occurs when the data generating process is defined through a Gumbel copula with a high level of dependence ($\tau = 0.7$). Apart from such a case which would require more detailed investigations, the relative bias and the relative root mean squared error have a similar value for the two versions of the method in all the scenarios. By comparing the above results with those in [4] it is clear that the performance of the new semiparametric version overcomes that of classical imputation methods (the donor and the EM). We can conclude with confidence that the slight performance loss that occurs when passing from the parametric to the semiparametric version of the copula-based imputation method is by far compensated by the gain in flexibility and practical applicability.

Table 2: Comparison between the copula-based imputation method in its analytical version and in its semiparametric version.

Copula Model	Performance Measures	Kendall's τ	Analytical Method	Semiparametric Method
Clayton	MARE	0.35	1.09	1.09
		0.7	1.28	1.31
	RB	0.35	-0.44	-0.43
		0.7	-0.71	-0.72
	RRMSE	0.35	0.45	0.45
		0.7	0.71	0.72
Gumbel	MARE	0.35	2.50	2.56
		0.7	0.78	1.58
	RB	0.35	0.045	-0.103
		0.7	0.061	-0.382
	RRMSE	0.35	0.08	0.112
		0.7	0.092	0.386
Frank	MARE	0.35	6.03	6.66
		0.7	4.03	5.05
	RB	0.35	0.017	-0.018
		0.7	-0.012	0.001
	RRMSE	0.35	0.129	0.157
		0.7	0.072	0.072

4 Discussion

In this paper we have proposed a semiparametric imputation method based on copula function that extends the framework in Di Lascio et al. [4]. Being based on copula functions, the method allows to model complex multivariate distributions by separating the influence of the margins from that of the multivariate dependence. For these reasons the imputation of missing data preserves the joint structure of the data generating process. Moreover, by employing local polynomials to fit the marginal distributions, it avoids any assumptions on the margins and overcomes the difficulties arising from the analytical derivations of the conditional densities.

The alternative approach based on the fully conditional specification due to [11] is semiparametric and flexible too but, since it specifies a conditional model for each incomplete variable there can be theoretical compatibility problems and the stationary distribution of the Gibbs sampler may not exist. On the contrary, the theoretical framework of copula functions coupled with non-parametric density estimation allows to derive conditional distributions of potentially any complex multivariate model without any incompatibility of conditionals [1]. The Monte Carlo study shows that the semiparametric version is very similar to the parametric version in terms of goodness of the imputed data as well as of preservation of the dependence level.

Acknowledgement

The first author acknowledges the support of Free University of Bozen-Bolzano, School of Economics and Management via the projects “Multivariate analysis techniques based on copula function” and MODEX.

Bibliography

- [1] Arnold, B.C., Castillo, E. and Sarabia, J.M. (1999) Conditional specification of statistical models. Springer.
- [2] Cherubini, U., Luciano, E. and Vecchiato, W. (2004). Copula methods in finance. John Wiley & Sons Inc., Chichester, West Sussex.
- [3] Bianchi, G., Di Lascio, F.M.L., Giannerini, S., Manzari, A., Reale, A. and Ruocco, G. (2009). Exploring copulas for the imputation of missing nonlinearly dependent data. *Proceedings of the VII Meeting del Classification and Data Analysis Group of the Italian Statistical Society (Cladag)*, Cleup, p. 429–432.
- [4] Di Lascio, F.M.L., Giannerini, S. and Reale, A. (2014). Exploring Copulas for the Imputation of Complex Dependent Data. *Submitted revision*.
- [5] Di Lascio, F.M.L. and Giannerini, S. (2014). A multivariate technique based on conditional copula specification for the imputation of complex dependent data. *Working paper*.
- [6] Joe, H. and Xu, J. (1996). The estimation method of inference functions for margins for multivariate models. *Technical Report 166, Department of Statistics, University of British Columbia*.
- [7] Loader, C.R. (1996). Local likelihood density estimation. *The Annals of Statistics*, **24(4)**, 1602–1618.
- [8] Nelsen, R.B. (2006). Introduction to copulas. Springer, New York.
- [9] Schafer, J.L. (1997). Analysis of Incomplete Multivariate Data. Chapman & Hall, London.
- [10] Sklar, A. (1959). Fonctions de répartition à n dimensions et leurs marges, *Publications de l’Institut de Statistique de L’Université de Paris*, **8**, p. 229–231.
- [11] van Buuren, S., Brand, J.P.L., Groothuis-Oudshoorn, C.G.M. and Rubin, D.B. (2006). Fully conditional specification in multivariate imputation, *Journal of Statistical Computation and Simulation*, **76(12)**, p. 1049-1064.
- [12] van Buuren, S. (2007). Multiple imputation of discrete and continuous data by fully conditional specification, *Statistical Methods in Medical Research*, **16**, p. 219-242.
- [13] Zimmer, D.M. and Trivedi, P.K. (2006). Using trivariate copulas to model sample selection and treatment effects: Application to family health care demand, *Journal of Business & Economic Statistics*, **24**, p. 63–76.

Cyclic coordinate for penalized Gaussian graphical models with symmetry restrictions

Antonino Abbruzzo, *University of Palermo*, antonino.abbruzzo@unipa.it

Luigi Augugliaro, *University of Palermo*, luigi.augugliaro@unipa.it

Angelo M. Mineo, *University of Palermo*, angelo.mineo@unipa.it

Ernst C. Wit, *University of Groningen*, e.c.wit@rug.nl

Abstract. In this paper we propose two efficient cyclic coordinate algorithms to estimate structured concentration matrix in penalized Gaussian graphical models. Symmetry restrictions on the concentration matrix are particularly useful to reduce the number of parameters to be estimated and to create specific structured graphs. The penalized Gaussian graphical models are suitable for high-dimensional data.

Keywords. Factorial dynamic Gaussian graphical models, Gaussian graphical models, graphical lasso, cyclic coordinate descent methods.

1 Introduction

In recent years research has been focused on estimating the concentration matrix of Gaussian graphical models (GGMs) in high-dimensional setting [4, 7, 13]. GGMs are used to estimate conditional independence structures among a set of p random variables $\mathbf{X} = (X_1, \dots, X_p)'$ under the assumption that \mathbf{X} follows a multivariate normal distribution. These conditional independences are represented by a graph which is defined as a pair $\mathcal{G} = (V, E)$, where V is a finite set of vertices (vertex-set), and E is a subset of cartesian product $\mathcal{V} \times \mathcal{V}$ (edge-set). In particular, we shall consider undirected graphs. An undirected graph is a graph with only undirected edges. Edge $(i, j) \in E$ is called undirected if also $(j, i) \in E$ otherwise it is called directed. The problem of drawing edges in the graph is equivalent to the problem of identifying non zero entries in the concentration matrix which is the inverse of the variance-covariance matrix. A structured concentration matrix is a concentration matrix with constraints on the parameters. Imposing constraints is equivalent to impose structures on the conditional independence graph. Examples of GGMs with structured precision matrix are given in [9] and

[1]. In the former paper the authors worked on maximum likelihood approaches and impose structures on the concentration matrix. In the latter paper the authors worked on ℓ_1 -penalized maximum likelihood models applied to longitudinal high-dimensional data. Adding symmetry restriction to the concentration matrix is useful when parsimony is needed since it reduces the number of parameters to be estimated. Using ℓ_1 -penalty to penalized the likelihood function has the advantage of avoiding problem with multiple statistical tests since it selects a subset of important variables in the optimization step. It becomes important to have both parsimony and regularization when estimating covariance matrix of large dimensions with relatively few observations. This paper introduces two efficient cyclic coordinate algorithms in order to deal with new types of GGMs which have structured concentration matrices. The `sglasso` package, available under general public license (GPL ≥ 2) from the Comprehensive R Archive Network at <http://CRAN.R-project.org/package=sglasso>, implements the proposed algorithms.

2 Preliminary and Notation

Assume N independent and identically distributed random variable \mathbf{X} , where $\mathbf{X} \sim \mathcal{N}(0, \boldsymbol{\Sigma})$. Let $\mathbf{K} = \boldsymbol{\Sigma}^{-1}$ be the concentration matrix. The log-likelihood function, up to an additive constant, is:

$$2\ell(\mathbf{K}) = N\{\log \det \mathbf{K} - \text{tr}(\mathbf{R}\mathbf{K})\},$$

where $\mathbf{R} = N^{-1} \sum_{i=1}^N \mathbf{X}_i \mathbf{X}_i'$ is an estimator of the covariance matrix. Since \mathbf{R} is a singular matrix, in the high-dimensional setting, this estimator cannot be used to obtain $\hat{\mathbf{K}}$. Whereas, the graphical lasso (glasso) estimator is suitable for high-dimensional data (see [7]). The estimate is obtained by optimizing the following objective function:

$$\hat{\mathbf{K}} = \arg \max_{\mathbf{K}} \left[\ell(\mathbf{K}) - \rho \sum_{i,j} |k_{ij}| \right], \quad (1)$$

where ρ is a tuning parameter used to control the amount of shrinkage.

Several algorithms have been proposed to solve the optimization problem in (1). A block coordinate descent algorithm was proposed in [7]. This algorithm is based on the original approach proposed in [4] which uses the dual of the maximization problem (1). The interested reader is referred to [11, 16]. The glasso estimator is one of the most important estimator, proposed in the literature, to make sparse inference in high-dimensional GGMs. However, there are several important applications in which structured concentration matrix can improve data analysis. Two motivating examples are given in [9]. Next, we illustrate a motivating example. Suppose that we have collected longitudinal data at T time points. Let $\mathbf{X}_t = (X_{1t}, \dots, X_{pt})'$ be the p -dimensional random variable at time t . Assume that $\mathbf{X} = (\mathbf{X}'_1, \dots, \mathbf{X}'_T)'$ follows a multivariate normal distribution, the concentration matrix has the following block structure

$$\mathbf{K} = \begin{bmatrix} \mathbf{K}_{1,1} & \mathbf{K}_{1,2} & \cdots & \mathbf{K}_{1,T} \\ \mathbf{K}'_{1,2} & \mathbf{K}_{2,2} & \cdots & \mathbf{K}_{2,T} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{K}'_{1,T} & \mathbf{K}'_{2,T} & \cdots & \mathbf{K}_{T,T} \end{bmatrix},$$

where $\mathbf{K}_{t,t} = (k_{it,jt})$ gives information about the conditional independence structure among the p random variables at time t , and $\mathbf{K}_{t,t+h} = (k_{it,j(t+h)})$ gives information about the conditional

i.	$k_{it,i(t+h)} = \theta^{s_h}$	constant effect	i.	$k_{it,j(t+h)} = \theta^{n_h}$	constant effect
ii.	$k_{it,i(t+h)} = \theta_t^{s_h}$	time effect	ii.	$k_{it,j(t+h)} = \theta_t^{n_h}$	time effect
iii.	$k_{it,i(t+h)} = \theta_i^{s_h}$	unit effect	iii.	$k_{it,j(t+h)} = \theta_{ij}^{n_h}$	unit effect
iv.	$k_{it,i(t+h)} = \theta_{i,t}^{s_h}$	interaction effect	iv.	$k_{it,j(t+h)} = \theta_{ij,t}^{n_h}$	interaction effect

Table 1: Equality constrains on the entries of $\mathbf{S}_{t,h}$ and $\mathbf{N}_{t,h}$.

independence structure between \mathbf{X}_t and \mathbf{X}_{t+h} . An interpretation of the elements of the submatrices $\mathbf{K}_{t,t+h}$ brings to the notion of natural structure, i.e:

$$\mathbf{K}_{t,t+h} = \begin{bmatrix} k_{1t,1(t+h)} & 0 & \dots & 0 \\ 0 & k_{2t,2(t+h)} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & k_{pt,p(t+h)} \end{bmatrix} + \begin{bmatrix} 0 & k_{1t,2(t+h)} & \dots & k_{1t,p(t+h)} \\ k_{2t,1(t+h)} & 0 & \dots & k_{2t,p(t+h)} \\ \vdots & \vdots & \ddots & \vdots \\ k_{pt,1(t+h)} & k_{pt,2(t+h)} & \dots & 0 \end{bmatrix}$$

$$= \mathbf{S}_{t,h} + \mathbf{N}_{t,h},$$

where the entries of the matrix $\mathbf{S}_{t,h}$ are called self-self conditional dependences at temporal lag h and represent the (negative) self-similarity of a given random variable across different time points. The entries of the matrix $\mathbf{N}_{t,h}$ are the conditional dependencies among the p random variables with time lag h . Equality constraints reported in Table 1 can be imposed for each lag h (see [1]) to reduce the number of the parameters and to make interpretation of the results more relevant. This is similar to ANOVA models. The flexibility of this new class of models allows the user to take into account time dynamics, to known present or absent links, and to impose particular autoregressive structures. We call this new class of Gaussian graphical models penalized $RCON(\mathcal{V}, \mathcal{E})$ model.

3 The weighted ℓ_1 -penalized $RCON(\mathcal{V}, \mathcal{E})$ model

We use coloured graphs to impose structure on the graph. Specifically, colouring the vertices with $R \leq |V|$ different colours induce a partition of V in $\mathcal{V}_1, \dots, \mathcal{V}_R$ disjoint sets which we call vertex colour classes. All vertices belonging to the same vertex colour class have the same colour. Similarly, there is a partition of the edge-set into $S \leq |E|$ disjoint subsets $\mathcal{E}_1, \dots, \mathcal{E}_S$, which we call edge colour classes. All the undirected edges belonging to the same edge colour class are labeled with the same colour. We shall call $\mathcal{V} = \{\mathcal{V}_1, \dots, \mathcal{V}_R\}$ vertex colouring, $\mathcal{E} = \{\mathcal{E}_1, \dots, \mathcal{E}_S\}$ edge colouring and the pair $(\mathcal{V}, \mathcal{E})$ a coloured graph. This is implicitly referred to the undirected graph $\mathcal{G} = (V, E)$ with partitions $V = \mathcal{V}_1 \cup \dots \cup \mathcal{V}_R$ and $E = \mathcal{E}_1 \cup \dots \cup \mathcal{E}_S$. The $RCON(\mathcal{V}, \mathcal{E})$ model is a GGM obtained using a coloured graph to specify equalities constraints among the elements of the concentration matrix.

Let us consider the coloured graph $(\mathcal{V}, \mathcal{E})$, than the $RCON(\mathcal{V}, \mathcal{E})$ is specified by the following restrictions

- i) $k_{ii} = \eta_n$ for any $i \in \mathcal{V}_n$,
- ii) $k_{ij} = \theta_m$ for any $(i, j) \in \mathcal{E}_m$.

The concentration matrix can be specified as

$$\mathbf{K}(\boldsymbol{\psi}) = \sum_{n=1}^R \eta_n \mathbf{D}_n + \sum_{m=1}^S \theta_m \mathbf{T}_m, \tag{2}$$

where $\boldsymbol{\psi} = (\boldsymbol{\eta}', \boldsymbol{\theta}')$, \mathbf{D}_n is a diagonal matrix with entries $D_{ii}^n = 1$ if $i \in \mathcal{V}_n$ and zero otherwise, \mathbf{T}_m is a symmetric matrix with entries $T_{ij}^m = 1$ if the undirected edge (i, j) belongs to the edge colour class \mathcal{E}_m and zero otherwise. The objective function is

$$\hat{\boldsymbol{\psi}} = \arg \max_{\boldsymbol{\psi} \in \mathbb{R}^{(R+S)}} \log \det \mathbf{K}(\boldsymbol{\psi}) - \text{tr}\{\mathbf{R}\mathbf{K}(\boldsymbol{\psi})\} - \rho \sum_{m=1}^S w_m |\theta_m|, \quad (3)$$

where w_m are positive weights used to improve the edge selection behavior of the proposed estimator.

4 Two cyclic coordinate algorithms

In this section we give some details of the proposed cyclic coordinate algorithms to solve the optimization problem in (3). The first algorithm is a cyclic coordinate minimization (CCM) algorithm which is based on the idea to compute $\hat{\boldsymbol{\psi}}$ by cycling maximizing the objective function $\ell_p(\boldsymbol{\psi}) = \log \det \mathbf{K}(\boldsymbol{\psi}) - \text{tr}\{\mathbf{R}\mathbf{K}(\boldsymbol{\psi})\} - \rho \sum_{m=1}^S w_m |\theta_m|$. This algorithm is described by the pseudo-code reported in Algorithm 4.1. The second algorithm is a cyclic coordinate descent (CCD) algorithm obtained substituting the objective function $\ell_p(\boldsymbol{\psi})$ with an one-dimensional approximation. The second algorithm can be useful for higher-dimensional problems since it reduces the computational burden. The main idea underlying this family of algorithms is to choose, at each iteration, an index and then to optimize the objective function with respect to the corresponding parameter keeping all the remaining indexes fixed. The index of the parameter that will be updated can be selected by several rules. For example, a greedy rule is proposed in [17]. This rule consists of updating the parameter with the most negative directional derivative. The simpler cycling rule is used in [6] to define a cyclic coordinate descent (CCD) method for the lasso estimator [15]. The CCD algorithm was also extended to generalized linear models in [8] and to Cox's proportional hazard model in [14]. Next, we denote $\ell(\theta_m)$ the log-likelihood function which is a function of the parameter θ_m . The same meaning is used to $\ell(\eta_m)$.

Algorithm 4.1.

Pseudo-code of the proposed CCM algorithm

Step 1 initialize $\boldsymbol{\psi}$ to a starting value

Step 2 repeat

Step 3 for $m = 1$ to S

Step 4 maximize $\ell_p(\boldsymbol{\psi})$ with respect to θ_m keeping all the remaining parameter fixed

Step 5 end for

Step 6 for $n = 1$ to R

Step 7 maximize $\ell_p(\boldsymbol{\psi})$ with respect to η_n keeping all the remaining parameter fixed

Step 8 end for

Step 9 until a convergence criterion is met

Suppose that we have computed the estimator $\hat{\psi}$ for a given value of the tuning parameter, say ρ' , and we want to compute a new estimate for a value of the tuning parameter, say ρ , with $\rho < \rho'$. If ρ is close enough to ρ' , the one-dimensional log-likelihood function $\ell(\theta_m)$ can be approximated by standard Taylor expansion, with respect to θ_m , around the old estimate $\hat{\psi}$. By straightforward algebra, it is easy to see that $\ell_p(\theta_m)$ can be approximated as follows

$$\begin{aligned}\ell_p(\theta_m) &\approx \ell(\hat{\psi}) - \rho \sum_{n \neq m}^S w_n |\hat{\theta}_n| + \frac{\partial \ell(\hat{\psi})}{\partial \theta_m} (\theta_m - \hat{\theta}_m) + \frac{1}{2} \frac{\partial^2 \ell(\hat{\psi})}{\partial \theta_m^2} (\theta_m - \hat{\theta}_m)^2 - \rho w_m |\theta_m| = \\ &= C(\hat{\psi}) + \frac{1}{2} \frac{\partial^2 \ell(\hat{\psi})}{\partial \theta_m^2} (\theta_m - \hat{\vartheta}_m)^2 - \rho w_m |\theta_m|,\end{aligned}\quad (4)$$

where $C(\hat{\psi}) = \ell(\hat{\psi}) - \rho \sum_{n \neq m}^S w_n |\hat{\theta}_n| - \frac{1}{2} \{\partial^2 \ell(\hat{\psi}) / \partial \theta_m^2\}^{-1} \partial_m \ell(\hat{\psi})^2$ is a constant with respect to θ_m and $\hat{\vartheta}_m = \hat{\theta}_m - \{\partial^2 \ell(\hat{\psi}) / \partial \theta_m^2\}^{-1} \partial_m \ell(\hat{\psi})$. Using approximation (4), the original maximization problem specified in Step 4 can be locally substituted by the simpler problem

$$\min_{\theta_m \in \mathbb{R}} \frac{1}{2} I_m(\hat{\psi}) (\theta_m - \hat{\vartheta}_m)^2 + \rho w_m |\theta_m|,\quad (5)$$

where $I_m(\hat{\psi}) = -\partial^2 \ell(\hat{\psi}) / \partial \theta_m^2$ is the Fisher information for θ_m evaluated at $\hat{\psi}$. Problem (5) can be solved in closed form (see [6]), i.e. $\hat{\theta}_m = S(\hat{\vartheta}_m; w_m I_m^{-1}(\hat{\theta}) \rho)$, where $S(x; \lambda) = \text{sign}(x)(|x| - \lambda)_+$ is the soft-thresholding operator. The proposed CCD algorithm is summarized in the pseudo-code reported in Algorithm 4.2.

Algorithm 4.2.

Pseudo-code of the second proposed CCD algorithm

Step 1 initialize $\hat{\psi} = (\hat{\eta}^T, \hat{\theta}^T)^T$ to a given starting value

Step 2 $\mathbf{K}(\hat{\psi}) \leftarrow \sum_{n=1}^R \hat{\eta}_n \mathbf{D}_n + \sum_{m=1}^S \hat{\theta}_m \mathbf{T}_m$

Step 3 $\Sigma(\hat{\psi}) \leftarrow \mathbf{K}^{-1}(\hat{\psi})$

Step 4 repeat

Step 5 for $m = 1$ to S

Step 6 $\partial_m \ell(\hat{\psi}) \leftarrow \text{tr}\{\mathbf{T}_m(\Sigma(\hat{\psi}) - \mathbf{R})\}$

Step 7 $I_m(\hat{\psi}) \leftarrow \text{tr}\{\mathbf{T}_m \Sigma(\hat{\psi}) \mathbf{T}_m \Sigma(\hat{\psi})\}$

Step 8 $\vartheta_m \leftarrow \hat{\theta}_m + I_m^{-1}(\hat{\psi}) \partial_m \ell(\hat{\psi})$

Step 9 $\hat{\theta}_m \leftarrow S(\vartheta_m; w_m I_m^{-1}(\hat{\theta}) \rho)$

Step 10 $\Sigma(\hat{\psi}) \leftarrow (\mathbf{K}(\hat{\psi}) + \hat{\theta}_m \mathbf{T}_m)^{-1}$

Step 11 end for

Step 12 for $n = 1$ to R

Step 13 $\partial_n \ell(\hat{\psi}) \leftarrow \text{tr}\{\mathbf{D}_n(\Sigma(\hat{\psi}) - \mathbf{R})\}$

Step 14 $I_n(\hat{\psi}) \leftarrow \text{tr}\{\mathbf{D}_n \Sigma(\hat{\psi}) \mathbf{D}_n \Sigma(\hat{\psi})\}$

Step 15 $\hat{\eta}_n \leftarrow \hat{\eta}_n + I_n^{-1}(\hat{\psi}) \partial_n \ell(\hat{\psi})$

Step 16 $\Sigma(\hat{\psi}) \leftarrow (\mathbf{K}(\hat{\psi}) + \hat{\eta}_n \mathbf{D}_n)^{-1}$

Step 17 *end for*

Step 18 *until a convergence criterion is met*

At each inner loop $O(p^3)$ operations are needed to compute the inverse of the structured concentration matrix. We use the iterative algorithm proposed in [12] to reduce the computational burden. This involves step 10 and 16 which require the inversion of the sum of two matrices and requires $O(p^2)$ operations.

Pathway solution and definition of the weights of sglasso

The algorithms proposed in the previous section are developed for a fixed value of the tuning parameter. However, we seldom know the optimal value of ρ . Usually, the optimal value of ρ is found by computing the solution for a path of its decreasing values. Then, every fitted model is evaluated by a measure of goodness-of-fit. Warm-start or continuation methods use the solution found at the previous step as initial guess for the solution corresponding to the new value of the tuning parameter. Firstly, we begin with a value of the tuning parameter sufficiently large to ensure that all the penalized parameters are set to zero. Secondly, we decrease ρ , using a multiplicative grid, until we arrive next to the unconstrained solution. The largest value of the parameter ρ can be found as consequence of the following Karush-Kuhn-Tucker conditions:

$$\begin{aligned} \frac{\partial \ell(\hat{\psi})}{\partial \eta_m} &= \text{tr}\{\mathbf{D}_n(\Sigma(\hat{\psi}) - \mathbf{R})\} = 0, \\ \frac{\partial \ell(\hat{\psi})}{\partial \theta_m} &= \text{tr}\{\mathbf{T}_m(\Sigma(\hat{\psi}) - \mathbf{R})\} = \rho w_m \hat{\gamma}_m, \end{aligned}$$

where $\hat{\gamma}_m = \text{sign}(\hat{\theta}_m)$, if $\hat{\theta}_m \neq 0$, and $\hat{\gamma}_m \in (-1, 1)$ if $\hat{\theta}_m = 0$. Since at the starting point each $\hat{\theta}_m$ is equal to zero, it is easy to see that $\rho_{max} = \max_m |\text{tr}\{\mathbf{T}_m \mathbf{R}\}|$, and consequently, the index of the first $\hat{\theta}_m$ that will become different from zero is

$$\arg \max_m |\text{tr}(\mathbf{T}_m \mathbf{R})|.$$

The previous result shows that the variable selection behavior of the proposed estimator can be compromised if the S elements of the score vector, corresponding to the ℓ_1 -penalized parameters, are not comparable in terms of variability. This problem, that also affects the classical lasso estimator, was solved in [2] introducing the notion of dglars estimator, which is an extension of the least angle regression method [5] for regression models based on the exponential family. Using the differential geometrical structure of a generalized linear model, the authors relate the variable selection behavior of the dglars estimator to the well known Rao score test statistics, which are defined as the elements of the score vector divided by the corresponding asymptotic estimates of the variances. Similarly to the approach proposed in [2], in this paper we propose to use the weights $w_m = \text{tr}(\mathbf{T}_m \mathbf{R} \mathbf{T}_m \mathbf{R})$ which are the asymptotic estimates of the variance of the elements of the score vector evaluated at the starting point.

5 Conclusions

In this paper we have introduced a new estimator to make sparse inference on a high-dimensional $RCON(\mathcal{V}, \mathcal{E})$ model, which is based on the idea to use a weighted ℓ_1 -norm defined on the parameters of a structured concentration matrix. Moreover, we have proposed two cyclic coordinate algorithms, which are suitable for large data sets, to compute the proposed estimator. The `sglasso` estimator and the two proposed algorithms are implemented in the R package `sglasso` which is available at <http://CRAN.R-project.org/package=sglasso>.

Bibliography

- [1] Abbruzzo, A. and Wit, E. C. (submitted) *Dynamic factorial graphical models for dynamic networks*. Network Sci.
- [2] Augugliaro, L., Mineo A. M. and Wit E. C. (2013) *Differential geometric least angle regression: a differential geometric approach to sparse generalized linear models*. J. Roy. Statist. Soc. Ser. B, **75**(3), 471–498.
- [3] Augugliaro, L. (2014) *sglasso: Lasso method for $RCON(V,E)$ models*.
- [4] Banerjee, O., El Ghaoui, L. and d’Aspremont A. (2008) *Model selection through sparse maximum likelihood estimation for multivariate gaussian or binary data*. J. Mach. Learn. Res., **9**, 485–515.
- [5] Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. *Least angle regression*. Ann. Statist., **32**(2), 407–499.
- [6] Friedman, J., Hastie, T., Höfling, H. and Tibshirani, R. (2007) *Pathwise coordinate optimization*. Ann. Appl. Statist., **1**(2), 302–332.
- [7] Friedman, J., Hastie, T. and Tibshirani, R. (2008) *Sparse inverse covariance estimation with the graphical lasso*. Biostatistics, **9**(3), 432–441.
- [8] Friedman, J., Hastie, T. and Tibshirani, R. (2010) *Regularized paths for generalized linear models via coordinate descent*. J. Stat. Softw, **33**(1), 1–22.
- [9] Højsgaard, S. and Lauritzen, S. L. (2008) *Graphical Gaussian model with edge and vertex symmetries*. J. R. Statist. Soc. B, **70**(5), 1005–1027.
- [10] Lauritzen, S. L. (1996) *Graphical Models*. Oxford: Clarendon.
- [11] Mazumder, R. and Hastie, T. (2012) *The graphical lasso: new insights and alternatives*. Electron. J. Stat., **6**, 2125–2149.
- [12] Miller, K. S. (1981) *On the inverse of the sum of matrices*. Mathematics Magazine, **54**(2), 67–72.
- [13] Meinshausen, N. and Bühlmann, P. (2006) *High-dimensional graphs and variable selection with the lasso*. Ann. Statist., **34**(3), 1436–1462.

- [14] Simon, N., Friedman, J., Hastie, T. and Tibshirani, R. (2011) *Regularized paths for Cox's proportional hazards model via coordinate descent*. J. Stat. Softw., **39**(5), 1–13.
- [15] Tibshirani, R. (1996) *Regression shrinkage and selection via the lasso*. J. Roy. Statist. Soc. Ser. B, **58**(1), 267–288.
- [16] Witten, D. M., Friedman, J. and Simon, N. (2011) *New insights and faster computation for the graphical lasso*. J. of Comp. and Graph. Statist., **20**(4), 892–900.
- [17] Wu, T. T. and Lange K. (2008) *Coordinate descent algorithms for lasso penalized regression*. Ann. Appl. Statist., **2**(1), 224–244.
- [18] Yuan, M. and Lin, Y. (2007) *Model selection and estimation in the Gaussian graphical model*. Biometrika, **94**(1), 19–35.

The optimal number of lags in variogram estimation in spatial data analysis

Sujung Kim, *Okayama University*, sjkim1021@s.okayama-u.ac.jp

Kuniyoshi Hayashi, *Okayama University*, k-hayashi@ems.okayama-u.ac.jp

Koji Kurihara, *Okayama University*, kurihara@ems.okayama-u.ac.jp

Abstract. The variogram plays an important role in spatial data analysis. Geostatistical spatial data are analyzed in three stages: estimation of the variogram, model fitting for the estimated variogram, and fitting the chosen model to the estimated variogram model parameters. The proper estimate of the variogram is important since it affects the next two stages.

To estimate the variogram, we must first decide on the ‘number of lags k ’. Semivariogram estimation is strongly influenced by number of lags k , which serves as a smoothing parameter. This means that k could significantly influence the least square estimator and kriging predictor. However, there is no established rule for selecting the number of lags when estimating variograms. The selection of a proper k value is important, but few studies have been done in this regard. In this paper, we propose a method for choosing the optimal number of lags based on leave-one-out cross-validation (LOOCV) and the Akaike information criterion (AIC).

Keywords. Akaike information criterion, lag increment, leave-one-out cross-validation

1 Introduction

The variogram plays a central role when analyzing spatial data. A valid variogram model must first be selected, and the parameters of the model estimated before kriging (spatial prediction) is performed.

In general, the variogram is estimated with a method of moment estimator [6], and the lag increment or number of lags must be chosen as it is being estimated. In practical simulation analysis, a data analyst estimates the variogram using several different numbers of lags, and then selects the best number of lags value among them. This method is subjective and can sometimes result in reposterous variogram estimation values. This paper proposes a method

for choosing the optimal number of lags when estimating variograms based on the given spatial data.

We conduct a simulation study to demonstrate our procedure. For simplicity, we assume that the underlying process of the observed spatial data is stationary and isotropic. In all data analyses, we used the environment of R.

We describe the spatial statistics approach used in Section 2. In Section 3, we present our simulation results showing the optimal number of lags in variogram estimation. We present our concluding remarks in Section 4.

2 Spatial prediction

Spatial data can be considered to be a realization of a stochastic process $Z(s)$, i.e.,

$$\{Z(s) : s \in D \subset R^d\}, \quad (1)$$

where s indicates a location in D and R^d ($d = 1, 2, 3$) is a d -dimensional Euclidean space. The basic form of spatial data is expressed as (z_i, s_i) , $i = 1, \dots, n$, where z_i is the i -th observation of a phenomenon of interest at location s_i .

Assume that this process satisfies the hypothesis of intrinsic stationarity:

- (a) $E(Z(s)) = \mu$, for all $s \in D$,
- (b) $Cov(Z(s_i), Z(s_j)) = C(h) = C(s_i - s_j) < \infty$, for all $s_i, s_j \in D$,
- (c) $Var(Z(s_i) - Z(s_j)) = 2\gamma(s_i - s_j) = 2\gamma(h)$, for all $s_i, s_j \in D$,

where $2\gamma(h)$ is the variogram, and $C(h)$ is the covariance for pairs of points separated by Euclidean distance h (the covariogram). In this paper, we suppose that $2\hat{\gamma}(h)$ is a variogram estimator for a given lag h , based on a sample $\{Z(s_1), \dots, Z(s_n)\}$ of the spatial process; let h_1, \dots, h_k be the vector lags defined by $h_i = ih / \|h\|$, $i = 1, \dots, k$, where $1 \leq k \leq K$, and K is the maximum possible distance between data in the direction h divided by $\|h\|$. [3]

Estimation of the variogram

The first step in spatial analysis is estimating the variogram $\gamma(h)$ using the observed data. When we assume the variogram to be isotropic, we can calculate an estimator for the variogram, called the sample variogram [6], using

$$\hat{\gamma}(h) = \frac{1}{2|N(h)|} \sum_{N(h)} (z(s_i) - z(s_j))^2, \quad (2)$$

where $N(h)$ is the set of all pairwise Euclidean distances $s_i - s_j = h$ and $|N(h)|$ is the number of the distinct pairs in $N(h)$. $z(s_i)$ and $z(s_j)$ are the data values at spatial locations s_i and s_j , respectively. In this formulation, h represents a distance measure with only magnitude.

When the variogram is isotropic, we can compute the directional sample variogram using the same formula by replacing h with vector \mathbf{h} . In practice, to calculate the variogram values using Eq. (2), we first select the lag distances h , then calculate the variogram values by regarding pairs with distance within $h \pm$ lag tolerance as the pairs in $N(h)$. The lag tolerance, which establishes distance bins for the lag increments, accommodates for unevenly spaced observations. The

lag increment defines the distances at which the variogram is calculated, and the number of lags in conjunction with the size of the lag increment will define the total distance over which the variogram is being calculated. To estimate the variogram, we next have to choose the lag increment or the number of lags.

Variogram model fitting

The next stage in spatial analysis is fitting a model that gives the best dependence (auto-correlation structure) in the underlying stochastic process. Most variogram models contain three parameters sill, range, and nugget (or nugget effect). Sill is a variogram threshold for lag distances. Range is the lag distance at which the variogram reaches the sill value. The nugget represents the variability at distances smaller than the typical sample spacing, including the measurement error. In theory, the variogram value at the origin (0 lag) should be zero. Thus far, several variogram models have been proposed according to their forms; for example, gaussian, exponential, and spherical models as bounded and power and linear models as unbounded variogram models. In this paper, we describe only the two models used in our study the exponential and spherical [1].

The exponential model is as follows:

$$\gamma_{exp}(h; \theta) = \begin{cases} 0, & h = 0, \\ c_n + c_s \left\{ 1 - \exp\left(-\frac{\|h\|}{c_r}\right) \right\}, & h > 0, \end{cases} \quad (3)$$

for $\theta = (c_n, c_s, c_r)'$, $c_n \geq 0$, $c_s \geq 0$, and $c_r \geq 0$.

The spherical model is

$$\gamma_{sph}(h; \theta) = \begin{cases} 0, & h = 0, \\ c_n + c_s \left\{ \frac{3}{2} \left(\frac{\|h\|}{c_r}\right) - \frac{1}{2} \left(\frac{\|h\|}{c_r}\right)^3 \right\}, & 0 < h \leq c_r, \\ c_n + c_s, & h > c_r, \end{cases} \quad (4)$$

for $\theta = (c_n, c_s, c_r)'$, $c_n \geq 0$, $c_s \geq 0$, and $c_r \geq 0$.

Kriging

Kriging is a linear interpolation method that allows predictions of unknown values in a random function from observations at known locations. There are a few type of kriging for spatial prediction problems in spatial statistics, including simple kriging, ordinary kriging, and universal kriging. In our simulation, we perform only ordinary kriging, which is often associated with the best linear unbiased estimator (BLUE). Ordinary kriging is based on a random function model of spatial correlation for calculating a weighted linear combination of available samples to predict a nearby unsampled location. Weights are chosen to ensure zero average error for the model and to minimize the model's error variance [4]. Ordinary kriging [5, 7] refers to spatial prediction under the following two assumptions. First, the model assumption is as follows:

$$Z(s) = \mu + \delta(s), \quad s \in D, \quad \mu \in R, \quad (5)$$

where μ is unknown. The second is the predictor assumption:

$$Z_{OK}^*(s_0) = \sum_{\alpha=1}^n w_{\alpha} Z(s_{\alpha}). \quad (6)$$

To minimize the error variance under the constraint $\sum_{\alpha=1}^n w_{\alpha} = 1$, we set up a system that minimizes Q , comprising the error variance and an additional term involving the Lagrange parameter, μ_{OK} :

$$Q = E \left[\left(Z^*(s_0) - Z(s_0) \right)^2 \right] + 2\mu_{OK} \left(1 - \sum_{\alpha=1}^n w_{\alpha} \right). \quad (7)$$

This minimization with respect to the Lagrange parameter forces the constraint to be obeyed:

$$\begin{cases} \frac{\partial Q}{\partial w_{\beta}} = -2 \sum_{\alpha=1}^n w_{\alpha} \gamma(s_{\alpha} - s_{\beta}) + 2\gamma(s_{\beta} - s_0) - 2\mu = 0, \\ \frac{\partial Q}{\partial \mu} = 1 - \sum_{\alpha=1}^n w_{\alpha} = 0. \end{cases} \quad \beta = 1, \dots, n,$$

In this case, the system of equations for the kriging weights is

$$\begin{cases} \sum_{\beta=1}^n w_{\beta}^{OK} \gamma(s_{\alpha} - s_{\beta}) + \mu_{OK} = \gamma(s_{\alpha} - s_0), \\ \sum_{\beta=1}^n w_{\beta}^{OK} = 1, \end{cases} \quad \alpha = 1, \dots, n, \quad (8)$$

where $\gamma(\cdot)$ is the covariance function for the residual component of the variable.

Once the kriging weights (and the Lagrange parameter) are obtained, the error variance for the ordinary kriging is given by

$$\sigma_{OK}^2 = \mu_{OK} - \gamma(s_0 - s_0) + \sum_{\alpha=1}^n w_{\alpha}^{OK} \gamma(s_{\alpha} - s_0). \quad (9)$$

3 Simulation study

The selection of lag size has significant effects on the sample semivariogram. For example, if the lag size is too large, short-range autocorrelation may be masked. If the lag size is too small, there may be many empty bins, and sample sizes within the bins will be too small to determine the bins representative averages. However, if the data are acquired using an irregular or random sampling scheme, a suitable lag size selection is not at all straightforward.

[5] suggests the following two practical rules for choosing the lag increment and number of lags: (i) the sample variogram should only be considered for distances h for which the number of pairs is greater than 30, and (ii) the distance of reliability for a sample variogram is $h < D/2$, where D is the maximum distance over the field of data. However, in practice, these rules are ambiguous when choosing the number of lags or the lag increment. In this paper, we focus on the number of lags denoted by symbol k because the above two rules are mutually reciprocal. Our main interest thus becomes finding the optimal number of lags among possible k values.

Model		Exponential			Spherical		
Number of lags	Sample size			Number of lags	Sample size		
	100	200	300		100	200	300
2	1.0801	1.0188	0.9065	2	2.3746	1.7157	1.7523
3	1.0193	0.9930	0.9105	3	2.2646	1.5523	1.4381
4	0.8970	0.8377	0.7439	4	2.1245	1.3444	1.1612
5	0.7079	0.6566	0.7001	5	1.9845	1.1996	1.0182
6	0.6611	0.6010	0.4275	6	1.9038	1.1152	0.7815
7	0.6497	0.5720	0.4102	7	1.8287	1.0637	0.7447
8	0.6322	0.5529	0.3983	8	1.8060	1.0291	0.7147
9	0.6297	0.5390	0.3892	9	1.7799	0.9971	0.6954
10	0.6174	0.5322	0.3829	10	1.7733	0.9780	0.6815
11	0.6113	0.5080	0.3777	11	1.4172	0.7637	0.5648
12	0.6107	0.5024	0.3733	12	1.4109	0.7489	0.5521
13	0.6097	0.4972	0.3698	13	1.3893	0.7390	0.5435
14	0.6062	0.4928	0.3672	14	1.4171	0.7361	0.5383
15	0.6064	0.4867	0.3652	15	1.3775	0.7329	0.5324
16	0.6044	0.4883	0.3629	16	1.3841	0.7234	0.5257
17	0.6043	0.4851	0.3622	17	1.3824	0.7200	0.5229
18	0.6043	0.4838	0.3618	18	1.3724	0.7145	0.5210
19	0.6041	0.4821	0.3614	19	1.3581	0.7158	0.5203
20	0.6040	0.4819	0.3607	20	1.3511	0.7120	0.5181

Table 1: Results of using LOOCV for choosing the optimal number of lags.

The optimal number of lags for LOOCV

In this paper, we consider the exponential and spherical models, which each contain three parameters (sill, range, and nugget), and we restrict the scope of the number of lags to be from 2 to 20 when selecting the optimal k .

As mentioned above, the simulation data are fixed in the two models and their three parameters, and the generated datasets (with sample sizes of 100, 200, and 300) include positions as well as the data values. When a theoretical variogram model is fitted to the number of lags k from 2 to 20, the optimal k can be selected on the basis of leave-one-out cross-validation (LOOCV). The selection of the optimal k can be explained as below.

Step 1 For a fixed lag k ($2 \leq k \leq 20$), estimate the variogram using $n - 1$ observations excepting the i -th one and obtain the predicted value $\hat{Z}(s_{-i})$ at the i -th location based on the estimated variogram.

Step 2 For every i ($i = 1, \dots, n$), calculate $Z(s_i) - \hat{Z}(s_{-i})$ based on the Step 1 from 1 to n ($n = 100, 200, \text{ and } 300$).

Step 3 For the fixed lag k ($2 \leq k \leq 20$), calculate the LOOCV statistic $\frac{1}{n} \sum_{i=1}^n \left| (Z(s_i) - \hat{Z}(s_{-i})) \right|^2$.

Step 4 Calculate the LOOCV for every k ($2 \leq k \leq 20$), and select the optimal k which minimizes the LOOCV.

The LOOCV result for given numbers of lags is presented in Table 1. The LOOCV [2] values in Eq. (10), are calculated as follows. The LOOCV uses a single observation from the original sample as the validation data, and the remaining observations as the training data. This is repeated such that each observation in the sample is used once as the validation data:

$$\frac{1}{n} \sum_{i=1}^n \left| (Z(s_i) - \hat{Z}(s_{-i})) \right|^2, \quad (10)$$

where $Z(s_i)$ and $\hat{Z}(s_{-i})$ represent the observed and predicted values, respectively.

From Table 1, we can see that the results based on the sample sizes of 100, 200, and 300 in the exponential variogram model show similar values from $k = 6$ to $k = 20$, respectively. In addition, we can see that the results based on the sample sizes of 100, 200, and 300 in the spherical variogram model show similar values from $k = 11$ to $k = 20$, respectively.

The optimal number of lags for AIC

A satisfactory compromise between goodness of fit and complexity of the model can be achieved by applying the Akaike information criterion (AIC). For a given set of data, the variable part of the AIC is estimated by

$$\hat{A} = -2n \ln \hat{R} + 2p, \quad (11)$$

where n is the number of variogram cloud, \hat{R} is the value of R which maximizes the likelihood (R is a vector of m parameters of covariogram model), and p is the number of parameters in the variogram model. The model to choose is the one for which \hat{A} is least.

Similarly, when applying the AIC, the simulation data are fixed in the two models and their three parameters, and the generated datasets (with sample sizes of 100, 200, and 300) include positions and the data values. When a theoretical variogram model is fitted to the number of lags k from 2 to 20, the optimal number of lags k can be selected on the basis of the AIC. The optimal k is defined to be the value that minimizes AIC. The selection of the optimal k can be explained as below.

Step 1 Calculate the \hat{R} with the given data Z ($Z = Z(s_1), Z(s_2), \dots, Z(s_n)$) and parameters of covariogram model R .

Step 2 Calculate the AIC for variogram model for every k ($2 \leq k \leq 20$).

Step 3 Select the optimal k which minimizes the AIC.

From Table 2, for the sample size of 100 in the exponential variogram model, the minimum AIC value is achieved at $k = 5$. For sample sizes of 200 and 300, the minimum AIC values are achieved at $k = 7$. For the sample size of 100 in the spherical variogram model, the minimum value of AIC is achieved at $k = 5$; for sample sizes of 200 and 300, the minimum values of AIC are achieved at $k = 6$ and $k = 7$, respectively.

Model		Exponential			Spherical		
Number of lags	Sample size			Number of lags	Sample size		
	100	200	300		100	200	300
2	871.93	1504.93	2148.32	2	871.79	1500.28	2152.07
3	867.50	1479.62	2108.39	3	862.63	1474.49	2110.97
4	865.26	1476.76	2100.05	4	860.67	1470.30	2094.99
5	865.13	1475.28	2094.52	5	859.89	1469.23	2096.20
6	866.57	1474.24	2094.13	6	861.22	1468.11	2095.40
7	868.43	1473.09	2092.02	7	862.79	1468.42	2092.13
8	868.83	1475.09	2092.87	8	864.56	1468.15	2093.30
9	870.19	1477.29	2093.34	9	865.90	1470.65	2096.14
10	871.16	1477.72	2093.41	10	866.52	1472.52	2095.69
11	874.27	1479.66	2093.56	11	868.29	1471.98	2097.30
12	875.17	1481.74	2095.99	12	869.97	1474.22	2098.08
13	875.76	1482.48	2096.34	13	871.39	1476.40	2099.62
14	877.40	1483.90	2098.17	14	872.74	1476.99	2099.43
15	878.95	1483.82	2099.03	15	874.47	1478.59	2102.89
16	881.18	1485.92	2101.32	16	874.30	1480.46	2102.77
17	881.78	1488.60	2102.84	17	878.09	1481.32	2105.35
18	882.28	1489.51	2104.42	18	878.13	1482.73	2105.40
19	882.47	1490.40	2105.34	19	879.42	1484.28	2107.99
20	887.59	1491.95	2107.46	20	880.95	1486.11	2109.07

Table 2: Results of applying the AIC for choosing the optimal number of lags.

4 Conclusion

In this paper, we examined the performance of a variogram estimator in spatial models, focusing on a piecewise constant estimator for an isotropic variogram. We proposed a method for selecting the optimal number for the estimator using LOOCV and AIC in the spatial data analysis. The proposed method's usefulness is established through a simulation study. In the future, to validate our proposed procedure's usefulness in detail, we have to perform various simulation studies. In addition, we have to apply our method for finding the optimal number of lag to many real spatial data analysis.

Bibliography

- [1] Cressie, N. (1993) *Statistics for spatial data*. John Wiley & Sons.
- [2] Devijver, P. A. and Kittler, J. (1982) *Pattern recognition: A statistical approach*. Prentice Hall.
- [3] Genton, M. G. (1998) *Variogram fitting by generalized least squares using an explicit formula for the covariance structure*. *Mathematical Geology*, **30**, 323–345.

- [4] Isaaks, E. H. and Srivastava, R. M. (1989) *An introduction to applied geostatistics*. Oxford University Press.
- [5] Journel, A. G. and Huijbregts, C. J. (1978) *Mining geostatistics*. Academic Press.
- [6] Matheron, G. (1963) *Principles of geostatistics*. *Economic geology*, **58**, 1246–1266.
- [7] Matheron, G. (1971) *The theory of regionalized variables and its application*. Cahiers du Centre de Morphologie Mathématique, **NO.5**. Fontainebleau.

GRID for variable selection in high dimensional regression

Francesco Giordano, *University of Salerno*, giordano@unisa.it

Soumendra Nath Lahiri, *University of North Carolina*, snlahiri@ncsu.edu

Maria Lucia Parrella, *University of Salerno*, mparrella@unisa.it

Abstract. Given a nonparametric regression model, we assume that the number of covariates may increase infinitely but only some of these covariates are relevant for the model. Our goal is to identify the relevant covariates and to obtain some information about the structure of the model. We propose a new nonparametric procedure, called GRID, having the following features: (a) it automatically identifies the relevant covariates of the regression model, also distinguishing the nonlinear from the linear ones (a covariate is defined *linear/nonlinear* depending on the marginal relation between the response variable and such a covariate); (b) the interactions between the covariates (mixed effect terms) are automatically identified, without the necessity of considering some kind of stepwise selection method. In particular, our procedure can identify the mixed terms of any order (two way, three way, ...) without increasing the computational complexity of the algorithm; (c) it is completely data-driven, so being easily implementable for the analysis of real datasets. In particular, it does not depend on the selection of crucial regularization parameters, nor it requires the estimation of the nuisance parameter σ^2 (self-scaling). The acronym GRID derives from *Gradient Relevant Identification Derivatives*, meaning that the procedure is based on testing the significance of a partial derivative estimator.

Keywords. Variable selection, model selection, high dimension, nonparametric regression.

1 Introduction

Much work has been made in the context of variable selection. There are two main approaches to this problem. The first one is based on the idea of LASSO, which uses a penalized regression with additive models (see [7], [10] and [8] among others). The appeal of this approach is the fast rate of convergence, which essentially derives from the imposition of an additive model. On the other side, a serious drawback is given by the difficulty of implementation on real datasets given the necessity of setting crucial regularization parameters. The second approach, which has inspired this work, is based on a general regression function of dimension d , without imposing any additive restriction on the model (see [4]). The main advantage of this approach is its flexibility

and simplicity of implementation on real datasets (there are not regularization parameters). At the same time, it suffers from a low rate of convergence that makes it unsuitable for the analysis of high dimensional datasets. Our aim here is to deal with the last drawback. We propose a new procedure based on a variant of the local linear estimator which remarkably improves the rate of convergence of the selection procedure and makes it suitable for high dimensional cases, in particular when the dimension p is greater than the sample size n . This improvement is pursued by taking separated the stage of variable selection from the stage of function estimation.

We assume that the number of covariates p of the regression model may tend to infinity but only some of these covariates are relevant. Given that the function is completely unknown, our goal is to identify the relevant covariates and to obtain some information about the structure of model (1). In particular, we want to classify the covariates into disjoint sets: 1) the set of nonlinear covariates, which includes those variables having a nonlinear effect on the dependent variable (*i.e.*, with not constant gradient); 2) the set of linear covariates, which includes the variables having a linear effect on the response variable (*i.e.*, with constant gradient); 3) the set of irrelevant covariates, collecting the variables with gradient equal to zero. Denote with C , A and U the correspondent index sets and let $\Xi = C \cup A \cup U$ represent the set of regressors $\{1, \dots, d\}$. Secondly, we want to detect the interactions among the covariates, identifying the mixed effects. More specifically, the information on the interaction terms is given in the following way. Let I^j be the set of covariates mixed with the j -th covariate, for $j \in \Xi$. A convention used here is that $j \notin I^j$, which means that self-interaction is ignored in practice. We propose a procedure called GRID that gives a consistent estimation of the sets C , A and I^j , for $j \in C \cup A$. Other sets can be derived easily by known relationships. In particular, $I_C^j = I^j \cap C$ is the set of nonlinear covariates which are mixed with the j -th covariate and $I_A^j = I^j \cap A$ is the set of linear covariates mixed with the j -th covariate. Then $I^j = I_C^j \cup I_A^j$. Moreover, the set $C_c = \cup_{j \in C} I_C^j$ (or $C_a = \cup_{j \in A} I_C^j$) collects the nonlinear covariates which are mixed with nonlinear (or linear) covariates. Similarly, we have $A_c = \cup_{j \in C} I_A^j$ (or $A_a = \cup_{j \in A} I_A^j$). Finally, the sets $C_p = C \setminus (C_c \cup C_a)$ and $A_p = A \setminus (A_c \cup A_a)$ include the pure (*i.e.*, not mixed) nonlinear and linear covariates, respectively.

Given the short length of this paper, in the following sections we only present the main results and the basic structure of the GRID procedure. The acronym GRID derives from *Gradient Relevant Identification Derivatives*, because the procedure is based on testing the significance of a partial derivative estimator. Section 2 introduces the estimator and gives the notation used in this paper. In section 3 we present the theoretical foundations of the selection procedure. The GRID procedure is briefly described in section 4. Finally, section 5 reports the results of a simulation study and some final comments.

2 The estimator

In this paper we consider the following nonparametric regression model

$$Y_t = m(X_t) + \varepsilon_t, \quad (1)$$

where X_t represents the \mathbb{R}^d -valued covariate and ε_t is the error term with zero mean and variance σ^2 . Both X_t and ε_t are assumed *i.i.d.*, and the errors ε_t are supposed to be independent from X_t . Here $m(X_t) = E(Y_t|X_t) : \mathbb{R}^d \rightarrow \mathbb{R}$ is the multivariate conditional mean function. We use the notation $X_{(j)}$ to refer to the univariate covariates, for $j = 1 \dots, d$. We indicate with $f_X(\cdot)$ the

multivariate density function of X_t , having support $supp(f_X) \subseteq \mathbb{R}^d$, and with $f_\varepsilon(\cdot)$ the density of the error term. In order to improve the rate of convergence, we propose to base our identification procedure on a variant of the local linear estimator. In fact, the local linear estimator used in the RODEO is not well defined when $d > n$, given the necessity of inverting the regression matrix. To overcome this, we propose the estimator

$$M(x; H) = \frac{1}{n} \text{diag}(1, H^{-2}) \Gamma^T W \Upsilon \equiv \begin{pmatrix} M_0(x; H) \\ M_1(x; H) \end{pmatrix}, \tag{2}$$

where H is the diagonal strictly positive bandwidth matrix (of dimension $d \times d$) and

$$\Upsilon = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}, \quad \Gamma = \begin{pmatrix} 1 & (X_1 - x)^T \\ \vdots & \vdots \\ 1 & (X_n - x)^T \end{pmatrix}, \quad W = \begin{pmatrix} K_H(X_1 - x) & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & K_H(X_n - x) \end{pmatrix},$$

for some multivariate Kernel function $K(\cdot)$. Note that $M_0(x; H)$ is a scalar while $M_1(x; H)$ is a vector of length d . Our selection method is based on the derivatives of (2) w.r.t. the different bandwidths. So,

$$\begin{aligned} \dot{M}_{0j} &= \frac{\partial M_0(x; H)}{\partial h_j} & j = 1, \dots, d \\ \dot{M}_{1j} &= \frac{\partial M_1(x; H)}{\partial h_j} \equiv \{\dot{M}_{1j}^{(i)}\}_{i=1, \dots, d}, \end{aligned} \tag{3}$$

whose explicit expressions derive from

$$\frac{\partial M(x; H)}{\partial h_j} = \frac{\partial}{\partial h_j} \left[\frac{1}{n} \begin{pmatrix} 1 & 0 \\ 0 & H^{-2} \end{pmatrix} \Gamma^T W \Upsilon \right] = \frac{1}{n} O_j \Gamma^T W \Upsilon + \frac{1}{n} \begin{pmatrix} 1 & 0 \\ 0 & H^{-2} \end{pmatrix} \Gamma^T \frac{\partial}{\partial h_j} W \Upsilon,$$

where O_j is a matrix with $d + 1$ rows and $d + 1$ columns, with all zeros except the element in position $(j + 1, j + 1)$ which is equal to $-\frac{2}{h_j^3}$.

Since W is a diagonal matrix with elements

$$K_H(X_t - x) = \frac{1}{|H|} \prod_{k=1}^d K\left(\frac{X_{tk} - x_k}{h_k}\right),$$

its derivative with respect to h_j is

$$\frac{\partial}{\partial h_j} K_H(X_t - x) = K_H(X_t - x) \left(-\frac{1}{h_j} + \frac{\partial}{\partial h_j} \log K\left(\frac{X_{tj} - x_j}{h_j}\right) \right).$$

So

$$\frac{\partial}{\partial h_j} W = W L_j$$

where $L_j = \text{diag}\left(\frac{\partial \log K((X_{1j} - x_j)/h_j)}{\partial h_j} - \frac{1}{h_j}, \dots, \frac{\partial \log K((X_{nj} - x_j)/h_j)}{\partial h_j} - \frac{1}{h_j}\right)$. Finally, we propose the following estimator

$$\frac{\partial M(x; H)}{\partial h_j} = \frac{1}{n} \left[O_j \Gamma^T W + \begin{pmatrix} 1 & 0 \\ 0 & H^{-2} \end{pmatrix} \Gamma^T W L_j \right] \Upsilon \equiv \begin{pmatrix} \dot{M}_{0j} \\ \dot{M}_{1j} \end{pmatrix}. \tag{4}$$

3 Theoretical foundations

In this paper, we consider the following assumptions.

- A1) The bandwidth H is a diagonal and strictly positive definite matrix, $H = \text{diag}(h_1, \dots, h_d)$, with $h_j = O(1)$ for $j = 1, \dots, d$.
- A2) The d -variate Kernel function K is a product kernel, with compact support and zero odd moments. Therefore, the following moments exist bounded (we assume that $\mu_0 = 1$)

$$\mu_r = \int u_1^r K(u_1) du_1, \quad \nu_r = \int u_1^r K^2(u_1) du_1 \quad r = 0, 1, \dots, 4.$$

Moreover, we assume that $K \in C^1[-a, a]$ for some $a > 0$.

- A3) All the partial derivatives of the function $m(x)$ up to and including fifth order are bounded.
- A4) The density f_X is uniform on the unit cube.

The rationale of our selection procedure is based on the following result (see [3] for a proof).

Theorem 3.1. *Under model (1) and assumptions (A1)-(A4), the following result holds*

$$E \{ \dot{M}_{0j} \} = \begin{cases} \theta_{0j}^m \neq 0 & \text{if and only if } j \in C \\ \theta_{0j}^m = 0 & \text{otherwise;} \end{cases} \quad (5)$$

$$E \{ \dot{M}_{1j}^{(i)}, i \neq j \} = \begin{cases} \theta_{ij}^m \neq 0 & \text{if and only if } i \in I^j, j \in C \\ \theta_{ij}^m = 0 & \text{otherwise.} \end{cases} \quad (6)$$

Remark 3.1: Theorem 3.1 can be used to detect the nonlinear effects in model (1). In fact, basing on the (5), the derivatives \dot{M}_{0j} can be used in order to identify the nonlinear covariates, obtaining C . Basing on the (6), the derivatives $\dot{M}_{1j}^{(i)}$ can be used in order to identify the interactions for the nonlinear covariates, obtaining I^j , for $j \in C$.

Remark 3.2: The values of the bandwidths are not crucial in our procedure, because at this stage we are not interested in the estimation of the function $m(x)$, but only in variable selection. Note that our identification procedure is based on evaluating the bias of the estimator (2) (*i.e.*, the values θ_{ij}), which is zero for linear and irrelevant covariates. Therefore, we suggest to use a bandwidth matrix that produces a very high bias. This means to take very large bandwidths, for example $h = 0.9$, also improving the efficiency of the estimator.

Using Theorem 3.1, we cannot identify the linear covariates in $A_u = A_a \cup A_p$ and the *linear mixed effects* in I_A^j , for $j \in A$. Anyway, a convenient solution is to consider an auxiliary regression with some of the covariates transformed, so that the linear covariates of the original model become *nonlinear* in the auxiliary model. In particular, if we consider model (1) under the partition $\{C, A_c, A_u, U\}$, with U including the not relevant covariates, it must necessarily be

$$m(x) = m_1(x_C, x_{A_c}) + m_2(x_{A_u}).$$

Now, let us define a transformation $z = \phi(x)$ and its inverse $x = \phi^{-1}(z)$ as follows (component-wise)

$$z = \phi(x) = (x_C, x_{A_c}^{1/2}, x_{A_u}^{1/2}, x_U^{1/2}), \quad x = \phi^{-1}(z) = (x_C, z_{A_c}^2, z_{A_u}^2, z_U^2). \quad (7)$$

We can consider the following auxiliary regression

$$Y_t = m(\phi^{-1}(Z_t)) + \varepsilon_t = g(Z_t) + \varepsilon_t, \quad t = 1, \dots, n,$$

where the new regression function can be written as

$$g(z) = g_1(x_C, z_{A_c}) + g_2(z_{A_u}).$$

Note once again that we use the same index partition considered in the first regression. Thanks to the transformation in (7), it appears immediately that the function $g_2(\cdot)$ depends only on the covariates in A_u in a nonlinear manner.

Given that we are not interested in the exact estimation of the function $g(z)$ but only in variable selection, we can exclude the nonlinear covariates in C from the auxiliary regression. Note that, when we consider the auxiliary regression with the transformed covariates $Z_t = \phi(X_t)$, the density f_Z does not satisfy the assumption A4, so Theorem 3.1 cannot be applied. The following theorem covers this case.

Theorem 3.2. *Using model (1), assumptions (A1)-(A4) and the transformed random variables*

$$Z_t = \{\phi(X_{(s)}), s \in \overline{C}\}$$

with ϕ defined in (7), the following result holds for the estimator defined in (2)

$$E \{ \dot{M}_{0j} \} = \begin{cases} \theta_{0j}^g \neq 0 & \text{if and only if } j \in A \\ \theta_{0j}^g = 0 & \text{otherwise} \end{cases} \quad (8)$$

$$E \{ \dot{M}_{1j}^{(i)}, i \neq j \} = \begin{cases} \theta_{ij}^g \neq 0 & \text{if } i \in I^j, j \in A \\ \theta_{ij}^g = 0 & \text{if } j \in U. \end{cases} \quad (9)$$

Remark 3.3: Given the (8), the derivatives $\dot{M}_{0j} = \partial M_0(z; H) / \partial h_j$ calculated with the transformed covariates Z can be used in order to identify the linear covariates, obtaining the set A . Moreover, we can use the (9) in order to identify the *linear mixed effects* in I^j , for $j \in A$.

4 The GRID procedure

Here we present the algorithm for estimating and testing the values of θ_{ij} , in order to classify the covariates of model (1). Let $X_{(j)}$ represent a uniform covariate while $Z_{(j)}$ stands for the same covariate after applying the transformation (7). For brevity, we report only the first stage of the GRID procedure, useful to identify the relevant covariates. A second stage can be organized in a similar way in order to identify the mixed terms (both the stages of the procedure have been implemented in the simulation study).

- O. Set the bandwidth matrix to a high value (for example, $H = 0.9I_d$). Let $R = C \cup A$ denote the set of relevant covariates. Initialize all the sets $(C, A, R, R_X, R_Z, \dots)$ to the empty set \emptyset .

I. **First stage** (*identifying the relevant covariates*):

- For $j = 1, \dots, d$, do:
 - using the covariates $X_{(j)}$, $j \in \Xi$, compute the (univariate) statistic \dot{M}_{0j} defined in (4)
 - using the Empirical Likelihood technique, compute the threshold γ_0
 - if $\dot{M}_{0j} > \gamma_0$ then (*relevant covariate*)
 - insert the index j in the set R_X
 - using the covariates $Z_{(j)}$, $j \in \Xi$, compute the (univariate) statistic \dot{M}_{0j} defined in (4)
 - using the Empirical Likelihood technique, compute the threshold γ_0
 - if $\dot{M}_{0j} > \gamma_0$ then (*relevant covariate*)
 - insert the index j in the set R_Z
- $R = R_X \cup R_Z$.
- For $j \in R$, do:
 - using the covariates $X_{(j)}$, $j \in R$, compute the (univariate) statistic \dot{M}_{0j} defined in (4)
 - using the Empirical Likelihood technique, compute the threshold γ_1
 - if $\dot{M}_{0j} > \gamma_1$ then (*nonlinear covariate*) then insert the index j in the set C
 - otherwise (*linear covariate*) insert the index j in the set A .
- Output C, A

Remark 3.4: the Empirical Likelihood is used in order to test the relevance of the covariates. This nonparametric inferential technique has several advantages, among which the *self-scaling* property. Therefore, it is not necessary to estimate any nuisance parameter in order to do the test. Refer to [3] for more details on this aspect.

5 Simulation results

We evaluate the finite dimension performance of our procedure by means of a simulation study. We consider two different models, and for each one we simulate 200 Monte Carlo replications under different configuration of settings. In particular, we use three sample sizes $n = (300, 500, 1000)$ and three dimensions $d = (20, n/2, 2n)$. Therefore, the last value of d is such that $d > n$. Following the suggestion in [7], the additive terms in each model are standardized so that no one of them dominates the variance of the model.

As a first example, we consider the following **model 1**:

$$Y = X_{(6)}^3 X_{(7)}^3 + X_{(10)} + \varepsilon, \quad \varepsilon \sim N(0, 1)$$

where there are two nonlinear covariates, with a mixed term, and one pure linear covariate. In table 1, we report the rates of classification of a given covariate in the sets: R , as a relevant covariate; C , as a nonlinear covariate; $I(6, 7)$, as an interaction term between the two covariates 6 and 7. We can derive the percentages for the set A as a difference between the values of R and C . All the values less than 0.025 are not shown, and they are reported with the symbol “*”.

As a second example, we use the following **model 2**:

$$Y = X_{(1)} X_{(2)} + X_{(1)} X_{(7)}^3 + \varepsilon, \quad \varepsilon \sim N(0, 1)$$

		n	$d = 20$			$d = n/2$			$d = 2n$		
			R	C	$I(6, 7)$	R	C	$I(6, 7)$	R	C	$I(6, 7)$
X_6	300	0.975	0.330	0.900	0.855	0.335	0.720	0.630	0.330	0.365	
	500	1.000	0.610	1.000	0.990	0.595	0.985	0.810	0.480	0.620	
	1000	1.000	0.910	1.000	1.000	0.915	1.000	0.910	0.835	0.815	
X_7	300	0.940	0.325	-	0.875	0.335	-	0.580	0.250	-	
	500	1.000	0.370	-	0.995	0.635	-	0.765	0.515	-	
	1000	1.000	0.935	-	1.000	0.890	-	0.835	0.815	-	
X_{10}	300	1.000	*	-	1.000	*	-	0.995	0.035	-	
	500	1.000	*	-	1.000	*	-	1.000	*	-	
	1000	1.000	*	-	1.000	*	-	1.000	*	-	

Table 1: Results for model 1. The values show the rates of classification of a given covariate in the sets R , as a relevant covariate, and C , as a nonlinear covariate. The column $I(6, 7)$ reports the observed rates for the interaction term between the two covariates $X_{(6)}$ and $X_{(7)}$. The symbol “*” means that the value is ≤ 0.025 while “-” is used for “Not applied”.

		n	$d = 20$				$d = n/2$				$d = 2n$			
			R	C	$I(1, 2)$	$I(1, 7)$	R	C	$I(1, 2)$	$I(1, 7)$	R	C	$I(1, 2)$	$I(1, 7)$
X_1	300	1.000	*	0.365	0.675	1.000	*	0.305	0.700	0.995	*	0.165	0.705	
	500	1.000	*	0.635	0.910	1.000	*	0.555	0.905	1.000	*	0.475	0.930	
	1000	1.000	*	0.945	1.000	1.000	*	0.925	1.000	1.000	*	0.860	1.000	
X_2	300	0.945	*	-	-	0.855	*	-	-	0.560	*	-	-	
	500	1.000	*	-	-	0.985	*	-	-	0.795	*	-	-	
	1000	1.000	*	-	-	0.995	*	-	-	0.905	*	-	-	
X_7	300	1.000	0.645	-	-	1.000	0.665	-	-	0.970	0.645	-	-	
	500	1.000	0.910	-	-	1.000	0.880	-	-	0.990	0.875	-	-	
	1000	1.000	1.000	-	-	1.000	0.995	-	-	1.000	0.995	-	-	

Table 2: Results for model 2. The values show the rates of classification of a given covariate in the sets R , as a relevant covariate, and C , as a nonlinear covariate. The column $I(1, 2)$ reports the observed rates for the interaction term between the two covariates $X_{(1)}$ and $X_{(6)}$. The same is made for column $I(1, 7)$. The symbol “*” means that the result is ≤ 0.025 while “-” is used for “Not applied”.

where there is a nonlinear covariate mixed with a linear covariate and one linear covariate mixed with another linear covariate. Table 2 reports the results for this model.

From tables 1 and 2, we can note that all the values increase for increasing sample sizes, showing the consistency of our selection procedure. Of course, the performance is worse in the last column, where the dimension is remarkably greater than the sample size. Anyway, even in this case the performance is quite satisfactory for the nonlinear variable identification, whereas there is some difficulty in identifying the linear mixed effects.

In conclusion, we want to stress here that our selection procedure does not need any regularization parameter. Under this point of view, the implementation of the procedure is very simple. Moreover, the procedure gives a simultaneous classification of the relevant covariates between linear/nonlinear and pure/mixed, without the necessity of imposing an additive structure in the model and without using any stepwise selection method. Comparing the GRID procedure with the LASSO based procedures or the RODEO procedure, our procedure shows a very good performance notwithstanding the generality of the model (the results are available in [3]). On

the other side, this good performance is paid in terms of function estimation, because the GRID procedure does not give a simultaneous estimation of the regression function, like the LASSO based procedures. Anyway, the model structure identified through the GRID procedure can be used with a GAM estimator in order to obtain the final estimation of the regression function.

Bibliography

- [1] Bertin, K. and Lecue, G. (2008) *Selection of variables and dimension reduction in high-dimensional non-parametric regression*. Electronic Journal of Statistics, **2**, 1224–1241.
- [2] Chen, X.S., Peng, L. and Qin, Y.L. (2009) *Effects of data dimension on empirical likelihood*. Biometrika, **96**, 711–722.
- [3] Giordano, F., Lahiri, N.S and Parrella, M.L. (2014) *GRID for variable and model selection in high dimensional regression*. Working paper.
- [4] Lafferty, J. and Wasserman, L. (2008) *RODEO: sparse, greedy nonparametric regression*. The Annals of Statistics, **36**, 28–63.
- [5] Lu, Z.Q. (1996) *Multivariate locally weighted polynomial fitting and partial derivative estimation*. Journal of Multivariate Analysis, **59**, 187–205.
- [6] Masry, E. (1996) *Multivariate local polynomial regression for time series: uniform strong consistency and rates*. Journal of Time Series Analysis, **17**, 571–599.
- [7] Radchenko, P. and James, G.M. (2010) *Variable selection using adaptive nonlinear interaction structures in high dimensions*. Journal of American Statistical Association, **105**, 1541–1553.
- [8] Storlie, C.B., Bondell, H.D., Reich, B.J. and Zhang, H.H. (2011) *Surface estimation, variable selection, and the nonparametric oracle property*. Statistica Sinica, **21**, 679–705.
- [9] Ruppert, D. and Wand, P. (1994) *Multivariate locally weighted least squares regression*. Annals of Statistics, **22**, 1346–1370.
- [10] Zhang, H.H., Cheng, G. and Liu, Y. (2011) *Linear or nonlinear? Automatic structure discovery for partially linear models*. Journal of American Statistical Association, **106**, 1099–1112.

Unsupervised Learning using Gaussian Mixture Copula Model

Sakyajit Bhattacharya, *Xerox Research Centre India*, sakyajit.bhattacharya@xerox.com
Vaibhav Rajan, *Xerox Research Centre India*, vaibhav.rajan@xerox.com

Abstract. Gaussian Mixture Copula Models (GMCM) use Gaussian mixtures to model the dependence structure in data. They are useful in modeling heterogeneous multimodal data with complex dependence structures common in many real-world datasets. In this paper, we present a modified Expectation-Maximization algorithm for estimating the number of components and the parameters of a GMCM, along with a proof of its convergence. We demonstrate the efficacy of our algorithm, in clustering and unsupervised classification tasks, on a variety of simulated and real datasets.

Keywords. Gaussian Mixture Copula, Expectation Maximization, Clustering

1 Introduction

The Gaussian copula is widely used but is found to be of limited use in multimodal datasets with dependence across different modes. This was the motivation for proposing Gaussian mixture copulas [9] wherein the dependence is obtained from a Gaussian Mixture Model (GMM):

$$\mathcal{C}(\boldsymbol{\vartheta}, \mathbf{y}_i) = \frac{\sum_{g=1}^G \pi_g \phi(\mathbf{y}_i | \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g)}{\prod_{j=1}^p \psi_j(y_{ij})}.$$

Here we assume there are n observations $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n)$ with p dimensions each from a G component Gaussian Mixture Copula; $y_{ij} = \Psi_j^{-1}(u_{ij})$, are the inverse cumulative distribution values, where $u_{ij} = F_j(x_{ij})$ and F_j is the unknown marginal distribution for the j -th dimension, ψ_j is the marginal density of the GMM along the j -th dimension, ϕ is a multivariate Gaussian density and $\boldsymbol{\vartheta} = (\pi_1, \dots, \pi_G, \boldsymbol{\mu}_1, \dots, \boldsymbol{\mu}_G, \boldsymbol{\Sigma}_1, \dots, \boldsymbol{\Sigma}_G)$ is the (unknown) parameter set representing mixing proportions (π), mean vectors ($\boldsymbol{\mu}$) and covariance matrices ($\boldsymbol{\Sigma}$). Note that the log-likelihood of $\boldsymbol{\vartheta}$ given $(\mathbf{u}_1, \dots, \mathbf{u}_n)$ is $\log \mathcal{L}(\boldsymbol{\vartheta} | \mathbf{u}_1, \dots, \mathbf{u}_n) = \sum_{i=1}^n \log \mathcal{C}(\boldsymbol{\vartheta}, \mathbf{y}_i)$. The authors in [9] present a gradient-descent based heuristic to obtain a maximum-likelihood estimate of $\boldsymbol{\vartheta}$ which lacks theoretical justification and does not estimate the number of components in the mixture.

