Book of Abstracts

COMPSTAT 2014

21st International Conference on

Computational Statistics

hosting the **5th IASC World Conference**

Geneva, Switzerland

August 19–22, 2014

SPONSORS

Exhibitors

Springer (www.springer.com)

Taylor & Francis (www.taylorandfrancis.com)

Wiley (www.wiley.com)

NAG (www.nag.co.uk)

PROGRAMME AND ABSTRACTS

21st International Conference on

Computational Statistics (COMPSTAT 2014)

hosting the 5th IASC World Conference

http://www.compstat2014.org

International Conference Centre Geneva, Switzerland
19–22 August 2014

http://iasc-isi.org

Scientific Program Committee:

Ex-officio:
COMPSTAT 2014 organiser and Chairperson of the SPC: Manfred Gilli
Past COMPSTAT organiser: Erricos John Kontoghiorghes
Next COMPSTAT organiser: Ana Colubi
IASC-ERS Chairman: Vincenzo Esposito Vinzi

Members:
Alessandra Amendola, Ivette Gomes, Sandra Paterlini, Anne Philippe,
Elvezio Ronchetti and Marieke Timmerman

Consultative Members:
Representative of the IFCS: Anuška Ferligoj
Representative of the ARS of IASC: Jung Jin Lee
Representative of ERCIM WG CMS: Stefan Van Aelst

COMPSTAT2014 Proceedings Management Committee:
Manfred Gilli, Gil Gonzalez-Rodriguez and Alicia Nieto-Reyes

ERS-IASC Young Statisticians Competition Jury:
Ana Colubi, Manfred Gilli, Elvezio Ronchetti and Vincenzo Esposito Vinzi

Local Organizing Committee:
Gérard Antille, Manfred Gilli, Dietmar Maringer, Stephan Morgenthaler, Marc Paolella,
Jacques Savoy and Stefan Sperlich

Local Organization Support Committee:
Gerda Cabej

Dear conference participants,

Welcome to the 21st International Conference on Computational Statistics (COMPSTAT 2014), which this year hosts also the 5th IASC World Congress. This edition coincides with the 40th anniversary of a biennial event which started in 1974 in Vienna and has been organized all over Europe. The last three venues were Porto, Paris and Cyprus. The Geneva edition seems to pursue 'the success story' with more than 400 participants and 370 presentations. According to COMPSTAT tradition, proceedings have been edited with 82 papers and 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.

I have to thank the numerous actors who contributed to set up this conference, the members of the SPC, the local team and most importantly all participants who are the soul of the conference.

The next edition of COMPSTAT will take place in Oviedo, Spain in 2016 and will be organized by Prof. Ana Colubi. We wish her the best success and hope that you all will be present.

Enjoy the conference and your stay in Geneva.

Manfred Gilli
(Chair)

SCHEDULE

Tuesday, 19 Aug 2014

8:15 – 9:45		Registration
9:45 – 10:00		Opening
10:00 – 10:45		Keynote: Peter Bühlmann
10:45 – 11:15	Coffee	
11:15 – 13:00	A	Parallel sessions
13:00 – 14:00	Lunch	IASC Executive Committee meeting
14:00 – 15:45	B	Parallel sessions
15:45 – 16:15	Coffee	
16:15 – 18:00	C	Parallel sessions
18:30 – 20:00		Cocktail

Wednesday, 20 Aug 2014

9:15 – 10:45	D	Parallel sessions
10:45 – 11:15	Coffee	10:30–16:30 Posters
11:15 – 13:00	E	Parallel sessions
13:00 – 14:00	Lunch	IASC Council meeting
14:15 – 15:00		Keynote: Anthony Davison
15:15 – 16:45	F	Parallel sessions
17:30 – 21:00		Boat trip

Thursday, 21 Aug 2014

8:45 – 10:00	G	Parallel sessions
10:00 – 10:30	Coffee	
10:30 – 11:15		Keynote: Xuming He
11:30 – 13:00	H	Parallel sessions
13:00 – 14:00	Lunch	ERS BoD meeting
14:00 – 15:45	I	Parallel sessions
15:45 – 16:15	Coffee	
16:15 – 18:00	J	Parallel sessions
18:00 – 19:00		IASC General Assembly
19:30 – 23:00		Conference dinner

Friday, 22 Aug 2014

9:00 – 10:45	K	Parallel sessions
10:45 – 11:15	Coffee	
11:15 – 12:45	L	Parallel sessions
13:00 – 13:15	Room: 1	Best paper award for the ERS-IASC Young Statisticians Competition and Closing
13:15 – 14:15	Lunch	

TUTORIALS, MEETINGS AND SOCIAL EVENTS

TUTORIALS

The tutorials will take place in room 3 during the conference and in parallel with the invited, organized and contributed sessions. The first tutorial is given by Dietmar Maringer (Heuristic methods for model selection and estimation) on Tuesday 19.8.2014 at 14:00–15:45. The second is given by Stefan Van Aelst (Robust estimation, inference and prediction) on Thursday 21.08.2014, 14:00-15:45.

SPECIAL MEETINGS by invitation to group members

- IASC Executive Committee meeting, *room 14*, Tuesday 19th August 2014, 13:00 - 14:00.
- IASC Council Meeting, *room 3*, Wednesday 20th August 2014, 13:00 - 14:00.
- ERS BoD Meeting, *room 1*, Thursday 21th August 2014, 13:00 - 14:00.
- IASC General Assembly, *room 3*, Thursday 21th August 2014, 18:00 - 19:00.

SOCIAL EVENTS

- *Coffee breaks.* Service will last 40 minutes. The location is on the 1st level, same as the registration desks.
- *Light Lunch* will be served at the M!P restaurant in the conference building. You must have your Lunch ticket of the appropriate day in order to attend the lunch.
- *Welcome Reception, Tuesday 19th of August, 18:30-20:00*. The Welcome Reception is open to all registrants (for free) who have reserved a place and non-registered accompanying persons who have purchased a reception ticket. It will take place at the terrace of the M!P restaurant in the conference building. Conference registrants and accompanying persons must bring their reception tickets in order to attend the reception.
- *Boat trip, Wednesday 20th of August, 17:30-21:00*. The excursion is open to all registrants and non-registered accompanying persons who have purchased a ticket. Boarding takes place on *Débarquement de Genève Pâquis, Quai du Mont Blanc 10, Genève* . The boat will not wait for late arrivals.
- *Conference Dinner, Thursday 21th of August, 19:30*. The conference dinner will take place at Restaurant Château de Penthes, Pavillon Gallatin, Chemin de l'Impératrice 18, Pregny/Chambesy. The conference dinner is optional and registration is required. You must have your conference dinner ticket in order to attend the conference dinner. The restaurant is about 15 minutes walking time from the conference venue.

Address of venue

The Conference venue is the Centre International de Conférences, Genève - CICG, Rue de Varembé 17, 1211 Geneva, Switzerland.

Registration and exhibitors

Registration will open on Tuesday 19th August, from 8:15-9:45.The registration desk will be located on the fist floor of the CICG. Exhibitors will be based in the open space of the same floor.

Lecture rooms

Lecture rooms are situated as follows: rooms 3 and 4 on floor 0; room 1 on floor 1;room 13 on floor 2 and rooms 5,6, 19 and 20 on floor 3. See plan hereafter for the exact location. The opening and keynote talks will take place in room 1. Poster presentations will be located in the open space of floor 1 where also the coffee breaks will take place.

Presentation instructions

The lecture rooms will be equipped with a PC and a computer projector. The session chairs should obtain copies of the talks on a USB stick before the session starts (use the lecture room as the meeting place), or obtain the talks by email prior to the start of the conference. Presenters must provide the session chair with the files for the presentation in PDF (Acrobat) or PPT (Powerpoint) format on a USB memory stick. This must be done at least ten minutes before the beginning session. The PC in the lecture rooms should be used for presentations. The session chairs are kindly requested to have a laptop for backup. Please note that Switzerland has plugs/power outlets which differ from those in the rest of Europe and beyond. We cannot provide adapters, so please do not forget to take your adapters if needed. IT technicians will be available during the conference and should be contacted in case of problems. The posters should be displayed only during their assigned session. The authors will be responsible for placing the posters in the poster panel displays and removing them after the session. The maximum size of the poster is A1.

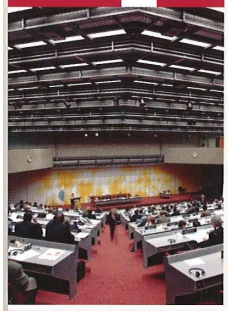
Internet

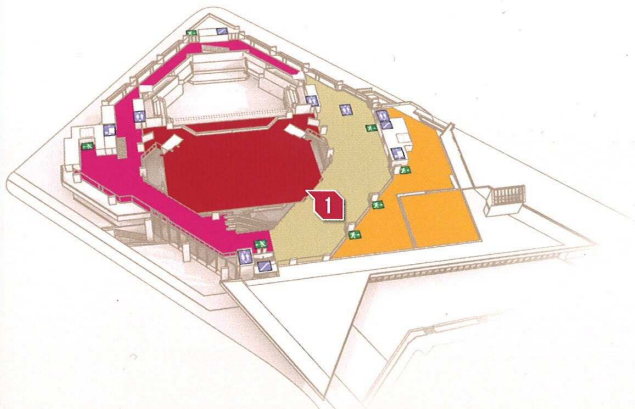
Throughout the conference site there will be wireless Internet connection. The username is: *UNIGE* and password: *COMPSTAT*. There will be a laptop connected to a printer available at the reception desk.

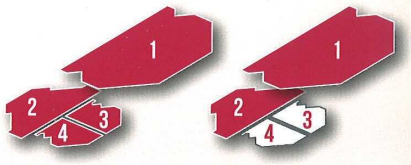
Information and messages

You may leave messages for each other on the bulletin board by the registration desks. General information about restaurants, useful numbers, etc. can be obtained at the stand of the Geneva Tourism Office on the first floor.

1







Room combination and capacity

Rooms 1 + 2 + 3 + 4 > 2'200
Rooms 1 + 2 > 1'660

Auditorium
Room 1 (combinable)
Seating capacity: 940
Can be combined with ground-floor rooms (level 0) for large plenary sessions, accommodating up to 2'200 people.

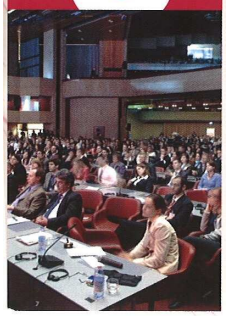
Multi-purpose area
Gallery, usable for exhibition, reception, cloakroom or in addition to the catering space.

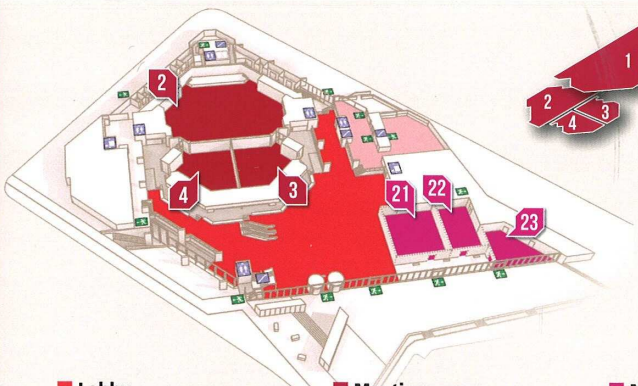
Dining area
1'470 square meters / seats 600 to 800. Large bar.

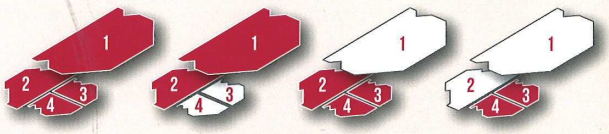
Terrace
800 square meters / seats 200

Emergency exits

0







Room combination and capacity

Rooms 1 + 2 + 3 + 4 > 2'200
Rooms 1 + 2 > 1'660
Rooms 2 + 3 + 4 > 1'220
Rooms 3 + 4 > 456

Lobby
1'280 m². Located at the entrance of the conference centre with direct access to the main meeting rooms and other levels.

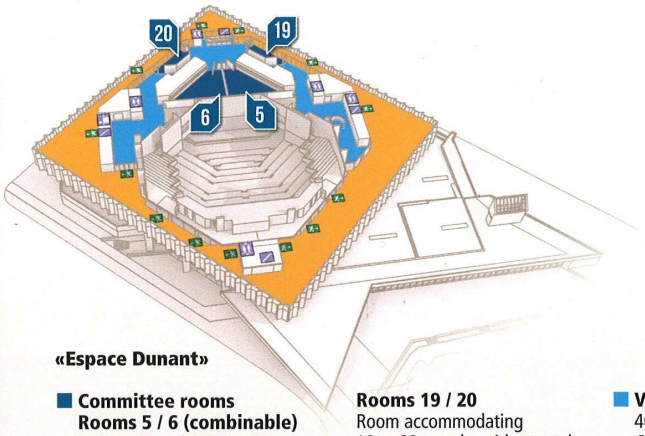
Meeting rooms
Rooms 2 / 3 / 4
(combinable)
3 rooms measuring 227 to 770 square meters, each seating 108 to 720. Can be combined with each other and with room 1, the auditorium on level 1. Simultaneous translation up to 20 languages.

Multi-purpose area
Rooms 21 / 22 / 23
(combinable)
590 square meters
3 rooms seating 26 to 240.
Ideal for exhibitions, cocktails, receptions.
Large bay window opening onto patio area.

«Espace Léman»
Bar offering beverages and snacks, and lounge space.

Emergency exits

3



Room combination and capacity

Rooms 5 + 6 **> 280**

«Espace Dunant»

Committee rooms
Rooms 5 / 6 (combinable)
Seating capacity: 60 to 140 seats. Spacious, offering various combinations, with master control room and six interpreters' booths per room.

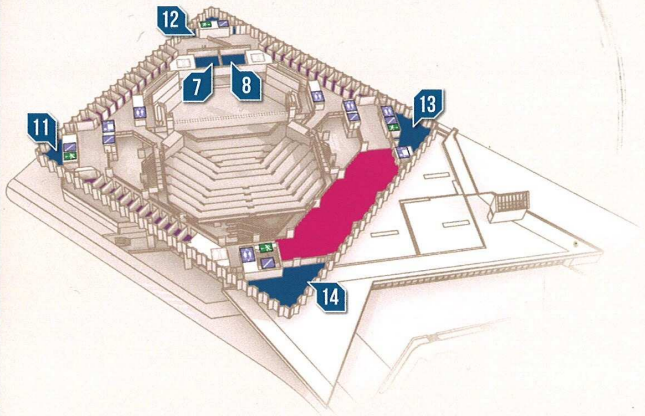
Rooms 19 / 20
Room accommodating 18 to 33 people, with natural light.

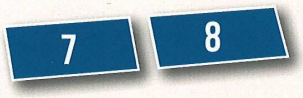
VIP Suite
40 square meters
Combinable, featuring bar, private bathroom and terrace access.

Terrace
800 square meters
Private outdoor patio across from VIP suite.

Emergency exits

2





Room combination and capacity

Rooms 7 + 8 **> 80**

Committee rooms
Rooms 7 / 8 (combinable)
2 rooms with seating capacity for 20 to 35.
Any number of combinations are possible by merging or dividing the rooms into separate units.

Rooms 11 / 12 / 13 / 14
Can accommodate 12 to 70 people.
Can be used as committee rooms or executive suites as you see fit.

Multi-purpose area (Motta)
630 square meters
Suitable for installing landscaped offices, exhibition halls and other purposes.

Offices
29 offices, including 6 double offices, i.e. 60 workstations with direct telephone, fax and Internet access.

Emergency exits

Contents

Tuesday 19.08.2014 10:00-10:45 Room: 1 Chair: Manfred Gilli Keynote Talk 1

Inhomogeneous large-scale data: maximin effects and their statistical estimation

Speaker: **Peter Bühlmann, Swiss Federal Institute of Technology, Zurich**

Large-scale or "big" data usually refers to scenarios with potentially very many variables (dimension p) and very large sample size n. Such data is most often of "inhomogeneous" nature, i.e., neither being i.i.d. realizations from a distribution nor being generated from a stationary distribution. We propose a new methodology for some class of large-scale inhomogeneous data, in terms of so-called maximin effects which optimize performance in the most adversarial constellation. The advocated procedure is computationally efficient and under certain circumstances orders of magnitudes faster than standard penalized regression estimators, and we provide statistical accuracy guarantees for scenarios where n and/or p are large.

Wednesday 20.08.2014 14:15-15:00 Room: 1 Chair: Gil Gonzalez-Rodriguez Keynote Talk 2

Statistics of complex extremes

Speaker: **Anthony Davison, Swiss Federal Institute of Technology, Zurich**

Statistics of complex extremes, used for example in space-time modelling of rainfall, or extreme river levels and flooding, has developed very rapidly in recent years. A wide variety of techniques are used to fit such models, ranging from MCMC methods, to composite likelihood ideas, to non- and semi-parametric models. In this talk I shall give a rapid review of the area, with particular emphasis on computational aspects.

Thursday 21.08.2014 10:30-11:15 Room: 1 Chair: Elvezio Ronchetti Keynote Talk 3

Scalable Bayesian model selection methods

Speaker: **Xuming He, University of Michigan, USA**

Bayesian model selection faces challenges both in theory and in computation when the number of potential covariates p is large. We propose a Bayesian variable selection method for logistic regression that adapts to both the sample size n and the number of potential covariates p with two important features. First, it has strong model selection consistency even when p is large. Second, we propose a new Gibbs sampler that does not require $p2$ operations in each of its iterations. In contrast with the standard Gibbs sampler which requires sampling from a p dimensional multivariate normal distribution with a non-sparse covariance matrix, our new algorithm is much more scalable to high dimensional problems, both in memory and in computational efficiency. We compare our proposed method with several leading variable selection methods through a simulation study to show that our proposed approach selects the correct model with higher probabilities than most competitors. The talk is based on ongoing work with Naveen Narisetty and Juan Shen.

Tuesday 19.08.2014 11:15 - 13:00 Parallel Session A

IS05 Room 3 ADVANCES IN FILTERING METHODS Chair: Elvezio Ronchetti

C1363: **Recursive Bayesian computation**
Presenter: **Nicholas Polson**, University of Chicago, United States
A recursive approach to Bayesian computation is developed. The framework provides state filtering and sequential parameter learning. The methodology is illustrated for mixture Kalman filter and non-Gaussian statistical regularization problems.

C1524: **Robust filtering**
Presenter: **Veronika Czellar**, EMLYON Business School, France
Co-authors: Laurent Calvet, Elvezio Ronchetti
Filtering methods are powerful tools to estimate the hidden state of a state-space model from observations available in real time. However, they are known to be highly sensitive to the presence of small misspecifications of the underlying model and to outliers in the observation process. In this paper, we show that the methodology of robust statistics can be adapted to sequential filtering. We define a filter as being robust if the relative error in the state distribution caused by misspecifications is uniformly bounded for every state and observation. Since standard filters are nonrobust even in the simplest cases, we propose robustified filters which provide accurate state and parameter inference in the presence of model misspecifications. In particular, the robust particle filter naturally solves the degeneracy problems that plague the bootstrap particle filter (Gordon, Salmond and Smith, 1993) and its many extensions. We illustrate the good properties of robust filters in several examples, including linear and nonlinear state-space models.

C1713: **Asymptotic behaviour of approximate Bayesian estimators**
Presenter: **Sumeetpal Singh**, Cambridge University, United Kingdom
Co-authors: Thomas Dean
Although approximate Bayesian computation (ABC) has become a popular technique for performing parameter estimation when the likelihood functions are analytically intractable there has not as yet been a complete investigation of the theoretical properties of the resulting estimators. We give a theoretical analysis of the asymptotic properties of ABC based parameter estimators for hidden Markov models and show that ABC based estimators satisfy asymptotically biased versions of the standard results in the statistical literature.

OS09 Room 4 MULTI-SET AND MULTI-WAY MODELS I Chair: Marieke Timmerman

C1307: **Second-order calibration**
Presenter: **Rasmus Bro**, The Royal Veterinary and Agricultural University, Denmark
Multi-way analysis provides an essential tool for what is called second order calibration – a special field of analytical chemistry. The basic theory behind second order calibration and the unique opportunities it provides in bringing analytical chemistry into more up-to-date state are described. The fundamental problems in PARAFAC (and PARAFAC2) that hinders more widespread application of second order calibration are also described.

C1477: **Multi-way Compressed Sensing and Parallel Computation of the CANDECOMP-PARAFAC Decomposition**
Presenter: **Nikos Sidiropoulos**, University of Minnesota, United States
Co-authors: E. Papalexakis, C. Faloutsos
CANDECOMP-PARAFAC (CP) decomposition algorithms are geared towards small- to moderate-size data that can comfortably fit in main memory. In many emerging applications, this is no longer the case. A parallel algorithm for CP decomposition that is especially well-suited for big data is presented in this talk. The new algorithm is based on parallel processing of a set of randomly compressed, reduced-size 'replicas' of the big tensor. Each is independently decomposed, and the results are joined via a master linear equation per tensor mode. The approach enables parallelism with guaranteed identifiability properties: if the CP decomposition of the big tensor is identifiable from the full data, then the rank-one factors of the big tensor will be exactly recovered from the analysis of the reduced-size replicas. The proposed algorithm is proven to yield memory / storage and complexity gains of order up to IJ/F for a big tensor of size $I \times J \times K$ of rank F with $F \leq I \leq J \leq K$.

C1455: **GINDCLUS: generalized INDCLUS including external information**
Presenter: **Laura Bocci**, Sapienza University of Rome, Italy
Co-authors: Donatella Vicari
Typical two-mode three-way proximity data consist of several symmetric matrices of pairwise proximities between N objects coming from different individual subjects (or equivalently occasions, experimental conditions, or other sources of data). Such three-way data generally contain a wide range of information, which is usually complex and hard to comprehend. Specific methodologies are needed to extract the relevant features and a way to achieve this goal is to synthesize the data by reducing one or two modes to a small number of homogeneous classes. Starting from the individual differences clustering (INDCLUS) model, the present paper presents the GINDCLUS model for the simultaneous reduction of the two modes (objects and subjects) of the proximity data which generalizes INDCLUS by incorporating possible external information on objects and/or subjects. Specifically, we assume that a number of unobserved classes of subjects does exist, each having a different weight structure for the common groups of objects. To better capture the common behaviour of the classes of subjects in evaluating the pairwise proximities, the class-conditional set of weights is in turn linearly related to the external variables on objects and/or subjects. The model is fitted in a least-squares framework and an efficient alternating least squares algorithm is provided.

C1446: **Mixture simultaneous factor analysis for modeling multivariate multilevel data**
Presenter: **Kim De Roover**, KU Leuven, Belgium
Co-authors: Jeroen K. Vermunt, Marieke E. Timmerman, Eva Ceulemans
Multivariate multilevel data consist of multiple data blocks that all involve the same set of variables. For instance, one may think of personality measures of inhabitants from different countries. The associated research questions often pertain to the underlying covariance structure (e.g., which dimensions underlie the individual scores), and whether this structure holds for each data block (e.g., do inhabitants of different countries vary on the same personality dimensions). To answer such research questions, we present mixture simultaneous factor analysis (MSFA) which performs a clustering of the data blocks according to their covariance structure. MSFA, which is the stochastic counterpart of clusterwise simultaneous component analysis, is a multilevel latent variable model with continuous latent variables at the observation level (common factor models) and a discrete latent variable at the block level (according to mixture model). In other words, we assume that the data blocks are sampled from a mixture of multivariate normal distributions

with different covariance matrices (note that existing multilevel mixture models assumed the covariances to be identical). MSFA can be applied by means of latent GOLD.

C1443: **Probabilistic latent feature models with rater differences in feature selection**
Presenter: **Michel Meulders**, KU Leuven, Belgium
Co-authors: Jeroen Vermunt, Iven Van Mechelen

Using a basic latent class model for the analysis of binary three-way three- mode data (i.e. raters who judge whether or not objects have certain attributes) to cluster raters is often problematic because the number of conditional probabilities increases rapidly when extra latent classes are added. To solve this problem we propose a constrained multilevel latent class model in which object-attribute associations are explained on the basis of binary latent features. In addition, rater differences are introduced by assuming that raters only consider each of the latent features with a class-specific probability. For parameter estimation, an EM-algorithm is developed to estimate the posterior mode(s) of the model and a Gibbs sampling algorithm is derived to compute a sample of the posterior distribution. As an illustration, the model is applied to two real data sets: First, the models are used to study individual differences in hostile behavior. Second, the models are used to analyze patient-symptom judgments of different clinicians to study the structure of psychiatric syndromes.

CS42 Room 6 CONTRIBUTIONS TO STATISTICS AND OPTIMIZATION IN FINANCE I Chair: Dietmar Maringer

C1310: **Algorithms for optimal control of econometric models**
Presenter: **Reinhard Neck**, Alpen-Adria Universität Klagenfurt, Austria
Co-authors: Dmitri Blueschke, Viktoria Blueschke-Nikolaeva

A series of algorithms called OPTCON for the optimal control of nonlinear econometric models is presented. They can be applied to obtain approximate numerical solutions to control (dynamic optimization) problems with a quadratic objective function for non-linear econometric models with additive and multiplicative stochastics. The algorithms were programmed in MATLAB and allow for deterministic and stochastic control, the latter with open-loop and passive learning information patterns, both without and with forward-looking expectations. The applicability of the algorithms is demonstrated by an application to a policy problem using a small quarterly macroeconometric model for Slovenia. This shows the convergence and the practical usefulness of the algorithms for a problem of stabilization policy in small-sized econometric models.

C1384: **Automating transition functions: a way to improve trading profits with recurrent reinforcement learning**
Presenter: **Jin Zhang**, University of Basel, Switzerland
Co-authors: Dietmar Maringer

An application of the logistic smooth transition function and recurrent reinforcement learning is presented for designing financial trading systems. A trading system is proposed which is an upgraded version of the regime-switching recurrent reinforcement learning (RS-RRL) trading system referred to the literature. In the proposed system (RS-RRL 2.0) an automated transition function is used to model the regime switches in equity returns. Unlike the original RS-RRL trading system, the dynamic of the transition function in this trading system is driven by utility maximization, which is in line with the trading purpose. Volume, relative strength index, price-to-earnings ratio, moving average prices from technical analysis, and the conditional volatility from a GARCH model are considered as possible options for the transition variable in RS-RRL type trading systems. The significance of Sharpe ratios, the choice of transition variables, and the stability of the trading system are examined by using the daily data of 20 Swiss SPI stocks for the period April 2009 to September 2013. The results from the experiment show that the proposed trading system outperforms the original RS-RRL and RRL trading systems suggested in the literature in terms of better Sharpe ratios recorded in three consecutive out-of-sample periods.

C1544: **Market-based valuations in an emerging market from a biplot perspective.**
Presenter: **Niel Le Roux**, Stellenbosch University, South Africa
Co-authors: Soon Nel, Wilna Bruwer

The benchmark approach is commonly used for analyzing the relationship between market prices and accounting information, as contained in financial statements, and to assess the valuation accuracy of market-based valuations. This study investigates the ability of the market-based models to predict actual share prices in an emerging market. The actual share prices, as reflected on the South African JSE securities exchange, are benchmarked as “correct”. The value estimates obtained from several market-based valuations, are subsequently compared to these benchmarked prices. Sixteen market-based valuations are investigated individually and as composites for the period 2001 to 2010. The chief objective is to optimize the weight allocations of the individual market-based models for inclusion in the composite model. Therefore, the minimization of the median of the valuation errors (MVE) was employed as objective function for optimization. The local optima problem was addressed using (i) random starts and (ii) employing the solution set for minimizing the sum of absolute valuation errors (SAVE) as starting values in the MVE procedure. The latter two-step procedure produced the most accurate results. Given the multi-dimensional nature of the data, the use of biplots and correlation monoplots proved to be particularly adept at displaying the behaviour of the various individual and composite models for market-based valuations.

C1701: **Early warning systems and stress testing for macroprudential policy**
Presenter: **Savas Papadopoulos**, Democritus University, Greece
Co-authors: George Papadopoulos

Policy makers started using extensively macroprudential tools for regulation after the inception of financial crisis. Useful tools for financial stability are early warning systems (EWS) and stress testing (ST). We have been developing an early warning system for Greece and another for EU15 countries based on banking, macro and market variables. We use principal component analysis to reduce the number of the variables and after removing the trend we estimate the probability of default via binary statistical models. We also present econometric models that are used to link macroeconomic and financial variables and transmit the shock imposed in the former to the latter. As a consequence, the performance of these models should be tested in order to ensure that it is satisfactory for stress testing purposes. For that reason we constructed several goodness-of-fit measures that outperform classical ones in terms of interpretability, intuitiveness and comparability among various models.

CS58 Room 19 COMPUTATIONAL STATISTICS I Chair: Giampiero Marra

C1291: **Semiparametric principal components Poisson regression on clustered data**
Presenter: **Kristina Celene Manalaysay**, University of the Philippines Diliman, Philippines
Co-authors: Erniel Barrios

In modelling count data with multivariate predictors, problems with clustering of observations and interdependency of predictors are often encountered. Principal components of predictors are proposed to mitigate the multicollinearity problem. In order to abate infor-

mation losses due to dimension reduction a semiparametric link between the count dependent variable and the principal components is postulated. Clustering of observations is accounted into the model as a random component and the model is estimated via the backfitting algorithm. A simulation study illustrates the advantages of the proposed model over standard poisson regression in a wide range of simulation scenarios.

C1295: **Estimation of isotonic spatio-temporal model with clustering**
Presenter: **Michael Daniel Lucagbo**, University of the Philippines, Philippines
Co-authors: Erniel Barrios

The concept of contagion has gained a foothold in mathematical epidemiology, economics, and environmental science, among others. A data-generating model like those that account for spatial-temporal dependencies is crucial in the understanding and appropriate mitigation of such contagious phenomenon. This study proposes a spatio-temporal model for contagion, with possible clustering of spatial units. It further considers monotonicity of spatial dependence-related covariates in the estimation through the backfitting algorithm embedded with FD-GMM bootstrap and monotone regression via integrated B-splines. Predictive ability of the model is evaluated through a simulation study. The procedure performs best when there are few clusters, the time series is long or the cluster sizes are much greater than the time series length. Furthermore, if the model is properly specified, predictive ability is fairly robust to temporal stationarity.

C1714: **Probability density estimation using Monte Carlo**
Presenter: **Aneta Karaivanova**, IICT-BAS, Bulgaria
Co-authors: Sofiya Ivanovska, Todor Gurov

The problem of reconstruction of unknown density based on a given sample is considered. We present a method for density reconstruction which includes B-spline approximation, least squares method and Monte Carlo method for computing integrals. The error analysis is provided. The method is compared numerically with other statistical methods for density estimation and shows very promising results.

C1556: **The penalized analytic centre estimator**
Presenter: **Keith Knight**, University of Toronto, Canada

In a linear regression model, the Dantzig selector minimizes the L_1 norm of the regression coefficients subject to a bound λ on the L_∞ norm of the covariances between the predictors and the residuals; the resulting estimator is the solution of a linear program, which may be non-unique or unstable. We propose a regularized alternative to the Dantzig selector. These estimators (which depend on λ and an additional tuning parameter r) minimize objective functions that are the sum of the L_1 norm of the regression coefficients plus r times the logarithmic potential function of the Dantzig selector constraints, and can be viewed as penalized analytic centres of the latter constraints. The tuning parameter r controls the smoothness of the estimators as functions of λ and, when λ is sufficiently large, the estimators depend approximately on r and λ via r/λ^2.

C1503: **Testing the hypothesis of exogeneity in regression spline bivariate probit models**
Presenter: **Giampiero Marra**, University College London, United Kingdom
Co-authors: Rosalba Radice, Panagiota Filippou

Bivariate probit models can deal with a problem usually known as endogeneity. This issue is likely to arise in observational studies when confounders are unobserved. We are concerned with testing the hypothesis of exogeneity (or absence of endogeneity) when using regression spline recursive and sample selection bivariate probit models. To this end, gradient and likelihood ratio tests are proposed and their empirical properties investigated and compared with those of the Lagrange multiplier and Wald tests through a Monte Carlo study. The reasonable performance of the likelihood ratio test for the challenging scenarios in which it is not possible to impose an exclusion restriction and the assumption of correct distributional specification does not hold makes it particularly attractive for applied analysis. The tests are also illustrated using two datasets (HIV and health care) in which the hypothesis of exogeneity needs to be tested.

CS63 Room 5 SPARSE MODELS Chair: Maria-Pia Victoria-Feser

C1438: **A Prediction divergence criterion for model selection**
Presenter: **Maria-Pia Victoria-Feser**, University of Geneva, Switzerland
Co-authors: Stephane Guerrier

The problem of model selection is a crucial part of any statistical analysis. In fact, model selection methods become inevitable in an increasingly large number of applications involving partial theoretical knowledge and vast amounts of information, like in medicine, biology or economics. These techniques are intended to determine which variables are “important” to “explain” a phenomenon under investigation. The terms “important” and “explain” can have very different meanings according to the context and, in fact, model selection can be applied to any situation where one tries to balance variability with complexity. For example, these techniques can be applied to select “significant” variables in regression problems, to determine the number of dimensions in principal component analyses or simply to construct histograms. In this respect, we introduce a new class of error measures and of model selection criteria. Moreover, a novel criterion, called the Prediction Divergence Criterion Estimator, is derived from these two classes and we demonstrate that, under some regularity conditions, it is asymptotically loss efficient and can also be consistent. This new criterion is shown to be particularly well suited in “sparse” settings which we believe to be common in many research fields such as Genomics and Proteomics. Our selection procedure is developed for linear regression models.

C1478: **Penalty-free sparse PCA**
Presenter: **Kohei Adachi**, Osaka University, Japan
Co-authors: Nickolay Trendafilov

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.

C1508: **Mixture of regression models with latent variables and sparse coefficient parameters**
Presenter: **Shu-Kay Ng**, Griffith University, Australia
Co-authors: Geoffrey McLachlan
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.

C1642: **Reduced K-means with sparse loadings**
Presenter: **Ryo Takahashi**, Osaka University, Japan
For a data matrix of objects by variables, De Soete and Carroll's 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. 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.

C1665: **Function constrained sparse linear discriminant analysis**
Presenter: **Tsegay Gebrehiwot Gebru**, The Open University, United Kingdom
Co-authors: Nickolay Trendafilov
The high-dimensionality of the data is the major burden of the modern multivariate analysis. Luckily, most of the original variables are irrelevant or redundant, which suggests to look for sparse solutions involving only small portion of them. Additional challenge is the relative small number of available observations in the typical contemporary applications, as gene engineering, climate data analysis, etc. In such situations, the classical linear discriminant analysis (LDA) cannot be applied due to the singularity of the within-group covariance matrix. In this work, we propose a function constrained sparse LDA (FC-SLDA) which does not rely on the original within-group covariance matrix. FC-SLDA finds sparse discriminant functions (vectors) by minimizing their ℓ_1-norm and maximizing their discrimination power simultaneously. Hence, our method selects only few important variables for the purpose of classification. The method is compared with other approaches to high-dimensional-small-sample LDA by performing experiments on both simulated and real data sets. As a result, our method reveals easily interpretable sparse discriminant functions while improving classification performance.

CS67 Room 13 CLASSIFICATION Chair: Gilbert Saporta

C1598: **Supervised classification and experimental designs**
Presenter: **Gilbert Saporta**, CNAM, France
In supervised classification, data generally comes from a basic sampling scheme: simple random, or a stratified sampling plan according to the categories of the response variable. In this paper we deal with the case of a binary response and quantitative predictors where data are collected by means of an orthogonal or near orthogonal design, built on the predictors (eg. fractional factorial, response surface design, ...), often with repeated measurements. In this framework, the estimation and use of the usual classifiers, such as Fisher's LDA, logistic regression, quadratic or SVM, have some specific features we will develop. The presentation will be illustrated with a real data set coming from industry.

C1285: **Multiple-class classification**
Presenter: **Yuan-chin Chang**, Academia Sinica, Taiwan
Co-authors: Yu-Chia Chang
Multiple-class classification is a very common problem in many applications. Many methods have been proposed in the literature, and most of them use binary classifiers as building blocks, such as one-verses-one, one-verses-others based algorithms. For a one-verses-one based algorithm, all possible binary classifiers are constructed and the final result is based on a kind of "voting" scheme. Hence, it will usually suffer from computational efficiency. Although, this kind of computational burden will not happen to the one-verses-others based algorithm, it is usually suffer from the class-imbalanced issues, which usually largely increases the construction of the base binary classifiers. Moreover, the labels of a multiple-class data can be ordinal, categorical or even the mixture of these two types. That makes multiple-class classification problems different from the binary ones. A novel likelihood based multiple-class classification method is presented. It is computational efficient with a very satisfactory performance in terms of several commonly used performance measures. In order to diminish the "curse of dimensionality" when the dimension of the measurement vector is high and size of subjects is relatively small, an ROC curve based dimension reduction method is applied. In addition, the method can fully take advantage of parallel computing techniques. Some basic statistical properties are discussed. Numerical results using both synthesized and real data sets are presented.

C1323: **Optimal ranking in multi-label classification using local precision rates**
Presenter: **Ci-Ren Jiang**, Academia Sinica, Taiwan
Co-authors: Chun-Chi Liu, Xianghong J. Zhou, Haiyan Huang
Multi-label classification is increasingly common in such modern applications as medical diagnosis and document categorization. One important issue in multi-label classification is the existence of statistical difference of classifier scores among different classes. When not accounted for properly, such differences can lead to poor classification decisions on some classes. This issue is addressed by developing a strategy based on a new concept, Local Precision Rate (LPR), under the assumption that classifiers learned for each class are given and corresponding classifier scores for a set of training objects and a set of objects to be classified are available. Under certain conditions, it is shown that transforming the classifier scores into LPRs and making classification decisions by comparing LPR values for all objects against all classes can theoretically guarantee the maximum of precision at any recall rate. It is also shown that LPR is mathematically equivalent to 1-ℓFDR, where ℓFDR stands for local false discovery rate. This equivalence and the Bayesian interpretation of ℓFDR provide an alternative justification for the theoretical optimal property of LPR. A new estimation method is proposed for 1-ℓFDR (or LPR) based on the formulation of LPR, since the original formulation of 1-ℓFDR has limitations for estimation

when data are noisy. Numerical studies are conducted based on both simulation and data to demonstrate the superior performance of LPR over existing methods.

C1511: **Data mining methods for subgroup identification and prediction**

Presenter: **James Chen**, US Food and Drug Administration, United States

Recent advances in biotechnology have generated great interest in the development of statistical methods and data mining techniques to analyze massive amounts of biomedical data. A data set can be expressed in a two-way data matrix with columns representing samples and rows representing measured predictor variables. This presentation describes data mining methods to classify samples into distinct subgroups and identify subsets of variables that can characterize individual subgroups for prediction of future samples. Several data mining techniques will be covered, including clustering and biclustering analysis, classification and prediction, and topic modeling. These techniques will be illustrated by applications to the development of prognostic, predictive models, and predictive enrichment classifiers in personalized medicine, and identification of drug subgroup to adverse event subgroup association in the FDA adverse event reporting system database.

C1674: **A bivariate cost-sensitive classifier performance index**

Presenter: **Luca Frigau**, University of Cagliari, Italy

Co-authors: Claudio Conversano, Francesco Mola

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.

Tuesday 19.08.2014 14:00 - 15:45 Parallel Session B

IS06 Room 4 DEVELOPMENTS IN MULTI-SET MODELING Chair: Marieke Timmerman

C1485: **What is hampering measurement invariance? Detecting outlying variables using clusterwise simultaneous component analysis**
Presenter: **Eva Ceulemans**, University of Leuven, Belgium
Co-authors: Kim De Roover, Jozefien De Leersnyder, Batja Mesquita, Marieke Timmerman
The issue of measurement invariance is ubiquitous in the behavioral sciences nowadays as more and more studies yield multivariate multigroup data. When measurement invariance cannot be established, often only a few items are hampering the invariance in that their loadings differ across groups. Within the multigroup CFA framework, methods have been proposed to trace such non-invariant items, but these methods have some disadvantages in that they require researchers to run a multitude of analyses and in that they imply assumptions that are often questionable. In this paper, we propose an alternative strategy which builds on clusterwise simultaneous component analysis (SCA). Being an exploratory technique, clusterwise SCA assigns the groups under study to a few clusters based on differences and similarities in the covariance matrices and thus on the component structure of the items. Next, non-invariant items can be traced by comparing the cluster-specific component loadings via congruence coefficients, which is far more parsimonious than comparing the component structure of all separate groups. To do this in a consistent way, four heuristics are presented and evaluated in this paper. Afterwards, one can return to the multigroup CFA framework and check whether removing the non-invariant items or removing some of the equality restrictions for these items, yields satisfactory invariance test results. An empirical application concerning cross-cultural emotion data is used to demonstrate that this novel approach is useful and can co-exist with the traditional CFA approaches.