We present a novel Expectation Maximization (EM) based algorithm, along with a proof of its convergence, to estimate the parameters and number of components of GMCM. We demonstrate

the efficacy of our algorithm on a wide variety of simulated and real datasets. The algorithm significantly outperforms not only the previous gradient descent based method (for GMCM) but also other clustering methods such as K-Means, Spectral clustering and GMM-based clustering.

We refer the reader to [6] and [8] for more details on copulas and to [3] and [4] for extensive reviews on GMMs.

2 An Algorithm for Estimating the Parameters of GMCM

To design an EM Algorithm for a GMCM with the dataset $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n)$, we have to estimate the parameter set $\boldsymbol{\vartheta}$ that maximizes the log likelihood:

$$\log \mathcal{L}(\boldsymbol{\vartheta} \mid \mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_n) = \sum_{i=1}^n \log \frac{\sum_{g=1}^G \pi_g \phi(\mathbf{y}_i \mid \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g)}{\prod_{j=1}^p \psi_j(y_{ij})} = \sum_{i=1}^n \log \mathcal{C}(\boldsymbol{\vartheta}, \mathbf{y}_i) \quad (1)$$

where the notations are defined as in Section 1. Note that the copula mixture in the equation is defined in terms of the inverse distribution values ($y_{ij} = \Psi_j^{-1}(u_{ij})$) and not the inputs (as in GMM). The conventional EM algorithm of a GMM, where the inputs remain fixed in each iteration, cannot be directly used here as the inverse distribution values change in every iteration.

Our method, called EM-GMCM, is presented below. Superscript (t) on the parameters denotes values in the t -th iteration. The algorithm runs iteratively with the standard Expectation (E) and Maximization (M) steps and an additional step in each iteration. In this additional step, the inverse distribution values (y_{ij}) are estimated from the current parameter values. We derive an approximate expression for y_{ij} in terms of the CDF u_{ij} (in lemma 1 below). In addition, we apply another crucial condition on the inverse distribution values: in any iteration, if the computed y_{ij} value is greater than a predefined value Γ_{ij} (defined below), we set the value of y_{ij} to Γ_{ij} , as shown in the algorithm. This step is added to ensure that the log likelihood does not decrease at any step: the proof of convergence clarifies this choice of Γ_{ij} .

-
- **Initialize:** Standardize the matrix $X = (\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_p)$.

Set $\boldsymbol{\vartheta}^{(0)}$ randomly or by K-means clustering under the constraints that $\pi_g^{(0)} > 0$, $\sum_{g=1}^G \pi_g^{(0)} = 1$ and $\boldsymbol{\Sigma}_g^{(0)}$ is positive definite, and set $\delta_i = \text{Min}_{g,j} |y_{ij}^{(0)} - 2\kappa^{(0)} \left([\boldsymbol{\Sigma}_g^{(0)} + I]^{-1} \boldsymbol{\Sigma}_g^{(0)} \mathbf{1} \right)_j|$. Set $u_{ij} = F_j(x_{ij})$.

- **Repeat** the following steps **until** $\|L^{(t+1)} - L^{(t)}\| < \gamma$.

- Set $y_{ij}^{(t)} = \min \left(\Lambda_{ij}^{(t)}, \Gamma_{ij}^{(t)} \right)$ where $\Lambda_{ij}^{(t)} = \left(\sum_{g=1}^G \frac{\pi_g^{(t)}}{\sigma_{g,jj}^{(t)}} \right)^{-1} \left[u_{ij} + \frac{1}{\sqrt{2\pi}} \sum_{g=1}^G \frac{\pi_g^{(t)} \mu_{gj}^{(t)}}{\sigma_{g,jj}^{(t)}} - \frac{1}{2} \right]$,

$$\Gamma_{ij}^{(t)} = \kappa^{(t)} \left(\left[S_i^{(t)} + I \right]^{-1} S_i^{(t)} \mathbf{1} \right)_j - \frac{\delta_i}{2} \left(3 - \frac{p}{m_i^{(t)} + p} \right), \quad \kappa^{(t)} = \max_{g,j} (\mu_{gj}^{(t)}), \quad S_i^{(t)} = \sum_{g=1}^G z_{ig}^{(t-1)} \boldsymbol{\Sigma}_g^{(t)},$$

and $m_i^{(t)}$ is the sum of all elements of $S_i^{(t)}$.

- **E-Step:** $z_{ig}^{(t)} = \frac{\pi_g^{(t)} \phi(\mathbf{y}_i^{(t)} \mid \boldsymbol{\mu}_g^{(t)}, \boldsymbol{\Sigma}_g^{(t)})}{\sum_{g=1}^G \pi_g^{(t)} \phi(\mathbf{y}_i^{(t)} \mid \boldsymbol{\mu}_g^{(t)}, \boldsymbol{\Sigma}_g^{(t)})}$.

- **M-Step:** $\pi_g^{(t+1)} = \frac{\sum_{i=1}^n z_{ig}^{(t)}}{n}$, $\boldsymbol{\mu}_g^{(t+1)} = \frac{\sum_{i=1}^n z_{ig}^{(t)} \mathbf{y}_i^{(t)}}{\sum_{i=1}^n z_{ig}^{(t)}}$, $\boldsymbol{\Sigma}_g^{(t+1)} = \frac{\sum_{i=1}^n z_{ig}^{(t+1)} (\mathbf{y}_i^{(t)} - \boldsymbol{\mu}_g^{(t+1)})^T (\mathbf{y}_i^{(t)} - \boldsymbol{\mu}_g^{(t+1)})}{\sum_{i=1}^n z_{ig}^{(t)}}$

- **Likelihood:** $\mathcal{L}^{(t+1)} = \prod_{i=1}^n \sum_{g=1}^G \pi_g^{(t+1)} \frac{1}{\sqrt{\det(2\pi \boldsymbol{\Sigma}_g^{(t+1)})}} \times \exp -\frac{1}{2} (\mathbf{y}_i^{(t)} - \boldsymbol{\mu}_g^{(t+1)})^T \boldsymbol{\Sigma}_g^{(t+1)^{-1}} (\mathbf{y}_i^{(t)} - \boldsymbol{\mu}_g^{(t+1)})$

Note that, since F_j is unknown, it is generally estimated via non-parametric methods (see [9]), if necessary for the application. The running time, in addition to the standard E and

M steps is mainly due to the matrix inversion required to compute the value of $\mathbf{\Gamma}$, thus a cubic order computation in each iteration. To reduce the computational cost of the estimation, special families of covariance structures have been introduced that impose constraints upon the constituent parts of the decomposition of the component covariance matrices such as the Parsimonious Gaussian Mixture (PGMM) family [5], which we use in this work. Once the general covariance matrix is estimated, as shown in the algorithm, noise and factor loading matrices for the corresponding PGMM family can be estimated as described in [5]. For a family of mixture models, we select the model having the maximum Bayesian Information Criteria ([7]), given by $\text{BIC} = 2 \log \mathcal{L}(\hat{\boldsymbol{\theta}} \mid \mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_n) - \rho \log n$. Note that this criterion is approximate because the parameter values used are from the final iteration of the EM-GMCM algorithm.

3 Theoretical Analysis

In this section we derive the approximation to the inverse distribution values that is used in our algorithm and discuss the accuracy of the approximation (lemma 1). We then analyze and prove the convergence of our algorithm (theorem 1).

Lemma 1.

For each i and j , with probability more than 0.9975, y_{ij} can be approximated as

$$y_{ij} \approx \left(\sum_{g=1}^G \frac{\pi_g}{\sqrt{\sigma_{g,jj}}} \right)^{-1} \left[u_{ij} + \frac{1}{\sqrt{2\pi}} \sum_{g=1}^G \frac{\pi_g \mu_{gj}}{\sqrt{\sigma_{g,jj}}} - \frac{1}{2} \right].$$

Proof Sketch. Using the definition of y_{ij} , we have $u_{ij} = \Psi_j(y_{ij}) = \sum_{g=1}^G \pi_g \Phi(y_{ij} \mid \mu_{gj}, \sigma_{g,jj})$. Using Taylor series expansion around the mean upto the second derivative, we have $\Psi_j(y_{ij}) = \sum_{g=1}^G \pi_g \Phi(z \mid \mu_{gj}, \sigma_{g,jj}) \approx \sum_{g=1}^G \pi_g \sum_{k=0}^{\infty} (y_{ij} - \mu_{gj})^k / k! \partial^k \Phi(x \mid \mu_{gj}, \sigma_{g,jj}) / \partial x^k \big|_{x=\mu_{gj}}$. It is to be noted that the above Taylor series expansion consists upto the fourth term, because all the subsequent terms are 0. Let us denote the fourth term as λ . Mathematical calculations and use of the 3- σ rule show that, with probability 0.9975, $|\lambda| \leq \kappa_1 \sum_{g=1}^G \pi_g \sigma_{g,jj}^{-1/2}$ where $\kappa_1 = 1.7948$. Using the update $\pi_g^{(t)}$ and $\sigma_{g,jj}^{(t)}$ as shown in the M-step of our algorithm, we conclude that at each iteration, with probability 0.9975, $|\lambda| \leq \kappa_1 \sum_{g=1}^G \sum_{i=1}^n z_{ig}^{(t)2} / n^2 \rightarrow 0$ as $n \rightarrow \infty$, since this is of order n^{-1} . Note the main idea behind this approximation: even when two means are far from each other, and any point is distant from either of them, after dividing by the standard deviation, the distance is bounded with very high probability.

We only keep the remaining terms until the first derivative because the second derivative of the above Taylor series is 0 owing to the fact that a Gaussian distribution has equal mean, median and mode. Thus we have $\Psi_j(y_{ij}) = 1/2 + \sum_{g=1}^G \pi_g [(y_{ij} - \mu_{gj}) \phi(\mu_{gj} \mid \mu_{gj}, \sigma_{g,jj})]$. From this expression, and using the fact that $\phi(\mu_{gj} \mid \mu_{gj}, \sigma_{g,jj}) = (2\pi\sigma_{g,jj})^{-1/2}$, the lemma is proved. \square

Proof of Convergence

We first prove a technical lemma that is used in the final proof of convergence. In the following, \mathbf{z}_i is the i -th latent variable; $z_{ig} = 1$ if the i -th observation belongs to the g -th cluster, and 0 otherwise. We use superscript t to denote the t -th iteration of our algorithm.

Lemma 2.

If $\text{Max}_j | y_{ij}^{(t+1)} - y_{ij}^{(t)} | \leq \delta_i$, then

$$\log \mathcal{C}(\boldsymbol{\vartheta}^{(t+1)}, \mathbf{y}_i^{(t+1)} | \mathbf{z}_i^{(t)}) \geq \log \mathcal{C}(\boldsymbol{\vartheta}^{(t+1)}, \mathbf{y}_i^{(t)} | \mathbf{z}_i^{(t)}). \tag{2}$$

Proof Sketch. Since z_{ig} is not known we use the update of \mathbf{z}_i as stated in our algorithm. Thus the complete-likelihood of $\boldsymbol{\vartheta}^{(t)}$ at the t -th iteration (i.e. with $\mathbf{y}^{(t)}$) is

$$\prod_{i=1}^n \prod_{g=1}^G \left[\pi_g^{(t+1)} \phi(\mathbf{y}_i^{(t)} | \boldsymbol{\mu}_g^{(t+1)}, \boldsymbol{\Sigma}_g^{(t+1)}) / \prod_{j=1}^p \psi_j(y_{ij}^{(t)}) \right]^{z_{ij}^{(t)}} \text{ and similarly for } \boldsymbol{\vartheta}^{(t)} \text{ with } \mathbf{y}^{(t+1)} .$$

First, note that for any two real numbers x and y , $y^2 - x^2 = (y - x)^2 + 2x(y - x) \geq 2x(y - x)$. So, if $|y - x| \leq \epsilon$, $y^2 - x^2 \geq -2x\epsilon$. We shall use this inequality in the following. For notational simplicity, we shall use ϵ instead of ϵ_i .

Suppose $|y_{ij}^{(t+1)} - y_{ij}^{(t)}| \leq \epsilon$ for all j . We can set $\epsilon = \text{Max}_j |y_{ij}^{(t+1)} - y_{ij}^{(t)}|$. In such a case,

$$\begin{aligned} & \log \mathcal{C}(\boldsymbol{\vartheta}^{(t+1)}, \mathbf{y}_i^{(t+1)} | \mathbf{z}_i^{(t)}) - \log \mathcal{C}(\boldsymbol{\vartheta}^{(t+1)}, \mathbf{y}_i^{(t)} | \mathbf{z}_i^{(t)}) \\ & \geq -\epsilon^2 \mathbf{1}^T \left(\sum_{g=1}^G z_{ig}^{(t)} \boldsymbol{\Sigma}_g^{(t+1)-1} \right) \mathbf{1} - 2\epsilon \mathbf{1}^T \sum_{g=1}^G z_{ig}^{(t)} \boldsymbol{\Sigma}_g^{(t+1)-1} (\mathbf{y}_i^{(t)} - \boldsymbol{\mu}_g^{(t+1)}) - 2\epsilon \sum_{j=1}^p y_{ij}^{(t)} \\ & \geq -\epsilon^2 \mathbf{1}^T S_i^{(t+1)} \mathbf{1} - 2\epsilon \mathbf{1}^T (S_i^{(t+1)} + I) \mathbf{y}_i^{(t)} + 2\epsilon \kappa^{(t+1)} \mathbf{1}^T S_i^{(t+1)} \mathbf{1} \quad \left(\text{since } \sum_{g=1}^G z_{ig} = 1 \right). \end{aligned}$$

Equating the above to zero and rearranging the terms, we note that to satisfy inequality 2, we must have $\epsilon \leq \mathcal{T}$ where $\mathcal{T} = 2\kappa^{(t+1)} \mathbf{1}^T S_i^{(t+1)} \mathbf{1} - 2 \mathbf{1}^T (S_i^{(t+1)} + I) \mathbf{y}_i^{(t+1)} / 3 \mathbf{1}^T S_i^{(t+1)} \mathbf{1} + 2p$. From the lower bound (Γ_{ij}) we set for y_{ij} in our algorithm, we obtain a bound on the numerator of \mathcal{T} : $\kappa^{(t+1)} \mathbf{1}^T S_i^{(t+1)} \mathbf{1} - \mathbf{1}^T (S_i^{(t+1)} + I) \mathbf{y}_i^{(t+1)} \geq (3m_i^{(t+1)} + 2p)\delta_i$. We also know that the denominator, $3 \mathbf{1}^T S_i^{(t+1)} \mathbf{1} + 2p = 3m_i^{(t+1)} + 2p$. From the above two expressions, we see that δ_i is lesser than \mathcal{T} . Thus, the condition $\epsilon \leq \delta_i$ is sufficient for inequality 2 to hold. This completes the proof. \square

We now prove that in our algorithm the likelihood increases at each iteration after a specific iteration t_0 . Convergence of our algorithm then follows from the convergence of the EM algorithm.

Theorem 1.

There exists t_0 such that $\mathcal{L}(\boldsymbol{\vartheta}^{(t+1)} | \mathbf{u}_1, \dots, \mathbf{u}_n) \geq \mathcal{L}(\boldsymbol{\vartheta}^{(t)} | \mathbf{u}_1, \dots, \mathbf{u}_n)$ for $t \geq t_0$.

Proof Sketch. We compare the complete data log-likelihoods taking into account both the observed \mathbf{y} values and the latent variables z_{ig} which is 1 when the i -th observation belongs to group g and 0 otherwise. We know that, if $\log \mathcal{C}(\boldsymbol{\vartheta}^{(t+1)}, \mathbf{y}_i^{(t+1)} | \mathbf{z}_i^{(t)}) \geq \log \mathcal{C}(\boldsymbol{\vartheta}^{(t)}, \mathbf{y}_i^{(t)} | \mathbf{z}_i^{(t)})$, then, marginalizing over the latent variables and summing over all the observations, we obtain $\log \mathcal{L}(\boldsymbol{\vartheta}^{(t+1)} | \mathbf{u}_1, \dots, \mathbf{u}_n) = \sum_{i=1}^n \sum_{\mathbf{z}_i^{(t)}} \log \mathcal{C}(\boldsymbol{\vartheta}^{(t+1)}, \mathbf{y}_i^{(t+1)} | \mathbf{z}_i^{(t)}) \geq \sum_{i=1}^n \sum_{\mathbf{z}_i^{(t)}} \log \mathcal{C}(\boldsymbol{\vartheta}^{(t)}, \mathbf{y}_i^{(t)} | \mathbf{z}_i^{(t)}) = \log \mathcal{L}(\boldsymbol{\vartheta}^{(t)} | \mathbf{u}_1, \dots, \mathbf{u}_n)$.

First, note that our approximation for y_{ij} is a uniformly continuous function of μ_{gj} for each g as evident from lemma 1. Also, from the convergence of EM algorithm we know that as $t \rightarrow \infty$, $|\mu_{gj}^{(t+1)} - \mu_{gj}^{(t)}| \rightarrow 0$. This fact, coupled with uniform continuity leads to the property that for a given ϵ and for each i , $\exists t_{0i} | \forall t \geq t_{0i}, \text{Max}_j |y_{ij}^{(t+1)} - y_{ij}^{(t)}| \leq \epsilon$. So, for $\epsilon \leq \delta_i$, we obtain

a t_{0i} that satisfies the above property. But Lemma 2 shows that if $\epsilon \leq \delta_i$ then Equation 2 holds. Thus by choosing $t_0 = \text{Max}_i(t_{0i})$ we claim that, for all $t \geq t_0$, $\sum_{i=1}^n \log \mathcal{C}(\boldsymbol{\vartheta}^{(t+1)}, \mathbf{y}_i^{(t+1)} \mid \mathbf{z}_i^{(t)}) \geq \sum_{i=1}^n \log \mathcal{C}(\boldsymbol{\vartheta}^{(t+1)}, \mathbf{y}_i^{(t)} \mid \mathbf{z}_i^{(t)})$. This fact, and the convergence properties of the conventional EM algorithm, after marginalizing over the latent variables, show that for all $t \geq t_0$, $\log \mathcal{L}(\boldsymbol{\vartheta}^{(t+1)} \mid \mathbf{u}_1, \dots, \mathbf{u}_n) \geq \sum_{i=1}^n \log \mathcal{C}(\boldsymbol{\vartheta}^{(t+1)}, \mathbf{y}^{(t)}) \geq \log \mathcal{L}(\boldsymbol{\vartheta}^{(t)} \mid \mathbf{u}_1, \dots, \mathbf{u}_n)$ which completes the proof. \square

4 Simulation Studies

We empirically evaluate the performance of our estimation algorithm, in terms of its ability to estimate the number of clusters and classify the data into meaningful clusters. We generate 200 simulations, each with 2-dimensional data points in four clusters. The clusters are chosen such that there is dependency within each cluster and data for each cluster is chosen from different distributions.

Each data point is a product of a sample from a Multivariate Normal (MVN) distribution and another distribution as outlined in table 1. We intentionally choose the same MVN parameters in clusters 1 and 4, to study whether our algorithm can accurately distinguish the two clusters. In each simulation we generate the clusters by sampling from the distributions, f_i . Note that in the case of f_3 , the full covariance matrix Δ needs to be estimated by EM-GMM and EM-GMCM. In the full covariance matrix Δ , autocorrelation is taken to be 0.9 in order to have a strong linear dependence structure between the features.

$f_1 = \text{MVN}(-5.5, I_2/2) \times \text{Unif}(0, 1)$
$f_2 = \text{MVN}(2, D_2) \times t(\text{df} = 9)$
$f_3 = \text{MVN}(3, \Delta) \times C(\text{loc} = 0, \text{sc} = 1)$
$f_4 = \text{MVN}(-5.5, I_2/2) \times \Gamma(\text{sh} = 0.5, \text{rt} = 1)$

Table 1: Parameter settings in each cluster; f_i : distribution in cluster $i = 1, 2, 3, 4$, I_2 : 2×2 Identity matrix with, D_2 : 2×2 diagonal matrix with unequal diagonal elements and Δ : matrix with (i, j) -th element $0.9^{|i-j|}$. Abbreviations – C: Cauchy, loc: location, sc: scale, sh: shape, rt: rate.

Sim	$ c_1 $	$ c_2 $	$ c_3 $	$ c_4 $	Data size
I	400	300	300	500	1500
II	300	500	300	400	1500
III	300	300	500	400	1500
IV	500	400	300	300	1500
V	1200	900	900	1500	4500
VI	900	1500	900	1200	4500
VII	900	900	1500	1200	4500
VIII	1500	1200	900	900	4500

Table 2: Parameter settings for each simulation set: $|c_i|$ denotes size of cluster c_i for $i = 1, 2, 3, 4$.

We vary the data size (the total number of data points) as well as the relative sizes of the clusters as shown in table 2 and generate 25 simulations for each set of parameter values (simulation sets I through VIII). In each simulation set, the ratio of cluster sizes (e.g. 4:3:3:5 in set I) are chosen such that the proportion of data points is higher in two clusters and the remaining data is equally divided in the other two clusters. This is done to generate different variance-covariance structures on varying sample sizes.

We denote our new algorithm by EM-GMCM, the EM algorithm for GMMs by EM-GMM and the previous gradient descent based method of Tewari et al [9] by HEU-T. Figure 1 shows the accuracy of both EM-GMCM and EM-GMM in determining the right number of clusters ($G = 4$). We assume the range of G (the number of clusters) to be $(2, 3, 4, 5)$ for both the mixture models. Accuracy is measured as the proportion of the total number of times (25),

the algorithm detects the right number of clusters. We observe that our algorithm outperforms EM-GMM in every run. Notice that in simulation sets III and VII, the performance of both the algorithms deteriorates. This is because in these two settings (see tables 1 and 2) the component corresponding to distribution f_3 (with full covariance matrix Δ) contains the maximum number of datapoints. With limited data size and the “curse” of a large number of parameters, BIC tends to prefer smaller number of clusters where the number of parameters is low.

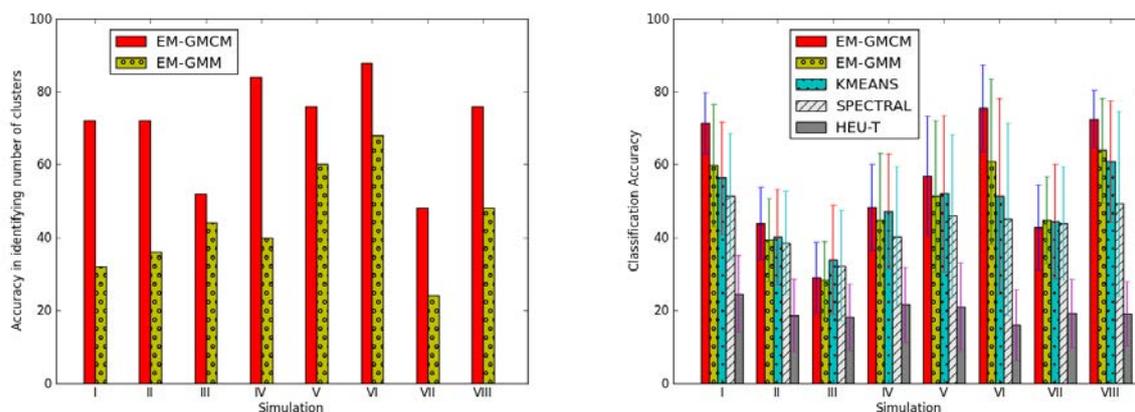


Figure 1: Left: Accuracy (in percentage) over 25 runs in determining the right number of clusters ($G = 4$) of EM-GMCM and EM-GMM for each simulation set. Right: Average classification accuracy (in percentage) over 25 runs for EM-GMCM, EM-GMM, HEU-T, Spectral and K-means clustering for each simulation set. The error bars indicate respective standard deviations.

Figure 1 also shows the classification accuracy of our algorithm in comparison with EM-GMM, HEU-T, Spectral clustering and K-means. Classification accuracy is measured by the proportion of the data points accurately identified in the right clusters. Note that the number of clusters ($G = 4$) is provided as input to HEU-T, Spectral and K-means algorithms. To make a fair comparison, we average the results over those simulations where both EM-GMCM and EM-GMM algorithms correctly identify the number of clusters.

Algorithm EM-GMCM has the lowest variance and outperforms K-means in six out of the eight settings. In seven out of eight settings it is better than EM-GMM, which in general performs poorly. The worst performer is HEU-T, which produces the lowest accuracy in most of the datasets we tested. As discussed earlier, EM-GMCM deteriorates in cases III and VII where the average accuracy is surpassed by K-means by a considerable margin.

5 Experiments with Real Data

We test our new algorithm, comparing it with other algorithms, on binary classification tasks in two datasets from different domains. One is an image dataset from physical sciences and another is a clinical dataset from healthcare, both from the UCI repository [1]. We measure the accuracy of the classification as the proportion of datapoints correctly predicted in the cluster: we choose the label of an estimated cluster to be the label of the majority of its constituent data points.

The first dataset used is the Cleveland Heart Disease dataset from the UCI repository [1]. We extract five numerical features (age, resting blood pressure, serum cholesterol, maximum heart rate, ST depression induced by exercise relative to rest) for 297 individuals in the dataset. The task is to classify the individuals into two groups: those with and those without heart disease. To initialize EM-GMCM and EM-GMM algorithms, we set the range of G (the number of clusters) and q (number of latent factors as defined in [5]) in the algorithms to (1, 2, 3) and (1, 2) respectively. For algorithms which do not estimate the number of clusters, such as K-means, Spectral clustering and HEU-T, we provide the correct number of clusters (2) as input.

While EM-GMCM detects the correct number of clusters, $G = 2$, with the CCU dependency structure and $q = 1$, EM-GMM incorrectly predicts 3 clusters with UCU dependence structure and $q = 2$ where the dependency structures are defined as in [5]. The accuracy achieved by five algorithms tested and the corresponding classification tables are shown in table 3. EM-GMCM shows the best results, outperforming the second best method, K-Means, by 13% in accuracy.

Algorithm	EM-GMCM	EM-GMM	K-Means	Spectral	HEU-T
Accuracy	70%	15%	57%	35%	50%

EM-GMCM		EM-GMM			K-Means		Spectral		HEU-T						
<i>A</i>	<i>B</i>	<i>A</i>	<i>B</i>	<i>C</i>	<i>A</i>	<i>B</i>	<i>A</i>	<i>B</i>	<i>A</i>	<i>B</i>					
<i>A</i>	137	23	<i>A</i>	19	69	72	<i>A</i>	110	50	<i>A</i>	64	96	<i>A</i>	114	46
<i>B</i>	67	70	<i>B</i>	12	26	99	<i>B</i>	76	61	<i>B</i>	98	39	<i>B</i>	103	34

Table 3: Above: Classification accuracy of the algorithms tested on the Cleveland Heart Disease Data. Below: Classification tables of the best models chosen by EM-GMCM and EM-GMM and of K-means, Spectral clustering and HEU-T for the Cleveland Heart data. *A*: group without heart disease, *B*: group with heart disease. EM-GMM erroneously estimates a third group *C*. True labels: horizontal, Estimated labels: vertical.

Algorithm	EM-GMCM	EM-GMM	K-Means	Spectral	HEU-T
Accuracy	72%	59%	57%	54%	51%

EM-GMCM		EM-GMM		K-Means		Spectral		HEU-T						
<i>A</i>	<i>B</i>	<i>A</i>	<i>B</i>	<i>A</i>	<i>B</i>	<i>A</i>	<i>B</i>	<i>A</i>	<i>B</i>					
<i>A</i>	814	186	<i>A</i>	731	269	<i>A</i>	857	143	<i>A</i>	1000	0	<i>A</i>	668	332
<i>B</i>	370	630	<i>B</i>	542	458	<i>B</i>	712	288	<i>B</i>	993	7	<i>B</i>	645	355

Table 4: Above: Classification accuracy of the algorithms tested on the Gamma Telescope Data. Below: Classification tables of the best models chosen by EM-GMCM and EM-GMM and of K-means, Spectral clustering and HEU-T. *A*: gamma signal, *B*: hadron showers (background). True labels: horizontal, Estimated labels: vertical.

The second dataset consists of features extracted from a preprocessed image of reconstructed radiation showers. Cherenkov radiation (of visible to UV wavelengths) leaks through the atmosphere and gets recorded in a detector, allowing reconstruction of the gamma signal. The task is to statistically discriminate between two signals within the image: the primary gamma signal in Cherenkov radiation and background hadronic shower signal from cosmic rays in the upper atmosphere. The signals on the detector are processed to form the final image which is typically

elliptical in shape. Ten numerical geometric features of the image are used. We select 2000 labeled samples from the dataset where 1000 samples belong to the gamma signal and 1000 belong to the background hadronic showers. We set the range of G (the number of clusters) and q (number of latent factors) in the algorithms to $(1, 2, 3)$ and $(1, 2, 3, 4)$ respectively.

Both EM-GMCM and EM-GMM detect the correct number of clusters $G = 2$ with the CCC dependency structure and number of latent factors, $q = 3$. The accuracy achieved by five algorithms tested and the corresponding classification tables shown in table 4. Here also, EM-GMCM outperforms all the other clustering methods with a difference in accuracy of at least 13%.

6 Concluding Remarks

We present a new algorithm, EM-GMCM, for estimating the parameters of a Gaussian Mixture Copula Model, and a proof of its convergence. Our algorithm works well with clusters that are not well separated, when the components are not Gaussian and when there are dependencies between the components: cases where GMM-based methods often fail. We demonstrate its efficacy in discovering the number of clusters and in unsupervised classification. In our experiments EM-GMCM significantly outperforms EM-GMM, K-Means, Spectral clustering and the previous GMCM-based heuristic. An implementation of our algorithm in R is available upon request.

Two limitations of our method, that are also limitations in the case of GMMs, are its inability to efficiently deal with non-numerical and high dimensional data. In the future we would like to address these limitations. The use of penalized log-likelihood that leads to consistent model selection criteria for high-dimensional cases would be worth exploring. Using Lasso-based penalization and its variants leading to an Oracle-proof criterion [2] could also be investigated.

Bibliography

- [1] K. Bache and M. Lichman. UCI machine learning repository, 2013.
- [2] J. Fan and R. Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456):1348–1360, 2001.
- [3] G. McLachan and K. Basford. *Mixture Models: Inference and Applications to Clustering*. Marcel Dekker Inc., 1988.
- [4] G. McLachan and D. Peel. *Finite Mixture Model*. John Wiley & Sons, Inc, 2002.
- [5] P. D. McNicholas and T. Murphy. Parsimonious Gaussian mixture models. *Statistics and Computing*, 18(3):285–296, 2008.
- [6] R. B. Nelsen. *An Introduction to Copulas (Lecture Notes in Statistics)*. Springer, 1998.
- [7] G. Schwartz. Estimating the dimensions of a model. *Annals of Statistics*, 6:461–464, 1978.
- [8] A. Sklar. Fonctions de rpartition n dimensions et leurs marges. *Publ. Inst. Statist. Univ. Paris*, 8:229–231, 1959.
- [9] A. Tewari, M. J. Giering, and A. Raghunathan. Parametric characterization of multimodal distributions with non-gaussian modes. In *Proceedings of the 2011 IEEE 11th International Conference on Data Mining Workshops*, pages 286–292, 2011.

A comparison of some estimation methods for latent Markov models with covariates

Francesco Bartolucci, *University of Perugia (IT)*, bart@stat.unipg.it

Giorgio E. Montanari, *University of Perugia (IT)*, giorgio.montanari@unipg.it

Silvia Pandolfi, *University of Perugia (IT)*, pandolfi@stat.unipg.it

Abstract. We compare different estimation methods for latent Markov models with covariates. These models represent a powerful tool for the analysis of longitudinal categorical data when the interest is to represent the evolution of a latent characteristic of a sample of units over time. In applications to complex data, with a large number of observed response variables and latent states, estimation of these models may present some critical aspects. These are mainly due to the presence of many local maxima of the model log-likelihood and to the slowness to converge of the Expectation-Maximization algorithm, which is typically used for parameter estimation. In such a context, alternative methods which allow us to overcome the drawbacks of the full maximum likelihood approach, with an advantage also in terms of computational cost, are of interest. In particular, we focus on estimation methods which may be seen as modified versions of the three-step approach for the latent class model with covariates. The behavior of these alternative approaches is investigated by means of a Monte Carlo simulation study on the basis of a wide set of model specifications.

Keywords. Expectation-Maximization algorithm, hidden Markov models, latent class models

1 Introduction

Latent Markov (LM) models [1] represent a powerful tool for the analysis of longitudinal categorical data. These models have a great potential of application in several fields, such as psychology, economics, and medicine, where the characteristic of interest is not directly observable (e.g., quality-of-life, health conditions, etc.). Under these models, occasion specific response variables are assumed to depend only on a discrete latent variable which follows a first order Markov chain. LM models are also related to the latent class (LC) model, which is typically used to cluster subjects on the basis of a series of categorical response variables. In particular,

an LM model may be seen as an extension of the LC model in which the subjects are allowed to move between the latent classes during the period of observation. An important extension of this class of models consists in the inclusion of individual covariates so that they affect the distribution of the latent process. Under this formulation, we assume that the response variables measure and depend on a latent trait, which may evolve over time. Then, we are interested in modeling the effect of the covariates on the latent trait distribution.

In such a context, and in application to complex and high-dimensional data characterized by a large number of response variables and latent states, estimation of these models may present some critical aspects. In particular, the model likelihood may present many local maxima, requiring an excessive number of different starting values to find the global maximum. This problem is very similar to that discussed by [6] in the context of finite mixture models. Moreover, the full maximum likelihood estimation process, typically based on the Expectation-Maximization (EM) algorithm [3, 5], may be particularly slow to converge, as observed in real applications.

Aim of the paper is to investigate the behavior of alternative estimation methods for this class of models, which may be useful when the full maximum-likelihood (FML) approach is difficult to implement. All these methods may be seen as modified versions of the three-step estimation method for the LC model with covariates of [7]. When applied to LM models, the first step consists of a preliminary fitting of the basic LC model, in which the responses of the same unit to different time occasions are considered as coming from separate units. Since the latent state to which every sample unit is assigned may change across time, it is also possible to estimate the parameters of the latent Markov process. Then, the alternative methods under comparison are based on different specifications of the steps beyond the first. In particular, we compare the performance of the three-step estimation (3S) method and its improved version (3S-IMP) introduced by [2]. We also propose to extend, to LM models, the modified three-step approach based on the method proposed by [4] (named BCH approach) which is aimed at taking into account the classification error introduced in the allocation of the units to latent classes; see also [7]. We finally consider a direct two-step (2S) approach which is based on performing a standard EM algorithm while keeping the estimated conditional response probabilities fixed from the first step. These methods may be seen as a stable alternative of the FML approach, since they allow us to decompose the maximization problem into simpler subproblems for which there is a higher chance to find the global maximum. The discussed estimation methods are also typically faster than the FML approach.

In the following sections we review the estimation methods under comparison and we evaluate their performance on the basis of a Monte Carlo simulation study.

2 Latent Markov model with individual covariates

Let consider the multivariate case, in which we observe a vector $\mathbf{Y}^{(t)}$ of r categorical response variables, $Y_j^{(t)}$, with c_j categories, labeled from 0 to $c_j - 1$, $j = 1, \dots, r$, which are available at the t -th time occasion, $t = 1, \dots, T$. Let also $\tilde{\mathbf{Y}}$ be the vector made up of the union of the vectors $\mathbf{Y}^{(t)}$ which has rT elements. When available, we also denote by $\mathbf{X}^{(t)}$ the vector of individual covariates at the t -th time occasion and by $\tilde{\mathbf{X}}$ the vector of all covariates obtained by stacking the vectors $\mathbf{X}^{(1)}, \dots, \mathbf{X}^{(T)}$.

The general formulation of the model assumes the existence of a latent process, denoted by $\mathbf{U} = (U^{(1)}, \dots, U^{(T)})$, which affects the distribution of the response variables. Such a process is

assumed to follow a first-order Markov chain with state space $\{1, \dots, k\}$, where k is the number of latent states. Under the *local independence* assumption, the response vectors $\mathbf{Y}^{(1)}, \dots, \mathbf{Y}^{(T)}$ are assumed to be conditional independent given the latent process \mathbf{U} . Moreover, the elements $Y_j^{(t)}$ within $\mathbf{Y}^{(t)}$, $t = 1, \dots, T$, are conditionally independent given $U^{(t)}$. Parameters of the measurement model, that is, the distribution of the response variables given the latent process, are the conditional response probabilities

$$\phi_{jy|u} = p(Y_j^{(t)} = y | U^{(t)} = u), \quad j = 1, \dots, r, \quad y = 0, \dots, c_j - 1, \quad u = 1, \dots, k, \quad t = 1, \dots, T.$$

Note that we assume the hypothesis that the measurement model is time-homogeneous, that is, $\phi_{jy|u}^{(t)} = \phi_{jy|u}$, $t = 1, \dots, T$, according to which the distribution of the responses depends only on the corresponding latent variable and there is no dependence on time. These probabilities are equal for all subjects and are collected in the matrix Φ_j of dimension $c_j \times k$, for $j = 1, \dots, r$.

In this paper, we consider the case in which the covariates are included in the latent model. In this context, the assumption of local independence and the assumption that the latent process is of first order still hold. Moreover, the initial and transition probabilities depend on the individual covariates through a multinomial logit parametrization

$$\log \frac{p(U^{(1)} = u | \mathbf{X}^{(1)} = \mathbf{x})}{p(U^{(1)} = 1 | \mathbf{X}^{(1)} = \mathbf{x})} = \beta_{0u} + \mathbf{x}'\boldsymbol{\beta}_{1u}, \quad u \geq 2, \tag{1}$$

$$\log \frac{p(U^{(t)} = u | U^{(t-1)} = \bar{u}, \mathbf{X}^{(t)} = \mathbf{x})}{p(U^{(t)} = \bar{u} | U^{(t-1)} = \bar{u}, \mathbf{X}^{(t)} = \mathbf{x})} = \gamma_{0\bar{u}u} + \mathbf{x}'\boldsymbol{\gamma}_{1\bar{u}u}, \quad t \geq 2, \quad \bar{u} \neq u. \tag{2}$$

In the above expressions, $\boldsymbol{\beta}_u = (\beta_{0u}, \boldsymbol{\beta}'_{1u})'$ and $\boldsymbol{\gamma}_{\bar{u}u} = (\gamma_{0\bar{u}u}, \boldsymbol{\gamma}'_{1\bar{u}u})'$ are parameter vectors to be estimated which are collected in the vector $\boldsymbol{\beta}$ and in the matrix $\boldsymbol{\Gamma}$.

When we deal with individual covariates, the *manifest distribution* of the response variables corresponds to the conditional distribution of $\tilde{\mathbf{Y}}$ given $\tilde{\mathbf{X}}$, which we denote by $p(\tilde{\mathbf{y}}|\tilde{\mathbf{x}})$. It is important to note that computing $p(\tilde{\mathbf{y}}|\tilde{\mathbf{x}})$ involves a sum extended to all the possible k^T configurations of the vector \mathbf{u} ; this typically requires a considerable computational effort. In order to efficiently compute such a probability, we can use a forward recursion as in [3]; see [1] for details.

3 Estimation of latent Markov models with covariates

In the presence of individual covariates, the observed data correspond to the vectors $\tilde{\mathbf{x}}_i$ and $\tilde{\mathbf{y}}_i$, for $i = 1, \dots, n$. The vector of covariates $\tilde{\mathbf{x}}_i$ may be decomposed into the time-specific subvectors $\mathbf{x}_i^{(1)}, \dots, \mathbf{x}_i^{(T)}$, whereas $\tilde{\mathbf{y}}_i$ is made up of the subvectors $\mathbf{y}_i^{(1)}, \dots, \mathbf{y}_i^{(T)}$. Then, the model log-likelihood assumes the following expression

$$\ell(\boldsymbol{\theta}) = \sum_{i=1}^n \log p(\tilde{\mathbf{y}}_i | \tilde{\mathbf{x}}_i),$$

where $\boldsymbol{\theta}$ is the vector of all free parameters affecting $p(\tilde{\mathbf{y}}_i | \tilde{\mathbf{x}}_i)$. In the following, we briefly introduce the estimation methods under comparison.

FML estimation method

The log-likelihood function can be maximized through the FML estimation method which is typically performed by means of the EM algorithm [5]. The latter is based on the complete data log-likelihood that, for the multivariate categorical data, has the following expression

$$\begin{aligned} \ell^*(\boldsymbol{\theta}) = & \sum_{i=1}^n \left[\sum_{j=1}^r \sum_{t=1}^T \sum_{u=1}^k \sum_{y=0}^{c_j-1} a_{ijuy}^{(t)} \log \phi_{jy|u} + \sum_{u=1}^k b_{iu}^{(1)} \log p(U^{(1)} = u | \mathbf{x}_i^{(1)}) \right. \\ & \left. + \sum_{t=2}^T \sum_{\bar{u}=1}^k \sum_{u=1}^k b_{i\bar{u}u}^{(t)} \log p(U_i^{(t)} = u | U_i^{(t-1)} = \bar{u}, \mathbf{x}_i^{(t)}) \right], \end{aligned} \quad (3)$$

where $a_{ijuy}^{(t)}$ is the indicator variable for subject i responding by y at occasion t to response variable j and belonging to latent state u at the same occasion, $b_{iu}^{(t)}$ is the indicator variable for subject i being in latent state u at occasion t , whereas $b_{i\bar{u}u}^{(t)}$ is the indicator variable for the transition from state \bar{u} at occasion $t-1$ to state u at occasion t . The EM algorithm alternates the following two steps until convergence:

- **E-step:** compute the posterior expected value of each indicator variable involved in (3) by suitable forward-backward recursions [3];
- **M-step:** maximize the complete data log-likelihood expressed as in (3), with each indicator variable substituted by the corresponding expected value. How to maximize this function depends on the specific formulation of the model.

Even if the EM algorithm is typically used to estimate LM models, in applications involving complex data the FML approach may have some drawbacks, as illustrated in Section 1.

3S and 3S-IMP estimation methods

The 3S approach proposed by [2] represents an extended version of the three-step approach for LC model with covariates [7], and it is based on the following steps:

- **Step 1:** fit a basic LC model for the set of response variables in which the responses provided by the same sample unit at different occasions are considered as coming from separate units. On the basis of this preliminary fitting, we obtain the final estimates of the conditional response probabilities, $\hat{\phi}_{jy|u}$, and the “temporary” estimates of the marginal probabilities of the latent states, $\hat{\rho}_u = \hat{p}(U_i^{(t)} = u)$.
- **Step 2:** for each subject i , compute the posterior expected value of $b_{iu}^{(t)}$ and $b_{i\bar{u}u}^{(t)}$ (see equation (3)) only on the basis of the results of the first step as

$$\begin{aligned} \tilde{b}_{iu}^{(t)} & \propto \hat{\rho}_u \hat{p}(\mathbf{Y}_i^{(t)} = \mathbf{y}_i^{(t)} | U_i^{(t)} = u) = \hat{\rho}_u \prod_j \hat{\phi}_{jy_{ij}^{(t)}|u}, \quad t = 1, \dots, T, \quad u = 1, \dots, k, \\ \tilde{b}_{i\bar{u}u}^{(t)} & = \tilde{b}_{i\bar{u}}^{(t-1)} \tilde{b}_{i\bar{u}u}^{(t)}, \quad t = 2, \dots, T, \quad \bar{u}, u = 1, \dots, k. \end{aligned}$$

- **Step 3:** maximize the components of the complete data log-likelihood involving the latent structure parameters

$$\tilde{\ell}_1^*(\boldsymbol{\beta}) = \sum_i \sum_u \tilde{b}_{iu}^{(1)} \log p(U_i^{(1)} = u | \mathbf{x}_i^{(1)}), \tag{4}$$

$$\tilde{\ell}_2^*(\boldsymbol{\Gamma}) = \sum_i \sum_{t \geq 2} \sum_{\bar{u}} \sum_u \tilde{b}_{i\bar{u}u}^{(t)} \log p(U_i^{(t)} = u | U_i^{(t-1)} = \bar{u}, \mathbf{x}_i^{(t)}). \tag{5}$$

Note that the second step of the procedure estimates the joint probabilities from the marginal, assuming independence. As a consequence, if we suppose that the estimated posterior probabilities are very concentrated on a given latent state, so that they are very close to 1 for a given state and to 0 otherwise, the product of these posterior probabilities will be close to 1 for a certain transition between states and to 0 otherwise. Moreover, when we deal with the extended LM model with covariates, the last step consists of estimating $\boldsymbol{\beta}$ and $\boldsymbol{\Gamma}$, defined in (1) and (2), by fitting multinomial logit models based on a weighed likelihood as defined in (4) and (5); the corresponding estimates are denoted by $\tilde{\boldsymbol{\beta}}$ and $\tilde{\boldsymbol{\Gamma}}$, respectively.

In order to overcome some limitations in the estimation of the parameters of the latent process, [2] proposed an improved version of the 3S approach, termed as 3S-IMP, in which the second and the third steps are iterated until convergence, while keeping the results from the first step fixed; see [2] for further details.

The 3S and 3S-IMP approaches produce consistent estimates of the conditional response probabilities. Moreover, since the first step is based on fitting a basic LC model, several starting values may be easily tried and the global maximum of its likelihood may be found with a reasonable effort. With respect to the FML approach, estimation of the parameters in $\boldsymbol{\beta}$ and $\boldsymbol{\Gamma}$ is fast and there are no problems of multiple solutions. Finally, as already illustrated in [2], the behavior of the estimators of the latent structure parameters improves as the number of response variables increases.

The extended BCH estimation method

As demonstrated by [4], the standard three-step approach for LC model underestimates the relationship between covariates and class membership. In practice, this means that the larger the amount of classification error introduced in the second step of the three-step approach, the larger the size of the bias in the parameter estimates. In order to overcome this drawback, the authors developed a bias-corrected method, based on this classification error, termed as BCH approach. [7] also proposed a modified BCH procedure to overcome some limitations of the original approach. Here, we propose to extend this modified approach to the case of LM models with covariates.

More in detail, in the second step the sample units are assigned to latent classes on the basis of the posterior class membership probabilities estimated during the first step, $\hat{p}(U_i^{(t)} = u | \mathbf{Y}_i^{(t)} = \mathbf{y}_i^{(t)})$. The assigned class membership of subject i at time t is denoted by $W_i^{(t)}$. Let $d_{uv}^{(t)} = p(W_i^{(t)} = v | U_i^{(t)} = u)$ denote the amount of classification error which is based on the conditional probability of the estimated value given the true assignment. Let the elements $d_{uv}^{(t)}$ be collected in the matrix $\mathbf{D}^{(t)}$. Let also denote by $d_{uv}^{(t)\dagger}$ the elements of the inverse of matrix $\mathbf{D}^{(t)}$. Accordingly, the third step of the proposed extended BCH approach consists of maximizing the components of the complete data log-likelihood involving the latent structure parameters,

expressed by (4) and (5), by replacing $\tilde{b}_{iu}^{(1)}$ and $\tilde{b}_{i\bar{u}u}^{(t)}$ with $\tilde{b}_{iu}^{(1)\dagger}$ and $\tilde{b}_{i\bar{u}u}^{(t)\dagger}$, where $\tilde{b}_{i\bar{u}u}^{(t)\dagger} = \tilde{b}_{i\bar{u}}^{(t-1)\dagger} \tilde{b}_{iu}^{(t)\dagger}$ and $\tilde{b}_{iu}^{(t)\dagger} = \sum_s \tilde{b}_{is}^{(t)} d_{su}^{(t)\dagger}$. This allows us to take into account the classification error and to reduce the bias in the estimates of the parameters in β and Γ .

2S estimation method

We also consider the 2S estimation method, based on a more direct two-step approach which allows for a maximum likelihood estimation of the parameters of the latent process. More in detail, after implementing the first step, aimed at estimating the conditional response probabilities, we propose to perform the second step by estimating an LM model with covariates with known measurement model. This can be directly done through the EM algorithm, while keeping the estimated conditional response probabilities fixed from the first step. It can be proven that this method produces efficient estimates of the parameters of the latent process, even with a few observed response variables. Moreover, it can be easily implemented in software for LM estimation which allows for parameters restriction.

4 Simulation study

We rely on a Monte Carlo simulation study aimed at evaluating the behavior of the FML, 3S, 3S-IMP, BCH, and 2S estimation algorithms under different scenarios. In particular, we generate sample data with certain population parameters and evaluate the performance of the estimation methods in terms of bias, standard error (se), root mean square error (rmse), and relative efficiency (eff), compared to the FML approach. The latter is computed as $\text{eff} = \text{rmse}(3\text{S}) / \text{rmse}(\text{FML})$.

We simulate 100 random samples according to a variety of settings: $n = 500, 1000$, $T = 5, 8$, $c_j = 2$, $r = 5, 10, 30$, and $k = 2, 3$. For every sample unit, we also consider two covariates, which are generated from an AR(1) process with autoregressive parameter equal to 0.5 and a Gaussian noise process with variance equal to 1. For the parameters affecting the distribution of the initial probabilities, we let $\beta_{0u} = 0$ and $\beta_{1u} = (0.5, 1)'$ for $u \geq 2$. Moreover, we evaluate the behavior of the estimation algorithms under different levels of uncertainty in the allocation of units to the latent states, by considering different structures of the matrix Φ_j . When $k = 2$, we let

$$\Phi_j^A = \begin{pmatrix} 0.92 & 0.08 \\ 0.08 & 0.92 \end{pmatrix}, \quad \Phi_j^B = \begin{pmatrix} 0.75 & 0.25 \\ 0.25 & 0.75 \end{pmatrix},$$

so as to allow more or less separated states. Furthermore, when $k = 3$, we let

$$\Phi_j = \begin{pmatrix} 0.92 & 0.08 & 0.75 \\ 0.08 & 0.92 & 0.25 \end{pmatrix}.$$

We also assume different degrees of persistence of the Markov chain, by letting the parameters in Γ to be equal to $\gamma_{0\bar{u}u} = \log(0.1/0.9)$, so as to include a high level of persistence in the latent Markov chain, and $\gamma_{0\bar{u}u} = \log(0.4/0.6)$, so as to allow a lower persistence of the chain. In both cases we fix $\gamma_{1\bar{u}u} = (0.5, 1)'$, for $\bar{u}, u = 1 \dots, k$ with $\bar{u} \neq u$ and $k = 2, 3$.

From the estimation results we observe that the conditional response probabilities are consistently estimated by the alternative approaches under all scenarios, as already demonstrated by [2]. Moreover, the bias of the 3S, 3S-IMP, and BCH estimators of β is always negligible and their root mean square errors decrease as the number of response variables increases. This

behavior is little affected by the level of persistence of the chain or by the degree of separation of the states. The 2S estimator of β performs quite well in almost all scenarios. As also showed by [2], estimation of parameters affecting the transition probabilities presents the most challenging issues. In this respect, for reason of space, we report here summary results for the most sensible scenarios, computed as the median value among the elements of the estimated matrix $\tilde{\Gamma}$ (for the bias we consider the median of the absolute value). We also report the computing time (in seconds) required to run the algorithms.

From the results reported in Table 1 we observe that the behavior, in terms of bias and root mean square error, of the 3S, 3S-IMP, and BCH estimators improves with: (i) number of response variables, (ii) separation between latent states, and (iii) as the level of persistence of the latent Markov chain decreases. In these contexts, the efficiency of the estimators tends to that of the FML approach. In terms of computational cost, we note that the computing time of the 3S and BCH algorithms is always significantly lower than that required by the FML approach. Overall, taking into account the computational cost, the BCH method outperforms the 3S and the 3S-IMP approaches. The 2S approach performs very well in all scenarios, even with a few response variables, with a negligible bias and an efficiency almost identical to that of the FML approach. However, in these simplified scenarios the computational cost is similar to that required by the FML algorithm.

In summary, according to our empirical evidence, these methods may represent a valid alternative to the FML approach when applied to complex real data, since they allow us to deal with the problem of multimodality of the model likelihood in a simplified framework, and to reach very similar performance in terms of parameter estimates with a lower computational effort.

Bibliography

- [1] Bartolucci, F., Farcomeni, A., and Pennoni, F. (2013). *Latent Markov Models for Longitudinal Data*. Chapman and Hall/CRC press.
- [2] Bartolucci, F., Montanari, G., and Pandolfi, S. (2014). Three-step estimation of latent Markov models with covariates. *arXiv:1402.1033v1*.
- [3] Baum, L., Petrie, T., Soules, G., and Weiss, N. (1970). A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *Annals of Mathematical Statistics*, 41:164–171.
- [4] Bolck, A., Croon, M., and Hagenaaars, J. (2004). Estimating latent structure models with categorical variables: One-step versus three-step estimators. *Political Analysis*, 12(1):3–27.
- [5] Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm (with discussion). *Journal of the Royal Statistical Society, Series B*, 39:1–38.
- [6] McLachlan, G.J. and Peel, D. (2000). *Finite Mixture Models*. Wiley.
- [7] Vermunt, J. (2010). Latent class modeling with covariates: Two improved three-step approaches. *Political Analysis*, 18:450–469.

		median(bias($\tilde{\Gamma}$))	median(se($\tilde{\Gamma}$))	median(rmse($\tilde{\Gamma}$))	median(eff($\tilde{\Gamma}$))	time
$k = 2, c_j = 3, \Phi_j^A, \gamma_{0\bar{u}u} = \log(0.1/0.9)$						
	FML	0.004	0.101	0.102	-	8.079
	3S	0.134	0.087	0.159	1.566	1.530
$r = 5$	3S-IMP	0.049	0.096	0.108	1.066	16.083
	BCH	0.010	0.104	0.104	1.067	1.616
	2S	0.004	0.101	0.102	1.000	7.350
	FML	0.005	0.104	0.104	-	6.928
	3S	0.005	0.104	0.104	1.000	1.946
$r = 30$	3S-IMP	0.005	0.104	0.104	1.000	4.128
	BCH	0.004	0.104	0.104	1.001	2.046
	2S	0.005	0.104	0.104	1.000	6.676
$k = 2, c_j = 3, \Phi_j^B, \gamma_{0\bar{u}u} = \log(0.1/0.9)$						
	FML	0.008	0.151	0.151	-	13.387
	3S	0.785	0.047	0.786	5.203	1.506
$r = 5$	3S-IMP	0.510	0.088	0.518	3.426	53.391
	BCH	0.013	0.215	0.215	1.483	1.690
	2S	0.003	0.150	0.150	0.993	11.995
	FML	0.019	0.113	0.115	-	7.870
	3S	0.025	0.106	0.111	0.970	2.198
$r = 30$	3S-IMP	0.006	0.111	0.111	0.977	13.345
	BCH	0.019	0.116	0.117	1.008	2.252
	2S	0.019	0.113	0.114	1.000	7.744
$k = 2, c_j = 3, \Phi_j^B, \gamma_{0\bar{u}u} = \log(0.4/0.6)$						
	FML	0.010	0.103	0.103	-	8.073
	3S	0.319	0.048	0.322	3.516	0.755
$r = 5$	3S-IMP	0.190	0.080	0.202	2.205	22.367
	BCH	0.016	0.129	0.137	1.221	0.810
	2S	0.007	0.103	0.103	0.993	5.991
	FML	0.007	0.078	0.079	-	4.151
	3S	0.010	0.077	0.078	0.986	1.253
$r = 30$	3S-IMP	0.004	0.078	0.079	0.997	6.772
	BCH	0.009	0.080	0.080	1.014	1.280
	2S	0.007	0.078	0.079	1.000	4.121
$k = 3, c_j = 2, \Phi_j, \gamma_{0\bar{u}u} = \log(0.4/0.6)$						
	FML	0.066	0.293	0.300	-	54.150
	3S	0.255	0.068	0.271	1.063	1.908
$r = 5$	3S-IMP	0.192	0.131	0.228	0.892	224.636
	BCH	0.079	0.253	0.290	0.929	2.056
	2S	0.050	0.224	0.225	0.781	24.361
	FML	0.008	0.125	0.126	-	6.678
	3S	0.138	0.093	0.160	1.452	2.278
$r = 30$	3S-IMP	0.076	0.105	0.137	1.084	49.337
	BCH	0.017	0.132	0.132	1.052	2.321
	2S	0.007	0.124	0.126	0.995	5.324

Table 1: Estimation results for parameters in Γ under different scenarios, with $n = 500$, and $T = 5$, together with the median, among the simulated sample, of the computing time (in seconds) required to run the algorithms

Estimation of spatially correlated ocean temperature curves including depth dependent covariates

Rosaura Fernández-Pascual, *University of Granada*, rpascual@ugr.es
Rosa M. Espejo, *University of Granada*, rosaespejo@ugr.es
Maria Dolores Ruiz-Medina, *University of Granada*, mruiz@ugr.es

Abstract. This paper addresses the problem of functional estimation of spatially correlated ocean temperature curves depending on depth. Specifically, a least-squares regression framework for the estimation of the scaling function coefficients is considered to approximate their functional trend. While a Bayesian inference approach for the estimation of the wavelet coefficients, approximating their local variation properties at different resolution levels, is adopted in a hierarchical model context. Spatial functional covariates are incorporated through depth dependent regression coefficients in the spatial functional linear model studied. A real-data example is considered for illustration of the derived results, in terms of spatial functional prediction of ocean temperature curves to detect global warming effects.

Keywords. Bayesian framework, covariates, spatially correlated depth-dependent curves, spatial functional regression, wavelet bases

1 Introduction

In the last few decades, statistical modeling, based on long record of observations at different spatial locations, has gained popularity in atmosphere and temperature ocean studies for the detection of global warming and climate change effects. Indeed, the application of functional statistical methodologies allows the identification of key features of the ocean and atmosphere that often occurred prior to subsequent changes in sea temperature at different depths. In absence of a proper spatial physical model, based, for example, on the theory of evolution equations (see, for instance, [10] and [11]), curves located at different weather stations in ocean are usually assumed to be uncorrelated for simplifications purposes.

The approach presented in this paper allows the statistical analysis of spatially correlated curve data, incorporating the information of depth dependent functional covariates. Recent

developments in the field of *Spatial Functional Statistics* provide useful tools for spatially dependent curve data processing. We refer to a few recent papers in relation to the inference problem from spatial correlated functional random variables. Specifically, [8] propose a new class of spatial functional regression models based on bivariate splines, in terms of which the surface defining the explanatory random variables is approximated. Such an approximation allows the construction of least squares estimators of the regression function with or without a penalization term. In [1], a fully Bayesian and Markov Chain Monte Carlo based approach is derived for the analysis of the spatially correlated functional data arising from different locations of biological structures called colonic crypts. Hierarchical clustering of spatially correlated functional data is performed in [7]. In the non-parametric context, the statistical properties of kernel-based density estimators, formulated in the context of spatial functional random variables, are studied in [2]. Functional statistical tools for the analysis of spatial correlated oceanology temporal curve data is considered in [13] within the spatial functional linear model context (see also [12]). Standard co-kriging is applied in terms of the projections of the functional data on a selected basis in [6]. In the context of spatial functional autoregressive series, we refer to the reader to the papers by [14] and [15].

On the other hand, for large spatiotemporal data sets new effective analytical tools with reasonable computational costs are of great interest (see [16]). Smoothing techniques such as function basis approximations and dynamical Bayesian modeling have been considered by [9] and [4]. A Bayesian hierarchical spatiotemporal estimation, based on Markov Chain Monte Carlo (MCMC) methods, which allows the efficient generation of samples from the posterior distribution is proposed in [5] and [9], in a mixed-effect spatiotemporal context. A space-time version of a Gaussian predictive process is considered in [4] for Bayesian dynamical modeling of large spatiotemporal data sets.

This paper proposes the combination of classical (least-squares regression), and Bayesian approaches in the estimation of spatially correlated ocean temperature curves at different depths, in a hierarchical functional modeling framework. Specifically, the wavelet transform is applied to discriminate between large scale smoothing (associated with the coarser scale described in terms of scaling function coefficients), and local smoothing (associated with the detail wavelet coefficients at different resolution levels). A multiple functional regression model with covariates is fitted by least-squares in the coarser scale of the wavelet domain, in terms of the scaling function coefficients. A Bayesian framework is adopted in the estimation of the parameters characterizing the distribution of wavelet coefficients at high resolution levels. The approach presented constitutes an alternative to the one presented in [17] for the estimation of spatially correlated functional responses depending on depth in an Hierarchical Bayesian framework, from projection into Empirical Orthogonal Function bases (EOF bases). Specifically, the proposed estimation methodology for spatially correlated curves is more flexible than the one considered in [17], since it allows the fitting of spatially heterogeneous trends, and local variation properties at different scales, combining classical and Bayesian inference frameworks. Note that the nonparametric approach adopted in [17] in a hierarchical modeling context only contemplates Bayesian estimation, without possible discrimination between large-scale and small-scale property approximations, which usually require different hierarchical modeling frameworks, respectively.

The wavelet-based spatial functional estimation methodology in this paper is illustrated with a real-data example in relation to investigation of ocean temperature changes at different depths. This study is motivated by the detection of possible effects of global warming and climate change, inducing increasing temperature differences between two closed ocean coastal areas, that could

produce flows strengthening altering marine biodiversity.

2 The model

Let us consider the following spatial functional linear model inspired in [17]

$$Y(\mathbf{s}; d) = \int X(\mathbf{s}; \mathbf{u})\beta(\mathbf{u}; d)d\mathbf{u} + \sum_{j=1}^J Z_j(\mathbf{s}; d)\delta_j(d) + G(\mathbf{s}; d), \tag{1}$$

where $\{Y(\mathbf{s}; d), \mathbf{s} \in D \subset \mathbb{R}^2, d \in \mathcal{I} \subset \mathbb{R}\}$ is the functional response, $\{X(\mathbf{s}; d), \mathbf{s} \in D \subset \mathbb{R}^2, d \in \mathcal{I} \subset \mathbb{R}\}$ is the functional regressors and $\{Z_j(\mathbf{s}; d), \mathbf{s} \in D \subset \mathbb{R}^2, d \in \mathcal{I} \subset \mathbb{R}\}, j = 1, \dots, J$, are the functional covariates. Here, D could be a compact domain in \mathbb{R}^2 , but in our case is a subset of \mathbb{Z}^2 , given by the latitudes and longitudes defining the locations of the studied ocean weather stations. Set \mathcal{I} is continuous, and in our case it represents a real closed finite interval of depths. The functional variables $Y(\mathbf{s}; \cdot), X(\mathbf{s}; \cdot)$ and $Z_j(\mathbf{s}; \cdot), j = 1, \dots, J$, are assumed to be in a separable Hilbert space H , for each $\mathbf{s} \in D \subset \mathbb{R}^2$. In the subsequent development, we will consider $H = L^2(\mathcal{I})$, the space of square integrable functions on the interval \mathcal{I} .

An estimation methodology, based on wavelets, is firstly proposed for the least-squares projection estimation of functional parameter $\beta(\mathbf{u}; d)$, which is assumed to be a Hilbert-Schmidt kernel defining the regression operator corresponding to $X(\mathbf{s}; \mathbf{u})$. This methodology also provides the least-squares functional estimator of regression coefficients $\delta_j(d), j = 1, \dots, J$, respectively corresponding to covariate curves $Z_j(\mathbf{s}; d), j = 1, \dots, J$, at each spatial location $\mathbf{s} \in D$. The spatial functional process $G(\mathbf{s}; d)$ is a zero-mean Gaussian H -valued noise in the strong sense, reflecting small-scale spatial and depth local variation.

Remark 8.

Note that here we have considered spatial functional regressors $X(\mathbf{s}, \cdot), \mathbf{s} \in D$, depending on depth, and a vector of spatial functional depth-dependent covariate vector with H -valued components $Z_j(\mathbf{s}; d), j = 1, \dots, J$, which is slightly different from the case studied in [17], where they consider J functional covariates $X_j(\mathbf{s}; \mathbf{u}_j), j = 1, \dots, J$, with $\mathbf{u}_j = (d, w_j)$ depending on depth and wavelength, respectively.

Let us denote by $\Phi_s^*, \Phi_d^*, \Psi_s^*$ and Ψ_d^* the projection operators into the scaling basis in space, the scaling basis in depth, the wavelet basis in space, and the wavelet basis in depth, respectively. Hence, Φ_s, Φ_d, Ψ_s and Ψ_d respectively denote their adjoint and inverse operators such that

$$\begin{aligned} \Phi_s^* \Phi_s &= I_{V_0^s}, & \Phi_d^* \Phi_d &= I_{V_0^d} \\ [\Psi_s^*]_m [\Psi_s]_m &= I_{W_m^s}, & [\Psi_d^*]_k [\Psi_d]_k &= I_{W_k^d}, \end{aligned} \tag{2}$$

for $m = 0, 1, \dots, M_s$, and $k = 0, 1, \dots, M_d$, where M_s and M_d respectively denote the number of resolution levels considered in the application of the discrete compactly supported wavelet transform in space and depth. Here, $[\Psi_s^*]_m$ denotes projection into the space W_m^s generated by spatial wavelet functions at resolution level m , for $m = 0, 1, \dots, M_s$. Similarly, $[\Psi_d^*]_k$ denotes projection into the space W_k^d generated by one-dimensional wavelet functions depending on depth at resolution level k , for $k = 0, 1, \dots, M_d$. Note that it is well-known (see, for example, [3]) that

the wavelet bases considered respectively provide a multiresolution analysis of the spaces $L^2(D)$ and $L^2(\mathcal{I})$ as follows:

$$\begin{aligned} L^2(D) &= V_0^s \bigoplus \sum_{m=1}^{\infty} W_m^s \\ L^2(\mathcal{I}) &= V_0^d \bigoplus \sum_{k=1}^{\infty} W_k^d. \end{aligned} \quad (3)$$

Applying Φ_s^* to the left-hand side of equation (1) and Φ_d to the right-hand side, we obtain

$$\Phi_s^* \mathbf{Y} \Phi_d = \Phi_s^* \mathbf{X} \Phi_d \Phi_d^* \beta \Phi_d + \sum_{j=1}^J \Phi_s^* \mathbf{Z}_j \Phi_d \delta_j \Phi_d + \Phi_s^* \mathbf{G} \Phi_d, \quad (4)$$

which defines our regression model, whose scaling parameter coefficients $\Phi_d^* \beta \Phi_d$ and $\Phi_d^* \delta_j$ are estimated by applying least-squares methodology.

Additionally, for given multiresolution levels $m = 0, 1, \dots, M_s$, and $k = 0, 1, \dots, M_d$, we apply $[\Psi_s^*]_m$ to the left-hand side of equation (1) and $[\Psi_d^*]_k$ to the right-hand side of equation (1), leading to

$$\begin{aligned} [\Psi_s^*]_m \mathbf{Y} [\Psi_d]_k &= [\Psi_s^*]_m \mathbf{X} [\Psi_d]_k [\Psi_s^*]_m \beta [\Psi_d]_k \\ &+ \sum_{j=1}^J [\Psi_s^*]_m \mathbf{Z}_j [\Psi_d]_k \delta_j [\Psi_d]_k + [\Psi_s^*]_m \mathbf{G} [\Psi_d]_k, \end{aligned} \quad (5)$$

which reflects the small-scale random variations and spatial and depth dependence in our model. The Bayesian framework is adopted for the estimation of the parameters characterizing the distribution of Gaussian wavelet coefficients. In particular, conjugate priors for the univariate Gaussian likelihood are proposed for the estimation of the variance of spatial and depth Gaussian wavelet coefficients of process G . While MCMC methods are implemented for approximation of the mean of the posterior, given the proposed multivariate priors for the wavelet coefficients of β and δ_j , $j = 1, \dots, J$, as well as for the covariance matrix of the spatial and depth wavelet coefficients of \mathbf{Y} with associated multivariate Gaussian likelihood.

Acknowledgement

This work has been supported in part by project MTM2012-32674 (co-funded with FEDER) of the DGI, MEC, Spain.

Bibliography

- [1] Baladandayuthapani, V., Mallick, B., Hong, M., Lupton, J., Turner, N., Carroll, R. (2008). Bayesian hierarchical spatially correlated functional data analysis with application to colon carcinogenesis. *Biometrics* 64, 64–73.
- [2] Basse, M., Diop, A., Dabo-Niang, S. (2008). Mean square properties of a class of kernel density estimates for spatial functional random variables. *Annales De L'I.S.U.P. Publications de l'Institut de Statistique de l'Université de Paris*.

- [3] Cohen, A., Daubechies, I. and Vial, P. (1994). Wavelets on the interval and fast wavelet transforms. *J. Appl. Comput. Harmon. Anal* 1, 54-81.
- [4] Finley, A.O., Banerjee, S. and Gelfand, A.E. (2012). Bayesian dynamic modeling for large space-time datasets using Gaussian predictive processes. *Journal of Geographical Systems* 14, 29-47.
- [5] Ganggang, Xu, Faming, L. and Genton, M. (2013). A bayesian spatio-temporal geo-statistical model with an auxiliary lattice for large datasets. *Statistica Sinica*. (preprint doi:10.5705/ss.2013.085w).
- [6] Giraldo, R., Delicado, P. and Mateu, J. (2010). Continuous time-varying kriging for spatial prediction of functional data: an environmental application. *Journal of Agricultural, Biological, and Environmental Statistics* 15, 66-82.
- [7] Giraldo, R., Delicado, P. and Mateu, J. (2012). Hierarchical clustering of spatially correlated functional data. *Statistica Neerlandica* 66, 403-421
- [8] Guillas, S. and Lai, M.J. (2010). Bivariate splines for spatial functional regression models. *J. Roy. Stat. Soc. Ser. B* 22, 477-497.
- [9] Katzfuss, M. and Cressie, N. (2012). Bayesian hierarchical spatio-temporal smoothing for very large datasets. *Environmetrics, Special Issue: Spatio-Temporal Stochastic Modelling* 23, pages 94-107.
- [10] Kelbert, M., Leonenko, N.N. and Ruiz-Medina, M.D. (2005). Fractional random fields associated with stochastic fractional heat equations. *Advances in Applied Probability* 108, 108-133.
- [11] Leonenko, N.N. and Ruiz-Medina, M.D. (2006). Scaling laws for the multidimensional Burgers equation with quadratic external potencial. *Journal of Statistical Physics* 124, 191-205.
- [12] Monestiez, P. and Nerini, D. (2008). A cokriging method for spatial functional data with applications in oceanology. *Functional and Operational Statistics. Contributions to Statistics* 36, 237-242.
- [13] Nerini, D., Monestiez, P. and Manté, C. (2010). Cokriging for spatial functional data. *J. Multiv. Anal.* 101, 409-418.
- [14] Ruiz-Medina, M.D. (2011). Spatial autoregressive and moving average Hilbertian processes. *J. Multiv. Anal.* 102, 292-305.
- [15] Ruiz-Medina, M. D. and Espejo, R. (2013). Integration of spatial functional interaction in the extrapolation of ocean surface temperature anomalies due to global warming. *International Journal of Applied Earth Observation and Geoinformation* 22, 27-39.
- [16] Sun, Y., Li, B. and Genton, M.G. (2012). Geostatistics for Large Datasets, in *Space- Time Processes and Challenges Related to Environmental Problems*, Porcu, E., Montero, J.M., Schlather, M. (eds), Springer, 207, Chapter 3, 55-77.