C1409: **Three-way generalized structured component analysis**
Presenter: **Hwang Heungsun**, McGill University, Canada, Canada
Three-way data consist of three different types of entities simultaneously (e.g., subjects, variables, and occasions). These data abound in various fields. Generalized structured component analysis (GSCA) is a component-based approach to structural equation modeling. It involves the specification of three sub-models to specify a structural equation model: measurement, structural and weighted relation models. GSCA combines the three sub-models into a single equation. It estimates parameters by minimizing a single least squares criterion. An alternating least squares (ALS) algorithm is used to minimize the criterion. GSCA provides overall model fit measures which can be used for assessing the variance of the data explained by a given model and for comparing different models. GSCA has been extended to deal with various issues and topics in structural equation modeling. However, GSCA is currently geared for the analysis of two-way data. Thus, in this paper, GSCA is extended to analyze three-way data. This proposed approach is called three-way GSCA. Its model specification and parameter estimation will be discussed. An application to a real data set will be provided to demonstrate the empirical usefulness of the proposed approach.

C1512: **Scaling in ANOVA-simultaneous component analysis**
Presenter: **Marieke Timmerman**, University of Groningen, Netherlands
Co-authors: Huub C.J. Hoefsloot, Age K. Smilde, Eva Ceulemans
In many experiments, data are collected on a large number of variables. Typically, the manipulations involved yield differential effects on subsets of variables. The key challenge is to unravel the nature of those differential effects and the associated subsets of variables. An effective method to achieve this goal is ANOVA-simultaneous component analysis (ASCA). ASCA accounts for the experimental design, and hence explicitly identifies sources of variance due to the experimental manipulations. The core idea of ASCA is to decompose the observed data matrix into a series of additive effect matrices, according to the experimental design, and subsequently perform a Principal Component Analysis on the effect matrices of interest. An issue neglected so far is that the ASCA results heavily depend on the scaling applied. In this paper, we aim at offering the tools for a rational selection of scaling in ASCA. On the basis of a small experimental design and simulated data set, we show the basic principles relevant for scaling. The thus shown principles are generalizable to any, more complicated, experimental design. We illustrate the implications of the resulting guidelines with a real-life metabolomics data set.

TS1 Room 3 TUTORIAL 1 Chair: Manfred Gilli

C1719: **Heuristic methods for model selection and estimation**
Presenter: **Dietmar Maringer**, University of Basel, Switzerland
Model selection and estimation involve optimization to find the ideal solution. More often than not, there is no closed-form solution, and one has to rely on numerical procedures. The underlying optimization problems, however, are rarely well-behaved: the search spaces can be discontinuous, have multiple optima, or pose massive combinatorial problems. This undermines the quality of traditional numerical search and optimization techniques of the kind usually used for optimization problems in statistical and econometric software. Numerical procedures typically consist of many iterations, each of which includes a “creation step” in which one or more new candidate solutions are created, and an “acceptance step” where a decision is made whether the new candidate is replacing the previous solution or not. Traditional methods put all their efforts into the creation step: exploiting the (assumed) properties of the search space and construction an improvement based on the characteristics of the current candidate(s). An actual improvement is then accepted, while failure of finding an improved new solution in one step results in stopping, assuming that the actual optimum has been found. These methods hinge on the creation step: if underlying assumptions about the search space are violated, then they can stop prematurely, or the creation can be misguided and never find the solution. New methods therefore put more emphasis on the acceptance step and relax the construction phase. Typically, non-deterministic elements are added to the guided construction or even replace them, while the acceptance step enforces a preference for improvements, but also facilitates overcoming local optima. Many of these “heuristic methods” draw inspiration from nature; prominent members of this group are evolutionary methods such as genetic algorithms or differential evolution. Although they usually require more iterations than traditional methods, they still can be faster than some traditional methods as the creation step is often much faster, and they are substantially faster than simple Monte Carlo methods. They can tackle much more demanding search problems (i.e., are suitable for more sophisticated statistical and econometric models), and, best of all, they can often be shown to converge to the global optimum. This tutorial presents some of the more popular heuristic methods and demonstrates how they can be used for model selection and estimation, including applications to time series analysis, multivariate data analysis, and robust statistics.

CS60 Room 6 COMPUTATIONAL STATISTICS II **Chair: Philip Reis**

C1616: **Bootstrap-based uncertainty measures for empirical best predictors in generalized linear mixed models**
Presenter: **Daniel Antonio Flores Agreda**, Université de Geneve, Switzerland
Co-authors: Eva Cantoni

The problem of uncertainty estimation in prediction for random effects in mixed models is studied. On a first stage, we review the evaluation and estimation of the mean squared error (MSE) of the empirical best predictor (EBP) based on second-order correct approximations. Resampling procedures, and specially empirical bootstrap, provide an attractive way of estimating MSE by either computing it directly or by providing some bias correction in conjunction with the approximation-based approach. We explore bootstrap schemes in mixed models for hierarchical data, previously used for estimation of the distribution of model parameters, and adapt them to the problem of MSE estimation. We further propose a non-parametric algorithm for estimation the MSE of the empirical best predictors of the random effects, based on the generalized bootstrap for estimating equations adapted for (Gaussian) linear mixed models. We apply this procedure in the framework of generalized linear mixed models and the EBP and we illustrate the properties of our proposal with simulation studies.

C1593: **Inflated bootstrap**
Presenter: **Naoto Niki**, Tokyo University of Science, Japan
Co-authors: Yoko Ono, Hiroki Hashiguchi

By introducing an additional parameter α, an extension of bootstrap is proposed based on a partition distribution, which is conjugate to Pitman sampling formula, and its accompanying urn model. The ordinary non-parametric bootstrap is the case of $\alpha = 0$, while the sampling without replacement is realized as $\alpha = -1$. For $\alpha > 0$, inflation or self-exciting happens in the meaning that those already resampled larger times before are sampled more frequently than less appearing ones. The joint and marginal probability mass functions for the frequency of each observation contained in a resample are given in explicit forms, and the mean and variance of resampling moments are compared with those of sampling distribution from population. In addition, discussion is made on computer algorithms for inflated ($\alpha \geq 0$) and deflated ($\alpha < 0$) bootstraps.

C1529: **The moment-based approximation with a skew-normal polynomial: a numerical comparison**
Presenter: **Hidetoshi Murakami**, National Defense Academy, Japan

Calculating the exact critical value of the test statistic is important in nonparametric statistics. However, to evaluate the exact critical value is difficult when the sample sizes are moderate to large. Under these circumstances, to consider more accurate approximation for the distribution function of a test statistic is extremely important. On testing the hypothesis in a two-sample problem, the Mood test is often used for testing the scale parameters. Various modified Mood tests have been proposed and discussed by many authors over the course of many years. Herein, we performed the moment-based approximation with a skew-normal polynomial in the upper tails for the modified Mood test under finite sample sizes. We then compared the skew-normal polynomial approximation with various approximations and investigated the accuracy of the approximation by a numerical comparison. The table of critical values was extended by using the suggested approximation for the moderate to large sample sizes.

C1547: **Performance of the skew normal distribution type symmetry model for analyzing square ordinal tables**
Presenter: **Kouji Yamamoto**, Osaka University, Japan
Co-authors: Hidetoshi Murakami

For analyzing square contingency tables, many statistical models have been proposed. The normal distribution type symmetry model was considered as an appropriate model if it is reasonable to assume an underlying bivariate normal distribution. As an extension of the model, the skew normal distribution type symmetry (SNDS) model, which may be appropriate for square ordinal tables if it is reasonable to assume an underlying bivariate skew normal distributions, was proposed. It is very important to evaluate the performance of the SNDS model on a variety of conditions because we often face the problem that we cannot assume an underlying normal distribution for analyzing square table data. So, in this talk, we evaluate the performance of the SNDS model in detail. In addition, for a generalization of the SNDS model, i.e., the GNDS model, we also investigate the properties and performance of the model.

C1656: **Cyclic coordinate for penalized Gaussian graphical models with symmetry restrictions**
Presenter: **Antonino Abbruzzo**, University of Palermo, Italy
Co-authors: Luigi Augugliaro, Angelo M. Mineo, Ernst C. Wit

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

CS62 Room 5 HIGH-DIMENSIONAL STATISTICS **Chair: Peter Buehlman**

C1304: **Computation of regularized linear discriminant analysis**
Presenter: **Jan Kalina**, Institute of Computer Science of the Academy of Sciences of the Czech Republic, Czech Republic
Co-authors: Zdenek Valenta, Jurjen Duintjer Tebbens

The focus is on regularized versions of classification analysis and their computation for high-dimensional data. A variety of regularized classification methods have 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 efficient algorithms for their computation.

C1325: **Generalized information criterion for sparse, high-dimensional logistic regression**
Presenter: **Hubert Szymanowski**, Polish Academy of Sciences, Poland
Co-authors: Jan Mielniczuk

A novel method of model selection for logistic regression is presented based on Generalized Information Criterion (GIC) designed for large number p_n of possible predictors. The consistency of a procedure which consists in calculation of GIC for all subsets of predictors of size not exceeding a certain preset number k_n and choosing the subset corresponding to its minimum is discussed. As k_n is allowed to grow with n this generalizes recent known results. In order to cope with computational complexity of the method, the following two-stage modifications are considered. When p_n is smaller than number of observations n we order predictors according to the value of the deviance for the logistic model when all variables except the considered one are taken into account. Then GIC is minimized on subfamily of the first k_n members of thus obtained nested family. For the case when number of predictors exceeds n a

tournament version of the approach is considered. The results of numerical experiments for both modifications in case of artificially generated and real data sets are discussed.

C1505: **Simultaneous dimension reduction and variable selection in modeling high dimensional data**
Presenter: **Joseph Ryan Lansangan**, University of the Philippines, Philippines
Co-authors: Erniel Barrios

High dimensional input in regression is usually associated with multicollinearity and with other estimation problems. As a solution, a constrained optimization method to address high-dimensional data issues is developed. The method simultaneously considers dimension reduction and variable selection while keeping the predictive ability of the model at a high level. The method uses an alternating and iterative solution to the optimization problem, and via soft thresholding, yields fitted models with sparse regression coefficients. Simulated data sets are used to assess the method for both p»n and n>p cases (where p is the number of inputs and n is the number of observations). Results show that the method outperforms the SPCR, LASSO and EN procedures in terms of predictive ability and optimal selection of inputs (independent variables). Results also indicate that the method yields reduced models which have smaller prediction errors than the estimated full models from the PCR or the PCovR. That is, the method identifies a smaller set of inputs that captures the dimensionality (hidden factors) of the inputs, and at the same time gives the most predictive model for the response (dependent variable).

C1514: **Nonparametric estimation of a switching regression model**
Presenter: **Erniel Barrios**, University of the Philippines, Philippines
Co-authors: Ruffy Guilatco

High dimensional data often exhibits multicollinearity leading to unstable estimates of regression coefficients. We postulate a switching regression model with high dimensional predictors. Principal components were extracted to mitigate the multicollinearity caused by high dimensional predictors. The principal components are specified in a nonparametric framework into the switching regression model to mitigate the decline in predictive ability of the model due to lost information in using principal components instead of the original predictors. Simulation studies indicated that nonparametric principal component switching regression model yields better predictive ability than the parametric counterpart while mitigating the adverse effect of multicollinearity. The predictive ability of the model is also robust to the nature of switch (endogenous or exogenous) between the two regimes.

C1594: **Pairwise comparisons among mean vectors in high dimension under non-normality**
Presenter: **Takahiro Nishiyama**, Senshu University, Japan
Co-authors: Masashi Hyodo

The problem of testing mean vectors is a part of many procedures of multivariate statistical analysis, such as multiple comparisons, MANOVA and classification. Recently, when the dimension is larger than the total sample size, many procedures for testing equality of mean vectors are proposed. In this talk, we consider the multivariate multiple comparison procedure among mean vectors for high-dimensional data under non-normality. Especially, we discuss the pairwise comparisons, and for this problem, we propose a Dempster type test statistic and derive its approximate distribution based on the asymptotic normality of concerned statistic. In addition, we construct approximate simultaneous confidence intervals based on this statistic. Using numerical simulations, we evaluate the accuracy of approximation.

CS55 Room 20 CONTRIBUTIONS TO FINANCIAL TIME SERIES Chair: Richard Gerlach

C1490: **Wavelets and their use in the analysis of financial time series**
Presenter: **Milan Basta**, University of Economics - Prague - Faculty of Informatics and Statistics, Czech Republic

Wavelet-based methods are useful methods of time series analysis. We discuss three situations where they could be applied as a part of the analysis of financial time series in particular. Firstly, we discuss the application of wavelets in the analysis of relationships between two financial time series. We show that wavelets are capable of revealing interesting scale-specific relationships. Secondly, we discuss a wavelet-based approach of forecasting financial time series. This approach is based on decomposing the input time series into several components – each associated with a different frequency band – and forecasting each component separately. Thirdly, we illustrate the usefulness of wavelets for detection of jumps in financial time series. Wavelet-based denoising techniques are used for this purpose. In all three cases, our discussion is accompanied by illustrations on artificial datasets as well as financial time series of stock log returns and volatility.

C1468: **Change-point adaptive multiscale decomposition and forecasting of nonstationary financial time series**
Presenter: **Anna Louise Schroder**, LSE, United States
Co-authors: Piotr Fryzlewicz

Financial returns can be modelled as centered around piecewise-constant trend functions which change at certain points in time. Given a set of change points for such a return time series, we propose a new stochastic time series framework with the goal of constructing a short-term forecast of cumulative daily returns using a multiscale decomposition of the trend function. The method introduced here is derived from the observation that a piecewise-constant trend can be reconstructed data-adaptively from a basis of Unbalanced Haar (UH) wavelets. The resulting model enables easy simulation and provides interpretable decomposition of nonstationarity into short- and long-term components. The model permits consistent estimation of the multiscale change-point-induced basis via binary segmentation, which results in a variable-span moving-average estimator of the current trend, and allows for short-term forecasting of the average return.

C1433: **Modeling the term structure with sparse cointegration**
Presenter: **Ines Wilms**, KU Leuven, Belgium
Co-authors: Christophe Croux

Cointegration analysis is often used to investigate the long-run equilibrium relations between several time series in levels. The coefficients of these long-run equilibrium relations are the cointegrating vectors. We provide a sparse estimate of the cointegrating vectors. For this purpose, we combine a penalized estimation procedure for vector autoregressive models with sparse canonical correlations analysis. In this talk, we examine whether the expectations hypothesis of the term structure of interest rates (EHT) holds in practice. The EHT implies that the long-term interest rate can be expressed as an average of current and market-expected future short-term interest rates plus a constant risk premium. Estimating the cointegration space sparsely is especially useful here since the cointegration space has a sparse structure under the EHT. The sparse cointegration approach is applied to monthly US interest rate data from January 1962 to February 2014. In a simulation study we show that the sparse cointegration procedure provides a significantly more precise estimate of the cointegration space compared to the traditional cointegration approach, in particular the Johansen approach.

C1316: **A sparse generalized DCC-GARCH model and its application to international stock returns**
Presenter: **Jianbin Wu**, KU Leuven, Belgium
Co-authors: Geert Dhaene

The dynamic conditional correlation (DCC) GARCH model assumes correlation dynamics that are identical in the cross-section dimension. For large-dimensional systems, this assumption is felt to be too strong. A sparse generalized DCC-GARCH model that allows for idiosyncratic correlation dynamics and correlation spillovers is proposed. The L_1 regularization technique is employed to cope with the issue of high dimensionality and adopt the coordinate descent algorithm to optimize the penalized log-likelihood function. A simulation experiment shows that this sparse model can be estimated and uncover the underlying sparse structure reasonably well. An application to 25 global stock returns from 1994 to 2013 shows that the sparse DCC model outperforms the standard DCC and the diagonal DCC models in and out of sample, both for weekly and daily data. In the weekly data it is shown that the Greek stock market appears to be an important source of correlation spillover.

C1376: **Commonality in liquidity dimensions: a generalized dynamic factor model approach**
Presenter: **Julia Reynolds**, Vienna Graduate School of Finance, Austria

The application of factor model methods to financial data has introduced key insights into asset pricing and risk analysis, particularly in the analysis of liquidity risk. Recent studies have made use of factor analysis to determine commonality in observable liquidity measures, typically corresponding to information about price, volume, and time, to find evidence for an unobservable market liquidity factor. The study builds on this growing literature by addressing two common limitations, and by contributing evidence on the variation in market liquidity commonality over time in response to changing liquidity conditions. First, by extending recent results obtained from the generalized dynamic factor model (GDFM), the limiting assumptions imposed by the use of static factor models are likewise avoided. Secondly, by modeling the time dimension of liquidity as a mean-reversion parameter in an Ornstein-Uhlenbeck process of changes in daily stock prices, high-frequency financial data is used to construct an empirical measure of this oft-overlooked dimension, thus reconciling all three liquidity dimensions in a single GDFM analysis. Lastly, the comparison of GDFM analyses for time periods before, during, and after the 2007-2008 financial crisis provides consistent evidence that commonality in market liquidity dimensions increases following periods of extreme liquidity conditions.

CS80 Room 19 APPLIED STATISTICS AND DATA ANALYSIS I Chair: Shu-Kay Ng

C1545: **Intervention analysis for volatility of stock returns based on the GARCH model**
Presenter: **Masaki Nagashima**, Chuo University, Japan
Co-authors: Norio Watanabe

The GARCH model is commonly used for analyzing returns of average stock prices or stock indices. When economical shocks like the Lehman Crash occurred and times series was influenced, it is expected to investigate the influence of shocks. The intervention analysis based on the ARIMA model by Box and Tiao is a method for such an investigation. However, there are few studies of intervention analysis for volatility of returns, though some researches have done for returns. Therefore we introduce a GARCH model with the intervention term for volatility. The usability of the proposed model is demonstrated by applying to series of TOPIX in japan, FTSE in England, and S&P500 in USA, which include the influence of the Lehman Crash at September 15, 2008.

C1654: **Varying levels of anomie in Europe: a multilevel analysis based on multidimensional IRT models**
Presenter: **Annalina Sarra**, of Chieti Pescara, Italy
Co-authors: Lara Fontanella, Simone Di Zio, Pasquale Valentini

Recent years have seen increased attention to monitor variations of level of social anomie on dependency on micros and macro factors. In this paper we endorse the approach of social anomie as a complex, multidimensional and multilevel phenomenon. To ensure a rigorous measurement of the varying levels of social anomie among European countries, the current study relies on a multilevel multidimensional IRT model. This class of IRT models explicitly integrates the multidimensional IRT representation for measuring the social anomie with the multilevel data structure. The available data set is drawn from Round 5 (Year 2010) of the European Social Survey. In trying to examine social anomie and its determinants we consider items related to different underlying dimensions. Two distinguishing features can be ascribed to our study. Firstly, to our knowledge, no research have to date attempted to carry out a differentiated country comparison of social anomie using the rigorous framework of IRT modeling. Besides, the methodology adopted is desirable to address the challenging issues arising with cross-national comparative studies. The estimates of parameters of interest are obtained within a Bayesian framework, using MCMC algorithms.

C1624: **Anomaly detection in CONTINENTAL data**
Presenter: **Nedjmeddine Allab**, University Pierre et Marie Curie with EDF Labs, France

CONTINENTAL is an extensive tool powered by EDF which, for a hundred electricity demand and climatic scenarios, simulates the optimal production plan in Europe. The simulation outputs 54 time series, indicating when and which production units must be turned on with their relevant costs. This helps in predicting the price of electricity, choosing what power station to invest in and negotiating the CO2 emission licensing. In this communication, we explore CONTINENTAL data from the point of view of anomaly detection to automatically identify significant peaks, regime or volatility changes in prices, odd electricity demands and blackouts. Several statistical techniques are implemented into one procedure to identify different types of anomalies such as abrupt peaks, variance or mean changes. We consider that a series in the normal regime is a stationary ARMA process. An anomaly is defined by its initial impact and the way it propagates through the series. The procedure iteratively estimates the series parameters, identifies one anomaly and removes it. We run our procedure both on simulated anomalies and on real CONTINENTAL data, namely the demand and price of electricity after eliminating the tendency and seasonality, determining the ARMA order, applying white noise tests and selecting the best ARMA estimator.

C1370: **Unravel: a method and a program to analyze contingency tables, unveiling confounders**
Presenter: **Helmut Vorkauf**, retired, Switzerland

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

C1509: **Goodness-of-fit measures in linear mixed-effects models**
Presenter: **Nadege Jacot**, University of Geneva, Switzerland
Co-authors: Eva Cantoni, Paolo Ghisletta

Linear mixed-effects models are widely used in modeling repeated-measures data in different disciplines, such as psychology, biomedicine, pharmacokinetics, forestry, agriculture or economics. In order to check the goodness-of-fit of a model, several measures have been proposed in the literature both in a frequentist and in a Bayesian framework. We select a large set of these indices, in particular those that can be interpreted in an absolute sense without being compared with another model. Such indices allow us to quantify the quality of the model at hand and some additionally allow us to select the most appropriate model among those proposed.

Through an extensive simulation study, we evaluate the sensitivity of the selected measures to the modification of the values of some parameters of a varying-intercept and varying-slope model. We also check how well one can choose the correct model, from a series of nested alternative models, based on these indices. Finally we illustrate the use of these indices in a famous data example considering dental measurements on children.

CS76 Room 13 MULTIVARIATE STATISTICS I Chair: Peter Filzmoser

C1389: **Local analysis of structural anisotropy**
Presenter: **Jaromir Antoch**, Charles University, Czech Republic
Analysis of images often includes measurement of structural anisotropy or directional orientation of object systems. This contribution deals particularly with the problem of estimating the main orientation of fiber systems (filament structures). The methods considered are based on the two-dimensional discrete Fourier transform combined with the method of moments. It is suggested to abandon the currently used global, i.e. all-at-once, analysis of the whole image, which typically leads to just one estimate of the characteristic of interest, and advise to replace it with a "local analysis". This means to split the image into (many) small, non-overlapping pieces, and to estimate the characteristic of interest for each piece separately and independently of the others. As a result many estimates of the characteristic of interest are obtained, one for each sub-window of the original image, and - instead of averaging them to get just one value – it is suggested to analyze the distribution of the estimates obtained for the respective sub-images. Proposed approach seems especially appealing when analyzing nanofibrous layers, nonwoven textiles, filament structures and dynamic of cell evolution, etc., in which may often exhibit quite a large anisotropy of the characteristic of interest.

C1540: **Sparse exploratory factor analysis**
Presenter: **Sara Fontanella**, The Open University, United Kingdom
Co-authors: Nickolay Trendafilov, Kohei Adachi
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.

C1571: **Tracking changes in scientific bibliography: a novel statistical approach using a multifactorial analysis**
Presenter: **Daria Micaela Hernandez Ramirez**, Universitat Politecnica de Catalunya, Spain
Co-authors: Monica Becue-Bertaut
The aim is to provide a global vision of the scientific publications related with the Systemic Lupus Erythematosus (SLE), taking as starting point their abstracts. Over the years, these abstracts have been evolving towards higher concern and complexity, which makes necessary the use of sophisticated statistical methods, for creating a meaningful big picture of the phenomenon. Textual analysis, correspondence analysis, multiple factor analysis for contingency tables and characteristic words are used for extracting relevant information from 506 abstracts obtained from 115 different journals with high-impact factors and covering a 18 years period. Results show the most relevant topics related with the SLE, how the vocabulary has been evolving over the time, and detect the pioneers papers. These statistical methods can be replicated on any scientific field and tackle the complexity on big textual datasets.

C1487: **Same, but better: comparing centers-coupled-with-radii and vertices principal component analyses for symbolic data**
Presenter: **Anuradha Roy**, The University of Texas at San Antonio, United States
Co-authors: Chengcheng Hao, Yuli Liang
Centers principal component analysis (C-PCA) and vertices principal component analysis (V-PCA) are two methods that are commonly used to explain the total variances of a set of interval variables. V-PCA computes the principal components using all vertices of the hyper-rectangle defined by the intervals of all variables whereas, C-PCA computes the principal components using the centers of only two vertices having lower and upper bounds of the interval variables respectively. Existing simulation study shows that C-PCA and V-PCA work almost similarly, but C-PCA works uniformly slightly better. In this article, this conclusion has been verified theoretically. The vertices belong to the same observation unit are not independent, rather equally correlated. We assume that the correlation matrices between any two vertices of the same observation unit are same. In particular, when the variance-covariance matrix of all the vertices for the same observation unit has the blocked compound symmetry (BCS) covariance structure, we show that V-PCA is mathematically equivalent to centers-coupled-with-radii PCA (CCR-PCA), and propose a more efficient method, CCR-PCA, which provides greater explanatory power. The proposed method is implemented with a real data set.

C1578: **Outlier detection in interval data**
Presenter: **Peter Filzmoser**, Vienna University of Technology, Austria
Co-authors: Paula Brito, A. Pedro Duarte Silva
The aim is to identify outliers in multivariate observations that are consisting of interval data. Interval data occur in multiple different situations, e.g., when describing ranges of variable values, as daily stock prices, or from the aggregation of huge data bases, when real values describing the individual observations result in intervals in the description of the aggregated data. Each variable of the multivariate interval data information can be represented by a mid-point and a range. Parametric models have been proposed which rely on multivariate normal or skew-normal distributions for the mid-points and log-ranges of the interval-valued variables. Different parameterizations of the joint variance-covariance matrix allow taking into account the relation that might or might not exist between mid-points and log-ranges of the same or different variables. Here we use the estimates for the joint mean and covariance for multivariate outlier detection. The Mahalanobis distances based on these estimates provide information on how different individual multivariate interval data are from the mean with respect to the overall covariance structure. A critical value based on the chi-square distribution allows distinguishing outliers from regular observations. The outlier diagnostics is particularly interesting when the covariance between the mid-points and the log-ranges in restricted to be zero. Then, Mahalanobis distances can be computed separately for mid-points and log-ranges, and the resulting distance-distance plot identifies outliers that can be due to deviations with respect to the mid-point, or with respect to the range of the interval data, or both.

Tuesday 19.08.2014 16:15 - 18:00 Parallel Session C

OS10 Room 6 LONGITUDINAL EXPLORATORY DATA MINING Chair: John J. McArdle

C1341: **Introduction to longitudinal exploratory data mining**
Presenter: **John McArdle**, University of Southern California, United States
Co-authors: Gilbert Ritschard, Paolo Ghisletta, Andreas Brandmaier
Longitudinal data is often encountered and ideas about how this data should be analyzed are known. Nevertheless, it is not often to know enough to fit models with multiple Factors that are Invariant over Time (MFIT) with Latent Change Scores (LCS). These "confirmatory" techniques are indeed quite elegant but they do not allow us enough freedom to be as wrong as we often are. The techniques discussed allow us to proceed with the longitudinal Data analysis. The discussion will be based on four previous talks: "Introduction to Longitudinal Exploratory Data Mining", "Experiences with some Longitudinal Exploratory Data Mining Problems", "Survival Analysis as an Longitudinal Exploratory Data Mining Device" and "Quantifying Variable Importance with Structural Equation Model Trees".

C1366: **Variable importance in structural equation model forests**
Presenter: **Andreas Brandmaier**, Max Planck Institute for Human Development, Germany
Co-authors: John J. Prindle, John J. McArdle, Ulman Lindenberger
Structural equation model (SEM) trees allow researchers to discover predictors and their interactions explaining heterogeneity with respect to parameter estimates in a hypothesized SEM. They provide a mean to detect variables that had not been modeled but should potentially be included in the model to increase its predictive power. A set of decision trees inferred from bootstrapped data is typically referred to as random forest. Such forests are an instance of random subspace methods, which induce controlled variation by repeatedly sampling variables and/or observations from the original dataset to increase robustness of the resulting aggregated model. Following the logic of random forests, forests of SEM trees can be employed to determine rankings of variable importance for a parametric SEM, and thus, inform researchers in the process of model modification. Like SEM trees, SEM forests offer a non-parametric search to refine initial hypotheses in the form of parametric SEM by combining theory-guided and data-driven modeling into theory-guided exploration of the model space.

C1465: **Longitudinal data mining to predict survival in a large sample of adults**
Presenter: **Paolo Ghisletta**, University of Geneva, Switzerland
Co-authors: Stephen Aichele, Pat Rabbitt
We applied data mining techniques to explore survival in a sample of 6203 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.

C1404: **Experiences with some longitudinal exploratory data mining problems**
Presenter: **Gilbert Ritschard**, University of Geneva, Switzerland
Co-authors: Matthias Studer, Emmanuel Rousseaux
Event sequence mining has been successfully used in many different fields ranging from device quality control to the analysis of web log and longitudinal customer behaviors. Here we are primarily concerned with the exploration of sequences of life events such as family events or occupational events. First we consider frequent subsequences and address the importance of taking time and content constraints into account in life course analysis and stress the interest of frequent maximal subsequences. Secondly, we discuss the identification of subsequences that best discriminate between groups as between birth cohorts or between women and men. Lastly we address the question of finding association rules between subsequences, where an association rule is a couple of subsequences such that when the first one is observed in a sequence it increases the likeliness to see the second one occur in the same sequence. The addressed methods are accessible through our TraMineR R toolbox and are illustrated using data describing Swiss cohabitational and occupational trajectories.

CS15 Room 4 REGRESSION MODELS I Chair: Ana M. Aguilera

C1284: **Model averaging for Poisson regression**
Presenter: **Jianhong Zhou**, City University of Hong Kong, China
Model averaging is a desirable approach to deal with model uncertainty, which, however, has rarely been explored for Poisson regression. A model averaging procedure based on an unbiased estimator of the expected Kullback-Leibler distance for the Poisson regression is proposed. Simulation study shows that the proposed model average estimator outperforms some other commonly used model selection and model average estimators in some situations. The proposed methods are further applied to a real data example and the advantage of the method is demonstrated again.

C1299: **Frequentist model averaging for ordered probit and nested logit models**
Presenter: **Longmei Chen**, City University of Hong Kong, China
Co-authors: Tze-Kin Alan Wan, Kwok Fai Geoffrey Tso
Ordered probit and nested logit techniques are two extensively applied quantitative techniques in many research areas. When applying this two techniques, it is quite usual that the practitioners choose a final "best" model from a multitude of models by various combinations of regressors, and then proceed their analysis as if the "best" model had been decided upon a priori. However, model selection has been criticized for its ignorance of the uncertainty associated with the selection process and the risk of taking poor models. The idea of combining models as an alternative to selecting a single model is advocated, and extant different weight choice mechanisms for the ordered probit model and the nested logit model are compared. A Monte Carlo study will show that model averaging delivers more accurate forecasts than model selection and the leave-one-out cross validation mechanism based on the

probability forecasting errors shows distinct advantage in producing more accurate forecast results and the performance of equal weighted method can be improved by deleting quite poor models before combining.

C1372: **Inflated discrete beta regression model for Likert and discrete rating scale outcomes**
Presenter: **Cedric Taverne**, Université catholique de Louvain, Belgium
Co-authors: Philippe Lambert

Discrete ordinal responses such as Likert or rating scales are regularly proposed in questionnaires and used as dependent variable in modeling. The response distribution for such scales is always discrete, with bounded support and often skewed. In addition, one particular level of the scale is frequently inflated as it cumulates respondents who invariably choose that particular level (typically the middle or one extreme of the scale) without hesitation with those who chose that alternative but might have selected a neighboring one. The inflated discrete beta regression (IDBR) model addresses those four critical characteristics that have never been taken into account simultaneously by existing models. The mean and the dispersion of rates are jointly regressed on covariates using an underlying beta distribution. The probability that choosers of the inflated level invariably make that choice is also regressed on covariates. Simulation studies suggest that the IDBR model produces more precise predictions than competing models. The ability to jointly model the location and dispersion of an ordinal response, as well as to characterize the profile of subjects selecting an “inflated” alternative are the most relevant features of the IDBR model. It is illustrated on a set of questions from the European Social Survey.

C1543: **Weighted estimation for prediction in the linear model**
Presenter: **Mark Hannay**, University of Geneva, Switzerland
Co-authors: Elvezio Ronchetti

Working in the linear model y=Xβ, while using the OLS estimate of β, the mean square error (MSE) of prediction is independent of the size of the β_i. We propose a new estimator, which performs better in the cases of some small β_i, while bounding the maximal MSE of prediction. This estimator depends on a parameter γ, which is affinely related to the maximal MSE of prediction allowed. We propose to pick γ based on cross validation. In a simulation study, we compare the MSE of prediction of this new estimator with that of the OLS estimator from a model found by backward selection.

C1500: **Copula regression spline models for binary outcomes**
Presenter: **Rosalba Radice**, Birkbeck University of London, United Kingdom
Co-authors: Giampiero Marra, Malgorzata Wojtys

A framework for estimating the effect that a binary treatment has on a binary outcome in the presence of unobserved confounding is introduced. The method is applied to a case study which uses data from the 2008 Medical Expenditure Panel Survey and whose aim is to estimate the effect of private health insurance on health care utilization. Unobserved confounding arises when variables which are associated with both treatment and outcome are not available. Also, treatment and outcome may exhibit a dependence that cannot be modelled using a linear measure of association, and observed confounders may have a non-linear impact on the responses. The problem of unobserved confounding is addressed using a two-equation structural latent variable framework, where one equation describes a binary outcome as a function of a binary treatment whereas the other equation determines whether the treatment is received. Non-linear dependence between treatment and outcome is dealt with by using copula functions, whereas covariate-response relationships are flexibly modelled using a spline approach. Related model fitting and inferential procedures are developed.The findings of our empirical analysis suggest that the issues discussed in this paper are present when estimating the impact of private health insurance on health care utilization, and that useful insights can be gained by using the proposed framework.

CS17 Room 5 MULTIVARIATE STATISTICS II Chair: Henry Wynn

C1480: **On the estimator of the canonical parameter in the exponential family**
Presenter: **Haruhiko Ogasawara**, Otaru University of Commerce, Japan

Asymptotic cumulants of the maximum likelihood estimator of the canonical parameter in the exponential family are obtained up to the fourth order with the added higher-order asymptotic variance. In the case of a scalar parameter, the corresponding results with and without studentization are given. These results are also obtained for the estimators by the weighted score, especially for those using the Jeffreys prior. The asymptotic cumulants are used for reducing bias and mean square error to improve a point estimator and for interval estimation to have higher-order accuracy. It is shown that the kurtosis to squared skewness ratio of the sufficient statistic plays a fundamental role.

C1456: **Confidence bands for impulse responses: Bonferroni versus Wald**
Presenter: **Peter Winker**, University of Giessen, Germany
Co-authors: Anna Staszewska-Bystrova, Helmut Luetkepohl

In impulse response analysis estimation, uncertainty is typically displayed by constructing bands around estimated impulse response functions. These bands may be based on frequentist or Bayesian methods. If they are based on the joint distribution in the Bayesian framework or the joint asymptotic distribution possibly constructed with bootstrap methods in the frequentist framework, often individual connfidence intervals or credibility sets are simply connected to obtain the bands. Such bands are known to be too narrow and have a joint confidence content lower than the desired one. If instead the joint distribution of the impulse response coefficients is taken into account and mapped into the band it is shown that such a band is typically rather conservative. It is argued that a smaller band can often be obtained by using the Bonferroni method. While these considerations are equally important for constructing forecast bands, we focus on the case of impulse responses in this study.

C1640: **Confidence interval of log odds ratio for the posterior probabilities for large dimension**
Presenter: **Takayuki Yamada**, Nihon, Japan
Co-authors: Tetsuto Himeno, Tetsuro Sakurai

The confidence interval of log odds ratio for the posterior probabilities of the two homoscedastic normal groups is considered. When the prior probabilities are assumed to be equal, the confidence interval is identified to the one for the linear discriminant function. The results proposed in the literatures are based on the large sample asymptotic theory. Generally, the precision for the large sample asymptotic approximation become worth when the dimension is comparable with the sample size, and one way of the improvement is to derive approximation under high-dimensional asymptotic framework that the sample size and dimensionality go to infinity together. We propose an approximated confidence interval based on the high-dimensional asymptotic framework. The precision is checked by simulation study.

C1651: **Confidence sets based on the positive part James-Stein estimator with the asymptotically constant coverage probability.**
Presenter: **Sujitta Suraphee**, Mahasarakham University, Thailand
Co-authors: S. Ejaz Ahmed, Iskander Kareev, Igor Volodin, Andrei Volodin

The asymptotic expansions for the coverage probability of a confidence set centered at the James-Stein estimator show that this probability depends on the noncentrality parameter τ^2 (the sum of the squares of the means of normal distributions). We show how these expansions can be used for a construction of confidence region with constant confidence level, which is asymptotically (the same formula for both case $\tau \to 0$ and $\tau \to \infty$) equal to some fixed value $1-\alpha$. We establish the shrinkage rate for the confidence region according to the growth of the dimension p and also the value of τ for which we observe quick decreasing of the coverage probability to the nominal level $1-\alpha$. When $p \to \infty$ this value of τ increases as $O(p^{1/4})$. The accuracy of the results obtained is shown by Monte-Carlo statistical simulations.

C1722: **Classification and variable selection in logistic regression via random subspace method**
Presenter: **Joanna Karlowska-Pik**, Polish Academy of Sciences, Poland
Co-authors: Jacek Koronacki, Jan Mielniczuk, Pawel Teisseyre

Classification in case of a high dimensional feature space and limited number of observations (so called "small n large p problem") is a very important challenge in biological sciences, especially genomic and proteomics (the typical example are genome-wide associations studies). The aim of the research is not only a construction of an effective classifier but, first of all, the selection of most informative features. One of possible solutions is random subspace method for logistic regression, where a random subset of features having cardinality smaller than a number of potentially useful regressors is chosen and the problem is solved with a reduced feature space of selected predictors. Features are assigned weights based on the Wald statistics. The procedure is repeated and weights are cumulated, finally the features are ordered according to the assigned weights. We would like to present this method paying particular attention to the specific requirements of the Wald statistics and indicating some solutions which may make this method more efficient.

CS44 Room 13 CONTRIBUTIONS TO COPULA-BASED MODELING I Chair: F. Marta L. Di Lascio

C1417: **Spill over effects and contagion in financial turbulence- evidence from global financial crisis**
Presenter: **Syeda Rabab Mudakkar**, Lahore School of Economics, Pakistan
Co-authors: Jamshed Uppal

The global financial crisis of 2007-09 (GFC) had a harsh impact on the financial markets of emerging and developed economies. The contagion effect as reflected by the different estimated parameters for pre and post GFC periods is found for majority of the countries across the globe. The phenomenon of financial-market crises spilling over to other countries has also been a major concern. The aim is to investigate the following two main issues i.e. Firstly it examines the volatility structure of few leading emerging and developed economies during GFC and how the distributions has changed for pre and post GFC periods. Secondly, we examine spilling over effects by using a copula approach for both pre and post crisis periods.

C1517: **Generalized additive models for conditional copulas**
Presenter: **Thibault Vatter**, University of Lausanne, Switzerland
Co-authors: Valérie Chavez

The aim is to develop a generalized additive modeling (GAM) framework for taking the effect of covariates on the dependence structure between two variables into account. In this context, the flexibility of the copula approach is an obvious advantage, as it allows to dissociate the specification of conditional margins and dependence structure. By letting the copula parameter(s) depend on covariates in a parametric, non-parametric or semi-parametric way, we develop a theory for the dependence structure similar to that of the GAM framework for univariate data. Fitting is by maximum penalized log-likelihood estimation, as is usual for roughness penalty-based methods, and some specific asymptotic properties are derived. Details of the model and smoothing parameter estimation procedure are also discussed, along with a measure of the penalized model's effective dimension, namely the equivalent degrees of freedom. Finally, selected simulations and two applications using real datasets are presented. In particular, we study the expression of genes involved in the development of breast cancer and the cross-sectional dynamics of intraday asset returns.