- [17] Yang, W.H., Wikle, C.K., Holan, S.H., Sudduth, K., and Meyers, D.B. (2014). Bayesian analysis of spatially-dependent functional responses with spatially-dependent multi-dimensional functional predictors. *Statistica Sinica* (to appear). doi:10.5705/ss.2013.245w

On the bootstrap methodology for the estimation of the tail sample fraction

Frederico Caeiro, *FCT and CMA, Universidade Nova de Lisboa*, fac@fct.unl.pt
M. Ivette Gomes, *FCUL and CEAUL, Universidade de Lisboa*, ivette.gomes@fc.ul.pt

Abstract. In statistics of extremes we are usually interested in the estimation of parameters of extreme events. Such estimation is usually based on the largest $k + 1$ order statistics or on the excesses over a high level u . In this paper, we consider the adaptive estimation of either k or u through the nonparametric bootstrap methodology. We shall introduce an improved version of Hall's bootstrap methodology and compare it with the double bootstrap methodology. The comparison of such methodologies is performed for simulated data sets.

Keywords. Bootstrap Methodology, Extreme Value Index, Heavy tail, Tail sample fraction.

1 Introduction

Let $\underline{X}_n = (X_1, \dots, X_n)$ denote a sample of either independent, identically distributed (i.i.d.) or even weakly dependent random variables (r.v.'s) from an underlying distribution function F . We shall assume that we are in the max-domain of attraction of the Extreme Value distribution $EV_\xi(x) := \exp(-(1 + \xi x)^{-1/\xi})$, $1 + \xi x > 0$, where the shape parameter ξ is the well known extreme value index (EVI). We shall consider $\xi > 0$, i.e., models with a heavy right tail. Then, the quantile function $U(t) := F^{\leftarrow}(1 - 1/t) = \inf\{x : F(x) \geq 1 - 1/t\}$, $t > 1$, is a regularly varying function with a positive index of regular variation equal to ξ , i.e.,

$$\lim_{t \rightarrow \infty} \frac{U(tx)}{U(t)} = x^\xi . \quad (1)$$

For a heavy tailed model, the classic semi-parametric Hill estimator of ξ , introduced in [17], is

$$H(k) \equiv \hat{\xi}_n^H(k) := \frac{1}{k} \sum_{i=1}^k (\ln X_{n-i+1:n} - \ln X_{n-k:n}), \quad k = 1, 2, \dots, n-1, \quad (2)$$

the average of the log excesses over the high threshold $X_{n-k:n}$, where $X_{i:n}$ denotes the i -th ascending order statistic of the sample of size n . Consistency is achieved for intermediate k , i.e. for sequences of integers $k = k_n$, $1 \leq k < n$, such that

$$k \rightarrow \infty \quad \text{and} \quad k/n \rightarrow 0, \quad \text{as} \quad n \rightarrow \infty. \quad (3)$$

To obtain the asymptotic distributional behaviour of the Hill and other semi-parametric EVI-estimators, we need to assume a second-order condition, that measures the rate of convergence in the first-order condition, i.e. the way $\ln U(tx) - \ln U(t)$ approaches $\xi \ln x$,

$$\lim_{t \rightarrow \infty} \frac{\ln U(tx) - \ln U(t) - \xi \ln x}{A(t)} = \begin{cases} (x^\rho - 1)/\rho, & \text{if } \rho < 0, \\ \ln x, & \text{if } \rho = 0, \end{cases} \quad (4)$$

for every $x > 0$, where ρ (≤ 0) is a second-order parameter that rules the rate of convergence and $|A|$ is compulsory a regular varying function with index ρ . For technical simplicity, we shall assume $\rho < 0$. Under the second-order condition in (4) the Hill estimator has usually a high asymptotic bias and recently several authors have considered different ways of reducing the bias. A simple class of second-order minimum-variance reduced-bias (MVRB) EVI estimators is the one in [2], given by

$$CH(k) \equiv \hat{\xi}_n^{CH}(k) := \hat{\xi}_n^H(k) \left(1 - \hat{\beta}(n/k)^{\hat{\rho}} / (1 - \hat{\rho})\right), \quad k = 1, 2, \dots, n-1, \quad (5)$$

with $(\hat{\beta}, \hat{\rho})$ adequate estimators of the second-order parameters (β, ρ) such that $A(t) = \gamma \beta t^\rho$, $\rho < 0$. This estimator has an asymptotic variance equal to that of the Hill EVI-estimator, but an asymptotic bias of smaller order, and thus beats the classical estimators for all k . For a reliable estimation of the EVI, some attention should be given to the choice of the number k , or equivalently to the threshold $X_{n-k:n}$. Recent overviews of statistics of univariate extremes were recently published (see [1, 6, 12], among others).

In section 2 of this paper we present several known results that allow us to compute the theoretical optimal level of the EVI-estimators in (2) and (5). In Section 3, we discuss the estimation of the second-order parameters ρ and β . In Section 4 we shall use bootstrap computer-intensive resampling methods for the choice of k , not only for the use of $H(k)$, but also for the use of $CH(k)$. We introduce a new bootstrap method, based on Hall's methodology, and present the double bootstrap algorithm. Finally, we provide an application to simulated data sets.

2 Asymptotic Properties

If we assume the validity of the second-order framework in (4), $\hat{\xi}_n^H(k)$ is asymptotically normal, provided that $\sqrt{k}A(n/k) \rightarrow \lambda$, finite, as $n \rightarrow \infty$. Indeed, we have, with $\mathcal{N}_{\mu, \sigma^2}$ denoting a normal random variable with mean value μ and variance σ^2 , and $b_1 = 1/(1 - \rho)$,

$$\sqrt{k}(\hat{\xi}_n^H(k) - \xi) \stackrel{d}{=} \mathcal{N}_{0, \xi^2} + b_1 \sqrt{k}A(n/k) + o_p(\sqrt{k}A(n/k)), \quad \text{as } n \rightarrow \infty. \quad (6)$$

The bias $b_1 \sqrt{k}A(n/k) = \xi \beta \sqrt{k} (n/k)^\rho / (1 - \rho)$ can be very large, moderate or small, and increases as k increases. And since the variance decreases with k , we have usually a very sharp mean square error (MSE) pattern, as a function of k . Under the same conditions as before, $\sqrt{k}(\hat{\xi}_n^{CH}(k) - \xi)$ is asymptotically normal with variance also equal to ξ^2 but with a null mean value.

To obtain information on the bias of MVRB EVI-estimators it is common to slightly restrict the class of models in (4), further assuming a third-order condition, ruling now the rate of convergence in the second-order condition in (4). We shall consider the third-order condition used in [3], which guarantees that for all $x > 0$,

$$\lim_{t \rightarrow \infty} \frac{\frac{\ln U(tx) - \ln U(t) - \xi \ln x}{A(t)} - \frac{x^\rho - 1}{\rho}}{B(t)} = \frac{x^{\rho+\rho'} - 1}{\rho + \rho'}, \tag{7}$$

where $|B|$ is a regular varying function with index ρ' . Further details can be found in [11].

The full asymptotic behaviour of $\hat{\xi}_n^{CH}(k)$ is provided in the following theorem.

Theorem 2.1. *If under the validity of the second-order condition in (4), we estimate β and ρ consistently through $\hat{\beta}$ and $\hat{\rho}$, in such a way that $\hat{\rho} - \rho = o_p(1/\ln n)$, the asymptotic distributional representation $\sqrt{k}(\hat{\xi}_n^{CH}(k) - \xi) \stackrel{d}{=} \mathcal{N}_{0,\xi^2} + o_p(\sqrt{k}A(n/k))$ holds. Under the validity of equation (7), we can guarantee*

$$\sqrt{k}(\hat{\xi}_n^{CH}(k) - \xi) \stackrel{d}{=} \mathcal{N}_{0,\xi^2} + b_2 \sqrt{k}A^2(n/k) (1 + o_p(1)), \tag{8}$$

for adequate k values such that $\sqrt{k}A^2(n/k) \rightarrow \lambda_A$, finite and $b_2 = (\omega/(1 - 2\rho) - (1 - \rho)^{-2})/\xi$ with $\omega = B(n/k)/A(n/k)$.

Regarding the choice of k , an usual approach is to minimize the MSE of the EVI-estimator. With AMSE standing for ‘asymptotic MSE’, on the basis of (6) and (8), and with the notation $\hat{\xi}_n^{(1)} = \hat{\xi}_n^H$, $\hat{\xi}_n^{(2)} = \hat{\xi}_n^{CH}$, we get $\text{AMSE}(\hat{\xi}_n^{(c)}(k)) = \xi^2/k + b_c^2 A^{2c}(n/k)$, $c = 1, 2$ and

$$k_0^{(c)}(n) := \arg \min_k \text{AMSE}(\hat{\xi}_n^{(c)}(k)) = \left(\frac{n^{-2c\rho}}{(-2c\rho)b_c^2 \xi^{2(1-c)} \beta^{2c}} \right)^{1/(1-2c\rho)}, \quad c = 1, 2. \tag{9}$$

3 Estimation of the second-order parameters

We have used particular members of the class of estimators of the second-order parameter ρ proposed in [10]. Such a class of estimators has been first parameterized by a tuning parameter $\tau \geq 0$, that can be straightforwardly considered as a real number, and is defined as

$$\hat{\rho}_\tau(k) := \min \left\{ 0, \frac{3(T_n^{(\tau)}(k) - 1)}{T_n^{(\tau)}(k) - 3} \right\}, \quad T_n^{(\tau)}(k) := \frac{(M_n^{(1)}(k))^\tau - (M_n^{(2)}(k)/2)^{\tau/2}}{(M_n^{(2)}(k)/2)^{\tau/2} - (M_n^{(3)}(k)/6)^{\tau/3}}, \quad \tau \in \mathbb{R},$$

with the notation $a^{b\tau} = b \ln a$ if $\tau = 0$ and where $M_n^{(j)}(k) := \frac{1}{k} \sum_{i=1}^k \{\ln X_{n-i+1:n} - \ln X_{n-k:n}\}^j$, $j = 1, 2, 3$. Interesting alternative ρ -estimators have recently been introduced in [5, 8]. Here we consider the same type of criterion used in [15] for the adaptive estimation of ρ : Consider a sample with n positive values, compute $\{\hat{\rho}_\tau(k)\}_{k \in \mathcal{K}}$, with $\mathcal{K} = (\lfloor n^{0.995} \rfloor, \lfloor n^{0.999} \rfloor)$, compute their median, denoted η_τ , and compute $I_\tau := \sum_{k \in \mathcal{K}} (\hat{\rho}_\tau(k) - \eta_\tau)^2$, $\tau = 0, 1$. Next choose $\tau^* = 0$ if $I_0 \leq I_1$; otherwise, choose $\tau^* = 1$ and compute $\hat{\rho} \equiv \hat{\rho}_{\tau^*} = \hat{\rho}_{\tau^*}(k_1)$, with $k_1 = \lfloor n^{0.995} \rfloor$.

The estimate of the scale second-order parameter β is given by $\hat{\beta} = \hat{\beta}_\rho(k_1)$ with $\hat{\beta}_\rho(k)$ the estimator in [13], given by

$$\hat{\beta}_\rho(k) := \left(\frac{k}{n} \right)^{\hat{\rho}} \frac{D_{\hat{\rho},0}(k) D_{0,1}(k) - D_{\hat{\rho},1}(k)}{D_{\hat{\rho},0}(k) D_{\hat{\rho},1}(k) - D_{2\hat{\rho},1}(k)}, \quad D_{\alpha_1,\alpha_2}(k) := \frac{1}{k} \sum_{i=1}^k \left(\frac{i}{k} \right)^{-\alpha_1} U_i^{\alpha_2},$$

with $U_i := i(\ln X_{n-i+1:n} - \ln X_{n-i:n})$ the rescaled log-spacings and dependent on the estimator $\hat{\rho}$, suggested before.

4 The bootstrap methodology

A method based on Hall's single bootstrap

The bootstrap methodology for the selection of the threshold k was first introduced by Hall ([16]). To avoid the underestimation of the bias, it is necessary to use smaller resamples of size $n_1 = o(n)$, where n is the size of the initial sample. Let $\underline{X}_{n_1}^* = \{X_1, \dots, X_{n_1}\}$ denote a resample of size $n_1 = o(n)$ taken with replacement. Hall's considered the minimization of the bootstrap estimate of the MSE of $\hat{\xi}_{n_1}^H(k)$,

$$MSE(n_1, k) = E \left[\left\{ \hat{\xi}_{n_1}^H(k) - \hat{\xi}_{n_1}^H(k_{aux}) \right\}^2 \mid \underline{X}_{n_1}^* \right] \quad (10)$$

where k_{aux} is an initial threshold such that $\hat{\xi}_{n_1}^H(k_{aux})$ is consistent for ξ . Next we choose the value $k_0^*(n_1)$ that minimizes (10). The bootstrap estimate of the tail fraction is then

$$k_0^*(n) = k_0^*(n_1)(n/n_1)^\alpha.$$

Hall suggested $\alpha = 2/3$, which is equivalent to say that our model is under the second-order condition with $\rho = -1$. Also, as noticed by [14], the method is very sensitive to the choice of k_{aux} . Here we shall consider again an auxiliary statistic of the type of the one considered in [14], directly related to the EVI-estimator under consideration, but going to the known value zero,

$$T_n^{(c)}(k) := \hat{\xi}_n^{(c)}(\lfloor k/2 \rfloor) - \hat{\xi}_n^{(c)}(k), \quad k = 2, \dots, n-1, \quad c = 1, 2. \quad (11)$$

Notice that if $c = 1$ this approach is equivalent to consider $k_{aux} = \lfloor k/2 \rfloor$ in (10). On the basis of the results similar to the ones in [14], we can get for $T_n^{(c)}(k)$, in (11), the asymptotic distributional representation,

$$T_n^{(c)}(k) \stackrel{d}{\approx} \frac{\xi^2}{\sqrt{k}} Q_k + b_c(2^{c\rho} - 1) A(n/k)(1 + o_p(1)),$$

with Q_k asymptotically $\mathcal{N}_{0,1}$, and b_c , $c = 1, 2$ given in Section 2. Then, the AMSE of $T_n^{(c)}(k)$ is minimal at a level $k_{0|T}^{(c)}(n)$, such that

$$k_0^{(c)}(n) = k_{0|T}^{(c)}(n)(2^{c\rho} - 1)^{\frac{2}{1-2c\rho}}$$

Based on Hall's method we now introduce a new bootstrap algorithm:

Algorithm 4.1.

Let $\hat{\xi}_n^{(c)}(k)$ denote any of the EVI-estimators in (2) ($c = 1$) or in (5) ($c = 2$). We now proceed with the description of the algorithm for the adaptive estimation of the optimal threshold $k_0^{(c)}(n)$ and the adaptive estimation of ξ .

Step 1 Given a sample (x_1, x_2, \dots, x_n) , compute the estimates $\hat{\rho}$ and $\hat{\beta}$ of the second-order parameters ρ and β as described in Section 3.

Step 2 Next, consider a sub-sample size $n_1 = o(n)$. For l from 1 until B , generate independently B bootstrap samples $(x_1^*, x_2^*, \dots, x_{n_1}^*)$ of size n_1 , from the empirical d.f. $F_n^*(x) = \frac{1}{n} \sum_{i=1}^n I_{\{X_i \leq x\}}$ associated with the observed sample (x_1, x_2, \dots, x_n) .

Step 3 Denoting $T_{n_1}^{(c)*}(k)$ the bootstrap counterpart of $T_{n_1}^{(c)}(k)$, defined in (11), obtain $t_{n_1,l}^*(k)$, $1 \leq l \leq B$, the observed values of $T_{n_1}^{(c)*}(k)$. For $k = 2, \dots, n_1 - 1$, compute $MSE^*(n_1, k) = \frac{1}{B} \sum_{l=1}^B (t_{n_1,l}^*(k))^2$, and obtain $\hat{k}_{0|T}^*(n_1) := \arg \min_{1 < k < n_1} MSE^*(n_1, k)$.

Step 4 Compute the threshold estimate

$$\hat{k}_0^*(n) \equiv \left\lceil (1 - 2^{c\hat{\rho}})^{\frac{2}{1-2c\hat{\rho}}} \hat{k}_{0|T}^*(n_1) (n/n_1)^{\frac{-2c\hat{\rho}}{1-2c\hat{\rho}}} \right\rceil + 1.$$

If $\hat{k}_0^*(n) > n - 1$ go back to **Step 2**, being careful not to generate the same samples.

Step 5 Obtain $\hat{\xi}^* \equiv \hat{\xi}_n^{(c)}(\hat{k}_0^*(n))$.

The double bootstrap method

The assumptions in Hall’s methodology were overpassed with the use of a double bootstrap method. This method was first used in [9] for the general max-domain of attraction and in [7, 14] for heavy tailed models. More recently, [15] modified the double bootstrap algorithm for an adaptive choice of the thresholds for second-order corrected-bias estimators. The next algorithm follows closely the bootstrap method in [15].

Algorithm 4.2.

Let $\hat{\xi}_n^{(c)}(k)$ denote any of the EVI-estimators in (2) ($c = 1$) or in (5) ($c = 2$). We now proceed with the description of the algorithm for the adaptive estimation of the optimal threshold $k_0^{(c)}(n)$ and the adaptive estimation of ξ .

Step 1 Equal to **Step 1** in Algorithm 4.1.

Step 2 Next, consider a sub-sample size $n_1 = o(n)$ and $n_2 = \lfloor n_1^2/n \rfloor + 1$. For l from 1 until B , generate independently B bootstrap samples $(x_1^*, \dots, x_{n_2}^*)$ and $(x_1^*, \dots, x_{n_2}^*, x_{n_2+1}^*, \dots, x_{n_1}^*)$, of sizes n_2 and n_1 , respectively, from the empirical d.f. $F_n^*(x) = \frac{1}{n} \sum_{i=1}^n I_{\{X_i \leq x\}}$ associated with the observed sample (x_1, x_2, \dots, x_n) .

Step 3 Denoting $T_{n_i}^{(c)*}(k)$ the bootstrap counterpart of $T_{n_i}^{(c)}(k)$, in (11), obtain $t_{n_i,l}^*(k)$, $1 \leq l \leq B$, $i = 1, 2$ the observed values of $T_{n_i}^{(c)*}(k)$. For $k = 2, \dots, n_i - 1$, and $i = 1, 2$ compute $MSE^*(n_i, k) = \frac{1}{B} \sum_{l=1}^B (t_{n_i,l}^*(k))^2$, and obtain $\hat{k}_{0|T}^*(n_i) := \arg \min_{1 < k < n_i} MSE^*(n_i, k)$.

Step 4 Compute the threshold estimate

$$\hat{k}_0^*(n) \equiv \left\lceil (1 - 2^{c\hat{\rho}})^{\frac{2}{1-2c\hat{\rho}}} (\hat{k}_{0|T}^*(n_1))^2 / \hat{k}_{0|T}^*(n_2) \right\rceil + 1.$$

If $\hat{k}_0^*(n) > n - 1$ go back to **Step 2**, being careful not to generate the same samples.

Step 5 Obtain $\hat{\xi}^* \equiv \hat{\xi}_n^{(c)}(\hat{k}_0^*(n))$.

Remarks:

- The use of the sample $(x_1^*, x_2^*, \dots, x_{n_2}^*)$, and of the extended sample $(x_1^*, \dots, x_{n_2}^*, \dots, x_{n_1}^*)$, $n_2 < n_1$, lead us to an increased precision of the result with the same number B of bootstrap samples generated in **Step 2**. This is quite similar to the use of the simulation technique of “Common Random Numbers” in comparison problems.
- Bootstrap confidence intervals are easily obtained, through the replication of this algorithm r times. The replication can also provide us more precise estimates, if we consider the estimate given by the mean or the median of the r bootstrap estimates.
- A few practical questions may be raised under the set-up developed: How does the asymptotic method work for moderate sample sizes? Is the method strongly dependent on the choice of n_1 ? Although aware of the theoretical need to have $n_1 = o(n)$, what happens if we choose $n_1 = n$? We will try to answer those questions in the next section.

Applications to Simulated Data Sets

Here we shall present an illustration of the performance of the algorithms to simulated samples, cases where we know the value of ξ . We have simulated one random sample of size $n = 500$, from a Burr model with d.f. $F(x) = 1 - (1 + x^{-\rho/\xi})^{1/\rho}$, $x > 0$, $\xi > 0$, $\rho < 0$ with $\xi = 0.25$ and $\rho = -0.75$ and one Student's- t_4 sample of size $n = 1000$ ($\xi = 0.25$, $\rho = -0.5$).

Conclusions: Bootstrap estimates of the optimal sample fractions, $\hat{k}_0^*(n)/n$ and of the EVI, $\hat{\xi}^*$, as functions of n_1 , for $\lfloor n^{0.85} \rfloor \leq n_1 \leq n$, are pictured in Figs. 1-2. Since we know the true value of ξ , and we can easily assess the reliability of the estimates provided by the Algorithms, immediately coming to the conclusion that Algorithm 4.2 provides a quite reliable EVI-estimation, even with $n_1 = n$. Algorithm 4.2 can be very sensitive to the choice of n_1 (see Fig. 1). We noticed that we can have some volatility in the estimates as function of n_1 and such volatility only decreases substantially with the replication of the algorithm $r = 25$ times. For the Burr sample, Algorithm 4.1 is sensitive to the choice of n_1 . These results claim obviously for a simulation study of the Algorithms and its application to real data sets. These are however topics that can only be covered in a full-length paper.

Acknowledgement

Research partially supported by FCT - Fundação para a Ciência e a Tecnologia, projects PEst-OE/MAT/UI0006/2014 (CEAUL), PEst-OE/MAT/UI0297/2014 (CMA/UNL).

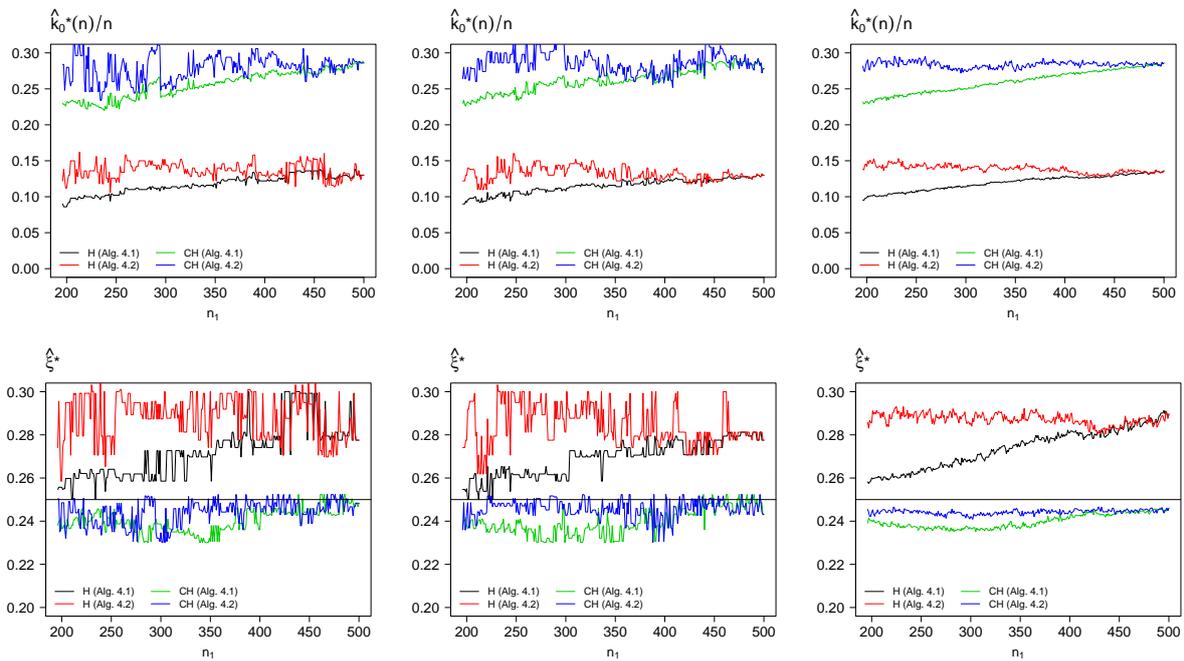


Figure 1: Adaptive estimates of $\hat{k}_0^*(n)/n$ (above) and $\hat{\xi}^*$ (below), as function of n_1 , with $B = 250$ (left), $B = 1000$ (center) and mean of $r \times B = 25 \times 250$ (right), for the Burr simulated sample.

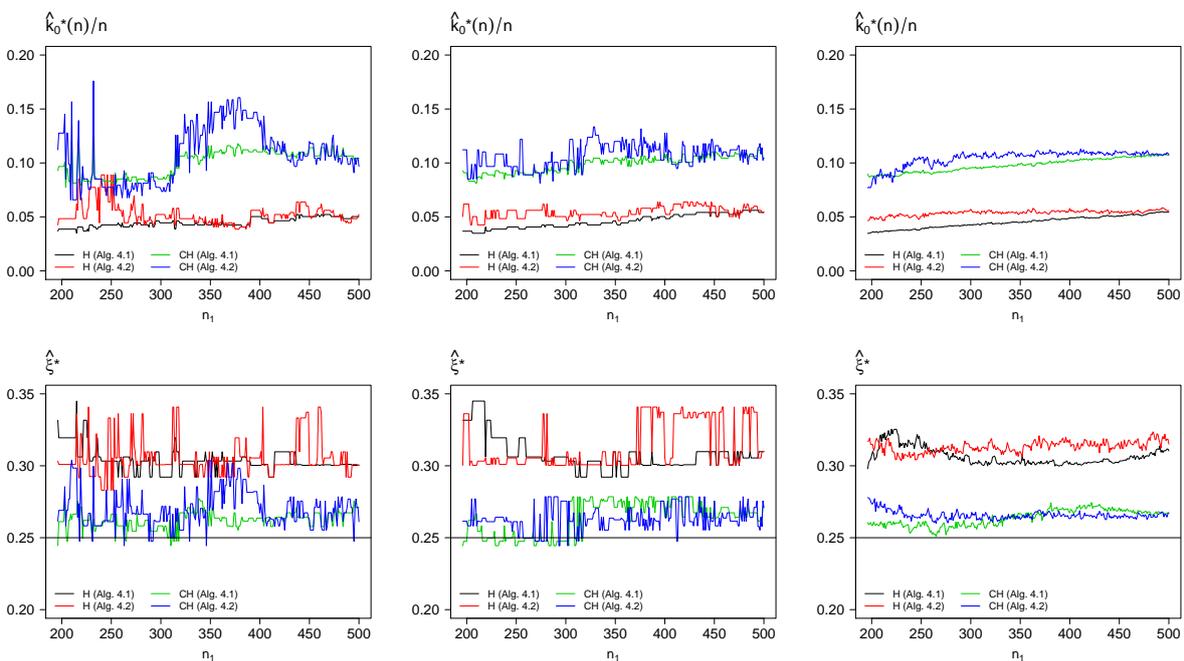


Figure 2: Adaptive estimates of $\hat{k}_0^*(n)/n$ (above) and $\hat{\xi}^*$ (below), as function of n_1 , with $B = 250$ (left), $B = 1000$ (center) and mean of $r \times B = 25 \times 250$ (right), for Student's-t sample.

Bibliography

- [1] Beirlant, J., Caeiro, F. and Gomes, M.I. (2012) *An overview and open research topics in statistics of univariate extremes*. *Revstat*, **10**:1, 1–31.
- [2] Caeiro, F. Gomes, M.I. and Pestana, D.D. (2005) *Direct reduction of bias of the classical Hill estimator*. *Revstat*, **3**:2, 111–136.
- [3] Caeiro, F. and Gomes, M.I. (2011) *Asymptotic comparison at optimal levels of reduced-bias extreme value index estimators*. *Statistica Neerlandica*, **65**:4, 462–488.
- [4] Caeiro, F. and Gomes, M.I. (2013) *Asymptotic comparison at optimal levels of minimum-variance reduced-bias tail index estimators*. In Lita da Silva, J. et al. (Eds.), *Advances in Regression, Survival Analysis, Extreme Values, Markov Processes and Other Statistical Applications*, Studies in Theoretical and Applied Statistics, 83–91, Springer Berlin Heidelberg.
- [5] Caeiro, F. and Gomes, M.I. (2014) *A semi-parametric estimator of a shape second order parameter*. In Pacheco, A., Santos, R., Rosário Oliveira, M. and Paulino, C.D. (Eds.), *New Advances in Statistical Modeling and Applications*, 137–144, Springer.
- [6] Charras-Garrido, M. and Lezard, P. (2013) *Extreme Value Analysis: an Introduction*, *Journal de la Société Française de Statistique*, **154**:2, 66–97.
- [7] Danielsson, J., Haan, L. de, Peng, L., and de Vries, C.G. (2001) *Using a bootstrap method to choose the sample fraction in the tail index estimation*. *J. Multivariate Anal.*, **76**, 226–248.
- [8] Deme, E.H., Gardes, L. and Girard, S. (2013) *On the estimation of the second order parameter for heavy-tailed distributions*. *Revstat* **11**:3, 277–299.
- [9] Draisma, G., Haan, L. de, Peng, L., and Pereira, T.T. (1999) *A bootstrap-based method to achieve optimality in estimating the extreme-value index*. *Extremes*, **2**, 367–404.
- [10] Fraga Alves, M.I., Gomes, M.I. and de Haan, L. (2003) *A new class of semiparametric estimators of the second order parameter*. *Portugaliae Math.*, **60**:2, 193–213.
- [11] de Haan, L. and Ferreira, A. (2006) *Extreme Value Theory: An Introduction*. Springer, New York.
- [12] Gomes, M.I. and Guillou, A. (2014) *Extreme Value Theory and Statistics of Univariate Extremes: A Review*. *International Statistical Review*, doi:10.1111/insr.12058
- [13] Gomes, M.I. and Martins, M.J. (2002) *“Asymptotically unbiased” estimators of the tail index based on external estimation of the second order parameter*. *Extremes*, **5**:1, 5–31.
- [14] Gomes M.I. and Oliveira O. (2001) *The bootstrap methodology in Statistics of Extremes: choice of the optimal sample fraction*. *Extremes*, **4**:4, 331–358.
- [15] Gomes, M.I., Figueiredo, F. and Neves M.M. (2012) *Adaptive estimation of heavy right tails: resampling-based methods in action*. *Extremes 2012*, **15**:4, 463–489.
- [16] Hall, P. (1990) *Using the bootstrap to estimate mean squared error and select smoothing parameter in nonparametric problems*. *J. Multivariate Anal.*, **32**, 177–203.
- [17] Hill, B.M. (1975) *A simple general approach to inference about the tail of a distribution*. *Ann. Statist.* **3**:5, 1163–1174.

Local likelihood estimation for multivariate directional data

Marco Di Marzio, *University of Chieti–Pescara*, `mdimarzio@unich.it`

Stefania Fensore, *University of Chieti–Pescara*, `stefania.fensore@unich.it`

Agnese Panzera, *University of Firenze*, `a.panzera@disia.unifi.it`

Charles C. Taylor, *University of Leeds*, `charles@maths.leeds.ac.uk`

Abstract. Our aim is to extend local likelihood methodology to circular density estimation. The idea lies in optimizing a spatially weighted version of the log-likelihood function, where the logarithm of the density is approximated by a polynomial.

Advantages of such an approach would amount to more flexibility near the boundary and bias reduction when the polynomial degree increases, especially for heavy tailed distributions (as it is often the case for directional models) in higher dimensions. The use of d -fold products ($d \geq 1$) of von Mises densities as weight functions facilitates the computational burden, specifically it makes possible to avoid numerical integration by exploiting the properties of Bessel functions.

Our findings consist of theoretical reasoning along with simulation experiments.

Keywords. Bessel functions, Circular data, Density estimation, Product kernels, Toroidal data, von Mises density

1 Introduction

A directional observation can be considered as a point on the circle of unit radius (or a unit vector in the plane) and represented by an angle in $[-\pi, \pi)$ after both an origin and an orientation have been chosen. Typical examples include flight direction of birds from a point of release, wind, and ocean current direction. A circular observation is periodic, i.e. the single observation $\theta \in [-\pi, \pi)$ can be represented by whatever element in the set $\{2m\pi + \theta, m \in \mathbb{Z}\}$. This sets apart circular statistical analysis from standard real-line methods.

Multidimensional versions of such data, which lie in spaces like the torus or the sphere, also occur frequently.

Concerning toroidal data, i.e. data lying in $[-\pi, \pi)^d$ where $d > 1$, in the study of wind directions over a time period, the need to model bivariate circular data naturally arises. In zoology countless instances arise. For example, [1] considers the orientations of the nests of 50 noisy scrub birds (θ) along the bank of a creek bed, together with the corresponding directions

(ϕ) of creek flow at the nearest point to the nest. Here the joint behaviour of the random variable (ϕ, θ) is of interest. In evolutionary biology it is of interest to study paired circular genomes. Data on the (two-dimensional) torus are commonly found in descriptions of protein structure. Here, the protein backbone is given by a set of atom co-ordinates in \mathbb{R}^3 which can then be converted to a sequence of conformation angles.

Basically, when parametric modeling seems too restrictive, a nonparametric approach is preferred. Nonparametric estimation is also employed for exploratory purposes, as a preliminary step. Circular, multivariate nonparametric density estimation has been pursued by [3].

In this paper we propose local likelihood for such types of data, as an extension of the method of [4], who considered euclidean data. The idea lies in optimizing a spatially weighted version of the log-likelihood function, where the logarithm of the density is approximated by a polynomial. In the standard theory higher polynomial degrees give smaller order bias. Consequently, it is showed that the method is equivalent to higher order kernel estimation. Therefore, our contribution can be regarded also as an attempt to build higher order density estimation for directional data, a field, until now, almost unexplored. Now observe that, as it is well known, higher order kernels reduce the bias, whereas a way to represent the *curse of dimensionality* in density estimation setting is observing that the classical bias-variance tradeoff is subject to failure since the optimal bin widths must be large, and are generally too wide to avoid substantial bias. Therefore, higher order estimation should have the potential of improving the efficiency, especially in the regions where the bias is severe.

As a byproduct, we also give a method for estimating the derivative of a toroidal density, recalling that when the interpolating polynomial is p the p -th coefficient of the interpolating polynomial estimates the p -th derivative. Major issues to be resolved in the implementation are the choice of the kernel, which has to be a periodic function, and the local approximation of functions that needs to be specific to the definition domain, in our case the torus.

In Section 2 we give some theory for the case of circular data, whereas in Section 3 we explore the case of toroidal data, when the approximating polynomial has order 0 or 1. In Section 4 we argue that some numerical aspects could be greatly simplified if d -fold products ($d \geq 1$) of von Mises densities are used as kernels. Finally, Section 5 is devoted to numerical experiments.

2 Circular densities

Given a random sample of angles $\theta_1, \dots, \theta_n$, $\theta_i \in [-\pi, \pi)$ for $i \in \{1, \dots, n\}$, from the unknown density f , the log-likelihood function is defined as

$$\mathcal{L}(f) := \sum_{i=1}^n \log(f(\theta_i)) - n \left(\int_{-\pi}^{\pi} f(\alpha) d\alpha - 1 \right). \quad (1)$$

Letting K_κ be a *circular kernel* (see [3]) with concentration parameter $\kappa \in (0, \infty)$, a local version of (1), at $\theta \in [-\pi, \pi)$, can be defined as

$$\mathcal{L}(f, \theta) := \sum_{i=1}^n K_\kappa(\theta_i - \theta) \log(f(\theta_i)) - n \int_{-\pi}^{\pi} K_\kappa(\alpha - \theta) f(\alpha) d\alpha. \quad (2)$$

Now, assume that $\log(f(\theta))$ is smooth enough to be approximated, for α in neighborhood of $\theta \in [-\pi, \pi)$, and $p \in \mathbb{Z}^+$, as

$$\log(f(\theta)) \approx \mathcal{P}_p(\alpha - \theta), \tag{3}$$

with

$$\mathcal{P}_p(\alpha - \theta) := \sum_{j=0}^p \frac{\sin(\alpha - \theta)^j a_j}{j!},$$

see [2] for details. Hence, a local polynomial approximation of (2) is given by

$$\mathcal{L}_p(f, \theta) = \sum_{i=1}^n K_\kappa(\theta_i - \theta) \mathcal{P}_p(\theta_i - \theta) - n \int_{-\pi}^\pi K_\kappa(\alpha - \theta) \exp(\mathcal{P}_p(\alpha - \theta)) d\alpha. \tag{4}$$

Denoting as $\{\hat{a}_0, \dots, \hat{a}_p\}$ the solution of the maximization, over $\{a_0, \dots, a_p\}$, of (4), the *local likelihood density estimator* of f at $\theta \in [-\pi, \pi)$ is defined as

$$\hat{f}(\theta) := \exp(\hat{a}_0). \tag{5}$$

If no maximizer exists, then $\hat{f}(\theta) = 0$. Notice that, differently from the standard setting where if x does not lie in the support of f , then $\hat{f}(x) = 0$, here, due to the periodic nature of f , we have $\hat{f}(\theta) = \hat{f}(\theta \bmod(2\pi))$.

Clearly, the \hat{a}_j s have to satisfy

$$\frac{1}{n} \sum_{i=1}^n \mathcal{A}(\theta_i - \theta) K_\kappa(\theta_i - \theta) = \int_{-\pi}^\pi \mathcal{A}(\alpha - \theta) K_\kappa(\alpha - \theta) \exp(\mathcal{P}_p(\alpha - \theta)) d\alpha, \tag{6}$$

where $\mathcal{A}(\theta_i - \theta) := (1 - \sin(\theta_i - \theta) \dots \sin(\theta_i - \theta)^p)'$, for $i \in \{1, \dots, n\}$.

Then, when $p = 0$, one has the standard circular kernel density estimator

$$\hat{f}(\theta) = \frac{\sum_{i=1}^n K_\kappa(\theta_i - \theta)}{n \int_{-\pi}^\pi K_\kappa(\alpha - \theta) d\alpha}, \tag{7}$$

while, for $p > 0$, $\hat{f}(\theta)$ does not generally have a closed form.

Notice that this circular local likelihood estimate closely recalls the matching of localized sample linear moments up to order p (LHS of (6)) with the corresponding localized population moments when the log-polynomial density approximation is employed.

3 Toroidal densities

Assume that f is a toroidal density, i.e., up to a 2π -periodic behaviour, it is defined on $[-\pi, \pi)^d$, $d > 1$. Here the local likelihood function has to be defined using a toroidal weight (see [3]), say $\mathbf{K}(\cdot; \kappa_1, \dots, \kappa_d)$, which is the d -fold product of circular kernels. If the circular kernels have the same concentration parameter κ , we have:

$$\mathbf{K}(\boldsymbol{\beta}; \kappa_1, \dots, \kappa_d) = \prod_{j=1}^d K_\kappa(\boldsymbol{\beta}^{(j)}),$$

where $\boldsymbol{\beta}^{(j)}$ stands for the j -th co-ordinate of $\boldsymbol{\beta}$.

Hence, a local constant fit of f at $\boldsymbol{\theta} \in [-\pi, \pi)^d$ is defined as (7) with \mathbf{K} in place of K_κ .

Moreover, a local linear fit is still estimator (5), but now \hat{a}_0 is obtained as solution of

$$\begin{aligned} & \frac{1}{n} \sum_{i=1}^n \mathcal{A}(\boldsymbol{\theta}_i - \boldsymbol{\theta}) \mathbf{K}_\kappa(\boldsymbol{\theta}_i - \boldsymbol{\theta}; \kappa_1, \dots, \kappa_d) \\ &= \int_{[-\pi, \pi]^d} \mathcal{A}(\boldsymbol{\alpha} - \boldsymbol{\theta}) \mathbf{K}_\kappa(\boldsymbol{\alpha} - \boldsymbol{\theta}; \kappa_1, \dots, \kappa_d) \exp\left(a_0 + \sum_{j=1}^d \mathbf{a}_1^{(j)} \sin(\boldsymbol{\alpha}^{(j)} - \boldsymbol{\theta}^{(j)})\right) d\boldsymbol{\alpha}, \quad (8) \end{aligned}$$

where, for a point $\boldsymbol{\beta}$, $\mathcal{A}(\boldsymbol{\beta} - \boldsymbol{\theta}) := (1 \ \sin(\boldsymbol{\beta}^{(1)} - \boldsymbol{\theta}^{(1)}) \dots \sin(\boldsymbol{\beta}^{(d)} - \boldsymbol{\theta}^{(d)}))'$.

4 Computational aspects

Clearly, formulas (6) and (8) need to be solved numerically in order to obtain estimates of the coefficients a_0 and $\mathbf{a}_1^{(j)}$. This appears to require numerical integration, which would be rather cumbersome. In this section we indicate a way to avoid numerical integration based on the properties of Bessel functions when products of von Mises densities are used as kernels. In the sequel we will treat in detail only the case $d = 2$, extensions to higher dimensions are easily feasible.

Estimation of $\mathbf{a}_1^{(1)}$ and $\mathbf{a}_1^{(2)}$

When $d = 2$ and $p = 1$ equations in (8) are

$$\begin{aligned} & \frac{1}{n} \sum_{i=1}^n K_\kappa(\boldsymbol{\theta}_i^{(1)} - \boldsymbol{\theta}^{(1)}) K_\kappa(\boldsymbol{\theta}_i^{(2)} - \boldsymbol{\theta}^{(2)}) \\ &= \int_{-\pi}^{\pi} \int_{-\pi}^{\pi} K_\kappa(\boldsymbol{\alpha}^{(1)} - \boldsymbol{\theta}^{(1)}) K_\kappa(\boldsymbol{\alpha}^{(2)} - \boldsymbol{\theta}^{(2)}) \\ & \quad \times \exp\left(a_0 + \mathbf{a}_1^{(1)} \sin(\boldsymbol{\alpha}^{(1)} - \boldsymbol{\theta}^{(1)}) + \mathbf{a}_1^{(2)} \sin(\boldsymbol{\alpha}^{(2)} - \boldsymbol{\theta}^{(2)})\right) d\boldsymbol{\alpha}^{(1)} d\boldsymbol{\alpha}^{(2)}, \quad (9) \end{aligned}$$

$$\begin{aligned} & \frac{1}{n} \sum_{i=1}^n K_\kappa(\boldsymbol{\theta}_i^{(1)} - \boldsymbol{\theta}^{(1)}) K_\kappa(\boldsymbol{\theta}_i^{(2)} - \boldsymbol{\theta}^{(2)}) \sin(\boldsymbol{\theta}_i^{(1)} - \boldsymbol{\theta}^{(1)}) \\ &= \int_{-\pi}^{\pi} \int_{-\pi}^{\pi} K_\kappa(\boldsymbol{\alpha}^{(1)} - \boldsymbol{\theta}^{(1)}) K_\kappa(\boldsymbol{\alpha}^{(2)} - \boldsymbol{\theta}^{(2)}) \sin(\boldsymbol{\alpha}^{(1)} - \boldsymbol{\theta}^{(1)}) \\ & \quad \times \exp\left(a_0 + \mathbf{a}_1^{(1)} \sin(\boldsymbol{\alpha}^{(1)} - \boldsymbol{\theta}^{(1)}) + \mathbf{a}_1^{(2)} \sin(\boldsymbol{\alpha}^{(2)} - \boldsymbol{\theta}^{(2)})\right) d\boldsymbol{\alpha}^{(1)} d\boldsymbol{\alpha}^{(2)} \quad (10) \end{aligned}$$

and

$$\begin{aligned} & \frac{1}{n} \sum_{i=1}^n K_\kappa(\boldsymbol{\theta}_i^{(1)} - \boldsymbol{\theta}^{(1)}) K_\kappa(\boldsymbol{\theta}_i^{(2)} - \boldsymbol{\theta}^{(2)}) \sin(\boldsymbol{\theta}_i^{(2)} - \boldsymbol{\theta}^{(2)}) \\ &= \int_{-\pi}^{\pi} \int_{-\pi}^{\pi} K_\kappa(\boldsymbol{\alpha}^{(1)} - \boldsymbol{\theta}^{(1)}) K_\kappa(\boldsymbol{\alpha}^{(2)} - \boldsymbol{\theta}^{(2)}) \sin(\boldsymbol{\alpha}^{(2)} - \boldsymbol{\theta}^{(2)}) \\ & \quad \times \exp\left(a_0 + \mathbf{a}_1^{(1)} \sin(\boldsymbol{\alpha}^{(1)} - \boldsymbol{\theta}^{(1)}) + \mathbf{a}_1^{(2)} \sin(\boldsymbol{\alpha}^{(2)} - \boldsymbol{\theta}^{(2)})\right) d\boldsymbol{\alpha}^{(1)} d\boldsymbol{\alpha}^{(2)}. \quad (11) \end{aligned}$$

Now, assume that the von Mises kernel is employed, i.e. $K_\kappa(\theta) := \{2\pi I_0(\kappa)\}^{-1} \exp(\kappa \cos(\theta))$, where $I_s(\cdot)$ stands for the modified Bessel function of the first kind and order s . Then the RHS in equation (9) becomes

$$\begin{aligned} \exp(a_0) \int_{-\pi}^{\pi} \frac{1}{2\pi I_0(\kappa)} \exp(\kappa \cos(\boldsymbol{\alpha}^{(1)} - \boldsymbol{\theta}^{(1)})) \exp(\mathbf{a}_1^{(1)} \sin(\boldsymbol{\alpha}^{(1)} - \boldsymbol{\theta}^{(1)})) d\boldsymbol{\alpha}^{(1)} \\ \times \int_{-\pi}^{\pi} \frac{1}{2\pi I_0(\kappa)} \exp(\kappa \cos(\boldsymbol{\alpha}^{(2)} - \boldsymbol{\theta}^{(2)})) \exp(\mathbf{a}_1^{(2)} \sin(\boldsymbol{\alpha}^{(2)} - \boldsymbol{\theta}^{(2)})) d\boldsymbol{\alpha}^{(2)}, \end{aligned}$$

and, expressing the integrals as Bessel functions, this is equal to

$$\exp(a_0) \left(\frac{1}{I_0(\kappa)}\right)^2 I_0\left(\|\kappa \mathbf{a}_1^{(1)}\|\right) I_0\left(\|\kappa \mathbf{a}_1^{(2)}\|\right).$$

As for the integral on the RHS of (10), it can be expressed as

$$\begin{aligned} \exp(a_0) \left(\frac{1}{2\pi I_0(\kappa)}\right)^2 \int_{-\pi}^{\pi} \exp(\kappa \cos(\boldsymbol{\alpha}^{(1)} - \boldsymbol{\theta}^{(1)})) \exp(\mathbf{a}_1^{(1)} \sin(\boldsymbol{\alpha}^{(1)} - \boldsymbol{\theta}^{(1)})) \\ \times \sin(\boldsymbol{\alpha}^{(1)} - \boldsymbol{\theta}^{(1)}) d\boldsymbol{\alpha}^{(1)} \int_{-\pi}^{\pi} \exp(\kappa \cos(\boldsymbol{\alpha}^{(2)} - \boldsymbol{\theta}^{(2)})) \exp(\mathbf{a}_1^{(2)} \sin(\boldsymbol{\alpha}^{(2)} - \boldsymbol{\theta}^{(2)})) d\boldsymbol{\alpha}^{(2)} \end{aligned}$$

and, representing the integrals as Bessel functions, we have

$$\exp(a_0) \left(\frac{1}{I_0(\kappa)}\right)^2 I_1\left(\|\kappa \mathbf{a}_1^{(1)}\|\right) \sin(\text{atan2}(\mathbf{a}_1^{(1)}, \kappa)) I_0\left(\|\kappa \mathbf{a}_1^{(2)}\|\right),$$

where $\text{atan2}(y, x)$ is the angle in radians between the x -axis and the vector from the origin to (x, y) . In the same way, the RHS of (11) becomes

$$\exp(a_0) \left(\frac{1}{I_0(\kappa)}\right)^2 I_1\left(\|\kappa \mathbf{a}_1^{(2)}\|\right) \sin(\text{atan2}(\mathbf{a}_1^{(2)}, \kappa)) I_0\left(\|\kappa \mathbf{a}_1^{(1)}\|\right)$$

so, the ratio between the integrals in the RHSs of (10) and (9) can be expressed as

$$\frac{I_1\left(\|\kappa \mathbf{a}_1^{(1)}\|\right) \sin(\text{atan2}(\mathbf{a}_1^{(1)}, \kappa))}{I_0\left(\|\kappa \mathbf{a}_1^{(1)}\|\right)},$$

this quantity, in order to obtain $\mathbf{a}_1^{(1)}$, has to be set equal with the ratio between the sums on the LHSs of (10) and (9). Now, the ratio between the integrals of (11) and (9) is

$$\frac{I_1\left(\|\kappa \mathbf{a}_1^{(2)}\|\right) \sin(\text{atan2}(\mathbf{a}_1^{(2)}, \kappa))}{I_0\left(\|\kappa \mathbf{a}_1^{(2)}\|\right)}$$

which, in order to obtain $\mathbf{a}_1^{(2)}$, has to be set equal with the ratio between the sums on the LHSs of (11) and (9). Given the numerical solutions $\hat{\mathbf{a}}_1^{(1)}$ and $\hat{\mathbf{a}}_1^{(2)}$ obtained on the basis of above equations, we are finally able to obtain $\hat{f}(\boldsymbol{\theta})$. In particular, from (9) we get

$$\exp(\hat{a}_0) = \frac{\frac{1}{n} \sum_{i=1}^n \exp(\kappa \cos(\boldsymbol{\theta}_i^{(1)} - \boldsymbol{\theta}^{(1)})) \exp(\kappa \cos(\boldsymbol{\theta}_i^{(2)} - \boldsymbol{\theta}^{(2)}))}{4\pi^2 I_0\left(\|\kappa \hat{\mathbf{a}}_1^{(1)}\|\right) I_0\left(\|\kappa \hat{\mathbf{a}}_1^{(2)}\|\right)}.$$

5 Simulations

We consider the performance of the proposed method in the case $d = 1$ and $d = 2$, comparing the standard kernel density estimate (which corresponds to $p = 0$) with the solution for $p = 1$. The solutions for $\hat{\mathbf{a}}_1$ and \hat{a}_0 were found using the calculations of Section 4.

We simulate 200 samples of n observations from a von Mises density with concentration λ . Firstly, we report the integrated squared error averaged over the simulations. Figure 1 shows the estimated $\log(\text{IMSE})$ for $\lambda = 1$ and $n = 100$ and $n = 500$. As expected, a larger κ (corresponding to less smoothing) is required for larger sample sizes, and it can be seen that the standard case ($p = 0$) performs slightly better in this setting.

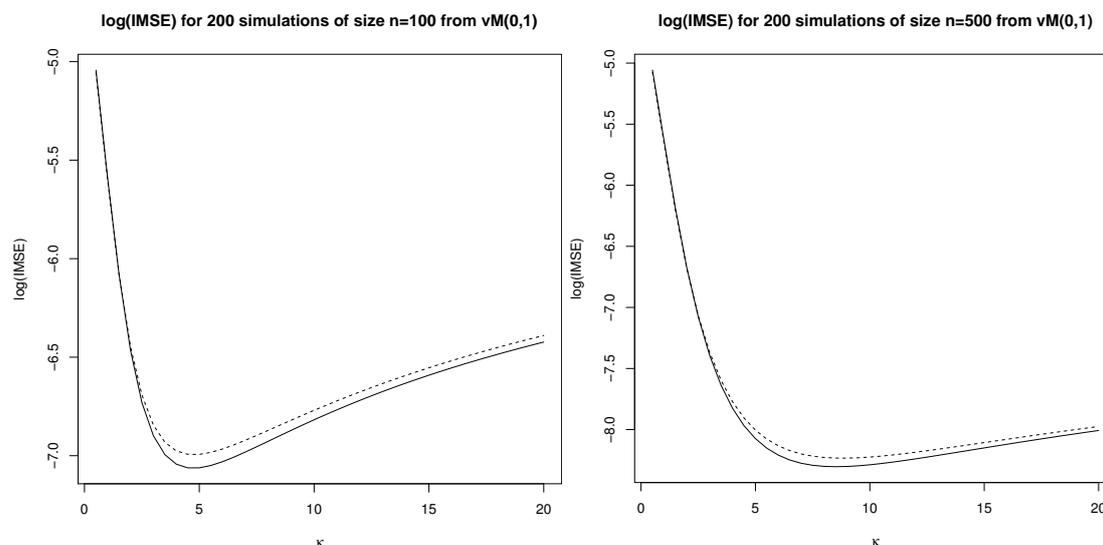


Figure 1: $\log(\text{IMSE})$ for a range of smoothing parameters κ for $p = 0$ (continuous) and $p = 1$ (dashed) for 200 samples of size $n = 100$ (left) and $n = 500$ (right).

Given that the polynomial approximation is expected to be useful in the tails of the density, we also considered the average mean squared error for various values of θ . Here we consider experiments in $d = 1$ and $d = 2$ dimensions, again taking 200 samples — here of size $n = 500$ — from a von Mises distribution with $\lambda = 1$. The results are shown in Figure 2, where it can be seen that the case $p = 1$ performs better than the case $p = 0$ in the tails of the distribution. This is as expected from the linear setting.

For the case $d = 2$ we use a bivariate von Mises distribution with independent components, each with $\lambda = 1$. In this case we estimate f at various pairs $(\theta^{(1)}, \theta^{(2)})$. These results are shown in Figure 3 in which the improvement in the tails can again be seen.

We note that there are numerical difficulties for some cases when $p > 0$. Specifically, the solution(s) for \mathbf{a}_1 ($d = 1$) and $\mathbf{a}_1^{(j)}$, $j = 1, 2$ ($d = 2$) become unstable (due to the computation of the Bessel functions) in the case that κ is large and n is small.

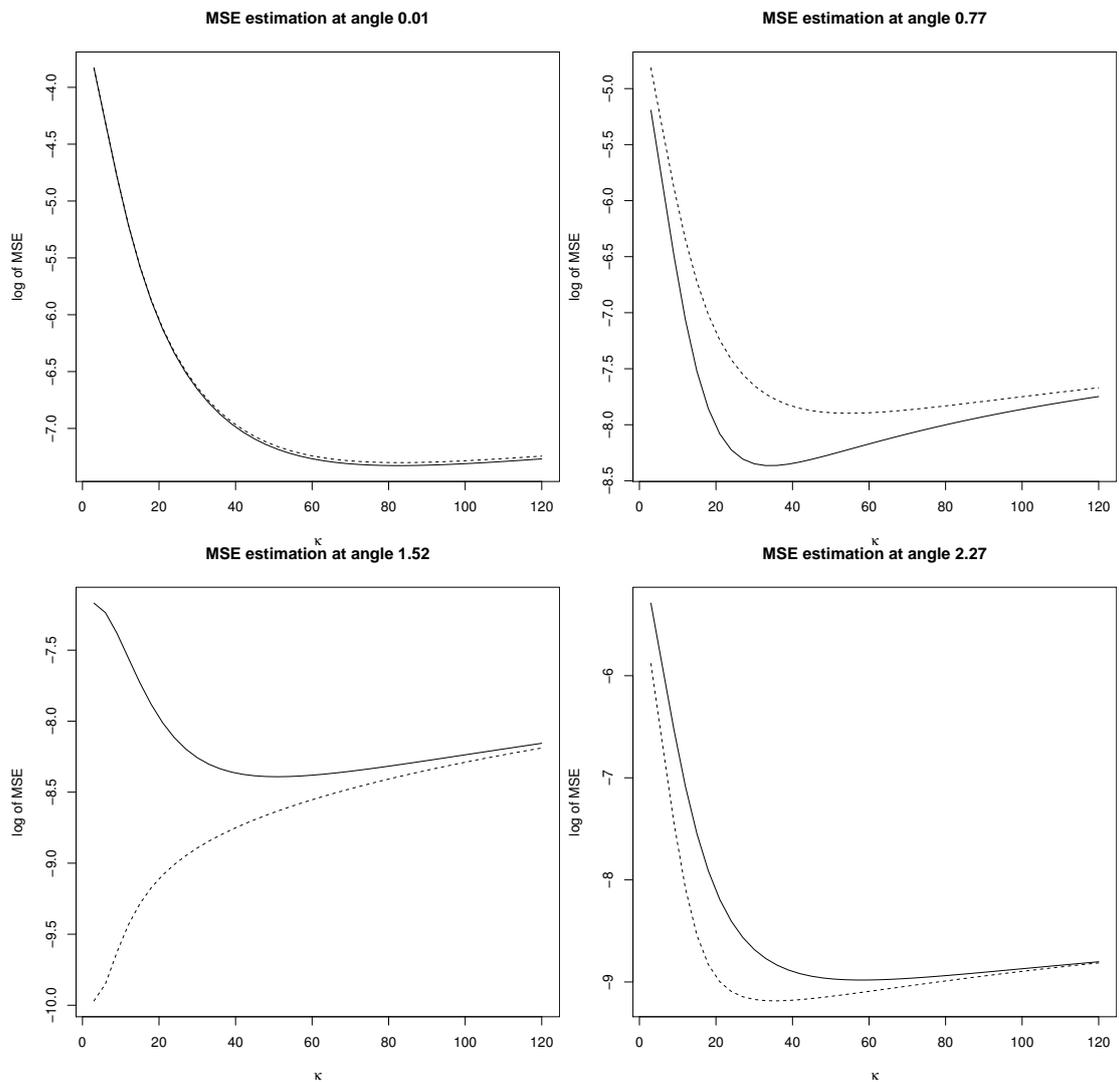


Figure 2: $\log(\text{MSE})$ for a range of smoothing parameters κ for $p = 0$ (continuous) and $p = 1$ (dashed) for 200 samples of size $n = 500$ computed at various θ .

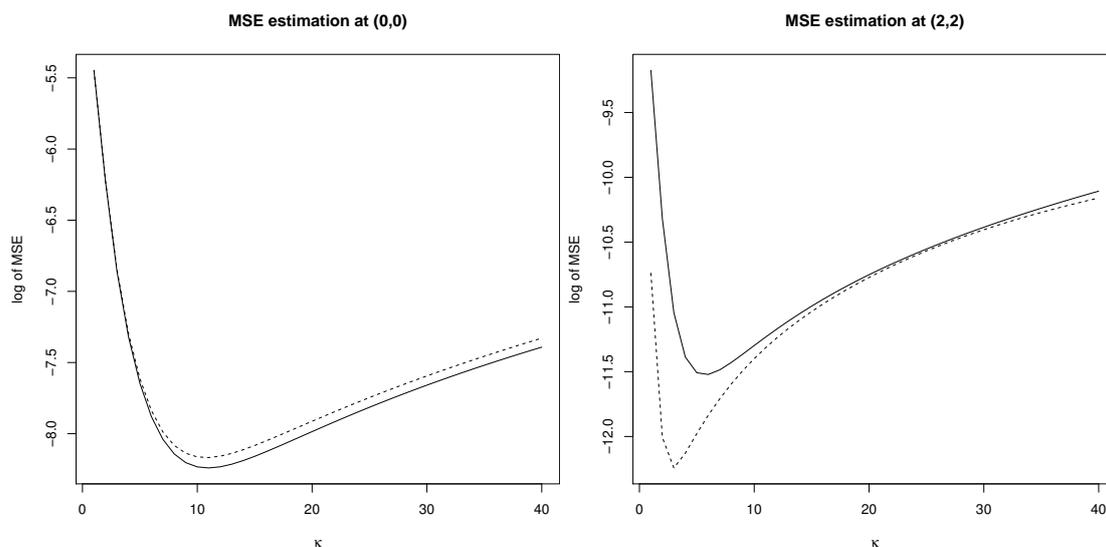


Figure 3: $\log(\text{MSE})$ for a range of smoothing parameters κ for $p = 0$ (continuous) and $p = 1$ (dashed) for 200 samples of size $n = 500$ from a bivariate von Mises distribution with independent components and $\lambda_1 = \lambda_2 = 1$. This is computed at various $(\boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}^{(2)})$, as shown in the titles.

Bibliography

- [1] Fisher, N. I. (1993) *Statistical Analysis of Circular Data*. Cambridge University Press.
- [2] Di Marzio, M., Panzera, A. and Taylor, C.C. (2009) *Local polynomial regression for circular predictors*. *Statistics & Probability Letters*, **79**, 2066–2075.
- [3] Di Marzio, M., Panzera, A. and Taylor, C.C. (2011) *Kernel density estimation on the torus*. *Journal of Statistical Planning and Inference*, **141**, 2156–2173.
- [4] Loader, C.R. (1996) *Local likelihood density estimation*. *The Annals of Statistics*, **24**, 1602–1618.

Estimation of Discrete Partially Directed Acyclic Graphical Models in Multitype Branching Processes

Pierre Fernique, *University of Montpellier 2, I3M and CIRAD, UMR AGAP and Inria, Virtual Plants*, pierre.fernique@inria.fr

Jean-Baptiste Durand, *University of Grenoble Alpes, Laboratoire Jean Kutzmann and Inria, Mistis*, jean-baptiste.durand@imag.fr

Yann Guédon, *CIRAD, UMR AGAP and Inria, Virtual Plants*, guedon@cirad.fr

Abstract. We address the inference of discrete-state models for tree-structured data. Our aim is to introduce parametric multitype branching processes that can be efficiently estimated on the basis of data of limited size. Each generation distribution within this macroscopic model is modeled by a partially directed acyclic graphical model. The estimation of each graphical model relies on a greedy algorithm for graph selection. We present an algorithm for discrete graphical model which is applied on multivariate count data. The proposed modeling approach is illustrated on plant architecture datasets.

Keywords. Partially directed graphical model, graph selection, multivariate discrete distribution, tree pattern, branching process, plant architecture, multivariate count data

1 Introduction

We consider discrete-state stochastic processes indexed by a rooted tree. Our aim is to introduce parametric models that can be efficiently estimated on the basis of data of limited size and that are easily interpretable. These models rely on local dependency assumptions between parent and child vertices and belong to the family of multitype branching processes (MTBPs). In our practical setting of plant architecture analysis, the combinatorics induced by the variable and high number of child vertices in each state induces an inflation in the number of model parameters. We thus introduce parametric MTBPs incorporating parsimonious discrete graphical models for each generation distribution. In order to have interpretable results, we propose to focus on a family of multivariate discrete generation distributions such that:

- child states that tend to appear simultaneously or on the contrary to be incompatible can be identified,
- multivariate parametric distributions can be used since the direct estimation of probability masses on the basis of multivariate counts is unreliable except for very large data sets.
- these multivariate parametric distributions can have zero-inflated and right-skewed marginals, so that multivariate Gaussian distributions are not appropriate.
- these multivariate parametric distributions can be easily simulated and probability masses can be easily computed in order to investigate hypotheses on generation distributions and long range pattern formation in trees.

To achieve this goal, an approach based on probabilistic graphical models [8] to represent the conditional independence relationships for each generation distribution is considered. Three kinds of graphical models are usual: undirected (UG), directed acyclic (DAG), and partially directed acyclic graphical (PDAG) model.

Methods for graph identification were proposed for UGs, using either frequencies to directly estimate probability masses (so-called nonparametric estimation) or mutual information – see [10] and references therein. Under a multivariate Gaussian distribution assumption, an approach based on a L_1 penalization (Lasso) was proposed in [5], with some extension to Poisson distributions and more generally to GLMs [13].

Specific models and methods were developed for DAGs. Most methods for graph identification in DAGs are based on exploring the set of possible graphs using some heuristic (e.g. hill climbing [1]) and by scoring the visited graphs (e.g. using BIC), the graph with highest score being selected – see [8] for a review.

PDAGs, which generalize both UGs and DAGs, have been considered less often in the literature. In such models, both marginal independence relationships and cyclic dependencies between quadruplets of variables (at least) can be represented. A family of such models was proposed using conditional Gaussian distributions, but the problem of graph identification was not addressed [2]. We choose here to use discrete parametric PDAGs to model generation distribution in MTBPs and present a graph identification procedure for PDAGs.

2 Discrete PDAG modeling of generation distributions in MTBPs

Data of interest are tree-indexed sets $\mathbf{x} = (x_t)_{t \in \mathcal{T}}$ where $\mathcal{T} \subset \mathbb{N}$ is the set of vertices of a rooted tree graph $\tau = (\mathcal{T}, \mathcal{A})$ and $\mathcal{A} \subset \mathcal{T} \times \mathcal{T}$ the set of directed edges representing lineage relationships between vertices. By convention, the root of the tree graph has index 0. Let $x_t \in \mathcal{V} = \{0, \dots, K - 1\}$ denote the label of vertex t . Let $pa(\cdot)$ denote the parent of a vertex, $ch(\cdot)$ the children set of a vertex, $an(\cdot)$ the ancestor set of a vertex and $de(\cdot)$ the descendant set of a vertex. These notations also apply to set of vertices – see [8] for graph terminology. We here assume that x_t (resp. \mathbf{x} , τ) is the outcome of a discrete random variable X_t (resp. discrete random vector \mathbf{X} , random rooted tree T).

MTBPs are based on local dependency assumptions between parent and child vertices, more precisely on the following Markovian property – children are independent of their non-

descendants given their parent

$$\forall t \in \mathcal{T}, \mathbf{X}_{ch(t)} \perp\!\!\!\perp \mathbf{X}_{\mathcal{T} \setminus de(t)} \mid X_{pa(t)},$$

and a permutation invariance property – see [6] for details – in order to obtain a more parsimonious model. As a consequence, the joint distribution can be factorized as follows

$$P(T = (\mathcal{T}, \mathcal{A}), \mathbf{X} = \mathbf{x}) \propto P[X_0 = x_0] \prod_{t \in \mathcal{T}} P(\mathbf{N}_t = \mathbf{n}_t \mid X_t = x_t), \tag{1}$$

where $\mathbf{N}_t \mid X = x_t$ is the discrete random vector of the number of children of t in each state given x_t . Therefore the outcome to model is a discrete random vector \mathbf{N}_t for each vertex

$$\mathbf{n}_t = (|\{s \in ch(t) \mid X_s = k\}|)_{k \in \mathcal{V}}.$$

MTBPs are thus specified by K discrete multivariate generation distributions.

We here propose an extension to PDAGs to model these generation distributions. This extension is based on an enlarged family of discrete parametric distributions incorporating multivariate generalizations of the classical univariate discrete parametric distributions: multinomial, negative multinomial and multivariate Poisson [7] distributions and corresponding regressions. Since we focus on a single generation distribution, we will omit in the following the tree indexing and parent state conditioning of each factor in (1). The class of considered PDAGs is such that the generation distribution factorizes as [9]

$$P(\mathbf{N} = \mathbf{n}) = \prod_{c \in \mathcal{C}} P(\mathbf{N}_c = \mathbf{n}_c \mid \mathbf{N}_{pa(c)} = \mathbf{n}_{pa(c)}), \tag{2}$$

where \mathcal{C} denotes a partition of \mathcal{V} such that in each subset, the induced subgraph – so-called chain component – is a connected undirected graph and each subset is connected – if connected – by directed edges.

Usually for each c in \mathcal{C} , $P(\mathbf{N}_c = \mathbf{n}_c \mid \mathbf{N}_{pa(c)} = \mathbf{n}_{pa(c)})$ can be factorize as a product of clique factors [9]. But in the case of multinomial, negative multinomial and multivariate Poisson distributions or regressions, each chain component is complete. PDAGs where chain components are not cliques could be introduced using the UG framework [13]. In such UGs, the graph is in fact a cyclic bidirected graph. This renders far more difficult and less reliable the exploration of long-range patterns in such models as many normalization constants have to be computed (one for each predictor value for a given clique). Therefore we chose to consider PDAGs such that chain components are complete.

Definition 2.1. *A clique directed acyclic graph (CDAG) is a PDAG such that:*

- *each chain component is a clique,*
- *each vertex of a clique has the same parent set,*
- *each parent set belongs to the power set of cliques.*

A probabilistic PDAG model is defined by a PDAG and a specification of the factors in (2). Our approach to identify such models relies on efficient methods for CDAG search and for variable selection in regression. Proposition 2.1 establishes a connection between probabilistic PDAG, CDAG and regression models.

Proposition 2.1.

A probabilistic PDAG model such that:

- each source vertex of the graph is associated with some univariate distribution chosen among the binomial, negative binomial and Poisson distributions and mixtures of such distributions.
- each non-singleton source component of the graph is associated with some multivariate distribution chosen among diverse extensions of the multinomial distribution, the multivariate Poisson distribution and mixtures of such distributions,
- each component of the graph with at least one parent is associated with the corresponding families of univariate and multivariate regression models defined above in the case of source components,

has the same distribution as a CDAG associated with the same parametric families such that for each edge in the CDAG that is not in the PDAG, the corresponding regression coefficient is null.

Proof. Let $G = (\mathcal{V}, \mathcal{E})$ be a PDAG and $\tilde{G} = (\mathcal{V}, \tilde{\mathcal{E}})$ be a CDAG with $\tilde{\mathcal{E}} = \mathcal{E}' \cup \mathcal{E}''$ – where $\mathcal{E}' \cap \mathcal{E}'' = \emptyset$ – such that

$$\mathcal{E}' = \{(u, v) \in \mathcal{E} \mid (v, u) \in \mathcal{E}\} \quad (3)$$

and

$$\mathcal{E}'' = \{(s, t) \in \mathcal{V} \times \mathcal{V} \mid \exists (u, v) \in ne(s) \times ne(t) \cap \mathcal{E} \setminus \mathcal{E}'\} \quad (4)$$

where $ne(\cdot)$ is denoting the set of neighbors of a vertex. Because of equation (3), \tilde{G} has the same chain components as G , since \mathcal{E}' is the set of undirected edges in both \mathcal{E} and $\tilde{\mathcal{E}}$. Equation (4) implies that the set of directed edges in G is included in $\tilde{\mathcal{E}}$: only edges from the neighbors of a parent of a child clique are added to every child clique vertices. As setting the regression coefficient to 0 does not change the conditional distribution, the two models are equivalent. \square

As a consequence of proposition 2.1, given a CDAG and using ML estimators combined with Lasso type estimators [12] for parametric regressions, we select among all PDAGs sharing the same CDAG a sparse PDAG solution with the previously introduced parametric distributions. Therefore the PDAG estimation task is performed using a graph search within a CDAG space which has a cardinal a little bit higher than the DAG space one but far less important than the PDAG space one (see table 1). This graph search can be achieved as for previous algorithms presented in [8] for DAGs using hill climbing, greedy search, first ascent or simulated annealing algorithms. For defining such an algorithm, lemma 2.2 specify how DAG operators (add/remove/reverse directed edges) can be applied to each CDAG. Since the space search graph is not connected using these 3 operators – chain components remain unchanged – two operators specific to CDAGs have been added: chain merging and splitting:

- A pair (c, c') of chain components of \mathcal{C} such that

$$[pa(c) = pa(c') \setminus c] \wedge [ch(c) \setminus c' = ch(c')]$$

will be merged in one chain component c'' which results from the removal of one chain component.

- A vertex from a chain component c can be removed and set to be a parent or a child of c resulting into the addition of one chain component.

Lemma 2.2. *Let \mathcal{M} be a vertex set and let $DAG(\mathcal{M})$ (resp. $CDAG(\mathcal{M})$) denote the set of DAGs (resp. CDAGs) with vertex set \mathcal{M} and $Part(\mathcal{M})$ denote the set of partitions of \mathcal{M} . There is a one-to-one mapping between $CDAG(\mathcal{V})$ and $\{DAG(p) \mid p \in Part(\mathcal{V})\}$.*

Proof. For $\mathcal{G} \in CDAG(\mathcal{V})$, let $\mathcal{C}(\mathcal{G}) \in Part(\mathcal{V})$ be the set of chain components of \mathcal{G} . Let $\sigma(\mathcal{G})$ be the DAG with vertex set $\mathcal{C}(\mathcal{G})$ and such that (c, c') is an edge if there exists an edge from $a \in c$ to $b \in c'$ in \mathcal{G} . It is easily seen that σ is a bijection from $CDAG(\mathcal{V})$ to $\{DAG(p) \mid p \in Part(\mathcal{V})\}$ since every chain component c of $\mathcal{G} \in CDAG(\mathcal{V})$ is a clique, and since all vertices in c have the same parents. □

Proposition 2.2.

Let b_K (resp. a_K) be the number of labeled CDAGs (resp. DAGs) of K vertices. One have:

$$b_K = \sum_{k=1}^K \left\{ \begin{matrix} K \\ k \end{matrix} \right\} a_k \tag{5}$$

where $\left\{ \begin{matrix} K \\ k \end{matrix} \right\}$ denote the Stirling number of second kind.