C1648: **Modeling and prediction of exchange rate and billion gold price of Thailand**
Presenter: **Pimpan Amphanthong**, Rajamangala University of Technology Suvarnabhumi, Thailand
Co-authors: Piyapatr Busababodhin

The aim of this article is to investigate a correlation of the dependence structure between THB/USD exchange rate and golden price of Thailand by using extreme value copula. The selection and estimation of the copula extreme value theory is based on maximum likelihood estimation (MLE) method, and behavior of dependence was determined by the dependence function. The procedure is suggested for the measurement of the copula function to recover the joint tailed distribution by comparing five parametric models of extreme value copulas. The results of this analysis denote that the Aneglog and Asy.log copulas analysis are the most appropriate method to best fit extreme value copula for bivariate generalized extreme value distribution (BGEV) and bivariate generalized Pareto distribution (BGPD), respectively. Furthermore, value at risk (VaR) is applied to calibrate the probability of the joint tail that may occur in block maxima and over the threshold. The neglog and aneglog copula are found to stand the maximum risk of block maxima and exceeding the threshold. These results could be beneficial for business and policy makers to predict the possibility of extreme economical fluctuation in the future.

C1647: **Modeling of rainfall and temperature: extreme value copula analyses**
Presenter: **Piyapatr Busababodhin**, Mahasarakham University, Thailand
Co-authors: Pimpan Amphanthong

Rainfall and temperature are important climatic inputs for agricultural production, especially in the context of climate change. However, accurate analysis and simulation of the joint distribution of rainfall and temperature are difficult due to possible dependence between them. The aim of this article is to investigate the dependence structure between rainfall and temperature in central northeast region of Thailand, using four parametric models of extreme value copula by combining the bivariate extreme value theorem and copula. The selection and estimation of the copula extreme value theory is based on maximum likelihood estimation (MLE) method, and a behaviour of dependence was determined by the dependence function. Historical climatic data of region is used to demonstrate the modeling process. Heteroscedasticity and autocorrelation of sample data are also considered to eliminate the possibility of observation error. The results indicate that there are high correlations between rainfall and temperature for the months from April to July and October. The extreme value copula is found to be most suitable to model the bivariate distribution of rainfall and temperature based on the Akaike information criterion (AIC). The resulting models can be integrated with research on agricultural production and planning to study the effects of changing climate on crop yields.

C1401: **Optimal designs for copula models**
Presenter: **Elisa Perrone**, Johannes Kepler University of Linz, Austria
Co-authors: Werner Mueller

Copula modelling is widely employed in many areas of applied statistics. However, the design of related experiments is still a neglected

aspect. In this work we analyze the relationship between optimal design theory and copula theory, with the goal of highlighting the influence of stochastic dependence in the optimal design domain. To this end, we provide a framework for the optimal design of copula models. The framework is provided with an extension of the classical equivalence theorem allowing one to formulate efficient design algorithms that quickly check whether designs are optimal or reasonably efficient. We investigate the parameter robustness, focusing on the way it relates to different types of copulas we considered. This approach is innovative also in the way we went through the issue of whether the estimation of copula parameters can be enhanced by optimizing experimental conditions. For this aim, particular importance is given to the role played by the copula parameter as one of the parameters to be estimated. Finally, examples are provided to illustrate that in practical situations considerable gains in design efficiency can be achieved. A natural comparison between different copula models with respect to design efficiency is shown as well.

CS66 Room 20 STATISTICAL PROCESS CONTROL Chair: Fernanda Figueiredo

C1345: **A cusum average loss control chart for monitoring process variation**
Presenter: **Su-Fen Yang**, National Chengchi University, Taiwan
Co-authors: Jeng-Sheng Lin, Chung-Ming Yang
A single chart to simultaneously monitor the process mean and variability would reduce the required time and effort. A number of studies have attempted to find such charts. The quality of processes/products and loss of productivity are crucial factors among competitive companies in industries. Loss function is widely used in the industry to measure the loss caused by deviation of the quality variable from its target value. The existing current loss-function-based control charts assumed that the in-control process mean was equal to the target value. However, in practice, the in-control process mean may not be the process target. The cumulative sum chart is an effective alternative to the Shewhart control chart, and may be used when small shifts occur in the process parameter. The aim is to propose a new cumulative sum average loss chart to effectively monitor the process variation, which is equivalent to monitor the changes in the difference of the process mean and target or the increase in process variability, or both. The performance of the cumulative sum average loss chart is measured using average run length. Data analyses illustrated that the cumulative sum average loss chart outperformed the joint Shewharts charts in detecting out-of-control mean and variance simultaneously if the adopted expected shift scales in mean and/or variance are larger. Therefore, the cumulative sum average loss chart is recommended.

C1405: **Sampling inspection by (Gaussian) variables via estimation of the lot fraction defective: a computational approach**
Presenter: **Miguel Casquilho**, University of Lisbon, Portugal
Co-authors: Elisabete Carolino
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) 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 distribution, for the computing of which we prepared a computer program and an open web page. 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.

C1539: **Behaviour of the quality index in acceptance sampling by variables: computation and Monte Carlo simulation**
Presenter: **Fatima Rosa Coelho**, University of Lisbon, Portugal
Co-authors: Miguel Casquilho
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 web site. 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.

C1606: **Monitoring the shape parameter of a Weibull distribution**
Presenter: **Fernanda Otilia Figueiredo**, Faculdade de Economia da Universidade do Porto and CEAUL, Portugal
Co-authors: M. Ivette Gomes, Adelaide Figueiredo
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.

C1626: **Monitoring the process variability using STATIS**
Presenter: **Adelaide Maria Figueiredo**, Faculdade de Economia da Universidade do Porto and LIAAD INESC Porto, Portugal
Co-authors: Fernanda Figueiredo
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.

CS79 Room 19 APPLIED STATISTICS AND DATA ANALYSIS II Chair: Heungsun Hwang

C1659: **The optimal number of lags in variogram estimation in spatial data analysis**
Presenter: **Sujung Kim**, Okayama University, Japan
Co-authors: Kuniyoshi Hayashi, Koji Kurihara
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. 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).

C1716: **The applications of symbolic data analysis in regional and spatial research**
Presenter: **Justyna Wilk**, Wroclaw University of Economics, Poland

Regional and spatial studies are focused on explaining the situations and interactions between regions. One of the most problematic issues of the research is an incomplete or insufficient precision of phenomena description and disregarding internal diversification of units. Symbolic data analysis performs a useful statistical tool to improve the quality of regional and spatial research. The approach consists in presenting phenomena in the form of symbolic data and applying symbolic data analysis methods. The aim is to apply symbolic data analysis in regional and spatial research. The first part of the paper presents the concept of symbolic data and the set of statistical methods to deal with. The second part examines the economic disparities between regions on the basis of symbolic data. The third part uses symbolic data analysis in an econometric gravity model of population migrations.

C1646: **Spatial dependence monitoring over distributed data streams**
Presenter: **Antonio Irpino**, Second University of Naples, Italy
Co-authors: Antonio Balzanella, Rosanna Verde

A strategy for monitoring spatial dependence in multiple, spatially located, data streams is proposed. 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.

C1676: **Effects of sampling methods on prediction quality: The case of classifying land cover using decision trees.**
Presenter: **Ronald Hochreiter**, WU Vienna University of Economics and Business, Austria
Co-authors: Christoph Waldhauser

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.

C1423: **Estimation of household income based on asset ownership in Georgia**
Presenter: **Nino Mushkudiani**, CBS, Netherlands

The National Statistics Office of Georgia (GeoStat) and Statistics Netherlands have carried out a survey study to improve the quality of the Integrated Household Survey (IHS) of GeoStat. One of the issues is that the incomes of the households obtained from the IHS are not reliable and are not representative for the population. In order to obtain information related to household income, we developed an asset ownership questionnaire. The questionnaire was applied to a small group of households with known income. Using a linear regression model we wanted to find the set of covariates that would lead to the optimal model according to the model AIC score. If we consider all possible combinations of our 23 covariates we will have $2^{23} - 1 = 8\,388\,607$ models. Instead we defined a core model of the 10 most significant variables and tried to improve it by adding each possible combination of the remaining covariates. The statistical program R had no trouble calculating these $8\,191$ models. Using the "optimal" linear regression model we estimated the household income for the IHS data frame in the Tbilisi (Georgia) area. Next we defined a score function for stratification of the IHS according to household income.

Wednesday 20.08.2014 09:15 - 10:45 Parallel Session D

IS04 Room 3 APPLIED BAYESIAN STATISTICS **Chair: Anne Philippe**

C1550: **BiiPS: software for inference in Bayesian graphical models with sequential Monte Carlo methods**
Presenter: **Adrien Todeschini**, Inria Bordeaux - Sud-Ouest, France
Co-authors: François Caron, Marc Fuentes, Pierrick Legrand, Pierre Del-Moral

The main factor in the success of Markov chain Monte Carlo methods is that they can be implemented with little efforts in a large variety of settings. Many types of software have been developed such as BUGS and JAGS, which helped to popularize Bayesian methods. These softwares allow the user to define the statistical model in a so-called BUGS language, and then run MCMC algorithms as a black box. Although sequential Monte Carlo methods have become a very popular class of numerical methods over the last 20 years, there is no such "black box software" for this class of methods. The BiiPS software, which stands for Bayesian Inference with Interacting Particle Systems, aims at bridging this gap. From a graphical model defined in BUGS language, it automatically implements sequential Monte Carlo algorithms and provides summaries of the posterior distributions. In this talk, we will highlight some of the features of the BiiPS software, its R and MATLAB interfaces, and illustrations of its application to various models in financial econometrics, object tracking or systems biology.

C1521: **A hierarchical Bayesian approach for dating in archaeology.**
Presenter: **Anne Philippe**, Nantes, France

Bayesian statistical tools are widely used by archaeologists in particular for building chronologies with help of the softwares Bcal (Buck et al., 1999) and OxCal (Bronk Ramsey, 1995). In the same context, we propose a new model based on "Fact" model. It provides a priori archaeological information about the contemporaneity of events. This is incorporated into the model using a hierarchical structure with individual effects on the variance. As in previous models, the prior distribution takes also into account the knowledge on relative dating from stratigraphic sequences, archaeological phases and information about calendar dates. The observations are the measurements obtained by dating laboratories (radiocarbon or luminescence TL/OSL ages, archaeomagnetic measurements, etc.), and so a step of calibration is needed to convert them into calendar years. The method is implemented in new software called ChronoModel. The outputs are, for instance, the chronologies with HPD regions; the predictive distributions of archaeological phases, etc. ... Some applications to the chronology of prehistoric sites and to palaeoenvironmental sequences are presented to illustrate the performances of our approach. We show in particular that the "Fact" model makes the calendar date estimates robust to the presence of outliers.

C1402: **Mixture model of Gaussian copulas to cluster mixed-type data**
Presenter: **Matthieu Marbac**, University of Lille, France
Co-authors: Christophe Biernacki, Vincent Vandewalle

A mixture model of Gaussian copulas is proposed to cluster mixed data. This approach allows us to straightforwardly define simple multivariate intra-class dependency models while preserving classical distributions for one-dimensional margin of each component in order to facilitate the model interpretation. In addition, the intra-class dependencies are taken into account by the Gaussian copulas providing one robust correlation coefficient per couple of variables and per class. This model generalizes different existing models defined for homogeneous or mixed variables. The inference is performed via a Metropolis-within-Gibbs sampler in a Bayesian framework. The model is illustrated by a real data set clustering.

OS08 Room 5 DEPENDENCE MODELS **Chair: F. Marta L. Di Lascio**

C1373: **Weighted least square inference for multivariate copulas based on dependence coefficients**
Presenter: **Gildas Mazo**, Grenoble, France
Co-authors: Stephane Girard, Florence Forbes

Copulas are a useful tool to model multivariate distributions, as they permit to impose a dependence structure on pre-determined marginal distributions. Given a sample of the copula, the estimation of the copula parameter vector must be carried out. When the copula is not differentiable, likelihood methods are not available. To deal with this issue, a weighted least square estimator based on dependence coefficients is presented. The consistency and asymptotic normality of the estimator are shown and several examples involving non-differentiable copulas are considered.

C1411: **The univariate conditioning and tail dependence; two ways of studying the extreme events**
Presenter: **Piotr Jaworski**, University of Warsaw, Poland

The interest in the construction of multivariate stochastic models describing the dependence among several variables has significantly grown in the last years. In particular, the recent financial crisis emphasized the necessity of considering models that can serve to estimate better the occurrence of extremal events. There are several ways to cope with this task. The distributional approach leads to the study of the behaviour of conditional copulas for extremal values of the first variable (called also tail-dependence or threshold copulas), i.e. the copulas of conditional distribution of $X = (X_1, \ldots, X_n)$ when X_1 is smaller than α-quantile or respectively greater than (1-α)-quantile. The other approach to study and model the interdependencies between extreme events is to study the tail (i.e. corner) behaviour of the corresponding copula. This is based on the tail expansion of copulas near the vertices of the unit multicube and the tail dependence functions. The goal of my talk is to show that there is a strong link between the limiting properties of the conditional copulas when α tends to 0 and the tail dependencies of the copulas.

C1340: **Efficient iterative maximum likelihood estimation of high-parameterized time series models**
Presenter: **Ostap Okhrin**, Humboldt-University Berlin, Germany
Co-authors: Nikolaus Hautsch, Alexander Ristig

An iterative procedure to efficiently estimate models with complex log-likelihood functions and the number of parameters relative to the observations being potentially high is proposed. Given consistent but inefficient estimates of sub-vectors of the parameter vector, the procedure yields computationally tractable, consistent and asymptotic efficient estimates of all parameters. The asymptotic normality is shown and the estimator's asymptotic covariance is derived in dependence of the number of iteration steps. To mitigate the curse of dimensionality in high-parameterized models, the procedure is combined with a penalization approach yielding sparsity and reducing model complexity. Small sample properties of the estimator are illustrated for two time series models in a simulation study. In an empirical application, the proposed method is used to estimate the connectedness between companies by extending a previous approach to a high-dimensional non-Gaussian setting.

C1422: **Optimal dependence structures for integrals involving the sine**
Presenter: **Maria Rita Iaco**, Technische Universität Graz, Austria
Co-authors: Vladimir Balaz, Markus Hofer, Otto Strauch, Robert Tichy
The aim is to provide the copula which gives the optimal upper bound for a two-dimensional integral involving the sine function. This result extends the one found by the presenting author and the third author in a previous work, where they conjectured which is the shape of the maximal copula for the problem at hand.

OS30 Room 4 INFINITE DIMENSIONAL DATA ANALYSIS Chair: Alicia Nieto-Reyes

C1473: **Random walk testing for time series of functions**
Presenter: **Juan Romo**, Universidad Carlos III de Madrid, Spain
Co-authors: Rosa Lillo, Nicola Mingotti
Given a time series of functions, we asume a functional autoregressive model of order one and test if the time series is a random walk in the space of squared integrable functions against the alternative hypothesis of stationarity. The test is based on the Hilbert-Schmidt norm of the empirical covariance operator. A bootstrap resampling strategy allows us to approximate the norm distribution and construct the corresponding rejection region. An extensive simulation study illustrates the good power properties of the test in small samples. Finally, the test is applied to several real functional time series to illustrate its behaviour.

C1458: **Functional theorem of Glivenko-Cantelli with application to data depth for functions**
Presenter: **Stanislav Nagy**, Charles University Prague, Belgium
In the contribution we provide a uniform strong law of large numbers for the collection of cross-sectional empirical distribution function processes computed from a random sample of functional data. The result is applied to the problem of consistency of depth for functions. Counterexamples illustrating the fact that in general the conditions of the theorem cannot be dropped are given.

C1437: **Bayesian non-parametric spectral density estimation: with application to under-sampled time series**
Presenter: **Benedict Powell**, University of Bristol, United Kingdom
Our motivation for this work is the simple and commonly-asked question: “have I been sampling this time series frequently enough?” The methodology needed to formalize the question centres on the spectral density function for the process generating the series. This function, for which there is rarely a natural parametric form, encodes the distribution of power in the process at different frequencies. Since it is not directly observable, we model our beliefs for its values according to the Bayesian paradigm. Constraints on the spectral density function associated with smoothness, positivity and compatibility with data taken at different sampling rates, make for a posterior density which is difficult to navigate and summarize effectively. The challenge of doing so constitutes the first part of this project. The second part is the formulation of a decision problem which allows us to rationalize an answer to the motivating question. Further, we investigate the potential to offer advice in regard to the length of a trial period, at a high sampling rate, that would allow us to provide a better answer to the motivating question.

C1425: **A study of Gaussianity in stationary processes**
Presenter: **Alicia Nieto-Reyes**, Universidad de Cantabria, Spain
Co-authors: Juan Cuesta-Albertos, Fabrice Gamboa
Goodness-of-fit tests have being widely studied in the literature. Very often, observed data are a finite path of real temporal phenomena modelled as a second order stationary process. Adding the Gaussianity assumption, the process possesses a lot of beneficial properties as regards their statistics or prediction and, in particular, it becomes strictly stationary. Regarding the study of Gaussianity tests for stationary processes, the existing tests only verify the Gaussianity of a marginal at a fixed finite order, generally order one. Therefore, they do not reject stationary non-Gaussian processes with the one-dimensional Gaussian marginal. A consistent test is proposed for Gaussianity of stationary processes when a finite sample path of the process is observed. This test is inspired by the fact that if we take at random a one-dimensional projection of a non-Gaussian distribution, then, with probability one, this projection will be non-Gaussian. Thus, decision rules are applied to the whole distribution of the process and not only on its marginal distribution at a fixed order, as in previous tests. The main idea is to test the Gaussianity of the one-dimensional marginal distribution of some random linear transformations of the process. Note that testing the one-dimensional marginal distribution can be done with previous tests of Gaussianity for stationary processes.

CS53 Room 19 CONTRIBUTIONS TO COMPUTATIONAL METHODS IN FINANCE I Chair: Peter Winker

C1318: **Fast detection of structural breaks**
Presenter: **Paul Fischer**, Denmarks Technincal University, Denmark
Co-authors: Astrid Hilbert
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.

C1553: **Forecasting forward price curves in electricity markets: a factor model**
Presenter: **Paolo Foschi**, University of Bologna, Italy
A dynamic factor model for the cross section for Italian electricity forward prices is presented. This market is characterised by a small number of quotations for each forward contract and by a large bid-ask spread. Moreover, it is not uncommon to observe inconsistencies in the cross section of forward quoted prices. The aim of the proposed model is twofold. Firstly, this approach allows to filter out noise and to help in identifying inconsistencies. Secondly, it provides a tool to fill-in missing quotations and to unpack forward contract with long delivery into smaller ones (i.e. a quarter or a calendar into its monthly components). The model is built using monthly contracts are used as basic entities and their dynamics is modeled as follows. The cross-section of those contracts is decomposed in two sets of factors: the first set, which depends on the time-to-delivery, allows to financial structure of the market. The second set, which depends on the actual delivery month, will capture the seasonal component. The parameters of those factors are allowed to be slowly varying to achieve maximal flexibility.

C1595: **Forecasting residual demand time series in electricity markets: a functional approach**
Presenter: **Jose Portela**, Universidad Pontificia Comillas, Spain
Co-authors: Antonio Munoz, Estrella Alonso

A new forecasting method for functional time series is proposed. The new model has been tested with the residual demand curves of the Spanish day-ahead electricity market and compared with other functional reference models. Electricity generators and retailers trading in electricity markets can take advantage of residual demand curves forecasts as tools for optimizing their bidding strategies. The model is aimed at extending the ARH model (Autoregressive Hilbertian model) to the SARH model in which the seasonality of the series is taken into account. This is a significant improvement, as high-frequency time series generated in the context of electricity markets show seasonal dynamics. Therefore, this model is built following a two steps procedure. In the first step, the structure of the model has to be identified. By means of a functional autocorrelation plot, significant autocorrelations in the functional time series are found, and initial values for the regular and seasonal autoregressive orders are inferred. Secondly, a seasonal autoregressive functional linear model is estimated using the inferred structure. While the functional parameter is usually estimated using a functional principal component basis, in this paper we propose a Gaussian estimator based on neural network techniques.

CS51 Room 20 CONTRIBUTIONS OF COMPUTATIONAL STATISTICS TO ENVIRONMENTAL AND LIFE SCIENCES Chair: Raphael Huser

C1667: **Estimation of spatially correlated ocean temperature curves including depth dependent covariates**
Presenter: **Rosaura Fernandez Pascual**, University of Granada, Spain
Co-authors: Rosa M. Espejo, Maria Dolores Ruiz Medina

The purpose is to study the 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 hierachical 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.

C1679: **Efficiency of sequential Monte Carlo and genetic algorithm in Bayesian estimation of the atmospheric contamination source**
Presenter: **Anna Wawrzynczak-Szaban**, National Centre for Nuclear Research, Poland
Co-authors: Anna Wawrzynczak, Piotr Kopka, Marcin Jaroszyski, Mieczyslaw Borysiewicz

Abrupt releases of hazardous material into the atmosphere pose a great threat to the human health and the environment. It is crucial to develop an 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.

C1685: **Parametric inference for stochastic SIS epidemic model in random environment**
Presenter: **Tewfik Kernane**, University of Sciences and Technology Houari Boumediene, Algeria
Co-authors: Sidali Becheket, Hamid El Maroufy

The problem of parametric inference for a stochastic SIS epidemic model in random environment is considered. This extension of the simple SIS epidemic model allows for switching between different sets of parameters by assuming a Markovian switching process. The SIS model is often used to model diseases for which there are no immunity, including for example gonorrhea and pneumococcus. We obtain diffusion approximation of the stochastic SIS epidemic model in random environment and for parameter estimations we use Pseudo-Maximum Likelihood Estimation. Computer simulations are performed to illustrate our approach.

C1564: **Functional data modeling to measure exposure to ozone**
Presenter: **Maeregu Woldeyes Arisido**, University of Padova, Italy

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 us to estimate the effect of ozone as a function of daily hour, which allows us 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.

CS65 Room 13 CLUSTERING I Chair: Luca Bagnato

C1683: **Clustering ordinal data using binary decision trees**
Presenter: **Pierre Michel**, Aix Marseille University, France
Co-authors: Badih Ghattas

The aim is to introduce an extension of CUBT (clustering using unsupervised binary trees) to ordinal data. CUBT is a hierarchical clustering method for continuous data inspired from CART and 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. We extend here this approach to ordinal data using mutual information and entropy criteria. Different simulations show the efficiency of our approach.

C1599: **To split or to mix? Tree vs. mixture models for detecting subgroups**
Presenter: **Hannah Frick**, Universität Innsbruck, Austria
Co-authors: Carolin Strobl, Achim Zeileis

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 is 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.

C1615: **Variable-wise kernel-based clustering algorithms for interval-valued data with kernelization of the metric**
Presenter: **Francisco de Assis Tenorio de Carvalho**, Universidade Federal de Pernambuco - UFPE, Brazil
Co-authors: Marcelo Rodrigo Portela Ferreira

This presentation gives partitioning kernel clustering algorithms for interval-valued data with kernelization of the metric based on adaptive distances. These adaptive distances are obtained as sums of squared Euclidean distances between interval-valued data computed individually for each interval-valued variable by means of kernel functions. The advantage of the proposed approach over the conventional kernel clustering approaches for interval-valued data is that it allows learning the relevance weights of the variables during the clustering process, improving the performance of the algorithms. Experiments with synthetic and real interval-valued data sets show the usefulness of these kernel clustering algorithms.

C1568: **Comparison of selected similarity measures used for clustering of nominal variables**
Presenter: **Zdenek Sulc**, University of Economics in Prague, Czech Republic
Co-authors: Hana Rezankova

The contribution deals with hierarchical clustering of nominal variables. This kind of clustering is based on a proximity matrix, which contains dissimilarities among all pairs of variables. We compare two approaches to express these dissimilarities. The first one consists of using association measures for nominal variables, e.g. contingency coefficients. The second approach deals with similarity measures determined for objects characterized by nominal variables, e.g. the simple matching coefficient. Several other similarity measures have been proposed in recent years. These measures take into account more characteristics regarding the dataset, such as distribution of frequencies of categories; therefore, they should provide better results. The similarity measures, introduced in the second approach, have one limitation though. All input variables must have the same number of categories and their categories must have the same meaning. Still, there are lots of areas, where clustering using these similarity measures can be applied, e.g. in batteries of questions. For the research, several real datasets from social surveys were used. The quality of created clusters was evaluated from aspects of both the within-cluster variability and the substantive interpretation. The results show that some of the examined measures provide considerably better results than the standardly used ones.

PS01 Room Espace Motta POSTER SESSION I Chair: Francisco de Asss Torres Ruiz

C1280: **Sample size determination for inferences based on functions of second moments using nonlinear minimization techniques**
Presenter: **Jen-pei Liu**, National Taiwan University, Taiwan
Co-authors: Chieh Chiang

Statistical inferences for evaluation of quality of biopharmaceutical products are often based on functions of the second moments of the normal distributions. These include individual bioequivalence for generic drug products, the in vitro bioequivalence for generic nasal aerosols and nasal sprays, the within-device precision for in vitro diagnostic devices, and many others. Most of testing procedures proposed for the inference with respect to these criteria are based on the upper confidence limit derived from the modified large-sample method. However, scarce literature exists for the sample size determination for the inference based on the second moments. It is shown that the upper confidence limit is asymptotically normal. Given the pre-specified significance level and power, the method of sample size determination for the functions of second moments can be formulated as an optimization problem of minimizing a continuous nonlinear function of decision variables with equality constraints which can be evaluated by the quasi-Newton method. Results of simulation studies demonstrate that the sample sizes of our proposed methods provide sufficient and yet over excessive power. Real data illustrate applications of the proposed methods.

C1351: **Regression modelling to explain adverse events after acute myocardial infarction**
Presenter: **Carla Henriques**, School of Technology and Management - Polytechnic Institute of Viseu, Portugal
Co-authors: Ana Matos, Davide Moreira

Based on 1006 records of patients admitted in a Portuguese Coronary Care Unit, due to acute myocardial infarction, the prognostic factors to in-hospital mortality, one-year mortality and one-year major event occurrence are studied. These outcomes are more prevalent in the elderly population, but research studies have been questioning the age as being an independent risk factor, investigating, in particular, if the effect of age is due to different therapeutic approaches (in general less invasive therapeutic approaches are administrated to elderly patients). Three age groups are considered: <65, 65-74 and $\geq$75 years. The cases were classified, according to the admission diagnosis, as acute myocardial infarction with ST elevation (STEMI) or without ST segment elevation (NSTEMI). Resorting to multiple regression modelling it cannot be concluded that age is not an independent risk factor for in-hospital mortality; instead, its effect depends on whether cardiac catheterization (an invasive strategy) is done. The STEMI proved to be a risk factor for in-hospital mortality and one-year major event occurrence. Also, cardiac catheterization revealed to be a protective factor in any age group. Other clinical variables were investigated as possible influential factors. Regression models included all significant main effects and interaction terms. The goodness of fit was evaluated.

C1377: **Forecasting mortality: some analytical, experimental and empirical evidences**
Presenter: **Taku Yamamoto**, Nihon University, Japan
Co-authors: Hiroaki Chigira, Chisako Yamamoto

Forecasting mortality has been a vital issue in demography and actuarial science. Forecasting methods for mortality are evaluated in the framework of cointegrated time series analysis. The Lee-Carter method has been regarded as the benchmark for forecasting mortality. However, its forecasting accuracy has been known to be particularly poor for short-term forecasts, while it is well for long-term forecasts. Recently, a new method called the MTV method, which explicitly satisfies cointegration restrictions of the series, has been proposed. The aim is to propose an alternative forecasting method which generalizes the Lee-Carter method in order to improve short-term forecasts. Forecasting accuracy of the proposed method with the Lee-Carter method and the MTV method are compared in the framework of cointegrated time series analysis, and further in the Monte Carlo experiment and in an empirical application of Swedish male. It is shown that the proposed method is evidently more accurate than the Lee-Carter method and equally well with the MTV method. Since the proposed method is simpler than the MTV method, it may be useful for practical use.

C1479: **Chemometric analysis of macro and micronutrients in soils and associated medicinal plants from Maasai region, Kenya**
Presenter: **David Maina**, University of Nairobi, Kenya
Co-authors: Benson Namwiba, John Onyari, Johan Boman, Keith Shepherd
Despite the presence of modern medicines, some residents of the Maasai region still consult traditional health providers. In most cases, the medicines are obtained from local plants. The macro and microdensity profiles of these local plants may have an influence on their choice as medicinal plants. Soil samples and medicinal plants were obtained from Kajiado and Narok counties in the Maasai region and the concentrations of calcium, magnesium, manganese, iron, copper and zinc analysed using X-ray fluorescence analytical technique. STATIS, Parafac2 and Tucker3 multiway methods were used to evaluate how the macro and micronutrient correlate with their use as medicinal plants. The results indicate that different species have different trace element density profiles even when grown under similar conditions. In addition, there was a significant difference in the macro and micronutrient density profiles for the two counties. Finally, the parts of the plants with the highest elemental density were also the parts used for medicinal purposes.

C1692: **A course on introductory statistics using interactive educational tools**
Presenter: **Kazunori Yamaguchi**, Rikkyo University, Japan
Co-authors: Takaaki Ohkawauchi, Kotaro Ohashi
The aim is to introduce a new e-learning course for principles and methods of introductory statistics, which is developing for all students in Rikkyo University. The course consists of the following contents; Usages and linkages to the official statistics in Japan, videos titled statistics for daily life, and interactive learning contents. For this course, we have developed Japanese versions of interactive Java applets for understanding statistical concepts and a tool for the simulation and data analysis. We expect that combination of these tools and e-learning contents make students easy to understand basic concepts of statistics.

C1331: **On the modification of the non-parametric test for comparing locations of two populations**
Presenter: **Grzegorz Konczak**, University of Economics in Katowiceue, Poland
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. The modification is based on the Hao L. and Houser D. adaptive test. In the mentioned test 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 analyzed in the Monte Carlo study.

C1710: **Bootstrap technique and number of PLS or PLSGLR components selection.**
Presenter: **Myriam Maumy-Bertrand**, Université de Strasbourg, France
Co-authors: Jeremy Magnanensi, Nicolas Meyer, Frederic Bertrand
The extraction of the correct number of PLS components is a real challenge. The problem of finding relevant degrees of freedom (DoF) for the PLS was solved recently by N. Kramer and M. Sugiyama. They also adapted the AIC and BIC criteria with these new DoF and applied them to the selection of the number of PLS components. They compared these criteria with the 10-fold cross-validated Q2 criterion. They concluded in comparable prediction accuracy between the BIC and the Q2 criteria. We adapted the so-called bootstrapping pairs technique in order to test the significance of the successive PLS components and compare this criterion with the more commonly used ones either based on cross-validation or on information criteria. An extension of the PLS regression to generalized linear regression has been developed by Bastien et al. In this case, even fewer criteria can be used to select the number of components since Dof are unknown and Q2-like criteria fail to select a reasonable number of components. Our experimentation show a better stability of our bootstrap criterion and a globally better predictive accuracy compared to all the others criterion, either in classic PLS framework or in PLS-logistic case.

C1709: **Cross-validated partial least squares models and their extensions with censored data**
Presenter: **Frederic Bertrand**, Université de Strasbourg, France
Co-authors: Philippe Bastien, Myriam Maumy-Bertrand
When cross-validating standard or extended Cox, the commonly used criterion is the cross-validated partial loglikelihood using a naive or a van Houwelingen scheme -to make efficient use of the death times of the left out data in relation to the death times of all the data-. Quite astonishingly, we will show, using a strong simulation study involving three different data simulation algorithms, that these two cross-validation methods fail with the extensions, either straightforward or more involved ones, of partial least squares regression to the Cox model. This is quite an interesting result for at least two reasons. Firstly, statisticians commonly use these extensions and usually select their hyperparameters using cross-validation. Secondly, they are almost always featured in benchmarking studies to assess the performance of a new estimation technique used in a high dimensional context. We carried out a vast simulation study to evaluate more than a dozen of potential cross-validation criteria, either AUC or prediction error based. Several of them lead to the selection of a reasonable number of components. Using these newly found cross-validation criteria to fit extensions of partial least squares regression to the Cox model, we performed a benchmark reanalysis that showed enhanced performances of these techniques.

C1699: **Identification of molecular targets and signaling networks associated with radiation sensitivity**
Presenter: **Agata Michna**, Helmholtz Zentrum Muenchen, Germany
Co-authors: Herbert Braselmann, Horst Zitzelsberger, Kristian Unger
Individual radiation sensitivity plays an important role in radiation oncology. With regard to the success of the therapy, stratification of patients and tumors, concerning their radiation sensitivity or resistance, is required. Therefore, mechanistic understanding of radiation resistance is crucial. It is necessary to identify signaling networks driving radiation resistance that can be targeted with pharmacological agents in order to modulate the radiation sensitivity. Here, we present a systems biology approach by which we examine the radiation response of two lymphoblastoid cell lines derived from young lung cancer patients with normal and reduced sensitivity. Total RNA, isolated from sham and 1Gy gamma-irradiated cell lines, was labeled and hybridized to Agilent human gene expression arrays. The time-course data set of global gene expression was subjected to partial correlation based reconstruction of radiation-sensitivity associated gene regulatory networks. The behavior of gene expressions over time was approximated with spline regression. For the selection of genes to be considered in the network reconstruction we applied differential expression analysis combination with the GeneRank algorithm. Identified interactome reflects networks already known from interactome databases. Furthermore, it shows specific networks that point to the involvement of e.g. the P53, MAPK and extracellular matrix organization pathways. Further network analysis (e.g. betweeness centrality) will be done to define network modules that could serve as potential

targets in radiochemotherapy. The preliminary results demonstrate the feasibility and the added value of our approach towards an understanding of the mechanisms associated with radiation sensitivity.

C1687: **Partial correlation based on L-comoments**
Presenter: **Helena Pickova**, Charles University in Prague, Czech Republic
Co-authors: Jan Picek
Partial correlation coefficient is a classical measure of the linear dependence between two random variables in the case where the influence of a set of controlling variables is eliminated. We propose an alternative estimator based on L-comoments. The results of previous studies showed that when sample sizes are small and underlying distributions are heavy-tailed that the approach based on L-moments has advantages over conventional approach based on usual moments. This issue is illustrated on simulated and educational data.

PS02 Room Espace Motta POSTER SESSION II Chair: Francisco de Asis Torres Ruiz

C1397: **A general Weibull diffusion process**
Presenter: **Francisco Torres-Ruiz**, Granada, Spain
Co-authors: Antonio Barrera-Garcia, Patricia Roman-Roman
The use of mathematical models to analyze and forecast dynamical phenomena has become commonplace in current research for several fields of knowledge. For the purpose of explaining real processes, the random influence produced by internal and external conditions must be considered. In this regard, the theory of stochastic processes is a useful and accurate research tool. For instance, models based on Weibull curves are successfully used in areas such as biology, finance, industry, or weather forecasting, and one of its recent extensions, the hyperbolastic growth model of type III, has shown great performance in cell growth dynamics. Nevertheless, the increasing number of models, in addition to their high degree sophistication, and the complex structure required to apply them to very specific areas, restricts their application and makes the research process useless in absence of global concepts. For these reasons, a generalized viewpoint employing mathematical abstraction is required. To this end, we have established a theoretical framework aiming to define a functional generalization of the Weibull model, in addition to its stochastic extension, to finally construct a unique and generalized diffusion process which could help characterize several Weibull-based models, such as the hyperbolastic one.

C1418: **Some estimation approaches on smoothness of a stationary Gaussian random field**
Presenter: **Wei-Ying Wu**, National Dong Hwa University, Taiwan
Co-authors: Chae Young Lim
For a stationary Gaussian random field, the decay rate of the spectral density as the frequency becomes large determines the smoothness of the random field. The decay rate of the spectral density is also related to the fractal dimension, which is used to measure the surface smoothness of a random field. An estimator of the decay rate using periodogram when the observations are on a grid is proposed and the asymptotic properties under the fixed domain asymptotic setting are investigated. A bias-reduced estimate is proposed based on the theoretical property of the estimator. A simulation study and a real data example are presented and compared with other available methods. In addition, some feasible ways for the non-grid data are also discussed.

C1489: **Classification of functional data with covariate adjustment**
Presenter: **Pai-Ling Li**, Tamkang University, Taiwan
A covariate adjusted subspace projected functional data classification (SPFC) method for classifying response curves with taking into account the additional covariate information is proposed. Curves of each cluster are embedded in the cluster subspace spanned by a mean function and eigenfunctions of the covariance kernel. We assume that the mean function may depend on covariates, and curves of each cluster are represented by the conditional Karhunen-Loeve expansion. Under the assumption that all the groups have different mean functions and eigenspaces, an observed curve is classified into the best predicted class by minimizing the distance between the observed curve and predicted functions via covariate adjusted subspace projection among all clusters. The proposed covariate adjusted SPFC method that accommodates additional information of other covariates is advantageous to improving the classification error rate. Numerical performance of the proposed method is examined by simulation studies, with an application to a data example.

C1486: **L1-penalized ordinal regression with applications to consumer preference data and survey data**
Presenter: **Ya-Ting Chang**, University of Waterloo, Canada
Co-authors: Mu Zhu
Polychotomous ordinal response data with covariates can be analyzed by imposing an ordinary regression model on an underlying latent variable that is assumed to be continuous. Such models are easily fitted with standard Markov Chain Monte Carlo (MCMC) techniques. We incorporate an L1-penalty into the latent regression step. Doing so inside an MCMC algorithm allows us to rank the importance of the covariates naturally by their respective selection probabilities. Other than simulations, we will show some preliminary (but interesting) results using data from the MovieLens Project and the World Values Survey.

C1504: **A geographically weighted autoregressive regression technique to model spatial dependence and nonstationarity**
Presenter: **Vivian Yi-Ju Chen**, Tamkang Universiy, Taiwan
Geographically weighted regression (GWR) and spatial autoregressive models have become the commonplace for empirical studies in geography and spatial science. The former is mainly used to study spatial nonstationarity and the latter is often employed to handle spatial dependence. While many disciplines have experienced growth in investigating either spatial dependence or spatial nonstationarity, they have been slow to explore them simultaneously. In this study, we attempt to combine the spatial lag model and GWR, constructing a spatial analytic tool, named geographically weighted autoregressive regression, which can account for spatial nonstationarity and spatial dependence at the same time. We first formulate the modeling specification, and then develop bootstrap methods to conducting the inference of model parameters. As an illustration, we apply the approach to a dataset of prenatal care utilization in the United States.

C1528: **On decomposition of refined estimator of measure for symmetry in square contingency tables**
Presenter: **Kouji Tahata**, Tokyo University of Science, Japan
Co-authors: Ryohei Auchi, Sadao Tomizawa
For the analysis of square contingency tables, the measures that represent the degree of departure from global symmetry and conditional symmetry have been proposed. Also, the refined estimator of measure for symmetry has been suggested. The present paper gives the refined estimators of the measures for global symmetry and conditional symmetry and gives the relationship between proposed estimators and the refined estimator of measure for symmetry. The proposed estimators can be obtained by using the

second-order term in the Taylor series expansion. The proposed estimators approach to the true values faster than the former estimators as the sample size becomes larger. These are shown by the simulation studies.

C1530: **Bayesian regression models to address replications of recordings for disease of Parkinson tracking**
Presenter: **Carlos Javier Perez**, University of Extremadura, Spain
Co-authors: Lizbeth Naranjo, Yolanda Campos-Roca, Jacinto Martin

Several investigations have recently considered the use of acoustic parameters extracted from speech recordings as an objective and non-invasive tool to perform diagnosis and monitoring of Parkinson's disease (PD). PD monitoring is often performed by applying the Unified Parkinson Disease Rating Scale (UPDRS). Some authors have achieved approximations to the UPDRS by applying regression models based on acoustic parameters. These previous works are based on the Parkinson Telemonitoring Dataset from UCI Machine Learning Repository. The experiment design consisted in the collection of repeated measurements (six per week at the same time) in a longitudinal study (six months). However, all the applied regression methods considered the characteristics extracted from recordings of any patient at any time as independent measures. This artificially increases the sample size and provides non-realistic results. In this work, repeated measurements and longitudinal properties are captured. Specifically, two Bayesian regression models are developed and implemented. The first one aggregates the weekly obtained measurements to achieve a single measurement for each patient at each time, then a longitudinal Bayesian model is applied. The second one introduces latent variables to directly address the repeated measurements in a longitudinal framework. Accurate approximations for the total and motor UPDRS have been obtained.