Proof. Consider a set of K vertices. The Stirling number of second kind gives the number of ways of partitioning such vertex set into k non-empty cliques. For each of these partitions, a DAG can be defined (see lemma 2.2) and there are a_k such labeled DAGs. We then just need to consider that the number of cliques can vary from 1 to K for CDAGs of K vertices to prove proposition 2.2. □

a_K	b_K	c_K	K
1	1	1	1
3	4	4	2
25	34	50	3
543	715	1,688	4
29,281	35,381	142,624	5
3,781,503	4,258,357	28,903,216	6
1,138,779,265	1,222,487,933	13,663,125,680	7
783,702,329,343	816,625,721,787	14,762,428,500,992	8

Table 1: Number a_K of DAGs, b_K of CDAGs and c_K PDAGs [11] from 1 to 8 vertices (see proposition 2.2)

3 Characterizing the apple tree irregular bearing phenomenon using MTBP and CDAG models

Recently, statistical indices have been proposed to characterize alternation in flowering at whole plant scale with a yearly time step for different apple tree cultivars [4]. A correlation has

been highlighted between synchronicity of flowering within plants, alternation along axes, and alternation at whole plant scale. However, little is known about structural factors that may induce heterogeneity in the fates (vegetative or flowering) of sibling shoots, and thus improve regularity at whole plant scale despite alternation along axes. Considering the methodology described in [3], a tree structure (see fig. 1) was built from the apple tree dataset provided by E. Costes (UMR AGAP, AFEF Team, Inra, Montpellier, France) in order to illustrate the interest of MTBPs to investigate this phenomenon.

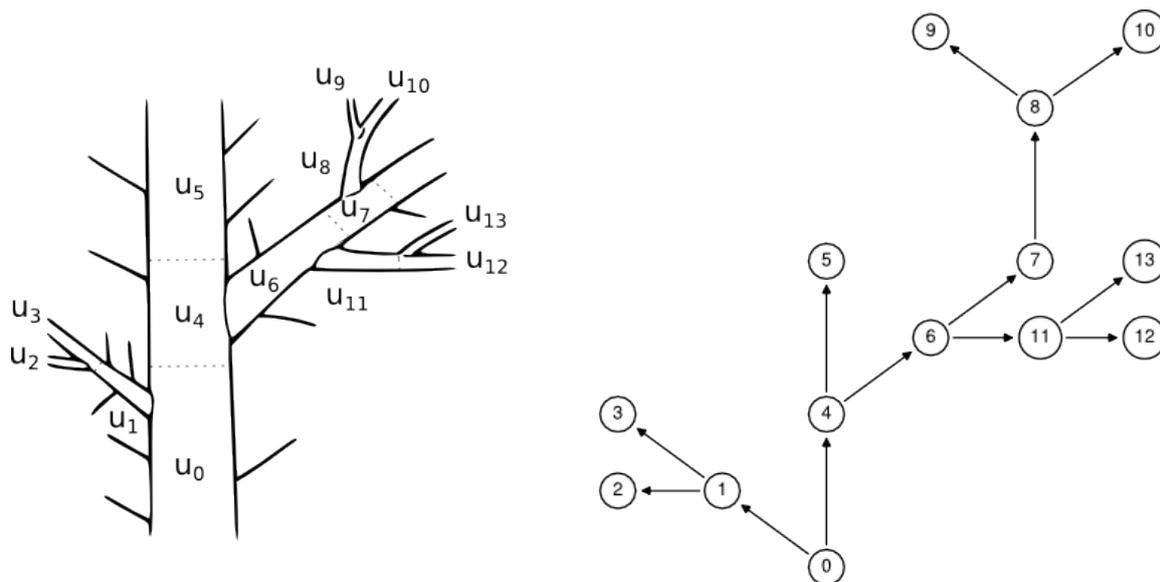


Figure 1: The tree is a formal representation of the plant topological information (drawing issued from [3]). Each label of this tree is the nature of the annual shoot.

MTBPs are used to model the number of flowering and vegetative shoots for parent shoots of different natures defined by their length and fate (see table 2). The aim is to identify parent states associated with homogeneous child fates from parents that may have heterogeneous child fates. As the dataset is composed of two trees per cultivar the objective is also to compare the two cultivars Fuji and Braeburn that have different behaviors regarding the irregular bearing phenomenon.

State	Length	Fate
0	Long	Vegetative
1	Long	Flowering
2	Medium	Vegetative
3	Medium	Flowering
4	Short	Vegetative
5	Short	Flowering

Table 2: Shoots state space and corresponding lengths and fates

CDAG-based generation distributions better fitted the data than DAG-based generation

distributions according to BIC. In the worst case, we obtained the same fit with CDAG-based and DAG-based generation distributions (see fig. 2).

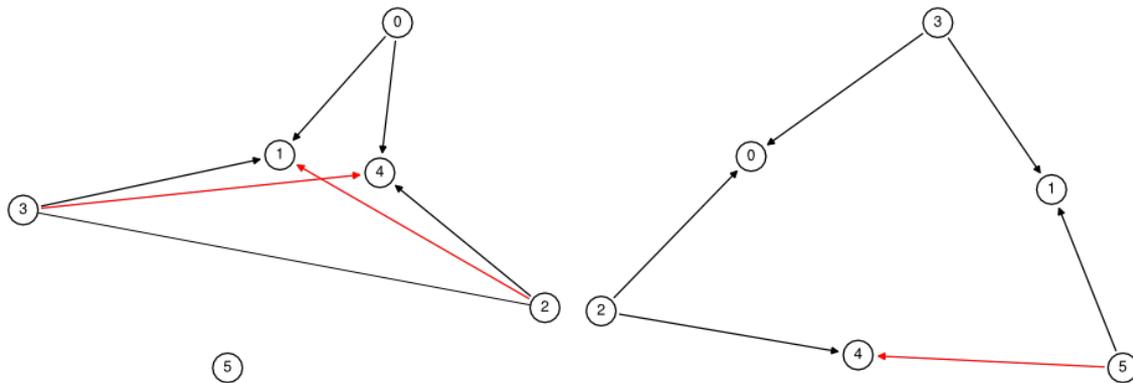


Figure 2: CDAG and DAG selected for the parent states 0 (left hand) and 1 (right hand) for the Braeburn cultivar. Edges associated with negative (resp. positive) covariances are in red (resp. black).

We obtained very contrasted graphs for the different parent states of a given cultivar. We also obtained different graphs for the two cultivars for some parent states. This was very informative for cultivar comparison (see fig. 2 and 3) The exam of the different graphs for a given cultivar highlights the more or less regular bearing behavior at the whole plant scale. Moreover comparing the graphs for the two cultivars leads to a better understanding of the biological functions underlying bearing behavior. This approach seems therefore promising to highlight pattern formation such as irregular bearing in tree structure development.

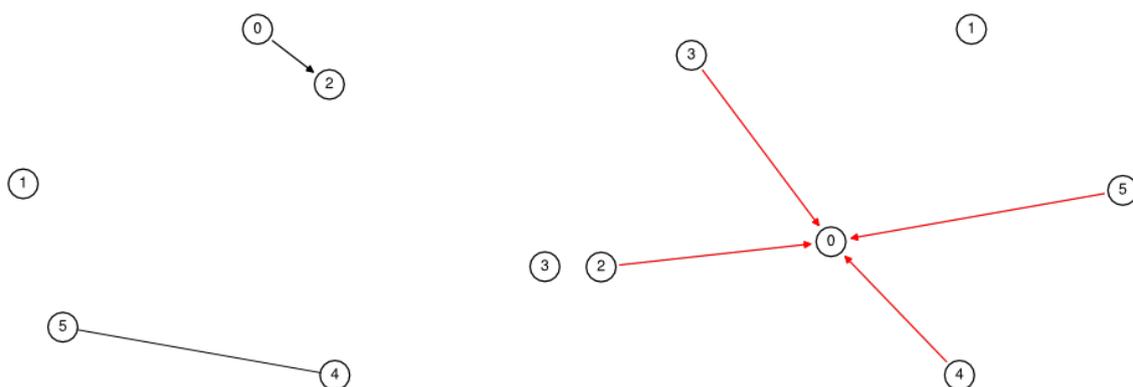


Figure 3: CDAG selected for the parent state 3 for the Braeburn (left hand) and the Fuji cultivar (right hand). Edges associated with negative (resp. positive) covariances are in red (resp. black).

Bibliography

- [1] D.M. Chickering. Learning equivalence classes of bayesian-network structures. *The Journal of Machine Learning Research*, 2:445–498, 2002.
- [2] Mathias Drton and Michael Eichler. Maximum likelihood estimation in Gaussian chain graph models under the alternative markov property. *Scandinavian journal of statistics*, 33(2):247–257, 2006.
- [3] J-B Durand, Y Guédon, Y Caraglio, and E Costes. Analysis of the plant architecture via tree-structured statistical models: the hidden Markov tree models. *New Phytologist*, 166(3):813–825, 2005.
- [4] Jean-Baptiste Durand, Baptiste Guitton, Jean Peyhardi, Yan Holtz, Yann Guédon, Catherine Trottier, and Evelyne Costes. New insights for estimating the genetic value of segregating apple progenies for irregular bearing during the first years of tree production. *Journal of experimental botany*, 64(16):5099–5113, 2013.
- [5] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Sparse inverse covariance estimation with the graphical Lasso. *Biostatistics*, 9(3):432–441, 2008.
- [6] Patsy Haccou, Peter Jagers, and Vladimir A Vatutin. *Branching processes: variation, growth, and extinction of populations*. Cambridge University Press, 2005.
- [7] D. Karlis. An EM algorithm for multivariate Poisson distribution and related models. *Journal of Applied Statistics*, 30(1):63–77, 2003.
- [8] D. Koller and N. Friedman. *Probabilistic graphical models: principles and techniques*. MIT press, 2009.
- [9] S.L. Lauritzen. *Graphical models*, volume 17. Oxford University Press, USA, 1996.
- [10] P.E. Meyer, F. Lafitte, and G. Bontempi. minet: A R/Bioconductor Package for Inferring Large Transcriptional Networks Using Mutual Information. *BMC bioinformatics*, 9(1):461, 2008.
- [11] Bertran Steinsky. Enumeration of labelled chain graphs and labelled essential directed acyclic graphs. *Discrete Mathematics*, 270(1):267–278, 2003.
- [12] Robert Tibshirani. Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.
- [13] Eunho Yang, Pradeep Ravikumar, Genevera Allen, and Zhandong Liu. Graphical models via generalized linear models. In *Advances in Neural Information Processing Systems 25*, pages 1367–1375, 2012.

Statistical modelling in time series extremes: an overview and new steps

Manuela Neves, *University of Lisbon and CEAUL*, manela@isa.ulisboa.pt
Clara Cordeiro, *University of Algarve and CEAUL*, ccordei@ualg.pt

Abstract. Unlike most traditional central statistical theory, which typically examines the usual (or the average) behaviour of a process, extreme value theory deals with models for describing unusual behaviour or rare events. The heart of extreme value theory is the reliable extrapolation of values beyond the observed range of sample data. Modelling rare events of univariate time series is an area of important research. Dealing with extremes of a time series needs specific statistical procedures based on the behaviour of extremes. For modelling and forecasting time series, `Boot.EXPOS` is a computational procedure built in R environment that has revealed itself to perform quite well in a large number of forecasting competitions. A modification of that algorithm is proposed in this work to model time series extreme values. An heuristic study of that procedure was performed and usual accuracy measures were calculated.

Keywords. Extreme values, modelling, resampling procedures, time series.

1 Introduction, motivation and scope of the paper

Statistical analysis of extreme values was traditionally applied to hydrology and insurance. Nowadays, there is a quite large variety of fields of application of extreme value theory such as Climatology, Material Science, Ocean Engineering, Structural Engineering, Material Strength, Environment and Biology.

Extreme value models were initially obtained through arguments that assumed an underlying process consisting of a sequence of independent and identically (i.i.d.) random variables. However in many situations where extreme value models are of great interest to be applied, temporal independence is unrealistic. The most natural generalization of a sequence of independent random variables is a stationary setup. In the last decades many progresses have been made in parameter estimation of extreme values in time series, with relevance to asymptotic results. However, for finite samples, limiting results provide approximations that can be poor. Computer intensive methods, among which we refer to Generalized Jackknife and Bootstrap methodologies [8, 3], have recently shown to improve results in parameter estimation in statistics of extremes [7]. Regarding modelling and forecasting time series, [1] have developed a computational pro-

cedure, Boot.EXPOS, built in R environment. It is based on exponential smoothing methods jointly with the bootstrap methodology. That procedure showed competitive results compared with the best procedures available, see [2].

The main motivation of this paper is to present some challenging steps for modelling time series extreme values. The paper is structured as follows. In Section 2 some notations and main results in extreme value theory both for independent and for dependent sequences are introduced. Particularly when a stationary sequence verifies some conditions, a new parameter can appear in the limit law of the maximum of the sequence. Its definition, some classical estimators and a recent reduced bias estimator based on Jackknife methodology are also presented. Resampling techniques and their application together with exponential smoothing methods for modelling and prediction a time series are reviewed in Section 3. In this section a modification of that computational procedure is proposed and an heuristic study is performed. This study is illustrated through an application to a data set included in the R database. The objective of the approach here introduced is to go further to improve the performance of the estimators addressed in Section 2 through more efficient bootstrap procedures. This paper ends with a few comments and notes on work in progress.

2 Notation and main results in extreme value theory

The classical limiting results in Extreme Value Theory (EVT) were derived for i.i.d. random variables (X_1, \dots, X_n) with unknown distribution function (d.f) F . One wants to know the distribution of $M_n \equiv X_{n:n} := \max(X_1, \dots, X_n)$ or $m_n \equiv X_{1:n} := \min(X_1, \dots, X_n)$. Given the “kind of symmetry” between the maximum and the minimum, $\min(X_1, \dots, X_n) = -\max(-X_1, \dots, -X_n)$, theory was usually developed for the maxima.

First results for the existence of a non-degenerate limit law for i.i.d. variables date back to the twenties but were completely established by [4]. However in many practical applications extreme values often persist over several consecutive observations, i.e. the random variables are no longer independent. In such a situation it appears a new parameter that needs to be taken into account in any inferential procedure.

Limiting results in i.i.d. situation

[4] gave necessary and sufficient conditions for the existence of sequences $\{a_n\} \in \mathbb{R}^+$ and $\{b_n\} \in \mathbb{R}$ such that,

$$\lim_{n \rightarrow \infty} P\left(\frac{M_n - b_n}{a_n} \leq x\right) = \lim_{n \rightarrow \infty} F^n(a_n x + b_n) = EV_\gamma(x), \quad \forall x \in \mathbb{R}. \quad (1)$$

EV_γ is designated by Extreme Value d.f. and given by

$$EV_\gamma(x) = \begin{cases} \exp[-(1 + \gamma x)^{-1/\gamma}], & 1 + \gamma x > 0 & \text{if } \gamma \neq 0 \\ \exp[-\exp(-x)], & x \in \mathbb{R} & \text{if } \gamma = 0. \end{cases} \quad (2)$$

The EV_γ incorporates the three Fisher-Tippett types: the Gumbel type, $\Lambda(x) \equiv EV_0 = \exp(-\exp(-x))$, $x \in \mathbb{R}$, ($\gamma = 0$); the Fréchet type, $\Phi_\alpha(x) \equiv EV_{1/\alpha}(\alpha(x-1)) = \exp(-x^{-\alpha})$, $\alpha > 0$, ($x > 0$, $\gamma = 1/\alpha > 0$) and the Weibull type, $\Psi_\alpha(x) \equiv EV_{-1/\alpha}(\alpha(x+1)) = \exp(-(-x)^\alpha)$, $\alpha > 0$ ($x < 0$, $\gamma = -1/\alpha < 0$).

The shape parameter, γ , determines the tail behaviour of a distribution and is directly related to the weight of the right tail, $\bar{F} := 1 - F$, of the underlying model F . Its estimation is then of primordial importance.

A function F for which limit (1) holds is said to be in the max-domain of attraction of EV_γ , and we write $F \in D_M(EV_\gamma)$. This means that for large values of n we may consider the approximation $P[M_n \leq x] = F^n(x) \approx EV_\gamma((x - a_n)/b_n)$, for adequate $a_n > 0$ and $b_n \in \mathbb{R}$.

We can further consider location and scale parameters, $\mu \in \mathbb{R}$ and $\sigma > 0$, respectively, in the EV_γ d.f., denoting it by $EV_\gamma(x; \mu, \sigma) = EV_\gamma((x - \mu)/\sigma)$.

Whenever the original scheme is no longer identically distributed but it remains independent the limiting results referred to before may hold true. However, when it is not possible to assume independence, the limiting results are also verified under adequate dependence conditions, but a new parameter can appear.

Maxima of stationary sequences

Let us begin with the *example 1*:

Let us consider the following sequences: $\{X_n\}_{n \geq 1}$ i.i.d. variables from the model $F(x) = (1 - \exp(-x))^2$, $x \geq 0$ and the two-dependent sequence $\{Y_n\}_{n \geq 1}$ defined as $Y_n = \max(Z_{n+1}, Z_n)$, $n \geq 1$, where Z_n are exponential i.i.d., i.e, $H(z) = 1 - \exp(-z)$, $z \geq 0$. The underlying model for Y_n is then given by $F(y) = P[Z_{n+1} \leq y, Z_n \leq y] = (1 - \exp(-y))^2$ $y \geq 0$.

So, the i.i.d. sequence $\{X_n\}_{n \geq 1}$ and the two-dependent sequence $\{Y_n\}_{n \geq 1}$ present the same distribution. Fig. 1 shows the plot of several values obtained from the $\{X_n\}$ and $\{Y_n\}$ sequences. Clusters of exceedances of high levels, of size equal to 2, for the $\{Y_n\}$ sequence can be seen. Actually these clusters size are related with a new parameter, designated as the *extremal index*. It can be shown that, in this case, it is equal to $1/2$.

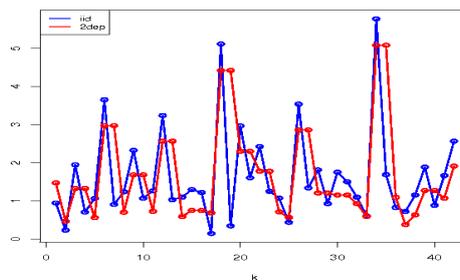


Figure 1: Plot of several values of an i.i.d. sequence (in blue) and a dependent sequence with the same distribution (in red). See clusters of size equal to 2 and the shrinkage of maximum values.

Within dependent situations, stationary series appear as the most natural generalization of independent sequences. For stationary series the characterization of the extreme behaviour was done by imposing some conditions on the dependence at extreme levels.

It is usual to assume a condition that limits the extend of long range dependence at extreme levels. Under that condition events $\{X_i > u\}$ and $\{X_j > u\}$ are approximately independent provided u is high enough and i and j are quite distant. This is expressed in the

$D(u_n)$ condition, see [13]. If $D(u_n)$ condition is satisfied the limit in (1) arises for the maxima of dependent data. [13] stated that if $\{X_n\}_{n \geq 1}$ is a stationary sequence with marginal distribution F , $\{\widetilde{X}_n\}_{n \geq 1}$ an i.i.d. sequence of r.v.'s with the same distribution F , $M_n := \max(X_1, \dots, X_n)$ and $\widetilde{M}_n := \max(\widetilde{X}_1, \dots, \widetilde{X}_n)$, then under the $D(u_n)$ condition, with $u_n = a_n x + b_n$, $Pr \{(\widetilde{M}_n - b_n)/a_n \leq x\} \xrightarrow{n \rightarrow \infty} G_1(x)$ as $n \rightarrow \infty$, for normalizing sequences $\{a_n > 0\}$ and $\{b_n\}$, where $G_1(\cdot)$ is a non degenerate d.f., if and only if, $Pr \{(M_n - b_n)/a_n \leq x\} \xrightarrow{n \rightarrow \infty} G_2(x)$ where $G_2(x) = G_1^\theta(x)$, for a constant θ such that $0 < \theta \leq 1$. This constant is *the extremal index*.

Since the extremal types theorem implies that the only possible non-degenerate limit law is an *EV* distribution, we have $G_2(x) \equiv G_1^\theta(x)$, an extreme value distribution with location, scale and shape parameters $(\mu_\theta, \sigma_\theta, \gamma_\theta)$ given by $\mu_\theta = \mu - \sigma(1 - \theta^\gamma)/\gamma$, $\sigma_\theta = \sigma\theta^\gamma$ and $\gamma_\theta = \gamma$, where (μ, σ, γ) are the location, scale and shape parameters of the extreme value distribution $G_1 \equiv EV_\gamma$.

The estimation of θ is then important not only by itself but because it affects the estimation of other parameters.

The Extremal Index: definition and estimators

This parameter, θ , is a measure of dependence related to the clusters of exceedances of high thresholds. If the sequence $\{X_n\}$ is independent, then $\theta = 1$, however the reciprocal is not true, i.e. for certain dependent series one can have $\theta = 1$.

The most common interpretation of θ is as being the reciprocal of the “mean time of duration of extreme events” what is directly related to the exceedances of high levels, i.e., $\theta = (\text{limiting mean size of clusters})^{-1}$, see [12]. To identify clusters of high level exceedances is then a key issue. One of the easiest way is to identify them by the occurrence of downcrossings or upcrossings. We can write $\theta = \lim_{n \rightarrow \infty} Pr[X_2 \leq u_n | X_1 > u_n] = \lim_{n \rightarrow \infty} Pr[X_1 \leq u_n | X_2 > u_n]$. This interpretation suggested the classical estimator of θ , the so-called Up-Crossing estimator (or Down-Crossing estimator), [14] and [5], defined as:

$$\widehat{\Theta}_n^{UC} := \frac{\sum_{i=1}^{n-1} I(X_i \leq u_n < X_{i+1})}{\sum_{i=1}^n I(X_i > u_n)} \equiv \frac{\sum_{i=1}^{n-1} I(X_i > u_n, X_{i+1} \leq u_n)}{\sum_{i=1}^n I(X_i > u_n)} := \widehat{\Theta}_n^{DC} \tag{3}$$

where $I(A)$ is the indicator function of A .

Other estimators have appeared in the literature, motivated by other forms of cluster identification, such as the *blocks estimator* and the *runs estimator*. As main references see [10, 16]. Given a normalized level u_n and defining a cluster as the set of exceedances of the threshold u_n that occur in an arbitrary block of length r_n , with $r_n = o(n)$, given that at least one exceedance occurs in the block, *blocks estimator* and *runs estimator*, [10] and [16], are defined as

$$\widehat{\Theta}_n^B := \frac{\sum_{i=1}^{k_n} I(\max(X_{(i-1)r_n+1}, \dots, X_{ir_n}) > u_n)}{\sum_{i=1}^n I(X_i > u_n)}; \quad k_n = [n/r_n], \tag{4}$$

$$\widehat{\Theta}_n^R := \frac{\sum_{i=1}^n I(X_i > u_n, \max(X_{i+1}, \dots, X_{i+r_n-1}) \leq u_n)}{\sum_{i=1}^n I(X_i > u_n)}. \tag{5}$$

Despite of having good asymptotic properties all those estimators present high variance for high levels, a high bias when the level decreases and a mean squared error (MSE) with a

very sharp pattern, showing then a strongly dependence on the high threshold u_n , for finite samples. Figure 2 shows a graphical display of these characteristics, for the sequence $\{Y_n\}$ given in Example 1 and for the *runs estimator* and the *blocks estimator*.

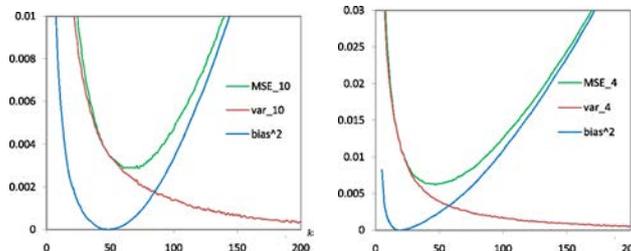


Figure 2: MSE, Bias² and Variance of the blocks estimator (left, $r_n = 10$) and the runs estimator (right, $r_n = 4$) for model in example 1 ($\theta = 0.5$).

Computational procedures have been used in statistics of extremes for dealing with those difficulties. The Generalized Jackknife methodology has the property of estimating the bias and the variance of any estimator, helping to build estimators with bias and mean squared error often smaller than those of an initial set of estimators.

Using Generalized Jackknife methodology [8], a reduced-bias Generalized Jackknife estimator of order 2, $\widehat{\Theta}^{GJ}$, was proposed by [6]. It is based on the estimator $\widehat{\Theta}^{UC}$ computed at the three levels, k , $[k/2] + 1$ and $[k/4] + 1$, ($[x]$ denotes, as usual, the integer part of x , and k is now a deterministic level, such that the upper order statistic $Y_{n-k:n} =: u_n$).

$$\widehat{\Theta}^{GJ}(k) := 5\widehat{\Theta}^{UC}([k/2] + 1) - 2(\widehat{\Theta}^{UC}([k/4] + 1) + \widehat{\Theta}^{UC}(k)). \tag{6}$$

The estimator $\widehat{\Theta}^{GJ}$ outperforms the associated classical estimator, see [6] among other authors. Figure 3 shows the simulated mean values for $\widehat{\Theta}^{UC}$ and $\widehat{\Theta}^{GJ}$ estimators (on left) and a sample path (on right), for the process $\{Y_n\}$ in the Example 1.

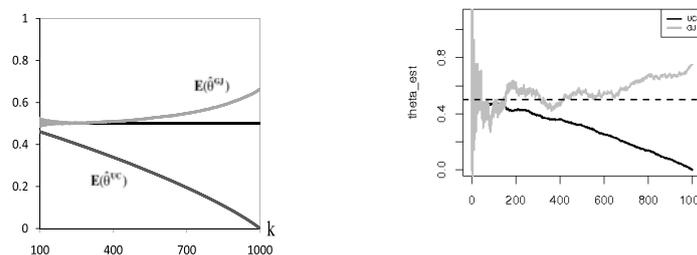


Figure 3: Simulated mean values for $\widehat{\Theta}^{UC}$ and $\widehat{\Theta}^{GJ}$ estimators in the Example 1 (left) and a sample path (right).

Given a sample one needs to choose the number k of upper order statistics to obtain the estimate. Recently the use of adequate bootstrap procedures allowed an improvement in the finite sample behaviour of the estimators, see for example [7].

3 Bootstrap under dependence

Classical bootstrap methodology, [3], has proven to be a powerful nonparametric tool when based on i.i.d. observations. But [15] showed that it could be inadequate under dependence. Nonparametric versions of the Bootstrap and Jackknife applicable to weakly dependent stationary sequences have appeared, considering to resample or to delete one-by-one whole blocks of observations to obtain consistent procedures for estimating a parameter of the stationary series' distribution. The motivation for this scheme is to preserve the dependence structure of the underlying model within each block. Several ways of blocking have been meanwhile pointed out, depending on the way how blocks of observations are defined. However for each procedure it is necessary to consider a block length $b \equiv b(n)$ to be used to resample. However, the accuracy of block bootstrap estimators depends critically on the block size for resampling [11].

Modelling and forecasting time series

Based on exponential smoothing methods and on bootstrap methodology, [1] built a computational procedure for modelling and forecasting time series. That procedure chooses, among a set of models, the one that best fits the data. Sieve bootstrap principle is applied to the residuals. A sketch of algorithm steps, in parallel with the sieve bootstrap procedure follows:

Sieve bootstrap	Boot.EXPOS
Step 0: Select the best $AR(p)$ model; AIC criterion;	Step 0: Select the best EXPOS model to fit the data by MSE; the residuals, r_i are obtained;
Step 1: Adjust an $AR(p)$ model; AIC criterion;	Step 1: Fit an $AR(p)$ to the residuals r_i by AIC criterion;
Step 2: Obtain the residuals, e_i ;	Step 2: Obtain the residuals, e_i ;
<i>For B replicates:</i>	<i>For B replicates:</i>
Step 3: Resample the centered residuals $\rightarrow e_i^*$;	Step 3: Resample the centered residuals $\rightarrow e_i^*$;
Step 4: Use AR coefficients (step 1) and e_i^* for obtaining a new series by recursion;	Step 4: Use AR coefficients (step 1) and e_i^* for obtaining a new series by recursion;
Step 5: Fit an $AR(p)$ to the new series;	Step 5: Use \hat{y} (step 0) to obtain a bootstrap series y^* ;
Step 6: Obtain the predicted values using the $AR(p)$ model fitted.	Step 6: Forecast using the previous EXPOS model fitted.

Table 1: Sketch of the computational procedure compared with Sieve bootstrap

Despite good results that Boot.EXPOS has shown, it was designed for modelling and forecasting the mean behaviour of the series rather than extremes. Extreme values are hardly estimated when Boot.EXPOS is applied.

Bootstrapping extremes

To resample the data for approximating the distribution of the k largest observations would not work because the "pseudo-samples" would never have values greater than the maximum, M_n . Given a sample of size n , [9] suggested to resample a subsample of size $n_1 = O(n^{1-\epsilon})$ with $0 < \epsilon < 1$ and to use the knowledge of the amount by which the two samples differ to estimate mean square error and to select the optimal smoothing parameter for deriving a bootstrap

estimator of a functional of $\{X_1, \dots, X_n\}$. [9] considered several cases: nonparametric density estimation, nonparametric regression and tail estimation. For example, for estimating k_0 , the optimal number of upper ordered statistics to be used in the estimators discussed before he showed that $k_0(n) \simeq cn^\beta$, $0 < \beta < 1$, with β known constant and c unknown. Under certain conditions and for a given class of models, he proposed $\hat{k}_0 \simeq \hat{k}_{1,0}(n/n_1)^\beta$, with $\beta = 2/3$ and $\hat{k}_{1,0}$ the value chosen in order to minimize the mean square error of the estimator using the sample of size n_1 .

We are now exploring that idea in modelling time series extremes, taking advantage of the good performance of Boot.EXPOS algorithm. The resample procedure consists of drawing subsamples of size $n_1 = \lceil n^{0.999} \rceil$ of the residuals obtained in Step 2 of the algorithm. Boot.EXPOS is performed and the final estimated values are multiplied by $(n/n_1)^\beta$. The series `UKDriverDeaths`, available in R, is used to show an application of that procedure. Figure 4 shows the resampled values of the series, calculated on the basis of 1000 replicates and using $\beta = 2/3$. Among several accuracy measures usually calculated, we mention the values of the *root mean squared error* (RMSE), *mean absolute error* (MAE) and *mean absolute percentage error* (MAPE).

The values obtained in this example were RMSE=128.99: MAE=102.97 and MAPE=6.26.

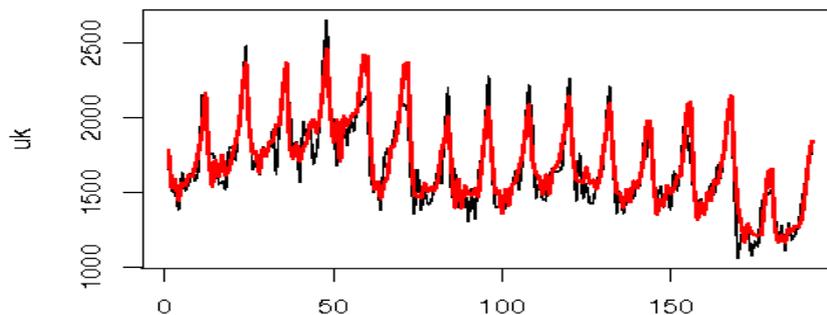


Figure 4: Observed series (in black), bootstrapped series (in red).

4 Concluding remarks

To resample data in a dependent situation needs to preserve the dependence structure and is a problem not yet completely solved. These are the first steps on extending Boot.EXPOS procedure to time series extremes in order to obtain “good” resampled extreme values and use them to improve parameter estimation. The idea is to build bootstrap estimators versions for parameters of extreme events in order to overcome the difficulties that classical estimators have shown. Forecasting extreme values is also another topic for future research.

Acknowledgement

Research partially supported by National Funds through **FCT**—Fundação para a Ciência e a Tecnologia, project PEst-OE/MAT/UI0006/2014 (CEAUL).

Bibliography

- [1] Cordeiro C. and Neves, M. (2009) *Forecasting time series with Boot.EXPOS procedure*. Revstat, **7**:2, 135–149.
- [2] Cordeiro, C. and Neves, M. (2010) *Boot.EXPOS in NNGC competition*. Proceedings of the IEEE World Congress on Computational Intelligence (WCCI 2010), 1135–1141.
- [3] Efron, B. (1979) *Bootstrap methods: another look at the jackknife*. Ann. Statist., **7**, 1–26.
- [4] Gnedenko, B. V.(1943) *Sur la distribution limite d'une série aléatoire*. Annals of Mathematics, **44**, 423-453.
- [5] Gomes, M.I. (1990) *Statistical inference in an extremal markovian model*. Compstat90, 257–262.
- [6] Gomes, M.I., Hall, A. and Miranda, C. (2008) *Subsampling techniques and the Jackknife methodology in the estimation of the extremal index*. J. Comput. Stat. and Data Analysis, **52**:4, 2022–2041.
- [7] Gomes, M.I., Figueiredo, F. and Neves, M.M. (2012) *Adaptive estimation of heavy right tails: resampling-based methods in action*. Extremes, **15**, 463–489.
- [8] Gray, H.L. and Schucany, W.R. (1972) *The Generalized Jackknife Statistic*. Marcel Dekker. New York.
- [9] Hall, P. (1990) *Using the bootstrap to estimate mean-square error and selecting smoothing parameter in non-parametric problems*. J. Mult. Anal., **32**, 177-203.
- [10] Hsing, T. (1993) *Extremal index estimation for a weakly dependent stationary sequence*. Ann. Stat., **21**, 2043–2071.
- [11] Lahiri, S., Furukawa, K. and Lee, Y-D. (2007) *Nonparametric plug-in method for selecting the optimal block lengths*. Statistical methodology, **4**, 292–321.
- [12] Leadbetter, M.R. and Nandagopalan, L. (1989) *On exceedance point process for stationary sequences under mild oscillation restrictions*. In Extreme Value Theory: Proceedings, Oberwolfach 1987, J. Hüsler and R.D. Reiss (eds.), Lecture Notes in Statistics **52**, 69-80. Springer-Verlag, Berlin.
- [13] Leadbetter, M.R., Lindgren, G. and Rootzén, H. (1983) *Extremes and related properties of random sequences and series*. Springer-Verlag, New York.
- [14] Nandagopalan, S. (1990) *Multivariate Extremes and Estimation of the Extremal Index*. PhD Thesis, University of North Carolina, Chapel Hill.
- [15] Singh, K. (1981) *On the asymptotic accuracy of the Efron's bootstrap*. Ann. Stat., **9**, 345–362.
- [16] Weissman, I. and Novak, S. (1998) *On blocks and runs estimators of the extremal index*. J. Stat. Plann. Inf., **66**, 281–288.

A bivariate cost-sensitive classifier performance index

Luca Frigau, *University of Cagliari*, frigau@unica.it

Claudio Conversano, *University of Cagliari*, conversa@unica.it

Francesco Mola, *University of Cagliari*, mola@unica.it

Abstract. A new approach is presented aimed at evaluating the performance of a classifier in terms of predictive accuracy and difference in distribution between predicted classes and observed ones. The output of the proposed three steps procedure allows us to consider classifier performance under two different perspectives: accuracy, measured through a cost sensitive (model-based) index; and similarity in distribution, measured through the Gini index which compares the cumulative distribution function of observed cases and predicted ones. Both index are defined in $[0, 1]$ so that their values can be graphically represented in a $[0, 1]^2$ space in order to allow the user to draw global information about the classifier performance. Results obtained on simulated data provide evidence about the effectiveness of the proposed approach.

Keywords. Cost-sensitive Classification, Beta Regression, Similarity in distribution, Predictive accuracy, Visualization.

1 Introduction

In a classification problem it is common practice testing a wide variety of learning algorithms by varying threshold values and by using different tuning parameters. In that way different classifiers are obtained which can be compared in order to evaluate their predictive ability. The comparison could concern different performance metrics: accuracy, sensitivity, specificity, speed, cost, readability, etc. Notationally, given a classification problem on ℓ classes observed on n cases, let \mathbf{Q} be a confusion matrix stemmed from a classifier X . In this framework rows of \mathbf{Q} refer to the true classes, and columns of \mathbf{Q} to the predicted ones. By checking rows, the elements c_{ij} indicate how many cases with true class x_i have been classified in each class. By checking columns, the elements c_{ij} indicate how many cases of each class have been classified in class \hat{x}_j . Starting from the confusion matrix \mathbf{Q} several measures and approaches have been proposed to evaluate classifier performance. Among these, the most known is accuracy. This measure is very plain, overlooking a lot of information about the costs of different elements of misclassification [1]. In order to use these information, a new measure based on the concept of entropy is proposed

in [2] as an index that compares different classifier performances using the misclassification cells. Many other measures based on confusion matrix were proposed, such as the global performance index [3], the entropy of a confusion matrix [4], the transmitted information of the classifier [5], or the relative classifier information [6].

The goal of this paper is to propose a new approach that enables us to compare performances of several classifiers in the framework of multi-class learning (i.e., when a new observation has to be classified into one, and only one, of ℓ non-overlapping classes). The output of the proposed approach is a bivariate classifier performance index obtained from two measures which refer to a cost-sensitive weighted classification accuracy index and to another index expressing the similarity in distribution between the n observations which are classified in one of the ℓ classes by a classifier with the original distribution of the n cases among the ℓ classes. Both indexes are defined in $[0, 1]$, so that a comparison of different classifier performance can be represented in a $[0, 1]^2$ space.

The rest of the paper is organized as follows. Section 2 presents the main features of the proposed approach and describes the three steps characterizing it. Section 3 presents the results of the performance of the proposed approach on artificial data and Section 4 ends the paper with some concluding remarks.

2 The bivariate classifier performance index

The bivariate classifier performance index derives from a three steps procedure to be carried out for each candidate classifier. The 3 steps can briefly identified with: 1) the model-based measurement of classification accuracy; 2) the measurement of the similarity in distribution between observed classes and predicted ones; 3) the visualization of the results of the previous steps in order to assess global classifier performance.

2.1 Model-based measurement of classification accuracy

Let $\pi \in [0, 1]$ be a misclassification level, so that $1 - \pi$ is the classification accuracy level. If k different classifiers are considered, k values of π can be observed and those values, defined in $[0, 1]$, can be modeled on the basis of other information related to each classifier. The model specified for π allows us to assess classifier performance through a model-based classification accuracy index.

In a regression modeling framework characterized by a continuous response variable Y defined in $[0, 1]$, data are usually transformed in order to map the domain of Y in the real line and then a standard linear regression analysis is applied. This approach has some shortcomings (see [7]), such as heteroskedasticity and difficulties in the interpretation of estimated parameters, which are expressed in terms of the transformed variable instead of the original one. Ferrari and Cribari-Neto [8] proposed a regression model for continuous variables that assumes values in $[0, 1]$, called *Beta Regression Model*. The assumption of this model is that the response variable is beta-distributed, $Y \sim Beta(a, b)$ with $a, b > 0$. Authors proposed a particular parameterization of the beta density in order to obtain a regression structure for the mean of the response along with a precision parameter. They showed that, through setting $\mu = a/(a+b)$ and $\phi = a+b$, it is possible to express expectation and variance of Y as $E(Y) = \mu$ and $VAR(Y) = \mu(1-\mu)/(1+\phi)$, respectively. The parameter ϕ conveys a rate of precision because, for larger ϕ , $VAR(Y)$ decreases.

The beta regression model introduced in [8] is applied in the framework of the present study in order to estimate π and, indirectly, $1 - \pi$. Specifically, the goal is to estimate a specific beta regression model using a large number of simulated confusion matrices weighted by some proximity measures and misclassification costs, in order to obtain estimated regression parameters and associated p values. Weighting is very important in this framework, because it conveys essential information to the model about the different importance attributed to possible different misclassification levels. Once the model is estimated, it is applied to the confusion matrix stemmed by each classifier in order to estimate a *cost-sensitive (model-based) weighted classification index*. For a classifier k and assuming $\pi_k \sim \text{Beta}(\mu_k, \phi)$, the beta regression model is defined as

$$g(\mu_k) = \sum_{i=1}^{\ell} \sum_{j=1}^{\ell} \beta_{ij} c_{ij}^k d(x_i, x_j) = \eta_k \quad (1)$$

where $d(x_i, x_j)$ is a proximity measure detailed below, c_{ij}^k are the frequencies of the confusion matrix stemmed by the classifier k , and β_{ij} are the model coefficients that express the contribution of each confusion matrix cell to global misclassification. Finally, $g(\cdot)$ is a link function: In Eq. (1) the probit distribution is chosen for specifying the link function $g(\cdot)$, so that the expectation of π_k can be defined as $\mu_k = g^{-1}(\eta_k) = \Phi(\eta_k)$, where $\Phi(\cdot)$ is the cumulative distribution function of a standard normal distribution.

As already mentioned, for estimating the β_{ij} in (1) a large number of confusion matrices are simulated. One-half with $\pi = 0$ and non-zero elements in the diagonal only, and the other half with random assigned non-zero off-diagonal elements, so that $\pi = 1$. The simulation of a random classified confusion matrix is quite simple to obtain. All confusion matrices stemmed by the classifiers have the same marginal row frequencies. In fact, if they come from the same dataset the number of true classes is fixed for all matrices. Hence, it is sufficient to simulate matrices with uniformly distributed rows, and then fix their marginal row frequencies equal to those of the confusion matrices stemmed by the classifiers. Next step consists in excluding diagonal cells from simulated matrices, leaving just cells that convey misclassification information.

Next, the cells of the simulated confusion matrices are weighted by some proximity measures, which are defined, for all entries c_{ij} (with $i \neq j$) corresponding to off-diagonal cells, as

$$d(x_i, x_j) = \begin{cases} \frac{x_\ell - x_1}{|x_i - x_j|} w_{ij} & \text{if } x \text{ is numerical} \\ \frac{\ell - 1}{|i - j|} w_{ij} & \text{if } x \text{ is ordinal} \\ w_{ij} & \text{if } x \text{ is nominal} \end{cases} \quad (2)$$

where w_{ij} is a weight, fixed by the researcher, that specifies the importance in terms of misclassification cost attributed to the proximity level between x_i and x_j . As such, weighting is motivated by the idea of adding information deriving from expert knowledge. Once the simulated matrices are weighted, the model could be fitted through them in order to derive the estimated value $\hat{\mu}_k$ of π_k for the k -th classifier as

$$\hat{\mu} = \Phi \left(\sum_{i=1}^{\ell} \sum_{j=1}^{\ell} \hat{\beta}_{ij} c_{ij}^k d(x_i, x_j) \right) \quad (3)$$

2.1 Similarity in distribution index

One of the main problem in the framework of classifier performance measurement is the choice of the best classifier once that two (or more) classifiers present the same value of the classification accuracy $1 - \pi$ but the latter derives from different confusion matrices. To define a classifier performance measure that also considers information about the difference in distribution among classifier confusion matrices, a *normalized similarity in distribution index* is considered. It derives from a dissimilarity index introduced by Gini and used, among others, in [9]. In general, for a M -class classification problem the Gini index of dissimilarity in distribution D is defined as

$$D = \sqrt{\frac{1}{M-1} \sum_{m=1}^{M-1} |F_m^x - F_m^y|^2} \tag{4}$$

where F_m^x and F_m^y are the cumulative frequencies in m of the vectors x and y , whereas $\sqrt{M-1}$ is equal to the maximum value of this index, and it is used to normalize it. D is defined in $[0, 1]$ and is susceptible to change in values as long as one or more observations are moved from class i to class j ($i \neq j$ and $i, j \in M$).

To introduce a similarity in distribution index, let us consider two confusion matrices, \mathbf{Q}_W and \mathbf{Q}_Z , corresponding to classifiers W and Z respectively. They refer to a situation in which the value of classification accuracy is the same for both classifiers, even if the two confusion matrices are clearly different. Measuring similarity between \mathbf{Q}_W and \mathbf{Q}_Z requires the comparison of each element of the two matrices with those of a common reference matrix \mathbf{U} . The latter is the matrix which refers to the situation of maximum accuracy so that all its non-zero elements are located in the diagonal, i.e., in the c_{ij} cells ($i = j$), and all predicted values correspond to observed ones. To make such a comparison, the first step consists in disassembling \mathbf{U} , \mathbf{Q}_W and \mathbf{Q}_Z , into a number of elements equal to the row number, as illustrated in Table 1. Next, these elements have to be placed side by side in an ordered way, as shown in Table 2, and reassembled to compose vectors u , q_W and q_Z concerning \mathbf{U} , \mathbf{Q}_W and \mathbf{Q}_Z , respectively. To compute the

	\hat{x}_1	\hat{x}_2	\cdots	\hat{x}_ℓ	
x_1	c_{11}	c_{12}	\cdots	$c_{1\ell}$	element 1
x_2	c_{21}	c_{22}	\cdots	$c_{2\ell}$	element 2
\vdots	\vdots	\vdots	\ddots	\vdots	\vdots
x_ℓ	$c_{\ell 1}$	$c_{\ell 2}$	\cdots	$c_{\ell \ell}$	element ℓ

Table 1: Confusion matrix disassembled.

element 1	element 2	\cdots	element n
$c_{1\ell}$ c_{12} \cdots $c_{1\ell}$	c_{21} $c_{2\ell}$ \cdots $c_{2\ell}$	\cdots	$c_{\ell 1}$ $c_{\ell 2}$ \cdots $c_{\ell \ell}$

Table 2: Elements of a confusion matrix reassembled into a vector.

similarity in distribution for \mathbf{Q}_W and \mathbf{Q}_Z , it is necessary to compare the distribution of q_W and

q_Z with the distribution of u . Considering the difference $1 - D$, where D has been defined in Eq. (4), for \mathbf{Q}_W and \mathbf{Q}_Z we define a similarity in distribution index whose values are in $[0, 1]$ as

$$S_{\mathbf{Q}_W} = 1 - \sqrt{\frac{1}{\ell^2 - 1} \sum_{m=1}^{\ell^2 - 1} |F_m^{q_W} - F_m^u|^2} \quad \text{and} \quad S_{\mathbf{Q}_Z} = 1 - \sqrt{\frac{1}{\ell^2 - 1} \sum_{m=1}^{m^2 - 1} |F_m^{q_Z} - F_m^u|^2} \quad (5)$$

2.3 Visualization

Once both values of the *cost-sensitive (model-based) weighted classification index* introduced in Section 2.1 and the *normalized similarity in distribution index* introduced in Section 2.2 are available for each classifier, their values can be projected in a $[0, 1]^2$ space in order to evaluate their performance from the perspective of both classification accuracy and similarity in distribution. The possibility of analyzing classifier performance in a two-dimensional space is very useful since it facilitates the comparison among different classifiers and allows the user to understand which of the two considered items (weighted classification and similarity in distribution) mostly influences classifier performance. Of course, the two-dimensional representation is particularly helpful when the number of considered classifiers is very large.

3 Example

Hereinafter, results obtained for the two-dimensional classification performance index on simulated data are presented. The simulation setting considers 6 confusion matrices (\mathbf{Q}_A to \mathbf{Q}_F) deriving from classifiers A to F (see Table 3), with respect to a classification problem involving 4 classes, x_1 to x_4 , of an ordinal response variable. It is possible to note that: a) the classifier

	\hat{x}_1	\hat{x}_2	\hat{x}_3	\hat{x}_4		\hat{x}_1	\hat{x}_2	\hat{x}_3	\hat{x}_4		\hat{x}_1	\hat{x}_2	\hat{x}_3	\hat{x}_4
x_1	20	4	2	22	x_1	28	3	12	5	x_1	12	10	12	14
x_2	4	10	1	0	x_2	3	9	2	1	x_2	5	5	1	4
x_3	0	3	5	0	x_3	2	0	4	2	x_3	1	2	3	2
x_4	11	7	8	3	x_4	4	4	6	15	x_4	4	10	8	7
(a) \mathbf{Q}_A					(b) \mathbf{Q}_B					(c) \mathbf{Q}_C				
	\hat{x}_1	\hat{x}_2	\hat{x}_3	\hat{x}_4		\hat{x}_1	\hat{x}_2	\hat{x}_3	\hat{x}_4		\hat{x}_1	\hat{x}_2	\hat{x}_3	\hat{x}_4
x_1	48	0	0	0	x_1	4	23	16	5	x_1	4	4	16	24
x_2	0	15	0	0	x_2	3	6	4	2	x_2	1	6	1	7
x_3	0	0	8	0	x_3	0	2	2	4	x_3	6	0	2	0
x_4	0	0	0	29	x_4	1	10	15	3	x_4	15	10	1	3
(d) \mathbf{Q}_D					(e) \mathbf{Q}_E					(f) \mathbf{Q}_F				

Table 3: Confusion matrices from 6 different classifiers.

D provides a perfect classification ($\pi_D = 0$); b) the classifiers E and F are characterized by the same accuracy, i.e., they present the same elements on the diagonal of the confusion matrix, but they differ with respect to off-diagonal entries; c) the confusion matrix obtained from C has quite

uniformly distributed rows, as it usually happens in random classification, and d) the confusion matrices obtained from A and B refer to a situation which can be considered as intermediate between that concerning C and D.

For each classifier, the proximity measure $d(x_i, x_j)$ introduced in Section 2.1 has been defined according to Eq. (2). In this example, we fix $w_{ij} = 1$ in order to refer to a situation in which the weight depends proportionally from the distance between observed class and predicted ones. As a result, more weights is attributed to cases which have been classified in class which is far from the original one (in our example x_1 classified as \hat{x}_4 or viceversa).

Next step is to simulate a large number of weighted confusion matrices. In this example 1,000 matrices are simulated as follows: 500 matrices refer to classifiers with complete accuracy (i.e. the best possible classification), so that they present all non-zero elements on the diagonal and $\pi = 0$; 500 matrices refer to cases deriving from random classified elements (i.e. the worst possible classification), so that they present uniformly distributed row elements and $\pi = 1$.

To apply the beta regression model specified in Eq. (1) information deriving from these confusion matrices has to be conveniently rearranged. Rearranging each simulated matrix in a row by row manner leads us to the $1,000 \times (4 \times 4)$ matrix represented in Table 4. It is possible to note that this matrix has the same row marginal frequencies. This is a common characteristic of all confusion matrices which can be derived from a classifier trained on the same dataset. The computation of a cost-sensitive (model-based) classification accuracy index as defined in Section 2.1 requires the elimination of the diagonal cells from the simulated confusion matrices since only cells that convey misclassification information are included in the beta regression model. Following this elimination, the simulated matrices with off-diagonal elements only are weighted by the proximity measures $d(x_i, x_j)$ in order to obtain the new (weighted) data matrix represented in Table 5. It allows us to put into the beta regression model information about the importance attributed to the possible different misclassifications.

#	p	c_{11}	c_{12}	c_{13}	c_{14}	c_{21}	c_{22}	c_{23}	c_{24}	c_{31}	c_{32}	c_{33}	c_{34}	c_{41}	c_{42}	c_{43}	c_{44}
1	0	48	0	0	0	0	15	0	0	0	0	8	0	0	0	0	29
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
500	0	48	0	0	0	0	15	0	0	0	0	8	0	0	0	0	29
501	1	11	10	13	14	2	9	3	1	1	2	4	1	3	11	3	12
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
1000	1	16	9	13	10	1	1	7	6	4	1	3	0	7	10	9	3

Table 4: Simulated confusion matrices with the diagonal cells highlighted.

The beta regression model is estimated using the above described weighted simulated matrices. Following the estimation of model parameters, the cost-sensitive (model-based) classification accuracy index is computed for classifiers A to F after the elimination of the diagonal cells from their confusion matrices. The index value is obtained by predicting the response value ($\hat{\pi}_k$, with $k = A, \dots, F$) on the basis of the classifier confusion matrix entries and the estimated β_{ij} .

As for the computation of the similarity in distribution index S , the six confusion matrices have been disassembled in order to form new variables as described described in Section 2.2. Then, the index S has been computed through the Eq. (5).

Results obtained for the two indexes are shown in Figures 1 and 2. The first plot refers to

#	p	c_{12}	c_{13}	c_{14}	c_{21}	c_{23}	c_{24}	c_{31}	c_{32}	c_{34}	c_{41}	c_{42}	c_{43}
1	0	0	0	0	0	0	0	0	0	0	0	0	0
\vdots													
500	0	0	0	0	0	0	0	0	0	0	0	0	0
501	1	30	26	14	6	9	2	2	6	3	3	22	9
\vdots													
1000	1	27	26	10	3	21	12	8	3	0	7	20	27

Table 5: Simulated confusion matrices after weighting.

the representation of the two considered indexes in a $[0, 1]^2$ space. The second plot refers to the distribution of the predicted classes and the observed ones, and it shows information about the decomposition of the similarity in distribution index S introduced in Eq. (5). Figure 1 shows that the best classifier is D. This result is not surprising since the considered classifier is the one providing a perfect classification ($\pi_D = 0$). As a consequence, for this classifier the distribution of the predicted classes corresponds to that of the observed ones so that the line in Figure 2 obtained for D overlaps the dotted one, which refers to the original distribution of cases among the 16 cells of the confusion matrix. Interesting considerations can be made about the other classifiers. The second best classifier is B, which presents highest values for the two considered indexes after D. This result depends on the fact that misclassified observations are not far from their true classes. Thus, the lower penalization of classifiers presenting extra-diagonal entries which are close to the diagonal ones is a peculiarity of the proposed bivariate index which correctly uses the higher cost of misclassified observations whose predicted class is far from the observed one. This consideration is enforced by the comparison of the performance of classifiers E and F which, as previously mentioned, have the same elements on the diagonal of the confusion matrix but they differ in the distribution of the extra-diagonal ones. In particular, comparing E and F it is possible to note that in the first case extra-diagonal frequencies are more close to the diagonal ones. The weighting system introduced in Eq. (5) causes the performance of F to be very poor in comparison to that of E.

4 Conclusion

Cost-sensitive classification is one of mainstream research topics in data mining and machine learning that induces models from data with an unbalanced class distribution and impacts by quantifying and tackling the unbalance. In this paper a bivariate index based on a model based accuracy measure and a similarity in distribution measure has been introduced. Results obtained on simulated data provide evidence on the effectiveness of our proposal, since the bivariate index appears as sensitive to misclassifications to which an highest cost is attributed. Future research efforts will be directed to the assessment of the reliability of the proposed bivariate index, as well as on the assessment of its performance in multiclass learning problems characterized by unbalanced distribution of the response classes and/or reduced data size.

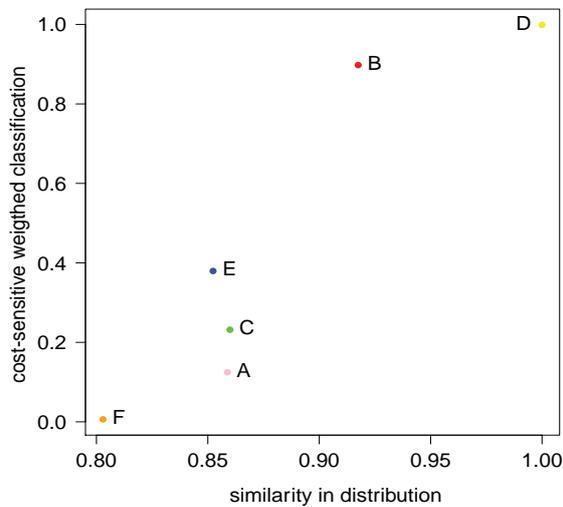


Figure 1: The bivariate cost-sensitive classification index.

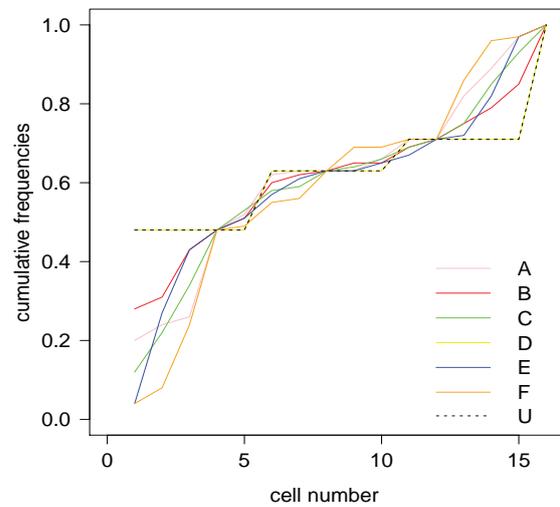


Figure 2: The cumulative distribution function for the observed confusion matrices.

Bibliography

- [1] Hand, D. J. and Till, R. J. (2001). *A simple generalisation of the area under the roc curve for multiple class classification problems*. *Machine Learning*, **45**(2), 171–186.
- [2] Wei, J.-M., Yuan, X.-J., Hu, Q.-H., and Wang, S.-Q. (2010). *A novel measure for evaluating classifiers*. *Expert Systems with Applications*, **37**(5), 3799–3809.
- [3] Freitas, C. O., De Carvalho, J. M., Oliveira Jr, J., Aires, S. B., and Sabourin, R. (2007). *Confusion matrix disagreement for multiple classifiers*. In *Progress in Pattern Recognition, Image Analysis and Applications*, pages 387–396. Springer.
- [4] Van Son, R. (1995). *A method to quantify the error distribution in confusion matrices*. In *Proceedings Eurospeech 95, Madrid*, pp. 2277–2280.
- [5] Abramson, N. (1963). *Information theory and coding*. **61**, McGraw-Hill New York.
- [6] Sindhvani, V., Bhattacharya, P., and Rakshit, S. (2001). *Information theoretic feature crediting in multiclass support vector machines*. In *Proceedings of the First SIAM International Conference on Data Mining*, pages 5–7. SIAM.
- [7] Cribari-Neto, F. and Zeileis, A. (2010). *Beta regression in R*. *Journal of Statistical Software*, **34**(2), 1–24.
- [8] Ferrari, S. and Cribari-Neto, F. (2004). *Beta regression for modelling rates and proportions*. *Journal of Applied Statistics*, **31**(7), 799–815.
- [9] Rachev, Svetlozar T. (1985). *The Monge-Kantorovich mass transference problem and its stochastic applications*. *Theory of Probability & Its Applications*, **29**(4), 647–676.

Effects of Sampling Methods on Prediction Quality. The Case of Classifying Land Cover Using Decision Trees.

Ronald Hochreiter, *WU Vienna University of Economics and Business*,
ronald.hochreiter@wu.ac.at

Christoph Waldhauser, *WU Vienna University of Economics and Business*,
christoph.waldhauser@wu.ac.at

Abstract. Clever sampling methods can be used to improve the handling of big data and increase its usefulness. The subject of this study is remote sensing, specifically airborne laser scanning point clouds representing different classes of ground cover. The aim is to derive a supervised learning model for the classification using CARTs. In order to measure the effect of different sampling methods on the classification accuracy, various experiments with varying types of sampling methods, sample sizes, and accuracy metrics have been designed. Numerical results for a subset of a large surveying project covering the lower Rhine area in Germany are shown. General conclusions regarding sampling design are drawn and presented.

Keywords. Machine Learning, Sampling, Decision Tree, Airborne Laser Scanning

1 Introduction

In this paper we seek to understand how clever sampling can be used to improve the handling of big data and increase its usefulness. Our case comes from remote sensing and contains airborne laser scanning point clouds representing different classes of ground cover. The aim is to derive a supervised learning model for the classification using CARTs. This paper is organized as follows. We will first briefly review the state of the art for the fields of airborne laser scanning, classification trees and stratified sampling. We will then describe our data set and the experimental setup. In Section 4 we present our results. After discussing them we conclude with suggestions for further research.

2 Classification of Airborne Laser Scan Point Clouds

The character of land surveying has changed dramatically in recent decades. The availability of powerful lasers scanning, surveying the ground from airborne platforms provides very accurate and timely measurements of ground cover. These airborne laser scans (ALS) produce geo-coded point clouds of laser return echos with resolutions in a decimeter scale. However, the resulting massive amounts of data – surveying planes can easily cover vast stretches of land – require specialized handling. We are now going to discuss the specifics of ALS. Then we review four sampling procedures in the context of ALS. Finally we explain the usage of the popular CART algorithm for automatically deriving classifications of ground cover.

Airborne Laser Scanning

Before the advent of ALS, land surveying was very cumbersome and involved traveling target areas, setting up measurement points. The situation improved somewhat due to the increased usage of aerial photography. These photos, however, are often ambiguous and it is difficult to tell three-dimensional structures from them, even when using stereo photography.

Airborne laser scanning, on the other hand offers a number of advantages. An airplane or helicopter flies over a designated area usually stripwise. A downward looking laser array emits beams and records the echos. Modern arrays use rotating mirrors to deflect the beam and scan an area perpendicular to the flight path. Also, these arrays are capable of recording the entire spectral characteristics of multiple echos.

From these raw return signal characteristics and the GPS coordinates of the plane, every signal is translated into points on (or above) the ground. Further point features can be computed from the signal using various neighborhood averaging techniques. For an overview and a systematic organization of available point features, see [9]. From this follows that every point is described by a high-dimensional feature vector.

These point clouds can be used instantly to compute digital terrain and surface models. In order to use them for other purposes, e.g. the identification of ground cover, classification models are needed. We are going to describe them in the next subsection.

Classifying ALS point clouds using CARTs

The deriving of types of ground cover from either aerial photos or ALS point clouds, is essentially a quite simple but time-consuming task. It involves investigating the aerial photos and use ones experience to classify a given area. For instance, looking at a certain shape in the point cloud and cross-referencing it with data from the photo, the human classifier can classify the related points to be coniferous forest or a road. Processing hundreds of acres in this manner is time consuming, expensive and error prone. It is therefore the aim of ongoing research to develop automatic classifiers. There has been some success in that matter. See e.g. [15] for an overview of current research. A summary of the state of ALS classification needs to point to still rather high misclassification rates and difficulties in classifying certain types of ground cover.

Automated classification tasks, in general, can be supervised or unsupervised. The latter works entirely autonomously and draws all information from the available data. The former requires manually created training data to learn its model from. An example of a supervised learning algorithm are classification and regression trees (CARTs) [2]. And it is CARTs that have been used with great success in the past [5, 10, 15]. There, a large number of manually

classified points are used to train and evaluate the classification of ground cover. For evaluation purposes, the data set is artificially split into training and test data. In practical applications, however, the training area needs to be determined using statistical sampling techniques. In the next subsection we are going to review four different sampling procedures that we find useful in this context.

Sampling in the ALS context

The most natural way of obtaining a slice of a data set is to simply take the first, say 10,000 cases. This, however, is not a random sample and it is rather unlikely that it will be representative enough to train a classifier from it. A better approach and perhaps the second most natural thing to do is a simple random sample. Since the extent of the entire data set is known when sampling, every point has the same chance of being included in the sample. The key advantage of this approach is obtaining a sample that is representative of the data set.

However, there is also a big disadvantage: rare classes that are only found sparingly throughout the data set might not be sampled at all. Then of course the classifier has no chance of learning the characteristics of these classes and will not be able to assign any data into these classes during classification. For instance, consider that a moderately small data set contains 2 million points and a single tree might be made up of as little as 500 points. For being able to classify this particular kind of tree, enough of its points need to end up in the sample. For simple random samples this means that large samples are required to ensure sampling of rare classes with reasonable probabilities.

Another approach originally invented for surveys among humans is stratified sampling [1, 4, 7]. There, a small simple random sample is taken from each class, no matter how rare it is. Obviously, the resulting sample is not representative of the population anymore. But presence of points from rare classes is guaranteed. In the ALS context, stratified sampling has been used numerous times e.g. [6, 12].

When using samples that are not representative of the population they were drawn from, care must be taken. The canonical method is to compute sampling weights that correct for the altered composition of the sample. As CARTs actively use the distribution of classes when estimating class probabilities, the sampling weights are used to compute correct priors. Alternatively, CARTs' computation of priors can be circumvented and the true priors be specified.

In order to examine the effect of these sampling plans, we used a data set from a real surveying project in a series of experiments. Both are being described in the next section.

3 Data & Method

In this section we are going to introduce the data set we used in our analysis as well as the experimental setup. Our data set is a small subset of a large surveying project covering the lower Rhine area in Germany. Our subset was selected for its representativeness of the region.²³ It contains 2,872,488 points. The data set was manually annotated using photo interpretation and ALS point clouds. Table 1 contains the distribution of points in the classes that were observed.

²³All computations were done in R [11] using *lasr* [14] for point cloud processing, *rpart* [13] for CARTs and [16] for visualization.

Class of ground cover	Frequency
undefined	959
ground	2401914
gravel	1903
asphalt	20301
decideous forest	175103
building roofs	1383
walls/buildings	13
water	61362
cars and other moving objects	3912
temporary objects	1411
bridges	173774
power poles	231
bridge cables	16240
road protection fence	11169
bridges construction	1819
cement/concrete	936
error class	58

Table 1: Distribution of ground cover classes in the data set

As can be seen from Table 1, the distribution is highly skewed with roughly 85% being ground. There are also some very rare classes that will be hard to sample and learn, e.g. walls.

In order to measure the effect of sampling methods on the classification accuracy we designed experiments with varying types of sampling methods, sample sizes, and accuracy metrics. The two types of sampling methods were simple random samples (without replacement) and stratified samples. While the first one is straight forward, the second type of sampling method deserves explanation. The sampling algorithm tried to sample s points from each class. If a class had less than twice as many points, it would only sample half of those points.

This sampling method then also dictated sample sizes. In total 8 different sample sizes were used. Table 2 gives an overview of the sample sizes used in the experiments.

Sample	Points
S1	84288
S2	72524
S3	57783
S4	39251
S5	18472
S6	4268
S7	906
S8	471

Table 2: Sample sizes used in experiments. The entire data set contains 2.9 million points.

In order to gain an estimate of the variability of our measurements, we drew fifty samples of each size using each type of sampling method.

These samples were then used to train a classifier using the CART framework introduced above. The so obtained classifier was then used to classify the remainder of the data set. When learning off stratified samples, the CART algorithm was given (a) no further information on class priors, (b) post-stratification case weights²⁴, and (c) the true class distributions. We therefore have four (simple random samples and stratified samples in the three fashions outlined above) different sampling methods to compare.

To gauge the quality of a classification we used three different metrics all based on cross classification error matrices M . The simplest metric is the total misclassification rate (MCR_{Total}) that is defined as

$$MCR_{Total} = 1 - \frac{\sum \text{diag}(M)}{N}$$

with N being the total number of points. A distinct property of the total MCR is that the misclassification of rare classes has only a small influence. Whether this is considered a bug or a feature depends on the context. In bids to accurately classify even rare classes, the total MCR can be misleading.

A approach that takes rare classes into account is the class-based MCR (MCR_{Class}). There, the misclassification is not averaged over the entire data set but per class:

$$MCR_{Class} = \sum 1 - P(a)$$

with $P(a)$ being the proportion of correctly classified points in that class. Another quality metric is Kohen's Kappa (κ) [12]. It is given as

$$\kappa = \frac{P(a) - P(e)}{1 - P(e)}$$

with $P(e)$ is the product of marginal proportions for that class. We computed each of these metrics for every sample size and sampling method combination. The results of these computations are given in the next section.

4 Results

In order to assess the effects of sampling methods on the classification accuracy, we conducted a series of experiments. In the following we will present the results from our computations. Figure 1 displays the relationships between sample size and classification accuracy for the four sampling methods. We depicted only the three smallest sample sizes as they—obviously—exhibit the largest differences, but larger sample sizes confirm the overall trend.

For the overall misclassification rate, simple random samples outperform all other classification approaches clearly. It is also notable, that sample size has no significant degrading effect on simple random samples using this metric.

Class-based misclassification rates reverse this picture. Using that metric, simple random samples perform very poorly and stratified samples provide better, consistent results. Among the stratified samples, interestingly, the method of not providing the CART algorithm with any further information on the priors or sample composition outperforms the other methods.

²⁴Post-stratification was computed using *survey* [8].

Finally, turning to κ , simple random samples once again deliver the best performance by achieving the greatest agreement. Among the stratification methods, post-stratification and prior specification both perform slightly better than their uninformed brother.

In this section, we presented the results from our experiments. They were 50 times bootstrapped each. It is remarkable to see that, while simple random sampling beats any other method by far in two of three metrics, this is not true when using the class-based misclassification metric. There, working with stratified samples not correcting for the stratification delivers best performance. In the next section we are going to discuss these findings in greater detail.

5 Discussion & Conclusion

We have compared the prediction quality using CARTs on land surveying laser scan data, given different sample sizes and sampling methods. Prediction quality was measured using three different metrics putting varying degrees of emphasis on overall correctness or correctness per class. Bootstrapping the results fifty times, we find that simple random samples provide the best overall classification quality. When using stratified samples in that context, it is important to specify either post-stratification weights or true class probabilities to achieve better results.

However, when defining classification quality as average per class, we find that stratified samples achieve much lower misclassification rates. Interestingly, this is even the case when the CART algorithm is not provided with true or assumed class probabilities. More precisely, specifying class probabilities or post-stratification weights to undo the stratification process lead to worse results. This is unexpected, as correcting the altered sample composition introduced by stratification is thought to be canonical [3] and has been shown to improve classification even with land surveying data above.

As CART-based analysis is becoming ever more widespread, also in survey contexts, our findings warrant caution when integrating the sampling design with the CART analysis.

Bibliography

- [1] A. A. Anganuzzi and S. T. Buckland. Post-stratification as a bias reduction technique. *The Journal of Wildlife Management*, 57(4):827–834, 1993.
- [2] L. Breiman, J. Friedman, C. J. Stone, and R. A. Olshen. *Classification and Regression Trees*. CRC press, 1984.
- [3] D. C. Doss, H. O. Hartley, and G. R. Somayajulu. An exact small sample theory for post-stratification. *Journal of Statistical Planning and Inference*, 3(3):235–247, 1979.
- [4] D. Holt and T. M. F. Smith. Post stratification. *Journal of the Royal Statistical Society. Series A (General)*, 142(1):33–46, 1979.
- [5] T. Hothorn, K. Hornik, and A. Zeileis. Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical Statistics*, 15(3):651–674, 2006.
- [6] S. M. Joy, R. M. Reich, and R. T. Reynolds. A non-parametric, supervised classification of vegetation types on the Kaibab National Forest using decision trees. *International Journal of Remote Sensing*, 24(9):1835–1852, 2003.

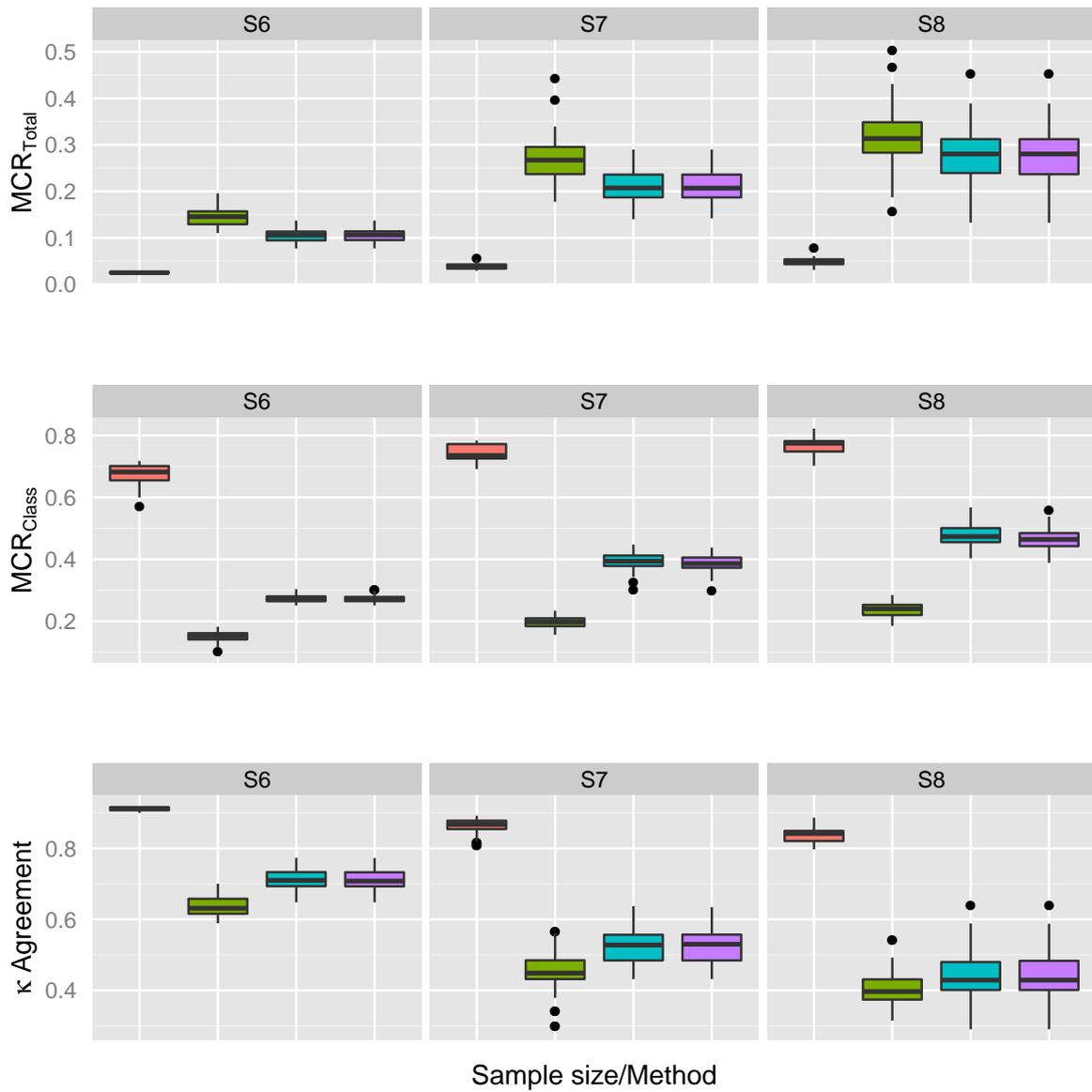


Figure 1: Quality of the obtained prediction for different sample sizes and sampling methods measured in three metrics. Results are 50 times bootstrapped each. Sampling methods per panel, from left to right: simple random sample, stratified sample, stratified with post-stratification weights, stratified with priors specified.

- [7] R. J. A. Little. Post-stratification: A modeler's perspective. *Journal of the American Statistical Association*, 88(423):1001–1012, 1993.
- [8] T. Lumley. Analysis of complex survey samples. *Journal of Statistical Software*, 9(1):1–19, 2004.
- [9] J. Otepka, S. Ghuffar, C. Waldhauser, R. Hochreiter, and N. Pfeifer. Georeferenced point clouds: A survey of features and point cloud management. *ISPRS International Journal of Geo-Information*, 2(4):1038–1065, 2013.
- [10] M. Pal and P. M. Mather. An assessment of the effectiveness of decision tree methods for land cover classification. *Remote Sensing of Environment*, 86(4):554–565, 2003.
- [11] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2014.
- [12] S. Stehman. Estimating the kappa coefficient and its variance under stratified random sampling. *Photogrammetric Engineering and Remote Sensing*, 62(4):401–407, 1996.
- [13] T. Therneau, B. Atkinson, and B. Ripley. *rpart: Recursive Partitioning and Regression Trees*, 2014. R package version 4.1-8.
- [14] C. Waldhauser and R. Hochreiter. *lasr: Tools for Working with Airborne LiDAR Data*, 2014. R package version 0.4-5.
- [15] C. Waldhauser, R. Hochreiter, J. Otepka, N. Pfeifer, S. Ghuffar, K. Korzeniowska, and G. Wagner. Automated classification of airborne laser scanning point clouds. In S. Koziel, L. Leifsson, and X.-S. Yang, editors, *Solving Computationally Extensive Engineering Problems: Methods and Applications*. Springer, 2014.
- [16] H. Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer New York, 2009.

β models for random hypergraphs with a given degree sequence

Despina Stasi, *Illinois Institute of Technology*, despina.stasi@gmail.com
Kayvan Sadeghi, *Carnegie Mellon University*, kayvans@andrew.cmu.edu
Alessandro Rinaldo, *Carnegie Mellon University*, arinaldo@stat.cmu.edu
Sonja Petrovic, *Illinois Institute of Technology*, sonja.petrovic@iit.edu
Stephen Fienberg, *Carnegie Mellon University*, fienberg@stat.cmu.edu

Abstract. We introduce the beta model for random hypergraphs in order to represent the occurrence of multi-way interactions among agents in a social network. This model builds upon and generalizes the well-studied beta model for random graphs, which instead only considers pairwise interactions. We provide two algorithms for fitting the model parameters, IPS (iterative proportional scaling) and fixed point algorithm, prove that both algorithms converge if maximum likelihood estimator (MLE) exists, and provide algorithmic and geometric ways of dealing the issue of MLE existence.

Keywords. Beta model, social networks, exponential random hypergraph model, ERGM, fixed point algorithm, IPS algorithm, likelihood function analysis.

1 Introduction

Social network models [8] are statistical models for the joint occurrence of random edges in a graph, as a means to model social interactions among agents in a population of interest. These models typically focus on representing only *binary* relations between individuals. As a result, when one is interested in higher-order (k -ary) interactions, statistical models based on graphs may be ineffective or inadequate. Examples of k -ary relations are plentiful, and include forum or committee membership, co-authorship on scientific papers, or proximity of groups of people in photographs. These datasets have been studied by replacing each k -dimensional group with a number of binary relations (in particular, $\binom{k}{2}$ of them, which form a clique), thus extracting binary information from the data, and then modeling and studying the resulting graph. Such a process inevitably causes information loss. For instance, let us consider statisticians Adam (A), Barbara (B), Cassandra (C), and David (D), see Figure 1. Suppose the authors wrote three papers in following groups: (A, B, C) , (A, D) , (C, D) . Representing this information as a graph with edges between any two individuals who have co-authored a paper provides a graph

G with edges $\{(A, B), (B, C), (C, D), (A, C)\}$. A hypergraph H representing this information would instead use the exact groups as hyperedges and, unlike G , would be able to represent additional properties of such interactions, including how many papers were coauthored by these four individuals; see Figure 1. If, in addition, A is more likely to write a 3-author paper than a 2-author paper, this requires modeling separately the probabilities of these collaborations. Despite the growing needs of practical values, models for random hypergraphs are relatively few and simple. Random hypergraphs have been studied ([7]) as generalizations of the simple Erdős-Rényi model [4] for networks; [5] considers an application of random tripartite hypergraphs to Flickr photo-tag data.

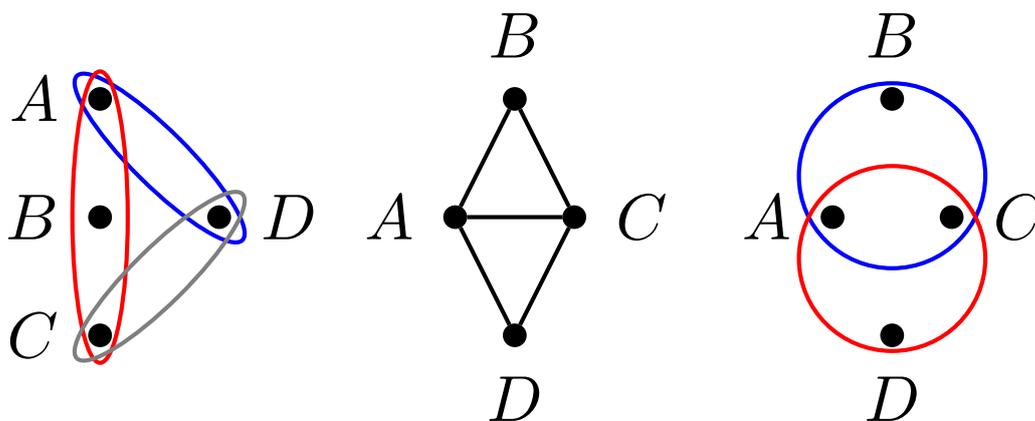


Figure 1: Distinct hypergraphs H and H' reduced to same graph G (left, right, middle).

In this paper we introduce a simple and natural class of statistical models for random hypergraphs, which we term hypergraph beta models, that allows one to model directly simultaneous higher-order (and not only binary) interactions among individuals in a network. As its name suggests, our model arises as a natural extension of the well-studied beta model for random graphs, the exponential family for undirected networks which assumes independent edges and whose minimal sufficient statistics vector is the degree sequence of the graph. It is a special class of the more general of p_1 models [6] which assume independent edges and parametrize the probability of each edge by the propensity of the two endpoint nodes. This model has been studied extensively; see [2, 3, 9, 10, 11], which give, among other results, methods for model fitting. Below we formalize the class of the beta models for hypergraphs. Just like the graph beta model, these are natural exponential random graph models over hypergraphs which postulate independent edges and whose sufficient statistics are the (hypergraph) degree sequences. Our contributions are two-fold: first we formalize three classes of linear exponential families for random hypergraphs of increasing degree of complexity and derive the corresponding sufficient statistics and moment equations for obtaining the maximum likelihood estimator (MLE) of the model parameters. Secondly, we design two iterative algorithms for fitting these models that do not require evaluating the gradient or Hessian of the likelihood function and can therefore be applied to large data: a variant of the IPS algorithm and a fixed point iterative algorithm to compute the MLE of the edge probabilities and of the natural parameters, respectively. We

show that both algorithms will converge if the MLE exists. Finally, we illustrate our results and methods with some simulations.

As our analysis reveals, the study of the theoretical and asymptotic properties of hypergraph beta models is especially challenging, more so than with the ordinary beta model. The complexity of the new models, in turn, leads to the problem of optimizing a complex likelihood function. Indeed, when the MLE does not exist, optimizing the likelihood function becomes highly non-trivial and, to a large extent, unsolved for our model as well as for many other discrete linear exponential families. To this end, we describe a geometric way for dealing with the issue of existence of the MLE for these models and gain further insights into this difficult problem with simulation experiments.

2 The hypergraph beta model: three variants

A *hypergraph* H is a pair (V, F) , where $V = \{v_1, \dots, v_n\}$ is a set of *nodes* (vertices) and F is a family of non-empty subsets of V of cardinality different than 1; the elements of F are called the *hyperedges* (or simply *edges*) of H . In a *k-uniform* hypergraph, all edges are of size k . We restrict ourselves to the set \mathcal{H}_n of hypergraphs on n nodes, where nodes have a distinctive labeling. Let $E = E_n$ be the set of all realizable hyperedges for a hypergraph on n nodes. While E can in principle be the set of all possible hyperedges, below we will consider more parsimonious models in which E is restricted to be a structured subset of edges. Thus we may write a hypergraph $x = (V, F) \in \mathcal{H}_n$ as the zero/one vector $x = \{x_e, e \in E\}$, where $x_e = 1$ for $e \in F$ and $x_e = 0$ for $e \in E \setminus F$. The degree of a node in x is the number of edges it belongs to; the degree information for x is summarized in the degree sequence vector whose i th entry is the degree $d_i(x)$ of node i in x .

Hypergraph beta models are families of probability distributions over \mathcal{H}_n which postulate that the hyperedges occur independently. In details, let $p = \{p_e : e \in \mathcal{E}_n\}$ be a vector of probabilities whose e th coordinate indicates the probability of observing the hyperedge e . We will assume $p_e \in (0, 1)$. Every such vector p parametrizes a beta-hypergraph model as follows: the probability of observing the hypergraph $x = \{x_e, e \in E\}$ is

$$\mathbb{P}(x) = \prod_{e \in E} p_e^{x_e} (1 - p_e)^{1-x_e}. \quad (1)$$

The graph beta model is a simple instance of this model, with $E = \{(i, j), 1 \leq i < j \leq n\}$. The $\binom{n}{2}$ edge probabilities are parametrized as $p_{i,j} = e^{\beta_i + \beta_j} / (1 + e^{\beta_i + \beta_j})$, for $i < j$ and some real vector $\beta = (\beta_1, \dots, \beta_n)$.

Various social network modeling considerations for node interactions require a flexible class of models adaptable to those settings. Thus, we introduce three variants of the beta model for hypergraphs with independent edges in the form of linear exponential families: *beta models* for *uniform hypergraphs*, for *general hypergraphs*, and for *layered uniform hypergraphs*. For each, we provide an exponential family parametrization in minimal form and describe the corresponding minimal sufficient statistics.

Uniform hypergraphs. The probability of a size- k hyperedge $e = i_1 \dots i_k$ appearing in the hypergraph is parametrized by a vector $\beta \in \mathbf{R}^n$ as follows:

$$p_{i_1, \dots, i_k} = \frac{e^{\beta_{i_1} + \dots + \beta_{i_k}}}{1 + e^{\beta_{i_1} + \dots + \beta_{i_k}}} \quad (2)$$

with $q_{i_1, \dots, i_k} = 1 - p_{i_1, \dots, i_k} = \frac{1}{1 + e^{\beta_{i_1} + \dots + \beta_{i_k}}}$, for all $i_1 < \dots < i_n$. In terms of odds ratios,

$$\log \frac{p_{i_1, \dots, i_k}}{q_{i_1, \dots, i_k}} = \beta_{i_1} + \dots + \beta_{i_k}. \tag{3}$$

In order to write the model in exponential family form, we abuse notation and define for each hyperedge $e \in F$, $\tilde{\beta}_e = \sum_{i \in e} \beta_i$. In addition, let $\binom{[n]}{k}$ be the set of all subsets of size k of the set $\{1, \dots, n\}$. By using (1), we obtain

$$\mathbb{P}_\beta(x) = \frac{\exp \left\{ \sum_{e \in \binom{[n]}{k}} \tilde{\beta}_e x_e \right\}}{\prod_{e \in \binom{[n]}{k}} 1 + e^{\tilde{\beta}_e}} = \exp \left\{ \sum_{i \in V} d_i(x) \beta_i - \psi(\beta) \right\},$$

where d_i is the degree of the node i in x . Then it is clear that the *sufficient statistics for the k -uniform beta model* are the entries of the degree sequence vector of the hypergraph, $(d_1(x), \dots, d_n(x))$, and the normalizing constant is

$$\psi(\beta) = \sum_{e \in \binom{[n]}{k}} \log(1 + e^{\tilde{\beta}_e}). \tag{4}$$

Layered uniform hypergraphs. Allowing for various size edges has the advantage of controlling the propensity of each individual to belong to a size- k group independently for distinct k 's. Let r be the (natural bound for the) maximum size of a hyperedge that appears in x . This model is then parametrized by $r - 1$ vectors in \mathbf{R}^n as follows:

$$p_{i_1, i_2, \dots, i_k} = \frac{e^{\beta_{i_1}^{(k)} + \beta_{i_2}^{(k)} + \dots + \beta_{i_k}^{(k)}}}{1 + e^{\beta_{i_1}^{(k)} + \beta_{i_2}^{(k)} + \dots + \beta_{i_k}^{(k)}}}$$

where, for each $k = 2, \dots, r$, $\beta^{(k)} = (\beta_1^{(k)}, \dots, \beta_n^{(k)})$. There are $(r - 1)n$ parameters in this parametrization. By using (1) again, we obtain

$$\mathbb{P}_\beta(x) = \prod_{k=2}^r \prod_{e \in \binom{[n]}{k}} \frac{e^{\tilde{\beta}_e^{(k)} x_e}}{1 + e^{\tilde{\beta}_e^{(k)}}} = \exp \left\{ \sum_{k=2}^r \sum_{i \in V} d_i^{(k)}(x) \beta_i^{(k)} - \psi(\beta) \right\},$$

where $d_i^{(k)}$ is the number of hyperedges of size k to which node i belongs in x . Notice that the vector of sufficient statistics in this case is $\mathbf{d} = (d_1^{(2)}(x), \dots, d_n^{(2)}(x), d_1^{(3)}(x), \dots, d_n^{(3)}(x), \dots, d_1^{(r)}(x), \dots, d_n^{(r)}(x))$, and the normalizing constant is

$$\psi(\beta) = \sum_{k=2}^r \sum_{e \in \binom{[n]}{k}} \log(1 + e^{\tilde{\beta}_e^{(k)}}). \tag{5}$$

General hypergraphs. In the third variant of the model we define one parameter for each node, controlling the propensity of that node to be in a relation of any size. The probability of observing a hypergraph x is thus

$$\mathbb{P}_\beta(x) = \frac{\exp \left\{ \sum_{k=2}^r \sum_{e \in \binom{[n]}{k}} \tilde{\beta}_e x_e \right\}}{\prod_{k=2}^r \prod_{e \in \binom{[n]}{k}} 1 + e^{\tilde{\beta}_e}} = \exp \left\{ \sum_{i \in V} d_i(x) \beta_i - \psi(\beta) \right\}.$$

The vector of sufficient statistics is then $\mathbf{d} = (d_1(x), \dots, d_n(x))$, where $d_i(x) = \sum_{k=2}^r d_i^{(k)}(x)$, and the normalizing constant is $\psi(\beta) = \sum_{k=2}^r \sum_{e \in \binom{[n]}{k}} \log(1 + e^{\tilde{\beta}_e})$.

3 Parameter estimation

Iterative proportional scaling algorithms. From the theory of exponential families, it is known that the MLE $\hat{\beta}$ satisfies the following system of equations:

$$\frac{\partial \psi(\hat{\beta})}{\partial \hat{\beta}_i} = \bar{d}_i, \quad \text{for } i \in \{1, \dots, n\}, \tag{6}$$

where \bar{d} is the average observed degree sequence. By using (4), we then obtain

$$\sum_{s \in \binom{[n] \setminus \{i\}}{k-1}} \frac{e^{\hat{\beta}_s + \hat{\beta}_i}}{1 + e^{\hat{\beta}_s + \hat{\beta}_i}} = \bar{d}_i, \quad \text{for } i \in \{1, \dots, n\}, \tag{7}$$

which is itself equivalent to $\sum_{s \in \binom{[n] \setminus \{i\}}{k-1}} \hat{p}_{s,i} = \bar{d}_i$, for $i \in \{1, \dots, n\}$.

Iterative proportional scaling (IPS) algorithms fit the necessary margins of a provided table, whose elements correspond to the mean-value parameters (in this case probabilities of observing an edge). We design the following IPS algorithm for computing \hat{p} .

Algorithm 2.

Define $A = (a_{i_1, \dots, i_k})$ to be an $n \times \dots \times n$ k -way table with margins $\bar{d}_1, \dots, \bar{d}_n$ for all its layers. Set the following structural zeros on the table: $a_{i_1, \dots, i_k} = 0$ if $i_a = i_b$ for at least one pair $a \neq b$, $1 \leq a, b \leq k$. (Note that there are $n(n-1) \dots (n-(k-1))$ non-zero elements in the table.) Place $2\bar{e}/(n(n-1) \dots (n-(k-1)))$ on all other elements of the matrix, where $2\bar{e} = \sum_{i=1}^n \bar{d}_i$. Then apply the following iterative $(t+1)$ st step for every element a_{i_1, \dots, i_k} : $a_{i_1, \dots, i_k}^{(t+1)} = a_{i_1, \dots, i_k}^{(t)} (F_{i_1}^{(t)} \dots F_{i_k}^{(t)})^{1/k}$, where $F_{i_b}(t) = d_{i_b} / \sum_{s \in \binom{[n] \setminus \{i_b\}}{k-1}} a_{i_b, i_s}^{(s)}$.

IPS algorithms are known to converge to elements of the limiting matrix $(\hat{p}_{i_1, \dots, i_k})$ which are unique and preserve all the marginals (see e.g. [1]). Solving the system (3) for every $1 \leq i_1 < \dots < i_k \leq n$ provides $\hat{\beta}$. Algorithm 2 can be adjusted for layered uniform and general hypergraph beta models.

For layered k -uniform hypergraphs, by using (6) and (5) we obtain for $i \in \{1, \dots, n\}$ and $k \in \{2, \dots, r\}$,

$$\sum_{s \in \binom{[n] \setminus \{i\}}{k-1}} \hat{p}_{s,i} = \bar{d}_i^{(k)}. \tag{8}$$

Therefore, we can apply Algorithm 2 to $(r-1)$ k -way tables similar to those of the k -uniform case, where k ranges from 2 to r .

For general hypergraphs, we similarly obtain

$$\sum_{k=2}^r \sum_{s \in \binom{[n] \setminus \{i\}}{k-1}} \hat{p}_{s,i} = \bar{d}_i, \quad \text{for } i \in \{1, \dots, n\}. \tag{9}$$

In this case we apply the IPS algorithm to the following table: Define $A = (a_{i_1, \dots, i_k})$ to be a k -way table of size $(n+1) \times (n+1) \times \dots \times (n+1)$ consisting of labels $(\emptyset, 1, 2, \dots, n)$ with margins $\bar{d}_\emptyset, \bar{d}_1, \dots, \bar{d}_n$ for all its layers, where \bar{d}_\emptyset does not need to be known or calculated. We also set the following structural zeros in the table: $a_{i_1, \dots, i_k} = 0$ if (1) $i_a = i_b \neq \emptyset$ for at least one pair $a \neq b$, $1 \leq a, b \leq k$; (2) $i_1 = \dots = i_k = \emptyset$ except possibly for one i_b . We apply Algorithm 2 as in the k -uniform case except the fact that we do not fit the \bar{d}_\emptyset margins. We read the elements of the limiting matrix of from, $\hat{p}_{\emptyset, s}$ as \hat{p}_s , which corresponds to a lower dimensional probability.

Fixed Point Algorithms. An alternative method for computing MLE is based on [3]. In the k -uniform case, for $i \in \{1, \dots, n\}$, Equation (7) can be rewritten as

$$\hat{\beta}_i = \log d_i - \log \sum_{s \in \binom{[n] \setminus \{i\}}{k-1}} \frac{e^{\hat{\beta}_s}}{1 + e^{\hat{\beta}_s + \hat{\beta}_i}} := \varphi_i(\hat{\beta}). \tag{10}$$

Therefore, in order to find $\hat{\beta}$, it is sufficient to find the fixed point of the function φ .

Algorithm 3.

Start from any $\hat{\beta}_{(0)}$ and define $\hat{\beta}_{(l+1)} = \varphi(\hat{\beta}_{(l)})$ for $l = 0, 1, 2, \dots$

Theorem 4.

If the MLE exists, Algorithm 3 converges geometrically fast; if the MLE does not exist there is a diverging subsequence in $\{\hat{\beta}_{(l)}\}$.

The proof is omitted due to space limitations. For the other models, the above theory can be easily generalized. For the layered models and general hypergraph models, we apply the same algorithm to obtain the fixed points of the following functions respectively for $i \in \{1, \dots, n\}$ and $k \in \{2, \dots, r\}$ and $i \in \{1, \dots, n\}$.

$$\varphi_i(\hat{\beta}^{(k)}) := \log d_i^{(k)} - \log \sum_{s \in \binom{[n] \setminus \{i\}}{k-1}} \frac{e^{\hat{\beta}_s^{(k)}}}{1 + e^{\hat{\beta}_s^{(k)} + \hat{\beta}_i^{(k)}}; \tag{11}$$

$$\varphi_i(\hat{\beta}) := \log d_i - \log \sum_{k=2}^r \sum_{s \in \binom{[n] \setminus \{i\}}{k-1}} \frac{e^{\hat{\beta}_s}}{1 + e^{\hat{\beta}_s + \hat{\beta}_i}}. \tag{12}$$

4 Simulations and Analysis

MLE. We use the fixed point algorithm to estimate the natural parameters for hypergraph beta models, examine non-existence of MLE and compare the layered and general variants of the model on simulated data. Note that most dense hypergraphs, when reduced to binary relations give the complete graph, for which the MLE does not exist. In contrast, MLE is expected to exist for the hypergraph beta model in this case.

Example 1.

We simulate a hypergraph $H = (V, F)$ drawn from the beta model for 3-uniform hypergraphs on 10 vertices with $\beta = (-5.05, -0.57, 2.87, 4.85, 1.98, -6.69, -3.95, 5.97, -6.61, -4.24)$. The average simulated degree sequence of hypergraphs drawn from this model is $\bar{d} = (6.28, 10.70, 17.59, 20.81, 16.55, 4.41, 7.47, 23.02, 4.50, 7.17)$, and the average simulated density of the corresponding hypergraph is 0.33. Algorithm 3 provides the following MLE estimate using \bar{d} as the sufficient statistic: $\hat{\beta} = (-4.94, -0.58, 2.81, 4.76, 1.94, -6.55, -3.86, 5.86, -6.48, -4.15)$. Note that $\|\beta - \hat{\beta}\|_\infty = 0.14$.

For a larger example, we select a β value giving rise to 3-uniform hypergraphs on 100 vertices with density 0.44, and obtain a closer estimate: $\|\beta - \hat{\beta}\|_\infty = 0.12$.

Example 2.

Theorem 4 guarantees that if $\hat{\beta}$ is the solution to the ML equations (10), (12), or (6), then the sequence of β -estimates that the fixed point algorithm produces will converge to $\hat{\beta}$; else there

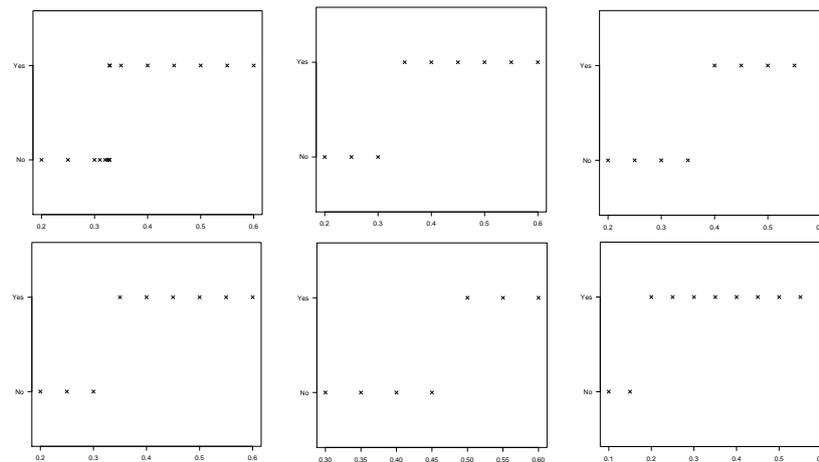


Figure 2: MLE existence against edge density; top row: 3-uniform beta on 25, 50, 100 vertices; bottom row: 4- and 2-uniform on 25 vertices, {2,3}-uniform on 50 vertices.

will be a divergent subsequence. To detect a divergent sequence in practice, we either look for a periodic subsequence, or for a number with large absolute value in the sequence that seems to be growing, sometimes quite slowly. From (2), since $e^{\beta_{i_k}} / (1 + e^{\beta_{i_k}})$ converges to 1 quickly ($e^{10} / (1 + e^{10}) > 0.9999$), for graphs with small number of nodes (*i.e.* far from the asymptotic behavior), it is plausible to conclude that the corresponding mean value parameter is approximately 0 or 1, and hence the MLE does not exist. Figure 2 demonstrates MLE existence against edge densities for random hypergraphs with a fixed edge-density. Interestingly, in this restricted class, our simulations give evidence of a transition from non-existence of the MLE to existence as the density of the hypergraphs increases. The transition point seems to depend on both the number of vertices and the edge sizes allowed in the model.

Model fitting: Layered versus general hypergraph beta models. Consider the two variants of the beta model for non-uniform hypergraphs: the general model, with one parameter β_i per node i , and the layered model, with one parameter $\beta_i^{(k)}$ per node i and edge size k . Since the former can be considered a submodel of the latter by setting certain constraints on $\beta_i^{(k)}$, $k \in \{1, \dots, r\}$, we compare the fit of these two models using the likelihood ratio test with test statistics $\lambda = 2 \log \mathcal{L}(\hat{\beta}_{\text{layered}}) - 2 \log \mathcal{L}(\hat{\beta}_{\text{general}})$. Our experiments indicate that the layered model fits significantly better than the general case. Using 100 random sequences on 10 vertices, with allowed edge-sizes 2 and 3, we obtain the average observed test statistics 53.649, in the critical region for 0.005 significance level, $(25.188, \infty)$, for chi-square with 10 degrees of freedom. The layered model fits significantly better for significance level 0.05 in all 100 cases, and 97 and 94 times better for significance levels 0.01 and 0.005, respectively.

Bibliography

- [1] Y. M. M. Bishop, S. E. Fienberg, and P. W. Holland. *Discrete Multivariate Analysis: Theory and Practice*. MIT Press, Cambridge, Mass.-London, 1975.
- [2] J. Blitzstein and P. Diaconis. A sequential importance sampling algorithm for generating random graphs with prescribed degrees. Available at

- www.people.fas.harvard.edu/~blitz/BlitzsteinDiaconisGraphAlgorithm.pdf, 2009.
- [3] Sourav Chatterjee, Persi Diaconis, and Allan Sly. Random graphs with a given degree sequence. *The Annals of Applied Probability*, 21(4):1400–1435, 2011.
 - [4] P. Erdős and A. Rényi. On random graphs. *Publicationes Mathematicae*, 6:290–297, 1959.
 - [5] Gourab Ghoshal, Vinko Zlatic, Guido Caldarelli, M. E. J. Newman. Random hypergraphs and their applications, 2009. <http://arxiv.org/abs/0903.0419>.
 - [6] Paul Holland, Samuel Leinhardt. An exponential family of probability distributions for directed graphs. *Journal of the American Statistical Association*, 76:33–50, 1981.
 - [7] Michal Karonski and Tomasz Luczak. The phase transition in a random hypergraph. *Journal of Computational and Applied Mathematics*, 142:125–135, 2002.
 - [8] Erik Kolaczyk: (2009): Statistical Analysis of Network Data: Methods and Models. Springer Series in Statistics.
 - [9] Alessandro Rinaldo, Sonja Petrović, and Stephen E. Fienberg. Maximum likelihood estimation in the beta model. *Annals of Statistics*, 41(3):1085–1110, 2013.
 - [10] T Yan, J Xu, and Y Yang. High dimensional Wilks phenomena in random graph models. Available at <http://arxiv.org/abs/1201.0058>, 2012.
 - [11] T Yan and J Xu. A central limit theorem in the β -model for undirected random graphs with a diverging number of vertices. Available at <http://arxiv.org/abs/1202.3307>, 2012.

Efficiency of Sequential Monte Carlo and Genetic algorithm in Bayesian estimation of the atmospheric contamination source

Anna Wawrzynczak, *National Centre for Nuclear Research, Poland,*
a.wawrzynczak@ncbj.gov.pl

Piotr Kopka, *1)National Centre for Nuclear Research, Poland; 2)Institute of Computer Science, Polish Academy of Sciences, p.kopka@ncbj.gov.pl*

Marcin Jaroszyński, *Inst. of Computer Sci., Siedlce University, Poland,*
marcinjaro89@gmail.com

Mieczyslaw Borysiewicz, *National Centre for Nuclear Research, Poland, manhaz@ncbj.gov.pl*

Abstract. Abrupt releases of hazardous material into the atmosphere pose a great threat to the human health and the environment. It is crucial to develop the emergency action support system which can quickly identify probable location and characteristics of the contamination source, by measuring concentration of certain substance using the sensors' network. Bayesian inference is a powerful tool able to combine observed data with prior knowledge, used to find the most probable values of the searched parameters. We apply the methodology combining Bayesian inference with Sequential Monte Carlo (SMC) and Genetic algorithm (GA) to the problem of the atmospheric contaminant source localization. Presented algorithms scan 5-dimensional parameters' space for the contaminant source coordinates (x, y) , release strength (Q) and atmospheric transport dispersion coefficients. In recent years the popularity of the nature inspired algorithms like GA increased, hence we compare the results given by SMC and GA algorithms. Performed tests show that both SMC and GA give comparable results, but GA estimates the correct parameters value faster, which results in the higher estimation probability.

Keywords. Bayesian inference, Event reconstruction, SMC methods, Genetic algorithm, Atmospheric contamination

1 Introduction

In the case of a sudden atmospheric release of chemical, radioactive or biological material, emergency responders need to quickly determine the location of dispersed substance's source. Thus, it is important to develop the emergency system that can estimate the most probable location of the atmospheric contamination source by measuring the concentration of dangerous substance using the network of sensors.

Knowing both gas source and wind field the appropriate atmospheric dispersion model can be applied to calculate the expected gas concentration for any downwind location. Conversely, given concentration measurements and knowledge of the wind field and other atmospheric air parameters, finding the location of the release source and its parameters is most improbable. This problem has no unique solution and can be considered only in the probabilistic frameworks. The issue boils down to the creation of the atmospheric dispersion model realistically reflecting the real situation, based only on a sparse point-concentration data. This task requires specification of set of model's parameters. In the framework of Bayesian statistics all quantities included in the model are modeled as random variables with joint probability distributions. This randomness can be interpreted as parameter variability, and is reflected in the uncertainty of the true values expressed in terms of probability distributions. Bayesian methods reformulate the problem into a search for a solution based on efficient sampling of an ensemble of simulations, guided by comparisons with data.

The problem of the source term estimation was studied in literature grounded both in the deterministic and probabilistic approach (e.g. [1]). [2] introduced dynamic Bayesian modeling, and the Markov Chain Monte Carlo (MCMC) sampling to reconstruct a contaminant source. The effectiveness of MCMC in the localizing of the atmospheric contamination source based on the synthetic data experiment was presented in e.g. [3]. The advantage of the Sequential Monte Carlo over the MCMC in the estimation of the probable values of the source coordinates presents [4]. The problem of finding the 'best fitting' model's parameters, for which a forward atmospheric dispersion model's output will reach agreement with real observations, can be considered the optimization problem. Consequently, metaheuristics, such as genetic algorithms (GA) can be applied. Since introduction [5] GA has been successfully applied, as an alternative optimization tool, in a variety of areas (e.g. [6]).

In this paper SMC and GA are applied to the problem of localizing the abrupt atmospheric contamination source based on point-concentrations reported by the network. Comparison between the performance of SMC and GA is based on the synthetic experiment data.

2 Bayesian inference

The Bayes' theorem, as applied to an abrupt release problem, can be stated as follows:

$$P(M|D) \propto P(D|M)P(M) \quad (1)$$

where M represents possible model configuration and D represents observed data. For our application, Bayes' theorem describes the conditional probability $P(M|D)$ of certain source parameters (model configuration M) given observed measurements (D) at sensor locations. This conditional probability $P(M|D)$ is also known as a *posteriori* distribution and is related to the probability of the data conforming to a given model configuration $P(D|M)$, and to the possible

model configurations $P(M)$, before taking into account the measurements. The probability $P(D|M)$, for fixed D , is called the *likelihood* function, while $P(M)$ -*a priori* distribution [7]. To estimate the unknown source parameters M using (1), the *posteriori* distribution $P(M|D)$ must be sampled. Value of *likelihood* for a sample is computed by running a forward dispersion model with the given source parameters M , and comparison of the model predicted concentrations C_i^M with actual observations C_i^E in the points of sensors location. This function compares the predicted from model with observed data at the sensor locations as:

$$\ln[P(D|M)] = - \sum_{i=1}^N [\log(C_i^M) - \log(C_i^E)]^2. \quad (2)$$

The closer the predicted values are to the measured ones, the higher is the likelihood of the sampled source parameters.

To obtain the posterior distribution $P(M|D)$ of the source term parameters, SMC and GA are used as the parameters sampling procedure. This way we completely replace the Bayesian formulation with a sampling procedure to explore the model's parameter space. The *posterior* probability distribution (1) is computed directly from the resulting (from SMC or GA) sets of parameters' values and is estimated as:

$$P(M|D) = \frac{1}{n} \sum_{i=1}^n \delta(M_i - M). \quad (3)$$

Eq. (3) represents the probability of a particular model configuration M giving results that match the observations. Thus, $\delta(M_i - M) = 1$ when $M_i = M$, and 0 otherwise. If, in the estimated set, numerous models have the same configuration $P(M|D)$ increases through the summation, increasing the probability of these source parameters.

3 Problem setup

Our goal is to select the efficient algorithm to conduct dynamic inference of an unknown atmospheric release. To test the proposed methods some concentration data are required. To satisfy this requirement we have performed the simulation using the atmospheric dispersion second-order Closure Integrated PUFF Model (SCIPUFF) [8]. Correspondingly, a forward model is necessary to calculate the concentration C_i^M for the tested set of model parameters M , at each algorithm iteration. Here, as forward model, we selected the fast-running Gaussian plume dispersion model (e.g. [9]).

Synthetic data

The algorithms input data were generated by the SCIPUFF model computing the time-dependent field of expected concentrations. We assumed 10 sensors distributed randomly over 15km x 15km area (Fig. 1). The atmospheric contamination source was located at $x = 3\text{km}$, $y = 8\text{km}$, $H = 25\text{m}$ within the domain. The simulated release was continuous with rate $Q = 8000\text{g/s}$ and started 1 hour before first sensors measurements. The wind was directed along x axis with speed of 5m/s . Further, in this paper, we assume that the only algorithm input data we have, are reported every 15 minutes (in subsequent time steps) during 1.5 hour concentrations of dispersed substance registered by 10 sensors (Fig. 1).

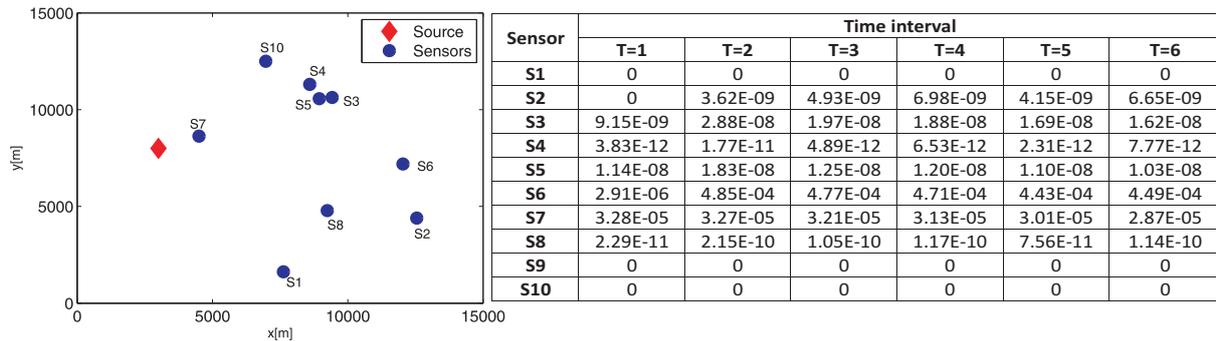


Figure 1: Distribution of the sensors and the release source within the considered domain, and concentrations $[g/m^3]$ reported by sensors in subsequent time intervals.

Forward dispersion model

The Gaussian plume e.g. [9] dispersion model for uniform steady wind conditions can be written out as follows:

$$C(x, y, z) = \frac{Q}{2\pi\sigma_y\sigma_z U} \exp\left[-\frac{1}{2}\left(\frac{y}{\sigma_y}\right)^2\right] \times \left\{ \exp\left[-\frac{1}{2}\left(\frac{z-H}{\sigma_z}\right)^2\right] + \exp\left[-\frac{1}{2}\left(\frac{z+H}{\sigma_z}\right)^2\right] \right\} \quad (4)$$

where $C(x, y, z)$ is the concentration at a particular location, U is the wind speed directed along x axis, Q is the emission rate and H is the height of the release; y and z are the distances along horizontal and vertical direction, respectively. In the equation (4) σ_y and σ_z are the standard deviation of concentration distribution in the crosswind and vertical direction. These two parameters were defined empirically for different stability conditions. In this work we restrict the diffusion to the stability class C. In scanning algorithm we assumed that we do not know exact behavior of the plume and consider these coefficients as unknown. Thus, the parameters σ_y and σ_z were taken as: $\sigma_y = z_1 \cdot x \cdot (1 + x \cdot 4 \cdot 10^{-5})^{-0.5}$, $\sigma_z = z_2 \cdot x$ where values z_1 and z_2 are sampled by algorithm within interval $[0.001, 0.35]$.

To summarize, in this paper the scanned model's parameter space is $M = (x, y, Q, z_1, z_2)$ where x and y are spatial coordinates of the source, Q release rate and z_1, z_2 the stochastic terms in the turbulent diffusion parametrization.

4 Reconstruction procedure

We run reconstruction algorithm, searching for the source location (x, y) , release rate (Q) and z_1 and z_2 , just after first sensors' measurements (data in time $T = 1$, Fig. 1). We assume that initially we have no *priori* information about the parameters' values. So, the initial value of each parameter is draw randomly from the predefined interval. Then, subsequent sets of parameters are evaluated by the SMC or GA algorithms, until the termination criterion is met. To reliably compare results reported by SMC and GA we have fixed the number of *likelihood* function value calculation to 6000 in each time step. As consequence, the termination criterion for GA is 40 generations and for the SMC 6 Markov Chains of length 1000. When the termination criterion is met the *posteriori* distributions of all parameters are calculated. Obtained *a posteriori* distributions are considered as the *priori* distributions in the subsequent time step. Consequently,

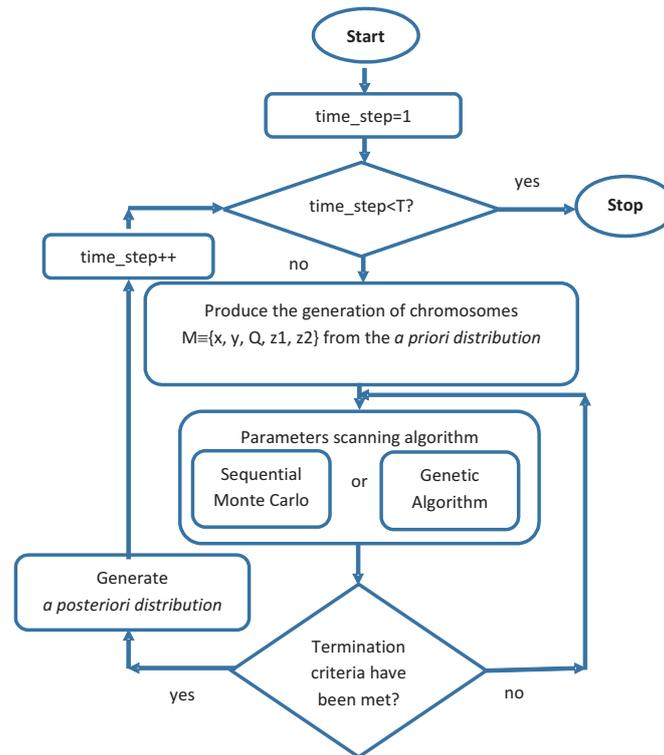


Figure 2: Flow chart of the reconstruction procedure.

in the next time step, when new data from the sensors arrive the initial population is drawn uniformly from the *priori* distribution i.e. *posteriori* distribution from previous time step. The flow chart of the reconstruction procedure presents Fig. 2.

Sequential Monte Carlo

In [4] we have shown that the SMC is definitely more effective than MCMC, in case of localizing the atmospheric contamination sources. SMC is designed to sample from dynamic posterior distributions, both in terms of use the dynamic nature of the model and also in terms of reusing previous calculations. SMC requires some set of samples to be initialized, thus in the initial phase the parallel MCMC Metropolis-Hastings procedure is applied. Then, state of all obtained chains, as a samples with weights, are passed to the SMC resampling procedure. The details of the applied algorithm are presented in [4].

Genetic algorithm

Algorithm starts with defining the initial population. The population is composed from the predefined number of chromosomes (here $n = 150$), $P(t) = x_1^t, \dots, x_n^t$, being initially randomly drawn from the admissible set of values. This set is explicitly defined by the space of explored parameters. GA chromosome is configured as binary value representing the real value of searched parameters M . The quality of each chromosome of current population is evaluated based on the objective/likelihood function (2). The 'improvement' of the current population is done by

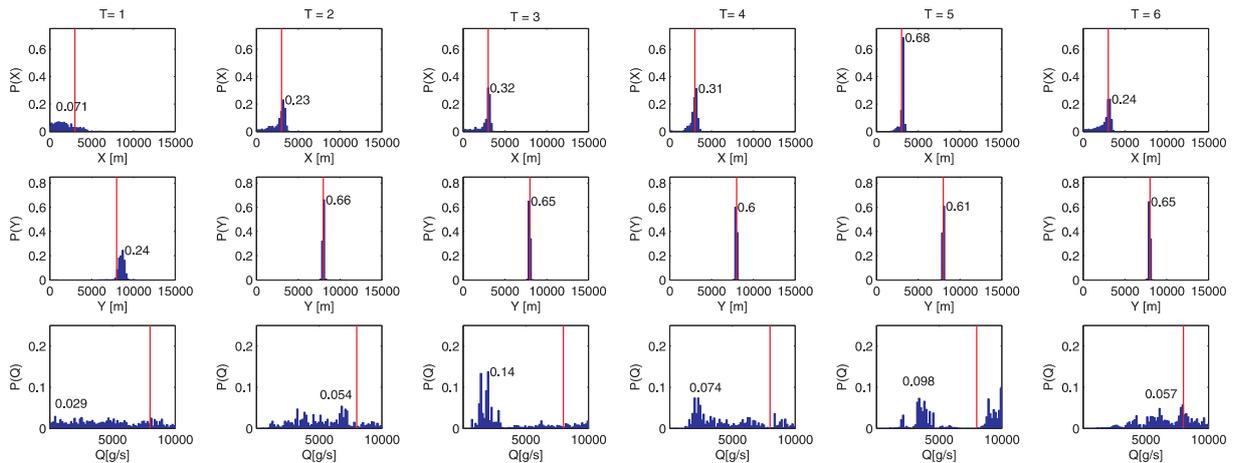


Figure 3: Resulted from SMC probability distributions of the models parameters x , y , and Q for the subsequent intervals. Red vertical lines represent the target value, the numbers represent the highest probabilities.

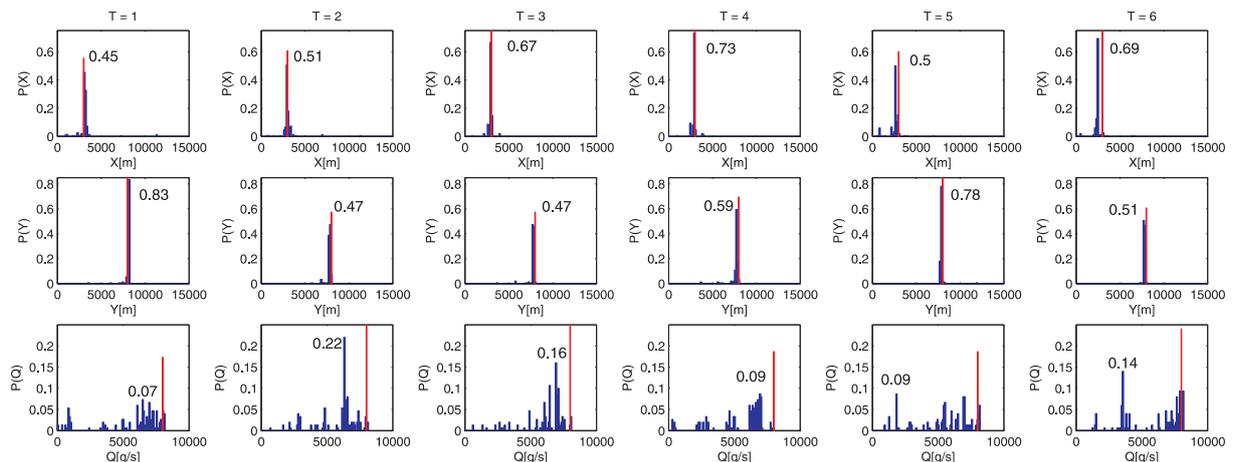


Figure 4: As in Fig. 3 but resulted from GA.

genetic operators.

Information about the quality of the each population's chromosome is used to perform selection. The hard tournament selection of size 2 was implemented. As the result, from each pair of the selected chromosomes one with the better likelihood function (Eq.2) value passes to the next population. Next, the crossover is performed. Crossover involves replacing parents by their children. We have applied the multi-point crossover with probability $CP = 0.75$, with 5 crossover points (corresponding to 5 searched parameters). Procedure begins with testing each chromosome for being the parent, in accordance with crossover probability CP . From the parents' population the unexploited pair is chosen, and then one crossover point for each parameter encoded in the chromosome is drawn. Parents are split, at the crossover points for each encoded parameter, then (in term of each encoded parameter) bits are swap resulting in two children. Subsequently, the current population is mutated by changing the chromosome's features. By

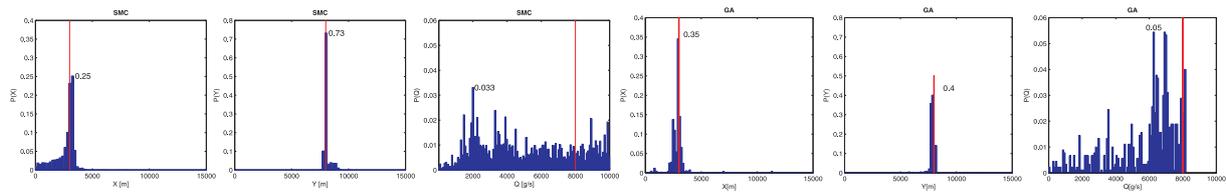


Figure 5: Resulted from SMC and GA cumulative probability distributions of the models parameters x , y , and Q . Red vertical lines represent the target value, the numbers represent the the highest parameter probabilities.

giving a chance of altering chromosome's individual bits, mutation allows to search for the entire solution's space, and not to converge to local extremes. The number of occurring mutations is determined by the mutation probability, here $MP = 0.02$. After performing the selection, crossover and mutation the new generation ($t+1$), being subject to the new evaluation, is established. After 40 generations the algorithm converges - it is expected that the best chromosome represents a near-optimum (reasonable) solution. More details of the GA operators can be found e.g. in [6].

5 Results and conclusions

In the problem presented in this paper the parameters M were limited by the intervals $x, y \in \langle 0, 15000m \rangle$, $Q \in \langle 1, 10000g/s \rangle$ and $z_1, z_2 \in \langle 0.001, 0.350 \rangle$. Figs. 3 and 4 presents the marginal probability distribution for x and y coordinates of source location and release rate Q , obtained with use of the SMC and GA, at each subsequent time interval. The exact source location and release rate, set up during creation of the testing synthetic data, is marked by the red vertical line. Probability distributions presented in Figs. 3 and 4 were obtained according to Eq. (3) based on the resampled samples in the case of SMC, and based on the 40th generation of chromosomes, in case of the GA. The estimated *posteriori* distributions obtained based on the data in given time interval were passed, as the *priori* distribution, to the succeeding reconstruction procedure iteration. We do not present the probability distributions for the z_1 and z_2 , as far the target value is not known. However, our previous works proved that 'freeing' the dispersion coefficients in some acceptable interval, allows to better fit the Gaussian plume to the 'real' data.

Comparing Figs. 3 and 4 one can see, that both SMC and GA algorithms finally (in $T = 6$) correctly pointed the target values of x and y parameters as the most probable. However, GA reached the near target value sooner than SMC. At first, in $T = 1$, the GA returned the probabilities: $P(x = 3075) = 0.45$, $P(y = 8175) = 0.83$ and $P(Q = 6520) = 0.07$, while the SMC pointed: $P(x = 1170) = 0.071$, $P(y = 8829) = 0.24$ and $P(Q = 496) = 0.029$. So, the GA estimations are closer to the target values i.e. $x = 3000$, $y = 8000$ and $Q = 8000$. In the subsequent reconstruction procedure iterations, the newly arrived data allowed to increase the performance of both methods, and in $T = 6$ the following parameters values were pointed as the most probable, GA : $P(x = 2475) = 0.69$, $P(y = 7725) = 0.51$ and $P(Q = 3560) = 0.14$, and the SMC: $P(x = 3297) = 0.24$, $P(y = 7978) = 0.65$ and $P(Q = 8013) = 0.057$. One can see, that SMC estimated the Q value better than GA. However, if we compare cumulative posteriori probabilities presented in Fig. 5, obtained based on distributions from all time steps, we observe that GA estimated the most probable Q value closer to the target one i.e. the GA

returned $P(x = 2925) = 0.35$, $P(y = 7875) = 0.40$ and $P(Q = 6280) = 0.05$, while SMC: $P(x = 3164) = 0.25$, $P(y = 8020) = 0.73$ and $P(Q = 2037) = 0.0033$.

We can conclude, that the performed test showed that GA is able to find the correct values of the contamination source coordinates quicker than SMC. Taking into account, that in the practical application time of response is crucial, the GA can be considered as more effective, than SMC which is usually applied in this type of problems.

Acknowledgement

This work was supported by the Welcome Programme of the Foundation for Polish Science operated within the European Union Innovative Economy Operational Programme 2007-2013 and by the EU and MSHE grant nr POIG.02.03.00-00-013/09.

Bibliography

- [1] Keats A., Yee E. and Lien F.S.,(2007),*Bayesian inference for source determination with applications to a complex urban environment*. Atmos. Environ. , **41**, 465–479.
- [2] Johannesson G., Chow F.K., Glascoe L., et al., (2005), *Sequential Monte-Carlo based framework for dynamic data-driven event reconstruction for atmospheric release*, Proc. of the Joint Statistical Meet., Minneapolis, American Stat. Ass. and Cosponsors, 73–80.
- [3] Borysiewicz M., Wawrzynczak A., Kopka P.,(2012), *Bayesian-Based Methods for the Estimation of the Unknown Model's Parameters in the Case of the Localization of the Atmospheric Contamination Source*, Found. of Computing and Decision Sci., **37**, 4, 253–270.
- [4] Wawrzynczak A., Kopka P., Borysiewicz M.,(2014), *Sequential Monte Carlo in Bayesian assessment of contaminant source localization based on the distributed sensors measurements*, Lecture Notes in Computer Sci. 8385, PPAM 2013, **II**, 407–417 .
- [5] Holland, J. H. (1992),*Adaptation in Natural and Artificial Systems*, Cambridge, MIT Press
- [6] Goldberg, D. E. (2006), *Genetic Algorithms in Search, Optimization and Machine Learning*, Addison Wesley Longman.
- [7] Gelman A., Carlin J., Stern H. and Rubin D.,(2003),*Bayesian Data Analysis*, Chapman and Hall/CRC.
- [8] Sykes, R.I., Parker S.F., Henn D.S., Cerasoli C.P. and Santos L.P. ,(1998) *PC-SCIPUFF Version 1.2PD Technical Documentation*, ARAP Report No. 718. Titan Corp.
- [9] Turner D. Bruce (1994) *Workbook of Atmospheric Dispersion Estimates*, Lewis Publishers, USA

Transfer of semiparametric single index model in binary classification

Muhammad-Anas Knefati, *LMA UMR CNRS 7348, Poitiers university,*
maknefati@hotmail.com

Farid Beninel, *LMA UMR CNRS 7348, Poitiers university,* fbeninel@gmail.com

Abstract. The semiparametric classification based on single index model is used in several domains of real life data engineering due to its flexibility. However, it has the same drawback as parametric classification: It is not suitable for the case where the training sample is derived from a certain subpopulation and the prediction sample from another one. The aim of this paper is to use the idea of transfer learning to reduce this drawback. Numerical experiments are performed and are intended to show the improvements from the prediction point of view.

Keywords. Supervised classification, Transfer learning, Parametric classification, Semiparametric classification, Single index model, Credit scoring, Morphometry.

1 Introduction

Classification is an important statistical field in many experimental sciences and real life applications. It aims to build predictive models to separate and classify data points in two or more groups. Here, we are interested in adapting or updating binary classification rules for some particular structure of data *i.e.*, the training sample and the prediction sample arise from two different subpopulations.

Classification methods are based on an estimate of $\mathbb{E}(Y|X = \mathbf{x})$ or more generally $g(\mathbb{E}(Y|X = \mathbf{x}))$ where g is a link function. These methods are classified into parametric, nonparametric and semiparametric.

Parametric methods assume that the function $\mathbb{E}(Y|X = \mathbf{x})$ is known up to a set of constant parameters that can be estimated from data. Several parametric methods have been proposed in this context such as discriminant analysis, logistic regression, see for instance [7] for a comprehensive review. Parametric methods have the advantage of being easily interpreted by practitioners, but rarely justified by theoretical or other *a priori* considerations related to the data design.

Nonparametric methods assume that the function of interest is unknown but smooth. No other assumptions about its shape or functional form are postulated. Therefore, they will be more flexible when data at hand does not fit strict classical statistical assumptions. In the

other hand, these methods have a serious drawbacks. One of them is that the estimation precision decreases rapidly as the dimension of the the covariate vector \mathbf{X} increases (curse of dimensionality). Another serious drawback is that they don't provide predictions of $E(Y|\mathbf{x})$ at points \mathbf{x} that are outside the considered support of the random variable \mathbf{X} .

Semiparametric methods are a trade off between the parametric and nonparametric ones. Their assumptions on the form of the function of interest are stronger than those of a nonparametric model but less restrictive than the assumptions of a parametric model, thereby reducing the possibility of specification error. Semiparametric methods give greater estimation precision than do nonparametric methods when X is multidimensional (see [11] for a review on semiparametric methods).

An approach that is very important in this domain is semiparametric single index models (SIM) that summarizes the effects of the feature measurements variable $\mathbf{X} = (X^1, \dots, X^p)^T$ within a single variable called the index or score for some specialists. In these models the conditional mean function has the form

$$\mathbb{E}(Y|\mathbf{X} = \mathbf{x}) = G(\beta^T \mathbf{x}), \quad (1)$$

where β is p -dimensional vector of real parameters and $G(\mathbb{R} \rightarrow \mathbb{R})$ a real function. These models mean that all the relevant information carried by X is contained in a linear combination of X components. Having the estimates of β and $G(\cdot)$, we can readily obtain the estimate of conditional mean from equation (1).

In this work, we deal with the binary classification *i.e.*, Y is a binary group label variable. The aim is to predict the group label value of a new individual, for which only the feature measurements $\mathbf{x} = (x^1, \dots, x^p)^T$ are known. We use the model

$$Y = f(\mathbf{X}) + \epsilon, \quad \epsilon \perp\!\!\!\perp \mathbf{X} \quad (2)$$

where $f(\mathbf{X}) = \mathbb{E}(Y|\mathbf{X})$ is estimated, using the training sample

$$\mathcal{S}_T = \{(Y_1, \mathbf{X}_1), \dots, (Y_n, \mathbf{X}_n)\}.$$

This problem is known as *supervised classification*. Several examples of such a problem are available such as in credit scoring, where we predict borrowers's behavior to pay pack loan by using information related to these customers. Another example in medicine, where we predict the risk of lung cancer recurrence for a patient previously treated, on the basis of used treatment for the first occurrence of the cancer and on some clinical and demographic measurements.

A main problem in supervised learning is that we assume that any individual to predict is supposed to be derived from the same statistical population as the training sample. Unfortunately, such assumptions are not realistic. For example, in credit scoring, to predict non customers behavior we use a training sample of costumers only. Also in medicine, the risk of lung cancer recurrence is learned from European patients and will be applied to Asian patients.

In order to avoid space limitations due to the available training sample, we use the transfer learning methodology which aims to transfer the knowledge from a source subpopulation to a target subpopulation.

This idea has been first proposed by Biernacki et al.[1] in the gaussian context, where they consider the case of two subpopulations slightly different. They establish that both subpopulations are linked through stochastic linear relationships. Estimation of the allocation rule (to be

applied on the non-labeled sample) is obtained by estimating parameters of this linear relationship, using several cases of constraints on this relation. They proved that this method is efficient and exhibits better performances than classical methods. Beninel and Biernacki[2] extended this approach to the multinomial logistic classification and proposed several additional links model in the case where the two studied subpopulations are gaussian ones. Beninel et al.[3] went in deep in the previous results with more tests and simulations in the context of credit scoring.

The semiparametric SIM in classification has a potential superiority over the classical classification methods. We develop here the idea of transfer learning to be applied in SIM as in the work of Beninel et al.[3].

This work is organized as follows: Section 2 is devoted to the building of the semi parametric single index model(SIM). The methodology of transfer learning and links models between source and target subpopulations are discussed in Section 3. The performance of the proposed method is assessed by means of a numerical experiments on two real examples in credit scoring and biology in Section 4.

2 Semiparametric single index model(SIM)

A semiparametric SIM has the form

$$Y = G(\beta^T X) + \epsilon, \quad \beta \in \mathbb{R}^p \tag{3}$$

where Y is the dependent variable, ϵ is the error such that $\mathbb{E}(\epsilon|\mathbf{X}) = 0$. The term $\beta^T \mathbf{X}$ is the *single index* or scoring.

For identification purpose on β and G , we suppose that \mathbf{X} must include at least one continuously distributed component whose associated β coefficient is non-zero. Also, we suppose that the model contains no intercept component. Last, we set the β coefficient of one component of X equal to one. This identification problem has been tackled by several authors. To name just a few, Manski[12] studied the identification of single index models for the case of binary response models. Ichimura[8] investigated the general case in which the response variable can be continue and he described a nonlinear least squares estimator of β . Klein and Spady[9] investigated the case of binary response models where they studied a semiparametric maximum likelihood estimator of β . Delecroix et al.[4] generalized the idea of Klein and Spady to arbitrary distributions for Y . Delecroix et al.[5] analyzed a large class of semiparametric M-estimators for single-index models, including semiparametric quasi-likelihood and semiparametric maximum likelihood estimators.

Now, we will review the methods of Ichimura[8] and Klein and Spady[9] to estimate β and G . These two methods use M-estimation as follows:

$$\hat{\beta} = \arg \min_{\beta \in \mathbb{R}^p} \sum_{i=1}^n \psi(Y_i, \hat{G}(\beta^T X_i; \beta)) \tau_n(X_i) \tag{4}$$

where

- $\hat{G}(t; \beta)$ is a nonparametric estimator of the regression function $E(Y|\beta^T \mathbf{X} = t)$. Here, we use the Nadaraya-Watson estimator *i.e.*,

$$\hat{G}(t; \beta) = \sum_{i=1}^n \frac{Y_i K(\frac{t - \beta^T \mathbf{X}_i}{h})}{\sum_{j=1}^n K(\frac{t - \beta^T \mathbf{X}_j}{h})} \tag{5}$$

where $K(\cdot)$ is a symmetric kernel and h is the smoothing parameter. In order to avoid degenerate problems, the version *leave-one-out* is considered to estimate G

$$\hat{G}^{(-i)}(\beta^T \mathbf{X}_i; \beta) = \sum_{k \neq i} \frac{Y_k K\left(\frac{\beta^T \mathbf{X}_i - \beta^T \mathbf{X}_k}{h}\right)}{\sum_{j \neq i} K\left(\frac{\beta^T \mathbf{X}_i - \beta^T \mathbf{X}_j}{h}\right)} \quad (6)$$

- $\tau_n(\cdot)$ is a trimming function. ψ is the loss function. In Ichimura[8] $\psi(y, r) = (y - r)^2$ and in Klein and Spady[9] $\psi(y, r) = -y \log(r) - (1 - y) \log(1 - r)$.

The estimator of β is very sensitive by the choice of h . Ichimura[8] proved that we can estimate, simultaneously, β and h from (4). Kong And Xia[10] defined a computational method that should be more efficient than the "classical" one, they called it "separated crossvalidation method". Given a threshold $s \in]0, 1[$, a new individual \mathbf{x}^* is allocated by $\hat{Y}(s) = \mathbb{1}_{\hat{G}(\hat{\beta}^T \mathbf{x}^*) \geq s}$.

3 Transfer learning

Methodology

We assume that the data consist of two samples: the first is $\mathcal{S} = \{(Y_1, \mathbf{X}_1), \dots, (Y_n, \mathbf{X}_n)\}$ with n points drawn from a source subpopulation \mathcal{U} and the second is $\mathcal{S}^* = \{(Y_1^*, \mathbf{X}_1^*), \dots, (Y_{n^*}^*, \mathbf{X}_{n^*}^*)\}$ with n^* points drawn from a target subpopulation \mathcal{U}^* . Here, the idea of the transfer is to allocate individuals from target subpopulation using both samples \mathcal{S} and \mathcal{S}^* .

From the training sample \mathcal{S} , we use the single index model to obtain the estimates $\hat{\beta}$ and \hat{G} . Then, we allocate individuals of \mathcal{S}^* using

$$\hat{E}(Y_j^* | \mathbf{X}_j^*) = \hat{G}(L(\mathbf{X}_j^*)), \quad j = 1, \dots, n^*, \quad (7)$$

where $L(\mathbf{X}_j^*) = c + \hat{\beta}^T \Lambda \mathbf{X}_j^*$, $c \in \mathbb{R}$ and $\Lambda = \text{diag}(\lambda_1, \dots, \lambda_p)$.

In order to estimate the $(p + 1)$ real parameters of (c, Λ) , we use the maximum likelihood function

$$\ell(c, \Lambda) = \sum_{j=1}^{n^*} Y_j^* \log(\hat{P}(Y_j^* = 1 | \mathbf{X}_j^*)) + (1 - Y_j^*) \log(1 - \hat{P}(Y_j^* | \mathbf{X}_j^*)) \quad (8)$$

By maximizing ℓ with respect to c and Λ after substituting $\hat{E}(Y_j^* | \mathbf{X}_j^*)$ from (7), we can obtain transfer parameters. In what follows, we discuss issues where the pairs (c, Λ) are unknown and based on the following postulated relationships:

Relationships

The estimation of parameters c and Λ can be done through several models depending on several possible situations for c and Λ that are:

- M_0 : No parameter to be estimated: $c = 0$ and $\Lambda = I_p$, where I_p is the identical matrix.
- M_1 : Here only c is to be estimated and $\Lambda = I_p$.
- M_2 : $c=0$ and $\Lambda = \lambda I_d$, where $\lambda \in \mathbb{R}$.

- M_3 : c is free and $\Lambda = \lambda I_d$ i.e., two parameters are to be estimated.
- M_4 : $c = 0$ and $\Lambda = \{\lambda_1, \dots, \lambda_p\}$, where $\lambda_1, \dots, \lambda_p \in \mathbb{R}$ are to be estimated.
- M_5 : The most complex model : c is free and $\Lambda = \text{diag}(\lambda_1, \dots, \lambda_p)$, where $\lambda_j \in \mathbb{R}$, $j = 1, \dots, p$.

4 Numerical experiments

The semiparametric transfer algorithm

First, we share \mathcal{S}^* into two samples: training sample \mathcal{S}_T^* and prediction sample \mathcal{S}_P^* . Here, the input of the algorithm are \mathcal{S} and \mathcal{S}_T^* , then the main steps of the transfer algorithm are as follow:

1. Calculate

$$L(X_j^*) = c + \hat{\beta}^T \Lambda X_j^*, \quad j = 1, \dots, n^* \quad (9)$$

2. Estimate the set of parameters $c \in \mathbb{R}$ and $\Lambda \in \mathbb{R}^p$ according to the chosen relationship by maximizing the empirical likelihood function given in (8) with respect to c and Λ .
3. Replace the estimated transfer parameters in (9) to obtain $L(\mathbf{X}^*)$ and then replace $L(\mathbf{X}^*)$ in (7) to allocate individuals in \mathcal{S}^* .
4. We predict the sample test $S_P^* = \mathcal{S}^* \setminus \mathcal{S}_T^*$ using the obtained estimator.
5. To measure the performance of our method, we calculate the error rate $e = \frac{1}{n_P^*} \sum_{j=1}^{n_P^*} \mathbb{1}_{Y_j^* \neq \hat{Y}_j^*}$, where n_P^* is the size of S_P^* and \hat{Y}_j^* is the predicted allocation of Y_j^* .

Biology Data

The data consist of three samples of seabirds that come from three subspecies of *Calanectris diomedea* species. These samples are: *Borealis* (n=206, 45% female) live in the Atlantic islands, *Diomedea* (n=35, 58% female) live in the Mediterranean islands, and *Edwardsii* (n=92, 52% female) live in the Cape Verde Islands.

Five morphological variables were measured to forecast the bird's sex. These variables are culmen, wing and tail lengths, tarsus and culmen depth. All simulations are replicated 50 times.

We use the *Borealis* subspecies as training sample (\mathcal{S}) to calculate $\hat{\beta}$ and \hat{G} which determinate the SIM estimator. First, We take the *Diomedea* subspecies as \mathcal{S}^* and the chosen sizes of samples \mathcal{S}_T^* are 10, 11, ..., 20. The left subfigure of the figure (1) illustrates the different obtained results. Second, we take the *Edwardsii* subspecies as \mathcal{S}^* and the chosen sizes of samples \mathcal{S}_T^* are 10, 15, ..., 70. The right subfigure of the figure (1) illustrates the different obtained results.

Credit scoring data

We consider a real data example in credit scoring on private loans from a southern German bank. The data set and the description of the variables are available at the web link http://www.stat.uni-muenchen.de/service/datenarchiv/kredit/kreditvar_e.html or see also for more description [6].

This data set consists in the description of 1000 consumers. For each consumer the binary response variable "creditability" is available ($Kredit=1$ for creditworthy and $Kredit=0$ otherwise). In addition, 20 covariates that are assumed to influence creditability were recorded. Here we are interested in the following six covariates :

laufkont *Balance of current account* with the following four categories:

1: no running account; 2: no balance or debit; 3: medium running account (less than 200 Deutsche Mark (DM)); 4: large running account (greater or equal to 200 DM or checking account for at least one year).

laufzeit: *Duration of credit in months*; sparkont: *Value of savings or stocks*;

moral: *Payment of previous credits*; weatkred: *Further running credits*.

beszeit: *Duration of employment* with five categories:

1: unemployed; 2: less than one year; 3: more than one year and less than four years;
4: more than four years and less than seven years; 5: more than seven years.

There are 700 observations with $Kredit = 1$ and 300 observations for $Kredit = 0$. For this experiment, we study the *borrowers non customers* behavior to pay back loans, we use the variable *Laufkont* to separate the available data set in two samples: the customers sample \mathcal{S} when $Laufkont > 1$ with 726 observations and the non customers sample \mathcal{S}^* when $Laufkont = 1$ with 274 observations.

We draw at random four training samples S_T^* of sizes: $n^* = 50, 100, 150, 200$ from the non costumers S^* . The figure (2) illustrates the different obtained results.

5 Conclusion

From these numerical experiments, we deduce the following conclusions:

First, the approach consisting in learning from the first data, to predict the label of individuals of another subpopulation (without any adaptation of the allocation rule ie, model M_0), leads to a high error rate. This is in contradiction with the assumptions underlying conventional methods of supervised classification. Also, this justifies the idea to use a maximum (possible) of individuals to adapt or update the allocation rule.

Second, the approach consisting in learning without exploiting the first data set is not satisfactory ; the results suffer from the small size of the second data set. Such problem of the size of the second data set is what generates the problem of transfer learning.

Third, the model (M_1) corresponds to existing practice and quite known among biologists (See Van Franeker and Ter Brack[13]) and credit scoring specialists. This practice consists in changing, empirically and without theoretical justification, the threshold value (or equivalently, the intercept of the linear score function or Anderson score) from which one affects to classes.

Finally, it is fairly clear that the best models (in the sense of the empirical error) ie, M_2 , M_3 and M_4 , are those exploiting the two data sets . The first is used to estimate a first allocation rule (appropriated to the prediction of individuals from the first sub-population) and the second to adapt and make it a rule to predict individuals of the second subpopulation.

Acknowledgement

We thank Prof. Marian Hristache(ENSAI-France) and the two anonymous referees for their constructive and helpful comments and suggestions.

Bibliography

- [1] Biernacki, C., Beninel, F. and Bretagnolle, V. (2002). *A generalized discriminant rule when training population and test population differ on their descriptive parameters*. *Biometrics*, **58**, 387–397.
- [2] Beninel, F. and C. Biernacki (2009). *Updating a logistic discriminant rule: Comparing some logistic submodels in credit-scoring*. International Conference on Agents and Artificial Intelligence, Porto, Portugal, pp. 267-274.
- [3] Beninel, F., Bouaguel, W. and Belmufti, G. (2012). *Transfer Learning Using Logistic Regression in Credit Scoring*. arXiv preprint arXiv:1212.6167.
- [4] Delecroix, M., Hardle, W. and Hristache, M. (2003). *Efficient estimation in conditional single-index regression*, *Journal of Multivariate Analysis* **86**, 213-226.
- [5] Delecroix, M., Hristache, M. and Patilea, V. (2006) *On semiparametric M-estimation in single-index regression* *Journal of Statistical Planning and inference* **136**, 730–769.
- [6] Fahrmeir, L. and Tutz, G. (2010). *Multivariate statistical modelling based on generalized linear models*. (2nd ed.). Springer.
- [7] Hastie, T., Tibshirani, R. and Friedman, J. H. (2011). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer.
- [8] Ichimura, H. (1993). *Semiparametric Least Squares (SLS) and Weighted SLS Estimation of Single-Index Models*. *Journal of Econometrics*, **58**, 71–120.
- [9] Klein, R.W. and Spady, R.H. (1993). *An efficient semiparametric estimator for binary response models*. *Econometrica*, **61**, 387–421.
- [10] Kong, E. and Xia, Y. (2007). *Variable selection for the single-index model*. *Biometrika*. **94**, 217–229.
- [11] Li, Q. and Racine, J.S. (2007) *Nonparametric econometrics: Theory and practice*. Princeton University Press.
- [12] Manski, C.F. (1988). *Identification of binary response models*. *Journal of the American Statistical Association*, **83**, 729–738.
- [13] Van Franeker, J. A. and Ter Brack, C. J. F. (1993). *A generalized discriminant for sexing fulmarine petrels from external measurements*. *The Auk* **110**, 492-502.