C1531: **Filtering algorithm in networked systems from uncertain observations with sensor random delays and packet dropouts**
Presenter: **Josefa Linares-Perez**, Universidad de Granada, Spain
Co-authors: Raquel Caballero-Aguila, Aurora Hermoso-Carazo

The signal estimation problem in multisensor systems, where sensor networks are used to obtain the full information on the signal and its estimation must be carried out from the observations provided by all the sensors, has become a broad and interesting research topic. Although the use of sensor networks offers several advantages, the communication channel imperfections usually cause problems during data transmission from the sensors to the fusion center, such as uncertain observations, random communication packet losses and random delays. This paper is concerned with the centralized least-squares linear estimation problem of discrete-time signals from noisy measurements coming from multiple sensors subject to these three sources of uncertainty. More specifically, it is assumed that the measured outputs, which may be only noise (uncertain observations), are transmitted from the different sensors to the processing center and one-step delays and/or packet dropouts may occur in the transmission. To model these random uncertainties at each sensor, different sequences of Bernoulli random variables are used. By using an innovation approach, a recursive linear filtering algorithm is derived without requiring the state-space model generating the signal, but only the mean and covariance functions of the signal and observation noises as well as the uncertainty probabilities.

C1634: **Statistical analysis platform based on webserver with R**
Presenter: **Miguel Angel Montero Alonso**, University of Granada, Spain
Co-authors: Antonio Lara-Aparicio, Juan de Dios Luna-del-Castillo

We present servidorR, a freely accessible statistical analysis environment, a tool for statistical computing and data analysis, that is fully accessible through a web interface and runs R scripts to perform the results. ServidorR is a platform that allows you to use the full power of R without the need to know anything about R, acting as an interface, without any restriction in the analysis capabilities and a great help to the teaching of statistics in general, and biostatistics in particular. Using web pages to interact with this system allows us to have an useful platform in any Internet browser, which is operating system independent and even mobile. The interface just allows you to enter your own data or upload it from a file, and outputs can be presented on screen and in different formats. We are at an early stage, with modules under development for use in introductory statistics courses and modules under binary diagnostic test.

PS03 Room Espace Motta POSTER SESSION III Chair: Cristian Gatu

C1655: **A search for new solutions of a to the p-1 equal to 1 modulo p squared with a probabilistic computation algorithm**
Presenter: **Ryuichi Sawae**, Okayama University of Science, Japan
Co-authors: Yoshiyuki Mori, Dasuke Ishii

Many computer scientists have searched the prime solution p of satisfying $a^{p-1} \equiv 1 \mod p^2$ for given a not a power. The reason some number-theoretic questions such as so famous Fermat's conjecture require primes p. Although Fermat's conjecture solved affirmatively, the questions inspired by these classes of primes still remain meaningful, and we also continue to turn our attention to the study of these primes. For example, the prime solution is very useful for a proof of the largest prime factor condition in odd perfect numbers, if it exists. Especially for $a = 2$, a solution is called as a Wieferich prime. Despite several intensive searches, until now, only two Wieferich primes are known: $p = 1093$ and $p = 3511$ for search up to 6.7×10^{15}. In our research, we adopt a new algorithm for search new solutions up to 10^{17} with a probabilistic computation algorithm.

C1638: **Modeling of text accumulation system for collective knowledge with text mining methods**
Presenter: **Ken Nittono**, Hosei University, Japan

As the spread of Internet communication services and mobile terminals, the importance of statistical approach for technology to process large amount of text data which are exchanged there has been increasing further. In this research, modeling of systematization for knowledge extraction from such bunch of text using text mining methods and issues for its implementation are considered. In order to extract significant information from obtained large-sized text which is observed as frequently exchanged sequential messages or its gathered set, applying statistical methods such as latent semantic analysis or association rule and automation of its extracting process become the keys to usability. And from the point of view of utilization of the extracted information as collective knowledge, systematic accumulation of the information and storing as tractable or easily viewable format become important. In this study, the process of the system achieving from extraction to accumulation automatically is modeled. In this case, web based system is assumed as a prototype for its implementation; however, the modeling approach is aimed to become functional internally for various kind of application programs or Internet cloud services.

C1627: **Bayesian hierarchical modelling of spatial extremes: an application to extreme low temperatures in Northern Finland**

Presenter: **Emeric Thibaud**, EPFL, Switzerland
Co-authors: Anthony Davison

Stimulated by the need for better risk assessment associated to rare climatic events, extreme value methods have greatly evolved during the last decade. Although methods to model extremes of univariate time series using generalized extreme-value and Pareto

distribution are well-known, flexible models for extremes of spatial processes are still needed. Brown-Resnick and extremal-t models have proven to be well-suited for modelling extremes of complex environmental processes, but their full density function cannot be calculated in general. Inference has therefore been based on composite likelihoods, resulting in a loss in efficiency compared to full likelihoods but also preventing the widespread use of these models in Bayesian inference. Recent advances in extreme value theory have shown how a likelihood function can be calculated in some particular cases. Using this, we propose the construction of a Bayesian hierarchical model for extreme low temperature in Northern Finland, allowing both the complex modelling of marginal distributions of extremes and an appropriate treatment of extremal dependence.

C1613: **A variable selection in linear regression models based on Tabu Search method**
Presenter: **Kannat Na Bangchang**, Thammasat University, Thailand

This research has proposed a variable selection method based on the Tabu search for multiple linear regression models. In this study two objective functions used in the Tabu search are mean square error (MSE) and the mean absolute error. The results of Tabu search are compared with the results obtained by stepwise regression method based on the hit percentage criterion. The simulations cover the both cases, without and with multicollinearity problems. Without multicollinearity problem, the hit percentages of the stepwise regression method and Tabu search using the objective function of MSE are almost the same but slightly higher than the Tabu search using the objective function of the mean absolute error. But with multicollinearity problem the hit percentages of Tabu search using the objective function of MSE or the mean absolute error are higher than the hit percentage of the stepwise regression method. Additionally, the correlation coefficients between the independent variables X1 and X4 are higher; yields the hit percentages are lower.

C1607: **An attitude survey of statistics and an analysis of the basic skills of mathematics in charge subjects**
Presenter: **Mie Fujiki**, Doshisha University, Japan

The idea is to consider how people that teach statistics should pay attention to attain the purpose of the university students to learn statistics. Japanese university students have a high tendency to feel weak in mathematics, especially liberal arts students. In Japan, it is important to learn how to gather data, interpret and analyze data through the statistics subjects, because required these skills. Therefore, we investigate how beginners of learning statistics have consciousness of statistics before a beginning of the lecture, and find out whether they bring about changes in their attitudes after the last lecture. For these investigations, we carry out a questionnaire survey. Moreover, we conduct a test on basic mathematics at the same time as the survey. We compare basic skills of mathematics and understandings of statistics in each class using correlation analysis and multiple comparisons. We show whether scores of these tests are affected by their comprehension and consciousness of statistics.

C1592: **The efficiency comparison of tests for equality of two variances**
Presenter: **Treerit Chotisathiensup**, Thammasat University, Thailand
Co-authors: Kamon Budsaba, Suprenee Lisawadi

Monte Carlo studies were conducted to compare on robustness and power of four statistical tests for equality of two variances. The equality of variance tests were (1) a nonparametric Wald test (called R test), (2) the Brown-Forsythe test, (3) the Levene test and (4) the F test where each test was analyzed under five distributions. The distributions were (1) normal, (2) chi-square, (3) exponential, (4) gamma and (5) Weibull distributions. The results from the Monte Carlo simulation studies demonstrated the Brown-Forsythe test was very good in terms of robustness, but it was poorly in terms of power when compared to the other tests. The F test performed well when the distribution is normal at level of significance 0.05. The R test outperformed the Levene test in terms of robustness and power at level of significance 0.01. The R test was nearly as robust as the Brown-Forsythe test at significance level 0.01 and nearly as power as the F test at significance level 0.05.

C1569: **Estimation based on covariances from multiple one-step randomly delayed measurements with noise correlation**
Presenter: **Raquel Caballero-Aguila**, University of Jaen, Spain
Co-authors: Hermoso-Carazo Aurora, Linares-Perez Josefa

The aim is to study 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.

C1408: **Estimation of the weighted kappa coefficient subject to case-control design**
Presenter: **Jose Antonio Roldan Nofuentes**, University of Granada, Spain

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.

C1715: **Spatial analysis of tuberculosis using space-time scan statistics to detect disease clusters**
Presenter: **Karuthan Chinna**, University Malaya, Malaysia
Co-authors: Alfred Aldrin, Yut Lin Wong

The aim of this population based study is to test a large set of TB cases for the presence of geographical clusters using the geographical information systems (GIS) and spatial scan statistics. A total of 7,071 TB cases were registered in Kuala Lumpur, the capital of Malaysia, between 2008 and 2012. In this study, first, all the cases were geocoded into 11 electoral constituency divisions based on residence at the time of diagnosis. The spatial and space-time scan statistics were then used to identify clusters of TB occurrence. In the purely spatial analyses, the most significant clusters were identified in density populated residential and business areas of Cheras, (2008, 2009, 2012) and Segambut (2008, 2009). Clusters were also identified in other areas within the city centre, Titiwangsa(2009), Lembah Pantai (in 2009), Bukit Bintang (2011) and Setiawangsa (2012). In the space-time analysis, the most likely cluster was the highly populous residential areas of Batu (2009) and Lembah Pantai (2011). The spatial and space-time scan statistics are effec-

tive ways of describing circular disease clusters. The spatial scan statistics methodology used in this study has a potential use in surveillance of tuberculosis for detecting the true clusters of the disease.

C1691: **Robust profiling of site index**
Presenter: **Manuela Souto de Miranda**, University of Aveiro, Portugal
Co-authors: Conceição Amado, Margarida Silva
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 Euclyptus 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.

PS04 Room Espace Motta POSTER SESSION IV Chair: Cristian Gatu

C1570: **Classification and biomarker identification for myopathic disorders.**
Presenter: **Markus-Hermann Koch**, Ruhr-University Bochum, Germany
Co-authors: Alexandra Maerkens, Rudolf Kley, Julian Uszkoreit, Matthias Vorgerd, Katrin Marcus, Martin Eisenacher
For skeletal muscle biopsies from 67 patients suffering from myofibrillar myopathies (mainly desminopathy, filaminopathy, myotilinopathy and titinopathy) our cooperation partners used microdissection and Orbitrap mass spectrometers to generate two quantitative spectral counting peptide datasets for each subject; one obtained from aggregate and one from aggregate free tissue. Using these data we built and validated nu-svm-based classifiers for the specific disorder subtypes as well as a detection for potential biomarker sets based on a workflow of hypothesis tests. Classification and detection was done in R.

C1581: **Extendint the exact test of Barnard to non-inferiority**
Presenter: **Felix Almendra-Arao**, UPIITA del Instituto Politecnico Nacional, Mexico
Co-authors: David Sotres-Ramos, Magin Zuniga-Estrada
George Alfred Barnard in 1945 presented an unconditional exact test to compare two independent proportions. By construction, critical regions of this test accomplish the very useful property of being Barnard convex sets. On the other hand, there are empirical findings suggesting that Barnard's test is the most generally powerful. Calculation of critical regions for this test is complicated due that they are constructed in an iterative form until is obtained a test size, as close as possible to the nominal significance level and less than or equal to it. The main goal of this work is to present an extension, to non-inferiority, of this very leading test. This extension was constructed for any dissimilarity measure, also were constructed tables for the difference between proportions. Besides were calculated the critical regions and test sizes for this extended test for several configurations of sample sizes, nominal significance levels and non-inferiority margins. A program written by the authors in the R environment was used for these calculations.

C1584: **New statistical function to discriminate metagenomic microbial community relative species abundance profiles**
Presenter: **Toni Monleon-Getino**, University of Barcelona, Spain
Co-authors: Jorge Frias-Lopez
There has been a great interest in associating specific groups of organisms with health/disease. We proposed that inflammatory diseases such as periodontitis and Crohn's disease not only alter the composition of the human microbiome but also its structure, focusing our interest in changes in relative species abundance patterns which have been widely used in ecology to describe the structure of living communities. We present the function dmcmetagen as a new R function for discriminate microbial community profiles of species abundance and we explain its use by means of an example of persons affected by Crohn's disease and periodontitis. This function first fits the metagenomic profile of abundance and richness distribution of each patient using a nonlinear regression, following an analysis of its biodiversity (Shannon/Simpson index) and finally a linear discriminant analysis of the data. The early results indicate that it is possible to discriminate between Crohn's disease (94.7% well classified) and Periodontitis (76.4% well classified) groups. The novelty of this work is that it confirms that the metagenomic microbial community species abundance distribution discriminates better than its composition and this issue can be helpful in the disease diagnosis.

C1587: **Stable variables in car insurance**
Presenter: **Amel Laouar**, University of Science and Technology Houari Boumediene USTHB, Algeria
Co-authors: Kamal Bouketala, Rachid Sabre
For a good modelling of the risk or claim process, it's important to know the distribution of claim amounts, as well as the frequency of claims. For this we propose to study real data from an Algerian insurance company. We assume that the claim amounts are infinite variance and we consider essentially the case of stable variables. Using various graphical and statistical tests we validate our hypotheses while estimating the four parameters of the corresponding distribution.

C1588: **An empirical comparison of homogeneity of variance tests**
Presenter: **Yada Pornpakdee**, Thammasat University, Thailand
Co-authors: Kamon Budsaba, Wararit Panichkitkosolkul
The objective of this research was to compare homogeneity tests of variances when sample sizes are equal. Analysis of means for variances (ANOMV), Samiuddin cube root test, and Bartlett's test were compared under normal distribution. Analysis of means for variances version of Levene's test (ANOMV-LEV), analysis of means for variances version of transformed ranks (ANOMV-TR), Levene's test, modified Levene's test, and trimmed mean Levene's test were compared under normal distribution, t distribution, lognormal distribution, double exponential distribution, gamma distribution, and logistic distribution. The comparison criterion was the capability to control type I error rate and its empirical power at significance level 0.05. The data were simulated by the Monte Carlo technique with 10,000 time replications. In the case of the normal population, Samiuddin Cube Root test and ANOMV test can control the type I error rate. However, Bartlett's test showed highest empirical powers. In the case of the non-normal population, modified Levene's test can control of type I error rate for all distributions. ANOMV-LEV test, ANOMV-TR test, Levene's test, modified Levene's test and trimmed mean Levene's test performed empirical power all distributions.

C1589: **The efficiency comparison of test for differences among several population means under heterogeneity of variances**

Presenter: **Uparittha Intarasat**, Thammasat University, Thailand
Co-authors: Kamon Budsaba, Saengla Chaimongkol
The purpose of this study is to compare the efficiency of the statistical tests for testing differences among several population means under heterogeneity of variances. Heteroscedastic analysis of means test (HANOM), analysis of mean test (ANOM), Welch test,

Brown-Forsythe test and Analysis of variance F-test (ANOVA-F) under 3 and 5 group means are investigated. The distributions of considered population are normal, beta, t and chi-square. The methods are compared by considering the ability to control the type I error rate and empirical power. The test is based on 0.05 levels of significance. In case of the 3 group and 5 group means, HANOM test can control type I error rate. HANOM, ANOM, Welch test, Brown-Forsythe test and ANOVA-F exhibit highest empirical powers. However, Welch test has lower empirical power test of normal distribution and beta distribution for all sample sizes in case the number of treatment groups is 5 ($k = 5$).

C1591: **Consideration of probability of failure in antiretroviral therapy using a stochastic model**

Presenter: **Takahiko Ueno**, St Marianna University School of Medicine, Japan

Co-authors: Shinobu Tatsunami

The purpose is to develop a simple dynamic model that expresses the importance of adherence to antiretroviral therapy. As we reported to the previous COMPSTAT in 2012, this model could describe the divergence of viral concentration when the interruption of drug administration occurs on two successive days between two periods with perfect adherence. In the present study, we revised some of minor terms in the equation, and tried another type of simulation. Our viral dynamics are composed of three parts: viral replication, suppression by immune activity, and elimination by drugs. Therefore, the simplest dynamic equation is containing three main variables of viral concentration, magnitude of the immune activity, and concentration of antiviral drug. The most important assumption in our formulation is that the probability of appearance of a drug-resistant virus depends on the time derivative of viral concentration. Under this assumption, the dynamic equation can describe the divergence of viral concentration even after the viral concentration attains the lower detectable limit. For example, if the interruption of drug administration occurs every other day in one week between long periods with perfect adherence, divergence of viral concentration within a month occurred 18 times in 10000 runs using different sequences of random numbers. The probability of the divergence was smaller compared to the two days of interruption under the same viral parametric conditions. The present results might be able to explain the difficulty of strategic treatment interruptions in antiretroviral therapy.

C1690: **Maximum simulated likelihood estimation of Thurstonian models**

Presenter: **Manuela Cattelan**, University of Padova, Italy

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 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.

C1730: **FEM-EVA framework for statistical regression analysis of extreme events**

Presenter: **Olga Kaiser**, Universita della Svizzera italiana, Switzerland

Co-authors: Illia Horenko

Data-based analysis of extreme events is a prominent problem in climate and weather research. Thereby, the main objective is the identification of (i) the statistical/stochastic models describing the dynamics of extremes, (ii) the spatio-temporal structure of model parameters and (iii) the most significant covariates that impact the probabilities of extremes. We propose a nonstationary and semiparametric framework for spatio-temporal statistical regression analysis for extremes accounting for systematic missing covariates, numerical instability and computational efficiency. Based on extreme value theory (EVA) and Finite Element time series analysis methods (FEM), the resulting FEM-EVA approach allows a well-posed problem formulation and goes beyond probabilistic a priori assumptions of methods for analysis of extremes based on, e.g., non-stationary Bayesian mixture models, smoothing kernel methods or neural networks. Further, FEM-EVA provides a pragmatic nonparametric, nonstationary and anisotropic description of the spatial dependence structure. We demonstrate the performance of FEM-EVA framework on test cases and real data with respect to systematically missing covariates, information content of the models, and interpretability of the models.

C1731: **FEM framework for time series analysis with missing data**

Presenter: **Dimitri Igdalov**, Universita della Svizzera italiana, Switzerland

Co-authors: Susanne Gerber, Illia Horenko

By approaching data analysis problems we are often confronted with missing values in observations. Thereby, following problems arise: estimation of model parameters in presence of missing values and reconstruction of missing values. Here we present a methodology for simultaneous model parameters estimation and missing values reconstruction beyond usual a priori assumptions. In particular, we extend the Finite Element Methods of time series analysis with Bounded Variation of model parameters (FEM-BV) towards handling of missing values. FEM-BV is a general purpose computational data analysis framework that can handle different datatypes (including real-valued and categorical data) and goes beyond stationarity and homogeneity assumptions. Exploiting information theory it finds the most descriptive model of the underlying dynamics. The underlying dynamics is described by a set of local stationary models and a nonstationary switching process. The optimal parameter set is obtained by constrained optimization of the corresponding objective function. To extend the FEM-BV framework towards dealing with missing data the objective function is appropriately reformulated. The overall procedure results in an alternating convergent optimization with respect to model parameters and missing values. We demonstrate the performance of FEM-BV in presence of missing data on test cases and on real data.

Wednesday 20.08.2014 11:15 - 13:00 Parallel Session E

CS16 Room 3 ROBUST REGRESSION Chair: Stefan Van Aelst

C1319: **Random start forward searches for detecting mixtures of regression models**
Presenter: **Anthony Atkinson**, London School of Economics, United Kingdom
Co-authors: Marco Riani, Andrea Cerioli, Domenico Perrotta
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. An example of international trade data is used to illustrate the diagnostic use of random start forward searches to reveal mixtures of regression models. The method identifies two models. Forward plots of residuals as increasing numbers of observations are used to estimate the parameters of the models are extremely informative about the structure. In contrast, a single robust fit to all the data completely fails to reveal the structure; robust estimation and outlier detection lead to removal of any evidence of inhomogeneity.

C1484: **A kernel based robust regression**
Presenter: **Eufrasio de Andrade Lima Neto**, Federal University of Paraiba, Brazil
Co-authors: Francisco de Assis Tenório De Carvalho, Marcelo Rodrigo Portela Ferreira, Pedro Monteiro Almeida Junior
The use of robust regression methods is common in practical situations due to the presence of outliers. This paper proposes a robust regression method that re-weighted the outliers observations considering kernel functions. The convergence of the parameter estimate algorithm is guaranteed with a low computational cost. A comparative study between the proposed kernel based robust regression method (KRR) against some classical robust approaches (WLS, M-Estimator, MM-Estimator, LTS, L1 regression) and the OLS method is considered. The performance of the methods has been evaluated in terms of the bias and MSE of the parameter estimates. We have considered synthetic datasets with X-axis outliers, Y-axis outliers and leverage points, in a Monte Carlo simulation framework with 10.000 replications, different sample sizes and percentage of outliers. The results have demonstrated that our approach presented a competitive performance or the best performance in simulation scenarios that are similar to those found in real problems. The KRR method has presented a similar performance to OLS method in the scenario without outliers. Moreover, the KRR method presented about half of the computational cost if compared with the methods LTS and MM-Estimator. An application with a real dataset has showed the usefulness of the proposed method.

C1575: **Robust test of restricted model**
Presenter: **Jan Amos Visek**, Charles University in Prague, Czech Republic
The purpose is to propose a test of restricted/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.

C1334: **Robust penalized regression estimator**
Presenter: **Kangmo Jung**, Kunsan National University, Korea, South
Penalized regression estimators have recently been spotlighted, because they achieve estimation of coefficients and variable selection simultaneously. However, regression outliers or leverage points can influence their performance. A robust penalized regression estimator to regression outliers and leverage points is proposed. For the loss function in a penalized regression model the least absolute deviation estimator is more robust than the least squares error estimator. However, the former is sensitive to leverage points of the predictors, even though it is robust to the regression outliers. A weighted version of least absolute deviation is proposed. For a penalty function the smoothly clipped absolute deviation is used because it has the oracle property. Since the penalty function is not convex, a local approximation algorithm is proposed and the tuning parameter is based on a Bayesian information criterion. Numerical simulations show that the proposed estimator is effective to analyze contaminated data.

C1618: **Robust and consistent variable selection for generalized linear and additive models**
Presenter: **Marco Avella-Medina**, University of Geneva, Switzerland
Co-authors: Elvezio Ronchetti
Generalized linear models (GLM) and generalized additive models (GAM) are popular statistical methods for modelling continuous and discrete data both parametrically and nonparametrically. In this general framework we consider the problem of variable selection through penalized methods by focusing on resistance issues in the presence of outlying data and other deviations from the stochastic assumptions. We propose robust penalized M-estimators and study their asymptotic properties. In particular we show that robust counterparts of the adaptive lasso and the nonnegative garrote satisfy the oracle properties. Our results extend the available theory from linear models to GLM and GAM, from classical to robust estimation and from fixed parameters to a high dimensional setting for GLM. Finally, we illustrate the finite sample performance of our method by a simulation study in a Poisson regression setting.

CS19 Room 6 METHODOLOGICAL STATISTICS I Chair: Jose E. Chacon

C1342: **Generalizing the multivariate normal distribution for accounting excess kurtosis: an application to model-based clustering**
Presenter: **Luca Bagnato**, Catholic University of the Sacred Heart - Milan, Italy
Co-authors: Antonio Punzo, Maria Grazia Zoia
For continuous multivariate data, statistical inference is commonly focused on elliptical distributions, the multivariate normal being the most classical choice. However, for many applied problems, the kurtosis of the normal distribution is lower than required. To overcome this problem, the multivariate leptokurtic-normal is proposed; it has a closed-form and is obtained as a generalization, by a particular class of orthogonal polynomials, of the classical multivariate normal distribution. The price to pay for this generalization is a single parameter which coincides with the excess kurtosis. Two parameter estimation techniques, the method of moments and the maximum likelihood, are discussed and mixtures of multivariate leptokurtic-normal distributions are also presented as a tool for robust model-based clustering and classification. An EM algorithm is presented to obtain maximum likelihood estimates of the mixture parameters and a simulation study is finally performed to evaluate parameter recovery and classification performance.

C1346: **A Gini-based stationarity test**
Presenter: **Amit Shelef**, Shamoon College of Engineering, Israel
A Gini-based statistical test for stationarity is suggested. This test is based on the well-known Dickey-Fuller test, where the ordinary least squares (OLS) regression is replaced by the semi-parametric Gini regression in modeling the autoregressive process. A residual-based bootstrap is used for finding critical values. The Gini methodology is a rank-based methodology that takes into account both

the variate values and the ranks. Therefore, it provides robust estimators that are rank-based, while avoiding loss of information. Furthermore, the Gini methodology relies on first-order moment assumptions, which makes it valid for a wide range of distributions. Simulation results validate the Gini-based test and indicate its superiority in some design settings, when compared to other available procedures. The Gini-based test opens the door for further developments such as a Gini-based cointegration test.

C1492: **Unifying approach to the shape and change-point hypotheses in a general exponential family**
Presenter: **Chihiro Hirotsu**, Meisei University, Japan
Co-authors: Harukazu Tsuruta

A unifying approach is presented in three ways. First it combines the shape and change-point hypotheses which have been developed in two different streams of statistics. Then the shape hypotheses such as monotone, convex and sigmoidal are approached in a systematic way. They are corresponding to step-type change, slope change and inflection point models, respectively. The relationship comes from that the corner vectors of the polyhedral cone defined by the shape hypotheses define the elements of the respective change-point models. The unification is important practically also since in monitoring the spontaneous reporting of adverse events it is useful to detect a change-point as well as increasing or downturn tendency. The well known isotonic regression has no obvious optimality to such a restricted parameter space and too complicated to extend beyond the monotone hypothesis in the normal model. Instead we propose maximal standardized cumulative, doubly cumulative and triply cumulative sum statistics directly derived from a complete class lemma for the test of restricted hypothesis, which leads to an elegant algorithm for probability calculation because of Markov properties of the subsequent component statistics. Finally because of simple structure the approach can be extended generally to an exponential family including discrete models.

C1451: **Change-point approach to multiple testing**
Presenter: **Zdenek Hlavka**, Charles University in Prague, Czech Republic
Co-authors: Marie Huskova

A rigorous approach to multiple hypotheses testing is needed in many real-life situations. Typically, a Bonferroni-type adjustment increases all p-values in order to control either the family-wise error rate or the false discovery rate. However, the structure of the observed data often calls for a more appropriate and powerful solution. Using gender-specific growth curves as a motivation, we propose a simple two-sample gradual change-point model in order to develop bootstrap-based tests and confidence intervals concerning the location of the unknown change-point. In this way, many two-sample t-tests are replaced by a single test concerning only the change-point and adjustments for multiple hypotheses testing thus become unnecessary.

C1548: **Sequential statistical sensitivity analysis for detecting a change point**
Presenter: **Kuniyoshi Hayashi**, Okayama University, Japan
Co-authors: Koji Kurihara

In statistics, assuming that population parameters of target datasets stay flat, we generally perform statistical diagnostics for a target statistical model based on these parameters or directly assess these parameters using statistical sensitivity analysis. However, when a new input datum sequentially comes into the existing dataset and the characteristics of the population do not match the parameters of the input datum, we cannot exactly evaluate and detect large influential observations, called outliers. Therefore, the aim is to extend the existing statistical sensitivity analysis based on influence functions, we propose an approach for detecting a change point: a time point for changing the population parameters of a new-comer or a turning point to re-optimize the tuning parameters for a target statistical model. With our proposed method and the traditional statistical diagnostics based on influence functions, we can detect not only an outlier but also a change point at each time point. Then, we can perform statistical diagnostics more flexibly than ever. Finally, we show the performance of our proposed approach through some understandable simulation studies.

CS41 Room 20 CONTRIBUTIONS TO STATISTICS OF EXTREME VALUES I Chair: Armelle Guillou

C1300: **Grouping seasonal time series using extreme value analysis**
Presenter: **Ann Maharaj**, Monash University, Australia
Co-authors: Pierpaolo D'Urso, Andres Alonso

The analyses of extreme temperatures and sea levels are important tasks in this era of awareness of the effects of climate change. Many authors have used extreme value analysis to study sea level extremes and temperature extremes. While previous studies using clustering methods focused on grouping together locations based on predictive distributions, the focus of this study is grouping the time series across the available record using both a variety of clustering and classification methods. Whereas previous studies focused mainly on specific applications, in this study, simulation studies are also conducted to evaluate the performance of the clustering and classification methodologies based on specific feature sets for more general use with seasonal time series. In particular, the feature sets considered are block maxima and parameter estimates of shape, mean and standard deviation, obtained from fitting the generalised extreme value density function to the block maxima. The series are clustered by using the conventional k-mean and k-medoids as well as fuzzy c-means, weighted fuzzy c-means, fuzzy c-medoids and weighted fuzzy c-medoids methods for which iterative solutions are obtained. Additionally, the series are classified using the K-nearest neighbors algorithm and pattern recognition neutral networks.

C1488: **Bayesian semiparametrics for modelling the clustering of extreme values**
Presenter: **Thomas Lugrin**, EPFL, Switzerland
Co-authors: Anthony Davison, Jonathan Tawn

Risk estimates for time series with short-range dependence are often obtained using the peaks over threshold method. Such estimates may be badly biased, however, because they do not properly account for dependence in clusters of extremes. An alternative characterization of cluster maxima allows separate consideration of both the marginal distribution of exceedances over the threshold and of the dependence structure. This characterization uses the extremal index at sub-asymptotic levels, for which a formulation can be derived from the conditional Heffernan–Tawn model, which involves specifying the conditional distribution and two stages of inference. Here we estimate this distribution by a semiparametric Bayesian approach using a dependent Dirichlet process, allowing us to fit a model for clusters of extremes in a single step. The ideas, illustrated using simulated and real data, can result in substantially improved estimates of high quantiles for time series.

C1519: **Kernel estimation of extreme risk measures for all domains of attraction**
Presenter: **Jonathan El Methni**, University of Geneva, Switzerland
Co-authors: Laurent Gardes, Stephane Girard

Value-at-risk, Conditional Tail Expectation, Conditional Value-at-risk and Conditional Tail Variance are classical risk measures. In statistical terms, the Value-at-risk is the upper α-quantile of the loss distribution where $\alpha \in (0,1)$ is the confidence level. Here, we focus on the properties of these risk measures for extreme losses (where $\alpha \downarrow 0$ is no longer fixed). To assign probabilities to extreme losses we assume that the distribution satisfies a von-Mises condition which allows us to work in the general setting, whether the

extreme-value index is positive, negative or zero *i.e.* for all domains of attraction. We also consider these risk measures in the presence of a covariate. The main goal of this communication is to propose estimators of the above risk measures for all domains of attraction, for extreme losses, and to include a covariate in the estimation. The estimation method thus combines nonparametric kernel methods with extreme-value statistics. The asymptotic distribution of our estimators is established and their finite sample behavior is illustrated on simulated data and on a real data set of daily rainfall.

C1673: **Statistical modelling in time series extremes: An overview and new steps**
Presenter: **Manuela Neves**, University of Lisbon and CEAUL, Portugal
Co-authors: Clara Cordeiro

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.

C1386: **Sloshing in the LNG shipping industry: risk modelling through multivariate heavy-tail analysis**
Presenter: **Antoine Dematteo**, Telecom ParisTech, France
Co-authors: Nicolas Vayatis, Mathilde Mougeot, Stephan Clémençon

In the liquefied natural gas (LNG) shipping industry, the phenomenon of sloshing can lead to the occurrence of very high pressures in the tanks of the vessel. The issue of modelling or estimating the probability of the simultaneous occurrence of such extremal pressures is now crucial from the risk assessment point of view. Heavy-tail modelling is applied to the study of sloshing. Multivariate heavy-tailed distributions are considered, with Sloshing pressures investigated by means of small-scale replica tanks instrumented with $d > 1$ sensors. When attempting to fit such nonparametric statistical models, one naturally faces computational issues inherent in the phenomenon of dimensionality. The primary purpose is to overcome this barrier by introducing a novel methodology. For d-dimensional heavy-tailed distributions, the structure of extremal dependence is entirely characterised by the angular measure, a positive measure on the intersection of a sphere with the positive orthant in $\mathbb{R}^d$. As d increases, the mutual extremal dependence between variables becomes difficult to assess. Based on a spectral clustering approach, a low dimensional approximation to the angular measure may be found. The nonparametric method proposed for model sloshing has been successfully applied to pressure data. The parsimonious representation thus obtained proves to be very convenient for the simulation of multivariate heavy-tailed distributions. Besides confirming its performance on artificial data, the methodology has been implemented on a real data set specifically collected for risk assessment of sloshing in the LNG shipping industry.

CS43 Room 13 CONTRIBUTIONS TO APPLIED BAYESIAN STATISTICS Chair: Anne Philippe

C1306: **Bayesian regression model with a random network**
Presenter: **Simon Cheung**, The Open University of Hong Kong, China
Co-authors: Tommy Cheung

Bayesian Model is an increasingly popular statistical method for the investigation of relationships between variables of interests. When the values of these variables are observed from individuals of a population, an underlying network often connects those individuals. The network is often dynamic in nature, either through re-wiring or natural growth. It typically follows a preferential attachment scheme which can be applied to direct or un-direct networks, and is the basis to the power-law degree distribution and small diameter properties of the network. A multivariate linear model is proposed to study relationships among variables from a population connected by such an underlying network. Together with the network generating mechanism, a Gibbs sampler is developed on the model to generate random numbers from the joint posterior distribution. At each Gibbs sampling step, the network is re-generated according to the updated posterior probabilities. It captures well the inherent rewiring and growth nature of real networks. Inference through the posterior distributions of parameters can be made to provide insights into the analysis.

C1552: **Bayes estimation of parameters for trend-renewal processes**
Presenter: **Ryszard Magiera**, Wroclaw University of Technology, Poland
Co-authors: Alicja Jokiel-Rokita

The problem of Bayes estimation of unknown parameters is considered for stochastic models determined by the trend-renewal process (TRP) which is defined to be a time-transformed renewal process, where the time transformation is given by a trend function. The TRP's, whose realizations depend on a renewal distribution as well as on a trend function, comprise the non-homogeneous Poisson processes and renewal processes and serve as useful reliability models for repairable systems. Some models of TRP's are considered for which statistical inferences are analytically intractable. One of the representatives of the TRP's is the TRP with a Weibull type renewal distribution and a power law trend function. Such TRP's, with distribution depending on three unknown parameters, will be called Weibull power law processes (WPLP's). In the Bayes estimation problems considered for the WPLP, various noninformative prior distributions are used and their influence on properties of the estimators obtained is examined. Under mean squared and absolute estimation errors and for some specific prior distributions we show advantage of the Bayes estimators over the maximum likelihood estimators, especially when the number of observed failures in the WPLP is small. The Markov Chain Monte Carlo methods used can also be applied to predict future failure time.

C1675: **Skewed Laplace approximation for censored data**
Presenter: **Ludger Evers**, University of Glasgow, United Kingdom

Many approximate inference techniques are based on a quadratic approximation to the loglikelihood / logposterior. Examples of this include asymptotic Wald-type confidence bands in the frequentist setting and Laplace approximations in the Bayesian setting. Both in the case of time-to-event data and data with detection limits, the loglikelihood, and thus the logposterior, can be heavily skewed, in which case these quadratic approximations yield very poor results and can lead to erroneous conclusions. This talk proposes a generalisation of the Laplace approximation which can account for the skewness in the logposterior and often yields results close to those obtained using full Bayesian inference involving MCMC. The results will be illustrated both using simulated and real-world examples.

C1663: **Noncompensatory multiple logistic regression model and its application**
Presenter: **Kensuke Okada**, Senshu University, Japan
Co-authors: Shin-ichi Mayekawa

Let us consider the analysis with binary outcome and multiple predictor variables. In conventional logistic regression model, the log odds (logit) of the outcome is related to the linear combination of the predictor variables (and possibly their interaction terms). This

means that a high score in one predictor compensates for the low score in others because weighted sum of predictors contributes to the logit of the outcome. Here, a different model is proposed. In this model, the outcome probability is represented as the product of logistic functions of each predictor. This means that the total outcome probability is the product of all the probabilities implied by the simple logistic regression of the outcome on each of the predictors. Therefore, a low probability from one predictor is not compensated by others. One- and two-parameter model can be considered for each of simple logistic regression components. The model is fitted using the Hamiltonian Monte Carlo algorithm. Our simulation study revealed that a widely applicable information criterion (WAIC) could be used to select the true model under the conditions that either existing or proposed model is correct. The proposed model is then applied to analyze the social science data.

C1688: **Bayesian blind source separation applied to the lymphocyte pathway**
Presenter: **Katrin Illner**, Helmholtz Research Center Munich, Germany
Co-authors: Christiane Fuchs, Fabian J. Theis

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.

CS56 Room 4 COMPUTATIONAL STATISTICS III Chair: Andreas Alfons

C1579: **Tree-based prediction on incomplete data**
Presenter: **Holger Cevallos**, Ghent University, Ecuador
Co-authors: Stefan Van Aelst

Appropriately handling missing data is an important issue in prediction problems. Therefore, 26 tree-based strategies have been compared in terms of prediction capability with data containing missing values. Techniques consisted of tree methods that either use surrogate decisions or are combined with an imputation method (single or multiple). To this end, a simulation study that introduces data MCAR, MAR and NMAR on two schemes and various fractions of missingness was performed on real-life and simulated datasets, covering both classification and regression problems. The imputation model is treated as part of the learning phase and is not allowed to use the response, so that the resulting procedures can predict individual cases. Simulations suggest that for data with a small amount of missingness it might be enough to fit an ensemble method (e.g. conditional random forests) in combination with surrogates or a previous single imputation. For data with larger amounts of missing values, combining a flexible multiple imputation method such as MICE with conditional random forests seems to be the safest strategy. Results also reveal that conditional bagging may emerge as a good and computationally cheaper alternative for the latter situation.

C1650: **Imputation of complex dependent data by conditional copulas: analytic versus semiparametric approach**
Presenter: **F. Marta L. Di Lascio**, Free University of Bozen-Bolzano, Italy
Co-authors: Simone Giannerini, Alessandra Reale

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 us 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.

C1498: **Weight choice by minimizing MSE for general likelihood averaging**
Presenter: **Ali Charkhi**, KU Leuven, Belgium
Co-authors: Gerda Claeskens, Bruce E. Hansen

In model averaging a weighted estimator is constructed based on a set of models, in contrast to model selection where a single model is selected based on 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 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.

C1495: **A general procedure to combine estimators**
Presenter: **Paul Rochet**, Nantes, France
Co-authors: Frédéric Lavancier

A general method to combine several estimators of the same quantity in order to produce a better estimate is proposed. The final estimator is expressed as a weighted average of the initial ones obtained under the minimal requirement that the weights sum to one. In this framework, the optimal weights, minimizing the quadratic loss, are entirely determined by the mean square error matrix of the vector of initial estimators. The solution is derived using an estimation of this matrix, which can be computed from the same dataset. We show that the solution satisfies a non-asymptotic oracle inequality and is asymptotically optimal, provided the mean square error matrix is suitably estimated. This method is illustrated on standard statistical problems: estimation of the position of a symmetric distribution, estimation in a parametric model, density estimation where it outperforms the initial estimators in most cases.

C1585: **Goodness of fit tests for the lognormal distribution**
Presenter: **Polychronis Economou**, University of Patras, Greece

Co-authors: Apostolos Batsidis, George Tzavelas

The lognormal distribution is not only one of the most widely used distributions in survival and reliability analysis but is also frequently used in many other fields. The lognormality assumption of the data is usually tested using the logarithmic data by applying one of the available tests for the normal distribution. In the present work two new goodness of fit tests for the lognormal distribution are proposed. The new procedures rely on a characterization property of the lognormal distribution which states that the Kullback Leibler measure of divergence between a p.d.f. of a non-negative r.v. X and its r-size weighted p.d.f.

$$f_r(x) = \frac{x^r}{/} E(X^r) f(x), x > 0$$

is symmetric only for the lognormal distribution. With a help of a simulation study the performance of the new procedures is compared with existing goodness of fit tests for the lognormal distribution. Finally, real data sets are used to illustrate the applicability of the proposed methods.

CS64 Room 5 CLUSTERING II Chair: Kim De Roover

C1344: **Time series clustering based on quantile autocovariances**
Presenter: **Borja Lafuente**, Universidade da Coruna, Spain
Co-authors: Jose Antonio Vilar

Time series clustering is an active research topic with applications in many fields. Unlike conventional clustering on static data objects, time series are inherently dynamic and hence the similarity searching must be governed by the behavior of the series over their periods of observation. In this study, a dissimilarity criterion aimed to compare quantile autocovariance functions is proposed to perform time series clustering. Results from an extensive simulation study show that the proposed metric outperforms a range of alternative dissimilarities proposed in the literature. Estimation of the optimal number of clusters is also discussed. In particular, a prediction-based resampling algorithm proposed by Dudoit and Fridlyand (2002) is adjusted to be applied in time series clustering based on quantile autocovariances. Several criteria to select the number of clusters are examined in new simulations.

C1705: **Consensus clustering of time series data**
Presenter: **Inci Batmaz**, Middle East Technical University, Turkey
Co-authors: Ayca Yetere Kursun, Cem Iyigun

The aim is 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.

C1631: **Tail-dependence clustering of financial time series**
Presenter: **Fabrizio Durante**, Free University of Bozen-Bolzano, Italy

In portfolio management a current practice for minimizing the risk consists of adopting some diversification techniques that are based, roughly speaking, on the selection of assets from sectors or regions that one believes to be weakly correlated. In fact, it is believed that diversification reduces the impact of simultaneous joint losses. To this end, cluster techniques for multivariate time series have been proposed in the literature that span from the use of correlation coefficient to the comparisons among the underlying univariate ARMA-GARCH. However, since diversification may fail when there is a change in the positive association among the markets in crisis period, it could be also useful to develop clustering procedures that reflect the behavior of different time series in extreme scenarios. Here, we propose some clustering procedures that focus their attention to the tail behavior of the joint distribution of involved time series. Specifically, we aim at creating groups of time series that tend to comove in their extreme values. Such methods are based on the use of suitable rank-correlation (i.e. copula based) measures of association. The performance of the proposed methodology will be illustrated in a simulation study and with empirical data.

C1574: **Bayesian cluster detection via adjacency modelling**
Presenter: **Craig Anderson**, University of Glasgow, United Kingdom
Co-authors: Duncan Lee, Nema Dean

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. A two stage approach is proposed, 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.

CS81 Room 19 CONTRIBUTIONS TO STATISTICS AND OPTIMIZATION IN FINANCE II Chair: Marguerita Giuzio

C1450: **A generalized description length approach for sparse and robust index tracking**
Presenter: **Margherita Giuzio**, EBS Universität fur Wirtschaft und Recht, Italy
Co-authors: Davide Ferrari, Sandra Paterlini

A new minimum description length criterion for index tracking is developed. It 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.

C1350: **A diagnostic criterion for approximate factor structure**
Presenter: **Elisa Ossola**, University of Lugano, Switzerland
Co-authors: Patrick Gagliardini, Olivier Scaillet

A simple diagnostic criterion for approximate factor structure in large cross-sectional equity datasets is built. Given a model for asset returns with observable factors, the criterion checks whether the error terms are weakly cross-sectionally correlated or share at least one unobservable common factor. It only requires computing the largest eigenvalue of the empirical cross-sectional covariance matrix of the residuals of a large unbalanced panel. The panel data model accommodates both time-invariant and time-varying factor structures. The theory for large cross-section and time-series dimensions is developed. No restriction is imposed on the relation between both dimensions. The empirical analysis runs on returns for about ten thousands US stocks from July 1964 to December 2012. Among several multi-factor models proposed in the literature, a model with zero factors in the errors cannot be selected.

C1516: **On singular and spurious solutions in finite mixture models**
Presenter: **Byungtae Seo**, Sungkyunkwan University, Korea, South
Co-authors: Daeyoung Kim

Finite mixture models have been widely used in modeling composite financial data. However, it is a challenging task to optimize the mixture likelihood especially with location-scale component densities. This is because the mixture likelihood is unbounded and has multiple modes. In this talk, we introduce several methods to avoid singular solutions. In addition, we investigate the nature of spurious solutions and propose some likelihood based methods to remove spurious solutions with examples.

C1680: **Transfer of semiparametric single index model in binary classification**
Presenter: **Muhammad-Anas Knefati**, Poitiers University, France
Co-authors: Farid Beninel

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 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.

C1702: **An empirical analysis of the Ross recovery theorem**
Presenter: **Markus Ludwig**, University of Zurich, Switzerland
Co-authors: Francesco Audrino, Robert Huitema

Implementing the recovery theorem requires the solution of two ill-posed problems. The first involves estimating the second derivative of the option pricing function from noisy and sparse market quotes to obtain state prices over a fixed domain. The second step entails the construction of a transition matrix that captures the dynamics of said state prices. Only then can recovery be obtained via spectral decomposition. We present a method based on Tikhonov regularized non-negative least squares to construct robust time-homogeneous Markov chains that not only yield an excellent fit to state prices at various maturity horizons, but pricing kernels that exhibit a sensible variation over time. Using daily snapshots of option prices on the S&P 500 index, we compute genuinely conditional and forward-looking risk-neutral and real-world moments and investigate their predictive information content. We find that changes in the recovered moments can be used to time the index, yielding strategies that not only outperform the market, but also are significantly less volatile.

Wednesday 20.08.2014 15:15 - 16:45 Parallel Session F

IS01 Room 3 VOLATILITY MODELLING AND FORECASTING Chair: Alessandra Amendola

C1321: **Option pricing with asymmetric heteroskedastic normal mixture models**
Presenter: **Jeroen Rombouts**, ESSEC Business School, France
An asymmetric GARCH in mean mixture model is proposed and a feasible way for option pricing within this general framework by deriving the appropriate risk neutral dynamics is provided. Out-of-sample prices of a large sample of options on the S&P 500 index from January 2006 through December 2011 are forecasted and dollar losses and implied standard deviation losses are computed. The results are compared with existing mixture models and other benchmarks like component models and jump models. Overall, the dollar root mean squared error of the best performing benchmark model is 28% larger than the best mixture model.

C1395: **A conservative test for the lag structure of assets realized volatility dynamics**
Presenter: **Francesco Audrino**, University of St Gallen, Switzerland
Co-authors: Lorenzo Camponovo, Constantin Roth
A conservative test is constructed to investigate the optimal lag structure for forecasting realized volatility dynamics. The testing procedure relies on the recent theoretical results showing the ability of the adaptive least absolute shrinkage and selection operator (adaptive lasso) in combining efficient parameter estimation, variable selection, and valid finite sample inference for time series regressions. In an application to several constituents of the S&P 500 index it is showed that (i) the optimal significant lag structure is time-varying and subject to drastic regime shifts that seem to happen across assets simultaneously; (ii) in many cases the relevant information for prediction is included in the first 22 lags, corroborating previous results about the accuracy and the difficulty to outperform out-of-sample the heterogeneous autoregressive (HAR) model; and (iii) some common features of the optimal lag structure can be identified across assets belonging to the same market segment.

C1471: **Combining information at different frequencies in multivariate volatility prediction**
Presenter: **Alessandra Amendola**, University of Salerno, Italy
Co-authors: Giuseppe Storti
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 issue 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, 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.

OS25 Room 4 FUNCTIONAL REGRESSION MODELS AND APPLICATIONS Chair: Ana M. Aguilera

C1390: **The functional linear array model estimated by boosting**
Presenter: **Sarah Brockhaus**, Ludwig-Maximilians-University Munich, Germany
Co-authors: Fabian Scheipl, Torsten Hothorn, Sonja Greven
The functional linear array model (FLAM) is introduced. It is a unified model class for functional regression models including scalar and functional responses in an additive model. Within the FLAM framework mean, quantile and generalized linear regression models are contained as special cases. The additive predictor can contain a variety of covariate effects. The current implementation supports linear, smooth and interaction effects of scalar and functional covariates. Taking advantage of the Kronecker product in the design matrix, the FLAM is represented as generalized linear array model to achieve computational efficiency. The array structure requires a common grid for all responses, but missing values are allowed. For estimation, a component-wise boosting algorithm is used that allows for numerous covariates and variable selection. The boosting algorithm minimizes the empirical risk and thus allows for many different model specifications by using an adequate loss function. Some features of the FLAM are illustrated using data on viscosity of resin over time depending on five experimental factors. A median regression model is fitted for the functional response depending on several scalar covariates. An implementation of these methods is provided in the R-package `FDboost` available on R-Forge.

C1466: **Generalized linear models for spatial functional data analysis**
Presenter: **Matthieu Wilhelm**, Université de Neuchâtel, Switzerland
Co-authors: Laura Sangalli
Adopting a functional data analysis approach, we propose a generalized linear regression model for the analysis of spatially distributed data from an exponential family distribution, able to efficiently deal with data occurring over irregularly shaped domains. The proposed generalized additive framework can handle all distributions within the exponential family, including binomial, Poisson and gamma outcomes. Specifically, we maximize a penalized log-likelihood function where the roughness penalty term involves a suitable differential operator of the spatial field over the domain of interest. In the simpler context of univariate smoothing problems, the idea of regularization with ordinary differential operators has already proved to be very effective and it is in general playing a central role in the functional data analysis literature. Space-varying covariate information is also be included in the model in a semi-parametric setting. The proposed model exploits advanced scientific computing techniques and specifically makes use of the finite element method, that provide a basis for piecewise polynomial surfaces.

C1407: **Varying-smoother models for functional responses**
Presenter: **Philip Reiss**, New York University, United States
Co-authors: Lei Huang, Huaihou Chen, Stan Colcombe
We consider estimation of a smooth function $f(t,s)$ when we are given functional responses of the form $f(t,\cdot)$ + error, but scientific interest centers on the collection of functions $f(\cdot,s)$ for different s. The motivation comes from studies of human brain development, in which t denotes age whereas s refers to brain locations. Analogously to varying-coefficient models, in which the mean response is linear in t, the "varying-smoother" models that we consider exhibit nonlinear dependence on t that varies smoothly with s. We focus on two spline-based approaches to estimating varying-smoother models: (i) methods that apply a tensor product penalty, and (ii) two-step methods consisting of an initial smooth with respect to t at each s, followed by a postprocessing step. For (i), we derive an

exact expression for a penalty proposed by Wood, and an adaptive penalty that allows smoothness to vary more flexibly with s. We also introduce "pointwise degrees of freedom," a new tool for studying the complexity of estimates of $f(\cdot, s)$ at each s. The different approaches to varying-smoother modeling are compared in simulations and with a diffusion tensor imaging data set.

C1441: **Linear discriminant analysis based on penalized functional PLS**
Presenter: **Ana M Aguilera**, University of Granada, Spain
Co-authors: M. Carmen Aguilera-Morillo

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.

OS29 Room 5 SURVEY SAMPLING Chair: Alina Matei

C1406: **Empirical likelihood confidence intervals for complex sampling designs using R**
Presenter: **Yves Berger**, University of Southampton, United Kingdom

The development of an R library which can be used for an empirical likelihood based inference, point estimation and confidence intervals, is studied. We will explain how to use this library and we will present the result of a simulation study which shows that the proposed empirical likelihood confidence interval may give better coverages than the approaches based on linearisation, bootstrap and pseudo empirical likelihood. Under complex sampling designs, estimators of interest may not have a normal sampling distribution. Hence standard confidence intervals based upon the central limit theorem may have poor coverages. We propose an empirical likelihood approach which gives design based confidence intervals. The proposed approach does not rely on the normality, variance estimates, design-effects, re-sampling, joint-inclusion probabilities and linearisation, even when the estimator of interest is not linear. It can be used to construct confidence intervals for a large class of complex sampling designs and complex estimators which are solution of an estimating equation. It can be used for means, regressions coefficients, quantiles, totals or counts even when the population size is unknown. It can be used with large and negligible sampling fractions. It also provides asymptotically optimal point estimators, and naturally includes calibration constraints. The proposed approach is computationally simpler than the pseudo empirical likelihood and the bootstrap approaches.

C1354: **Variance estimation for regression imputed quantiles and inequality indicators**
Presenter: **Eric Graf**, University of Neuchâtel, Switzerland

In a sample survey only a sub-part of the selected sample has answered (total non-response, treated by re-weighting). Moreover, some respondents did not answer all questions (partial non-response, treated through imputation). One is interested in income type variables. One further supposes here that the imputation is carried out by a regression. The idea presented by Deville and Särndal in 1994 is resumed, which consists in constructing an unbiased estimator of the variance of a total based solely on the known information (on the selected sample and the subset of respondents). While these authors dealt with a conventional total of an interest variable y, a similar development is reproduced in the case where the considered total is one of the linearized variables of quantiles or of inequality indicators, and that, furthermore, it is computed from the imputed variable y. By means of simulations on real survey data, one shows that regression imputation can have an important impact on the bias and variance estimations of inequality indicators. This leads to a method capable of taking into account the variance due to imputation in addition to the one due to the sampling design in the cases of quantiles.

C1534: **Adjustment for nonignorable nonresponse using latent homogeneous response groups**
Presenter: **Caren Hasler**, Université de Neuchâtel, Switzerland
Co-authors: Alina Matei

A setup in which nonignorable nonresponse is present in the survey is considered. In such a case, the unit response probabilities depend on the variable of interest. We assume that the variable of interest follows a mixture distribution (a typical example of such a variable is income). This allows us to highlight latent homogeneous response groups based on the variable of interest and auxiliary information. Two approaches are discussed. In both approaches, the membership group variable is unknown for the nonrespondents and is imputed using auxiliary information. In the first approach, the unit nonresponse is modelled using logistic regression including the membership group variable in the covariates. In the second approach, the unknown values of the variable of interest are imputed in each latent homogeneous group, based on available auxiliary information. The response probabilities are estimated using logistic regression with the variable of interest (imputed or observed) as a covariate. In both approaches, the 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.

C1555: **Multivariate outliers in incomplete survey data**
Presenter: **Beat Hulliger**, University of Northwestern Switzerland FHNW, Switzerland
Co-authors: Marc Bill

Multivariate outlier detection in incomplete survey data must take into account the often skew and sometimes semi-continuous nature of the distributions and the sample design as well as nonresponse adjustments. In addition missing values may hinder the methods. Usually outliers must be treated by imputation to enable a straightforward use of the data. Two outlier detection and imputation procedures which are implemented in an experimental package of the software R called modi are discussed: The BACON-EEM algorithm together with imputation based on the multivariate normal distribution and the epidemic algorithm for detection and imputation are applied to a data set from a business survey on various types of expenditures for environmental protection. The practical issues of dealing with the zero-inflated distributions and of choosing tuning constants for the parameters are also discussed.

OS36 Room 6 STATISTICS IN MEDICINE Chair: Valentin Rousson

C1393: **Flexible models for cure interval-censored data**
Presenter: **Vincent Bremhorst**, Université Catholique De Louvain, Belgium
Co-authors: Philippe Lambert

A common hypothesis in the analysis of survival data is that any observed unit will experience the monitored event if it is observed for a sufficiently long time. Alternatively, one can explicitly acknowledge that an unknown and unidentified proportion of the patient population under study is cured and will never experience the event of interest. The promotion time model, which is motivated using

biological mechanisms in the development of cancer, is one of the survival models taking this feature into account. The promotion time model assumes that the failure time of each subject is generated by the minimum of N independent latent event times with a common distribution independent of N. An extension is proposed which allows the covariates to influence simultaneously the probability of being cured and the time necessary for a cell to yield a detectable tumor. The latent distribution is estimated using a flexible Cox proportional hazard model where the logarithm of the baseline hazard function is specified using Bayesian P-splines. Moreover, in order to relax the linear assumption in the covariate structures, a flexible modelling of the effect of continuous variables is proposed using Bayesian P-splines. The context of interval censoring will be discussed.

C1453: **Using an intraclass odds-ratio as an alternative to kappa to assess the inter-rater reliability of binary measurements**

Presenter: **Isabella Locatelli**, University of Lausanne, Switzerland
Co-authors: Valentin Rousson

Inter-rater reliability of binary measurements is usually assessed using the concept of kappa coefficient (e.g. Cohen's kappa or Scott's kappa), which is known to be particularly difficult to interpret. One reason for this difficulty is that a kappa coefficient can be defined as a correlation between two exchangeable measurements made on a same subject, i.e. an intraclass correlation, a concept originally defined for continuous measurements. To measure an association between two binary variables it is in fact much more common to calculate an odds-ratio rather than a correlation, and we thus propose to assess inter-rater reliability of binary measurements by calculating the odds-ratio (instead of the correlation) between two exchangeable measurements made on a same subject, yielding the concept of "intraclass odds-ratio". In the same way that an intraclass correlation is smaller in magnitude than a classical correlation, the intraclass odds-ratio will be smaller than the classical odds-ratio, penalizing both systematic and random discrepancies between the raters. We explore the relationships of the intraclass odds-ratio with kappa, the probability of agreement and the marginal distribution of the data. In particular, a kappa value of 0.75, which is usually considered as the threshold for a good reliability, corresponds to an intraclass odds-ratio of at least 49, meaning a probability of concordance at least 49 times higher than the probability of discordance. This may suggest a more lenient approach to assess the inter-rater reliability of binary data, considering e.g. an intraclass odds-ratio of 25 as being already a good reliability. We also propose an extension of the concept of intraclass odds-ratio to the case of more than two raters per subject, and we develop an explicit formula to calculate a valid confidence interval. We finally illustrate the usefulness of the intraclass odds-ratio concept on medical data.

C1427: **Approximate Bayesian model selection with the deviance statistic**
Presenter: **Leonhard Held**, University of Zurich, Switzerland
Co-authors: Daniel Sabanes Bove

Bayesian model selection poses two main challenges: the specification of parameter priors for all models, and the computation of the resulting Bayes factors between models. There is now a large literature on automatic and objective parameter priors, which unburden the statistician from eliciting them manually in the absence of substantive prior information. One important class is g-priors, which were recently extended from linear to generalized linear models. We show that the resulting Bayes factors can conveniently and accurately be approximated by test-based Bayes factors using the deviance statistic. For the estimation of the hyperparameter g, we show how empirical Bayes estimates correspond to shrinkage estimates from the literature, and propose a conjugate prior as a fully Bayes alternative. Considerable computational gains are obtained which enable an exhaustive evaluation of the model space in moderate size variable selection problems without the need to employ MCMC methods. We illustrate the methods with the development of a clinical prediction model for 30-day survival in the GUSTO-I trial, and with variable and function selection in Cox regression for the survival times of primary biliary cirrhosis patients.

C1533: **Showing statistically the existence of disease subtypes using model based clustering: how feasible is it?**
Presenter: **Aziz Chaouch**, IUMSP-CHUV, Switzerland
Co-authors: Alex Randriamiharisoa, Valentin Rousson

A recurrent concern of medical research is to determine whether patients suffering from a disease form a homogeneous group or whether they form a heterogeneous group with different subtypes of the disease, implying possibly different treatments. One elegant way to tackle this question is to adopt a model based clustering approach where one fits patients data using different statistical models involving mixtures of normal distributions with varying number of components and compares them statistically using e.g. the BIC criterion. While attractive conceptually, this approach needs a large sample size to achieve a reasonable statistical power as illustrated with anthropometric data. In this presentation, we investigate via simulation whether the statistical power can be increased by increasing the dimension of the data under various constraints. While the answer was generally negative, we identified one favorable situation: the case of (within-cluster) uncorrelated bivariate data with a same (within-cluster) variance for the two variables. Interestingly, the power was decreasing when considering more than two variables. While emphasizing that detecting clusters in data is a difficult task in practice, these results should encourage researchers to carefully select their variables before entering them into a model based clustering algorithm.

CS48 Room 20 CONTRIBUTIONS TO LONGITUDINAL DATA ANALYSIS I Chair: Anuradha Roy

C1296: **Statistical analysis of multivariate longitudinal data**
Presenter: **Xinyuan Song**, Chinese University of Hong Kong, China

A hidden Markov latent variable model for analyzing multivariate longitudinal data is developed. The latent variable model is defined in a structural equation modeling framework, in which a measurement equation measures latent constructs through multiple longitudinal responses, and a structural equation examines the interrelationships among observed and latent variables. To reveal the dynamic patterns and possible heterogeneity of the above mentioned associations and interrelationships, a mixed hidden Markov model is introduced to model the transition probabilities across different latent states. A maximum likelihood procedure is developed to analyze the proposed model. The Monte Carlo expectation conditional maximization algorithm is employed to obtain the estimates of unknown parameters. The Gibbs sampler coupled with the forward-backward recursion sampling is implemented in the Monte Carlo expectation-step. The asymptotic properties of the parameter estimator and test statistics for testing the heterogeneity of model parameters are established. Simulation studies are conducted to assess the performance of the proposed methodologies. The model is applied to a longitudinal study of cocaine use.

C1536: **Robust estimation in joint mean-covariance regression model for longitudinal data**
Presenter: **Wing K Fung**, University of Hong Kong, China

We develop robust estimation for the mean and covariance jointly for the regression model of longitudinal data within the framework of generalized estimating equations (GEE). The proposed approach integrates the robust method and joint mean-covariance regression modeling. Robust generalized estimating equations using bounded scores and leverage-based weights are employed for the mean and covariance to achieve robustness against outliers. The resulting estimators are shown to be consistent and asymptotically

normally distributed. Simulation studies are conducted to investigate the effectiveness of the proposed method. As expected, the robust method outperforms its non-robust version under contamination. Finally, we illustrate by analyzing a hormone data set. By down-weighing the potential outliers, the proposed method not only shifts the estimation in the mean model, but also shrinks the range of the innovation variance, leading to a more reliable estimation in the covariance matrix.

C1563: **Comparison of block bootstrap testing methods of mean difference for paired longitudinal data**
Presenter: **Hirohito Sakurai**, National Center for University Entrance Examinations, Japan
Co-authors: Masaaki Taguri

The aim is to compare 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.

C1611: **Nonparametric dynamic screening system for monitoring longitudinal data**
Presenter: **Jun Li**, University of California - Riverside, United States

In many applications, including disease early detection and prevention, and performance evaluation of airplanes and other durable products, we need to sequentially monitor the longitudinal pattern of certain performance variables of a subject. A signal should be given as soon as possible once the pattern becomes abnormal. Recently, a new statistical method called dynamic screening system (DySS) has been proposed to solve this problem. It is a combination of longitudinal data analysis and statistical process control. However, the current DySS method can only handle cases when observations are normally distributed and within-subject observations are independent or follow a specific time series model (e.g., AR(1) model). In this talk, we propose a new nonparametric DySS method which can handle cases when the observation distribution and the correlation among within-subject observations are arbitrary. Therefore, it broadens the application of the DySS method greatly. Numerical studies show that the new method works well in practice.

CS14 Room 13 COMPUTATIONAL STATISTICS IV Chair: Paolo Foschi

C1538: **Finite-sample multivariate tests for ARCH in vector autoregressive models**
Presenter: **Paul Catani**, Hanken School of Economics, Finland
Co-authors: Niklas Ahlgren

The aim is to 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. The tests are based on standardised multivariate residuals. 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. The tests are applied to VAR models estimated on credit default swap (CDS) prices.

C1482: **Different approaches to differential entropy estimation**
Presenter: **Emilija Nikolic-Djoric**, University of Novi Sad, Serbia and Montenegro
Co-authors: Zagorka Lozanov Crvenkovic

Entropy is a measure of uncertainty that is useful in many applications. It quantifies the expected value of information contained in discrete distribution (Shannon entropy) or in continuous distribution (differential entropy). In this paper three approaches to estimating differential entropy are considered. The first one is nonparametric, based on discretization of a random variable, then estimation of empirical probability density function by a histogram, and calculating Shannon entropy. The second one is discretization of a random variable and estimating its probabilities by means of several algorithms, which are included in Entropy package in R. The last one is plugging maximum likelihood estimates of the distribution parameters in exact differential entropy formula. For several well known distributions (normal, Student, uniform, lognormal and exponential), we generate a thousand samples for each of the sample sizes 20, 50, 100, 200, 500, and calculate the differential entropy estimates using three different approaches. Using these data, we investigate the distribution of differential entropy estimates. In order to compare the performances of three applied procedures for estimation, we compare the distribution of differential entropy estimates with respect to mean value, variance, kurtosis, skewness, bias and mean squared error.

C1502: **Quantifying and localizing state uncertainty in hidden Markov models using conditional entropy profiles**
Presenter: **Jean-Baptiste Durand**, Université Grenoble Alpes, France
Co-authors: Yann Guedon

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.

C1562: **Dynamic stress-strength modeling with cumulative stress and strength degradation**
Presenter: **Prajamitra Bhuyan**, Indian Statistical Institute, India
Co-authors: Anup Dewanji

The stress-strength model is widely used in mechanical engineering, medicine, psychology and various other allied fields for reliability calculation. Ample amount of research work on stress-strength modelling, under the assumption that stress and strength are static random variables, are available in the literature. However, in most real life scenarios, strength degrades over time and stress is accumulation of damage due to shocks arriving at random time points. Quite surprisingly, strength degradation and damage accumulation has not been taken concurrently into consideration in reliability analysis. Reliability function is defined under suitable assumptions where strength degrades and stress accumulates over time. Methods for numerical evaluation of reliability are suggested under deterministic strength degradation and cumulative damage due to shocks arriving according to a Poisson process using simulation method and inversion theorem. These methods are specifically useful in the scenarios where damage distributions do not possess closure property under convolution. Results from inversion method is compared with known approximate methods and extended for non-identical, dependent damage distributions as well as for random strength degradation. As it turns out, the simulation method

seems to have an edge in terms of computational burden and has much wider domain of applicability.

CS50 Room 19 CONTRIBUTIONS TO DEPENDENCE MODELS Chair: Piotr Jaworski

C1358: **Regularized principal component analysis for spatial data**
Presenter: **Hsin-Cheng Huang**, Academia Sinica, Taiwan
Co-authors: Wen-Ting Wang

Nonstationary spatial modeling is considered using empirical orthogonal functions (EOFs) based on data observed at p spatial locations with n repeated measurements. Traditionally, EOFs are obtained using principal-component-analysis related approaches. However, when data are noisy or n is small, the leading eigen-vectors produced from PCA may have noisy patterns with weak physical interpretation. To obtain more precise estimates of dominant patterns, a regularization approach is proposed incorporating smoothness and sparseness of the dominant patterns while accounting for orthogonality of eigenvectors, even when data are observed at irregularly spaced locations. The resulting optimization problem is solved using the alternating direction method of multipliers. Some numerical examples are provided to demonstrate the effectiveness of the proposed method.

C1499: **Modeling dependence under censoring and truncation using multivariate mixtures of Erlangs**
Presenter: **Roel Verbelen**, KU Leuven, Belgium
Co-authors: Katrien Antonio, Gerda Claeskens

The estimation and use of multivariate mixtures of Erlangs (MME) to model dependent multivariate censored and truncated data is studied. MME form a *highly flexible* class of distributions, as they are dense in the space of positive continuous multivariate distributions. Moreover, the class is *analytically tractable*, has many quantities of interest with a closed-form expression and enjoys interesting closure properties. The use of MME should be regarded as semiparametric density estimation technique to model the dependence directly and hence forms a suitable alternative to copulas. We present an *estimation technique* for fitting MME using the EM algorithm to data that can be *censored and / or truncated*, which is often the case in applications such as loss modeling (finance and actuarial science), clinical experiments (survival / failure time analysis), mastitis studies (veterinary studies), and duration data (econometric studies). We demonstrate the effectiveness of the proposed algorithm and the practical use of MME on simulated data as well as on real data sets.

C1671: **Estimation of discrete partially directed acyclic graphical models in multitype branching processes**
Presenter: **Pierre Fernique**, INRIA, France
Co-authors: Jean-Baptiste Durand, Yann Guedon

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 which is applied on multivariate count data. The proposed modeling approach is illustrated on plant architecture datasets.

C1281: **Life tests of series systems under copula based lifetime distributions**
Presenter: **Tsai-Hung Fan**, National Central University, Taiwan
Co-authors: Tsung-Ming Hsu

The dependence among the components lifetimes of a series system by copula functions under a life tests is modelled. Different copulas such as Clayton copula, and Gumbel copula will be considered. Some commonly used lifetime distributions are adopted to model the marginal distributions of the componentślifetime in a series system and fundamental statistical inference is explored when masked data appear. In dealing with the identifiability issue of the unknown copula parameters, the profile likelihood by fixed parameter values in the copula function is considered to estimate the parameters of the underlying lifetime distributions. The copula parameter is selected based on some optimization criteria. The results will be demonstrated by numerical simulation.

Thursday 21.08.2014 08:45 - 10:00 Parallel Session G

OS13 Room 6 MULTIPLE NONPARAMETRIC REGRESSION Chair: Joachim Schnurbus

C1289: **Nonparametric kernel estimation: an exact parallel kd-tree approach**
Presenter: **Jeffrey Racine**, McMaster, Canada
Co-authors: Hayfield Tristen
Often practitioners require nonparametric estimates of density functions, regression functions, conditional PDF and CDF and quantile functions and so forth. Kernel methods have emerged as a leading approach for such problems. Real-world data often involves a mix of continuous, discrete, and categorical data, and there has been much recent activity in the area of nonparametric kernel estimation with this mix of datatypes. Unfortunately, the nature of these methods presents a computational barrier to their widespread adoption as the computational burden associated with these methods has blocked their application to all but small to moderately sized datasets. Rather than resorting to approximate approaches as is often done in computational circles, we instead pursue an exact approach based on kd-trees and exploitation of the obvious parallelism inherent to these methods reducing them to computational order $O(nlog(n))$ from $O(n^2)$, by way of illustration. These methods have been incorporated into the R np and npRmpi packages which focus on kernel methods appropriate for the mix of continuous, discrete, and categorical data often found in applied settings.

C1459: **Estimation of non-simplified vines using (nonparametric) trivariate copula constructions**
Presenter: **Christian Schellhase**, Bielefeld University, Germany
Co-authors: Fabian Spanhel
Recently, vine copulas (or pair-copula constructions) have become an important tool in high-dimensional dependence modeling. A commonly used assumption is that each bivariate conditional copula in the vine collapses to a bivariate unconditional copula, which is called the simplifying assumption. In this paper we consider ways and means to weaken the simplifying assumption. We show how bivariate conditional copula densities with one conditioning argument can be used to approximate bivariate conditional copula densities with an arbitrary number of conditioning arguments. We call this a trivariate copula construction since bivariate conditional copula densities with one conditioning argument can be recognized as trivariate copula densities with particular restrictions. Using trivariate copulas we obtain a vine copula that is still feasible in practice and gives a better approximation to the multivariate density. We also present the non-parametric estimation of conditional copulas with a penalized hierarchical B-splines approach. The great advantage of this approach is that we directly estimate a conditional copula with uniform margins, setting simple linear restrictions on the spline coefficients using quadratic programming. Thus, there is no need to extract the conditional copula from an unrestricted estimation as it would be the case for other nonparametric approaches.

C1430: **Smoothing matrix analysis for mixed kernel regression**
Presenter: **Joachim Schnurbus**, University of Passau, Germany
Co-authors: Harry Haupt
The mixed kernel approach provides a flexible regression framework for discrete and continuous covariates. As for any smoothing method a priori decisions have to be made on the nonparametric configuration: Choosing kernel type, mode of local regression, and bandwidth estimation approach. Crucial for the latter choices is the behavior of the smoothing matrix. The aim of this paper is twofold. First, carve out the non-standard properties of the smoothing matrix of mixed kernel regressions. Second, demonstrate the use of smoothing matrix analysis to select an appropriate nonparametric configuration. The exposition combines graphical and analytical tools to evaluate the impact of bandwidth values and single observations on the smoothing matrix. The proposed tools are illustrated using simulated data and a well-known data set on house prices.

OS22 Room 13 STATISTICS IN MEDICAL IMAGING: GEOMETRY, INFERENCE AND COMPUTATIONS Chair: Vic Patrangenaru

C1527: **Stochastic deconvolution for drift-estimation in photoactivated nanomicroscopy imaging of microtubulin networks**
Presenter: **Stephan Huckemann**, University of Goettingen, Germany
Co-authors: Alexander Hartmann, Jörn Dannemann, Oskar Laitenberger, Claudia Geisler, Alexander Egner, Axel Munk
In nano-biolgogy, stochastic marker switching nanoscale fluorescence microscopy is a powerful technique to allow for non-destructive imaging well below the diffraction barrier of visible light. The price to be paid is that images have to be reconstructed from sparse image sequences over considerable time intervals such that thermomechanical and cell locomotion effects may cause drifts. For drift and image estimation we propose a novel stochastic spatio-temporal white noise deconvolution model. We derive semiparametric estimators as well as their asymptotic behavior under suitable cutoff rates: Strong consistency and a central limit theorem with respect to the true drift and true image, respectively. The practicability of our method is demonstrated in a simulation study and in an application to PALMIRA (photoactivated localization microscopy with independently running acquisition) fluorescence microscopy.

C1561: **Clustering Approaches to Improved Activation Detection in fMRI**
Presenter: **Ranjan Maitra**, Iowa State University, United States
Co-authors: Wei-Chen Chen
The past two decades have seen the development of functional Magnetic Resonance Imaging (fMRI) as a tool to noninvasively study the spatial characteristics and extent of human brain function. Preliminary statistical analysis relates the time-course MR sequences at a voxel to a stimulus and assessing significance of this relationship based on the p-values. Incorporating spatial context in the significance assessment is computationally challenging. We propose modeling the p-values as a mixture of beta distributions, while incorporating spatial context via the voxel coordinates. Parallel and accelerated Expectation-Maximization (EM) methods are used to perform computations in a practical setting and the Bayes Information Criterion (BIC) is used to determine the number of components in the best-fitting solution. The resulting segmentation is analyzed to specify common regions of activation, with promising results on simulation datasets and also on data from an fMRI study.

C1608: **How far is the corpus callosum of an average individual from the corpus callosum ofAlbert Einstein?**
Presenter: **Vic Patrangenaru**, Florida State University, United States
Co-authors: Mingfei Qiu, Leif Ellingson
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.

OS23 Room 19 STATISTICS IN LIFE SCIENCES Chair: Athanassios Kondylis

C1369: **Dynamics of DNA minicircles in motion via Fourier analysis of functional time series**
Presenter: **Shahin Tavakoli**, EPFL, Switzerland
Co-authors: Victor Panaretos
The problem of studying the dynamics of DNA minicircles that are vibrating in solution is considered. At a large scale, DNA minicircles are modelled as elastic rods, and the problem of understanding their dynamics can be recasted into the problem of estimating the second order structure of a stationary functional time series (FTS). This problem is tackled by a frequency domain approach, where the spectral density operators (or spectra) of the DNA minicircle are estimated. Then hypothesis tests are carried out to compare the spectra of two specific DNA minicircles, and localise their differences both in frequencies and on the DNA minicircles.

C1424: **Analysis of incomplete functional data**
Presenter: **David Kraus**, University of Bern, Switzerland
Techniques of functional data analysis have been developed for analysing collections of functions, for example curves, surfaces or images. It is customary to assume that all functions are completely (or densely) observed on the same domain. In this work we extend the scope of application of functional data analysis to situations where each functional variable may be observed only on a subset of the domain while no information about the function is available on the complement. For this partial observation regime, we develop main tools of functional data analysis, such as estimators of the mean function and covariance operator and principal components analysis, and show how individual incomplete functions can be recovered from observed fragments. Our work is motivated by a data set from ambulatory blood pressure monitoring where only parts of temporal heart rate profiles are observed. This work was done at the University Hospital Lausanne.

C1717: **Predictive component-based multi-block path modeling**
Presenter: **Vincenzo Esposito Vinzi**, ESSEC Business School of Paris, France
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.

OS26 Room 5 LARGE SCALE PORTFOLIO ALLOCATION IN A NON-ELLIPTICAL UNIVERSE Chair: Simon Broda

C1475: **A fast and accurate method for value at risk and expected shortfall calculation using the noncentral t distribution**
Presenter: **Jochen Krause**, University of Zurich, Switzerland
Co-authors: Marc Paolella
Value at risk (VaR) and, more recently, expected shortfall (ES), are fundamental risk measures. In May 2012, the Basle Comittee on Banking Supervision announced its intention to replace VaR with ES in banks' internal models for determining regulatory capital requirements. Observing that the evaluation of ES first requires calculating the VaR, accurate methods for both VaR and ES are necessary. We develop an extremely fast method for VaR and ES prediction based on NCT-GARCH, and use two variants of the normal-mixture-GARCH model class as a benchmark for testing. Unfortunately, the method we employ to speed up the NCT-GARCH estimation is not straightforwardly applicable to the normal-mixture-GARCH framework, and so the latter is, relative to the proposed technique, extremely slow. Besides being fast and accurate, the new method can also be used in a portfolio optimization framework, in which the ES values corresponding to thousands of candidate portfolios need to be computed.

C1467: **Asset returns density forecasting with MCD algorithms**
Presenter: **Marco Gambacciani**, University of Zurich, Switzerland
Co-authors: Marc Paolella
A new approach for multivariate modeling of asset returns is proposed which accounts for all the major stylized facts, and also lends itself to portfolio optimization. It is based on a two-component multivariate normal mixture model, estimated using a new variation of the minimum covariance determinant (MCD) method. An empirical application demonstrates the viability of the proposed method, in terms of estimation speed and out of sample density prediction.

C1442: **Saddlepoint approximation of expected shortfall for transformed means**
Presenter: **Simon Broda**, University of Amsterdam, Netherlands
Co-authors: Marc Paolella, Jochen Krause
The Basle committee's proposed move from value at risk to expected shortfall as the mandated risk measure in its market risk framework necessitates practical methods for evaluating said measure. Defined as a partial expectation of the return distribution standardized by the tail probability, expected shortfall cannot be obtained in explicit form for many distributions of interest. The present paper derives a saddlepoint approximation for the expected shortfall associated with certain random variables that permit a stochastic representation in terms of some underlying random variables possessing a moment generating function. The new approximation can be evaluated quickly and reliably, and provides excellent accuracy. Examples in which the method is applicable include the (singly and doubly) noncentral Student's t, and the Azzalini skewed t (Azzalini, 2003). We illustrate the accuracy of the proposed methods using the former. We also establish a link between our proposed approximation and mean-expected shortfall portfolio optimization.

OS28 Room 4 SYMBOLIC/ALGEBRAIC METHODS IN COMPUTATIONAL STATISTICS I Chair: Elvira di Nardo

C1380: **Tree cumulants for latent two-state tree models**
Presenter: **Piotr Zwiernik**, University of California Berkeley, United States
Consider a graphical model on a tree such that only leaves of the tree are observed. The parametrization of the model requires integrating over the set of possible outcomes of the hidden variables. In the discrete case this results in a complicated map, which

makes it hard to understand the model structure. The focus is on the case when all variables in the given system are binary. It is shown that there exists an elegant change of coordinates which resembles cumulants, in which the model has a surprisingly simple description given by products of edge correlations and skewnesses for all the nodes in a tree.

C1392: **Performance analysis of an algorithm from computational algebra for implicit regression**
Presenter: **Eva Riccomagno**, University degli Studi di Genova, Italy
Co-authors: Claudia Fassino, Laura Torrente

Recently new class of algorithms has been developed that construct polynomials that almost vanish at a finite number of d-dimensional points $\mathcal{D}$. Such polynomials can be interpreted in terms of implicit regression models of the form $f(x_1,\ldots,x_d)=0$, where in the regression equation there is no distinction between dependent and independent factors. One such algorithm based on two steps is considered. In Step 1 it returns a polynomial f which almost vanishes at $\mathcal{D}$ and is based on a version of the Buchberger-Möller algorithm from Computational Commutative Algebra. In Step 2 by Newton's method it modifies the coefficients of f to obtain $\bar{f}$ so that each point in $\mathcal{D}$ is near to a zero of $\bar{f}$ within a fixed tolerance. An estimate of the goodness of this approximation of $\mathcal{D}$ by the zeros of $\bar{f}$ is based on Kantorovich theorem on the convergence of Newton's method. An auxiliary set of points $\bar{\mathcal{D}}$ needs to be computed, of which we shall give an interpretation. An evaluation and performance study of this algorithm is presented on a number of datasets with different characteristics and performance indices are discusses/introduced.