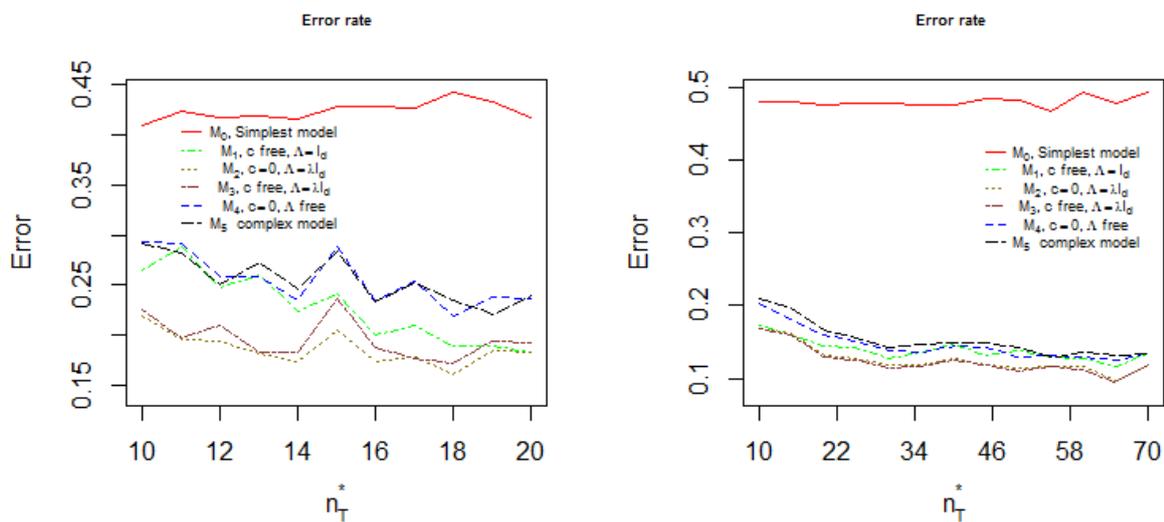


Figure 1: Borealis subspecies are training sample. In the left subfigure Diomedea subspecies are prediction sample while in the right subfigure Edwardsii subspecies are prediction sample.

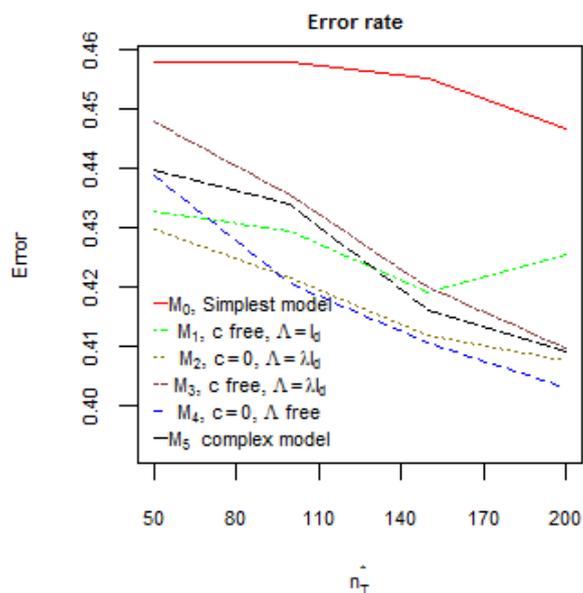


Figure 2: Customers are training sample and non customers are prediction sample.

Clustering ordinal data using binary decision trees

Badih Ghattas, *Aix Marseille University*, badih.ghattas@univ-amu.fr
Pierre Michel, *Aix Marseille University*, pierre.michel@univ-amu.fr

Abstract. We introduce here an extension of CUBT (Clustering using Unsupervised Binary Trees [1]) to ordinal data. CUBT is a hierarchical clustering method for continuous data inspired by CART that uses three steps to estimate an optimal partition of the data. The splitting process is based on a covariance type criterion. The pruning step uses a robust dissimilarity measure with Euclidean distance. Here, we extend this approach to ordinal data using mutual information and entropy criteria. Different simulations show the efficiency of our approach.

Keywords. CUBT, Clustering, Binary decision trees, Ordinal data, Mutual information

1 Introduction

CUBT [1] is a top-down hierarchical clustering method inspired by CART [2] that consists of three stages. The first step grows a “maximal tree” by recursively splitting the dataset into several subsets of observations and minimizing a heterogeneity criterion, the *deviance*, within the final clusters. This heterogeneity criterion is based on the trace of the covariance matrix within each subset of observations. The second step prunes the tree. For each pair of sibling terminal nodes (i.e. leaves with the same ascendant node), a measure of dissimilarity between the two nodes is computed. If this distance measure is lower than a certain threshold *mindist*, then the two nodes are aggregated into a single node (i.e. the parent node). The dissimilarity measure used by CUBT is based on the Euclidean distance. The final step, the *joining* is also a step of node aggregation, in which the constraint of node adjacency for aggregated clusters is not required. For the joining, either the deviance or the dissimilarity measure can be used.

CUBT shares various advantages with the CART method. It is flexible and efficient, i.e. produces good partitions for a large family of data structures; it is interpretable (because of the binary splits) and it has good convergence properties. In their original work Fraiman *et al.* compared CUBT with several clustering methods and showed that it produced quite satisfactory performance results.

Several applications of CUBT have been undertaken mainly in medicine and the social sciences; see, for example, [3]. This approach allowed clinicians to improve interpretability and

decision-making with regard to cut-off scores that define the membership of individuals in different clusters.

One of the limitations of the CUBT method is that the criteria used to grow and prune a tree (heterogeneity and dissimilarity criteria) are specific to continuous data.

We present here an extension of CUBT to ordinal data. For each step, growing and pruning, we suggest a new criterion based either on mutual information or on entropy. Section 2 describes the ordinal version of CUBT. Section 3 suggests certain simulation models and comparisons with other clustering methods. The final section gives the conclusion and proposes ideas for future work.

2 CUBT for ordinal data

We describe the three steps of CUBT using the new criteria for ordinal data. The first step grows the maximal tree, while the second and third steps (*pruning* and *joining*) prune the maximal tree.

Let $X \in \{1, \dots, m_j\}^p$ be a random p -dimensional discrete vector with coordinates X_j , $j \in \{1, \dots, p\}$, and let $m_j \in \mathbb{N}$ be the number of categories of the j^{th} variable. We have a set S of n random independent variables identically distributed as X , denoted X_i with $i \in \{1, \dots, n\}$. Finally, X_{ij} is the i^{th} observation of the j^{th} component of X . Similar notations are used with small letters to denote the realizations of these variables: x , x_i , x_j and x_{ij} .

For any node t (a set of observations), let $X^{(t)}$ be the restriction of X to node t i.e. $X^{(t)} = \{X|X \in t\}$, and we define $R(t)$, the heterogeneity measure of t , also called the *deviance*, as follows:

$$\begin{aligned} R(t) &= \text{trace}(\mathbf{MI}(X^{(t)})) \\ &= \sum_{j=1}^p H(X_j^{(t)}) \\ &= - \sum_{j=1}^p \sum_{i \in t} P(X_{i,j}) \log_2 P(X_{i,j}) \end{aligned}$$

where $\mathbf{MI}(X^{(t)})$ is the mutual information matrix of $X^{(t)}$, and $H(X_j^{(t)})$ is the Shannon entropy of $X_j^{(t)}$ within node t .

Growing stage

Initially, the primary node (the root) of the tree contains all the observations of S . The sample will be split recursively into two disjoint samples using binary splits with the form $x_j < a$, where $j \in \{1, \dots, p\}$ and a is a threshold over x_j . Thus, a split of a node t into two sibling nodes t_l and t_r is defined by a pair (j, a) . The nodes t_l and t_r are defined as follows:

$$\begin{aligned} t_l &= \{x \in \{1, \dots, m_j\}^p : x_j \leq a\} \\ t_r &= \{x \in \{1, \dots, m_j\}^p : x_j > a\} \end{aligned}$$

Let t be a node of a tree. Then, the best split of t into two sibling nodes t_l and t_r is defined by:

$$\text{argmax}_{(j,a) \in \{1, \dots, p\} \times \{1, \dots, m_j\}} \{\Delta(t, j, a)\}$$

with

$$\Delta(t, j, a) = R(t) - R(t_l) - R(t_r)$$

The stopping rules for growing the maximal tree are determined by two parameters $minsize \in \mathbb{N}$ and $mindev \in [0, 1]$. A node t having n_t observations is split no further whenever one of the two following criteria are satisfied:

1. $n_t < minsize$
2. $\Delta(t, j, a) < mindev \times R(S)$

Once the algorithm stops, a class label is assigned to each leaf of the maximal tree. A partition of the initial dataset is obtained and each leaf from the tree corresponds to a cluster.

Pruning stage

If the number of classes of the final partition k is known, the number of leaves of the maximal tree could be greater than k . Thus, it can be necessary to prune the tree by aggregating certain leaves. In this stage, a dissimilarity measure between two nodes is defined. Two nodes are aggregated if their dissimilarity measure is below a threshold ϵ . We introduce a new pruning criterion for the dissimilarity measure.

Let t_l and t_r be two adjacent nodes that share a direct ascendant node t . Let n_l (respectively n_r) be the size (number of observations) of the node t_l (respectively t_r) and $\alpha \in [0, 1]$. For each $X_i \in t_l$ and $X_j \in t_r$, with $i, j \in \{1, \dots, n\}$, we consider:

$$\tilde{d}_i = \min_{x \in t_l} d(x, X_i) \quad \text{and} \quad \tilde{d}_j = \min_{x \in t_r} d(X_j, x)$$

and their ordered versions d_i and d_j . Note that $d(X, Y)$ can be either the Manhattan distance or the mutual information between two random variables X and Y . For $\delta \in [0, 1]$, we define:

$$\bar{d}_l^\delta = \frac{1}{\delta n_l} \sum_{i=1}^{\delta n_l} d_i \quad \text{and} \quad \bar{d}_r^\delta = \frac{1}{\delta n_r} \sum_{i=1}^{\delta n_r} d_j$$

. Thus, the empirical dissimilarity measure between t_l and t_r is computed as follows:

$$d^\delta(l, r) = d^\delta(t_l, t_r) = \max(\bar{d}_l^\delta, \bar{d}_r^\delta).$$

At each step of the algorithm the leaves t_l and t_r are aggregated and replaced by their parent t if $d^\delta(l, r) \leq \epsilon$ with $\epsilon > 0$. The pruning stage requires two parameters, the proportion δ and a threshold ϵ corresponding to the minimal distance, called *mindist*.

Joining stage

The joining stage aggregates pairs of nodes that do not share the same ascendant (not sibling nodes) as in ascendant hierarchical clustering, successively joining the most similar pairs of clusters. Two joining criteria may be used for this step.

1. The first criterion used is the same as in the growing stage. Each pair of nodes t_l and t_r (sibling or not) is compared using the following measure (loss of deviance):

$$\Delta(t_l, t_r) = R(t_l \cup t_r) - R(t_l) - R(t_r)$$

The pairs of nodes with minimal loss of deviance are aggregated.

2. The second criterion is the same as the pruning stage. Pairs of nodes t_l and t_r are compared by computing $\Delta(t_l, t_r) = d^\delta(l, r)$. The pairs of nodes with minimal dissimilarity $\Delta(t_l, t_r)$ are aggregated.

For either criterion, let N_L be the number of leaves of the maximal tree. For each pair of values (i, j) , with $i, j \in \{1, \dots, N_L\}$ and $i \neq j$, we have $(\tilde{i}, \tilde{j}) = \operatorname{argmin}_{i,j} \{\Delta(t_i, t_j)\}$. The pair of nodes $t_{\tilde{i}}$ and $t_{\tilde{j}}$ are replaced by their union $t_{\tilde{i}} \cup t_{\tilde{j}}$ and $N_L = N_L - 1$.

There are two types of stopping rules depending on whether the number of clusters k is known or not.

- If k is known: The process is repeated until the number of tree leaves is below or equal to k . The stopping rule is the a priori number of classes.
- If k is unknown: The leaves are aggregated if $\Delta(t_l, t_r) < \eta$ where η can be a minimum threshold for the loss of deviance (*mindev*) or a minimum distance (*mindist*).

3 Experiments

In this section, we present certain simulations using different models. We consider only the case where the number of groups k is known a priori. We compare the results of CUBT with other methods such as hierarchical clustering using Manhattan distances, k -modes and k -medians. The misclassification error and the adjusted Rand index are used for these comparisons. We use the CUBT package [4] with **R** with the criterion described in the previous section.

Clustering methods

We compare our approach to three classical methods suited for ordinal data where the number of clusters needs to be known.

k -modes The k -modes algorithm [5] is a clustering method for categorical data that extends the k -means algorithm [6]. It seeks to partition the observations into k groups such that the distance from the observations to the cluster modes is minimized. We use the simple-matching distance to assess the dissimilarity between pairs of observations.

k -medians The k -medians approach is recommended in [7] for dealing with ordinal data. It is similar to the k -means algorithm except that it uses medians instead of means as centers for the clusters. We use the Manhattan distance to assess the dissimilarity between the observations and the cluster medians.

Hierarchical clustering We consider the classical agglomerative hierarchical cluster-analysis (HCA) method. We use the complete linkage option to aggregate clusters and the Manhattan distance as the dissimilarity measure of the observations.

Simulation models for ordinal data

We consider two data simulation models. We fix the number of groups $k = 3$ and, the dimension $p = 9$, and we test different sample sizes $n \in \{100, 300, 500, 1000\}$. For $j \in \{1, \dots, p\}$, we use the same number of levels for all the variables, $m_j = m = 5$. Each group has the same number of observations, $E(\frac{n}{k})$ where E is the floor function.

Frequentist simulation The first data simulation model uses a simple frequentist approach. Each variable X_j , $j \in \{1, \dots, p\}$ has $m = 5$ levels. We define three clusters, each characterized by a high frequency of one level. For observations from cluster 1, $P(X_j = 1) = q$, and a uniform probability is used for the other levels i.e. $P(X_j = l) = \frac{1-q}{m-1}$ for $l \neq 1$. For clusters 2 and 3, the frequent levels are 3 and respectively 5, using the same probabilities. We fix $q = 0.8$. This simulation model is very difficult for CUBT because $\sum_{j=1}^p X_j$ is a perfectly discriminating variable for the clusters, especially for high values of q ; $1 - q$ may be regarded as a clusters overlapping index.

IRT-based simulation The second simulation model uses item response theory (IRT) models. These models allow us to assess the probability of observing a level for each variable, given a latent trait factor. The latent trait is an unobservable continuous variable that defines the individuals' ability, measured by the observed variables. In the IRT framework, the variables are ordinal and are called items. The observations can be either binary or polytomous responses to the items. Here, we introduce a polytomous IRT model that helps us to generate data in a probabilistic way. The generalized partial credit model [8] (GPCM) is an IRT model that can address ordinal data. It is an extension of the 2-parameter logistic model for dichotomous data. The model is defined as follows:

$$p_{jx}(\theta) = P(X_{ij} = x|\theta) = \frac{\exp \sum_{k=0}^x \alpha_j(\theta_i - \beta_{jk})}{\sum_{r=0}^{m_j} \exp \sum_{k=0}^r \alpha_j(\theta_i - \beta_{jk})}$$

where θ is the latent trait and θ_i represents the latent trait level of the i^{th} individual. β_{jk} is a difficulty threshold parameter for the category k of the item j . For $j \in \{1, \dots, p\}$, β_j is a vector of dimension $m - 1$. α_j is a discrimination parameter represented by a scalar. We generate random datasets using the GPCM by simulating latent trait values for the three groups. For $c \in \{1, 2, 3\}$ we simulate a vector of latent trait values for each class c using $N(\mu_c, \sigma^2)$, $\mu = (-3, 0, 3)$ and $\sigma^2 = 0.1$. For $j \in \{1, \dots, p\}$, we fix $\alpha_j = 1.2$ and $\beta_j = (-1, -\frac{1}{3}, \frac{1}{3}, 1)$.

Tuning the method

We perform 100 replicates for each model. We compare our results with the results obtained using HCA, k -modes and k -medians. To apply CUBT, we fix values for the parameters involved at each stage of the algorithm (see Section 2). For the growing stage we use $minsize = E(\ln(n))$ and $mindev = 0.001$; for the pruning stage we fix $\delta = 0.3$ and $mindist$ as the fourth quintile of the distribution of distances between sibling nodes. We choose mutual information as the

measure of dissimilarity. Finally, the only parameter to be fixed in the joining stage is the number of classes $k = 3$.

As the true clusters are known we assess the performance of the different algorithms using the Adjusted Rand index and the matching error. Let y_1, \dots, y_n be the class labels of each observation, and let $\hat{y}_1, \dots, \hat{y}_n$ be the labels assigned to the n observations by a clustering algorithm. We denote by Σ the set of all possible permutations of the set of labels. The missclassification error rate, also called the “matching error” is defined as follows:

$$MCE = \min_{\sigma \in \Sigma} \frac{1}{n} \sum_{i=1}^n \mathbf{1}_{\{y_i \neq \sigma(\hat{y}_i)\}}$$

Results

Figure 1 shows an example of the trees we obtain for datasets drawn from our simulation models. Tables 1 and 2 show the results of the simulations. We report the matching error and the adjusted Rand index obtained for each clustering algorithm, namely HCA, k -modes, k -medians and CUBT, with four different sample sizes.

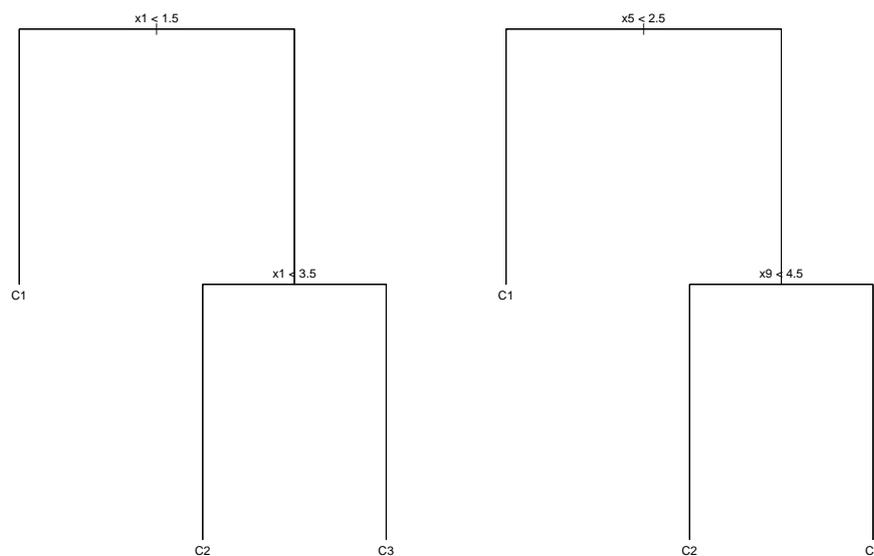


Figure 1: CUBT trees for the two models.

For the first model (frequentist) k -modes exhibits good performance, but it fails in the IRT-based model. The k -medians approach shows inferior results to k -modes in the first model but performs better in the second. HCA performs better in the second model, where its results are between the results of k -modes and k -medians, but fails in the first model. The results obtained by CUBT were similar to the results of HCA for the first model, but CUBT performs perfectly for the second model. CUBT yields better results than the other methods over the IRT-based

model	n	HCA	k -modes	k -medians	CUBT
Frequentist	100	0.064	0.008	0.084	0.150
	300	0.130	0.008	0.120	0.100
	500	0.160	0.022	0.074	0.150
	1000	0.150	0.005	0.075	0.160
IRT-based	100	0.023	0.062	0.009	0.005
	300	0.023	0.078	0.005	0.001
	500	0.025	0.062	0.005	0.001
	1000	0.024	0.077	0.005	0.001

Table 1: Simulations results: Matching Error for both frequentist and IRT models

model	n	HCA	k -modes	k -medians	CUBT
Frequentist	100	0.870	0.990	0.880	0.800
	300	0.800	0.990	0.840	0.740
	500	0.760	0.960	0.900	0.630
	1000	0.770	0.990	0.890	0.600
IRT-based	100	0.940	0.850	0.980	0.990
	300	0.940	0.820	0.980	1.000
	500	0.930	0.840	0.980	1.000
	1000	0.940	0.820	0.980	1.000

Table 2: Simulations results: Adjusted Rand Index for both frequentist and IRT models

model for all sample sizes. All these observations and results are retrieved and confirmed with regard to the Rand Index.

4 Conclusions

We have presented an ordinal version of the CUBT algorithm, which uses new criteria to handle this type of data. We have defined new criteria to use with ordinal data and compared this approach to other classical methods using simulations.

The results are quite satisfactory and even when CUBT does not outperform the other methods in all situations, it produces an interpretable clustering.

An extension of CUBT to nominal data can be performed directly, using the same criteria we introduce here. The only difference is the set of splits to explore. Continuous and qualitative data may be mixed using a mixing additive criterion for both types. These extensions are currently under consideration.

Bibliography

- [1] Fraiman, R., Ghattas, B. and Svarc, M. (2013) *Interpretable clustering using unsupervised binary trees*. *Data analysis and classification*, **7**, 125–145.

- [2] Breiman, L., Friedman, J., Stone, C.J. and Olshen, R.A. (1984) *Classification and regression trees*. Editions Chapman & Hall/CRC, Monterey, CA.
- [3] Michel, P., Boyer, L., Baumstarck, K., Fernandez, O., Flachenecker, P., Pelletier, J., Loundou, A., Ghattas, B. and Auquier, P. (2014) *Defining quality of life levels to enhance clinical interpretation in multiple sclerosis: application of a novel clustering method*. Medical Care Applied Methods (accepted).
- [4] Ghattas, B., Svarc, M. and Fraiman, R. (2013) *R-package for interpretable clustering using binary trees*. <http://lumimath.univ-mrs.fr/ghattas/CUBT.html>.
- [5] Huang, Z. (1998) *Extensions to the k-modes algorithm for clustering large data sets with categorical values*. Data Mining and Knowledge Discovery, **2(3)**, 283–304.
- [6] MacQueen, J. (1967) *Some methods for classification and analysis of multivariate observations*. Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Editions L. M. Le Cam & J. Neyman, **1**, 281–297.
- [7] Leisch, F. (2006) *A tool box for K-centroids cluster analysis*. Computational Statistics and Data Analysis, **51(2)**, 526–544.
- [8] Muraki, E. (1992) *A generalized partial credit model: application of an EM algorithm*. Applied Psychological Measurement, **16**, 159–176.

Bayesian blind source separation applied to the lymphocyte pathway

Katrin Illner, *Helmholtz Zentrum München and Technische Universität München*,
katrin.illner@helmholtz-muenchen.de

Christiane Fuchs, *Helmholtz Zentrum München and Technische Universität München*,
christiane.fuchs@helmholtz-muenchen.de

Fabian J Theis, *Helmholtz Zentrum München and Technische Universität München*,
fabian.theis@helmholtz-muenchen.de

Abstract. In many biological applications one observes a multivariate mixture of signals, where both the mixing process and the signals are unknown. Blind source separation can extract such source signals. Often the data have additional structure, i.e. the variables (e.g. genes) are linked by an interaction network. Recently, we developed the probabilistic method **emGrade** that explicitly uses this network structure as a Bayesian network and thus performs a more appropriate separation of the data than standard methods. Here, we consider the application of **emGrade** to gene expression data together with a literature-derived pathway. Thanks to the probabilistic modeling, we can use model selection criteria and demonstrate the relevance of the pathway information for explaining the data. We further use estimates of missing observations to identify the most appropriate microarray probe sets for two genes that were not uniquely annotated after standard filtering. Finally, we identify genes relevant for the dynamics underlying the data; these genes were not detected without the network information.

Keywords. expectation maximization, model selection, gene expression data, gene regulatory networks

1 Introduction

Blind source separation (BSS) is a widely used method to extract informative signals from a multivariate observed mixture. In many applications the data have additional structure that can be exploited to achieve a more appropriate signal separation. Recently, our group developed two algorithms that explicitly include the network structure – **Grade** (graph-decorrelation algorithm) [5] and its probabilistic extension **emGrade** (expectation maximization graph-decorrelation algorithm) [4]. In the latter the network structure is modeled as a Bayesian network and parameters and source signals are estimated using expectation maximization. In this manuscript we demonstrate the application of **emGrade** to gene expression data where the genes are linked by a gene

regulatory network. In Section 2, we briefly review the **emGrade** algorithm. As described in Section 3, we use publicly available microarray data for a lymphocyte pathway. In Section 4, we analyze the data using **emGrade**. The probabilistic framework enables us to use model selection criteria, and we find that the pathway information indeed improves our model. This has only been shown for synthetic data so far. Furthermore, we estimate missing observation values and determine the most appropriate microarray probe set for two genes that were not uniquely annotated after standard filtering. Finally, we characterize the estimated signals in terms of relevant genes and compare the gene sets from different observations. This leads the way to a biological interpretation of the estimated source signals. Section 5 concludes this paper.

Throughout the paper we use bold symbols to denote random variables and solid symbols to denote parameters and realizations of random variables, respectively.

2 The blind source separation method **emGrade**

In this section we shortly review the blind source separation method **emGrade** introduced in [4].

We assume observed Gaussian random variables $\mathbf{X} = (\mathbf{x}(i))_{i=1}^N$ with state space \mathbb{R}^m that are generated by the following linear mixing model:

$$\mathbf{x}(i) = A\mathbf{s}(i) + \mu + \varepsilon(i), \quad i = 1, \dots, N. \quad (1)$$

Here, $A \in \mathbb{R}^{m \times q}$ denotes the mixing matrix, $\mu \in \mathbb{R}^m$ is a common mean vector for all i , and $\varepsilon(i) \sim \mathcal{N}(0, \sigma^2 I)$ is additive measurement noise. The latent variables $\mathbf{S} = (\mathbf{s}(i))_{i=1}^N$ are normally distributed with state space \mathbb{R}^q ($q \leq m$). The components of these variables represent the source signals we are interested in, i. e. we have a source signal $\mathbf{s}_k = (\mathbf{s}_k(1), \dots, \mathbf{s}_k(N))$ for $k = 1, \dots, q$.

To define the (joint) distribution of the latent variables we assume a weighted directed acyclic graph $G = (V, E, \Lambda)$ that is determined a priori. Let $V = (v_1, \dots, v_N)$ be the set of nodes, $E \subset V \times V$ the set of edges, and let $\lambda_{ij} \in \mathbb{R}$ denote the weight assigned to the edge $(v_i, v_j) \in E$. We assume that the latent variables \mathbf{S} form a Bayesian network with respect to G , i. e. the latent variables are associated to the nodes V and the joint distribution decomposes as

$$(A0) \quad p(\mathbf{S}) = \prod_{i=n_0}^N p(\mathbf{s}(i) \mid \mathbf{Pa}(i)) \prod_{i=1}^{n_0-1} p(\mathbf{s}(i)).$$

Here, $\mathbf{Pa}(i)$ denotes the vector of all latent variables associated to the parent nodes of v_i , and we assume that v_1, \dots, v_{n_0} are root nodes. We then make the following stationarity and scaling assumptions where v_i and v_j are adjacent nodes:

$$\begin{aligned} (A1) \quad & \mathbb{E}[\mathbf{s}(i)] = \mathbf{0}_q, \\ (A2) \quad & \text{Cov}(\mathbf{s}(i), \mathbf{s}(i)) = I_q, \\ (A3) \quad & \text{Cov}(\mathbf{s}(i), \mathbf{s}(j)) = \lambda_{ij} D. \end{aligned}$$

We denote D as graph-delayed covariance, and we assume that it is diagonal. The assumptions (A0)-(A4) define a unique distribution of \mathbf{S} which is characterized by the conditional distributions in (A0). The parameter D occurs in (A0) as different rational terms.

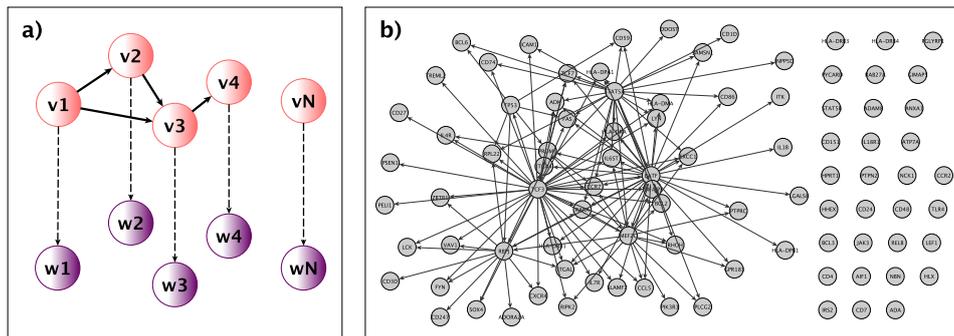


Figure 1: **Bayesian network for emGrade.** a) Graphical representation of the Bayesian network for emGrade with latent variables in red and observed variables in purple. The dependence between the latent variables is with respect to a known network structure, for instance a gene regulatory network. b) The pathway “lymphocyte activation” (net1) derived from the Genomatix database.

We now expand the Bayesian network and add nodes w_1, \dots, w_N that represent the observed variables (Figure 1a). For all $i = 1, \dots, N$ we insert an edge (v_i, w_i) , and the conditional distribution of the associated random variables is given as $\mathbf{x}(i) \mid \mathbf{s}(i) \sim \mathcal{N}(A\mathbf{s}(i) + \mu, \sigma^2 I)$. In the resulting Bayesian network we can estimate the latent variables \mathbf{S} and the model parameter $\theta = (A, \mu, \sigma^2, D)$ using expectation maximization. For the expectation step we use the Bayes net toolbox [6] and estimate the latent variables from their posterior distribution. If data points are missing (i.e. some variables $\mathbf{x}(i)$ are unobserved) we can simply treat them as additional latent variables. For the maximization step we derive explicit updates rules for A , μ and σ^2 and use numerical maximization for D . All (diagonal) entries of D can be estimated separately, and the domain depends on the network structure. Both steps are repeated alternately until convergence. Here we assume convergence if the change for all parameter entries is less than 10^{-8} , or if a maximum number of 10 000 iterations is achieved. The resulting values for the parameters and source signals then define the **emGrade** estimate.

3 The data

For the application of **emGrade** we consider gene expression data that are accessible online, and we use the Genomatix database [3] to derive a network structure for the gene interactions.

Gene expression data and pre-processing

In Calvano *et al.* [1] four healthy humans were treated with intravenous endotoxin, and gene expression measurements of blood leukocytes were taken at time points 0, 2, 4, 6, 9, and 24h after endotoxin administration. In a control study the leukocytes of four non-treated humans were taken at the same time points. After quantile normalization and filtering with the **limma** R-package [7] we get normalized expression values for 12 683 human genes. For source separation we consider a subset of $N \sim 100$ genes that are associated with a specific pathway. The derivation of the pathway is discussed in the next paragraph. We further divide the data into

the measurements for each individual. We thus have observations LPS1-4 from the four treated persons and observations PT1-4 from the non-treated persons. Each selected gene corresponds to an observed random variable, and since the measurements are taken at six time points we have $m = 6$ as the dimension of the observed variables.

In simulations we found that the performance of **emGrade** increases if the variance of the observed variables has a similar range compared to the variance of the unobserved variables. Since we assume $s(i) \sim \mathcal{N}(0, I)$ in (A2) we scale the variance of the components of $x(i)$ to 1, accordingly.

Literature-derived pathways

In our BSS method we assume an initially known network that describes the dependencies between the variables (genes). To derive such a network structure we use pathway information from the Genomatix Pathway System (GePS) [3]. Based on the expression data from [1] the database provides (amongst others) biological processes that are associated with changes between treatment and control group. One highly significant pathway is “lymphocyte activation” (net1) which consists of 91 genes and 138 edges (Figure 1b). For comparison we also investigate the less significant sub-pathway “cell proliferation” (net2). Since no further information about the strength of interaction is available, we fix all edge weights at $\lambda_{ij} = 1/\#\{\text{parents of } v_j\}$.

4 Results

In the following, we apply **emGrade** to the gene expression data and pathways from the previous section and present our main findings.

Comparison of different networks

In a first investigation we apply **emGrade** to all patients LPS1-4 and PL1-4 separately. As network structures we consider net1 and net2 and for comparison a network without any edges (net0). If we fix one data set, we can compare the BIC values for different network structures and different numbers of source signals $q = 1, 2, 3, 4$. Figure 2 indicates that for all data sets the informative net1 is more appropriate compared to net0 (lower BIC values). Furthermore, we find that for the treatment groups LPS1-4 a higher number of source signals is preferred.

Missing observation values

We now investigate the predictive power of our model for missing observation values. As already stated in Section 2, we can easily treat missing observations as additional latent variables in our Bayesian network. We therefore leave out the observation value for one gene in the data LPS1 and compare the estimate $\hat{x}(i) = \hat{A}\hat{s}(i) + \hat{\mu}$ to the true observation $x(i)$. We assume net1 to be the underlying network and use genes that are highly connected as well as genes that are not connected. Figure 3 shows the Euclidean distances $\|x(i) - \hat{x}(i)\|_2$ for 10 different missing genes and for $q = 1, 2, 3, 4$ estimated source signals. For comparison we estimate parameters and source signals from the complete data set. As expected, we find a better agreement of $\hat{x}(i)$ with $x(i)$ in this case.

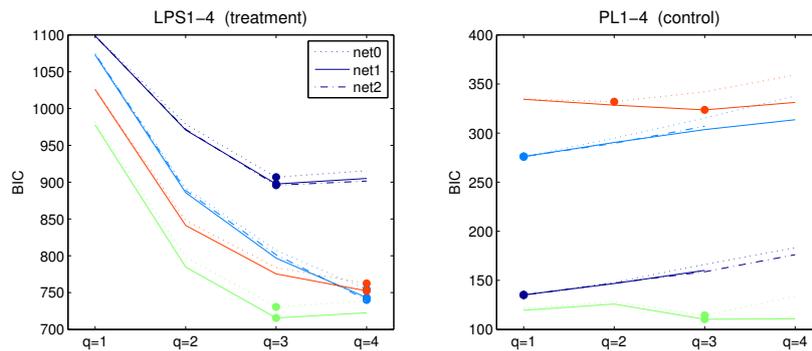


Figure 2: **Comparison of different networks.** The plots show the BIC values for patients LPS1-4 (left) and PL1-4 (right) in case of $q = 1, 2, 3$ and 4 source signals. As network structures we consider net0 and net1 and for LPS1-2 also net2. The different patients are coded in different colors and the dots indicate the value for q with the lowest BIC value.

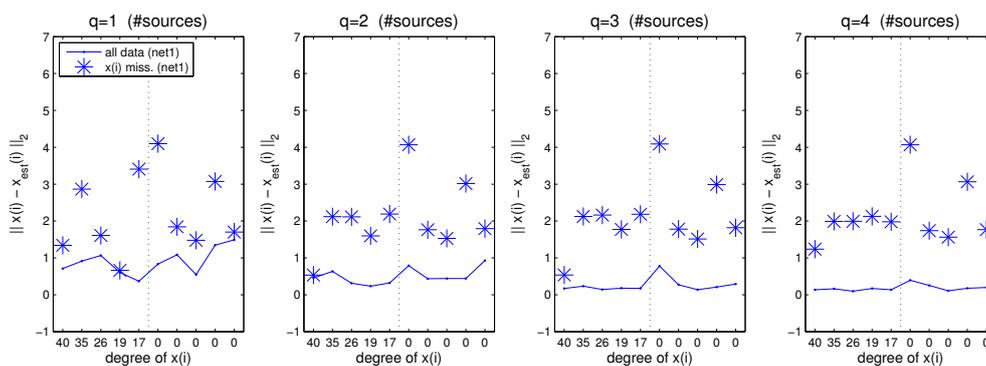


Figure 3: **Reconstruction of missing observation values.** We consider data LPS1 where we leave out the measurements of one gene, and we use net1 as network structures. We then apply **emGrade** with $q = 1, 2, 3, 4$ sources (left to right). The stars illustrate the Euclidean distances between the estimates $\hat{x}(i) = \hat{A}\hat{s}(i) + \hat{\mu}$ and the true observation values $x(i)$ where we consider different missing genes with high and low connectivity (degree). The solid lines show the corresponding distances when $\hat{x}(i)$ is estimated from the complete data set.

The estimation of missing observations provides a useful feature in the present data situation: The genes HLA-DRB1 and HLA-DRB3 from net1 are annotated to 5 and 2 probe sets of the microarray chip. Gene filtering performed with the **limma** R-package omits these genes and one does not know which probe set provides the most appropriate expression values. We therefore treat both genes as missing observations and compare our estimates to the measurements of the different probe sets. Table 1 shows the microarray measurements of all probe sets together with our estimates. The comparison suggests to use the 4th probe set as observation for HLA-DRB1 and the 2nd probe set as observation for HLA-DRB3.

time	HLA-DRB 1						HLA-DRB 3		
	obs.1	obs.2	obs.3	obs.4	obs.5	est.	obs.1	obs.2	est.
0h	4.99	5.23	5.26	5.20	2.46	5.44	5.20	2.46	2.52
2h	4.24	4.26	4.38	4.11	2.76	3.74	4.11	2.76	2.29
4h	4.26	4.65	4.47	4.43	2.54	4.66	4.43	2.54	2.81
6h	4.52	4.81	4.58	4.66	2.76	4.60	4.66	2.76	2.38
9h	4.66	5.21	5.22	5.04	2.62	5.15	5.04	2.62	2.61
24h	4.86	5.28	5.12	5.23	2.08	5.11	5.23	2.08	2.21

Table 1: **Identification of the most appropriate microarray probe set.** The table shows the microarray measurements from LPS1 at all probe sets that are linked to the genes HLA-DRB1 and HLA-DRB3. If we treat both genes as missing observations we get estimated observations (red). The comparison to the measurements identifies the most appropriate annotated probe set for both genes (bold symbols).

Genes associated to source signals

Next, we determine key genes associated with the estimated source signals, and we compare source signals that are estimated from different observations. Let $s_k = (s_k(1), \dots, s_k(N))$ be an estimated source signal. Based on a cut-off $c > 0$ we select all genes with absolute value larger than c . This yields a set of key genes that characterize s_k . With this we can compare key genes of source signals that are estimated from different observations. We compare the treatment groups LPS1-3 and the control groups PL1-3. Since the estimated source signals are unique only up to sign and permutation we first align the source signals and minimize

$$\min_{P_1, P_2} \{ \|P_1 S_1 - P_2 S_2\|_2 + \|P_1 S_1 - S_3\|_2 + \|P_2 S_2 - S_3\|_2 \}.$$

Here, P_1 and P_2 are matrices with one entry ± 1 per row and column and all other entries equal to zero. Figure 4 illustrates the alignment of source signals and Figure 5 indicates that we have a higher key gene agreement for the treatment groups LPS1-3 compared to the control groups PL1-3. For the control groups only the network without edges (net0) yields a source with high agreement of key genes.

5 Discussion and conclusion

In this manuscript we applied the recently developed blind source separation method **emGrade** to gene expression data. The method aims to separate multivariate data with known network structure into informative source signals. We discussed the pre-processing of publicly available microarray data consisting of treatment data LPS1-4 and control data PL1-4. From the Genomatix database we derived the “lymphocyte activation” pathway which reflects differences between the control and treatment group. We then applied **emGrade** to this data set.

In comparison to a network without edges, the pathway information improved our estimates and we found lower BIC values – this was true for the treatment and the control group. Nevertheless, the pathway information played a major role particularly for the treatment group where more source signals were preferred. We further investigated the estimation of missing observation values. For two genes from the lymphocyte pathway standard annotation to one unique

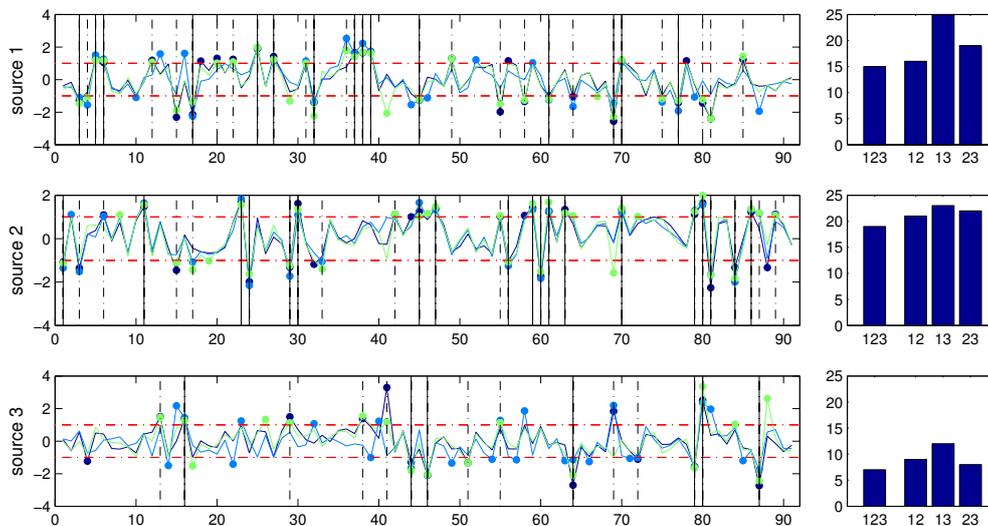


Figure 4: **Alignment of source signals and intersection of key genes for LPS.** For patients LPS1, LPS2 and LPS3 and network structure net1 we determine the *emGrade* source estimates. The plots on the left show the aligned source signals (different patients in different colors) together with the selected key genes (dots). The horizontal red lines show the cut-off for key gene selection. Solid vertical lines indicate genes that are key genes in the aligned sources from all patients, dashed vertical lines indicate genes that are key genes in the aligned sources from at least two patients. The bars on the right provide the counts of key genes in all estimates (123) and the counts of key genes in two estimates (12), (13) and (23).

microarray probe set failed. When treated as missing observations *emGrade* could identify the most appropriate annotated probe set in both cases. Finally, we characterized the estimated source signals in terms of key genes, i. e. genes with high absolute value in the respective source signals. We found a high number of key genes (per source) that were in agreement with LPS1-3. For PL1-3 these numbers were lower. This might again indicate that the “lymphocyte activation” pathway better explains the dynamics in the treatment group.

In our ongoing work we extend the proposed BSS model and consider different pathway structures for each source signal. We aim to separate the data into a pre-defined set of pathways and determine the impact of the estimated source signals in terms of the graph-delayed covariance.

Acknowledgements

The authors very much thank Steffen Sass and Nikola Müller for helping with the data pre-processing and the pathway extraction from the Genomatix database. We also thank Justin Feigelman for careful proofreading. Furthermore, the work was financially supported by the German Federal Ministry of Education and Research (BMBF) within the GerontoSys project

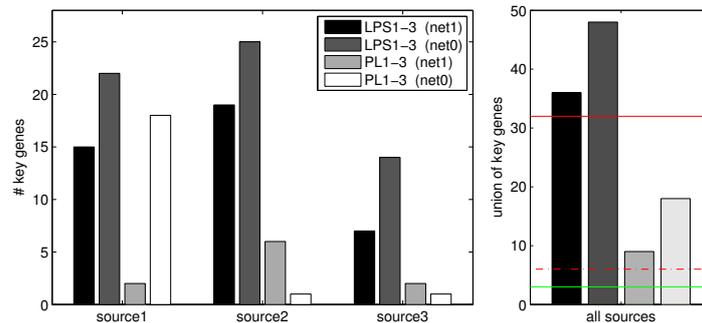


Figure 5: **Intersection of key genes for LPS and PL.** For patients LPS1-3 and PL1-3 and network structures net1 and net0 we determine the `emGrade` source estimates. The left figure shows the counts of genes that are key genes in all aligned source estimates from LPS1-3 or PL1-3, respectively. The right figure gives the total number of key genes in all sources. The solid red line indicates the count of identical key genes in LPS1-3 (net1) and LPS1-3 (net0), the dashed red line indicates the corresponding count for PL1-3. The green line is the count of identical key genes in all four groups.

“Stromal Aging” (Grant No. FKZ 0315576C), the German Research Foundation (DFG) within the project “InKoMBio” (Grant No. BO 3834/1-1), and the European Union within the ERC grant “LatentCauses”.

Bibliography

- [1] Calvano, S. E., et al. (2005). *A network-based analysis of systemic inflammation in humans*. Nature, **437**(7061), 1032–1037.
- [2] Friedman, N., Linial, M., Nachman, I., and Pe’er, D. (2000). *Using Bayesian networks to analyze expression data*. Journal of Computational Biology, **7**(3-4), 601–620.
- [3] Genomatix Pathway system (GePS), <http://www.genomatix.de/>
- [4] Illner, K., Fuchs, C. and Theis, F. J. (2012). *Blind source separation using latent Gaussian graphical models*. Ninth International Workshop on Computational Systems Biology, WCSB 2012, 34–46.
- [5] Kowarsch, A., Blöchl, F., Bohl, S., Saile, M., Gretz, N., Klingmüller, U., and Theis, F.J. (2010). *Knowledge-based matrix factorization temporally resolves the cellular responses to IL-6 stimulation*. BMC Bioinformatics, **11**, 585–598.
- [6] Murphy, K., et al. (2001). *The Bayes net toolbox for Matlab*. Computing Science and Statistics, **33**(2), 1024–1034.
- [7] Smyth, G. K. (2005). *Limma: linear models for microarray data*. In: Bioinformatics and Computational Biology Solutions using R and Bioconductor, Springer, 397–420.

Maximum simulated likelihood estimation of Thurstonian models

Manuela Cattelan, *University of Padova*, manuela.cattelan@unipd.it

Abstract. Thurstonian models are a class of models widely employed in psychometrics for the analysis of preference data. These models assume that when some items are presented to a subject, each of them elicits a continuous preference and the item with larger preference at the moment of the comparison is the preferred one. Moreover, Thurstonian models assume that the unobserved preferences are normally distributed in the population, and the main goal of the analysis is the estimation of the mean and the covariance matrix of the stimuli produced by the items compared. Such an estimation is awkward since it implies the computation of high dimensional multivariate normal integrals. To overcome this difficulty, in the psychometric literature a limited information estimation method, that uses only marginal univariate and bivariate probabilities, was proposed. We show that Thurstonian models for preference data can be estimated using maximum simulated likelihood via the Geweke-Hajivassiliou-Keane algorithm. An important advantage of this method is that the value of the likelihood function is available, hence it can be used for other inferential purposes as hypothesis testing and model selection.

Keywords. GHK algorithm, maximum simulated likelihood, Thurstonian models

1 Introduction

The models introduced by Thurstone [9] are widely employed in the analysis of choice behavior, with the aim of investigating preferences, attitudes and values of people. The items, which can be physical objects, statements, values, etc., are compared in couples, and Thurstonian models assume that each item presented to a judge elicits a continuous preference. The item that has a larger preference at the moment of the comparison is the preferred one. Moreover, these models assume that the unobserved preferences are normally distributed in the population. Since the same item is involved in many paired comparisons, a complex structure of cross dependence is originated. The difficulties associated with maximum likelihood estimation of Thurstonian models fostered the proposal of alternative estimation methods based on low order marginal distributions as limited information estimation [5, 6] and pairwise likelihood [1]. Here, we propose the use of maximum simulated likelihood, in which the value of the likelihood function is simulated via the Geweke-Hajivassiliou-Keane [10] algorithm. An important benefit of this

method is that a (simulated) value of the likelihood function is available, and this can be used for other inferential procedures.

In Section 2 Thurstonian models for paired comparison data are introduced and in Section 3 maximum simulated likelihood estimation through the Geweke-Hajivassiliou-Keane algorithm is illustrated. Section 4 shows the results of two applications while Section 5 concludes.

2 Thurstonian models

Let Y_{ijs} , $i = 1, \dots, n-1$, $j = i+1, \dots, n$ and $s = 1, \dots, S$ denote the result of the comparison between items i and j performed by subject s . $Y_{ijs} = 1$ if subject s prefers item i , while $Y_{ijs} = 0$ if item j is preferred. If n objects are considered, the number of possible paired comparisons is $N = \binom{n}{2}$. It is assumed that every subject performs all the N paired comparisons.

Thurstonian models assume that the preferences for the items compared, $\mathbf{T} = (T_1, \dots, T_n)$, follow a multivariate normal distribution with mean $\boldsymbol{\mu} = (\mu_1, \dots, \mu_n)$ and covariance matrix $\boldsymbol{\Sigma}_T$. Let t_i be a realization of T_i , then in each paired comparison item i is preferred to j if $t_i > t_j$, which is equivalent to $z_{ij} = t_i - t_j > 0$, where z_{ij} is a realization of $Z_{ij} = T_i - T_j$. This specification assigns probability zero to inconsistent choices, which occur for example when item i is preferred to item j , item j is preferred to item k , but item k is preferred to item i . Since inconsistencies, also called circular triads, are not uncommon in paired comparisons, Takane [8] proposes to add a vector of pair specific errors that give non-zero probabilities to circular triads. Let $\mathbf{Z}_s = (Z_{12s}, \dots, Z_{n-1ns})$ denote the vector of all latent variables pertaining to subject s , then

$$\mathbf{Z}_s = \mathbf{A}\mathbf{T} + \boldsymbol{\epsilon},$$

where $\boldsymbol{\epsilon} = (\epsilon_{12s}, \dots, \epsilon_{n-1ns})$ denotes the vector of pair specific normally distributed errors with mean $\mathbf{0}$ and covariance matrix $\boldsymbol{\Omega}$, and \mathbf{A} denotes the $N \times n$ design matrix, in which each row denotes a paired comparison and each column represents an item. For example, if $n = 4$, then

$$\mathbf{A} = \begin{pmatrix} 1 & -1 & 0 & 0 \\ 1 & 0 & -1 & 0 \\ 1 & 0 & 0 & -1 \\ 0 & 1 & -1 & 0 \\ 0 & 1 & 0 & -1 \\ 0 & 0 & 1 & -1 \end{pmatrix}.$$

Hence, the vector of latent variables related to subject s follows a multivariate normal distribution $\mathbf{Z}_s \sim MVN(\mathbf{A}\boldsymbol{\mu}; \mathbf{A}\boldsymbol{\Sigma}_T\mathbf{A}' + \boldsymbol{\Omega})$. Thurstone [9] proposes various models with different structures of the correlation matrix $\boldsymbol{\Sigma}_T$. In the unstructured model the elements of the covariance matrix are not restricted, while Case III assumes that $\boldsymbol{\Sigma}_T$ is a diagonal matrix, and Case V assumes $\boldsymbol{\Sigma}_T = \sigma^2\mathbf{I}_n$, where \mathbf{I}_n denotes the identity matrix of dimension n . In the applications illustrated in Section 4, an unstructured covariance matrix is employed and it is assumed that the pair specific errors have equal variance, hence $\boldsymbol{\Omega} = \omega\mathbf{I}_N$.

Let $\mathbf{Z}_s^* = \mathbf{D}(\mathbf{Z}_s - \mathbf{A}\boldsymbol{\mu})$ be the standardised version of the latent variable \mathbf{Z}_s , where $\mathbf{D} = [\text{diag}(\boldsymbol{\Sigma}_Z)]^{-1/2}$ and $\boldsymbol{\Sigma}_Z = \mathbf{A}\boldsymbol{\Sigma}_T\mathbf{A}' + \boldsymbol{\Omega}$ denotes the covariance matrix of \mathbf{Z}_s . Then, \mathbf{Z}_s^* follows a multivariate normal distribution with mean $\mathbf{0}$ and correlation matrix $\boldsymbol{\Sigma}_{Z^*} = \mathbf{D}\boldsymbol{\Sigma}_Z\mathbf{D}$. Object i is preferred to object j when $z_{ijs}^* \geq \tau_{ij}^*$, where the vector of the thresholds is given by $\boldsymbol{\tau}^* = -\mathbf{D}\mathbf{A}\boldsymbol{\mu}$.

Then, the probability of the paired comparisons observed for subject s is

$$L_s(\boldsymbol{\theta}; \mathbf{Y}_s) = \int_{R_{12s}} \cdots \int_{R_{n-1ns}} \phi_N(\mathbf{z}_s^*; \boldsymbol{\Sigma}_{Z^*}) d\mathbf{z}_s^*,$$

where $\mathbf{Y}_s = (Y_{12s}, \dots, Y_{n-1ns})$, $\boldsymbol{\theta}$ denotes the model parameters, $\phi_N(\cdot; \boldsymbol{\Sigma}_{Z^*})$ denotes the density function of an N -dimensional normal random variable with mean $\mathbf{0}$ and correlation matrix $\boldsymbol{\Sigma}_{Z^*}$ and

$$R_{ijs} = \begin{cases} (-\infty, \tau_{ij}^*) & \text{if } Y_{ijs} = 0 \\ (\tau_{ij}^*, \infty) & \text{if } Y_{ijs} = 1. \end{cases}$$

The likelihood function is the product of the probabilities of the observations for all subjects

$$L(\boldsymbol{\theta}; \mathbf{Y}) = \prod_{s=1}^S L_s(\boldsymbol{\theta}; \mathbf{Y}_s).$$

Note that this approach requires the approximation of S integrals whose dimension is equal to $N = n(n-1)/2$, the number of paired comparisons, so its growth is quadratic with the increase in the number of items.

Identification is an important issue in Thurstonian models [5, 12, 11, 7], indeed the model requires some restrictions since only the thresholds $\boldsymbol{\tau}$ and the tetrachoric correlations, which are the elements of the matrix $\boldsymbol{\Sigma}_{Z^*}$, can be identified. It can be shown [11] that $n+2$ constraints are needed in order to identify the unstructured model. A first restriction is necessary to identify the mean parameters; it is possible either to set one of them to zero, *i.e.* $\mu_n = 0$, or to set $\sum_{i=1}^n \mu_i = 0$. Many different identification restrictions can be used for the covariance matrix, an example is to set all the diagonal elements of $\boldsymbol{\Sigma}_T$ to 1 and one off-diagonal element to 0. However, the estimated parameters identify a class of covariance matrices since $\boldsymbol{\Sigma}_T$ and $\boldsymbol{\Sigma}_T + \mathbf{d}\mathbf{1}' + \mathbf{1}\mathbf{d}'$, where $\mathbf{1}$ is a vector of n ones and \mathbf{d} is a vector of n constants, are not distinguishable, see [5, 1].

3 Estimation

Maximum likelihood estimation of Thurstonian models requires the approximation of multivariate normal integrals which can also be of large dimension. Such an approximation is quite cumbersome, hence alternative estimating methods have been proposed. In the psychometric literature, [5] suggests a limited information estimation method which uses only univariate and bivariate marginals. This method is performed in three steps. In the first step the threshold parameters $\boldsymbol{\tau}$ are estimated from the univariate sample proportions, while in the second step, given the estimated thresholds, the tetrachoric correlations are estimated using the bivariate proportions. Finally, in the last step, the model parameters are estimated through minimisation of the Mahalanobis distance between the estimated thresholds and tetrachoric correlations and the corresponding model based quantities.

Another method based on low dimensional marginal distributions is proposed in [1], that suggests a composite likelihood [14] approach. In particular, Thurstonian models are estimated using the pairwise likelihood, which employs only marginal bivariate probabilities. The likelihood function is replaced with a pseudo-likelihood function which is the product of all marginal bivariate probabilities. Maximum pairwise likelihood estimation seems to perform well in a number of different Thurstonian models, see [1]. This method reduces noticeably the computational burden since only bivariate normal probabilities are computed.

Thurstonian models are a particular instance of the multivariate probit model, hence inferential methods developed for the probit model can be adapted to Thurstonian models. Multivariate normal probabilities can be simulated via the Geweke-Hajivassiliou-Keane (GHK) algorithm [10]. This algorithm approximates the joint distribution of all the outcomes by sequential simulation from univariate truncated normal distributions.

The implementation of the algorithm requires the assumption of an order for the preferences expressed by a person. We choose to arrange the comparisons made by a person in lexicographic order of the first and then the second item. The GHK algorithm is a sequential importance sampling algorithm based on drawing from the conditional density $p(z_{ijs}|y_{ijs}, \zeta_{ijs}; \theta)$, where ζ_{ijs} is the vector of latent variables Z_{kls} preceding Z_{ijs} in the chosen order. In other words, the GHK algorithm employs as importance density the normal density $p(z_{ijs}|\zeta_{ijs}; \theta)$ truncated over the interval $(\delta_{y_{ijs}}, \delta_{y_{ijs+1}}]$, where $\delta_0 = -\infty$, $\delta_1 = 0$ and $\delta_2 = \infty$.

Let m_{ijs} and v_{ijs} be the mean and the standard deviation, respectively, of the conditional density $p(z_{ijs}|\zeta_{ijs}; \theta)$. Then, a draw from the importance density $p(z_{ijs}|y_{ijs}, \zeta_{ijs}; \theta)$ is obtained by setting

$$z_{ijs}(u_{ijs}) = m_{ijs} + v_{ijs}\Phi^{-1}\{(1 - u_{ijs})l_{ijs} + u_{ijs}r_{ijs}\},$$

where u_{ijs} is a draw from a random variable uniformly distributed in the unit interval, and quantities l_{ijs} and r_{ijs} are defined as

$$l_{ijs} = \Phi\left(\frac{-\delta_{y_{ijs}} - m_{ijs}}{v_{ijs}}\right) \text{ and } r_{ijs} = \Phi\left(\frac{-\delta_{y_{ijs+1}} - m_{ijs}}{v_{ijs}}\right).$$

Denote by $z_{ijs}^{(b)}$ the b th draw from the above importance density ($b = 1, \dots, B$). The GHK algorithm approximates the likelihood of the proposed paired comparison model by the Monte Carlo sum

$$L_{\text{GHK}}(\theta; \mathbf{y}) = \frac{1}{B} \sum_{b=1}^B \left\{ \prod_{s=1}^S \prod_{\text{ord}(i,j)} \frac{p(z_{ijs}^{(b)}|\zeta_{ijs}^{(b)}; \theta)}{p(z_{ijs}^{(b)}|y_{ijs}, \zeta_{ijs}^{(b)}; \theta)} \right\},$$

where the product follows the predetermined comparisons order indicated by $\text{ord}(i, j)$. The GHK algorithm is popular in the econometric literature for approximate inference in multivariate probit models; more details and references can be found in [10].

In the applications illustrated in Section 4, the GHK algorithm implemented in the R package `gcmr` [4] is employed. The GHK algorithm provides an unbiased estimate of the likelihood function, but it is biased on the scale of the log-likelihood. Since it is more convenient to maximize the log-likelihood, it is opportune to correct the bias of its importance sampling approximation. For this purpose, in the `gcmr` package the correction suggested in [2] is implemented.

The performances of maximum simulated likelihood and limited information estimation are compared in a simulation study using the same setting employed in [13]. There are six paired comparisons involving four items, and every comparison is performed by 100 judges. The model is parametrised in terms of differences with respect to a reference item, hence means and variances of the differences $\tilde{T}_i = T_i - T_n$, $i = 1, \dots, n - 1$, are estimated. The assumed parameters are $\tilde{\mu} = (-0.2, 1, -1.5)$ while the covariance matrix is

$$\begin{pmatrix} 1.5 & 1 & 1.3 \\ 1 & 4 & 2.5 \\ 1.3 & 2.5 & 3 \end{pmatrix},$$

	True	LI			GHK		
	value	Mean	Median	Std. dev.	Mean	Median	Std. dev.
$\tilde{\mu}_1$	-0.2	-0.226	-0.214	0.205	-0.216	-0.207	0.190
$\tilde{\mu}_2$	1	1.068	1.007	0.417	1.018	0.995	0.326
$\tilde{\mu}_3$	-1.5	-1.594	-1.531	0.489	-1.540	-1.507	0.346
$\tilde{\sigma}_1^2$	1.5	2.058	1.579	1.967	1.690	1.492	1.040
$\tilde{\sigma}_2^2$	4	5.342	4.372	4.424	4.333	3.909	2.252
$\tilde{\sigma}_3^2$	3	3.913	3.190	3.170	3.282	2.914	1.785
$\tilde{\sigma}_{12}$	1	1.340	1.059	1.443	1.094	0.952	0.823
$\tilde{\sigma}_{13}$	1.3	1.716	1.387	1.476	1.414	1.246	0.906
$\tilde{\sigma}_{23}$	2.5	3.351	2.724	2.673	2.697	2.457	1.407
b	0.5	0.720	0.577	0.820	0.559	0.520	0.448

Table 1: Averages (**Mean**), medians (**Median**) and simulation standard deviations (**Std. dev.**) of parameters estimated by limited information estimation (**LI**) and maximum simulated likelihood via the GHK algorithm (**GHK**).

and $\tilde{\sigma}_{ij}$ is used to denote the element in row i and column j of the above reduced matrix. The covariance matrix used in [13] depends on a further parameter b whose value is set equal to 0.5. Table 1 shows means, medians and simulation standard deviations of limited information estimation and maximum simulated likelihood estimates employing the GHK algorithm. The maximum simulated likelihood estimates show a lower estimation bias and appear much more precise than the limited information estimates.

4 Applications

The first application concerns preferences for compact cars. The data are analyzed in [6] and they consist of paired comparisons among compact cars performed by Spanish college students in order to investigate their purchasing preferences. The cars considered in the study are Citroën AX, Fiat Punto, Nissan Micra, Opel Corsa, Peugeot 106, Seat Ibiza and Volkswagen Polo. All the paired comparisons were performed by 294 students, hence the computation of the likelihood function requires the approximation of 294 multivariate normal integrals of dimension 21. In [6] a limited information estimation is employed. The maximum simulated likelihood estimates, with standard errors in parentheses, of the mean and unstructured covariance matrix parameters are reported in Table 2. The identification restrictions used are those employed in [6], specifically the mean of the preference for Volkswagen Polo is set to 0, the variances of Citroën AX and Volkswagen Polo are set to 1 and all the covariances with Volkswagen Polo are set to 0. Mean estimates have to be interpreted relative to Volkswagen Polo, the reference car. Only Seat Ibiza has a mean significantly higher than Volkswagen Polo, while the mean for Peugeot 106 is not significantly different from that of Volkswagen Polo. The least preferred car is Citroën AX. The estimated common variance of pair specific errors is $\hat{\omega} = 0.31$, with standard error 0.05. Given the identification restrictions, significant positive covariances in Table 2 indicate that the association between the utility of the two cars is stronger than the association between Citroën AX and Volkswagen Polo. On the contrary, negative covariances denote an association which is weaker than that between Citroën AX and Volkswagen Polo.

	AX	Punto	Micra	Corsa	106	Ibiza	Polo	$\hat{\mu}$
Citroën AX	1 (fixed)							-1.41 (0.12)
Fiat Punto	-0.04 (0.15)	0.98 (0.31)						-0.53 (0.10)
Nissan Micra	0.31 (0.18)	0.10 (0.23)	1.90 (0.45)					-0.74 (0.12)
Opel Corsa	0.23 (0.12)	-0.06 (0.17)	0.05 (0.19)	0.67 (0.23)				-0.27 (0.09)
Peugeot 106	0.31 (0.15)	0.00 (0.20)	0.26 (0.24)	0.16 (0.19)	1.21 (0.36)			-0.10 (0.10)
Seat Ibiza	-0.34 (0.18)	0.21 (0.24)	-0.39 (0.22)	-0.06 (0.20)	0.08 (0.24)	1.44 (0.34)		0.29 (0.11)
Volkswagen Polo	0 (fixed)	0 (fixed)	0 (fixed)	0 (fixed)	0 (fixed)	0 (fixed)	1 (fixed)	0 (fixed)

Table 2: Maximum simulated likelihood estimates and standard errors of an unstructured Thurstonian model for paired comparisons among compact cars.

The second application considered is a paired comparison study that investigates which of five training delivery modes trainees prefer. The modes were computer-based (CO), TV-based (TV), paper-based (PA), audio-based (AU) and classroom-based (CL) training. The 198 study participants were unemployed persons in the labour market training of the Austrian labour market service. The data set is available in the R package `prefmod` [3]. We used the same identification restrictions as those employed in the previous application. Estimates and standard errors are shown in Table 3. Classroom-based training is the reference item; since all the other means are negative, and significantly different from zero, classroom-based training is the preferred training method, followed by computer-based, paper-based, TV-based and audio-based training. Again, the significant positive association between computer-based and TV-based training means that these two training methods have an association stronger than the association between computer-based and classroom-based training. The estimate of the variance of the pair specific errors is $\hat{\omega} = 0.38$, with standard error 0.08.

5 Conclusions

Thurstonian models are widely employed in the analysis of choice behavior. However, these models are quite difficult to estimate since maximum likelihood estimation requires the approximation of high dimensional integrals, even for a small number of items compared. To overcome this difficulty many authors have proposed alternative estimation methods that employ only low dimensional marginal distributions, as pairwise likelihood or limited information estimation. Here, we propose the use of maximum simulated likelihood, in which the value of the likelihood function is simulated via the GHK algorithm for multivariate normal probabilities. This method is computationally more expensive than estimating methods based on low dimensional distributions, but it has the advantage of computing a likelihood function, so standard methods for

	CO	TV	PA	AU	CL	$\hat{\mu}$
CO	1 (fixed)					-0.27 (0.12)
TV	0.59 (0.18)	1.66 (0.42)				-1.10 (0.16)
PA	-0.34 (0.16)	-0.37 (0.19)	0.00 (0.22)			-0.51 (0.10)
AU	0.72 (0.17)	1.11 (0.30)	-0.34 (0.19)	1.55 (0.37)		-1.44 (0.17)
CL	0 (fixed)	0 (fixed)	0 (fixed)	0 (fixed)	1 (fixed)	0 (fixed)

Table 3: Maximum simulated likelihood estimates and standard errors of an unstructured Thurstonian model for paired comparisons among training delivery modes.

hypothesis testing and model selection based on the likelihood can be employed. Moreover, this method can easily be extended to the case of ordinal data, when ties between items are allowed, or when the results of the paired comparisons can be, for example, strong preference for an item, mild preference for an item or indifference. The extension of maximum simulated likelihood estimation to instances in which subjects perform only a subset of the N paired comparisons is straightforward.

Acknowledgement

The Author acknowledges the financial support of the Università degli Studi di Padova for the research project “Inferential issues in regression models for dependent categorical and discrete data” carried out at the Department of Statistical Sciences, University of Padova.

Bibliography

- [1] Cattelan, M. (2012) *Models for paired comparison data: a review with emphasis on dependent data*. *Statistical Science*, **27**, 412–433.
- [2] Durbin, J. and Koopman, S.J. (1997) *Monte Carlo maximum likelihood estimation for non-Gaussian state space models*. *Biometrika*, **84**, 669–684.
- [3] Hatzinger, R. and Dittrich, R. (2012) *prefmod: an R package for modeling preferences based on paired comparisons, rankings, or ratings*. *Journal of Statistical Software*, **48**, 1–31.
- [4] Masarotto, G. and Varin, C. (2012) *Gaussian copula marginal regression*. *Electronic Journal of Statistics*, **6**, 1517–1549.
- [5] Maydeu-Olivares, A. (2001) *Limited information estimation and testing of Thurstonian models for paired comparison data under multiple judgement sampling*. *Psychometrika*, **66**, 209–228.

- [6] Maydeu-Olivares, A. and Böckenholt, U. (2005) *Structural equation modeling of paired-comparison and ranking data*. Psychological Methods, **10**, 285–304.
- [7] Maydeu-Olivares, A. and Hernández, A. (2007) *Identification and small sample estimation of Thurstone's unrestricted model for paired comparisons data*. Multivariate Behavioral Research, **42**, 323–347.
- [8] Takane, Y. (1989) *Analysis of covariance structures and probabilistic binary choice data*. In New Developments in Psychological Choice Modeling (G. De Soete, H. Feger and K. C. Klauer, eds.). North-Holland, Amsterdam.
- [9] Thurstone, L. L. (1927) *A law of comparative judgement*. Psychological Review, **79**, 281–299.
- [10] Train, K. E. (2009) *Discrete choice methods with simulation*. Cambridge University Press, New York.
- [11] Tsai, R.-C. (2003) *Remarks on the identifiability of Thurstonian ranking models: Case V, Case III, or neither?* Psychometrika, **68**, 361–372.
- [12] Tsai, R.-C. and Böckenholt, U. (2002) *Two-level linear paired comparison models: estimation and identifiability issues*. Mathematical Social Sciences, **43**, 429–449.
- [13] Tsai, R.-C. and Böckenholt, U. (2008) *On the importance of distinguishing between within- and between-subject effects in intransitive intertemporal choice*. Journal of Mathematical Psychology, **52**, 10–20.
- [14] Varin, C., Reid, N. and Firth, D. (2011) *An overview of composite likelihood methods*. Statistica Sinica, **21**, 5–42.

Robust profiling of Site Index

Manuela Souto de Miranda, *University of Aveiro*, manuela.souto@ua.pt

Conceição Amado, *University of Lisbon*, conceicao.amado@tecnico.ulisboa.pt

Margarida Silva, *Raiz - Portucel Soporcel Group*, margarida.silva@portucelsoporcel.com

Abstract. The main objective of the present study is to investigate how robust multivariate methods can contribute to characterizing relevant environmental conditions for the Site Index of the *Eucalyptus globulus*. Site Index is an important indicator of forest productivity and it is affected by environmental properties of each geographical location. In order to identify which environmental variables are more relevant, a robust principal components analysis was conducted. The use of the robust approach, when compared to the conventional one, resulted in a more realistic structure of variability and showed some advantages. Moreover, collected data suggested a grouping process. A cluster analysis was also accomplished considering both conventional and robust procedures. Some practical difficulties arose with robust clustering methods; however they resulted in some benefits, particularly in robust outliers detection.

Keywords. Cluster analysis, *Eucalyptus globulus*, outliers detection, principal components, robustness, Site Index.

1 Introduction

Site Index (SI) is an indicator used in forestry for the evaluation of the potential productivity at a particular location or site. It is determined by the average height of the dominant trees on the site at a given age and it expresses the quality of the tree stand (for instance, see [13]). The age of the trees depends on the biological species. For the *Eucalyptus globulus* herein considered it is ten years. Generally, SI is adopted for measuring the growth of the trees in the site, thus being an important tool in forestry management for determining the economic value of a forest or the interest in future stands (see [13, 16]).

SI depends on climate and environmental conditions, while it is barely affected by silviculture conditions as stated in [16]. Thus, before taking management decisions, it is of major importance to find the best set of available measures for characterizing the environment conditions in the locations.

Several studies (for instance, [9] or [5]) concerning the empirical distribution and the prediction of SI for other species of *Eucalyptus* are published in the literature, and also for different geographical regions as Australia, Bolivia or Brasil (see [15], [10] or [12]). However, only a few

papers (see [1]) are concerned with the *SI* of the *Eucalyptus globulus* in the European Union, particularly in the Iberian countries, where there exists a large area devoted to commercial forests. The present study aims to contribute to the knowledge of the main environmental variables that affect the *SI*'s distribution of the *Eucalyptus globulus* in the Portuguese forest, which represent about 26% of the Portuguese forest (according with the official report [14]). Other authors used a traditional Principal Components Analysis (*PCA*) with similar purposes, but those methods are not robust in sense that they are very sensitive to gross error observations, as well as to extreme values and other deviations from the models. In the last decades, robust statistical methods appeared as alternatives to the conventional statistic models and techniques. In general, robust procedures lead to better results, but they also have the counterparts of more computational complexity and less simplicity of interpretation. Thus they are not widely spread among application fields like forestry, in particular. Besides, many robust procedures are current fields of research and some future improvements will be welcome for simplicity of application.

In the following work we used conventional and robust multivariate methods. In the next section we present the data and we summarize the methodology. The set of original variables contained several correlated variables suggesting that a *PCA* should be used for reducing dimensionality. Also a preliminary data analysis pointed out for the existence of two groups of seats, associated with the available water and humidity conditions. A cluster analysis was consequently performed which confirmed the suspicious. When considering robust clustering it was used the proposal of [8] which is based on a trimming process.

The remaining sections of this paper are devoted to the description of the data set, followed by a brief review of the methods, a discussion of the results and some final comments.

2 Data description

The analyzed data was sampled in commercial pure *Eucalyptus globulus* stands in continental Portugal, between 2000 and 2010 and according with international rules on the subject. The data set consists of 3022 observations of *SI* and of 14 environmental variables, comprising climate indicators and soil characteristics. The choice of the variables took into account their contribution for water availability to plants and their easy real access in collecting data, thus ensuring operability of future modeling. Table 1 describes the variables whose measurements were available, their abbreviation and their units system.