C1410: **Confidence nets based on finite reflection groups**
Presenter: **Henry Wynn**, London School of Economics, United Kingdom

For a one sample mean/median, θ, with errors satisfying a special symmetry it is possible to build a confidence net, namely a set of random intervals with disjoint interiors covering the whole real line, and such that the probability of any interval covering θ has known probability. The symmetry is defined in terms of finite reflection groups and the fundamental cone of the groups plays a special role. The technique is to find the probability function of the coverage probabilities. A fundamental formula is derived for this which requires the study of the so-called length function of the Cayley graph of the relevant group. Specifically, the statistical problem imposes special restrictions on the Cayley graph which can be expressed in terms of the generating functions for certain length function distributions. Exact formulae are given for the reflection groups B_n, D_n, E_6, E_7 and E_8.

OS33 Room 3 NONPARAMETRIC METHODS FOR HIGH DIMENSIONS Chair: Maria Lucia Parrella

C1349: **Large precision matrix estimation via pairwise tilting**
Presenter: **Na Huang**, London School of Economics, United Kingdom
Co-authors: Piotr Fryzlewicz

A $tilting$-based method is proposed to estimate the precision matrix of a p-dimensional random variable, $\mathbf{X}$, when p is possibly much larger than the sample size n. Each 2×2 block indexed by (i,j) of the precision matrix can be estimated by the inversion of the pairwise sample conditional covariance matrix of X_i and X_j controlling for all the other variables. However, in the high dimensional setting, including too many or irrelevant controlling variables may distort the results. To determine the controlling subsets in high dimensional scenarios, the proposed method applies the $tilting$ technique to measure the contribution of each remaining variable to the variance-covariance matrix of X_i and X_j, and only puts the (hopefully) highly relevant remaining variables into the controlling subsets. Conditions under which it can successfully distinguish the highly relevant remaining variables from the rest are illustrated. The simulation results will be presented under different scenarios for the underlying precision matrix. Comparison with other competing methods will also be given.

C1670: **Local likelihood estimation for multivariate directional data**
Presenter: **Marco Di Marzio**, University of Chieti-Pescara, Italy
Co-authors: Stefania Fensore, Agnese Panzera, Charles C. Taylor

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.

C1660: **GRID for variable selection in high dimensional regression**
Presenter: **Francesco Giordano**, University of Salerno, Italy
Co-authors: Soumendra Nath Lahiri, Maria Lucia Parrella

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.

Thursday 21.08.2014 11:30 - 13:00 Parallel Session H

IS02 Room 3 COMPUTER INTENSIVE METHODS IN STATISTICS OF EXTREMES Chair: Ivette Gomes

C1324: **Robust and bias-corrected estimation of the coefficient of tail dependence**
Presenter: **Armelle Guillou**, Strasbourg, France
Co-authors: Christophe Dutang, Yuri Goegebeur

A robust and asymptotically unbiased estimator for the coefficient of tail dependence in multivariate extreme value statistics is introduced. The estimator is obtained by fitting a second order model to the data by means of the minimum density power divergence criterion. The asymptotic properties of the estimator are investigated. The efficiency of the methodology is illustrated on a small simulation study and by a real dataset from the actuarial context.

C1668: **On the bootstrap methodology for the estimation of the tail sample fraction**
Presenter: **Frederico Caeiro**, Universidade Nova de Lisboa, Portugal
Co-authors: M. Ivette Gomes

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. 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.

C1557: **Efficiency of partially reduced-bias mean-of-order-p versus minimum-variance reduced-bias extreme value index estimation**
Presenter: **Ivette Gomes**, University of Lisbon, Portugal
Co-authors: Frederico Caeiro

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.

OS11 Room 6 ADVANCES AND NEW METHODOLOGIES IN SURVIVAL AND RELIABILITY Chair: Juan Eloy Ruiz-Castro

C1374: **Robust spherical separation**
Presenter: **Immanuel Bomze**, Universität Wien, Austria
Co-authors: Annabella Astorino, Antonio Fuduli, Manlio Gaudioso

A robust spherical separation technique aimed at separating two finite sets of points is proposed. Robustness concerns the possibility to admit uncertainties and perturbations in the data set, which may occur when the data are corrupted by noise or are influenced by measurement errors. In particular, starting from the standard spherical separation under the assumption of spherical uncertainty, a model is proposed characterized by a non-convex non-differentiable objective function, which is minimized by means of a bundle type algorithm. Quite promising numerical results are provided on small and large data sets drawn from well-established test beds in literature.

C1472: **The Dagum distribution from a survival analysis point of view**
Presenter: **Mariangela Zenga**, Milano-Bicocca University, Italy
Co-authors: Filippo Domma

The Dagum distribution, introduced by Dagum in 1977, is a Burr III distribution with an additional scale parameter, it is closely related to the Burr XII distribution and, more generally, turns out to be a special case of the Generalized Beta distribution. Even if the Dagum model has been used in studies of income and wage distribution as well as wealth distribution, only recently it was introduced in the field of the survival analysis and the reliability. The hazard rate of this model is very flexible; in fact, it is proved that, according to the values of the parameters, the hazard rate of the Dagum distribution has a decreasing, or a Upside-down Bathtub, or Bathtub and then Upside- down Bathtub failure rate. Moreover some features of this distribution (as the reversed hazard rate, the mean and variance of the random variables residual life and reversed residual life and their monotonicity properties) were studied. In this work we will consider the observed heterogeneity on the hazard function of the Dagum distribution.

C1435: **The jackknife estimate of variance for transition probabilities in the non-markov illness-death model**
Presenter: **Leyla Azarang**, University of Vigo, Spain
Co-authors: Jacobo de Una-Alvarez

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.

C1371: **Preventive maintenance in a complex warm standby system. A transient analysis**
Presenter: **Juan Eloy Ruiz-Castro**, University Of Granada, Spain

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 the 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.

OS27 Room 4 COMPUTATIONAL CHALLENGES IN ENVIRONMENTAL STATISTICS Chair: Marc Genton

C1444: **Visualization of environmental data**
Presenter: **Marc Genton**, KAUST, Saudi Arabia

Many datasets from environmental sciences consist of functional observations, such as temporal curves, spatial surfaces or images. We describe functional boxplots to visualize such data and study various rankings of functional data based on band depth or tilting. We investigate the performance of these approaches via simulations, discuss computational issues, and report the outcome of functional data visualization from environmental problems. We also provide a natural extension of our method to multivariate functional data.

C1347: **Emulation of global 3D spatio-temporal temperature: a distributed computing approach to model one billion data points**
Presenter: **Stefano Castruccio**, KAUST, Saudi Arabia
Co-authors: Marc Genton

Statistics for emulation of climate model output has attracted increasing interest among geophysicists and policy makers. The use of statistical models for data generation rather than more time-consuming systems of partial differential equations is a promising avenue for providing simple, fast, and easy-to-use tools for decision making. The main challenge of fitting a statistical model to climate model output in space and time is the amount of data: covariance matrices describing interactions between all data points easily become too large to store in RAM memories, thus making data reduction or model simplification necessary. The aim is to show how the gridded geometry of the data and the use of computational facilities with a large memory and a large number of processors allow for emulation of global three-dimensional temperature fields over time without data reduction. By using designed and dedicated computational facilities and a spectral statistical model specifically defined to exploit parallelization, it becomes possible to fit models with data of unprecedented size. Here, it is shown how it is possible to fit a non-trivial model to a data set of one billion data points with a covariance matrix comprising of 10^{18} entries.

C1367: **Statistically and computationally efficient estimating equations for large spatial datasets**
Presenter: **Ying Sun**, KAUST, Saudi Arabia
Co-authors: Michael Stein

For Gaussian process models, likelihood based methods are often difficult to use with large irregularly spaced spatial datasets, because exact calculations of the likelihood for n observations require $O(n^3)$ operations and $O(n^2)$ memory. Various approximation methods have been developed to address the computational difficulties. New unbiased estimating equations are proposed based on score equation approximations that are both computationally and statistically efficient. The inverse covariance matrix that appears in the score equations is replaced by a sparse matrix to approximate the quadratic forms, then set the resulting quadratic forms equal to their expected values to obtain unbiased estimating equations. The sparse matrix is constructed by a sparse inverse Cholesky approach to approximate the inverse covariance matrix. The statistical efficiency of the resulting unbiased estimating equations are evaluated both in theory and by numerical studies. These methods are applied to nearly 90,000 satellite-based measurements of water vapor levels over a region in the Southeast Pacific Ocean.

C1415: **High-order composite likelihood for spatial extremes**
Presenter: **Raphael Huser**, King Abdullah University of Science and Technology, Saudi Arabia

Max-stable processes are natural models for spatial extremes, because they provide suitable asymptotic approximations to the distribution of maxima of random fields. However, classical approaches to inference are extremely computationally demanding, if not impossible in large dimensions, since the number of terms involved in the likelihood function is as large as the number of partitions of the set 1,. . . ,D, where D is the number of monitoring sites. As a result, in order to reduce the computational burden, the traditional approach to inference has been to use pairwise likelihoods as surrogates for the full likelihood. In my talk, I will discuss inference based on higher-order composite likelihoods, show the efficiency improvements, and discuss some of the computational challenges that are faced in practice.

OS32 Room 5 FINANCIAL TIME SERIES Chair: Cathy W.S. Chen

C1457: **Realized GARCH models incorporating intra-day and realized range data**
Presenter: **Richard Gerlach**, University of Sydney, Australia
Co-authors: Chao Wang

The aim is to build on the recent realized GARCH modelling framework by considering data sources including the intra-day range and the realized range, in addition to the standard choice of realized volatility. Including an observation equation for realized volatility, directly linking it contemporaneously to unobserved volatility, has been shown to increase predictive power over standard GARCH models. Our work shows that further increases in predictive likelihhood can be achieved, in several real return series over long forecast periods, by considering either the intra-day range or the realized range in a similar realized GARCH-type framework. In particular, the realized range observation equation with volatility gives rise to much lower measurement error variance, when compared to that for realized volatility, subsequently leading to marked increases in predictive power.

C1414: **A multivariate mixture stochastic covariance model**
Presenter: **Mike So**, The Hong Kong University of Science and Technology, China

A multivariate mixture stochastic covariance model based on Wishart distribution is proposed. A main feature of the mixture model is to provide a flexible dynamic for multiple returns by specifying a mixture of normal conditional distribution given the stochastic covariance. It is expected that this mixture model can exhibit complex dynamic found in many financial time series such as volatility clustering, large kurtosis and extreme observations. We model the conditional covariance of each component as the expectation of the stochastic covariance. Based on the assumption that the stochastic covariance is observable, we separate the modeling of the conditional covariance into modeling the dynamics of the correlation and variance. A GARCH-type process is chosen to model the two series making use of the realized correlation and realized variance in high-frequency financial data. Estimation of parameters is performed using Markov Chain Monte Carlo methods. Real financial returns are used to illustrate how to use the mixture model to forecast conditional covariance.

C1448: **Parameter change test for zero-inflated generalized Poisson autoregressive models**
Presenter: **Sangyeol Lee**, Seoul National University, Korea, South
Co-authors: Youngmi Lee, Cathy Chen

The aim is testing for a parameter change in zero-inflated generalized Poisson autoregressive (ZIGP) models. We first verify that the ZIGP process is stationary and ergodic and suggest cumulative sum (CUSUM) tests based on estimates and residuals. We then demonstrate that the conditional maximum likelihood estimator (CMLE) is strongly consistent and asymptotically normal and construct

the CMLE-based CUSUM test. It is shown that under regularity conditions, its limiting null distribution is a function of independent Brownian bridges. A simulation study and real data analysis are conducted for llustration.

C1454: **Bayesian assessment of dynamic quantile forecasts**
Presenter: **Cathy WS Chen**, Feng Chia University, Taiwan
Co-authors: Richard Gerlach, Edward Lin

Methods for Bayesian testing and assessment of dynamic quantile forecasts are proposed. Specifically, Bayes factor analogues of popular frequentist tests for independence of violations from, and for correct coverage of a time series of, quantile forecasts are developed. To evaluate the relevant marginal likelihoods involved, analytic integration methods are utilized when possible, otherwise multivariate adaptive quadrature methods are employed to estimate the required quantities. The usual Bayesian interval estimate for a proportion is also examined in this context. The size and power properties of the proposed methods are examined via a simulation study, illustrating favourable comparisons both overall and with their frequentist counterparts. An empirical study employs the proposed methods, in comparison with standard tests, to assess the adequacy of a range of forecasting models for Value at Risk (VaR) in several financial market data series.

CS45 Room 19 CONTRIBUTIONS TO COPULA-BASED MODELING II Chair: Ostap Okhrin

C1383: **Bounded-influence robust estimation of copulas**
Presenter: **Samuel Orso**, University of Geneva, Switzerland
Co-authors: Stéphane Guerrier, Maria-Pia Victoria-Feser

Copula functions are very convenient for modelling multivariate observations. Popular estimation methods are the (two-stage) maximum likelihood and an alternative semi-parametric with empirical cumulative distribution functions of the margins. Unfortunately, they can often be biased whenever relatively small model deviations occur at the marginal and/or copula levels. Two robust estimators that do not share this undesirable feature are proposed. Since skewed and heavy tailed parametric marginals are considered in many applications, also a bounded-bias robust estimator is proposed, for such distributions, that is corrected for consistency by means of indirect inference. In a simulation study it is shown that these robust estimators outperform the conventional approaches.

C1649: **Multivariate L-moment homogeneity test based on copula modeling**
Presenter: **Jan Picek**, Technical University of Liberec, Czech Republic
Co-authors: Jan Kysely, Tereza Simkova

The contribution is devoted to a study of the recent extension of univariate regional homogeneity tests based on L-moments to the multivariate case. We propose improvements to existing test and discuss the further development of the methodology, especially models of extreme events with time-dependent variables. The test statistics are based on the multivariate L-moments, L-comoments and their possible generalizations. Copula models (f.e. Clayton, Gumbel-Hougaard) are used to describe the statistical behavior of dependent variables. The methodology is illustrated on multivariate regional frequency analysis of extreme precipitation events in the Czech Republic.

C1662: **Unsupervised learning using Gaussian mixture copula model**
Presenter: **Sakyajit Bhattacharya**, Xerox Research Centre INDIA, India
Co-authors: Vaibhav Rajan

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. We present a modified expectation-maximization algorithm for estimating the number of components and the parameters of a GMCM, long 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.

C1669: **Efficiency of bivariate copula on the Shewhart control chart**
Presenter: **Sasigarn Kuvattana**, King Mongkut University of Technology North Bangkok, Thailand
Co-authors: Saowanit Sukparungsee, Piyapatr Busababodhin, Yupaporn Areepong

The objective of this article is to propose four types of copulas on the Shewhart control chart when observations are generated by exponential distribution with the means shifts. The Monte Carlo simulations are used and the performance of control chart which based on the Average Run Length (ARL) is compared for each copulas. Four types of copula function for specifying dependence between random variables are used and measured by Kendall's tau. The results show that ARL is depended on Kendall's tau. Finally, the efficiency of copula is also depended on level of dependence.

CS47 Room 13 CONTRIBUTIONS TO LONGITUDINAL DATA ANALYSIS II Chair: Juan Romo

C1653: **Tree-based algorithm for learning moderation in linear regression models**
Presenter: **Reto Buergin**, University of Geneva, Switzerland
Co-authors: Gilbert Ritschard

Moderated relationships, i.e. relationships between predictors and the response that depend on other covariates, are an important association structure in regression analysis. The effect of a clinical trial, for example, could depend on side diagnoses, and wage gaps between men and women could vary between labor sectors or depend on labor market policies. In this talk, a tree-based algorithm for learning such moderation, implemented in the R package vcrpart, is presented. This algorithm consist in incorporating a piecewise constant coefficient function of moderators in the linear predictor equation. The algorithm is scaleable for many moderators of possibly mixed scales, integrates interaction between moderators and can handle nonlinear moderation. The potential of the algorithm is illustrated with an application, using longitudinal data, that examines how the effect of an individual transition from employment to unemployment on self-reported happiness varies across individual characteristics and life circumstances.

C1666: **A comparison of some estimation methods for latent Markov models with covariates**
Presenter: **Silvia Pandolfi**, University of Perugia, Italy
Co-authors: Francesco Bartolucci, Giorgio E. Montanari

The aim is to 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 of these models. 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.

C1399: **A simulation study to assess statistical approaches for longitudinal count data**
Presenter: **Maria Salome Cabral**, CEAUL and FCUL, Portugal
Co-authors: Maria Helena Goncalves

The aim is to study 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.

C1330: **Incomplete longitudinal binary responses in marginal model**
Presenter: **Maria Helena Goncalves**, CEAUL and FCT of UAlg, Portugal
Co-authors: Maria Salome Cabral

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 presents in successive observations of response variable that are taken over time on each individual. In this paper we discuss the performance of the marginal models implemented in the R package `bild` 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 two illustrate that behaviour.

CS52 Room 20 CONTRIBUTIONS TO COMPUTATIONAL METHODS IN FINANCE II Chair: Robert Kunst

C1357: **Method of moments estimation and affine term structure models**
Presenter: **Leopold Soegner**, Institute for Advanced Studies, Austria
Co-authors: Jaroslava Jaroslava

Parameter estimation of affine term structure models is investigated by means of the generalized method of moments. Results obtained for m-polynomial processes in mathematical finance literature are used to derive the moments of the affine latent process driving the term structure, and after specifying the properties of the micro-structure noise, the moments of the yields observed. Equipped with these moments parameter estimations by means of the generalized method of moments (GMM) can be performed. To implement GMM estimation the number of moment restrictions has to be chosen carefully to obtain reliable parameter estimates. In addition, the minimization of the GMM distance function turned out to be complicated and computationally demanding. Markov chain Monte Carlo methods are used to minimize this distance function such that reliable estimates can be obtained. In addition, it is tested for different market price of risk specifications. After a simulation study, the estimation procedure is applied to empirical interest rate data.

C1559: **The SIML estimation of Integrated covariances and hedging coefficient under micro-market noise and random sampling**
Presenter: **Naoto Kunitomo**, University of Tokyo, Japan
Co-authors: Hiroumi Misaki

For estimating the integrated volatility and covariances by using high frequency data, Kunitomo and Sato (2008, 2011) have proposed the Separating Information Maximum Likelihood (SIML) method when there are micro-market noises. The SIML estimator has reasonable finite sample properties and asymptotic properties when the sample size is large under general conditions with non-Gaussian processes or volatility models. We shall show that the SIML estimation is useful for estimating the integrated covariances and hedging coefficient when we have micro-market noises and the financial high frequency data are randomly sampled. Thus it is useful for practical purposes of analysing financial high frequency (multivariate) data and risk managements.

C1630: **Estimation of Levy CARMA models in the yuima package: application on the financial time series**
Presenter: **Lorenzo Mercuri**, University of Milan, Italy
Co-authors: Stefano Maria Iacus

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 Levy model. The estimation of the parameters for the underlying Levy process makes yuima package appealing for modeling financial time series. Indeed, identifying the appropriate noise for a CARMA model allows us to capture asymmetry and heavy tails observed in the real data.

C1576: **Keeping a finger on the pulse of the economy: nowcasting Swiss GDP in real-time squared**
Presenter: **Boriss Siliverstovs**, KOF ETHZ, Switzerland

The aim is tp evaluate forecasting performance of a large-scale factor model involving 557 economic indicators in a genuine ex ante forecasting exercise. We perform our forecasts of GDP growth in Switzerland in real time using real-time data vintages collected at weekly frequency. This allows us to monitor how newly released economic and financial data influence our forecasts and hence capture prevailing tendencies in the current course of economic development

Thursday 21.08.2014 14:00 - 15:45 Parallel Session I

TS2 Room 3 TUTORIAL 2 **Chair: Gil Gonzalez-Rodriguez**

C1718: **Robust estimation, inference and prediction**
Presenter: **Stefan Van Aelst**, KU Leuven, Belgium

Statistical models are at best a good approximation of the often complex process that generated the available data. Moreover, individual observations may deviate from the process that produces the majority of the data due to rare events, technological failures and so on. To address these issues when analyzing the data robust methods need to be used that still yield reliable results for the majority of the observations while some fraction of the data deviates from the statistical model. The robust methods then allow us to detect the deviating observations for further inspection or may guide adjustments to the model to obtain a better approximation of reality. However, not all outlying observations are meaningful and in that case the model should not be adjusted to accommodate the outliers, but the inference should still be reliable in the presence of these outliers. We will start with an overview of robust methods to estimate the model parameters in regression and multivariate location-scale models. Both methods for low-dimensional and high-dimensional data will be discussed and the challenges encountered for both types of data are highlighted. Moreover, we then focus on methods for robust model selection, hypothesis testing and robust prediction models. Robust and computationally efficient resampling techniques will be discussed in this context.

OS38 Room 20 THE YSG-IASC SESSION: ROBUST METHODS IN REGRESSION ANALYSIS **Chair: Han-Ming Wu**

C1560: **Asymptotic properties of factorial k-means clustering**
Presenter: **Yoshikazu Terada**, National Institute of Information and Communications Technology, Japan

Factorial k-means (FKM) clustering is a method for clustering objects in a low-dimensional subspace. The advantage of this method is that the partition of objects and the low-dimensional subspace reflecting the cluster structure are obtained, simultaneously. In some cases that reduced k-means (RKM) clustering does not work well, FKM clustering can discover the cluster structure underlying a lower dimensional subspace. Conditions that ensure the almost sure convergence of the estimator of FKM clustering as the sample size increases unboundedly are derived. The result is proved for a more general model including FKM clustering. Moreover, a rough large deviation inequality for FKM clustering is also provided. Moreover, it is shown a surprising result that there exist some cases in which RKM clustering becomes equivalent to FKM clustering as the sample size goes to infinity.

C1396: **Robust regression in high-dimensions: sparse S and sparse MM**
Presenter: **Viktoria Oellerer**, KU Leuven, Belgium
Co-authors: Christophe Croux

Outliers in the data are a common problem in applied statistics. Estimators that give reliable results under contamination are called robust. Two commonly used robust regression estimators are the S- and MM-estimator. The former has a high breakdown point, thus it can deal with a high amount of contamination, while the latter combines a high breakdown point with a high efficiency. However, these estimators cannot be applied to high-dimensional data, data with more variables than observations. Adding an L_1-penalty to their objective function yields the *sparse S-* and *sparse MM-estimator*, two estimators suitable for such high-dimensional analysis that are both sparse and robust. The aim is to present their robustness properties, such as breakdown point and influence function. Furthermore, their performance will be compared in numerical experiments with the ordinary S- and MM-estimators, as well as with other sparse estimators. Additionally, a real data example is presented.

C1507: **An algorithm for robust groupwise least angle regression**
Presenter: **Andreas Alfons**, Erasmus University Rotterdam, Netherlands
Co-authors: Christophe Croux, Sarah Gelper

In many regression problems there exists a natural grouping among the predictor variables. Examples are groups of dummy variables that represent categorical variables, or linear and quadratic terms of numerical variables. Typically, the aim of model selection in such cases is to select groups of variables rather than individual covariates. However, outliers that do not follow the model of the majority of the data are a common problem in applied data analysis. We propose an extension of the popular least angle regression (LARS) procedure to groupwise variable selection together with strategies to reduce the influence of outlying data points. The excellent performance of robust groupwise LARS is demonstrated in numerical experiments.

C1523: **Cumulative slicing estimation for multiple nonlinear manifolds learning**
Presenter: **Han-Ming Wu**, Tamkang University, Taiwan

The isometric sliced inverse regression (ISOSIR) is a nonlinear extension of SIR. It has been shown that ISOSIR can recover the embedded geometric structure of the nonlinear manifolds data sets such as the Swiss roll. ISOSIR used K-means as a base clustering method to the pre-calculated isometric distance matrix of the data set so that the classical SIR algorithm can be applied. However, based on the results of K-means, ISOSIR ignored the ordering information of response both within and between the slices where the ordering information was one of the most important characteristics of a nonlinear manifold data set. In this study, we are motivated to settle this problem by using the cumulative slicing estimation. First, the proposed method computes the isometric distance between data points; the resulting distance matrix is then sorted by the rank-two ellipse seriation method, and the classical cumulative slicing estimation algorithm is applied. We conducted a pilot study and shown that the proposed method can reveal the geometric structure of a nonlinear manifold data set and the results were comparable to ISOSIR. We further applied it to the multiple nonlinear manifolds. We also investigated the applications of the found features for the regression problems to the real world data and microarray gene expression data.

C1724: **Robust Bayesian variable selection to mixture regression models**
Presenter: **Kuo-Jung Lee**, National Cheng-Kung University, Taiwan
Co-authors: Yi-Chi Chen

A Bayesian variable selection approach is applied to a finite mixture regression model with t-errors to address the robustness of the cross-country growth determinants. The proposed Bayesian method can simultaneously accommodate model uncertainty, population heterogeneity, and outliers. In particular, we adopt an alternative prior specification to the widely-used g-prior to circumvent the undesirable difficulties arise when the covariates are highly correlated, resulting in the singularity. A Monte Carlo simulation study is conducted to examine the ability of the proposed method to correctly identify the important variables present under a number of scenarios for linear relation among the covariates. For the empirical application of economic growth, we find that a number of important growth determinants identified. In addition, we notice a group of African countries that is considered as outliers despite the regional dummies are included. These results are compared with the previous cross-country growth studies, and the limitations of

this study, as well as future mixture modeling and variable selection research possibilities, are discussed.

CS18 Room 19 BAYESIAN METHODS **Chair: Leonhard Held**

C1365: **A hierarchical Bayesian approach to negative binomial regression**
Presenter: **Shuai Fu**, SUPSI, Switzerland
Co-authors: Giorgio Corani

There is a growing interest in establishing the relationship between the count data $\mathbf{y}$ and numerous covariates $\mathbf{x}$ through a generalized linear model (GLM), such as explaining the road crash counts from the geometry and environmental factors. The aim is to propose a hierarchical Bayesian method to deal with the negative binomial GLM. The Negative Binomial distribution is preferred for modeling nonnegative overdispersed data. The Bayesian inference is chosen to account for prior expert knowledge on regression coefficients in a small sample size setting and the hierarchical structure allows us to consider the dependence among the subsets. A Metropolis-Hastings-within-Gibbs algorithm is used to compute the posterior distribution of the parameters of interest through a data augmentation process. The Bayesian approach highly over-performs the classical maximum likelihood estimation in terms of goodness of fit, especially when the sample size decreases and the model complexity increases. Their respective performances have been examined in both the simulated and real life case studies.

C1412: **ABC methodology for a Y-linked two-sex branching model**
Presenter: **Cristina Gutierrez Perez**, University of Extremadura, Spain
Co-authors: Miguel Gonzalez Velasco, Rodrigo Martinez Quintana

The genetic causes of infertility in males or the history of paternal lineages are some relevant problems directly related to mutations in Y-linked genes. The interest of how these genes and their mutations evolve in a population leads us to introduce a new two-sex two-type branching process. The present work focuses on the development of Bayesian inference for this model, considering a parametric framework. The sample considered is given by the total number of females and males of each genotype (original allele and its mutations) up to some generation. Using the Approximate Bayesian Computation (ABC) methodology, we approximate the posterior distributions of the parameters of this model. The accuracy of the procedure is illustrated by way of simulated examples covering the different interactions between the parameters. We pay special attention to the case where a mutation gives rise to the reproductive incapacity of the individual. In this case, we introduce a modified ABC algorithm in order to estimate accurately the mutation rate and to discover the mutant males who do not generate offspring.

C1572: **New state-space modelling based inverse Bayesian learning, in the absence of training data**
Presenter: **Dalia Chakrabarty**, Univ of Leicester, United Kingdom

In Science, an unknown model function $\rho(X)$ that embodies a system property is often pursued. The function is sought, given noisy (and sometimes, missing) data V. We express the relation between the data and unknown as $V = \xi(\rho(X)$ and attempt writing the posterior density of the unknown given the data, except, $\xi(\cdot)$ is typically unknown. In principle, we can learn the unknown $\xi(\cdot)$ by modelling it with a Gaussian Process and training the parametrisation of the covariance structure of this GP using training data. However, often in real-life systems, training data is unavailable. Then, the unknown $\rho(X)$ can be embedded within the definition of the support of the state space pdf and the likelihood written in terms of this pdf. Measurement uncertainties are accounted for by convolving the resulting likelihood with the error distribution. In lieu of training data, the supports of the state space pdf, as well as the unknown model function are discretised, reducing the problem to the learning of a very large number of independent parameters, each of which represents the respective function over a chosen interval of the corresponding domain variable. Upon invoking suitable priors, the posterior density of these unknown parameters given the data is written and sampled from using an appropriate inferential tool. Applications to real and simulated data will be presented.

C1694: **Bayesian density regression for count data**
Presenter: **Charalampos Chanialidis**, University of Glasgow, United Kingdom
Co-authors: Ludger Evers, Tereza Neocleous

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.

C1308: **Bayesian bioequivalence test based on robust linear mixed effect model for bioavailability measures**
Presenter: **Yuh-Ing Chen**, National Central University, Taiwan
Co-authors: Chi-Shen Huang

Testing for the bioequivalence between two drugs is considered in a pharmacokinetic study under a 2x2 crossover design when the bioavailability measure, for example, the area under the drug concentration-time curve or the maximum concentration, individually estimated from each subject in the study is possibly outlying or skewed distributed. In the conventional linear mixed effect models (NLMEM) for the bioavailability measures, both the between-subject variation and measurement error are usually assumed to be normal variables. In practice, however, it occurs very often that the original or even the logarithm of the bioavailability measure may not be symmetrically distributed. To overcome the problem with skewed outliers, a more robust model is suggested, referred to as SNILMEM, where the two variables corresponding to the between-subject variation and measurement error are distributed according to different skew normal/independent distributions. Following the posterior probability of the bioequivalence between the two drugs based on a Bayesian analysis of the SNILMEM is computed and a Bayesian bioequivalence test which incorporates plausible prior distribution for the robust model is constructed. The results of a simulation study investigation of the posterior probability of bioequivalence are reported and discussed. Finally, a data set is illustrated based on the proposed Bayesian bioequivalence test based on the suggested robust model.

CS68 Room 6 VARIABLE SELECTION **Chair: Francesco Giordano**

C1315: **Structural and hierarchical variable selection using weighted regularization methods**
Presenter: **Samuel Mueller**, University of Sydney, Australia
Co-authors: Tanya Garcia, Raymond Carroll

The presentation is motivated through two dietary treatment studies in mice for which fecal microbial diversity was measures. Both studies used an obesity reversal paradigm and consisted of 30 mice equally and randomly assigned to one of three diets. For each mouse, data consisted of several phenotypes including Insulin and relative mRNA expression of CD68 in adipose, and microbial percentages from up to 186 microbes classified at the phylum, family, and genus levels. Strategies to use changes in microbiota composition to effect health improvements require knowing at which taxonomy level interventions should be aimed. Identifying these important levels is difficult, however, because most statistical methods only consider when the microbiota are classified at one taxonomy level, not multiple. Using L1 and L2 regularizations, a new variable selection method that identifies important features at multiple taxonomy levels is presented. The regularization parameters are chosen by a new, data-adaptive, repeated cross-validation approach which performed well. In simulation studies, the method outperformed competing methods: it more often selected significant variables, and had small false discovery rates and acceptable false positive rates.

C1491: **Variable selection in multivariate linear models for functional data**
Presenter: **Hidetoshi Matsui**, Kyushu University, Japan

Penalties with an ℓ_1 norm provide sparse solutions, and can be used for selecting variables in regression settings. We consider the problem of variable selection in multivariate linear models with the help of ℓ_1 regularization. In particular, we focus on functional linear models where multiple predictors are given as functions and multiple responses are scalars. Observed data corresponding to the predictors are supposed to be measured repeatedly at discrete time, and then they are treated as smooth functional data. Parameters included in the functional multivariate linear model are estimated by the penalized least squared method with the ℓ_1-type penalty. We construct the coordinate descent algorithm for the functional multivariate linear model. A tuning parameter which controls the degree of the regularization is chosen by Bayesian model selection criterion. We apply the proposed method to the analysis of real data, and then investigate the effectiveness of it.

C1388: **High-dimensional problems in ranking (ordinal regression) with Lasso**
Presenter: **Wojciech Rejchel**, Nicolaus Copernicus University, Poland

Variable selection is a fundamental challenge in statistical learning if one works with data sets containing huge amount of predictors. Finding significant (relevant) variables helps to better understand the problem and improves statistical inference. Methods based on penalized empirical risk minimization (for instance Lasso) have gained much attention recently. The main characteristic of these procedures is an ability to select significant variables and estimate unknown parameters simultaneously. The aim is applying these ideas in ranking (rank regression, ordinal regression) that is related to predicting or guessing ordering between objects on the basis of their observed features. The model is assumed to be "sparse" which means that there are only a few features that are significant in the model. The so called "oracle inequalities" are obtained, that is probabilistic inequalities comparing the risk of the estimator with the risk of the oracle that is the ranking rule that has the best balance between the risk and "sparsity" in the considered class.

C1320: **Model selection for misspecified logistic regression**
Presenter: **Pawel Teisseyre**, Polish Academy of Sciences, Poland
Co-authors: Jan Mielniczuk

The effect of model misspecification on variable selection when the logistic link is fitted is studied. The case of incorrect model specification is important as in real applications we have no prior knowledge of data generation process. Furthermore, one would like to have variable selection procedures which are resistant to departure from model assumptions. Under general conditions it is proven that the selection procedure based on Generalized Information Criterion picks variables pertaining to averaged Kullback-Leibler projection of vector of true parameters on the family of logistic models. Analogous result is proven for less computationally demanding two-stage procedure employing preordering of variables. The interplay between the set of relevant variables is studied for the true model and its projection and it is shown that when response function is monotone and not constant the latter is necessarily nonempty. Under certain condition on the distribution of x these two sets coincide, which implies that identifying relevant variables $\{k : \beta_k \neq 0\}$ under model misspecification is still possible. The results of simulation experiments are also presented in which how different link functions and distributions of explanatory variables influence the probability of correct model selection and the power of the significance tests are investigated.

C1637: **Adaptive penalized logistic regression for uncovering biomarker**
Presenter: **Heewon Park**, University of Tokyo, Japan
Co-authors: Yuichi Shiraishi, Seiya Imoto, Satoru Miyano

The penalized logistic regression based on L_1-type penalty (e.g., ridge, lasso, elastic net and etc.) has been widely used for identifying biomarkers and classifying samples based on microarray dataset. Although the existing penalized logistic regression performs well for uncovering biomarkers in disease classification, the existing L_1-type penalty suffers from inefficient of estimator and inconsistent of variable selection in regression modeling, due to the equally imposed penalty to all features. In order to incorporate the significance of each gene into gene selection, we propose an adaptive penalized logistic regression based on the Wilcoxon rank sum test. The proposed method effectively selects crucial features based on a discriminately imposed adaptive penalty to each gene depending on important degree of gene in classification, and thus we can improve the classification accuracy without noisy genes. We apply the proposed method to the Sanger dataset, and perform cancer cell lines classification based on not a single gene but gene modules via a principal component analysis. We can see through the Monte Carlo experiments and real world example that the proposed adaptive penalized logistic regression outperforms for feature selection and classification even in the high dimensional dataset (e.g., microarray dataset).

CS70 Room 4 ECONOMETRIC MODELS Chair: Joachim Schnurbus

C1700: **Improved parameter estimation and predictions using the generalized least squares in the presence of heteroskedasticity**
Presenter: **Edy Zahnd**, Forex and Economic Research, Switzerland
Co-authors: Valentin Rousson

In a linear regression (time-series) model, one usually assumes homoskedastic errors. In the presence of heteroskedastic errors, it is well known that the ordinary least squares (OLS) estimates of the regression coefficients remain consistent, but that statistical inference may not be valid because estimates of standard errors will then be biased. Improved estimates and valid inference can be achieved using generalized least squares (GLS) which consists in weighting each observation by the inverse of the corresponding error's variance. To achieve this, one has to model the variance error in function of the time, which may not be an easy task. If such a model is not available, we propose to use a two-stage procedure, which consists in a second stage to weight each observation by the inverse of its squared empirical residual obtained in a first stage. A similar procedure has been proposed by White in 1980 to get consistent estimates of standard errors, and is sometimes referred to as a "White-washing" procedure. However, White did not consider this procedure to update the regression coefficients estimates. We investigate via simulation the performance of such a

two-stage GLS estimate under classical settings in 3 cases: a linear regression model, a linear regression model with an autoregressive process of order 1 for the residuals, and finally an autoregressive process of order 1 alone, with various kinds of heteroskedastic errors in each case. We show that the mean squared error for the parameter estimation can be reduced significantly compared to the OLS estimate, which may further also lead to a decrease of the mean squared prediction error when one is using the two-stage GLS estimate to predict future observations.

C1501: **Endogeneity and varying coefficient models**
Presenter: **Stefan Sperlich**, Univsersity of Geneva, Switzerland
Co-authors: Raoul Theler

The aim is to study problems that occur when using linear models in combination with standard methods like OLS or IV estimation in the presence of heterogeneous returns. They are illustrated along theory, simulations and examples. We show that varying-coefficient models have a strong potentiality in econometrics and causal statistics. They offer a way to model heterogeneity in many interesting situations and can help to overcome typical endogeneity problems. Also alternative IV estimations are proposed when dealing with observable and unobservable dependent heterogeneity.

C1513: **Jackknife estimation in the presence of a near-unit root**
Presenter: **Maria Kyriacou**, University of Southampton, United Kingdom
Co-authors: Marcus Chambers

The use of the non-overlapping jackknife in a local-to-unity framework is studied. We show that in the presence of a near-unit root each sub-sample estimator has different limit distribution from the full-sample least squares estimator of the autoregressive parameter. In fact, we show that the limit distribution of each sub-sample estimator depends on both the sub-sample indicator and the number of non-overlapping sub-samples. These findings imply that the expansions for the bias used to derive the usual jackknife estimator are not correct under a near-unit root setting. To overcome this issue, we derive the joint moment generating function (MGF) of the relevant functionals of the Ornstein-Ulhenbeck process over the subinterval $[a,b]$. This allows us to define an optimal jackknife estimator which utilises optimal weights that achieve the aim of first-order bias removal under a near-unit root. Simulations indicate that our proposed jackknife estimator is capable of producing substantial bias and Root Mean Square Error (RMSE) reduction over its full-sample least squares estimator for all values of the localisation parameter and for a wide range of sample sizes.

C1476: **Fancy non-(non)-linearities in applied economics**
Presenter: **Setareh Ranjbar**, University of Geneva, Switzerland
Co-authors: Stefan Andréas Sperlich

Recent developments in the field of econometrics and statistics provide us with a vast variety of analytical tools. This calls for more caution when using or adopting them to applications of real life. In this paper we provide two examples that highlight some of the common mistakes in applying sophisticated methods in empirical studies. In the first example we use a class of ordered discrete response model which is often used in psychology or socio-economics to study the subjective well-being. We show that adopting generalised partial linear model for making inference on the non-parametric part of the model, will not free one from the correct specification of the parametric part. In contrary, misspecification in the parametric part will adversely affect the estimation of the non-parametric part. The second example refers to techniques that have been widely used in policy evaluation. It shows why global estimators are not an adequate tool for prediction. Especially in cases where the distributions of the confounders differ seriously over groups, a situation that is typically desired, they easily lead to severely biased estimates.