Some variables were initially transformed as recommended in specific literature. Namely, a logarithmic transform was applied to the variable *available water storage capacity*, and Box-Cox transforms were applied to the variables *altitude*, *annual average precipitation* and *annual average evapotranspiration* (see [9] or [5]). The multivariate analysis proceed with the transformed variables in place of the corresponding originals; they will be abbreviated, respectively, by *lnAusc*, *tAlt*, *tPrec*, *tEvap*.

An exploratory data analysis highlighted a set of potential outliers in the variable *SI* and, as expected, high correlations between some other variables. It is also important to notice that the empirical distribution of some variables like the precipitation along summer months, strongly suggested grouping data in two (or more) groups.

Variable	Description	Units
SI	Site Index	m
x	latitude	° (degrees)
y	longitude	° (degrees)
alt	altitude	m
dcl	local slope	%
exp	exposition	° (degrees)
Prof	depth site	cm
Pedreg	stoniness	%
prec	annual average precipitation	mm
dprec	number of days with precipitation greater than 1 mm	
prec678	average precipitation along summer months	mm
evap	annual average real evapotranspiration	mm
tmin	annual average minimal temperature	C°
tmax	annual average maximal temperature	C°
awsc	available water storage capacity	mm

Table 1: Description of the variables with available measurements.

3 Methodology review

In this section we briefly review the supporting methods. *PCA* is probably the most spread multivariate technique. Its main goal is to reveal relationships between observed p variables and find the best linear transform of those variables such that the new k variables are uncorrelated and can explain the most variability in data, reducing dimensionality if possible ($k \leq p$). The new variables are called the principal components (*PC*) and their mathematical derivation assures that they are ordered by decreasing importance in the explanation of the variability. Formally, if \mathbf{X} denotes an $n \times p$ data matrix, with (non singular) covariance matrix Σ with eigenvalues λ_j and eigenvectors \mathbf{e}_j ($j = 1, \dots, p$), each PC_j has the form

$$PC_j = \mathbf{e}_j^T \mathbf{X}, \quad (1)$$

where \mathbf{e}_j such that $\|\mathbf{e}_j\| = 1$ and $cov(PC_i, PC_j) = 0$ for every $i \neq j$, with $i, j = 1, \dots, p$. Among their properties it is convenient to remember that the variance of each PC_j equals the eigenvalue λ_j , also that summing those variances over j one obtains the trace of Σ and finally that the *PCs* are not invariant. Last property is especially important for explaining the robust counterpart. In fact, when data is standardized in a conventional way, the *PCs* are obtained by

$$PC_j^* = \mathbf{e}_j^T (\Sigma_{diag}^{1/2})^{-1} (\mathbf{X} - \boldsymbol{\mu}), \quad (2)$$

where $\boldsymbol{\mu}$ denotes the vector mean of the original data and the elements of the diagonal matrix $\Sigma_{diag}^{1/2}$ are the square roots of λ_j . The role of $\boldsymbol{\mu}$ and $\Sigma_{diag}^{1/2}$ is to correct the effects of location and scale introduced by the standardization.

The lack of robustness of the location-scale traditional pair (mean vector, covariance matrix) is well-known. The main idea for robustifying the *PCA* was to consider alternative robust pairs of location-scale measures. Nowadays there exist many robust statistics that can replace that

choice, like using a robust M-estimator for estimating the center of location together with a robust covariance matrix estimator as, for instance, improved forms of the *MCD* (*Minimum Covariance Determinant*) estimator. Other methods for robustifying *PCA* include those based on projection pursuit algorithms (e.g. [2]) or those based on convex optimization for recovering low-dimensional structure (e.g. [17]). In the statistical literature the main choices are summarized in [4], a paper that gives an interesting review of robust regression and robust *PCA* methods, which probably will contribute for a faster dissemination of statistical robustness.

As mentioned before, the study also included a clustering analysis. One of the most popular methods for grouping data is the so-called *k*-means, a non-hierarchical method for partition of data. Assuming *a priori* the existence of *k* groups, the method looks for the location center of each group \mathbf{m}_j and assigns each sample point $\mathbf{x}_i \in R^p$ to the nearest center, thus including the observation in that group. The minimization process that seeks the center points is accomplished with the Euclidean norm, as the solution of

$$\arg \min_{m_1, \dots, m_k} \sum_{i=1}^n \min_{j=1, \dots, k} \|\mathbf{x}_i - \mathbf{m}_j\|^2 . \quad (3)$$

The method is not robust, primarily due to the least squares criterion. Moreover, the choice of the number *k* of clusters is often a difficult decision that can be worsened when data contains outliers.

Looking for robust alternatives, one can use the trimming algorithm proposed by [8] and later improved as in [6]. This method is based on the original proposal [3] that combines conventional *k*-means with a trimming procedure, which was later improved by considering robust Mahalanobis distances (with an efficient robust scale estimator). It is a very attractive robust clustering method since it has good statistical properties and its computational counterpart is well systematized. Nevertheless, for practical purposes the investigator user can be faced with important troubles inherent in the method: besides the (almost) always subjective choices of the number of clusters *k* and the proportion α of trimming, it is also necessary to take decisions in advance about specific constrains related with shapes and structure of the covariance matrices of the clusters and relations among its eigenvalues. Actually, specific tools are implemented for helping in the choice of the pair (α, k) but the remainder decisions constitute a serious difficulty in many practical situations.

4 Results

As already mentioned, having in mind a fast evaluation of the conditions that might affect *SI*, as well as eventual future modeling of the distribution of this variable, was a main goal of the present case study to identify the most important environmental variables, so that a location can be characterized in terms of those indicators.

From empirical knowledge and a preliminary data analysis, two main suspicions were drawn: high correlation between some variables and possible need of considering at least two clusters of locations. Actually, correlation analysis was conducted with both the conventional empirical correlation matrix and using robust Fast-MCD estimator method; the correlations were generally low when relating *SI* with any other variables (less than 40%), against several great values in the correlations tables corresponding to some climate variables (greater than 90%, as expected). That scenario of dependence was strengthened by the robust approach.

The study continued with a *PCA* applied to the set of original and transformed variables (except *SI*), aiming the identification of the main variables. For the conventional approach, data was standardized in the usual way by the sample mean and standard deviation, while for robust approach the standardization was achieved with the pair median and median absolute deviation (*MAD*) (recommended, for instance, in [11]).

The two first components together explained more than 70% of the variability (72% and 75%, respectively, by conventional and robust techniques). Interpreting the loadings in the conventional *PCA*, one can point out two factors: the first one might traduce climate conditions related to water and humidity; and the second factor that might be associated with soil characteristics, in particular its capacity of retaining the water. Those conclusions were also valid when using robust approach; moreover, the latter still highlights the significant contribution of two more variables, namely, altitude in PC_1 and depth in PC_2 . Notice that these last two variables were not relevant by the conventional process, but the conclusions are coherent with empirical forestry knowledge and it seems to mean that robust approach describes better the structure of variability in data. Figure 1 and Figure 2 show the biplots with conventional and robust PC s.

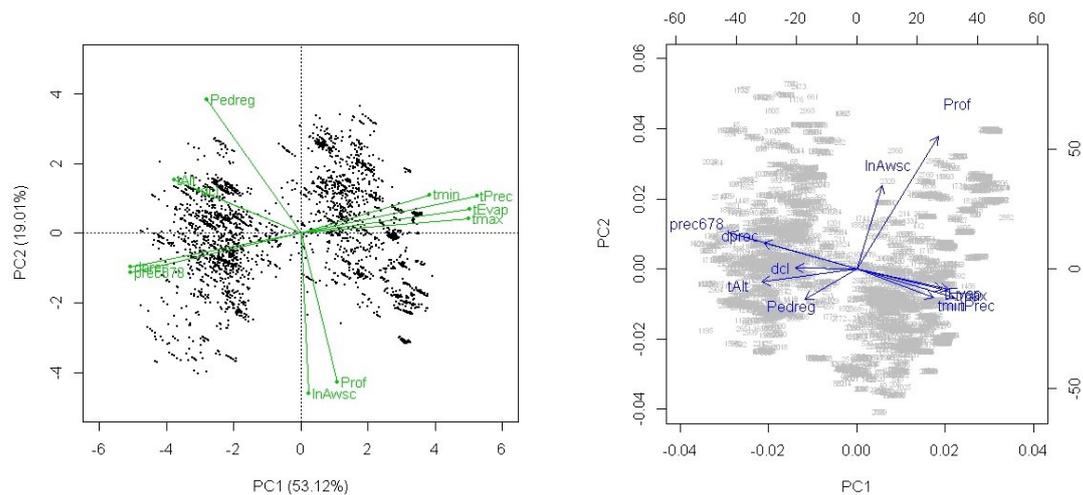


Figure 1: Biplots obtained by conventional *PCA* (on the right) and by robust (on the left) *PCA*.

Another advantage of the robust treatment was in outliers detection. Actually, since robust statistics are less sensitive to anomalous points, potential outliers are much more evidenced. The geographical locations corresponding to the outliers was verified and it was possible to confirm in the field the outlying environmental conditions.

Also notice that both images in Figure 1 seem to show two clusters. A cluster analysis was experimented with two to four groups. There was no clear advantage in considering more than two groups, so that was the option, moreover because the choice of two clusters is well supported by practice. Figure 2 shows the conventional corresponding clusters in the (PC_1, PC_2) plane.

Actually, a map can show that the clusters correspond to geographical distinct areas, namely, north, and center and south of the country, with different climatological conditions.

When considering robust clustering, the option was for trimmed clustering. As mentioned

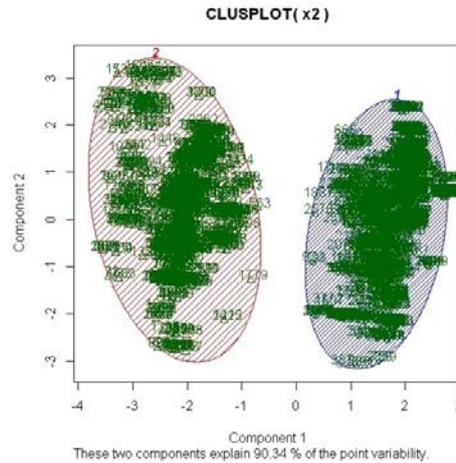


Figure 2: Clusters obtained by conventional k -means.

before, the process has some advantages but also some difficulties.

One of the main advantages from a practical point of view is the existence of a package produced for the method in the R software environment, the `tclust` package (see [7]), since all computations were accomplished with R software. Unfortunately the package is not available when writing the present text.

The problems related with the choice of the number of clusters k and the trimming proportion α are attenuated by an auxiliary tool, which graphically represents the *CTL* (*Classification Trimmed Likelihood*) curves; in the present case there was already the conjecture of 2 clusters, but the trimming proportion it was not pointed out by the CTL curves, for any number of clusters. In such a case the decision can be uncomfortably subjective. In this particular example it was decided to exclude 5% of the observations, since there was 4% of points identified as outliers in the *PCA*.

There are two more crucial practical options for using the method, which emerge from theoretical constrains and which are difficult to interpret for general users, especially because options must be taken before clustering: the question of similar structure of the clusters covariance matrices and shape of the clusters and, finally, the choice of a constant that should traduce a relative measure of clusters dispersion. In what concerns the shape of the cluster the basic configuration was used, which is based on the analysis of the eigenvalues. For deciding the constant value that controls the relative differences among scatter clusters, and since there was not previous nor auxiliary information, the clusters produced by k -means were used in the estimation of the eigenvalues of the robust clusters covariance matrices (in this case it returned the value $restr.fact = 43$).

Finally notice that the package creates an additional auxiliary cluster consisting of the detected outliers, thus presenting a more clear separation of the groups. Figure 4 maps the results of the robust clustering with two clusters. The resulting geographical clustering was very good in dividing the country into two regions almost disjoint. Detection of outliers was also very precise, since outlier points were confirmed and all of them were located in seats with special climate and/or altitude characteristics.

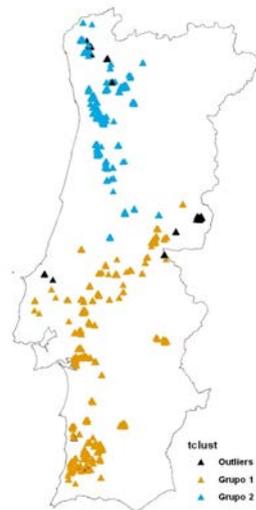


Figure 3: Geographical location of clusters points by the robust approach with outliers evidenced by dark triangles.

5 Concluding remarks

The *PCA* was very useful in profiling environmental properties that affect *SI*, particularly with a robust approach, which seems to describe in a more realistic way the relative importance of some variables. The first two *PC* can explain 75% of the variability, from which 50% is attributed to climate conditions. From those *PC*s it was possible identify two factors with real meaning, namely, water climatic disposability and capacity of soil for retaining the water. Robust clustering confirmed that it is worthwhile to consider two groups of seats associated with the two main regions of the country and shoed to have a good capacity for outlier detection.

From a practical point of view it is still not straightforward to use robust clustering methods and it is desirable to develop auxiliary statistical tools which might give more support to the user.

The main goal of the present study was to identify the most important environmental variables in the characterization of the sites. But there is an evident interest in the future modeling of the *SI*; thus, a robust *PC* regression will be the natural future step. Since present clustering was performed after *PCA*, conclusions herein obtained were prepared for constituting a first stage towards that future modeling.

Bibliography

- [1] Aertsen, W., Kint, V., van Orshoven, J., Ozkan, K. and Bart Muys (2010) *Comparison and ranking of different modelling techniques for prediction of site index in mediterranean mountain forests*. Ecological Modelling, **221**, 1119 - 1130.
- [2] Croux, C., Filzmoser, P. and Oliveira, M. (2007) *Algorithms for Projection-Pursuit robust principal component analysis*. Chemometrics and Intelligent Laboratory Systems, **87 (2)**, 218-225.

- [3] Cuesta-Albertos, J. A., Gordaliza, A. and Matrán, C. (1997) *Trimmed k-means. An attempt to robustify quantizers* Annals of Statistics, **25**, 553-576.
- [4] Filzmoser, P. and Todorov, V. (2011) *Review of robust multivariate statistical methods in high dimension*. Analytica Chimica Acta, **705 (1-2)**, 2-14.
- [5] Fontes, L., Tomé, M., Thompson, F., Yeomans, A., Luís, S. and Savill, P. (2003) *Modelling the douglas-fir (*pseudotsuga menziesii* (mirb.) franco) site index from site factors in Portugal*. Forestry, **76 (5)**, 491-507.
- [6] Fritz, H., García-Escudero, L. A. and Mayo-Isacar, A. (2013) *A fast algorithm for robust constrained clustering*. Computational Statistics and Data Analysis, **61**, 124-136.
- [7] Fritz, H., García-Escudero, L. A. and Mayo-Isacar, A. (2011) *tclust: An R package for a Trimming Approach to Cluster Analysis*. Journal of Statistical Software, **47**, 1-26.
- [8] García-Escudero, L. A., Gordaliza, A., Matrán, C. and Mayo-Isacar, A. (2010) *Exploring the number of groups in robust model-based clustering*. Statistics and Computing, **21**, 585-599.
- [9] Grant, J. C., Nichols, J. D., Smith, R. G. B., Brennan, P. and Vanclay, J. K. (2010) *Site index prediction of Eucalyptus dunnii maiden plantations with soil and site parameters in sub-tropical eastern Australia*. Australian Forestry, **73**, 234-245.
- [10] Guzman, G., Morales, M., Pukkala, T. and de-Miguel, S. (2012) *A growth model for Eucalyptus globulus in Bolivia*. Forest Systems, **21**, 205-209.
- [11] Maronna, R., Martin, R. D. and Yohai, V. (2006) *Robust statistics*. Jhon Wiley & Sons.
- [12] Scolforo, J. R., Maestri, R., Ferraz Filho, A., Mello, J. M., Oliveira, A. D. and Assis, A. L. (2013) *Dominant Height Model for Site Classification of Eucalyptus grandis Incorporating Climatic Variables*. International Journal of Forestry Research, vol. 2013, Article ID 139236, 7 pages, 2013. doi:10.1155/2013/139236.
- [13] Skovsgaard, J. P. and Vanclay, J. K. (2008) *Forest site productivity: a review of the evolution of dendrometric concepts for even-aged stands*. Forestry, **81 (1)**, 13-31.
- [14] Uva, J. S. (2013) *IFN6 Áreas dos usos do solo e das espécies florestais de Portugal continental. Resultados preliminares*. [pdf], 34 pp. Instituto da Conservação da Natureza e das Florestas. Lisboa.
- [15] Wang, Y., LeMay, V. M. and Baker, T. G. (2007) *Modelling and prediction of dominant height and site index of Eucalyptus globulus plantations using a nonlinear mixed-effects model approach*. Canadian Journal of Forest Research, **37 (8)**, 1390-1403.
- [16] West, P. W. (2009) *Tree and Forest Measurement, 2nd Edition*. Springer.
- [17] Wright, J., Peng, Y., Ma, Y., Ganesh, A. and Rao, S. (2009) *Robust Principal Component Analysis: Exact Recovery of Corrupted Low-Rank Matrices by Convex Optimization*. Neural Information Processing Systems, NIPS 2009.

Bayesian density regression for count data

Charalampos Chaniavidis, *University of Glasgow*, c.chaniavidis1@research.gla.ac.uk
Ludger Evers, *University of Glasgow*, ludger.evers@glasgow.ac.uk
Tereza Neocleous, *University of Glasgow*, tereza.neocleous@glasgow.ac.uk

Abstract. Despite the increasing popularity of quantile regression models for continuous responses, models for count data have so far received little attention. The main quantile regression technique for count data involves adding uniform random noise or “jittering”, thus overcoming the problem that the conditional quantile function is not a continuous function of the parameters of interest. Although jittering allows estimating the conditional quantiles, it has the drawback that, for small values of the response variable Y , the added noise can have a large influence on the estimated quantiles. In addition, quantile regression can lead to “crossing” quantiles. We propose a Bayesian Dirichlet process (DP)-based approach to quantile regression for count data. The approach is based on an adaptive DP mixture (DPM) of COM-Poisson regression models and determines the quantiles by estimating the density of the data, thus eliminating all the aforementioned problems. Taking advantage of the exchange algorithm, the proposed MCMC algorithm can be applied to distributions on which the likelihood can only be computed up to a normalising constant.

Keywords. Quantile regression, Bayesian density regression, Bayesian nonparametrics, Dirichlet processes, COM-Poisson distribution

1 Quantile regression

Quantile regression was introduced as a nonparametric method for modelling a variable of interest as a function of covariates [6]. By estimating the conditional quantiles rather than the mean, it gives a more complete description of the conditional distribution of the response variable than least squares regression, and is especially relevant in certain types of applications.

Consider a random variable Y with cumulative distribution function $F(y)$. The p th quantile function of Y is defined as

$$Q(p) = \inf\{y \in \mathbb{R} : p \leq F(y)\} \tag{1}$$

and can be obtained by minimising the expected loss $E[\rho_p(Y - u)]$ with respect to u , where $\rho_p(y) = |y(p - I(y < 0))|$. The p th sample quantile is obtained in a similar way by minimising $\sum_{i=1}^n \rho_p(y_i - u)$.

Suppose that the p th conditional quantile function, $Q_Y(p|X = \mathbf{x})$, is a linear function of the predictors so that $Q_Y(p|X = \mathbf{x}) = X'\beta_p$. The parameter estimates $\hat{\beta}_p$ are then obtained as

$$\hat{\beta}_p = \arg \min_{\beta_p \in \mathbb{R}^k} \sum_{i=1}^n \rho_p(Y - X'\beta_p). \quad (2)$$

A closed-form solution for this minimisation problem does not exist since the objective function is not differentiable at the origin, and it is solved using linear programming techniques [1].

Quantile regression for count data

The problem with applying quantile regression to count data is that the cumulative distribution function of the response variable is not continuous, resulting in quantiles that are not continuous, and which thus can not be expressed as a continuous function of the covariates. One way to overcome this problem is by adding uniform random noise (“jittering”) to the counts [7]. The general idea is to construct a continuous variable whose conditional quantiles have a one-to-one relationship with the conditional quantiles of the counts. Defining the new continuous variable $Z = Y + U$ where Y is the count variable and U is a uniform random variable in the interval $[0, 1)$, the conditional quantiles $Q_Z(p|X = \mathbf{x}) = p + \exp(X'\beta_p)$.

The variable Z is transformed in such a way that the new quantile function is linear in the parameters, i.e. $Q_{T(Z;p)}(p|X = \mathbf{x}) = X'\beta_p$ where

$$T(Z; p) = \begin{cases} \log(Z - p) & \text{for } Z > p, \\ \log(\varsigma) & \text{for } Z \leq p, \end{cases} \quad (3)$$

with ς being a small positive number. The parameters β_p are estimated by running a linear quantile regression of $T(Z; p)$ on x . Finally, the conditional quantiles of interest, $Q_Y(p|X = \mathbf{x})$ can be obtained from the previous quantiles as

$$Q_Y(p|X = \mathbf{x}) = \lceil Q_Z(p|X = \mathbf{x}) - 1 \rceil \quad (4)$$

where $\lceil p \rceil$ denotes the ceiling function which returns the smallest integer greater than, or equal to, p .

While the jittering approach eliminates the problem of a discrete response distribution, for small values of the response variable Y , the mean and the variance in the transformed variable Z will be mainly due to the added noise, resulting in poor estimates of the conditional quantiles $Q_Y(p|X = \mathbf{x})$. As an example, when $Y = 0$ the term $\log(Z - p) = \log(U - p)$ could go from $-\infty$ to 0, simply due to the added noise. In addition, quantile regression can suffer from the problem of crossing quantile curves, which is usually seen in sparse regions of the covariate space. This happens due to the fact that the conditional quantile curve for a given $X = \mathbf{x}$ will not be a monotonically increasing function of p .

Another approach would be to view the counts as ordinal variables with fixed thresholds and then model the new latent variable by an infinite mixture of normal densities [5]. Instead of using the aforementioned methods, we propose an adaptive Dirichlet process mixture approach which estimates the conditional density of the data. The approach is based on an adaptive Dirichlet Process mixture (DPM) of COM-Poisson regression models.

2 COM-Poisson distribution

The COM-Poisson distribution [2, 11] is a two-parameter generalisation of the Poisson distribution that allows for different levels of dispersion. The probability mass function of the COM-Poisson(λ, ν) distribution is

$$P(Y = y|\lambda, \nu) = \frac{\lambda^y}{(y!)^\nu} \frac{1}{Z(\lambda, \nu)} \quad \text{where } Z(\lambda, \nu) = \sum_{j=0}^{\infty} \frac{\lambda^j}{(j!)^\nu} \quad \text{and } y = 0, 1, 2, \dots \quad (5)$$

for $\lambda > 0$ and $\nu \geq 0$, where the normalisation constant does not have a closed form and has to be approximated numerically. The extra parameter ν allows the distribution to model under- ($\nu > 1$) or over-dispersed ($\nu < 1$) data, having the Poisson distribution as a special case ($\nu = 1$).

The above formulation of the COM-Poisson does not have a clear centering parameter since the parameter λ is close to the mean only when ν takes values close to 1, which makes it difficult to interpret for under- or over-dispersed data. Substituting the parameter λ with $\mu = \lambda^{\frac{1}{\nu}}$, where $[\mu]$ is the mode of the distribution

$$\mathbb{E}[Y] \approx \mu, \quad \mathbb{V}[Y] \approx \frac{\mu}{\nu} \quad (6)$$

and the new probability mass function is

$$P(Y = y|\mu, \nu) = \left(\frac{\mu^y}{y!}\right)^\nu \frac{1}{Z(\mu, \nu)} \quad \text{where } Z(\mu, \nu) = \sum_{j=0}^{\infty} \left(\frac{\mu^j}{j!}\right)^\nu \quad \text{and } y = 0, 1, 2, \dots \quad (7)$$

Mixtures of COM-Poisson distributions

The COM-Poisson is flexible enough to approximate distributions with any kind of dispersion in contrast to a Poisson or a mixture of Poisson distributions which can only deal with overdispersion.

The two parameters of the COM-Poisson distribution allow it to have arbitrary (positive) mean and variance; one can obtain a point mass by letting the variance parameter ν tend to infinity. Thus one can show that mixtures of COM-Poisson distributions can provide an arbitrarily precise approximation to any discrete distribution with support \mathbb{N}_0 , which is why COM-Poisson distributions are used by our method. All other generalisations of the Poisson distribution we are aware of do not have this property.

COM-Poisson regression

A regression model can be defined based on (7), in which both the mean and the variance parameter are modelled as a function of covariates:

$$\log \mu_i = \mathbf{x}_i^\top \boldsymbol{\beta} \quad (8)$$

$$\log \nu_i = \mathbf{x}_i^\top \mathbf{c} \quad (9)$$

where Y is the response variable being modelled, and $\boldsymbol{\beta}, \mathbf{c}$ are the regression coefficients for the centering link function and the shape link function respectively. The parameters in this

formulation have a direct link to either the mean or the variance, providing insight into the behaviour of the response variable. Notably,

$$\mathbb{E}[Y_i] \approx \exp(\mathbf{x}'_i \boldsymbol{\beta}), \quad \mathbb{V}[Y] \approx \frac{\exp(\mathbf{x}'_i \boldsymbol{\beta})}{\exp(\mathbf{x}'_i \mathbf{c})} = \exp(\mathbf{x}'_i (\boldsymbol{\beta} - \mathbf{c})). \quad (10)$$

The calculation of the normalisation constant of the COM-Poisson distribution is the computationally most expensive part of the proposed regression model. It can be seen, in the next subsection, that this calculation is redundant.

Exchange algorithm

Any probability density function $p(y|\theta)$ can be written as

$$p(y|\theta) = \frac{q_\theta(y)}{Z(\theta)} \quad (11)$$

where $q_\theta(y)$ is the unnormalised density and the normalising constant $Z(\theta) = \int p(y, \theta) dy$ is unknown. In this case the acceptance ratio of the Metropolis-Hastings algorithm is

$$\alpha = \min \left(1, \frac{q_{\theta^*}(y)\pi(\theta^*)Z(\theta)h(\theta|\theta^*)}{q_\theta(y)\pi(\theta)Z(\theta^*)h(\theta^*|\theta)} \right) \quad (12)$$

where $\pi(\theta)$ is the prior distribution of θ . The acceptance ratio in (12) involves computing unknown normalising constants. Introducing auxiliary variables θ^*, y^* and sampling from an augmented distribution

$$\pi(\theta^*, y^*, \theta|y) \propto p(y|\theta)\pi(\theta)p(y^*|\theta^*)h(\theta^*|\theta) \quad (13)$$

results in

$$\alpha = \min \left(1, \frac{p(y|\theta^*)\pi(\theta^*)p(y^*|\theta)h(\theta|\theta^*)}{p(y|\theta)\pi(\theta)p(y^*|\theta^*)h(\theta^*|\theta)} \right) \quad (14)$$

$$= \min \left(1, \frac{q_\theta(y^*)\pi(\theta^*)h(\theta|\theta^*)q_{\theta^*}(y)Z(\theta)Z(\theta^*)}{q_\theta(y)\pi(\theta)h(\theta^*|\theta)q_{\theta^*}(y^*)Z(\theta^*)Z(\theta)} \right) \quad (15)$$

$$= \min \left(1, \frac{q_\theta(y^*)\pi(\theta^*)q_{\theta^*}(y)}{q_\theta(y)\pi(\theta)q_{\theta^*}(y^*)} \right) \quad (16)$$

where the normalising constants cancel out and $h(\cdot)$ is a symmetric distribution [9, 8]. In order to be able to use this algorithm one has to be able to sample from the unnormalised density which in the case of the COM-Poisson distribution can be done efficiently using rejection sampling.

Updating the parameter μ of the COM-Poisson we have $\theta = (\mu, \nu)$ and $\theta^* = (\mu^*, \nu)$ where μ^* follows a Normal distribution centered at μ and

$$q_\theta(y^*) = \left(\frac{\mu_i^{y_i^*}}{y_i^{*!}} \right)^{\nu_i} \quad q_{\theta^*}(y) = \left(\frac{(\mu_i^*)^{y_i}}{y_i!} \right)^{\nu_i} \quad (17)$$

$$q_\theta(y) = \left(\frac{\mu_i^{y_i}}{y_i!} \right)^{\nu_i} \quad q_{\theta^*}(y^*) = \left(\frac{(\mu_i^*)^{y_i^*}}{y_i^{*!}} \right)^{\nu_i} \quad (18)$$

Likewise for updating the parameter ν .

3 Bayesian density regression

Density regression is similar to quantile regression in that it allows flexible modelling of the response variable Y given the covariates $\mathbf{x} = (x_1, \dots, x_p)'$. Features (mean, quantiles, spread) of the conditional distribution of the response variable, vary with \mathbf{x} , so, depending on the predictor values, features of the conditional distribution can change in a different way than the population mean. The difference between density regression and quantile regression is that density regression models the probability density function or probability mass function rather than directly modelling the quantiles.

Bayesian density regression for count data

This paper focuses on the following mixture of regression models:

$$f(y_i|\mathbf{x}_i) = \int f(y_i|\mathbf{x}_i, \phi_i)G_{\mathbf{x}_i}(d\phi_i) \text{ where } f(y_i|\mathbf{x}_i, \phi_i) = \text{COM-P}(y_i; \exp(\mathbf{x}_i'\mathbf{b}_i), \exp(\mathbf{x}_i'\mathbf{c}_i)) \quad (19)$$

the conditional density of the response variable given the covariates is expressed as a mixture of COM-Poisson regression models with $\phi_i = (\mathbf{b}_i, \mathbf{c}_i)$ and $G_{\mathbf{x}_i}$ is an unknown mixture distribution that changes according to the location of \mathbf{x}_i .

MCMC algorithm

Let $\boldsymbol{\theta} = (\theta_1, \dots, \theta_k)'$ denote the $k \leq n$ distinct values of ϕ and let $\mathbf{S} = (S_1, \dots, S_n)'$ be a vector of indicators denoting the global configuration of subjects to distinct values $\boldsymbol{\theta}$, with $S_i = h$ indexing the location of the i th subject within the $\boldsymbol{\theta}$. In addition, let $\mathbf{C} = (C_1, \dots, C_k)'$ with $C_h = j$ denoting that θ_h is an atom from the basis distribution, $G_{\mathbf{x}_j}^*$. Hence $C_{S_i} = Z_i = j$ denotes that subject i is drawn from the j th basis distribution.

Excluding the i th subject, $\boldsymbol{\theta}^{(i)} = \boldsymbol{\theta} \setminus \{\phi_i\}$ denotes the $k^{(i)}$ distinct values of $\boldsymbol{\phi}^{(i)} = \boldsymbol{\phi} \setminus \{\phi_i\}$, $\mathbf{S}^{(i)}$ denotes the configuration of subjects $\{1, \dots, n\} \setminus \{i\}$ to these values and $\mathbf{C}^{(i)}$ indexes the DP component numbers for the elements of $\boldsymbol{\theta}^{(i)}$.

Grouping the subjects in the same cluster and updating the prior with the likelihood for the data \mathbf{y} , we obtain the conditional posterior

$$(\phi_i|\mathbf{S}^{(i)}, \mathbf{C}^{(i)}, \boldsymbol{\theta}^{(i)}, \mathbf{X}, a) \sim q_{i0}G_{i,0} + \sum_{h=1}^{k^{(i)}} q_{ih}\delta_{\theta_h^{(i)}}, \quad (20)$$

where $G_{i,0}(\phi)$ is the posterior obtained by updating the prior $G_0(\phi)$ and the likelihood $f(y_i|\mathbf{x}_i, \phi)$:

$$G_{i,0}(\phi) = \frac{G_0(\phi)f(y_i|\mathbf{x}_i, \phi)}{h_i(y_i|\mathbf{x}_i)}, \quad (21)$$

$$q_{i0} = cw_{i0}h_i(y_i|\mathbf{x}_i), \quad q_{ih} = cw_{ih}f(y_i|\mathbf{x}_i, \theta_h), \quad (22)$$

$$w_{i0} = \sum_{j=1}^n \frac{ab_{ij}}{a + \sum_{l \neq i} \mathbf{1}(C_{S_l^{(i)}} = j)}, \quad w_{ih} = \frac{b_{i,C_h^{(i)}} \sum_{m \neq i} \mathbf{1}(S_m^{(i)} = h)}{a + \sum_{l \neq i} \mathbf{1}(C_{S_l^{(i)}} = C_h)}, \quad h = 1, \dots, k^{(i)} \quad (23)$$

where b_{ij} are weights that depend on the distance between subjects' predictor values and c is a normalising constant. Since there is no closed form expression for the posterior distribution, approximation of the probability $q_{i0} = cw_{i0}h_i(y_i|\mathbf{x}_i)$ is difficult.

We overcome this problem by bridging: i) an MCMC algorithm for sampling from the posterior distribution of a Dirichlet process model, with a non-conjugate prior, found in [10]; ii) the MCMC algorithm found in [3]; and iii) a variation of the MCMC exchange algorithm.

Algorithm 3.1.

The MCMC algorithm alternates between the following steps:

Step 1: Update S_i for $i = 1, \dots, n$, by proposing, from the conditional prior, a move to a new cluster or an already existing cluster with probabilities proportional to w_{i0} and w_{ih} for $h = 1, \dots, k^{(i)}$.

a) If the proposed move is to go to a new cluster we draw parameters (μ_0, ν_0) for that cluster from G_0 and at the same time sample an observation y^* from the COM-Poisson(μ_0, ν_0). The acceptance ratio of the Metropolis-Hastings algorithm is

$$\min \left(1, \frac{q_{\theta}(y^*)q_{\theta^*}(y)}{q_{\theta}(y)q_{\theta^*}(y^*)} \right) \quad (24)$$

If the proposal is accepted, $C_{S_i} \sim \text{multinomial}(\{1, \dots, n\}, \mathbf{b}_i)$.

b) If the proposed move is to an already existing cluster h , we sample an observation y^* from the COM-Poisson(μ_h, ν_h) and accept with the same probability as in (24). If the proposal is accepted $C_{S_i} = C_h$.

Step 2: Update the parameters θ_h , for $h = 1, \dots, k$ by sampling from the conditional posterior distribution

$$(\theta_h | \mathbf{S}, \mathbf{C}, \boldsymbol{\theta}^{(h)}, k, \mathbf{y}, \mathbf{X}) \sim \prod_{i:S_i=h} f(y_i | \mathbf{x}_i, \theta_h) G_0(\theta_h), \quad (25)$$

using the Metropolis-Hasting algorithm with acceptance probability as in (16).

Step 3: Update C_h , for $h = 1, \dots, k$, by sampling from the multinomial conditional with

$$(C_h | \mathbf{S}, \mathbf{C}^{(h)}, \boldsymbol{\theta}, k, \mathbf{y}, \mathbf{X}) \sim \frac{\prod_{i:S_i=h} b_{ij}}{\sum_{l=1}^n \prod_{i:S_i=h} b_{il}}, \quad j = 1, \dots, n \quad (26)$$

and location weights γ_j for $j = 1, 2, \dots, n$, using an approach used in [4].

4 Simulations and application

We consider two simulated data sets to compare the proposed discrete Bayesian density regression method to the “jittering” method. These are

$$Y_i | X_i = x_i \sim \text{Binomial}(10, 0.3x_i) \quad (27)$$

$$Y_i | X_i = x_i \sim 0.4\text{Pois}(\exp(1 + x_i)) + 0.2\text{Binomial}(10, 1 - x_i) + 0.4\text{Geom}(0.2) \quad (28)$$

where $x_i \sim \text{Unif}(0, 1)$. Table (1) shows the absolute mean errors obtained using both methods. If q_p is the true conditional quantile when $x = p$ and \hat{q}_p is the estimated conditional quantile, the mean absolute error is defined as $\mathbb{E}[|q_p - \hat{q}_p|]$. The discrete Bayesian density regression (BDR) estimates outperform the “jittering” method and in almost all cases the “jittering” method leads to crossing quantiles (except when $n = 500$).

Method	Number of Observations					
	Binomial			Mixture		
	20	100	500	20	100	500
Density Regression	0.5576	0.2820	0.2421	0.7435	0.5833	0.3589
Jittering (linear)	0.5256	0.8461	0.4765	1.1923	0.6666	0.4294
Jittering (splines)	0.7820	0.5128	0.3020	1.9487	0.8269	0.3910

Table 1: Mean absolute error obtained using the different density/quantile regression methods.

We apply the discrete density regression technique to data on housebreakings in Greater Glasgow (Scotland). The data consist of the number of housebreakings in each of the 127 intermediate geographies in Greater Glasgow in 2010. We aim to relate the number of housebreakings to the deprivation score of the intermediate geography area, as measured by the Scottish Index of Multiple Deprivation (SIMD). The deprivation score is standardised by considering the difference of each intermediate geography’s deprivation from the average deprivation in Greater Glasgow e.g. low values relate to affluent areas, large values to deprived areas. The solid and dashed lines in figure 1 show the quantiles (for $p = 0.1, 0.5, 0.95$) obtained for the standard Poisson regression model and the COM-Poisson model respectively. The first model is not able to capture the overdispersion of the data, nor the skewness of the distribution.

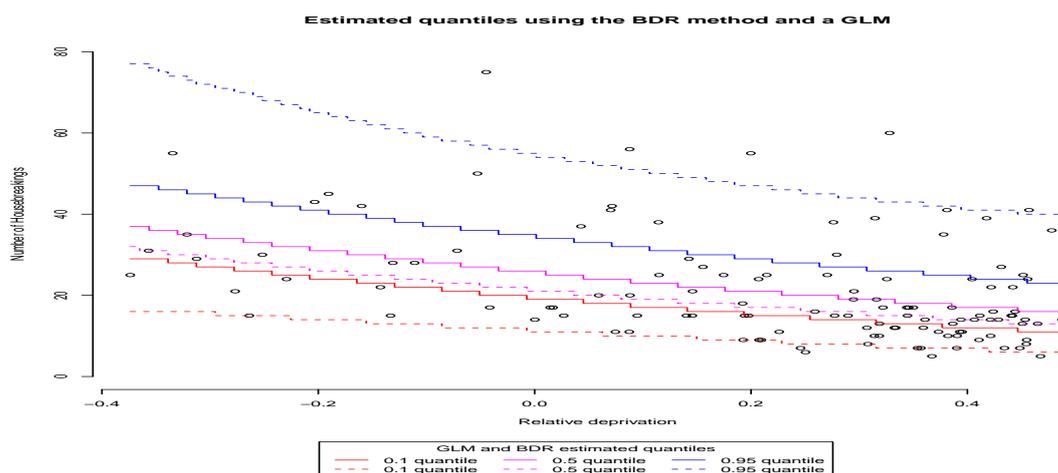


Figure 1: Estimated quantiles for housebreaking data, using discrete Bayesian density regression (dashed lines) and derived from a Poisson model.

5 Conclusions and further research

In this manuscript we have proposed a novel Bayesian density regression technique for discrete data which is based on mixing COM-Poisson distributions. The new method takes advantage of the exchange algorithm and updates the cluster allocations by drawing a new allocation for an auxiliary observation and then accepting or rejecting it. As a result the MCMC samples from the target distribution without the need to estimate the normalisation constant of the likelihood. The method overcomes the two main drawbacks of the “jittering” method for discrete quantile regression, namely that it does not require the addition of artificial additional noise and that it does not suffer from the problem of crossing quantiles. We have illustrated the method in a real

world application as well as simulated examples in which our method compared favourably to the “jittering” method. Further research efforts will be devoted in improving the computational speed and efficiency of the MCMC algorithm to make it an even more attractive alternative to “jittering”.

We would like to thank the anonymous referee for his suggestions and comments.

Bibliography

- [1] Buchinsky, Moshe (1998) *Recent advances in quantile regression models: A practical guideline for empirical research*. The journal of human resources, **33**, 88–126.
- [2] Conway, Richard W. and Maxwell, William L. (1962) *A queuing model with state dependent service rate*. Journal of industrial engineering, **12**, 132–136.
- [3] Dunson, David B. and Pillai, Natesh and Park, Ju-Hyun. (2007) *Bayesian density regression*. Journal of the royal statistical society: Series B, **69**, 163–183.
- [4] Dunson, David B. and Stanford, Joseph B. (2005) *Bayesian inferences on predictors of conception probabilities*. Biometrics, **61**, 126–133.
- [5] Karabatsos, George and Walker, Stephen G. (2012) *Adaptive-modal Bayesian nonparametric regression*. Electronic journal of statistics, **6**, 2038–2068.
- [6] Koenker, Roger and Bassett, Gilbert (1978) *Regression quantiles*. Econometrica, **46**, 33–50.
- [7] Machado, José António Ferreira and Santos Silva, João M.C.. (2005) *Quantiles for counts*. Journal of the american statistical association, **100**, 1226–1237.
- [8] Møller, J. and Pettitt, A. N. and Reeves, R. and Berthelsen, K. K. (2006) *An efficient Markov chain Monte Carlo method for distributions with intractable normalising constants*. Biometrika, **932**, 451–458.
- [9] Murray, Ian and Ghahramani, Zoubin and MacKay, David J. C. (2006) *MCMC for doubly-intractable distributions*. Proceedings of the 22nd Annual UAI Conference, 359–366.
- [10] Neal, Radford M. (2000) *Markov chain sampling methods for Dirichlet process mixture models*. Journal of computational and graphical statistics, **9**, 249–265.
- [11] Shmueli, Galit and Minka, Thomas P. and Kadane, Joseph B. and Borle, Sharad and Boatwright, Peter (2008) *A useful distribution for fitting discrete data: revival of the Conway-Maxwell-Poisson distribution*. Journal of the royal statistical society: Series C, **54**, 127–142.

Score Function of Distribution and Heavy-tails

Zdeněk Fabián, *Inst. of Computer Science ASCR*, `zdenek@cs.cas.cz`

Abstract. In this contribution we explain the recently introduced notion of the distribution-dependent scalar-valued score function of distribution. Function and its moments are used for description of continuous distributions and data samples generated from them. Each distribution including the heavy-tailed ones and random samples from them are described by a typical value and variability. Further, we discuss the generalized (score) moment estimates and a distribution-dependent score correlation coefficient for continuous random variables, and present results of simulation experiments with data generated from heavy-tailed distributions. Since score functions of distribution of heavy-tailed distributions are bounded, the point estimates as well as the sample score correlation coefficients are insensitive to outliers.

Keywords. Score function, Point estimation, Correlation, Heavy-tailed distributions.

1 Introduction

Statistical estimation is typically based on averaging functions generally called score functions. In the classical parametric setup are observed data x_1, \dots, x_n taken as realizations of random variables X_1, \dots, X_n iid according to F , where F is assumed to be a member of a regular parametric family $\mathcal{F}_{\mathcal{X}} = \{F_{\theta} : \theta \in \Theta \subseteq \mathbb{R}^m\}$ with support interval $\mathcal{X} \subseteq \mathbb{R}$ and densities $f(x; \theta)$. The result of the inference procedure is the density $f(x) \approx f(x; \hat{\theta}_n)$ where $\hat{\theta}_n$ is the estimate of the true θ with suitable properties. During the estimation procedure the observed data are 'treated' in accordance with the model, that is, inserted into some type of score functions describing relative influence of observations with respect to the estimated characteristic of the model distribution. Thus, there are not the random variables themselves entering into inference procedure, but the 'latent' values of the corresponding score functions. However, the vector nature of the Fisher (maximum likelihood) score functions of classical statistics does not enable a consistent use of this point of view and scalar-valued score functions of robust statistics are often in a loose relation to the assumed model.

As a remedy of this 'state of the world' we have constructed in [1], [2] and [6] a scalar-valued score function reflecting the model, with the mathematical form depending on the support interval \mathcal{X} . It was derived by means of transformations from the score function $S_G(y) = -g'(y)/g(y)$

of 'prototype' distribution G with support \mathbb{R} . In some cases the new function equals the Fisher score for the central parameter of the distribution, in other ones it is yet an unknown function.

2 Score function of distribution

By $\Pi_{\mathcal{X}}$ is denoted the set of distributions with (finite or infinite) interval support \mathcal{X} and regular in the sense stated later. Let $f(x)$ be the density of $F \in \Pi_{\mathcal{X}}$ and $\eta: \mathcal{X} \rightarrow \mathbb{R}$ be a smooth strictly increasing mapping. Let g be the density of some $G \in \Pi_{\mathbb{R}}$ such that

$$f(x) = g(\eta(x))\eta'(x) \quad (1)$$

where the Jacobian of the transformation $\eta'(x) = d\eta(x)/dx$. G is called the *prototype* of F .

We noticed that any distribution with support $\mathcal{X} \neq \mathbb{R}$ can be taken as a transformed prototype.

Definition 2.1. We say that $\eta: \mathcal{X} \rightarrow \mathbb{R}$ is Johnson's mapping if

$$\eta(x) = \begin{cases} x & \text{when } \mathcal{X} = \mathbb{R} \\ \log(x-a) & \text{when } \mathcal{X} = (a, \infty) \\ \log \frac{(x-a)}{(b-x)} & \text{when } \mathcal{X} = (a, b). \end{cases} \quad (2)$$

Theorem 2.2. For any $\mathcal{X} \in \mathbb{R}$, the decomposition of density $f(x)$ of any regular $F \in \Pi_{\mathcal{X}}$ into form (1) is unambiguous.

Proof. Either $\eta(x)$ and/or $\eta'(x)$ in formula (1) are clearly identifiable, or $f(x)$ is to be written in the form

$$f(x) = \frac{1}{\eta'(x)} f(x)\eta'(x) \quad (3)$$

where $\eta'(x)$ is the derivative of the Johnson's mapping for the given \mathcal{X} .

The 'right' $\eta(x)$ is called the *innate mapping*. Often, $\eta(x)$ and/or $\eta'(x)$ in the formula (1) can be easily recognized. For instance, distribution from $\Pi_{(-\pi/2, \pi/2)}$ with density $f(x) = \frac{1}{\sqrt{2\pi \cos^2 x}} e^{-\frac{1}{2} \tan^2 x}$ is in the form (1) with innate mapping $\eta(x) = \tan x$, the density of the inverse gamma distribution from $\Pi_{(1, \infty)}$ is $f(x) = \frac{c^\alpha}{\Gamma(\alpha)} (\log x)^{\alpha-1} \frac{1}{x^{c+1}} = \frac{c^\alpha}{\Gamma(\alpha)} (\log x)^\alpha \frac{1}{x^c} \frac{1}{x \log x}$ with innate mapping $\eta(x) = \log \log x$ and the density of the loglogistic distribution from $\Pi_{(0, \infty)}$, $f(x) = \frac{c}{x} \frac{(x/\tau)^c}{[(x/\tau)^c + 1]^2}$ has $\eta'(x) = 1/x$. The density of the exponential distribution from $\Pi_{(0, \infty)}$ cannot be decomposed into (1); by Theorem 2.2 it is to be written as $f(x) = e^{-x} = x e^{-x} \frac{1}{x}$. Johnson's mappings (2) as 'default innate mappings' are used not only due to the principle of parsimony (they are the simplest possible mappings $\eta: \mathcal{X} \rightarrow \mathbb{R}$), but, as it is seen from the Table 1, due to reasonable consequences as well: the score functions of distribution of the exponential and uniform distributions are, when using (2), linear. Another examples are given in [6].

Definition 2.3. Let $F \in \Pi_{\mathcal{X}}$ has density $f(x)$ and $\eta: \mathcal{X} \rightarrow \mathbb{R}$ be the innate mapping. Set

$$T_F(x) = -\frac{1}{f(x)} \frac{d}{dx} \left[\frac{1}{\eta'(x)} f(x) \right]. \quad (4)$$

Let the solution x^* to the equation

$$T_F(x) = 0 \quad (5)$$

be unique. x^* is called the score mean, and function

$$S_F(x) = \eta'(x^*)T_F(x) \quad (6)$$

is the score function of distribution F (sfd of F).

Let us explain the concepts introduced so far. If random variable Y has distribution $G \in \Pi_{\mathbb{R}}$, random variable $X = \eta^{-1}(Y)$ has distribution $F = G \circ \eta$ with density (1). The first term on the right hand side of (1) contains probabilistic information about X , whereas the second term $\eta'(x)$, the Jacobian of a (virtual) transformation, is carrying information about \mathcal{X} only and masking the genuine change of the statistical content of $f(x)$ and is to be removed before differentiation with respect to variable in (4). By [1], $T_F(x) = S_G(\eta(x))$.

3 Description of distributions by sfd

Let $F \in \Pi_{\mathcal{X}}$. The regularity conditions are: f is differentiable according to the variable a.e., the density g of prototype distribution G is differentiable and unimodal, and $ES_G^2 < \infty$.

The *score moments*

$$ES_F^k = \int_{\mathcal{X}} S_F^k(x) f(x) dx \quad (7)$$

in cases of regular distributions exist and are usually given by simple expressions, since S_F 'fits' the distribution F , as is apparent from (4) and (6).

The score mean is taken as the typical value of a distribution: score mean $y^* = \mu$ of a location prototype distribution $G_{\mu} \in \Pi_{\mathbb{R}}$ with sfd

$$S_G(y - \mu) = -\frac{g'(y - \mu)}{g(y - \mu)} = 0$$

is the mode, the sfd of $F \in \Pi_{\mathcal{X}}$ given by $F = G \circ \eta$ is

$$S_F(x; \tau) = S_G(\eta(x) - \eta(\tau))$$

where we set $\tau = \eta^{-1}(\mu)$. Parameter τ is actually the *transformed location* of the prototype. It was shown in [1] or [6] that for distributions with transformed location (considered by us as the 'central parameter' of the distribution) it holds

$$\frac{\partial}{\partial \tau} \log f(x; \tau) = S_F(x; \tau),$$

which entitles the removing of $\eta'(x)$ in (4). In a general case, x^* is the transformed mode y^* of the prototype, $x^* = \eta^{-1}(y^*)$, cf. [1], which may or may be not a component of θ , and sfd is the generalized Fisher score for x^* .

The value ES_F^2 is interpreted as *Fisher information for x^** or even as Fisher information of distribution F . Its reciprocal value, which we call the *score variance*,

$$\omega^2 = \frac{1}{ES_F^2}, \quad (8)$$

has been suggested in [2] as a measure of variability of F . The values $x^*(\theta)$ and $\omega^2(\theta)$ can be used instead of the mean value and variance (which may be infinite in cases of heavy-tailed distributions) as representatives of the center and variability of distributions.

4 Example: Description of two heavy-tailed distributions

Fréchet distribution from $\Pi_{(0,\infty)}$ with density

$$f(x; \tau, c) = \frac{c}{x} \left(\frac{x}{\tau}\right)^{-c} e^{-\left(\frac{x}{\tau}\right)^{-c}} \quad \tau, c > 0, \tag{9}$$

is an example of a distribution with transformed location parameter τ (its prototype is the extreme value distribution with location parameter). The distribution has mean $EX = \tau\Gamma(1 - 1/c)$ and variance $VarX = \tau^2[\Gamma(1 - 2/c) - (EX)^2]$, existing only if $c > 1$ and $c > 2$, respectively. By (4)-(6), its sfd

$$S_F(x; \tau, c) = \frac{c}{\tau} [1 - (\tau/x)^c]$$

equals the Fisher score function for τ . Fisher information for τ is $ES^2 = (c/\tau)^2$ so that the score variance $\omega^2 = \tau^2/c^2$ describing variability is proportional to the square of the score mean, similarly as the variance. However, it is finite.

Beta-prime distribution (beta distribution of the second kind) from $\Pi_{(0,\infty)}$ with density

$$f(x; p, q) = \frac{1}{B(p, q)} \frac{x^{p-1}}{(x+1)^{p+q}}, \quad p, q > 0 \tag{10}$$

where B is the beta function, is an example of a distribution without central parameter. Its mean $EX = p/(q - 1)$ and variance $VarX = p(p + q + 1)/[(q - 1)^2(q - 2)]$ exist only if $q > 1$ and $q > 2$, respectively. By (4), $T_F(x; p, q) = \frac{qx-p}{x+1}$, the score mean is thus $x^* = p/q$ and the sfd

$$S_F(x; p, q) = \frac{q^2}{p} \frac{x - x^*}{x + 1} \tag{11}$$

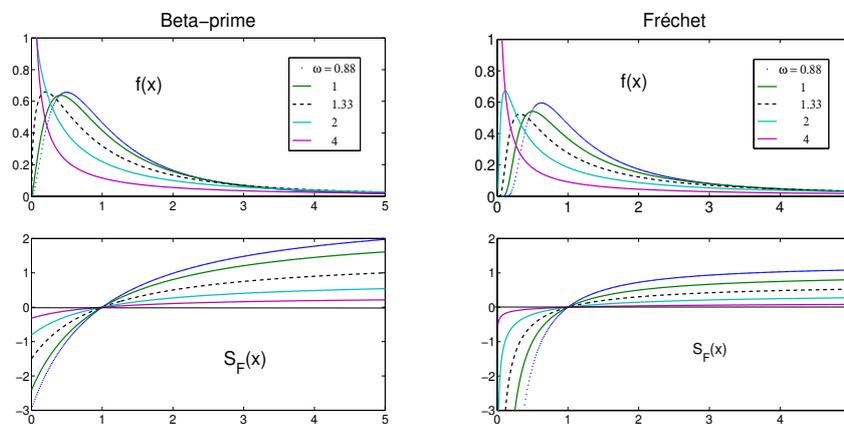


Fig. 1. Densities and sfd of two heavy-tailed distributions.

is different from both Fisher scores for p and q . (11) can be taken as the generalized Fisher score for the 'center' $x^* = p/q$ of the distribution. Since $ET_F^2 = pq/(p + q + 1)$, the variability of the distribution is described by the score variance $\omega^2 = \left(\frac{p}{q}\right)^2 \frac{(p+q+1)}{pq}$, proportional to $(x^*)^2$ again. The density (10) in the form $f(x; 1, c)$ appears to be the density of the 'shifted' Pareto distribution.

Densities and sfd of both distribution with $x^* = 1$ and the same various values of ω are in Fig. 1.

5 Point estimation

Let $\theta \in \Theta \subseteq \mathbb{R}^m$ and X_1, \dots, X_n be random variables iid according $F \in \text{mathcal{F}}_\S = \{F_\theta, \theta \in \Theta\}$. The finite counterpart of parametric score moments $ES_F^k(\theta)$ given by (7) are the generalized moment equations defining the *score moment estimator*

$$\hat{\theta}_n : \quad \frac{1}{n} \sum_{i=1}^n S_F^k(x_i; \theta) = ES_F^k(\theta), \quad k = 1, \dots, m, \tag{12}$$

which is a special form of the m-dimensional M-estimator.

In some particular cases equations (12) are especially simple. While score moment equations for parameters are to be solved by an iterative way, sfd of some (one-parameter or even two-parameter) distributions can be written in the form $S_F(x; x^*)$, where $x^* = x^*(\theta)$, so that the first score moment equation (12) sounds

$$\sum_{i=1}^n S_F(x_i; x^*) = 0 \tag{13}$$

and the estimate \hat{x}^* of x^* is efficient, cf. [7].

In a general case, score moment estimates are not efficient, but since $S_F(x; \theta)$ is a scalar-valued function (even if θ is a vector), they have a remarkable advantage with respect to the maximum-likelihood estimators: for distributions with bounded sfd are robust to outliers for all the components of θ ; in cases of unbounded sfd can be relatively easily modified by some of robust approaches, cf. [8], see [6]. It should be said that in cases of heavy-tailed distributions, an outlier means an extremely large value generated, however, in accordance with the model.

In the case of the beta-prime distribution with typical value $x^* = p/q$, the sample score mean (typical value of the sample) can be estimated from (13) having a simple form

$$\sum_{i=1}^n \frac{x_i - x^*}{x_i + 1} = 0,$$

that is, by an explicit formula. Some examples of other distributions with the sample score mean expressed by explicit formulas are given in Table 1.

\mathcal{X}	F	$f(x)$	$S_F(x)$	x^*	\hat{x}^*
\mathbb{R}	normal	$\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}$	$\frac{1}{\sigma} \frac{x-\mu}{\sigma}$	μ	\bar{x}
$(0, \infty)$	gamma	$\frac{\gamma^\alpha}{x\Gamma(\alpha)} x^\alpha e^{-\gamma x}$	$\frac{\gamma^2}{\alpha} (x - x^*)$	$\frac{\alpha}{\gamma}$	\bar{x}
$(0, 1)$	beta	$\frac{x^{p-1}(1-x)^{q-1}}{B(p,q)}$	$(p+q)(x - x^*)$	$\frac{p}{p+q}$	\bar{x}
$(0, \infty)$	lognormal	$\frac{1}{\sqrt{2\pi}x} e^{-\frac{1}{2}c \log^2(\frac{x}{\tau})^c}$	$c \log(\frac{x}{\tau})^c$	τ	$\prod_{i=1}^n x_i$
$(0, \infty)$	Weibull	$\frac{c}{x} (\frac{x}{\tau})^c e^{-(\frac{x}{\tau})^c}$	$\frac{c}{\tau} [(\frac{x}{\tau})^c - 1]$	τ	$(\frac{1}{n} \sum_{i=1}^n x_i^c)^{1/c}$
$(1, \infty)$	Pareto	c/x^{c+1}	$c^2(1 - \frac{x^*}{x})$	$\frac{c+1}{c}$	\bar{x}_H
\mathbb{R}	extr. value	$\frac{1}{\sigma} e^{-\frac{x-\mu}{\sigma}} e^{-e^{-\frac{x-\mu}{\sigma}}}$	$\frac{1}{\sigma}(1 - e^{-\frac{x-\mu}{\sigma}})$	μ	$-\sigma \log(\frac{1}{n} \sum_{i=1}^n e^{-x_i/\sigma})$

Table 1. Density, sfd, score mean and sample score mean of some distributions. \bar{x}_G and \bar{x}_H denotes the geometric and harmonic mean, respectively, σ in the case of extreme value and c in the case of the Weibull distribution are supposed to be known constant.

No matter by which method has the parameter vector been estimated, the *sample score mean* and the *sample score variance* can be constructed as $\hat{x}_n^* = x^*(\hat{\theta}_n)$ and $\hat{\omega}_n^2 = \omega^2(\hat{\theta}_n)$, respectively, giving a possibility to describe data samples by their typical value and dispersion under the assumption of the model \mathcal{F}_X . This approach guarantees an easy comparison of results of the estimation for various assumed models with a different type and a different number of parameters.

6 Score correlation coefficient

Let X and Y be random variables with supports \mathcal{X}_X and \mathcal{X}_Y , respectively, with joint distribution F_{XY} , marginal densities $f_X(x)$, $f_Y(y)$ and marginal sfds $S_X(x)$, $S_Y(y)$. Let us denote by $\rho_P(X, Y)$ Pearson's correlation coefficient.

The distribution-dependent measure of association of random variables, the *score correlation coefficient*, has been defined in [4] as

$$\rho_F(X, Y) = \rho_P(S_X, S_Y). \quad (14)$$

Given a random sample $(x_1, y_1), \dots, (x_n, y_n)$ from some F assumed to be member of a parametric family $F_{XY}(\theta_X, \theta_Y)$, the *sample score correlation coefficient* is given by

$$r_F = \frac{\sum S_X(x_i; \hat{\theta}_X) S_Y(y_i; \hat{\theta}_Y)}{\sqrt{\sum S_X^2(x_i; \hat{\theta}_X) \sum S_Y^2(y_i; \hat{\theta}_Y)}}, \quad (15)$$

where $\hat{\theta}_X$ and $\hat{\theta}_Y$ are the estimates of vectors of parameters of marginal distributions.

To model association of random variables X and Y , we set in simulation experiments

$$Y = \alpha X + (1 - |\alpha|)Z \quad (16)$$

where X and Z were generated independently from the same distribution. The theoretical value of the correlation coefficient is

$$\rho = \rho(X, Y; \alpha) = \alpha / \sqrt{2\alpha^2 - 2|\alpha| + 1}.$$

Samples were generated from different distributions with increasing variability of distributions described by ω by setting $\theta = \theta(x^*, \omega)$. For each sample (of length $n = 75$) were computed also Pearson's correlation coefficient, Spearman's rank coefficients, Kendal's tau and the robust correlation coefficient with Huber's score function (each computed by means of the Statistical toolbox from MATLAB), averaged after 10^4 replications for each ω and plotted against ω for data generated from various distributions. For distributions with 'mild' non-symmetry (that is, for distributions with support $\mathcal{X} = \mathbb{R}$ and the gamma, Rayleigh and Maxwell) were all averages roughly equal to the theoretical value: correlation properties overcome the structure of distributions. However, correlation coefficients in cases of highly non-symmetric distributions with support $\mathcal{X} = (0, \infty)$ are strongly dependent on the variability of the distribution described by ω . In cases of heavy-tailed distributions (as Fréchet, beta-prime and Pareto), Pearson's coefficients with quickly increasing MSEs with increasing ω are of no use as well as Kendal's tau with queer results even for small ω . Correlation coefficients capable to detect association of random variables with heavy-tailed distributions are the score correlation coefficient (r_F), the Spearman

rank coefficient (r_S) and, to a certain extent, robust correlation coefficients (r_R). A simulation study of the behavior of these three correlation coefficients with increasing variability of the beta-prime distribution is given in Fig. 2.

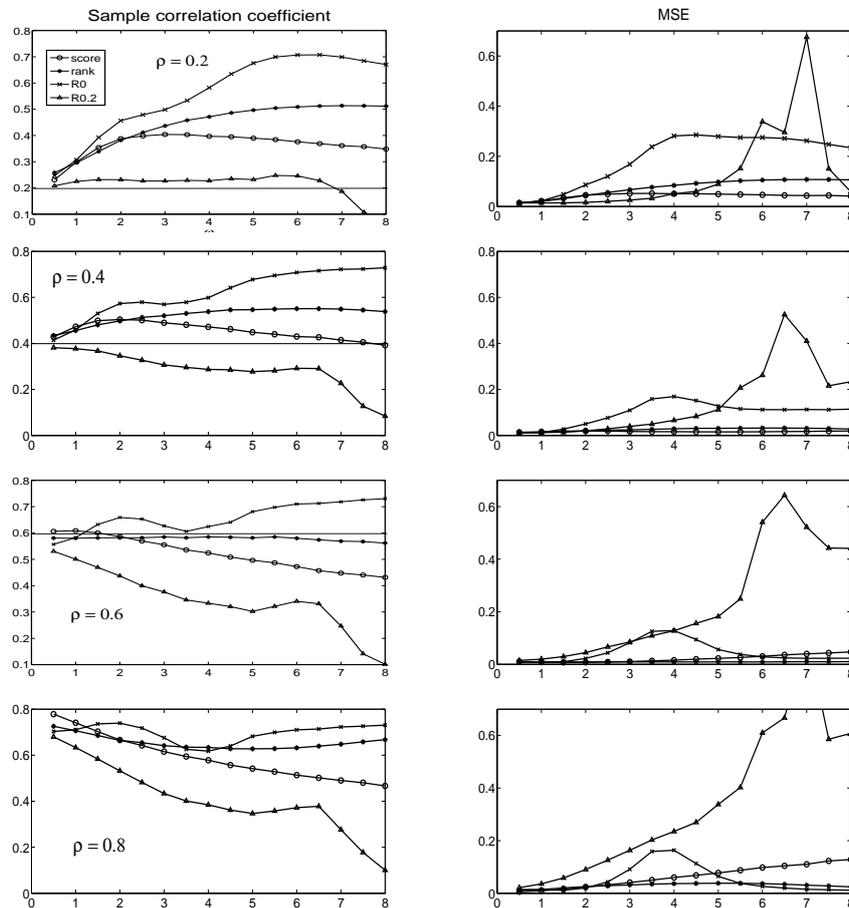


Fig. 2. Average sample correlation coefficients of heavy-tailed distributions in increasing variability of distributions.

The clearly apparent tendency with increasing variability of distributions is that both r_F and r_S possess a strong positive bias in cases of small values of ρ (with less biased r_F), and strong negative bias if ρ approaches to 1 (with less biased r_S), with splitting value about $\rho = 0.5$. This behavior is, unfortunately, unfavorable to attempts to find the true value of correlation of random variables with heavy-tailed distributions. In cases of heavy-tailed distributions we do not recommend robust correlation estimates since they are too dependent on the percent of trimming used according to [9]. By using two versions presented in Fig. 2, with score function $\psi(x) = const.$ if $|\psi(x - med(x))| \geq 3\omega$ without trimming (R0) and with 20% trimming (R0.2), we obtained quite dissimilar results.

7 Conclusions

By introducing the score function of distribution we obtained new characteristics of distributions, the score mean (center) and score variance (variability), finite even for heavy-tailed distributions.

It enables to view data from heavy-tailed distributions as data from regular distributions.

Acknowledgement

The work was supported by grant of the Czech Ministry of Education, Youth and Sports under contract No. LG12020.

Bibliography

- [1] Fabián, Z. (2001). *Induced cores and their use in robust parametric estimation*. Comm. Statist. Theory Methods, **30**, 537–556.
- [2] Fabián, Z. (2007). *Estimation of simple characteristics of samples from skewed and heavy-tailed distribution*. In Recent Advances in Stochastic Modelling and Data Analysis, Singapore, World Scientific, 43–50.
- [3] Fabián, Z. (2010). *Score moment estimators*. In Proc. of conference COMPSTAT 2010, Physica-Verlag, Springer.
- [4] Fabián, Z. (2010). *Score correlation*. Neural Network World, **20**, 793–798.
- [5] Fabián Z. (2013). *Score function of distribution and association of random variables*. Forum Statisticum Slovakum, **5**, 10-18.
- [6] Fabián, Z. (2014). *Score function of distribution and revival of the moment method*. Res. rep. <http://hdl.handle.net/11104/0225830> (to appear in Comm. Statist. Theory-Methods).
- [7] Fabián Z. (2009). *Confidence intervals for a new characteristic of central tendency of distributions*. Comm. Statist. Theory Methods **38** , 1804 - 1814.
- [8] Huber P. J., Ronchetti E. M. (2009). *Robust statistics*. Wiley.
- [9] Shevlyakov, G., Smirnov, P. (2011). *Robust estimation of the correlation coefficient: An attempt of survey*. Austrian J. of Statistics, **40**,147– 156.

Consensus Clustering of Time Series Data

Ayca Yetere Kursun, *Middle East Technical University*, e112987@metu.edu.tr

Cem Iyigun, *Middle East Technical University*, iyigun@metu.edu.tr

Inci Batmaz, *Middle East Technical University*, ibatmaz@metu.edu.tr

Abstract. In this study, we aim to develop a methodology that merges Dynamic Time Warping (DTW) and consensus clustering in a single algorithm. Mostly used time series distance measures require data to be of the same length and the distance between time series data mostly depends on the similarity of each coinciding data pair in time. DTW is a relatively new measure used to compare two time dependent sequences which may be out of phase or may not have the same lengths or frequencies. However, DTW is a similarity measure that is employed for single variable with standard clustering methods rather than consensus clustering. Thus our motivation is to create an algorithm that can combine the benefits of the DTW with benefits of consensus clustering, which will also provide a solution for multivariate applications. We present the results of our study both with simulated data and well known datasets from the literature.

Keywords. Consensus Clustering, Ensemble Clustering, Dynamic Time Warping, Time Series Clustering

1 Introduction and Motivation

Clustering is the activity of unsupervised grouping of data points into classes so that the similar objects will be in the same cluster. There are varieties of clustering methods extensively used in literature such as k -means, hierarchical clustering, graph partitioning and so on. Since each method depends on different rationale, the results obtained from their use usually may not be the same. This situation leads to some confusion regarding which one gives the best clustering result. The common practice is to find the overlapping classes generated by different clustering methods and determine the non-overlapping observations. However, we may not come up with a solid solution with this approach. Alternatively, domain knowledge if available can be utilized to resolve this problem.

Consensus clustering is an attempt to solve this problem objectively; it tries to combine multiple clusterings of a dataset into one consolidated clustering. Consensus clustering methodologies offers benefits such as improved quality of solution, improved robustness against wide

ranges of datasets, elimination of the model selection process, knowledge reuse, distributed clustering and effective consolidation of clusters depending on different views of data having multiple features [4].

Employing the clustering algorithms requires comparing two objects, thus one needs a distance (similarity) measure to define how much similar those two objects are. Commonly used similarity measures are Euclidean distance, Minkowski distance, Pearson's correlation coefficient and related distances, short time series distance and so on [7]. Mostly used time series distance measures require data to be of the same length and measure the distance between time series data mostly dependent on the similarity of each coinciding data pair in time. Dynamic Time Warping (DTW) is a relatively new measure used to compare two time dependent sequences with data which may be out of phase or may not have the same lengths or frequencies. DTW aligns two time series data so that the distance between them is minimized [7]. In literature DTW provided successful results when used for classification applications [5], while in this study it is used for clustering applications.

In this study, we aim to develop a methodology that merges DTW and consensus clustering in a single algorithm. DTW is a similarity measure that is employed for single variable with standard clustering methods rather than consensus clustering. Thus our motivation is to create an algorithm that can combine the benefits of the DTW with benefits of consensus clustering, which will also provide a solution for multivariate applications. In literature, time series clustering algorithms are used for several application areas like medicine, signal processing, economics, bio statistics and so on [6]. So we believe our approach with the DTW consensus clustering will also be applicable to those application areas.

2 Dynamic Time Warping

Dynamic Time Warping algorithm is a time-series similarity measure, which can be used for data that may be out of phase or may not have the same lengths or frequency for that matter. DTW algorithm can be summarized as follows;

Step 1. Generating the Cumulative Distance Matrix: The first step is to compare each point in one time series data with every other point in the second time series data, generating a matrix. So the cumulative distance between time series data points is calculated using dynamic programming technique.

Given two time series $X = (x_1, x_2, \dots, x_n)$ and $Y = (y_1, y_2, \dots, y_n)$ the cumulative distance matrix can be found using Equations (1), (2) and (3). In those equations, the Euclidian distance between two data points is normally used for defining $d(x_i, y_i)$.

$$dtw(1, j) = d(x_1, y_j) + dtw(1, j - 1) , \quad (1)$$

$$dtw(i, 1) = d(x_i, y_1) + dtw(i - 1, 1) , \quad (2)$$

$$dtw(i, j) = d(x_i, y_j) + \min\{dtw(i - 1, j - 1), dtw(i - 1, j), dtw(i, j - 1)\} \quad (3)$$

In the above formulation one step dynamic programming is used, however the depth (time window) of algorithm can be defined specific to the problem nature. The Euclidean distance measure can be seen as a special case of DTW with step size being equal to zero. However, this special case can only be defined when the two time series have the same length.

Step 2. Finding The Optimal Path: The optimal warping path is the minimum distance path on the cumulative distance matrix. The minimum distance path is a sequence of points $p = (p_1, p_2, \dots, p_Q)$, with $p_l = (i, j) \in [1 : n] \times [1 : m]$ for $l \in [1 : Q]$ (with Q being the total number of points in the path), $i \in \mathbb{Z}$, $j \in \mathbb{Z}$ and $l \in \mathbb{Z}$, satisfying the following conditions [9]:

- **Boundary condition:** The starting and ending points of the warping path must be the first and the last points of aligned sequences, $p_1 = (1, 1)$ and $p_Q = (n, m)$.
- **Monotonicity condition:** $n_1 \leq n_2 \leq \dots \leq n_Q$ and $m_1 \leq m_2 \leq \dots \leq m_Q$. This condition preserves the time-ordering of points.
- **Step size condition:** Limits the warping path making big shifts in time while aligning sequences. Step size condition can be formulated as $p_{l+1} - p_l \in \{(1, 1), (0, 1), (1, 0)\}$ for a single step size.

So starting in reverse order with $p_Q = (n, m)$ and finishing with $p_1 = (1, 1)$ the simple procedure for the optimal path is described in Equation (4) [10]:

$$p_l = \left\{ \begin{array}{ll} (1, j-1) & \text{if } i = 1, \\ (i-1, 1) & \text{if } j = 1, \\ \operatorname{argmin}\{dtw(i-1, j-1), dtw(i-1, j), dtw(i, j-1)\} & \text{otherwise.} \end{array} \right\} \quad (4)$$

3 Consensus Clustering

In literature, the idea of using several runs of one or more clustering algorithms, different parameters of an object or dataset resamples, to create better clusters is known as consensus clustering, clustering aggregation or in other words ensemble clustering. In this study we will use the term ‘consensus clustering’ to indicate this idea. However the only reason for using consensus clustering is not only to obtain better clustering. Ghosh et al. [4] and Ghaemi et al. [2] lists other reasons to use consensus clustering as: improved quality of solution, novelty, stability, elimination of model selection, knowledge reuse, multiview clustering, distributed computing (parallelization).