C1360: **An exchange rate model with market pressures and contagion effect**
Presenter: **Aleksander Welfe**, University of Lodz, Poland
Co-authors: Wojciech Grabowski

The exchange rate model proposed combines PPP and UIP hypotheses with the effect of risk aversion in financial markets, the impact of currency market pressures, and the contagion effect between countries of the same region. A new indicator was developed to identify the periods of pressure in the currency markets, which allows not only short-term but also long-term instabilities to be detected. Because in addition to being an explanatory variable (in the exchange rate equation) the indicator is also explained by the model variables and because the time series are generated by non-stationary stochastic processes, cointegration with dichotomous variables was chosen as an appropriate tool for inference. The asymptotic distribution of the trace statistics necessary to determine the size of the cointegrated space had to be derived and the critical values were simulated. The empirical results obtained with the Polish data confirm that instabilities in currency market are caused not only by fundamental factors, such as the volume of economic activity and the countryś balance of payments, but also by the contagion effect resulting from investors' inclination to perceive Poland and its neighbours, the Czech Republic and Hungary, as one group. On the other hand, it has been shown that the exchange rate is increased by rising inflation and interest rates (vis-á-vis the reference countries), higher risk attributed to the particular market (country) and market instabilities.

CS57 Room 13 COMPUTATIONAL STATISTICS V Chair: Veronika Czellar

C1617: **Application of Kalman Filter with alpha-stable distribution**
Presenter: **Pavel Mozgunov**, National Research University Higher School of Economics, Russia

The behavior of the Kalman filter state estimates in the case of distribution with heavy tails is studied. 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.

C1426: **A quadratic Kalman filter**
Presenter: **Guillaume Roussellet**, Banque de France - CREST, France
Co-authors: Alain Monfort, Jean-Paul Renne

A new filtering and smoothing technique for non-linear state-space models is proposed. Observed variables are quadratic functions of latent factors following a Gaussian VAR. Stacking the vector of factors with its vectorized outer-product, we form an augmented state vector whose first two conditional moments are known in closed-form. We also provide analytical formulae for the unconditional moments of this augmented vector. Our new quadratic Kalman filter (Qkf) exploits these properties to formulate fast and simple filtering and smoothing algorithms. A first simulation study emphasizes that the Qkf outperforms the extended and unscented approaches

in the filtering exercise showing up to 70% RMSEs improvement of filtered values. Second, we provide evidence that Qkf-based maximum-likelihood estimates of model parameters always possess lower bias or lower RMSEs that the alternative estimators.

C1661: **A functional Hodrick Prescott filter**
Presenter: **Hiba Nassar**, Linnaeus University, Sweden
Co-authors: Boualem Djehiche
The aim is to propose a functional version of the Hodrick-Prescott filter for functional data which take values in an infinite dimensional separable Hilbert space. We further characterize the associated optimal smoothing parameter when the associated linear operator is compact and the underlying distribution of the data is Gaussian.

C1678: **Beta models for random hypergraphs with a given degree sequence**
Presenter: **Despina Stasi**, Illinois Institute of Technology, United States
Co-authors: Kayvan Sadeghi, Alessandro Rinaldo, Sonja Petrovic, Stephen Fienberg
The beta model for random hypergraphs is introduced 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 in- stead 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.

C1658: **Enumeration and statistics of the prime pairs computation**
Presenter: **Disuke Ishii**, Okayama University of Science, Japan
Co-authors: Ryuichi Sawae, Yoshiyuki Mori
A twin prime $(p, p+2)$ is a prime number that has a gap 2, and a cousin prime $(p, p+4)$ is a prime number that has a gap 4. Exactly definable any gap even numbers, in other word, for the every natural number of k, there are infinitely many prime pairs $(p, p+2k)$. The count $\pi_2(x)$ of the number of twin prime pairs up to x, also $\pi_4(x)$ is cousin prime counting function, as well as the prime counting function $\pi(x)$. Hardy and Littlewood made a conjecture an asymptotic formula for the number of prime pairs. V.Brun was able to show that the sum of inverses of the twin primes was convergent. That value B_2 is called Brun's constant. Our computer results of the prime pairs are calculated up to 10^{15}, prime pairs gap $2k \leq 64$, using sieving method. Furthermore, taking into account the statistical data, we can surmise that Hardy-Littlewood conjecture is estimated true on our bound.

CS77 Room 5 METHODOLOGICAL STATISTICS II Chair: Keith Knight

C1625: **Log-linear multidimensional Rasch model for capture-recapture**
Presenter: **Elvira Pelle**, University of Milano-Bicocca, Italy
Co-authors: David J. Hessen, Peter G. M. van der Heijden
Traditional capture-recapture method assumes the homogeneity of the capture probability. However, differences of character or behaviour between individuals may occur and models that allow for varying susceptibility to capture through 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 view 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.

C1602: **Estimating the homogeneous population size via empirical Bayes method in a complex dual-record system**
Presenter: **Kiranmoy Chatterjee**, Indian Statistical Institute Kolkata, India
Co-authors: Diganta Mukherjee
Estimation of the size of a human population or the number of vital events occurred at a specified time is a problem that has attracted much attention specifically in the domain of epidemiology and official statistics. Dual-record system (equivalently two sample Capture-recapture experiment) is a common procedure assuming independence between the two capture probabilities. However, this assumption does not hold in many situations as individuals possess behavioral change after being captured first time. We focus on the dual-record system model with both the time as well as behavioral response variation in capture probabilities. The relevant model suffers from parameter identifiability problem which results unsatisfactory performances from two existing methodologies which are actually built in a spirit of multiple lists problem. Here two simple approaches are proposed from which empirical Bayes estimates can be obtained using the idea of Empirical Bayes Gibbs sampling. Extensive simulation studies are carried out to evaluate their performances and compare with few competitive approaches. These approaches are also illustrated by a real data example. Finally, the improvements and other features of our methods are explored depending upon the availability of knowledge on the nature of behavioral response effect.

C1698: **Score function of distribution and heavy tails**
Presenter: **Zdenek Fabian**, ASCR, Czech Republic
The purpose is to 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 heavy-tailed distribution and random samples from it are described by the typical value and variability. Further, we describe 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 robust.

C1632: **Empirical Likelihood approach to combining information from multiple surveys**
Presenter: **Ewa Kabzinska**, University of Southampton, United Kingdom
Co-authors: Yves G. Berger
It is often the case that several surveys carried out independently in the same population measure some common variables. It may be beneficial to combine this separately gathered information in order to increase precision and ensure that estimates are consistent across surveys. The classic approach relies on a GREG-family estimator. The Pseudo Empirical Likelihood methods have also been applied in this context. We propose to use the Empirical Likelihood (EL) approach. Empirical Likelihood is a non-parametric method that uses a data driven likelihood ratio function for inference. While it does not rely on distribution of the data, EL possesses some attractive features of parametric likelihood methods. It may be used to incorporate consistency and benchmark constraints. The EL approach is asymptotically equivalent to generalized regression estimation. However, it always gives positive weights, which is not the case for GREG-family estimators. We evaluate performance of the proposed estimator in a series of simulations. We also

consider the numerical aspects of optimization necessary to get the Empirical Likelihood estimator. We discuss some alternatives to the Newton-Raphson algorithm, which is commonly used in these settings.

C1352: **Root n estimates of vectors of integrated density partial derivative functionals**

Presenter: **Tiee-Jian Wu**, National Cheng-Kung University, Taiwan

Co-authors: Chih-Yuan Hsu, Huang-Yu Chen, Hui-Chun Yu

Based on a random sample of size n from an unknown d-dimensional density f, the nonparametric estimations of a single integrated density partial derivative functional as well as a vector of such functionals are considered. These single and vector functionals are important in a number of contexts. The purpose of this paper is to derive the information bounds for such estimations and propose estimates that are asymptotically optimal. The proposed estimates are constructed in the frequency domain by using the sample characteristic function. For every d and sufficiently smooth f, it is shown that the proposed estimates are asymptotically normal, attain the optimal $O_p(n^{-1/2})$ convergence rate and achieve the (conjectured) information bounds. In simulation studies the superior performances of the proposed estimates are clearly demonstrated.

Thursday 21.08.2014 16:15 - 18:00 Parallel Session J

OS39 Room 4 DIMENSION REDUCTION AND CLASSIFICATION Chair: Patrick Groenen

C1684: **Two-way clustering with least-squares matrix decompositions**
Presenter: **Pieter Schoonees**, Erasmus University Rotterdam, Netherlands
Co-authors: Patrick JF Groenen
Multivariate categorical data collected from surveys or questionnaires can be considered to contain two types of unobserved segments. We jointly consider clusters related to the popularity of the rating categories (response style segments) as well as substantive segments which attach different levels of importance to the set of items (variables). This is achieved by using least-squares matrix factorizations on the set of all individual rating by item indicator matrices. By assuming that the response style segments represent the attractiveness of the different rating categories irrespective of the item being answered, it is shown that the proposed loss function decomposes into two separable optimization problems. An algorithm for this problem is discussed. In the context of an empirical application, we show how the method results in low-dimensional biplots where the coordinates of the rating categories are constant across substantive segments but those of the items vary. The results are compared to that of a computationally intensive model-based approach.

C1672: **Genetic risk scores using ridge regression**
Presenter: **Ronald de Vlaming**, Erasmus School of Economics, Netherlands
Co-authors: Patrick Groenen
Genome-wide association studies (GWAS) are used to explore the genetic architecture of outcomes such as diseases and behaviour. In a GWAS the data often consist of $M > 1,000,000$ predictors called SNPs, observed for a limited number of respondents N (e.g., $N < 10,000$). In case the outcome variable is quantitative, classical multiple regression would be a natural method for predicting the outcome. However, ordinary least squares (OLS) is troubled by overfitting and multicollinearity. In fact, as $N \ll M$, the OLS estimator does not have a solution. The standard has become to estimate each SNP effect within a set of SNPs in a separate regression. The predicted outcome – the genetic risk score – is the linear sum, weighted according to the regression weights. Instead, we propose to return to one regression using all SNPs as predictors, adding a penalty to overcome the problem of multicollinearity. We choose ridge regression (RR), which has a penalty term consisting of the sum of squared regression weights. We provide a computationally efficient RR implementation suitable for GWAS. Furthermore, by applying the kernel trick we efficiently incorporate simple gene–gene interactions. We compare the predictive accuracy of RR and classical GWAS estimates.

C1652: **Cluster correspondence analysis**
Presenter: **Michel van de Velden**, Erasmus University Rotterdam, Netherlands
Co-authors: Alfonso Iodice D'Enza, Francesco Palumbo
A new method is proposed that combines dimension reduction and cluster analysis for categorical data. A least-squares objective function is formulated that approximates the cluster by variables cross-tabulation. Consequently, individual observations are assigned to clusters in such a way that the distributions over the categorical variables for the different clusters are optimally separated. In addition, we consider existing methods with similar objectives in a unified framework and derive a new algorithm for the GROUPALS method. A simulation study is used to appraise the methods in a structural fashion and compare the performance with respect to each other. Moreover, with respect to the clustering, we compare the results to k-means cluster analysis based directly on the full dimensional data. Our results show that, in general, all methods perform similarly. For nearly all considered scenarios, the joint dimension reduction and clustering methods outperform, with respect to the retrieval of the true underlying cluster structure, the full dimensional clustering. Furthermore, for some scenarios covered in our simulation study, certain methods appear to be more appropriate than others.

C1348: **An extended comparison of multiclass support vector machines**
Presenter: **Gertjan van den Burg**, Erasmus University Rotterdam, Netherlands
Co-authors: Patrick Groenen
The Support Vector Machine (SVM) has been shown to be very successful for binary classification tasks. In recent years, several extensions have been proposed to generalize the binary SVM to multiclass classification tasks while retaining the desirable properties of the binary SVM. These methods include heuristic approaches using the binary SVM, error correcting code approaches, and single-machine approaches, each with different merits. Previous comparisons of multiclass SVMs have shown conflicting results and do not include the most recently developed methods. In part, this may have been the result of not using proper evaluation measures or not using careful statistical analysis in comparing the methods. The aim is to provide solutions to these problems. Some results obtained from carefully controlled experiments on 13 empirical datasets are presented. A unified software library that has all methods implemented in C is proposed. This library allows a fair comparison in time and allows reproducibility of the obtained results. All methods are compared with respect to classification performance through the Adjusted Rand Index, and with respect to training time through CPU time. To compare different classifiers, performance profiles and rank tests are used. The results show that the One-versus-One heuristic, DAG procedure, and GenSVM are amongst the best performing and fastest SVM methods for multiclass classification.

C1636: **GenSVM: a general multiclass support vector machine**
Presenter: **Patrick Groenen**, Erasmus University Rotterdam, Netherlands
Co-authors: Gertjan van den Burg
A flexible multiclass SVM, GenSVM, for $K \geq 2$ classes, is proposed. In case $K = 2$, it simplifies to a standard binary SVM. The method has a simple geometric representation of the class boundaries (separating hyperplanes) in a $K-1$ dimensional simplex space. If an object in class k is at the proper side for each pair of classes j and k (with $j \neq k$) then the object is correctly classified and it receives error of zero. The error for an object is based on functions on the distance of the object to each of the separating hyperplanes with the other classes. GenSVM is flexible in the hinge function that is used for calculating the error. It uses the Huberized hinge errors that have as special cases the linear and quadratic hinges. It is also flexible in how these errors are added: it uses the L_p norm of the Huberized hinge error. This convex loss function has some existing multiclass SVM loss functions as a special case. By using the L_2 norm it becomes possible to use the Euclidean distance of an observation to the margin as the measure of error thereby making the error independent of the number classes where it has a misclassification.

OS34 Room 5 SYMBOLIC DATA ANALYSIS IN THE BIG DATA AGE Chair: Javier Arroyo

C1470: **The data accumulation graph (DAG): visualization of high dimensional complex data**
Presenter: **Paula Brito**, Universidade do Porto, Portugal
Co-authors: Manabu Ichino

This paper presents the data accumulation graph (DAG) to visualize multi dimensional data. For a data (N objects)$\times$(d features), the DAG plots N objects as parallel monotone line graphs. The DAG shows macroscopic properties of objects as the differences of the size of line graphs, and it shows also microscopic properties of objects as the differences of the shapes of the monotone line graphs. We present the DAGs for actual data sets including an interval-valued data, a histogram-valued data, and a high dimensional data in order to show the usefulness of the proposed method.

C1464: **Aggregated symbolic data description including real and categorical variables**
Presenter: **Junji Nakano**, The Institute of Statistical Mathematics, Japan
Co-authors: Yoshikazu Yamamoto, Nobuo Shimizu, Takeshi Fujiwara

When we have huge amount of data, it is difficult to focus on each individual data. Instead we are often interested in the relationships among meaningful groups of individual data. Then we summarize a group of individual data by several statistics and call them "aggregated symbolic data". We consider the case where each individual data has both real and categorical variables. We know that mean and variance are simple descriptive statistics for a real variable. For summarizing a categorical variable, we assign real values to categories to distinguish groups as much as possible. For aggregated symbolic data, we need to consider measures of association between variables in a group. A simple measure of association between two real variables is the correlation coefficient. We propose measures of association between categorical variables, and between real and categorical variables. We describe aggregated symbolic data by using them and propose an extended parallel coordinate plot to illustrate them. "Big" real estate data are analyzed by using these methods.

C1460: **Analysis of sensing data with moving functional methods**
Presenter: **Masahiro Mizuta**, Hokkaido University, Japan
Co-authors: Yusuke Matsui

One of the challenging tasks against big data is on the "volume". We could see many situations where data are recorded and stored into huge scale databases automatically. A typical case might be sensing data incorporated with machine to machine (M2M) system. That is, many sensors continuously observe the measurements and they are piled into databases via computer networks. For the data, the primary issue would be the "volume", and we also have to deal with the "variety" i.e. structured internal variation. In this paper, we discuss the analysis of sensing data from the viewpoint of symbolic data analysis. We regard each sensor as a concept and its description as a function of time. Our focus is to investigate the proximity between the concepts that are varied over the time intervals. We exploit the moving functional methods such as moving clustering and moving multidimensional scaling. We demonstrate the methods with the sensing data of radiation levels observed by about 4000 sensors in Japan. The sensors include various behaviors affected by climate conditions and geometrical features and so on. We analyze the proximities of sensors that are varied over the periods.

C1708: **Model based symbolic description for big data analysis**
Presenter: **Carlo Drago**, University of Rome Niccolo Cusano, Italy
Co-authors: Carlo Lauro, Germana Scepi

This contribution is in the context of big data analysis. In the big data analysis we encounter the problem of defining a suitable technique of representation in order to retain the structure of data and to allow the identification of an underlying model. In particular, in big data characterized by temporal correlations, the research of a model becomes crucial for forecasting aims. Here we propose to represent this type of data by means of a distribution based on a kernel density estimator which graphically generates a beanplot. This graphic representation can be considered a peculiar type of symbolic data, like histogram symbolic data, where all the observations are considered and the variability of the whole distribution is taking into account. Starting from big financial data, we choice initially to divide the data in several temporal intervals on the basis of the aim of the analysis and of a priori information. A beanplot for each period can be so generate and a time series of beanplot is obtained. Beanplot time series (BTS) describe situations where a distribution of values is available for each instant of time. This allows us to solve the representation problem by retaining a lot of fundamental information, like the trend, the intra-period and inter-periodal variations, the presence of outliers and break down point; information which could be lost otherwise. Furthermore, we decide to model each single beanplot by a peculiar mixture model. By retaining the main parameters of the different models it is possible to derive the different temporal components of the distributions arising in the mixture. It is important to underline that several relevant aspects must be taking into account for obtaining the "best models". For example, it is very important to choose an adequate window as temporal interval. At the same time the selection of the number of mixture is very important to accomplish a criteria both of simplicity that comparability and usability. The mixture models and, in particular, their parameter can be used in forecasting and clustering.

C1706: **SDA for mixed-type data and its application to analysis of environmental radio activity level data**
Presenter: **Yusuke Matsui**, Hokkaido University, Japan
Co-authors: Masahiro Mizuta

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. 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. Various processes collect the data. 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.

CS74 Room 13 TIME SERIES Chair: Jeroen Rombouts

C1283: **Time series outlier detection using singular spectrum analysis and robust principal component analysis**
Presenter: **Jacques De Klerk**, North-West University, South Africa

Singular Spectrum Analysis (SSA) is a powerful non-parametric time series technique with wide application in time series analysis. SSA is particularly powerful for time series exhibiting seasonal variation with/without trend components. SSA can be applied to time series found in market research, economics, meteorology and oceanology, to name but a few. SSA places a univariate time series into a multivariate framework by unfolding into a Hankel structured matrix. Outliers that might be present in time series can unduly influence model fitting, forecasting results and confidence intervals constructed using bootstrap methodology. The aim

is to compare outlier identification techniques in SSA by simulating time series from the broad spectrum of time series that SSA can handle. Specific attention is paid to modern robust principal component analysis techniques such as ROBPCA which employs projection pursuit combined with estimation of robust covariance matrices. The latter is employed to outlier maps, which represents multivariate data in a two dimensional plot consisting of projected orthogonal distances plotted against score distances, in order to identify outliers. Promising results are obtained by applying robust principal component analysis to SSA. A well-known time series with an additive outlier present is used to illustrate the usefulness of the techniques.

C1336: **Analysis for multivariate time series using local fractal dimension in n-dimensional space**
Presenter: **Kenichi Kamijo**, Toyo University, Japan

When three discrete time series for variate x, y and z are given, a time series consisting of these coordinates (x, y, z), which are individually extracted from the three time series at the corresponding discrete time, can be newly constructed in 3-dimensional space. This can be considered a kind of discrete orbit in 3-dimensional space. The same procedure can be easily expanded to general multivariate time series. The local fractal dimension, LFD, has been defined and calculated in a finite short processing window set on a constructed orbit in n-dimensional space. An analysis for the said discrete orbit has been proposed using LFD to evaluate the fractal structure in the space. The moving LFD can be easily obtained by sliding the said processing window along the line on the orbit. As an example, the logistic mapping has been selected at each coordinate, and a computer simulation has been dynamically performed. Results show that the probability distribution of the moving LFD becomes almost a normal distribution within a certain restricted range of the control parameter concerned with the occurrence of the so-called chaos.

C1639: **Modelling multivariate time series by structural equations modelling and segmentation approach**
Presenter: **Christian Derquenne**, EDF Research and Development, France

A method to build a complex model for irregular and multivariate time series is proposed. First, to take into account non-linearity, break, volatility between them, we chose to segment these series as linear segments to eliminate non-stationarity, standardizing the raw series with thereof. Then, we used an exploratory approach using free structural equation models to establish links between these standardized variables. The latter allows building blocks (groups) homogeneous time series, and then summarize these ones with latent variables, to seek significant links between them and finally apply the partial least squares approach on the proposed free model. In the application, we compared the results of two models, the first on differentiated series and the second on the standardized series using segmentation. This has shown a greater consistency of results in terms of stability coefficients, significant relationships and business aspects for the model in standardized series using segmentation. Our future work will involve comparison with other methods, 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.

C1586: **A combined nonparametric test for seasonal unit roots**
Presenter: **Robert Kunst**, Institute for Advanced Studies, Austria

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 the RURS (seasonal range unit roots) test statistic and a variant 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.

C1549: **Methods for estimating a time series of densities**
Presenter: **Thilaksha Tharanganie**, Monash University, Australia
Co-authors: Rob Hyndman

The problem of estimating a time series of density functions is considered. A data set comprising many observations is recorded at each time period, and the associated probability density function is to be estimated for each time period. It is assumed that the densities change slowly over time and that neighbouring densities are similar but not identical. We consider several methods for estimating a time series of densities: (1) A logspline approach applied to each data set separately, where each estimated density has common knots but different coefficients; (2) A logspline approach applied to all data simultaneously with common knots and kernel weights to account for time variation; (3) A conditional kernel approach with two new bandwidth selection approaches that account for the discrete conditioning variable (time). We apply our methods to a set of UK household income data for the period 1961-1991, with approximately 6000 observations per year. Probability integral transforms and proper scoring rules are used to compare the estimates. We conclude that the method involving logsplines with common knots and kernel weights gives the best results for this data set.

CS71 Room 6 REGRESSION MODELS II Chair: Stefan Sperlich

C1518: **Model selection for functional mixed model via Gaussian process regression**
Presenter: **Toshihiro Misumi**, Astellas Pharma Inc, Japan
Co-authors: Sadanori Konishi

In recent years, a functional mixed model (FMM) has attracted considerable attention in longitudinal data analysis, because of its flexibility. The FMM consists of a fixed effect or a population mean function and some subject-specific functional random effects. In this presentation, we introduce the FMM constructed by using a basis expansion technique and a Gaussian process regression, and consider the model evaluation and selection problem for the estimated model. When estimating the unknown parameters included in the FMM by the maximum penalized marginal likelihood method, the FMM is extremely sensitive to the choice of tuning parameters. In order to appropriately select them, we derive two model selection criteria for the FMM based on the perspective of information or Bayesian theories by using a marginalization approach. We conduct Monte Carlo simulations to investigate the effectiveness of our proposed modeling procedures. The proposed modeling procedures for the FMM are applied to the analysis of a longitudinal gene expression data.

C1558: **Data-driven wavelet resolution choice in multichannel box-car deconvolution with long memory**
Presenter: **Justin Wishart**, University of New South Wales, Australia

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.

C1356: **Regression rank scores and testing of heteroscedasticity**
Presenter: **Radim Navratil**, Technical University of Liberec, Czech Republic
Homoscedasticity is often tacitly assumed in the analysis of linear models, both classical and robust. To avoid a negative consequence of ignored heteroscedasticity, it is recommendable to either analyze its possibility before starting an inference on the parameters of the model, or look for an approach invariant to heteroscedasticity, if there is one. In the linear regression model with heteroscedastic errors, nonparametric tests for regression under nuisance heteroscedasticity, and tests for heteroscedasticity under nuisance regression are proposed. Both types of tests are based on suitable ancillary statistics for the nuisance parameters; hence they avoid their estimation, in contradistinction to tests proposed in the literature. These test statistics are linear forms of regression rank scores that are generalization of usual ranks for regression models. A simulation study, as well as an application of tests to real data, illustrate their good performance.

C1582: **Conditional quantile estimation using optimal quantization: a numerical study**
Presenter: **Isabelle Charlier**, Université Libre de Bruxelles and Université de Bordeaux, Belgium
Co-authors: Davy Paindaveine, Jérôme Saracco
The aim is to 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 us 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 ($N \to \infty$) 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.

C1317: **The extensively corrected score for measurement error models**
Presenter: **Yih-Huei Huang**, Tamkang University, Taiwan
Co-authors: Chi-Chung Wen, Yu-Hua Hsu
In measurement error problems, two major estimation methods are conditional score and corrected score. They are functional methods that require no parametric assumptions on true covariates. The conditional score requires that a suitable sufficient statistic for the true covariate can be found, while the corrected score requires that the object score function can be estimated without bias. These assumptions limit their ranges of applications. An extensively corrected score is proposed as an extension of the corrected score. It yields consistent estimations in many cases for which neither the conditional score nor corrected score is feasible. Its construction is demonstrated in generalized linear models and the Cox proportional hazards model. Its performance is assessed by simulations and its implementation is illustrated by a real example.

CS82 Room 20 APPLIED STATISTICS AND DATA ANALYSIS III Chair: Ludger Evers

C1619: **Combining sub(up)-approximations of different type to improve a solution**
Presenter: **Bernard Fichet**, Aix-Marseille University, France
The aim is to study 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.

C1723: **Joint analysis of closed and-open ended questions in a survey about the Tunisian revolution**
Presenter: **Rjiba Sadika**, University of Scientific Economics and Management Sousse, Tunisia
Co-authors: Mireille Summa Gettler, Saloua Benammou
Two major results are found. 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 disturbance 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.

C1664: **Variable selection to construct indicators of quality of life for data structured in groups**
Presenter: **Amaury Labenne**, IRSTEA, France
Co-authors: Marie Chavent, Vanessa Kuentz-Simonet, Saracco Jerome
The analysis and measurement of quality of life may be made via two complementary approaches. The first one, based on survey of individuals, concerns the analysis of levels of life satisfaction. We focus here on the second one, based on national data, which analyses living conditions of people. The aim is to create composite indices of living conditions. According to authors, the components of quality of life are related to different themes (groups of variables): "Family conditions", "Employment", "Housing", ... For this purpose, dimension reduction methods are particularly suitable. Multiple Factor Analysis (MFA) is a method designed to handle data structured into groups of quantitative variables. In our study, each theme is composed of a group of quantitative and/or categorical variables. Since our data are naturally structured in groups of variables, we develop an extension of MFA for mixed data type, called MFAmix. Thus the principal components from MFAmix are our composite indices for measuring quality of life. However, the creation of these indices raises two questions. How many principal components keep creating indices? How select a limited number of variables to get similar indices for easier interpretation? We propose answers to these questions in this communication.

C1610: **Statistical registration of frontal view gait silhouette with application to gait analysis**
Presenter: **Kosuke Okusa**, Kyushu University, Japan
Co-authors: Toshinari Kamakura
The problem of analyzing and classifying frontal view gait video data is studied. We focus on the shape scale changing in the frontal view human gait silhouette, we estimate scale parameters using the statistical registration and modeling on a video data. In the biometrics research area, many of researchers mainly using human silhouette shape with application to the human gait authentication. However, they did not focus on the scale registration in the frontal view gait analysis case. They just normalize the human silhouette and it applies to the gait authentication. It is reasonable to suppose that the normalization of the human silhouette lost a lot of gait information. 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

for the frontal view human gait authentication. As a result, our model shows good performance for the scale estimation and gait authentication.

C1551: **Estimation of space deformation models for non-stationary random functions**
Presenter: **Francky Fouedjio**, MINES ParisTech, France
Co-authors: Nicolas Desassis, Thomas Romary, Jacques Rivoirard

Stationary random functions have been successfully applied in geostatistical applications for decades. In some instances, the assumption of a homogeneous spatial dependence structure across the entire domain of interest is unrealistic. A useful approach for modeling and estimating non-stationary spatial dependence structure has been developed. It consists on reducing a non-stationary random function to stationary and isotropy via a bijective bi-continuous deformation of the index space. So far, this approach has been really working in the context of data from several independent realizations of a random function. In this work, we propose an approach for non-stationary geostatistical modeling using space deformation in the context of a single realization with possibly irregularly spaced data. The estimation method is based on a non-stationary variogram estimator which serves as dissimilarity measure between two locations in the geographical space. The proposed procedure combines aspects of kernel smoothing, multi-dimensional scaling and radial basis functions to transform the originally non-stationary random function towards a new deformed space where it is stationary and isotropic. Standard techniques for prediction and simulation can be applied in the deformed space. The predicted and simulated results are then mapped back into the original space. On a simulated data, the method is able to find the true deformation. A comparison scheme of ordinary kriging under stationary and non-stationary assumptions demonstrates that the proposed approach has better prediction performance on both a simulated and real datasets. Finally, we also show that the method can be seen as a visualisation tool of the non-stationarity.

CS59 Room 19 COMPUTATIONAL STATISTICS VI Chair: Jeffrey S. Racine

C1612: **A computational efficient strategy for estimating the block recursive simultaneous equations model**
Presenter: **Cristian Gatu**, Alexandru Ioan Cuza University of Iasi, Romania
Co-authors: Mircea Ioan Cosbuc, Erricos John Kontoghiorghes

A new computational efficient strategy for estimating the block recursive simultaneous equations model with non-full rank variance-covariance matrix is proposed. The estimation procedure uses the generalized singular value decomposition and derives a numerically-stable three-stage least squares estimator. The algorithm exploits the block-diagonal and banded structures of the matrices involved in the factorizations as well as the block recursive properties of the model. Experimental results illustrate the computational efficiency of the new estimation algorithm when compared with the equivalent method that ignores the block recursive structure.

C1497: **Recursive computation of higher order multivariate Gaussian density derivatives**
Presenter: **Jose E Chacon**, Universidad de Extremadura, Spain
Co-authors: Tarn Duong

Many developments in Mathematics involve the computation of higher order derivatives of Gaussian density functions. In the multivariate case, it is necessary to first establish a convenient formulation to assemble all the partial derivatives. Theoretically, it is possible to derive concise explicit expressions for the multivariate Gaussian density derivatives of an arbitrarily high order if we formally arrange all the partial derivatives into a high-dimensional vector. However, the huge matrices involved in these general expressions have traditionally made them of little practical use. In this talk we propose several recursive algorithms to overcome these difficulties that allow us to compute these higher order derivatives in a very efficient way.

C1573: **Density and Distribution Function estimation through iterates of fractional Bernstein Operators**
Presenter: **Claude Mante**, CNRS, France

A method for distribution function and density estimation with Bernstein polynomials is described. 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.

C1394: **Using storm for scaleable sequential statistical inference**
Presenter: **Simon Wilson**, Trinity College Dublin, Ireland
Co-authors: Tiep Mai, Peter Cogan, Arnab Bhattacharya, Oscar Robles Sanchez, Louis Aslett, Sean O'Riordain

We describe storm, an open-source, scaleable and fault tolerant environment for doing streaming data analysis, and discuss its use in both simple data processing examples - for which it was designed - and more sophisticated sequential learning algorithms. The differences between it and most computing environments, which focus on batch analysis, are investigated. 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. These illustrate how an algorithm is implemented in Storm and demonstrate the advantages of using it over a more batch analysis programming environment. To conclude, the difficulties of implementing iterative sequential learning algorithms, such as sequential Monte Carlo, are discussed and solutions are proposed.

C1537: **Performance of acceleration of ALS algorithm in nonlinear PCA**
Presenter: **Yuichi Mori**, Okayama University of Science, Japan
Co-authors: Masahiro Kuroda, Masaya Iizuka, Michio Sakakihara

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 (vε-ALS) which increases the speed of convergence, and have observed that computational costs by vε-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 vε-ALS by simulation, in which NLPCA with vε-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 vε-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 vε-ALS reduces the computational costs.

Friday 22.08.2014 9:00 - 10:45 Parallel Session K

CS40 Room 6 CONTRIBUTIONS TO STATISTICS OF EXTREME VALUES II Chair: Ivette Gomes

C1416: **Semiparametric exponential families for heavy-tailed data**
Presenter: **Stefan Wager**, Stanford University, United States
Co-authors: William Fithian

A semiparametric method for estimating the mean of a heavy tailed population given a relatively small sample from that population and a larger sample from a related background population is proposed. We model the tail of the small sample as an exponential tilt of the better-observed large-sample tail using a robust sufficient statistic motivated by extreme value theory, and give both theoretical and empirical evidence that our method outperforms the sample mean and the Winsorized mean. If the small and large samples are drawn from regularly varying distributions with the same tail index, we show under mild additional conditions that our method achieves a better rate of convergence than the optimally Winsorized mean. Applying our method to both simulated data and a large controlled experiment conducted by an Internet company (Facebook), we exhibit substantial efficiency gains over competing methods.

C1428: **A test for the portion of bivariate dependence in multivariate tail risk**
Presenter: **Carsten Bormann**, University Hannover, Germany
Co-authors: Mélanie Schienle, Julia Schaumburg

In practice, multivariate dependencies of extreme risks are often only assessed in a pairwise way. We propose a novel test to detect when bivariate simplifications produce misleading results. This occurs when a significant portion of the multivariate dependence structure in the tails is of higher dimension than two. Our test statistic is based on a decomposition of the stable tail dependence function, which is standard in extreme value theory for describing multivariate tail dependence. The asymptotic properties of the test are provided and a bootstrap based finite sample version of the test is suggested. An extensive simulation shows the satisfactory performance of the test. Two financial applications involving government bonds and commodity prices underline the test's practical usefulness.

C1641: **Study on the choice of regression quantile threshold in a POT model**
Presenter: **Martin Schindler**, Technical University of Liberec, Czech Republic

The peak over threshold (POT) method with a non-stationary threshold is adopted to estimate high quantiles. We use a regression quantile as the time-dependent threshold. We assume that a linear trend is present in the data and so we use a linear regression quantile as the threshold. 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 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.

C1333: **Comparison of techniques for extreme values using financial data**
Presenter: **Isabel Serra**, Universitat Autonoma de Barcelona, Spain
Co-authors: Joan del Castillo, Maria Padilla

The extreme value theory 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 is to compare the extreme value index using parametric, semi-parametric and non-parametric approach. The more classical parametrical approaches are generalized extreme value for maxima and generalized Pareto distribution for exceedances. Some semi-parametric models based on bias reduction techniques for heavy tails trough 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. The focus is on value-at-risk for log-returns arising in modeling extremes of four datasets in the field of finance, widely documented and studied that can be considered with heavy-tail. Applying extreme value statistics in finance requires accurate estimators on extreme value indices that can be around zero. New parametric models can still be of high interest for the analysis of extreme events, if associated with appropriate statistical inference methodologies, for instance, the full-tails gamma distribution.

C1644: **Estimation of the right endpoint of a light tailed distribution**
Presenter: **Claudia Neves**, CEAUL and University of Aveiro, Portugal
Co-authors: Isabel Fraga Alves

When modeling extreme events generated by a light tailed distribution function, estimation of the supposedly finite right endpoint becomes of utter importance. In this talk we shall present an estimator for the right endpoint of a distribution function belonging to some extreme domain of attraction which only uses the largest observations of the sample and does not require the estimation of the (non-positive) extreme value index. Large sample properties, i.e. consistency and asymptotic normality, are addressed under suitable but not too restrictive conditions. A comparative simulations study is presented on the basis of several distributions of independent interest.

CS49 Room 13 CONTRIBUTIONS IN SURVIVAL AND RELIABILITY Chair: Polychronis Economou

C1303: **Joint modeling of laboratory and field data with application to warranty prediction**
Presenter: **Sheng-T Tseng**, National Tsing-Hua University, Taiwan

Warranty policy is a useful tool for manufacturers to attract customers and to compete with other companies. Therefore, achieving good prediction for the product's field return rate during the warranty period is an important task not only for better understanding of product reliability but also for successful warranty management. In the literature, studies use laboratory reliability to calibrate the field reliability for predicting return rate during the warranty period, focusing on a single product with continuous measurements collected from the laboratory test. Our study, however, is motivated by the need to predict the field return rate for the case of multiple products with discrete-type laboratory data. A hierarchical failure model is proposed to incorporate all failure information from multiple products, under which the empirical Bayes inference is applied for parameter estimation. Because of using the information from similar products, the proposed method is generally more efficient in characterizing the laboratory reliability for each individual product, especially under the scenarios of few or even no testing failures. Furthermore, demonstrated by a real case study, this empirical Bayes approach provides better connections between laboratory and field reliability, leading to an improvement on prediction for the field return rates.

C1339: **Simplicial depth for growth models in construction engineering**
Presenter: **Christoph Kustosz**, TU Dortmund University, Germany
Co-authors: Christine Mueller

Due to specific properties of concrete, AR processes with non-standard error distributions can be considered to describe crack growth processes. Based on models from engineering we introduce a stochastic version of the Paris-Erdogan equation for crack growth and propose test statistics based on simplicial depth to account for specific physical properties. This leads to the analysis of robust estimators for explosive AR(1) processes with and without intercept. The limit distributions of these test statistics are derived. While the result for the model without intercept is a χ^2 distribution which can be derived by standard arguments for U-Statistics, the limit in the two parameter case is an integrated bivariate Gaussian process. Tests for $H_0 : \theta_1 = \theta_1^*$ and $H_0 : (\theta_0, \theta_1) = (\theta_0^*, \theta_1^*)$ are proposed and compared with tests based on simplified depth statistics and standard tests for AR processes. It is illustrated that for crack growth in prestressed concrete simplicial depth gives a considerable alternative to the existing methods. Since calculation of simplicial depth is computationally costly we also discuss the efficient implementation. Further the simulation of the non-Markov Gaussian process to calculate critical values for the two parameter test is discussed.

C1496: **Using mixture cure models with unobserved heterogeneity for the analysis of credit loan data.**
Presenter: **Lore Dirick**, KU Leuven, Belgium
Co-authors: Gerda Claeskens, Andrey Vasnev, Bart Baesens

Due to more strict regulations as a result of the Basel accords, survival analysis is becoming more popular in the field of credit risk analysis due to a large number of censored cases. As an extension to survival models, multiple event mixture cure models are used in order to model several event types for a credit loan (default, early repayment and maturity) jointly in one model. In our research, the multiple event mixture cure model is extended by allowing for unobserved heterogeneity within the event groups. This way, different parameter estimates are possible for different subject groups. A hierarchical EM-algorithm is used to model the (higher level) event types on one hand, and the unobserved heterogeneity on the other, resulting in parameter estimates through the maximization of the expected complete-data log likelihood. We perform a simulation study in which it is shown that allowing for heterogeneity can result in improved prediction accuracy compared to the multiple event mixture cure model without unobserved heterogeneity. Additionally, using a real life credit data example, we illustrate that there is indeed heterogeneity present in the group of subjects that repay their loans early.

C1359: **Adaptive warranty prediction for highly reliable products**
Presenter: **Nan-Jung Hsu**, National Tsing Hua University, Taiwan
Co-authors: Sheng-Tsaing Tseng, Ming-Wei Chen

The aim is warranty prediction for the field return rate. A reliability database is considered for multiple products of similar type in the sense that the products were developed in a series but equipped with similar functionality and manufactured under slight modifications. In order to have a timely prediction, lab reliability tests have been widely used in assessing the field performance before the product introduced to the market. But, due to high reliability associated with the modern electronic devices, the failure data in the lab tests are typically insufficient for each individual product, resulting in less accurate prediction for the field return rate. To overcome this issue, a hierarchical reliability model is suggested to efficiently integrate the information from multiple devices of similar type in the historical database. Under the Bayesian framework, the warranty prediction for a new product can be inferred adaptively along with the data collection process. The proposed methodology is applied to a case study in the information and communication technology industry for illustration. Bayesian prediction is demonstrated to be very effective. In particular, the prediction error rate based on this updating prediction scheme is significantly improved as more field data are collected, and achieves better than 20% after 3 months on field.

C1305: **Nonparametric hypothesis testing for clustered survival model**
Presenter: **John Eustaquio**, University of the Philppines - Diliman, Philippines
Co-authors: Erniel Barrios

A nonparametric test procedure based on the bootstrap in testing for the presence of clustering in survival data is developed. Assuming a model that incorporates the clustering effect into the Cox Proportional Hazards model, simulation studies indicate that the procedure is correctly-sized and powerful in a reasonably wide range of scenarios. The test procedure for the presence of clustering over time is also robust to model misspecification. With large number of clusters, the test is powerful even if the data is highly heterogenous and/or there is misspecification error.