Consensus clustering is generally a two-stage approach. First stage is to create diversity of clustering and the second stage is to obtain a consensus across those diverse solutions by utilizing an algorithm. Diversity can be achieved by several mechanisms [4] [2]: using different clustering algorithms, using different initialization points or parameters, using different subsets of data or creating resamples from the original data and using different features of data.

In this study, the consensus methodology developed by Monti et al. [8] is used. This methodology uses a resampling-based approach for class discovery and clustering validation. Monti et al.’s study provides a method to represent the consensus across multiple runs of clustering algorithms [8]. This approach is an coassociation matrix based approach for consensus clustering. Monti et al. [8] expressed their motivation for the proposed methodology as to increase the robustness and stability of clusters to sampling variability. They have also explained that their method can also be used to represent the consensus over multiple runs of a clustering algorithm with random restart so as to account for the sensitivity to the initial conditions. Even though

it was not mentioned in their paper, their approach is also suitable for obtaining a consensus result for different clustering algorithms, as it is suggested by Simpson [11]. Also it is suitable for multiview (multivariate) clustering, giving way for the utilization of different features of the data.

So Monti et al.s' proposed methodology, by using different clustering algorithms, different initialization points or parameters, different features of data and resamples from the original data, has the benefits of 'Improved Quality of Solution,' 'Novelty,' 'Robust Clustering' and 'Stability'. In this study, we use this proposed approach for achieving a consensus clustering result, by also including the usage of different clustering algorithms. This consensus clustering approach can simply be summarized as follows [8]:

For a selected bootstrapping technique with different clustering algorithms and number of clusters;

1. Resample the dataset for h iterations (in our case the square distance matrix developed by DTW will be resampled).
2. Select a number of clustering algorithms for the consensus solution,
 $K = \{k\text{-means, Agglomerative Nesting}\}$

Starting from the first clustering algorithm, $k=1$, repeat Step 3 and 4 for all the clustering algorithms:

3. Apply the clustering algorithm to each and every resampled dataset.
4. Compute the consensus clustering matrix using all the runs (each and every resampled dataset) from the same algorithm. Here, the consensus clustering matrix for the k^{th} clustering algorithm can be generated using the following equation:

$$C_k^*(i, j) = \frac{\sum_h C_k^h(i, j)}{\sum_h I_k^h(i, j)} \quad (5)$$

Here, C_h^k is the connectivity matrix corresponding to the h^{th} iteration of the k^{th} clustering algorithm, where

$$C_k^h(i, j) = \left\{ \begin{array}{ll} 1, & \text{if item } i \text{ and } j \text{ belongs to the same cluster,} \\ 0, & \text{otherwise.} \end{array} \right\} \quad (6)$$

Furthermore, I_h^k is the indicator matrix corresponding to the h^{th} iteration of the h^{th} clustering algorithm such that

$$I_k^h(i, j) = \left\{ \begin{array}{ll} 1, & \text{if item } i \text{ and } j \text{ are present in the same resampled dataset,} \\ 0, & \text{otherwise.} \end{array} \right\} \quad (7)$$

5. Combine the consensus clustering matrices obtained for each algorithm using weights (w_k) in order to form the following merged matrix C^*

$$C^*(i, j) = \sum_k w_k C_k^*(i, j) \quad (8)$$

6. Use the merged matrix C^* as a similarity matrix and obtain the final clustering solution.

4 Multivariate Problems

As mentioned earlier, one of the benefits offered by consensus clustering is to obtain a single consolidated partition by effectively combining all the clusterings of different aspects of the data. For the multivariate case there can be several different approaches to tackle the problem. This study deals with the following two approaches:

Combining the similarity matrices into a single merged similarity matrix and obtaining the clusters with the consensus clustering algorithm: As for each variable before creating a consensus solution with the algorithm defined in previous section, one can create an ensemble using the similarity matrices of each variable using the DTW methodology. This can be simply done by using Equation (9), where DTW_n represents the similarity matrix for the n^{th} variable:

$$DTW_Merged = \frac{\sum_{n=1}^N DTW_n}{N} \quad (9)$$

Combining the merged matrix of each variable and obtaining a final consensus clustering: It is also possible to use the same approach defined in the previous section to obtain the ensemble of variables using the merged matrices of the consensus algorithm (see Figure1).

In that case one can obtain the final merged matrix using Equation (10), where represents the merged matrix for the n th variable obtained by using the consensus clustering algorithm:

$$C^*(i, j) = \frac{\sum_{n=1}^N C_n^*(i, j)}{N} \quad (10)$$

5 Experimentation

In order to test our proposed approach we have experimented with four distinct datasets. Two of those datasets were created specific for this study. The other two datasets are used in the literature for testing of the classification algorithms for time series data. Those two datasets are Synthetic Control Dataset and Daily and Sports Activities Dataset [1].

In literature Agglomerative Hierarchical clustering and k -means clustering algorithms with Euclidian distance measure are the mostly used algorithms for time series data clustering analysis. Hence we have used those algorithms in order to compare the performance of theirs to that

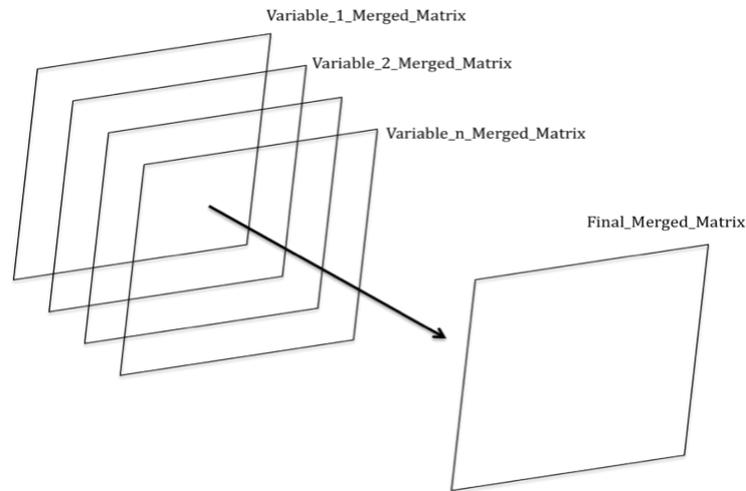


Figure 1: Final Merged Matrix

of our proposed algorithm. Initially we will discuss the performance of just DTW as a time series distance measure when compared to that of the Euclidian distance measure. For that purpose we will be using different window sizes (namely, window sizes 1, 2, 3, 4 and 5).

For obtaining consensus clustering merged matrices, we have utilized the R Package that was created by Simpson [11]. For all datasets we have used Agglomerative Nesting (Hierarchical Clustering), Partitioning Around Medoids, Divisive Analysis Clustering and k -means as different clustering algorithms within the consensus clustering algorithm. All four clustering algorithms were equally weighted for calculation of the final consensus clustering merged matrix.

For most of the cases we experimented with, DTW provides better results than the Euclidian Distance measure. Mostly usage of window size equal to one provides better results than the usage of other window sizes. However, consensus clustering with DTW is computationally very expensive when compared to the usage of Euclidian Distance and the conventional clustering algorithms. Thus this feature makes it harder to work with large datasets having too many time series samples and data points. Also in some cases the performance difference is around 1%, which makes it unnecessary to use both DTW and Consensus Clustering simultaneously.

But in all the cases we experimented with, when used with consensus clustering DTW performs better than Euclidian Distance measure, both regarding the errors and cluster discoveries. In addition, generally k -means (both as conventional clustering algorithm and final clustering algorithm for consensus clustering) is better in performance (errors and the number of clusters detected truly) compared to hierarchical clustering when DTW is used as a distance measure.

Also, dataset we have created backed up our initial expectation that DTW would perform better with data having phase shifts. This point is open for further experimentation with simulated datasets as the phase shift properties of real datasets are hard to observe if it wasn't considered in the data collection and mentioned in the dataset description.

Finally it should be mentioned that all these conclusions are dependent on the dataset's properties and need to be experimented with more data in detail in order to be expressed firmly.

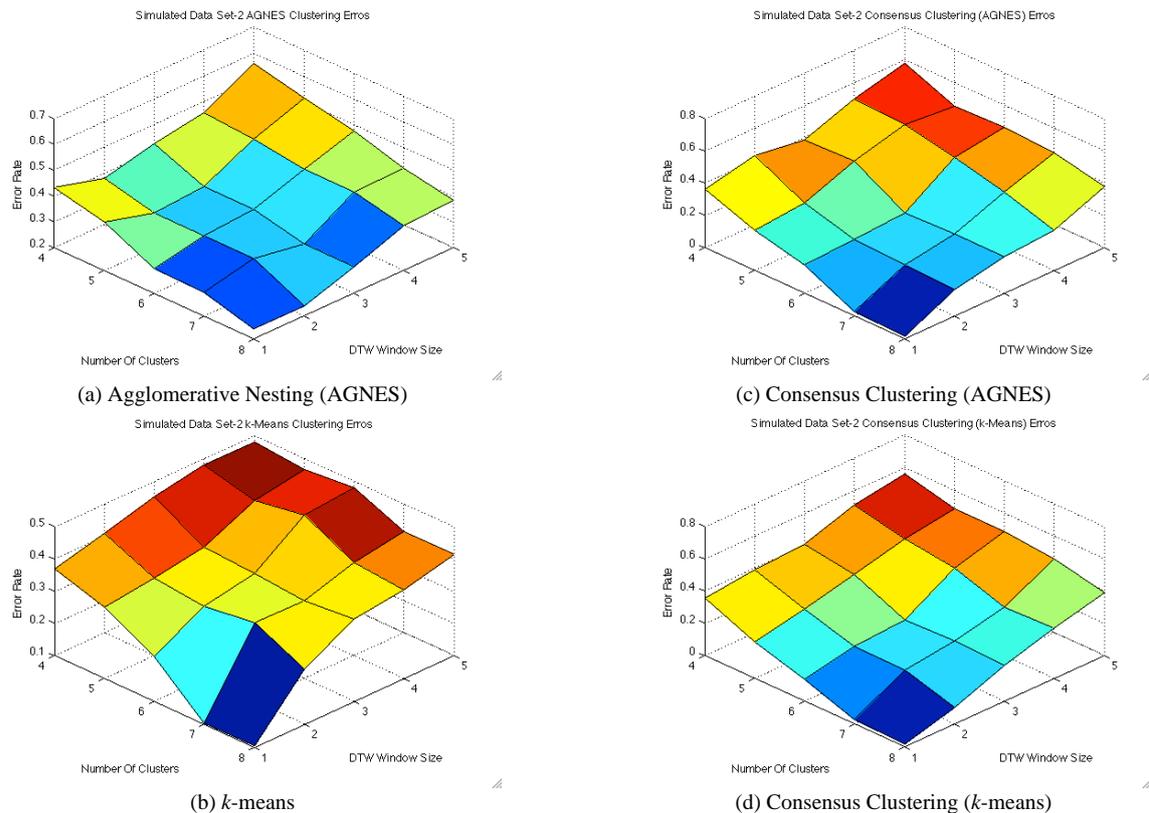


Figure 2: Example for Error Rates with Respect to Window Size and Number of Clusters

6 Conclusion and Future Work

Our initial suggestion was that, better clustering results can be obtained by using a similarity measure (DTW for our study) that is suitable for time series data, rather than simple distance measures (Euclidian Distance Measure etc.) used for common applications. So in this study we have discussed the use of DTW as a similarity measure for time series data in clustering applications and whether it performs better or not. Results of our experimentation showed that even though DTW is computationally very expensive, for most of the cases we experimented with DTW provides better results than the Euclidian Distance measure. As a feature work it will be also beneficial to use additional distance measures (cross-correlation etc.) to compare with DTW. In addition to our discussions with DTW we also discussed consensus clustering in this study. Regarding our experimentation, DTW with consensus clustering performs better than Euclidian Distance measure. However in some cases the performance difference was not beneficial enough to use both DTW and Consensus Clustering, due to time consuming computations. With the use of consensus clustering we also introduce to methodologies for multivariate clustering using DTW.

Bibliography

- [1] K. Bache and M. Lichman (2013) *UCI Machine Learning Repository*. [<http://archive.ics.uci.edu/ml>] Irvine, CA, University of California, School of Information and Computer Science.
- [2] R. Ghaemi, M. N. Sulaiman , H. Ibrahim , N. Mustapha (2009) *A Survey: Clustering Ensembles Techniques*. World Academy of Science, Engineering and Technology , **50**.
- [3] J. Ghosh, A. Strehl and S. Merugu (2001) *A Consensus Framework for Integrating Distributed Clusterings Under Limited Knowledge Sharing*. Proceedings of NSF Workshop on Next Generation Data Mining, Baltimore, USA, 99–108.
- [4] J. Ghosh and A. Acharya (2011) *Cluster Ensembles*. WIREs Data Mining and Knowledge Discovery, **1**, 305–315.
- [5] Y. Jeong, M. K. Jeong, O.A. Omitaomu (2011) *Weighted Dynamic Time Warping for Time Series Classification*. Pattern Recognition , **44**, 2231–2240.
- [6] E. Keogh, C. A. Ratanamahatana (2004) *Exact Indexing of Dynamic Time Warping*. Knowledge and Information Systems, Springer-Verlag London.
- [7] T.W. Liao (2005) *Clustering of Time Series Data â A Survey*. Pattern Recognition, **38**, 1857–1874.
- [8] S. Monti, P. Tamayo, J. Mesirov and T. Golub (2003) *Consensus Clustering: A Resampling-Based Method for Class Discovery and Visualization of Gene Expression Microarray Data*. Machine Learning , **1**, 91–118.
- [9] P. Senin (2008) *Dynamic Time Warping Algorithm Review*. Information and Computer Science Department University of Hawaii at Manoa Honolulu, USA, December.
- [10] M. Muller (2007) *Chapter 4: Dynamic Time Warping*. Informational Retrieval for Music and Motion, Springer Meinard.
- [11] T. I. Simpson (2011) *R Package-clusterCons*.

SDA for mixed-type data and its application to analysis of environmental radio activity level data

Yusuke Matsui, *Hokkaido University*, matsui@iic.hokudai.ac.jp
Masahiro Mizuta, *Hokkaido University*, mizuta@iic.hokudai.ac.jp

Abstract. A feature of Big Data is *variety*. We sometimes retrieve information based on many kinds of descriptions related to one's purposes. In SDA, *description* is a key issue for modeling objects. In this paper, we discuss the analysis for mixed-type data, which consist of different types of descriptions including sets of scalars, intervals, distributions and functions. We develop the method for analyzing relations among the descriptions, such as linearity, with PCA techniques. As an actual example, we analyze the monitoring data of environmental radio activity levels in Fukushima prefecture in Japan. The data are collected by various processes. We adopt the proposed method to the datasets from air borne monitoring survey, vehicle-borne survey, and stationary measurements on monitoring posts. We give the three *descriptions* for each city (as a concept): "Radio activity levels measured by air borne monitoring survey", "Radio activity levels measured by vehicle-borne survey" and "Radio activity levels measured on monitoring posts". We investigate the relations among them.

Keywords. BigData, Fukushima, environmental data, PCA, Correlation

1 Introduction

We come to utilize much more large and complex data than before. The emergence of cloud base systems bring us a potential to deal with Big Data. From the life science to the social science, the many have a great interest in it. Big Data is usually characterized by *3V* - *Volume*, *Velocity*, *Variety*. Typically, "Volume" claims inexhaustible storages and fast retrieval, "Velocity" demands efficient data processing and algorithms for the analysis, and "Variety" requires effective data modeling in the sense of the database architecture or the analyzing methodology.

From the view point of data analysis, "Variety" is challenging since there could be statistically new problems. In conventional data analysis, we deal with data which usually consists of scalar values. However, as application areas are growing, we need to treat the structured data instead of scalars, such as functional data, symbolic data including intervals, trees, modal, distributions etc. More recently, statistical analysis for topological manifold is studied.

In actual situations, it is often that we need to represent the objects by the various types of descriptions. However the treatment (and / or) interpretations of such data is usually difficult. One of the toughness stems from the nature of unstructured data, that is, we could not perform the usual operations such ordering and arithmetic etc. To deal with the difficulties, we pay attention to dissimilarities that are often commonly observable even when the data are unstructured.

In this paper, we focus on the dependency between descriptions and develop the method to analyze the their linear dependencies. The presented method could work with a various case of the mixed-type data. We also show the actual example, by applying the method to environmental radio activity level data related to the accident on Fukushima nuclear power plant.

SDA as a tool for Big Data

Here we briefly introduce the SDA related to Big Data and define some terminologies that we use later.

The reason why Big Data have *variety* is from *unstructured data*. That is, we could not deal with the data by the typical format of p variate n vectors. Actually, it is often the case that the data are intractable with missing and ill-formed. This force us perform the data "cleaning" where we usually "exclude" or "coerce" those values into simple alternatives.

In the frame work of SDA, we could naturally deal with complex data that does not have the form of p variate n vectors. The observations are considered to consist from two levels, called *1st level individual* and *2nd level individual*. 1st levels individuals are the most primitive observations such as points of p dimensional n vectors in Euclidean space. However, we could possibly consider observations equipped with more complex structures like intervals. Such objects are called 2nd level individuals.

In SDA, we often deal with the sets of 2nd level individuals such as categories, classes and concepts, and we call those sets *Concept space*. The concepts have structured variations in their own. To represent such data, we give *descriptions* to concepts, the mapping defined on concepts space to intervals, modal, distributions, functions and tree etc.

Under the framework, we think that we would naturally deal with the complex data and at the same time, we could also enhance the value of messy data by suitably abstracting the 1st level individuals.

In this paper, we deal with the concepts and consider the mixture of various descriptions, called "mixed-type data".

2 Analysis of mixed-type data

We assume that the concept space C and concepts are denoted as $c_i \in C$ for $i = 1, 2, \dots, n$. We define descriptions $Y_m : C \rightarrow S_m$ for $m = 1, 2, \dots, p$ where S_m are description spaces. We put description vector for each concept by $Y(c_i) = (Y_1(c_i), Y_2(c_i), \dots, Y_p(c_i))$.

For each description, we define dissimilarity measure $d_m : C \times C \mapsto \mathbf{R}^+$ for Y_m . We denote $d_{k,l}^{(m)} = d_m(c_k, c_l)$. We have dissimilarity matrix defined on Y_m . We put the matrix by $D^{(m)} = \{d_{k,l}^{(m)}; k, l = 1, 2, \dots, n\}$. We make the configurations in a space such as Euclidean space from the dissimilarity matrices $D^{(m)}(m = 1, 2, \dots, p)$.

In this paper, we consider the configurations in Euclidean space from each $D^{(m)}(m = 1, 2, \dots, p)$. We exploit Multidimensional scaling(MDS) technique and this time, we adopt the classical metric MDS. We put the number of dimensions to be configured as k_m and configurations as $X^{(m)} = (X_1^{(m)}, \dots, X_{k_m}^{(m)})$. As a whole, we recover the Euclidean space with $\mathbf{R}^M = \mathbf{R}^{\sum_{m=1}^p k_m}$, denoted by:

$$X := (X^{(1)}, X^{(2)}, \dots, X^{(p)}) = (X_1^{(1)}, X_2^{(1)}, \dots, X_{k_1}^{(1)}, X_1^{(2)}, X_2^{(2)}, \dots, X_{k_2}^{(2)}, \dots, X_1^{(p)}, X_2^{(p)}, \dots, X_{k_p}^{(p)}).$$

We adopt X as a *probe* for analyzing the structure of mixed-type data.

Here we focus on the dependent structure of mixed-type data through recovered space X . Although there are several ways for investigating the dependencies, we exploit PCA technique. The standard PCA technique is to project high dimensional data to linear subspaces with orthogonal lower dimensions, preserving as much variance in original spaces as possible. The resulting spaces are spanned with some eigenvectors of covariance matrix.

We define $(\sum_{m=1}^p k_m) \times (\sum_{m=1}^p k_m)$ covariance matrix

$$\Sigma = (X - E[X])^T (X - E[X]) \tag{1}$$

The eigen equation is

$$\gamma_l \Sigma = \gamma_l \mathbf{a}_l, \tag{2}$$

where $\gamma_l (1 \leq l \leq \sum_{m=1}^p k_m; \gamma_1 \geq \gamma_2 \geq \dots \geq \gamma_{k_m})$ is a eigenvalue, and \mathbf{a}_l is a corresponding eigenvector. Since it is known that γ_l corresponds to the variance in the space \mathbf{a}_l , the explained variance in the subspace spanned by first maximal N eigenvectors is given by

$$\nu_N = \frac{\sum_{l=1}^N \gamma_l}{\sum_{l=1}^M \gamma_l}, \tag{3}$$

where $0 \leq \nu_N \leq 1$.

Dissimilarity measures

In the procedure introduced in the previous section, we must primarily choose(or define) dissimilarity measure for each description. Important dissimilarity classes that would be applicable to a various type of descriptions are Gowda-Diday[5] and Ichino-Yaguchi[6]. The general definition of dissimilarity measure for compositional data was proposed[5], which consists of the measure for *location*, *span* and *contents*. [6] proposed the *cartesian space model* equipped with the operations of cartesian product and join on which *Generalized Minkowski distance* is introduced. The dissimilarity measures for the intervals, modal, histograms and taxonomical trees would fall into these two classes. As for other generalized dissimilarities measures such as *Generalized Hasudorff distance*, we refer [1, 2, 4]. Here, we show some typical examples.

Example 1.(Interval)

Suppose A, B are intervals, *i.e.*, $A = [l_a, u_a], B = [l_b, u_b]$. l_a, l_b are lower bound of A, B , and u_a, u_b are upper bound of A, B . Then $d_{A,B} = (|l_a - l_b| + |u_a - u_b| - 2|\max(l_a, l_b) - \min(u_a, u_b)|)$.

Example 2.(Distribution)

Suppose A, B are distributions. Wasserstein metric could be adopted. Suppose Y_i and Y_j are random variables of A and B , and corresponding distribution function are F_i and F_j . Then $d(A, B) = \int_0^1 |F_i^{-1}(q) - F_j^{-1}(q)|^2 dq$.

Example 3.(Function)

L_2 -norm is often used. When the function is defined on $L = [a, b]$ ($a, b \in \mathbf{R}$, $a < b$), for the $t \in L$: $d(f_i, f_j) = \int_a^b |f_i(t) - f_j(t)|^2 dt$.

3 Fukushima radio activity level data

Just after the accident of nuclear power plant in Fukushima prefecture in Japan that is triggered by the big earthquake in eastern area of Japan 2011.3.11, the government collect a huge amount of environmental data to evaluate the radio activity levels (<http://radioactivity.nsr.go.jp/en/>). A various types of measurements have been performed from view point of the many-sided. Some measurements are aimed at the spatially exhaustive investigation, or some are aimed at monitoring the levels in long term(or short term).

Comparing the datasets from the many measurements is one of the important task. However, in most case, we have serious difficulties to deal with the datasets since each measurement have been performed independently, the spatially covered ranges are different in each time, the measuring locations are different etc. Besides, we face with the situation such that some types of data are reasonably described by distributional type, however, some other data should be represented as functional type.

Here, we focus on the three datasets, *i.e.*, "Vehicle borne survey", "Air borne survey" and "Monitoring post", that all measure the radio activity level in Fukushima prefecture (http://radb.jaea.go.jp/mapdb_prev/en/). "Vehicle borne survey" measured the levels along with roads by the cars. "Air borne survey" has been conducted to measure the levels from the helicopters over the prefecture. "Monitoring post" is aimed at real time monitoring the levels with sensors, and the sensors are placed more than 3000 points. Those sensors send their data via the Internet every 10 minutes.

We assume the concepts are cities ($w_j \in \Omega$, $j = 1, 2, \dots, 55$) and we give the three descriptions for each city, $Y_1(w_j) = \text{Measurement by vehicle borne survey}$, $Y_2(w_j) = \text{Measurement by air borne survey}$ and $Y_3(w_j) = \text{Measurement by monitoring post}$.

Figure 1 shows the heat maps of radio activity levels, left side shows measurements by air borne around December 2012 and right side shows measurements by vehicle bores around the period from November to December 2012 (<http://ramap.jmc.or.jp/map/eng/>). Black lines in figure 1 left side shows the tracks of air borne where measurements points.

Figure 2 shows the generalized data by each descriptions. We describe Y_1 and Y_2 as distributions, and Y_3 as functions. As for monitoring post data, we extract period from November to end of December 2012. We randomly chose 10 sensors in each city and get mean functions. We would like to know the correlations between descriptions.

One of the way is to use means within concepts. The correlations using means are shown in table 1. However, this approach is not enough to take consider the dispersions within concepts.

We adopt the proposed method. We define the dissimilarity for Y_1, Y_2 by Wasserstein distance, and for Y_3 by L_2 -norm distance. Table 1 and figure 3 and show the result of MDS. From the each eigenvalues, each description could approximately be configured by the dimensions; $k_1 = 1, k_2 = 1$ and $k_3 = 2$. The monitoring post is represented by two dimensions in the euclid space although each of Y_1 and Y_2 is represented by one dimension. The correlations between sets of variables are not straightforward. Then PCA is adopted to analyze the linear dependency (Table 3, Figure 4). The result don't show the clear linearity between descriptions. It indicates that the three measurements performed on a city do not have almost nothing to do with each other. From the figure 4, we could find outliers such as naraha (top right), koriyama (bottom left), kawamata (top left). Investigating their data, narahara includes large values in air borne survey, koriyama includes them in vehicle borne surveys and kawamata does in monitoring post.

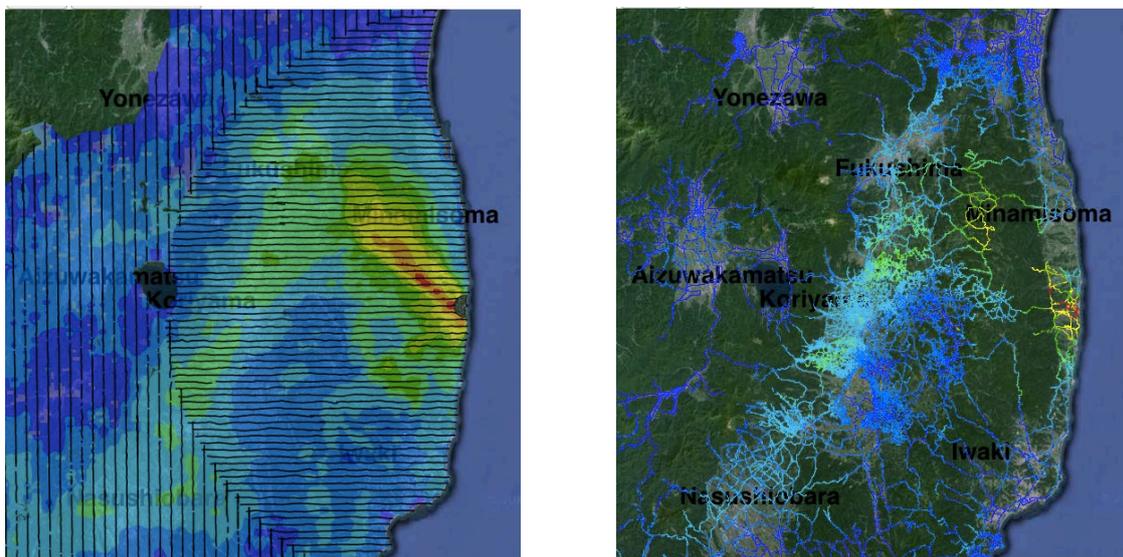


Figure 1: Heat maps of radio activity levels. The left side shows the measurements by air borne survey performed around December 2012. Black lines show the tracks of helicopters where the measurements are performed. The right side shows the measurements by vehicle borne survey. Both survey were performed in the period from November to December 2012.

4 Concluding Remarks

In the paper, we proposed the novel approach to mixed-type data that consist of various data representations such as functional and distributional data. We especially aimed at the correla-

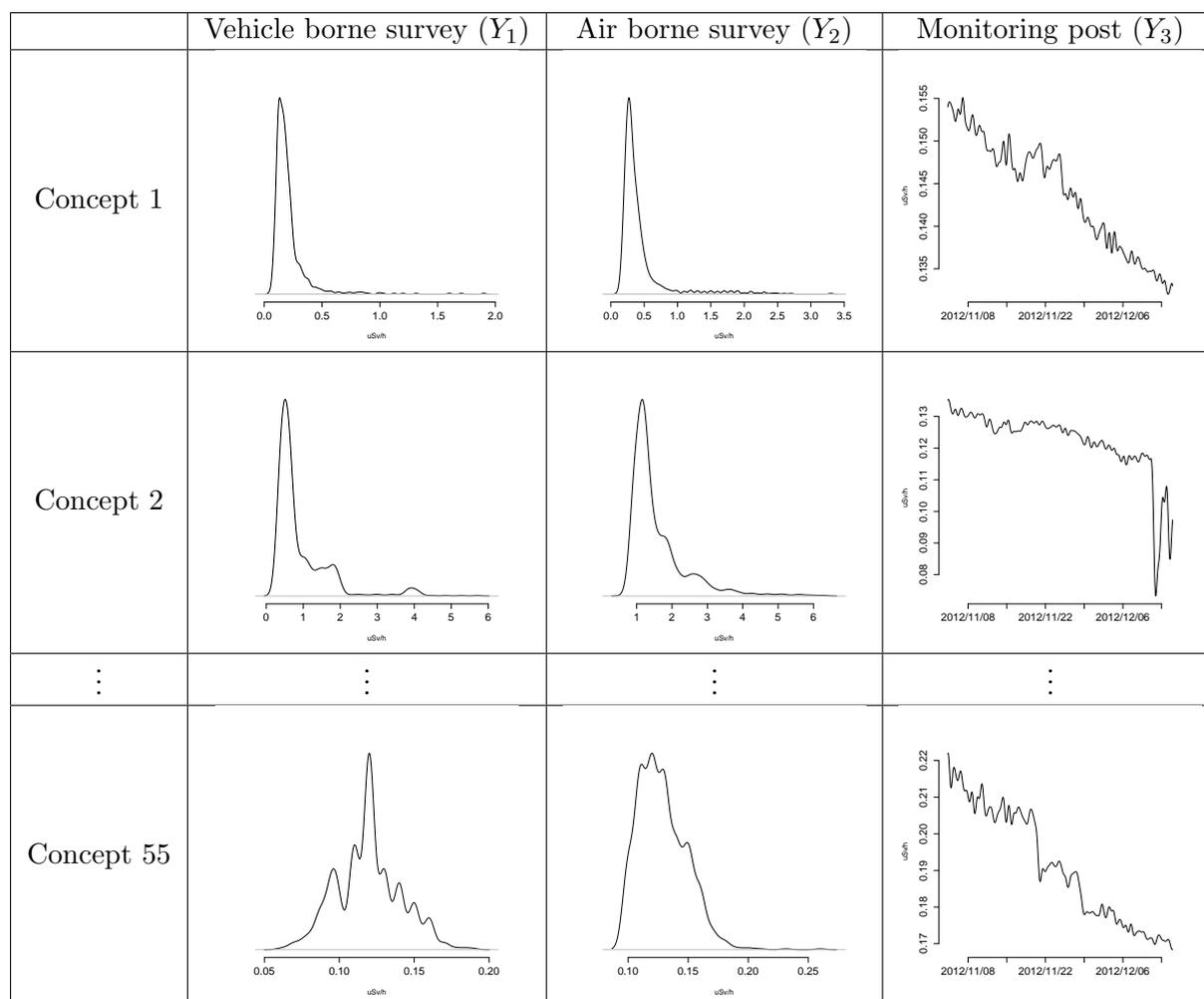


Figure 2: Symbolic data tabs consisted of three descriptions

Table 1: Correlation of the descriptions using means within concepts

	Vehicle borne	Air borne	Monitoring post
VehicleBorne	1.000	-0.041	-0.082
AirBorne		1.000	-0.046
MonitoringPost			1.000

tions between descriptions with different types.

Since it is difficult to directly get correlations of them, we firstly focus on dissimilarities of concepts for each description. Then, from the dissimilarities, we make inference on linear dependency using MDS and PCA technique.

On one hand, we could use means within concepts and estimating correlation coefficient. On the other hand, our approach is based on dissimilarities derived from each data representation, the resulting linearity is considered to include the information of dispersions in original data.

In the actual example, we only use distributional and functional data. However, dissimilar-

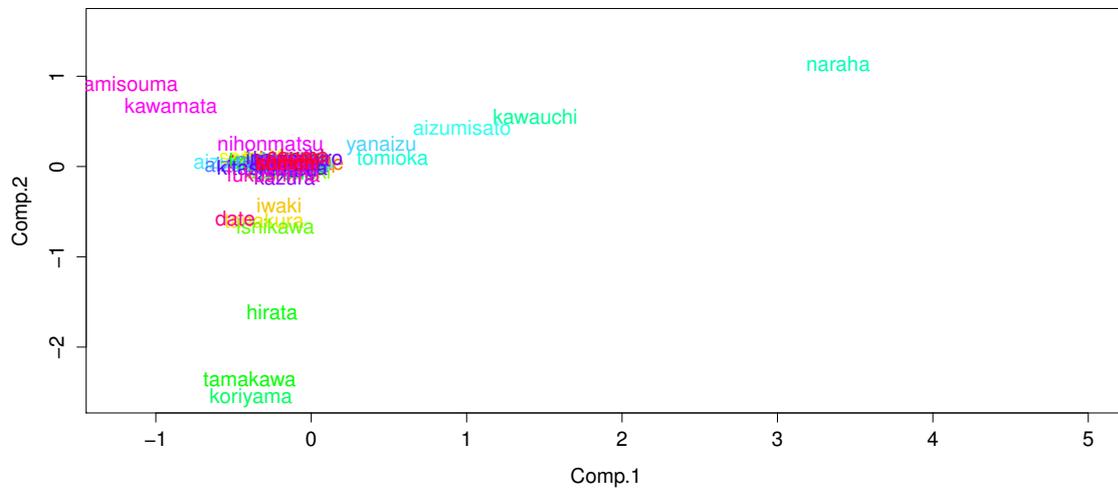


Figure 4: Principal component scores.

dissimilarity matrices[3].

Any Big Data should be structured in some way even when we deal with unstructured data from the viewpoint of the analysis. We'll face at the data with various representations at one time, *i.e.*, mixed-type data at the time. Our approach is effective for such case. As a future work, we study the analysis for various mixed-type data.

Bibliography

- [1] Billard, L. and Diday, E. (2006). *Symbolic Data Analysis: Conceptual Statistics and Data Mining*, Wiley, Chichester.
- [2] Bock, H. H. and Diday, E. (2000). *Analysis of Symbolic Data*, Springer, Berlin Heidelberg.
- [3] Borg, I., Groenen, P. J. F. (2005). *Modern Multidimensional Scaling* (2nd edition), Springer.
- [4] Diday, E. and Noirhomme-Fraiture, M. (2008). *Symbolic Data Analysis and the SODAS Software*, Wiley-Interscience.
- [5] Gowda, K. C. and Diday, E. (1999). *Symbolic clustering using a new dissimilarity measure*, Pattern Recognition, **24**(6), 567–578.
- [6] Ichino, M; Yaguchi, H. (1994). *Generalized Minkowski metrics for mixed feature-type data analysis,* Systems, Man and Cybernetics, IEEE Transactions on, **24**(4), pp.698–708.

Predictive Component-based Multi-block Path Modeling

Pasquale Dolce, *University of Naples “Federico II”*, pasquale.dolce@unina.it
Vincenzo Esposito Vinzi, *ESSEC Business School, Paris*, vinzi@essec.edu
Carlo Lauro, *University of Naples “Federico II”*, clauro@unina.it

Abstract. Partial Least Squares Path Modeling (PLSPM) is a method aimed to model a network of dependence relationships between blocks of variables where each block is summarized by a construct. It is known that PLSPM presents some inconsistencies in terms of coherence with the direction of the relationships specified in the path diagram. Even though PLSPM analyzes networks of dependence relationships among constructs, the estimation process analyzes and amplifies interdependence among them. PLSPM misses to distinguish between dependent and explanatory blocks in the inner model. We propose a more suitable nonsymmetric approach that aims at maximizing the explained variance of the dependent manifest variables in one block given the others (i.e., a redundancy-related criterion). In this perspective, we propose a new algorithm based on extracting and utilizing all the information in the blocks that is relevant to maximizing the explained variances of manifest variables in dependent blocks.

Keywords. PLS Path Modeling, Redundancy Analysis, PLS Regression, Component-Based Approach

1 Introduction

Partial Least Squares Path Modeling (PLSPM) is a method aimed to model a network of dependence relationships between blocks of variables [3] where each block is summarized by a construct, i.e. a linear composite of its own manifest variables.

In order to respect the direction of the relationship specified in the Path diagram (i.e. the path directions), the estimation process should implicitly assume that there is a network of dependence relationships among constructs. However, it is known that PLSPM presents some inconsistencies in terms of coherence with the direction of the relationships specified in the path diagram [5]. In the inner model, each construct is defined as a linear combination of all the connected constructs. Two constructs are connected if there exists a link between the two blocks: an arrow goes from one variable to the other in the Path diagram, independently of the direction. Thus, the directions of the links in the inner model do not play a role in the algorithm

apart from the specific case of the so-called *path weighting scheme* for the inner estimation. In the latter, the path direction is taken into account only in the way the inner weights are computed, but each construct is still defined in the inner step of the algorithm as a function of all the connected constructs.

PLSPM provides composite scores that are as much correlated as possible to each other while being somehow representative of each corresponding block of manifest variables. In the search for optimally correlated constructs, the estimation process amplifies interdependence among blocks, and as a consequence it misses to distinguish between dependent and explanatory blocks.

In order to show such inconsistencies of PLSPM, let us consider the case of three blocks of variables, consisting of two explanatory blocks \mathbf{X}_1 and \mathbf{X}_2 , and a block \mathbf{Y} to be explained. Whether we establish path directions from the two blocks \mathbf{X}_1 and \mathbf{X}_2 to the block \mathbf{Y} or from the block \mathbf{Y} to the two blocks \mathbf{X}_1 and \mathbf{X}_2 , the PLSPM algorithm produces the same results, in terms of weights and loadings linking manifest variables to their constructs.

We propose a more suitable non-symmetrical approach that aims at maximizing the explained variance of the dependent manifest variables in one block given the others, i.e. a new approach based on the optimization of a redundancy-related criterion in a multi-block framework [2].

The methodological core of our approach exploits multivariate explicative statistical methods, such as redundancy analysis [8], PLS2 regression, ridge regression, PCR and so forth so on, in order to inherit their prediction oriented objective [6, 7] as well as their non-symmetrical approach that takes the direction of relationships explicitly into account.

In this perspective, we propose a new algorithm based on extracting and utilizing all the information in the blocks that is relevant to maximizing the explained variances (i.e. improving the prediction) of manifest variables in dependent blocks.

2 Extensions of Redundancy Analysis

Given an explanatory block \mathbf{X} , and a block \mathbf{Y} to be explained, the redundancy analysis proposed by Wollenberg (1977) [8] derives successively orthogonal components of the predictors \mathbf{X} which optimally explain the variance of the \mathbf{Y} -variables. Redundancy analysis as developed by Wollenberg shows how the optimal \mathbf{X} -components should be chosen, but it does not provide \mathbf{Y} -components simultaneously with the \mathbf{X} -components.

Following this argument, Johansson (1981) [2] suggested two alternative transformations for the \mathbf{Y} set which are naturally associated with the transformation for the \mathbf{X} set. Given the weights \mathbf{w}_i defining the i -th \mathbf{X} -component, the corresponding \mathbf{Y} -component is defined via another vector of weights, \mathbf{v}_i , which satisfy desirable orthogonality properties. In particular, the two solutions proposed by Johansson are directly based on the optimal \mathbf{X} -components from redundancy analysis, for which the condition $\mathbf{w}'_i \mathbf{X}' \mathbf{X} \mathbf{w}_j = 0$ is fulfilled. The first solution, based on a least squares approach, satisfies the orthogonality condition between components of the same block so that $\mathbf{v}'_i \mathbf{Y}' \mathbf{Y} \mathbf{v}_j = 0, \forall i \neq j$, but not the orthogonality condition between components of different order across blocks so that $\mathbf{w}'_i \mathbf{X}' \mathbf{Y} \mathbf{v}_j \neq 0, \forall i \neq j$. The second solution, based on a restandardized procedure, fulfills opposite conditions, so that $\mathbf{v}'_i \mathbf{Y}' \mathbf{Y} \mathbf{v}_j \neq 0, \forall i \neq j$, and $\mathbf{w}'_i \mathbf{X}' \mathbf{Y} \mathbf{v}_j = 0, \forall i \neq j$.

We exploit the first solution proposed by Johansson because the specific orthogonality conditions are useful for interpretation purposes and then we generalize the analysis to more than two blocks.

3 The Method

Our approach is based on a multi-step algorithm. At first, we extract as many components as allowed by the ranks of blocks from each exogenous block, based on a redundancy analysis or a series of multivariate PLS2 regressions with respect to adjacent dependent blocks as defined by the inner model relationships. Then, the components of each dependent block are extracted by means of a redundancy analysis or multivariate PLS2 regressions applied to the manifest variables of the dependent blocks with respect to the components of the exogenous adjacent blocks extracted at the previous step. This is then repeated for the subsequent dependent blocks where the sequence is defined by the prediction flow specified in the inner model.

At the end of the second step, the manifest variables of all the blocks are replaced by components. Then, the same steps are applied by replacing the original manifest variables with the newly extracted components so as to update these components. As some of the blocks play a role of both exogenous and endogenous blocks, we apply redundancy analysis or multivariate PLS2 regressions considering these components firstly as dependent variables, then as explanatory variables. The procedure is repeated till convergence of the weights defining the components.

In this new approach, we overcome the theoretical difference between formative and reflective schemes for the measurement model. We only make a distinction between explanatory blocks and dependent blocks, in the sense of redundancy analysis. Furthermore, we do not assume unidimensionality within each block.

Upon converge, we apply a backward selection procedure on the set of the extracted components in order to remove noise but also to simplify interpretability and, finally, to yield final component scores.

We can also integrate useful variable selection features, since we can identify the manifest variables that do not improve the capability of the model in terms of variance explained on the dependent variables and thus implement the method as a sparsifying tool.

In order to assess the quality and validity of results, we provide a new goodness-of-fit index based on redundancy criterion and prediction capability together with a classical bootstrap-based inferential approach.

Finally, we show the functioning of the proposed algorithm (implemented in a R code) through a simulation study and a real data application in the area of healthcare performance.

The performance of the proposed method in terms of explained variability, predictiveness and interpretation is compared to classical PLSPM as well as to other component-based methods such as Generalized Canonical Correlation Analysis [4] and Generalized Structured Component Analysis [1], either on real or artificial data.

Currently, our research work is mainly focusing on fine tuning the methodological and computational aspects of the proposed method. The next step will concern the search for and the specification of a redundancy-based optimizing criterion.

Bibliography

- [1] Hwang, H. and Takane, Y. (2004) *Generalized structured component analysis*. Psychometrika, **69**, 81–99.
- [2] Johansson, J. (1981) *An extension of wollenbergâs redundancy analysis*. Psychometrika, **46**, 93–103.

- [3] Tenenhaus, M. and Esposito Vinzi, V. and Chatelin, M. and Lauro, C. (2005) *Pls path modeling*. Computational Statistics & Data Analysis, **48**, 159–205.
- [4] Tenenhaus, A. and Tenenhaus, M. (2011) *Regularized generalized canonical correlation analysis*. Psychometrika, **76**, 257–284.
- [5] Vittadini, G. and Minotti, S. C. and Fattore, M. and Lovaglio, P. G. (2007) *On the relationships among latent variables and residuals in PLS path modeling: The formative-reflective scheme*. Computational Statistics & Data Analysis, **51**, 5828–5846.
- [6] Wold, H. (1985) *Partial least squares*. in Encyclopedia of Statistical Sciences, S. Kotz and N. Johnson, eds., Wiley, New York **6**, 581–591.
- [7] Wold, S. and Martens, H. and Wold, H. (1983) *The multivariate calibration problem in chemistry solved by the pls method*. In Matrix Pencils, E Kagstrom, B. and Ruhe, A. (Eds.) Springer Berlin Heidelberg, **973**, 286–293.
- [8] Wollenberg, A. L. (1977) *Redundancy analysis an alternative for canonical correlation analysis*. Psychometrika, **42**, 207–219.

Joint analysis of closed and open-ended questions in a survey about the Tunisian revolution

Sadika Rjiba, *Faculté des Sciences Economiques et de Gestion, Tunisie,*
rjibasadika@yahoo.fr

Mireille Summa Gettler, *Université Paris Dauphine, CEREMADE, France,*
summa@ceremade.dauphine.fr

Saloua Benammou, *Faculté des Sciences Economiques et de Gestion, Tunisie,*
saloua.benammou@yahoo.fr

Abstract. In this work we find two major results. The first one is related to the validation of the exploratory analysis of contingency tables built from textual data including both closed and open-ended questions. On one hand we propose “non-lemmatization” as a perturbation of the responses, the effect of which is studied both through Correspondence Analysis and Clustering, on the other hand a bootstrap phase in order to check the quality of the lemmatization phase. The second important result is a new insight about the Tunisian revolution through a survey among young Tunisians. The study reveals that it is not the economical preoccupation as media programs claimed, that first guided the involvement in the revolution but the feeling of dignity about being a Tunisian citizen.

Keywords. Text Mining, Open-ended questions, Correspondence Analysis, Bootstrap, clustering

1 Introduction

The analysis of textual data was developed nearly half a century ago [2]. It implies that we assume the relevance of several dimensions to summarize the “bag of words” and to take into account the textual complexity.

Factorial analysis (FA) of contingency tables can view the lexical profiles by the factorial maps and synthesize them by the chosen compound variables [3]. Automatic classification reveals clusters that may make sense for the case study and the expert will put labels on them according to his linguistic domain.

Visualization [6] is a focal point for exploratory approaches. We can thus associate classification and factorial representation, by projecting the partition clusters onto the factorial mapping. The theoretical results of the positioning of additional elements help in the finding of interpretations for the results of the data mining process, but they also help in the starting of a modeling approach by statistical inference.

An external validation [4] (bootstrap, perturbation, etc.) can be a step to discriminate sub-clouds represented by their centroids which are projections of categories.

In our work, we will introduce structural factors in order to work in a supervised context [6]. We will also analyze the textual data before lemmatization and after lemmatization [8]. We implement these new perspectives in the study of a questionnaire that is detailed here below and that includes open-ended questions. Structural factors and lemmatization will help as validation procedures.

2 The case study questionnaire

The textual data set comes from a survey about what has been termed the Tunisian Revolution in 2011. Data were collected throughout a questionnaire targeted at students or young professionals, men and women of different origins, all enrolled in schools, institutes, universities of Sousse and Monastir. These institutes include about sixty thousands young Tunisians. We focused on students because they were at the very beginning of the revolution process (social networks) and went on being the major actors in it. This survey was conducted during the academic year 2012/2013, by direct face to face interviews. The duration of an interview is approximately about one hour. Sampling is up to now a very difficult challenge in the Tunisian context both for official or private companies. Particularly in our questionnaire which includes open-ended questions, major difficulties are first of all the languages in use (Arabic or French, answers had to be written in French), then the fact that people are globally not positive about participating in surveys and most of all that at that time particularly there was defiance and suspicion in the context of Tunisian post-revolution period.

We tried to be as much as possible in a probability sampling frame: first a geographical cluster sample including heterogeneous respondents and within each regions a stratified random sample including homogeneous respondents in the different grades of each universities. Moreover, few unplanned respondents were included when submitting the questionnaire, in passing on the university campus for example. We eventually collected about 600 students and the size of the final subset is 541 people because open-ended answers in Arabic were not used. Note that the classical sampling procedure may be non-relevant when collecting phrases and words instead of measuring a numerical variable in order to make inference on the target population.

The objective of closed questions is to identify the respondents in terms of age, sex, educational level, marital status, etc. and also throughout their answers about the core of the study (economic situation, etc.). Four open-ended questions were added: two primary questions are in the form of comments related to two closed questions coded with a Likert scale. One of them is “How proud are you about the Tunisian revolution” and the other one is “How proud are you about being a Tunisian citizen”. The third open question is an explanation of the closed question which classifies the major causes of the Tunisian revolution outbreak. The last open ended question is the “What opinion do you have about the Tunisian economical situation” after the Tunisian revolution. To be noted that all questions and all answers have to be in French.

3 Analysis of the closed questions

As the closed questions are categorized we process the data by Multiple Correspondence Analysis (MCA)[7]. The data table T1 includes 541 rows and 23 columns with 89 categories. Five questions are eventually selected as active variables: five categories for the respondent origin, five categories for the feeling upon the Tunisian revolution, five categories for the opinion about the Tunisian economical situation and two dummy variables about the participation to the revolution and the participation to the social networks.

Results and interpretations According to the modified eigenvalues histogram [1] three factorial axes are to be considered. Active and supplementary categories are projected in figure 1 which shows the first factorial plane where only strongly contributing points appear.



Figure 1: Multiple Corresponding Analysis of T1, Axis 1 x Axis 2, variables projections

The maximization of inertia reveals on axis 1 that the main structure among the respondents is an opposition between those who were involved in the Tunisian revolution and social networks versus those who were not involved in the Tunisian revolution nor in social networks. Moreover the participation to the events is linked more to the pride of being a Tunisian citizen than to the economical situation. In fact, the three categories slightly, moderately and very for the variable about the economical Tunisian situation are together projected in the same area, next to the students who have been active during the revolution. The secondary main structure among the respondents is made of those who are extremely negative: not at all proud of being a citizen Tunisian, very bad opinion about the economic situation and not at all proud about the revolution.

Another important element is revealed in the data by the first plane of the MCA: Sousse and Monastir are both next to the negative opinions about the Tunisian events whereas “Tunis and surroundings”, Mahdia and “South and Inlands” areas are supporters of the revolution.

4 Joint analysis of closed and open ended questions

Open-ended questions give opportunity for free responses. According to the sociologist Lazarsfeld

[5] using open-ended questions is essential for the understanding of a set of responses to closed questions. We process a contingency table T2 the rows of which are the 541 respondents, the active columns include non-lemmatized words. Out of the 25306 initial tokens there are 2497 distinct words, 492 have more than 12 occurrences, they are the ones which are kept as active columns of T2. Moreover we define table T3 which includes the same rows as T2 but the active columns are lemmatized words. The lemmatized list of words is a simplified set of words coming from distinct words included in the initial bag of words. We used the Tree Tagger software [14] in order to obtain lemmas: 296 words are finally kept. Example in the case study, a non-lemmatized list in French: “adore”, “aimable”, “aimer”, “aime”, “aimé” and “aimerai” (in English: “adore”, “lovely”, “love”, “loved”, “will love”), lemmatized in French as the single word “aimer” (English lemma “Love”).

Lebart [9] proposes to term as “supervised” the Correspondence Analysis (CA) performed on the table crossing the bag of words extracted out of the open-ended questions and the categories of one closed question. In our case study we process a contingency table T4 the rows of which include 492 words. According to the Multiple Corresponding Analysis of T3 results we selected three closed questions, those which mainly structure the information enclosed in the data set: “How proud are you about the Tunisian revolution”, “How proud are you about being a Tunisian citizen” and “What opinion do you have about the Tunisian economical situation” after the Tunisian revolution. These columns are structural factors kept as active columns in table T4 analysis. This framework can be considered as an extension of the supervised approaches studied by Lebart because there are more than one closed question in the lexical table.

Out of the 25306 initial tokens there are 2497 distinct words, 492 have more than 12 occurrences and finally 296 are kept after lemmatization. According to the multiple corresponding analysis of section 2 we selected three closed questions as columns, those which mainly structure the information enclosed in the data set: ‘actual level of pride about the Tunisian revolution’, ‘actual level of pride about the Tunisian citizenship’ and ‘personal opinion about the post revolution economical situation’. We also perform the Correspondence Analysis of the contingency table T5 which includes the same columns as T4 but with the 296 distinct words out of the bag of words after lemmatization. These two analysis are the ones that are then submitted to the bootstrap procedures.

The Correspondence Analysis on T4 produces a first axis with 19% of inertia and a second axis with 10% of inertia whereas on T5 percentages are respectively 25 and 11. These results are coherent with the decreasing of words in the lemmatized context. The results of major interest are the preservation of the oppositions when interpreting the axes (see figure 2). We verify that first axis in both lemmatized and non lemmatized corpus is about the “Feeling upon the Tunisian revolution”, “very proud” versus “not all proud” and that the second axis as well for both analysis is about “Opinion upon Tunisian economical situation”, “excellent” versus “very bad”.

Greenacre [11] proposes a first attempt of bootstrapping methodology in order to validate a Principal Component Analysis, the so called “Partial Bootstrap”. We performed a bootstrap for the Correspondence Analysis of table T2 and table T3 through the Dtm-vic software[13]. As far as lemmatization can be considered as a disturbance of the units, the comparison of bootstrap

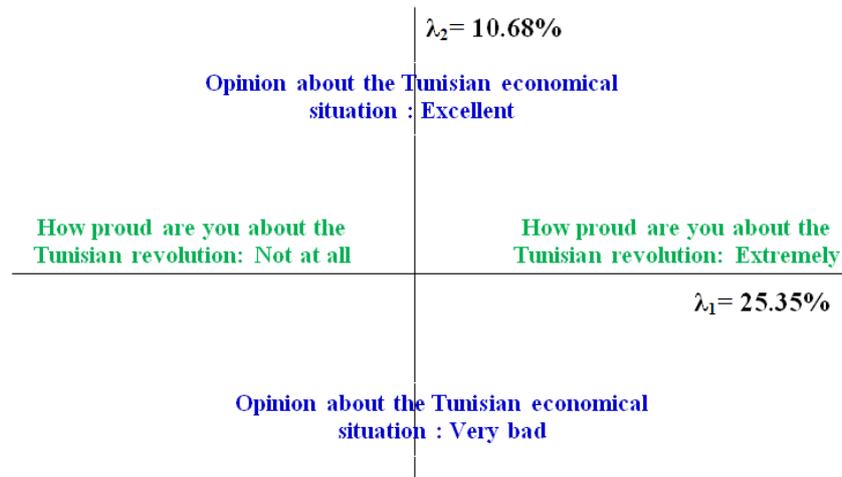


Figure 2: Interpreted first factorial mapping of both CA on tables T4 and T5

results before and after lemmatization may represent a tool to validate the quality of the chosen lemma set. The two bootstraps before and after lemmatization are presented on figure 3 for the example of the following set of words in French : “realiser”, “réalisés”, “réalisé”, “réaliser” (in English: “realise”, “realised”, “realized”, “realize”), French lemma “realise” (English lemma “realize”). In this case study the decreasing of confidence ellipses areas from non lemmatized situation to lemmatized situation, as expected by the mathematical framework for a proper reduction of words through lemmatization, proves that the lemmatization through Tree Tagger software is good for the root “réalise” (English “realize”). It is consistent with the main structures and main interpretations of the initial questionnaire responses.

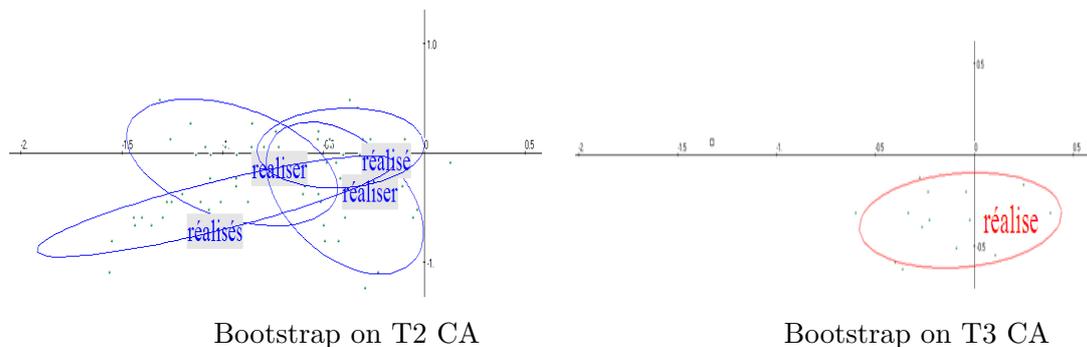


Figure 3: Confidence ellipses on Axis 1 x Axis 2

The next step consists of a comparison analysis between non-lemmatized and lemmatized words clusters.

Statistical comparison of clustering results before and after lemmatization We perform a clustering procedure after the Correspondence Analysis on both tables T4 (non-lemmatized context) and T5 (lemmatized context). The Tandem Analysis [12] proceeds through

the following steps: first Factorial Analysis procedures in order to eliminate noise in the data, only a subset of all CA axes is selected for the next step which is a K-means procedure. Input parameters are set with a large number of clusters (in our case study eighty). A Hierarchical Ascending Cluster (HAC) analysis is then performed on the K-means clusters centroids. By cutting the dendrogram at different levels different partitions are obtained. For each partition statistical indices and geometrical results are collected by projecting the centroids of the related clusters onto the first factorial plane of the two corresponding CAs.

Many indices such as between class-variance and within-class variance help in optimizing the final choice of clusters.

between-class variance We show on Table 1 an example for the partitions into 12 clusters. The inspection of the table that presents the between-class variances before (NL) and after lemmatization (L), allows us to observe that the number of iterations decreases after lemmatization. Initially, seven iterations were needed to achieve the maximum between-class variance (0.589612), after lemmatization only six are needed. A second remark is that the between-class variances, in the case of non-lemmatized data are lower than in the case of lemmatized data.

Itération	NL variance	L variance
0	0.537464	0.443747
1	0.579583	0.480360
2	0.585249	0.484078
3	0.587976	0.485556
4	0.588833	0.486063
5	0.589564	0.486234
6	0.589593	0.486384
7	0.589612	-

Table 1: Between-class variances before and after lemmatization.

This result is the same along the different partitions: the heterogeneity between classes is higher at the end of processing lemmatized data than non-lemmatized data (0.5896 versus 0.4864). The interpretations of the clusters are thus much clearer in the lemmatized context.

Within-class variances Through the double optimization of the between-variances and within-variances we find in our case-study that a 12 clusters partition is the best choice for both non-lemmatized and lemmatized context. These results are confirmed by the comparison of the partitions projections on the first factorial plane both for non-lemmatized and lemmatized context. As the number of clusters changes the visualization is more or less easy to interpret. We present on figure 4 four partition choices (4, 6, 12, 20) illustrated by the centroids on T4 and T5 CAs (left column: non-lemmatized context, right column: lemmatized context). Up to 12 clusters we can easily distinguish the projections: for the twenty classes partition we observe that the images are much more overlapping. The clusters indeed become increasingly close and sometimes contiguous.

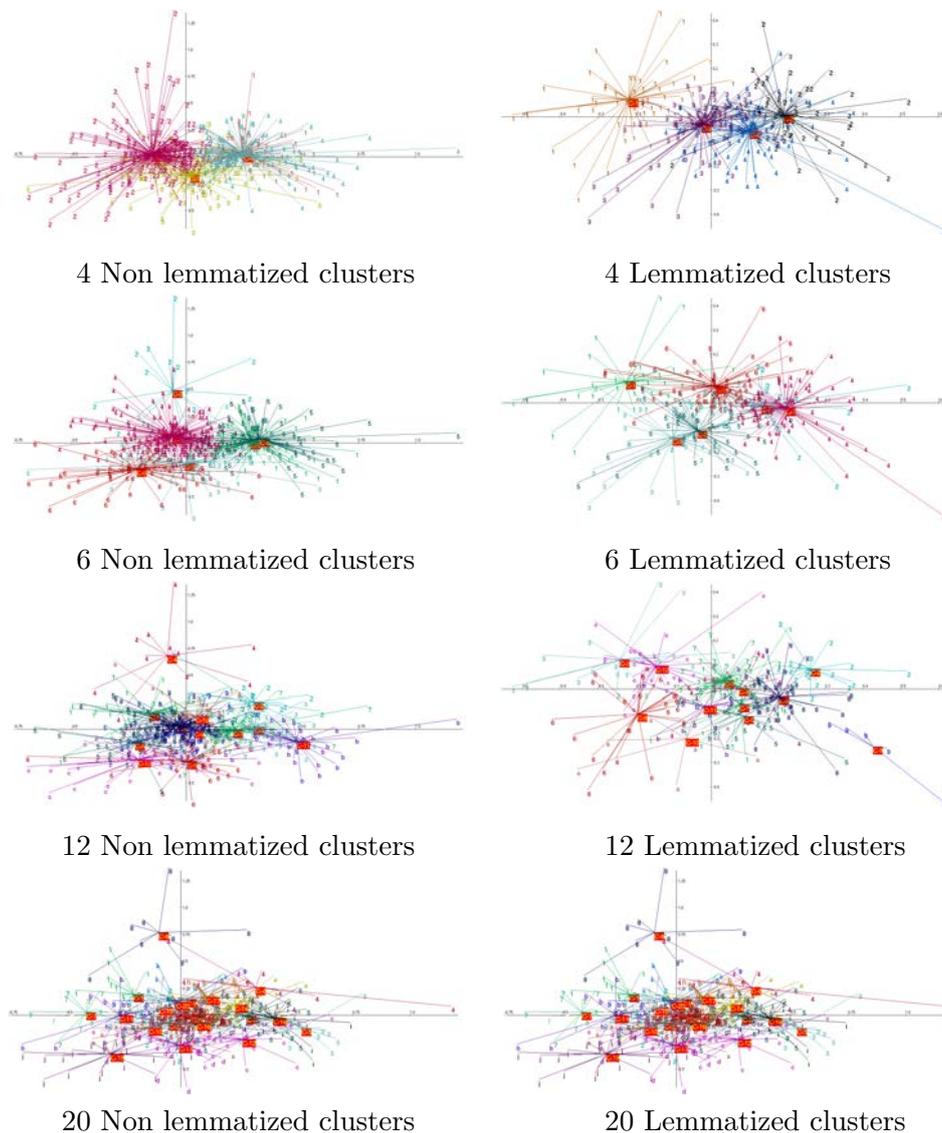


Figure 4: Between variance before and after lemmatization.

Conclusion In this work we find two major results. The first one is related to the validation of the exploratory analysis of contingency tables built from textual data including both closed and open-ended questions. On one hand we propose “non-lemmatization” as a perturbation of the responses, the effect of which is studied both through Correspondence Analysis and Clustering, on the other hand a bootstrap phase in order to check the quality of the lemmatization phase. The second important result is a new insight about the Tunisian revolution through a survey among young Tunisians. The study reveals that it is not the economical preoccupation as media programs claimed, that first guided the involvement in the revolution but the feeling of dignity about being a Tunisian citizen.

Further studies based on the same survey will consist in a complete textual analysis [10] by discovering the repeated phrases and by characterizing the respondents both through the closed

questions and the words coming from the open-ended questions. Moreover two important theoretical issues come out along the study: what does sampling means when the measurements on the population consist in texts, phrases and words, and how is the resampling to be precisely done for example in a bootstrap context.

Bibliography

- [1] Benzécri, J.P. (1979) *Sur le calcul des taux d'inertie dans l'analyse d'un questionnaire*. Les cahiers de l'analyse des données, tome 4, 377–378.
- [2] Lebart, L. and Salem, A. (1988) *Analyse Statistique des Données Textuelles, Questions ouvertes et lexicométrie*. Dunod, Paris.
- [3] Benzécri, J.P. (1981) *Pratique de l'Analyse des Données : linguistique et lexicologie* tome 3, Dunod, Paris.
- [4] Efron, B., Tibshirani, R.J. (1993) *An introduction to the bootstrap: Chapman and Hall*. New York.
- [5] Lazarsfeld, P.E. (1944) *The controversy over detailed interviews: An offer for negotiation*. Public opinion quarterly., vol(8), 38.
- [6] Le Roux, B. (2004) *Structured Data Analysis*. Blasius J. and Greenacre M. Visualization and Verbalization of Data. CRC Computer Science & Data Analysis, Chapman & Hall.
- [7] Le Roux, B. Bonnet, P. and Lebaron, F. (2011) *La notion de champ et l'analyse des correspondances multiples (ACM)*. Séminaire Résidentiel Méthodologique, 3–9.
- [8] Balbi, S. (1998) *Lo studio dei messaggi pubblicitari con l'analisi dei dati testuali*. Quaderni del Dipartimento di Scienze Economiche e statistiche, Linguistica e statistica.
- [9] Lebart, L. (2012) *L'articulation entre exploration et inférence en analyse statistique de textes*. JADT 2012 (11^{ème} Journée Internationale d'Analyse Statistiques des Données Textuelles), 708–715.
- [10] Murtagh, F., Ganz, A. and Reddington, J. (2011) *Semantics from narrative: state of the art and future perspectives*. in Gettler Summa, M., Bottou, L., Goldfarb, B. Murtagh F., Pardoux, C. and Touati, M. Editions Statistical Learning and Data Science, Chapman & Hall/CRC Press, 91–102.
- [11] Greenacre, M.J. (1984) *Cluster analysis in marketing research* Academic Press, London.
- [12] Arabie, P., Hubert, L., (1994) *Factorial k-means analysis for two-way data* Bagozzi, R.P. (Editor), Handbook of marketing research. Blackwell, Oxford.
- [13] Dtm Vic Software. <http://www.dtmvic.com/index.html>.
- [14] Tree Tagger.http://txm.sourceforge.net/installtreetagger_fr.html.

Author Index

- Abbruzzo, Antonino, 499
Adachi, Kohei, 197, 281
Aguilera, Ana M., 151
Aguilera-Morillo, M. Carmen, 151
Ahlgren, Niklas, 265
Aichele, Stephen, 167
Amado, Conceição, 641
Amendola, Alessandra, 187
Anderson, Craig, 343
Arisido, M., 319
Arpino, Bruno, 387
Aslett, Louis, 103
Atkinson, Anthony C., 17
Augugliaro, Luigi, 499
Azarang, Leyla, 143
- Balzanella, Antonio, 483
Bartolucci, Francesco, 531
Batmaz, Inci, 665
Benammou, Saloua, 685
Beninel, Farid, 609
Bhattacharya, Arnab, 103
Bhattacharya, Sakyajit, 523
Biernacki, Christophe, 119
Borysiewicz, Mieczyslaw, 601
- Caballero-Águila, Raquel, 327
Cabral, M. Salomé, 25, 111
Caeiro, Frederico, 289, 545
Cannas, Massimo, 387
Carolino, Elisabete, 127
Casquilho, Miguel, 127, 273
Castillo, Joan del, 45
Catani, Paul, 265
Cattelan, Manuela, 633
Cerioli, Andrea, 17
Chanialidis, Charalampos, 649
- Charkhi, Ali, 205
Charlier, Isabelle, 361
Claeskens, Gerda, 205
Cogan, Peter, 103
Comon, Pierre, 233
Conversano, Claudio, 577
Cordeiro, Clara, 569
- de Uña-Álvarez, Jacobo, 143
Dean, Nema, 343
Derquenne, Christian, 459
Di Lascio, F. Marta L., 491
Di Marzio, Marco, 553
Dolce, Pasquale, 681
Duintjer Tebbens, Jurjen, 1
Durand, Jean-Baptiste, 213, 561
- Ellingson, Leif, 403
Espejo, Rosa M., 539
Esposito Vinzi, Vincenzo, 681
Evers, Ludger, 649
- Fabián, Zdeněk, 657
Fastrich, Bjoern, 177
Fensore, Stefania, 553
Fernández-Pascual, Rosaura, 539
Fernique, Pierre, 561
Ferrari, Davide, 157
Fichet, Bernard, 427
Fienberg, Stephen, 593
Figueiredo, Adelaide, 395, 443
Figueiredo, Fernanda, 395, 443
Fischer, Paul, 9
Fontanella, Sara, 281
Frick, Hannah, 379
Frigau, Luca, 577
Fuchs, Christiane, 625

- Ghattas, Badih, 617
 Ghisletta, Paolo, 167
 Giannerini, Simone, 491
 Giordano, Francesco, 515
 Giuzio, Margherita, 157
 Gomes, M. Ivette, 289, 395, 545
 Gonçalves, M. Helena, 25, 111
 Guédon, Yann, 213, 561

 Hansen, Bruce E., 205
 Hasler, Caren, 241
 Hayashi, Kuniyoshi, 507
 Hermoso-Carazo, Aurora, 327
 Hessen, David J., 435
 Hilbert, Astrid, 9
 Hochreiter, Ronald, 585

 Iacus, Stefano M., 451
 Iizuka, Masaya, 257
 Illner, Katrin, 625
 Irpino, Antonio, 483
 Iyigun, Cem, 665

 Jaroszyński, Marcin, 601

 Kalina, Jan, 1
 Kamakura, Toshinari, 411
 Kashanchi, Faramarz, 97
 Kim, Sujung, 507
 Knefati, Muhammad-Anas, 609
 Konczak, Grzegorz, 35
 Kopka, Piotr, 601
 Kumar, Pranesh, 97
 Kunst, Robert M., 369
 Kurihara, Koji, 507
 Kuroda, Masahiro, 257
 Kyselý, Jan, 467

 Lafuente-Rego, Borja, 61
 Lahiri, Soumendra Nath, 515
 Lauro, Carlo, 681
 Lee, Duncan, 343
 Linares-Pérez, Josefa, 327
 Lourenço, Vanda, 53

 Mai, Tiep, 103
 Manté, Claude, 335
 Marbac, Matthieu, 119

 Matei, Alina, 241
 Matsui, Yusuke, 673
 McLachlan, Geoffrey J, 223
 Mercuri, Lorenzo, 451
 Michel, Pierre, 617
 Mineo, Angelo M., 499
 Mizuta, Masahiro, 673
 Mola, Francesco, 577
 Montanari, Giorgio E., 531
 Monteiro, Andreia, 53
 Mori, Yuichi, 257
 Mozgunov, Pavel, 419

 Neocleous, Tereza, 649
 Neves, Manuela, 569
 Ng, Shu Kay, 223

 O'Ríordáin, Seán, 103
 Okusa, Kosuke, 411

 Padilla, Maria, 45
 Paindaveine, Davy, 361
 Pandolfi, Silvia, 531
 Panzera, Agnese, 553
 Parrella, Maria Lucia, 515
 Paterlini, Sandra, 157, 177
 Patrangenaru, Vic, 403
 Pelle, Elvira, 435
 Perrotta, Domenico, 17
 Petrovic, Sonja, 593
 Phoa, Frederick Kin Hing, 69
 Picek, Jan, 467

 Qiu, Mingfei, 403

 Rabbitt, Pat, 167
 Rajan, Vaibhav, 523
 Reale, Alessandra, 491
 Riani, Marco, 17
 Rinaldo, Alessandro, 593
 Rjiba, Sadika, 685
 Robles Sánchez, Oscar, 103
 Rodrigues, Paulo Canas, 53
 Roetzer, Gernot, 103
 Roldan-Nofuentes, José Antonio, 135
 Rosa, Fátima, 273
 Ruiz-Castro, Juan Eloy, 89

- Ruiz-Medina, Maria Dolores, 539
- Sadeghi, Kayvan, 593
- Sahnoun, Souleyman, 233
- Sakakihara, Michio, 257
- Sakurai, Hirohito, 309
- Saracco, Jérôme, 361
- Schindler, Martin, 467
- Sekiya, Yuri, 249
- Serra, Isabel, 45
- Silva, Margarida, 641
- Souto de Miranda, Manuela, 641
- Stasi, Despina, 593
- Storti, Giuseppe, 187
- Strobl, Carolin, 379
- Summa Gettler, Mireille, 685
- Taguri, Masaaki, 309
- Takahashi, Ryo, 475
- Taneichi, Nobuhiro, 249
- Taylor, Charles C., 553
- Theis, Fabian J, 625
- Toyama, Jun, 249
- Trendafilov, Nickolay, 197, 281
- Valenta, Zdeněk, 1
- Van der Heijden, Peter G.M., 435
- Vandewalle, Vincent, 119
- Verde, Rosanna, 483
- Vilar, Jose Antonio, 61
- Visek, Jan Amos, 351
- Vorkauf, Helmut, 81
- Waldhauser, Christoph, 585
- Wawrzynczak, Anna, 601
- Wilson, Simon, 103
- Winker, Peter, 177
- Wishart, J.R., 299
- Wit, Ernst C., 499
- Yetere Kursun, Ayca, 665
- Zeileis, Achim, 379