CS73 Room 3 ROBUST METHODS Chair: Anthony C. Atkinson

C1335: **Robust singular value decomposition with application to statistical genetics**
Presenter: **Paulo Rodrigues**, Nova University of Lisbon, Portugal
Co-authors: Andreia Monteiro, Vanda Lourenco

The distribution of continuous variables is usually not normal, often showing heavy tails. Therefore, in such scenarios, the classical approach whose likelihood-based inference leans on the normality assumption may be inappropriate, having low statistical efficiency. Robust statistical methods are designed to accommodate for certain data deficiencies, allowing for reliable results under various conditions. They ought to be resistant to influent factors as outlying observations, non-normality and other model misspecifications. Moreover, if the model verifies the classical assumptions, robust methods provide results close to the classical ones. 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. In particular, interest resides in the development of a robust version of the AMMI model and the comparison between its performance and the performance of the classic AMMI model. This is achieved through Monte Carlo simulations where various contamination schemes are considered.

C1368: **Robust estimation of precision matrices under cellwise contamination**
Presenter: **Garth Tarr**, Australian National University, Australia
Co-authors: Samuel Mueller, Neville Weber

Standard robust procedures assume that less than half the observation rows of a data matrix are contaminated, which may not be a realistic assumption when the number of variables is large. The problem of estimating covariance and precision matrices under cellwise contamination is considered. In particular, robust pairwise covariance matrices are used as inputs to various regularisation routines, such as the graphical lasso, QUIC and CLIME. To ensure the input covariance matrix is positive semidefinite, a method that transforms a symmetric matrix of pairwise covariances to the nearest covariance matrix is used. The result is a potentially sparse precision matrix that is resilient to moderate levels of cellwise contamination and scales well to higher dimensions as it is not based on subsampling. A comparison between the pairwise approach and other standard robust techniques, such as the MCD, is made in terms of entropy loss, various matrix norms and Gaussian graphical discovery rates.

C1620: **Robust moment selection in GMM**
Presenter: **Carlos de Porres**, University of Geneva, Switzerland
Co-authors: Elvezio Ronchetti

Moment selection is an important issue in a generalized method of moments (GMM) setting. There exist several moment selection criteria (MSC), including MSC-BIC, MSC-HQIC, downward testing and upward testing, which provide consistent correct moment selection. However, the choice of the set of moments that maximises the number of orthogonality conditions that are asymptotically zero can be drastically affected by small departures from the model specification. This lack of robustness is due to unbounded orthogonality conditions for consistent set of moments and can lead to wrong moment selection when using standard GMM techniques. In this paper, we study the local robustness conditions for set of moments that give rise to both inconsistent and consistent estimates of the structural parameters. In addition, we propose a robust MSC that will have a stable behaviour under small deviations from the reference model. We illustrate the problem by providing explicit analytical results on a simple example, and we compare classical and robust procedures through Monte-Carlo simulations as well as in an empirical illustration.

C1621: **Comparison of robust detection techniques for local outliers in multivariate spatial data**
Presenter: **Marie Ernst**, University of Liege, Belgium
Co-authors: Gentiane Haesbroeck

Spatial data are characterized by statistical units, with known geographical positions, on which non spatial attributes are measured. Spatial data may contain two types of atypical observations: global and/or local outliers. The attribute values of a global outlier are outlying with respect to the values taken by the majority of the data points while the attribute values of a local outlier are extreme when compared to those of its neighbors. Usual outlier detection techniques may be used to find global outliers as the geographical positions of the data is not taken into account in this specific search. The detection of local outliers is more complex, especially when there are more than one non spatial attributes. This talk focuses on local detection with two main objectives. First, we will shortly review some of the local detection techniques that seem to perform well in practice. Among these, one can find robust "Mahalanobis-type" detection techniques and a wheighted PCA approach. We suggest an adaptation to one of these to further develop its local characteristic. Then, examples and simulations, based on linear model of co-regionalisation with Matern models, are reported and discussed in order to compare in an objective way the different detection techniques.

C1622: **Exact computation of the halfspace depth in arbitrary dimension**
Presenter: **Rainer Dyckerhoff**, University of Cologne, Germany
Co-authors: Pavlo Mozharovskyi

We present two algorithms for exact computation of the halfspace depth of a point z w.r.t. a data cloud $x_1,\dots,x_n$ in $\mathbb{R}^d$. The first algorithm projects the d-variate data into different $d-1$-dimensional subspaces and computes the halfspace depth w.r.t the projected data. This is done recursively and the recursion is stopped, when dimension $d=2$ is reached, in which case an existing algorithm for computing the bivariate halfspace depth is used. This algorithm has a time complexity of $O(n^{d-1}\log n)$. The second algorithm enumerates all possible hyperplanes that contain z and $d-1$ data points. For each hyperplane the data are projected in the direction of the normal vector of this hyperplane. The halfspace depth is then computed from these univariate projections. This algorithm obtains a time complexity of $O(n^{d-1}(d^3+nd))$. As our simulations show, both proposed algorithms prove to be very efficient. Furthermore both algorithms can also be applied to data in non-general position and even to data with ties.

CS72 Room 4 REGRESSION MODELS III Chair: Rosalba Radice

C1554: **To scale and how to scale: Interpreting the edges of a dynamic network**
Presenter: **Kirsten Bulteel**, KU Leuven, Belgium
Co-authors: Francis Tuerlinckx, Eva Ceulemans

Recently, a network perspective to psychopathology has been proposed, based on vector autoregressive (VAR) modeling of time series data. Contrary to the disease model that is operationalized in traditional psychometric analyses, the network approach focuses on the dynamic interplay between symptoms. To examine their vicious causal connections and formulate suggestions for treatment, one can inspect the VAR coefficients, which constitute the edges of the network. But what do these edges exactly mean? Which statements can or cannot be made? For instance, can the strength of the edges be directly compared? We will discuss several ways of scaling the VAR coefficients and list their pros and cons. It will be argued that different standardizations provide different angles on the same process. Moreover, we will discuss how the network representation can be enriched to facilitate a more nuanced interpretation of the edges.

C1580: **A further inspection of the role of the alpha parameter in principal covariates regression**
Presenter: **Marlies Vervloet**, KU Leuven, Belgium
Co-authors: Wim Van den Noortgate, Eva Ceulemans

Principal covariates regression is a method that combines dimension reduction with regression, in that the predictors ($\mathbf{X}$) are reduced to a few components, on which the criteria ($\mathbf{Y}$) are regressed. The extent to which both aspects are emphasized can be manipulated through a weighting parameter α, ranging between 0 (corresponding with reduced-rank regression) and 1 (corresponding with principal components regression). However, how the value of α can be optimally tuned, is not so obvious as well as how the number of components impacts the optimal α. Recently, we integrated the scattered findings on the impact of α and conducted a simulation study which verified the resulting hypothesis that the effect of α is most pronounced when the underlying components differ strongly in strength (i.e., explained variance in $\mathbf{X}$) and relevance (i.e., explained variance in $\mathbf{Y}$). Additionally, we proposed a couple of model selection techniques that combine the selection of α and the number of components. In the present study, we will evaluate the performance of these techniques, especially in conditions where model selection is challenging (e.g., presence of components that have a zero regression weight, multiple criteria ...). Specifically, we will study the recovery of the criteria and the underlying components.

C1385: **Linear regression models using L1, L2 and L infinity norms**
Presenter: **Pranesh Kumar**, University of Northern British Columbia, Canada
Co-authors: Faramarz Kashanchi

In modelling and forecasting applications, the L_2-norm based linear regressions which are known as the least-squares estimation (LSE) models, are often employed and 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, for instance, in non-Gaussian situations when errors follow distributions having fat tails and error terms possess a finite variance. We have investigated the L_1, L_2 and L_∞-norm estimation based linear models and noted that the LSE based models do not always perform the best. Therefore, for estimating the parameters of linear regression models, we consider to apply L_1, L_2 and L_∞ - norm based

linear regressions and use residual analysis based criteria to choose best fitted model. We discuss results on these L_p -norm based estimations in regression modelling by describing analysis of some real data sets which vary in size from small to large and follow different probability distributions.

C1623: **Partial least-squares logistic regression using F-measure**
Presenter: **Jun Tsuchida**, Doshisha University, Japan
Co-authors: Hiroshi Yadohisa

Applying logistic regression for binary classification to high-dimensional data to problems such as multicollinearity and difficulty with interpreting result. Additionally, if classes are heavily unbalanced, logistic regression assumes that all observations have the same class label. In this case, it is better to use the F-measure to evaluate the logistic regression model. The F-measure combines recall and precision into a global measure of utility for the small label. To address these problems, we propose partial least-squares logistic regression maximizing the F-measure using a modified partial least-squares generalized linear model. For the multicollinearity problem, independent variables are dimensionally reduced by partial leas-squares regression. For the second problem, the smaller labels of unbalanced classes are weighted by maximizing the F-measure. We apply the proposed method, partial least-squares logistic regression and logistic regression to wine quality data. Results show that our method gives the best F-measure among the three methods. Moreover, the dimensional reduction results are different between the proposed method and partial least-squares logistic regression.

C1481: **Combining several types of single-case experimental designs using three-level meta-analytic models**
Presenter: **Mariola Moeyaert**, Katholieke Universiteit Leuven, Belgium
Co-authors: Wim Van den Noortgate

As the number of published single-case studies is increasing at an astonishing rate during the last decade, there is a need to optimize the statistical techniques to quantify the research findings in an objective way. A single-case experimental study is " . . . a designed experiment in which one case (i.e., a unit which can be one subject or a small group of subjects) is observed repeatedly during a certain period under different levels (*treatments*) of at least one independent variable." When using data from multiple single-cases, a three-level structure becomes visible: measurement occasions are nested within subjects and subjects in turn are nested within studies. The raw single-case data can be synthesized across subjects and across studies using a multilevel model which is an extension of the regression approach. The purpose of the study is to further extend the multilevel meta-analysis of single-case data by including complex single-case designs such as alternating treatment designs and ABAB reversal designs in addition to the multiple-baseline across participants design (which are the most popular single-case designs). We suggest two univariate models and one multivariate multilevel model and illustrate the proposed methods using a published meta-analysis of single-cases including the three types of single-case designs.

CS75 Room 5 MULTIVARIATE STATISTICS III Chair: Simon Wilson

C1597: **Functional multiple-set canonical correlation analysis for square integrable stochastic processes**
Presenter: **Michio Yamamoto**, Kyoto University, Japan

Functional multiple-set canonical correlation analysis (FMCCA) is the extension of classical multiple-set canonical correlation analysis from a set of vectors to the case where a data sample consists of a set of curves. In this work, we propose FMCCA for a set of square integrable stochastic processes. A question concerning the existence of canonical variates in the proposed model is addressed, and sufficient conditions for the existence of canonical variates are discussed. We also discuss some other properties of the proposed FMCCA model.

C1535: **Bartlett adjustment of deviance statistic for three types of binary response models**
Presenter: **Nobuhiro Taneichi**, Kagoshima University, Japan
Co-authors: Yuri Sekiya, Jun Toyama

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 our presentation, using the continuous term of asymptotic expansion for deviance under the null hypothesis that each model is correct, we obtain a Bartlett-type 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.

C1520: **A quadratic discriminant analysis biplot**
Presenter: **Sugnet Lubbe**, University of Cape Town, South Africa

Principal component analysis (PCA) biplots have been proved very useful in the graphical representation of multivariate data. The PCA biplot is the most common form of biplot, but by no means the only variant. Specifically, linear discriminant analysis can be optimally represented in a Canonical Variate Analysis (CVA) biplot. When quadratic discriminant analysis (QDA) is applied, differing within class covariance matrices is assumed. The canonical transformation in CVA is based on estimating a single pooled within class covariance matrix. When the within class covariances differ, a single canonical transformation cannot be performed to graphically represent the QDA process. In this paper an alternative set of transformations is provided which can lead to an exact representation in two dimensions of the QDA for two groups and an approximate representation of more than two groups. Together with the transformed samples or objects, information on the variables is added, making the graphical representation a true biplot.

C1541: **An association rule miner for unbalanced data based on artificial bee colony optimization**
Presenter: **Emmanuel Rousseaux**, University of Geneva, Switzerland
Co-authors: Gilbert Ritschard

Association rule plays a major role for mining relevant associations within large data. Classical algorithms, as for example the apriori algorithm, require specifying a minimum support to find frequent itemsets. In the case of significantly unbalanced class distributions, this may lead to miss interesting rules about minority classes. On the other side, by lowering the minimum support we would get too many and uninteresting rules. We introduce a new approach based on a binary artificial bee colony optimization algorithm for mining rules involving low support classes. As this association rule miner is formulated as a combinatorial global optimization problem, it does not require specifying a minimum support. Discovered rules are pruned by a chi-squared test and the quality of a rule is assessed by its lift. First experiments have shown that the proposed algorithm is able discover relevant patterns on unbalanced data. Furthermore, our approach does not induce much more time complexity. The algorithm is developed in R and is available on demand.

CS78 Room 19 GRAPHICAL TOOLS AND VISUALIZATION Chair: Alicia Nieto-Reyes

C1326: **A contribution to the visualisation of three-way arrays**
Presenter: **Casper Albers**, University of Groningen, Netherlands

Co-authors: John Gower

Visualisations of two-way arrays are well-understood. In this presentation, some methods for the visualisation (low-rank approximations of) three-way arrays in two or three dimensions are presented. Two-dimensional visualisation, based on the theory of biplots, is possible when working with rank-two (approximations to) arrays. Furthermore, threeway interactions can nicely be visualised in three dimensions when one works with arrays of maximum rank 3. The value of the interaction is then proportional to the volume of a single tetrahedron (rank 1 or 2) or the sum of three tetrahedra (rank 3).

C1355: **A graphical user interface platform of the stepwise response refinement screener for screening experiments**

Presenter: **Frederick Kin Hing Phoa**, Academia Sinica, Taiwan

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 panal.

C1577: **Q-Q plots with confidence**

Presenter: **Wanpen Chantarangsi**, University of Southampton, United Kingdom

Co-authors: Wei Liu, Frank Bretz, Anthony Hayter, Seksan Kiatsupaibul

The importance of normal distribution is undeniable since it is an underlying assumption of many statistical tools. This is the reason why checking the assumption of normality is required prior to applying the normal model to data in hand. There are two types of procedures in assessing whether a population has a normal distribution based on a random sample: graphical methods (e.g., Q-Q plots) and non-graphical methods (e.g., Anderson-Darling test and Shapiro-Wilk test). The normal quantile-quantile plot (Q-Q plot), also called normal probability plot, is the most commonly used diagnostic tool for assessing whether a random sample is drawn from a normally distributed population. In this study we provide, on a normal probability plot, exact simultaneous intervals into which the points fall with probability $1-\alpha$ if the sample is taken from a population with normal distribution. These simultaneous intervals provide therefore an objective way to judge whether the plotted points fall close to a straight line. Several different sets of simultaneous intervals associated with Kolmogorov-Smirnov test (D test), Michael test (Dm test), Dn test, Dbeta test and Dnew test are investigated, including the power comparison among these graphical methods and with the non-graphical Anderson-Darling test and Shapiro-Wilk test.

C1686: **Visualization for reason-specified suicide data in Japan**

Presenter: **Takafumi Kubota**, Tama University, Japan

Co-authors: Hiroe Tsubaki

The study visualizes reason-specified suicide data in Japan. The available data on reason used in this study are categorized into eight reasons; family, health, economy, workplace, male-female, school and other relationships, and unknown. The proposed method can give users of the data some perspective of temporal trends of the reasons of suicide in Japan and comparisons among the prefectures in Japan by detecting some areal characteristics. At first, cumulative bar chart was used to describe period effects to cumulative frequency of the reasons, while line chart were used to compare with the reasons of suicide by time series data. Then, web application was developed by the R packages “shiny” to perform interactive handling. The graphs of several prefectures’ reason-specified suicide data will be shown to find out areal characteristics and the interactive handling web application will be demonstrated in presentation.

C1712: **Advantages of molding results in data analysis**

Presenter: **Yoshitomo Akimoto**, Chuo University, Japan

Co-authors: Takemi Yanagimoto, Toshinari Kamakura

As innovative improvements of visualizing results in data analysis, we would like to use the 3D printer apart from computer graphics. Data visualization with computer graphics is known as a method of understanding the property of data. To pursue the better understanding of the data, we proposed the method by 3D molding the result of data analysis. It provides us with the deep view point the property of the result. We will investigate usefulness of the 3D molding of Bayesian credibility in this article. The new technique of 3D molding may help us much greater with searching for small p-values in the framework of Bayesian credibility. Although it is difficult to get small values from non-convex function, the larger credibility values are to recognize comparatively easy with 3D molding structures. In this study, we also devised the utility programs that enable to communicate the R language and the 3D printer.

Friday 22.08.2014 11:15 - 12:45 Parallel Session L

IS03 Room 3 STATISTICS AND OPTIMIZATION IN FINANCE Chair: Sandra Paterlini

C1387: **Vine copulas: statistical inference and model selection**
Presenter: **Claudia Czado**, Technische Universität München, Germany
Standard multivariate copula classes such as the elliptical or Archimedean one are restricted in their tail and symmetry behavior, since they are closed under margins. These restrictions might not be satisfied in real data applications. In contrast vine copula models are very flexible. They are constructed using only bivariate copula building blocks called pair copulas. The full specification of a vine model requires the choice of vine tree structure, copula families for each pair copula term and their corresponding parameters. This class of copulas will be introduced and their statistical inference including model selection will be discussed. Approaches will be illustrated using financial data.

C1447: **Estimating time series models with heuristic methods: the case of economic parity conditions**
Presenter: **Dietmar Maringer**, University of Basel, Switzerland
Co-authors: Sebastian Deininger
Time series models are a common approach in economic and econometric analysis. Special cases are vector error correction (VEC) models where several economic variables are assumed to depend on their own and each other's recent developments. While they facilitate economically sound modeling, their actual application is often hampered by technical difficulties: Finding the optimal parameter values is usually based on maximizing some likelihood function or "information criterion" for which no closed-form solution exists. Even more importantly, the number of parameters increases quickly when allowing for more lags, i.e., including past observations — which is highly desirable, e.g., when seasonalities, delayed reactions, or long memory need to be catered for. In this case, it is desirable to keep the model still as parsimonious as possible to avoid over-fitting. Ideally, one can "cherry-pick" the parameters which one wants and doesn't want to include; this, however, makes parameter estimation even harder as it adds challenging combinatorial problems. In this paper, we investigate how differential evolution (DE), a nature-inspired search heuristic, can help to tackle the parameter selection and estimation problem simultaneously, which, in traditional approaches to econometric model selection, is not possible. This approach is applied to the case of parity conditions using data for the US, the Euro-Area and for Switzerland to investigate the concepts of uncovered interest rate parity, the expectation hypothesis of the term structure and the purchasing power parity. Referring to BIC as the information criterion, the results indicate that for the considered currencies and economic regions, only some, but not all of these parities hold. While a constant is rejected in any considered model, it can be observed that parameter values remain more or less the same across parity concepts. It is also found that different approaches can lead to conflicting conclusions, which emphasizes the importance of careful economic modeling and reliable methods.

C1469: **Penalized least squares for optimal sparse portfolio selection**
Presenter: **Sandra Paterlini**, European Business School Germany, Germany
Co-authors: Bjoern Fastrich, Peter Winker
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 allow us 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 then consider several alternative non-convex penalties, that allow us to improve on the 1-norm penalization approach. Empirical results on large 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.

OS54 Room 5 SYMBOLIC/ALGEBRAIC METHODS IN COMPUTATIONAL STATISTICS II Chair: Elvira di Nardo

C1677: **Everything you always wanted to know about moments (but were afraid to derive)**
Presenter: **Colin Rose**, Theoretical Research Institute, Australia
Co-authors: Murray Smith
Moment (and moments of moments) problems are usually simple to pose as a question, but can be highly intractable to solve, even if the solution has a neat "elegant" closed-form. Such problems are ideally suited to solving with a symbolic computer algebra system. We illustrate using the latest moment algorithms designed for the mathStatica add-on to the Mathematica computer algebra system. This provides a general automated approach to solving moment problems such as: (i) moments of random variables, for arbitrary (known) distributions, (ii) converting between any of: raw, central, cumulant, factorial moment, (iii) new symbolic moment operators e.g. find $Var(XYZ)$ in terms of the moments of X, Y and Z (unknown pdf), or find $Cov(XYZ, XY+YZ)$, (iv) moments of any rational algebraic symmetric function of random variables, such as: Cov $\left(\sum_{i=1}^{n}(X_i-\bar{X})^2, \sum_{i=1}^{n}X_i^2\right)$, (v) unbiased estimators of: raw moments, of central moments (h-statistics), and of cumulants (k-statistics), (vi) unbiased estimators of products of: raw moments (polyraws), central moments (polyaches), cumulants (polykays) and (vii) new conversion functions between power sums and symmetric functions: including to/from unitary/elementary symmetric functions, augmented symmetrics, monomial symmetrics etc ...

C1436: **Goodness of fit for log-linear network models: dynamic Markov bases using hypergraphs**
Presenter: **Sonja Petrovic**, Illinois Institute of Technology, United States
Co-authors: Elizabeth Gross, Despina Stasi
Social networks and other large sparse data sets pose significant challenges for statistical inference, as many standard statistical methods for testing model fit are not applicable in such settings. Algebraic statistics offers a theoretically justified approach to goodness-of-fit testing that relies on the theory of Markov bases and is intimately connected with the geometry of the model as described by its fibers. Most current practices require the computation of the entire basis, which is infeasible in many practical settings. We present a dynamic approach to explore the fiber of a model, which bypasses this issue, and is based on the combinatorics of hypergraphs arising from the toric algebra structure of log-linear models. We demonstrate the approach on the Holland-Leinhardt p1 model for random directed graphs that allows for reciprocated edges.

C1378: **Algebraic geometry meets causal inference**
Presenter: **Caroline Uhler**, IST Austria, Austria
Many algorithms for inferring causality are based on partial correlation testing. Partial correlations define hypersurfaces in the parameter space of a directed Gaussian graphical model. The volumes obtained by bounding partial correlations play an important role for

the performance of causal inference algorithms. By computing these volumes it is shown that the so-called “faithfulness assumption”, one of the main constraints of many causal inference algorithms, is in fact extremely restrictive, implying fundamental limitations for these algorithms. Thus an alternative method is proposed that involves finding the permutation of the variables that yields the sparsest DAG. In the Gaussian setting, the sparsest permutation (SP) algorithm boils down to determining the permutation with sparsest Cholesky decomposition of the inverse covariance matrix. It is proven that the constraints required for this SP algorithm are strictly weaker than faithfulness and are necessary for any causal inference algorithm based on conditional independence testing.

C1353: **Taking advantage of a symbolic representation of non-central Wishart distributions**
Presenter: **Elvira Di Nardo**, Basilicata, Italy

The computation of joint moments $E\left\{\mathrm{Tr}[W(n)H_1]^{i_1}\cdots\mathrm{Tr}[W(n)H_m]^{i_m}\right\}$, with $H_1,\ldots,H_m$ complex matrices, Tr the trace and $W(n)$ non-central Wishart distribution has different applications: in the study of asymptotic properties of the sample covariance matrix or in quantify the performance of multidimensional signal processing algorithms. Indeed for $H_1,\ldots,H_m$ sparse matrices, joint moments of entries of $W(n)$ can be recovered. This computation is a very general task and object of in-depth analysis, due to the complexity of the cumbersome involved formulae. Usually symbolic languages are required involving derivatives of vector of matrix functions. The aim is to introduce a different symbolic method allowing us to take advantage of the cyclic properties of traces. Thanks to this device, $\mathrm{Tr}[W(n)]$ is represented as convolution of its central component and a matrix of formal variables, whose entries are uncorrelated with those of the central component. Due to the cyclic properties of traces, the notion of necklace is fruitfully employed in the computation allowing us to set up an efficient symbolic procedure. The algorithm has been implemented in `Maple 12`. Comparisons with other techniques proposed for computing moments of $W(n)$ are given.

OS46 Room 4 MULTI-SET AND MULTI-WAY MODELS II Chair: Eva Ceulemans

C1421: **Structure-revealing unsupervised data fusion based on coupled matrix and tensor factorizations**
Presenter: **Evrim Acar**, University of Copenhagen, Denmark

Analysis of data from multiple sources has the potential to enhance knowledge discovery by capturing underlying structures, which are, otherwise, difficult to extract. Fusing data from multiple sources has already proved useful in many applications in social network analysis, signal processing and bioinformatics. However, data fusion remains a challenging task in need of data mining tools that can jointly analyze multi-relational and heterogeneous data sources. In order to address this challenge, data fusion has been formulated as a coupled matrix and tensor factorization (CMTF) problem. While the traditional CMTF formulation models only shared factors, we introduce a structure-revealing data fusion model that has the potential to automatically reveal shared and unshared components. As an algorithmic approach, coupled factorization problems have commonly been solved using alternating methods and, recently, unconstrained all-at-once optimization algorithms. Unlike previous studies, in order to have a flexible modeling framework, we use a general-purpose optimization solver that solves for all factor matrices simultaneously and is also capable of handling linear/nonlinear constraints with a nonlinear objective function. We formulate CMTF as a constrained optimization problem and develop accurate models more robust to overfactoring. The effectiveness of the proposed modeling/algorithmic framework is demonstrated on simulated and real data.

C1434: **Multiset analysis based on coupled tensor decompositions**
Presenter: **Mikael Sorensen**, KU Leuven, Belgium
Co-authors: Lieven De Lathauwer

It is now well-known that several problems in signal processing and statistics are inherently multilinear when the full diversity is exploited. In signal processing the canonical polyadic decomposition (CPD) and block term decomposition (BTD) models have proven to be useful for (single set) analysis. In recent years there has been a great interest in multiset analysis, e.g., multimodal data acquired by different types of equipment. To accommodate this demand we have recently extended the CPD/BTD modeling framework to coupled models, leading to uniqueness conditions and algorithms for tensor-based multiset data analysis. If time permits, we will also discuss connections to missing data problems and extensions to multiset data problems with a two-dimensional coupling.

C1439: **Clusterwise parafac with varying complexities in latent variable structure**
Presenter: **Tom Frans Wilderjans**, KU Leuven, Belgium
Co-authors: Eva Ceulemans

Recently, Wilderjans & Ceulemans (2013) proposed the clusterwise Parafac model to disclose qualitative differences in the component structure underlying three-way three-mode data (e.g., the anxiety-related reactions of persons in different situations). In this model, the elements of one of the three modes (e.g., the persons) are clustered and, simultaneously, a Parafac model is fitted to the data within each cluster. As such, elements assigned to different clusters are described by a different underlying component structure. Although clusterwise Parafac accounts for qualitative differences in the component structure by allowing components to be different across clusters, the complexity (i.e., number) of the component structure is constrained to be equal across groups, which can be considered as a rather restrictive assumption. Therefore, in this paper, the clusterwise Parafac model is extended to allow the number of components per cluster to vary across clusters. Further, to determine the clustering and the appropriate number of components per cluster, which appears to be a non-trivial task, a model selection tool that is based on CPOPT (Acar, Dunlavy and Kolda, 2011) is proposed. To evaluate the performance of this new Clusterwise Parafac algorithm, an extensive simulation study is conducted. Moreover, the strategy is applied to EEG data.

C1510: **Tensor polyadic decomposition for antenna array processing**
Presenter: **Pierre Comon**, CNRS, France
Co-authors: Souleymen Sahnoun

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.

OS37 Room 6 MIXTURE MODELING OF LONGITUDINAL DATA Chair: Andre Berchtold

C1452: **Finding Likelihood optimum by trials**
Presenter: **Zhivko Taushanov**, University of Lausanne, Switzerland

Optimizing the log-likelihood of a model containing a latent part, is often difficult, especially when re-estimation equations cannot be derived explicitly. The standard EM algorithm needs then to be replaced by a Generalised EM procedure. The main difference with the standard algorithm is the lack of equations ensuring an increase of the log-likelihood after the re-estimation step, so we need a procedure able to simultaneously find new parameters values and check their consistency. We propose such a procedure based on a "trial strategy" applied to the case of the Hidden Mixture Transition Distribution model (HMTD) for continuous variables. Our procedure is a development of an optimisation strategy successfully applied in the past to the discrete version of the MTD model. With many parameters and little prior knowledge about the structure of the solution space, we need to find an optimal approach of re-estimation in order to reach the (possibly local) optimum of the log-likelihood. Do we modify the parameters independently or simultaneously, should we use a fixed re-estimation step or not? These are some of the questions we explore in our work by comparing the speed of convergence and the efficiency of different kinds of parameter re-estimations.

C1431: **Hidden mixture distribution modeling: model selection and computation**
Presenter: **Danilo Bolano**, University of Geneva, Switzerland
Co-authors: André Berchtold

Analysis of longitudinal data presenting heterogeneous behaviors may be difficult to deal with. One possibility is to use the class of Hidden Mixture Transition Distribution models. The observed heterogeneity can then be induced by an underlying hidden process and each factor of the latent process is related to a different component of an observed mixture of Gaussian distributions. Possible refinements include the use of a hidden Markovian process of any order governing the transition between components, MTD approximation of the hidden transition process, time-dependent expectation and standard deviation for each component and covariates included both at the visible and hidden level. Because of its flexibility, a critical step lies in the correct specification of the model. Working in a multilevel framework, we introduce a hierarchical procedure for the selection and computation of an appropriate model regarding the available data. Our approach will be illustrated on two real datasets with an application in gerontology and an analysis of the U.S. Panel Study of Income Dynamics.

C1429: **DCMM for latent group identification with longitudinal data**
Presenter: **Pauline Adamopoulos**, University of Geneva, Switzerland
Co-authors: Andre Berchtold, Gilbert Ritschard

We show how the double chain Markov model (DCMM) can be extended to identify latent groups (clusters) in longitudinal data. This is a stochastic model-based approach that requires neither a (dis)similarity nor a distance measure. A given sequence is assigned to the latent group that would maximize the probability of observing a sequence with such transition patterns. To reflect these various transition patterns, hidden transition matrices may be unconstrained, diagonal, triangular or strict hierarchic. In order to account for the serial correlation of observations within a sequence, we allow the observed process to be non-stationary. As higher-order dependencies increase the number of parameters exponentially, we approximate the model parameters using the Mixture Transition Distribution (MTD) model as a means of parsimony. An R package in progress is extended to allow for the necessary computations in latent group identification, while the capacities of the R package TraMineR are used to visualize the sequences. A package allowing to visualize output using TraMineR graphics is also in progress.

C1419: **March: an R implementation of Markovian models**
Presenter: **Ogier Maitre**, EPFL, Switzerland

An R S4 package called March, which is designed to construct and compare several high order Markovian models on discrete weighted sequences, such as the homogeneous Markov chain, the mixture transition distribution model (MTD) and the double chain Markov model (DCMM), is presented. A mixture transition distribution is implemented into its standard form, as well as its multi-matrix form (MTDg), where a different transition matrix is used for each lag. For the DCMM, this package uses an evolutionary algorithm (EA), in order to avoid being stuck into local optima. EA operators are random initialization, SBX and Gaussian mutation. The log-likelihood is used to evaluate and compare possible solutions. Due to computation constraints and R performance the algorithm can be run and is efficient even on a small population and with a relatively low number of generations. To discover better solutions, the EA performs an optimization step on each child produce during a run, using an EM algorithm variant, namely the Baum-Welch algorithm. Results are presented on several test problems and compared with previously published results.

OS83 Room 20 APPLICATION OF EXTREME VALUE THEORY IN COMPUTING SYSTEMS Chair: Petar Radojkovic

C1727: **Rationale to the application and tailoring of EVT in context**
Presenter: **Code Lo**, INRIA, France

EVT can be regarded as the counterpart of Central Limit Theory: where the latter studies the bulk of the population of a given distribution, EVT studies the tail of it, in other words the extreme deviations from the median of probability distributions. By analysing a sample of observations of a given random variable, EVT determines the probability of extreme events to occur, where "extreme" refers to either end of the range of the value domain of those events. EVT is an important branch of statistics, which has found application in multiple disparate domains, including civil engineering, material testing, finance, and risk management. It provides many powerful theorems and tools, which might be of great interest to the computer science and engineering community. Its application in these fields is, however, currently marginal. EVT will be introduced explaining some basic understanding of its uses in several application domains. The main aspects considered for its application to the computer system domain will be discussed. A description of the input data (population) considered in the application domain in which we have applied EVT will be provided. Some of the used of EVT in the field of application will be summarized showing some of the problems in its application and how they have been solved.

C1726: **EVT in real-time systems: i.i.d properties emanating from the experimentation**
Presenter: **Francisco J Cazorla**, Barcelona Supercomputing Center and IIIA-CSIC, Spain

In the application domain, i.e. safety-critical real time systems (like the ones embedded in cars, airplanes, trains, satellites), current methods to derive the longest execution time of a program (also known as worst case execution time or WCET) focus on deriving the absolute maximum execution time. However, this produces a degree of pessimism where unknown states of the computing system during the execution of the program have to be modelled as having their worst consequences on the timing of the system. However, the consequences of these unknown states can be considered probabilistically, allowing to reason about the WCET probabilistically. In this respect, techniques from Extreme Value Theory (EVT) can be used to construct a worst-case probability distribution. This requires defining worst-case bounds with confidence levels. Confidence levels can be chosen to match the degree of uncertainty present in the rest of the system being analysed. An application of EVT to safety-critical real-time systems, etc will be introduced. The

different uses of EVT to derive upper bounds to the worst-case (longest) execution time of real-time software programs when running on a processor will be discussed. The considered scenario is that in which the properties required to apply EVT on the set of execution time observations collected are achieved as part of the methodology used to collect observations, i.e. the timing of end-to-end runs of the program of interest on the computing system. In this line, the focus will be on standard time-deterministic computing systems.

C1728: **EVT in real-time systems: i.i.d properties emanating from the computing system**
Presenter: **Tullio Vardanega**, University of Padova, Italy

The use of EVT to derive upper bounds to the worst-case (longest) execution time of real-time software programs running on a computing system will be illustrated. The use of EVT places requirements on the analysis procedure and on the nature of execution: what these requirements are and how they can be met using "time randomisation" will be explained. A novel analysis method called Measurement-Based Probabilistic Timing Analysis (MBPTA) will be described. MBPTA, which uses EVT as central building block, takes in input a set of observations made of actual end-to-end runs of the program of interest, on the target computing system. A clear differentiation will be made among the requirements that MBPTA inherits from its use of EVT and MBPTA's own requirements of the specific goal of finding the worst-case execution time of a software program running on a computing system. To conclude, results obtained with MBPTA with real applications from avionics domain will be described.

C1725: **EVT in non-real-time systems: Tackling NP-complete computer-science problems**
Presenter: **Petar Radojkovic**, Barcelona Supercomputing Center, Spain

How random sampling and EVT can be used to solve intractable NP-complete computer science problems will be shown. NP-complete problems in computer science are usually addressed using heuristics-based approaches, designed for a specific problem and metric. Since problems typically have a vast exploration space, it is infeasible to find the optimum using an exhaustive search. Therefore, the room for improvement of a given heuristics-based algorithm is also unknown. It is thus hard to decide whether to invest effort to try to improve an algorithm that may already be close to optimal. We will present a method that predicts the performance of the optimal solution for an intractable problem, i.e. the method that uses EVT to estimate population maximum (or minimum) based on a set of random observations. We will present theoretical background for our analysis and the application of EVT on two problems related to compilation and scheduling of multithreaded applications.

CS20 Room 19 APPLIED STATISTICS AND DATA ANALYSIS IV Chair: Stan Azen

C1361: **Compositional data approach for statistical process control of railway ballast**
Presenter: **Vera Hofer**, University of Graz, Austria

A statistical monitoring of railway ballast is introduced. Due to the particular way of measuring the explanatory variables that are used for predicting various technical properties, the measurements statistical modelling during the monitoring process is based on form compositional data. Compositional data are vectors whose components are the proportions of a total and thus sum to 1. Due to this structure, classic multivariate methods cannot be applied. Thus, modelling is carried out using a compositional data approach. Statistical monitoring of samples from daily production involves not only the estimation of technical properties such as resistance to wear and fragmentation, but also surveying the dynamics of explanatory variables by means of CUSUM charts.

C1364: **Nonparametric profile monitoring and fault diagnosis via functional data analysis**
Presenter: **Jyh-Jen Horng Shiau**, National Chiao Tung University, Taiwan
Co-authors: Yuchun Wu

In many practical situations, the quality of a process or product is characterized by a relationship (or profile) between a response variable and one or more independent variables instead of by the distribution of a univariate or multivariate quality characteristic. Most research work in the literature assumed fixed effects and/or parametric regression models to model profiles. The main purpose is to develop monitoring and diagnosis tools for profiles of more flexible shapes and under more general and practical situations. Under a random effects model and adopting nonparametric regression approach, a new profile monitoring scheme is proposed. In addition, assuming sample profiles for certain frequently-occurred out-of-control conditions (faults) are available, a fault diagnosis procedure is developed for practitioners to use. When a profile signals out-of-control, the procedure will determine which of the known faults contributes to it or identify it as coming from a novel fault.

C1379: **Nonparametric estimation of menarcheal age distribution based on recall data**
Presenter: **Sedigheh Mirzaei**, Indian Statistical Institute, India
Co-authors: Debasis Sengupta

Menarche, the onset of menses, plays major role in a womanś life. The most common nonparametric approach of estimating age at menarche is to use dichotomous ("Status quo") data on menarcheal status on the day of interview, and to use the Turnbull estimator (Turnbull, 1976) for interval censored data. On the other hand recall data on the actual age at menarche contain additional information, and are expected to produce better estimates. In the case of respondents who did not have menarche or could not recall the date of menarche, one can only identify a time interval containing the menarcheal age. However, the nature of censoring involved in gathering retrospective menarchial data is informative, which precludes the use of the Turnbull estimator. A non-parametric maximum likelihood estimator is provided, based on a model that makes use of the special nature of the data at hand. Monte Carlo simulations indicate that the performance of the proposed estimator has less bias than the Turnbull estimator based on incomplete recall data, less variance than the Turnbull estimator based on status quo data, and less mean squared error than both of them. The method is applied to menarcheal data from a recent Anthropometric study on adolescent and young adult females in Kolkata, India.

C1603: **Propensity score matching with clustered data: an application to birth register data**
Presenter: **Massimo Cannas**, University of Cagliari, Italy
Co-authors: Bruno Arpino

The implementation of propensity score matching for clustered data is considered. 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.

CS69 Room 13 COUNT DATA Chair: Eva Cantoni

C1329: **Analysis of discrete dependent variable models with spatial correlation**
Presenter: **Jan Vogler**, University of Cologne, Germany
Co-authors: Roman Liesenfeld, Jean-Francois Richard

An ML estimation for a broad class of parameter-driven models for discrete dependent variables with spatial correlation is considered. Under this class of models, which includes spatial discrete choice models, spatial Tobit models and spatial count data models, the

dependent variable is driven by a latent stochastic state variable which is specified as a linear spatial regression model. The likelihood is a high-dimensional integral whose dimension depends on the sample size. For its evaluation to use efficient importance sampling (EIS) is proposed. The specific spatial EIS implementation developed exploits the sparsity of the precision (or covariance) matrix of the errors in the reduced-form state equation typically encountered in spatial settings, which keeps numerically accurate EIS likelihood evaluation computationally feasible even for large sample sizes. The proposed ML approach based upon spatial EIS is illustrated with estimation of a spatial probit for US presidential voting decisions and spatial count data models (Poisson and Negbin) for firm location choices.

C1463: **Modelling and predicting clustered count data with excess zeros**
Presenter: **Eva Cantoni**, University of Geneva, Switzerland
Co-authors: Joanna Mills Flemming, Alan Welsh

Hurdle models (a.k.a. two-part, zero-altered or separated models) are a very popular choice for modelling count data with excess zeros. However models for these data when in addition, they are clustered, have received less attention. We present a general formulation for mixed effects hurdle models. A novel approach to the introduction of the random effects allows extensions beyond the usual multivariate normality assumption and facilitates inference about dependence between the two parts of the model. We obtain the fixed effects parameter estimates by maximum likelihood and develop empirical best predictors of the random effects and other cluster specific targets. We also address the unsolved issue of computing the mean squared error of these predictions. We pay careful attention to computational aspects, and use, for example, a fast bootstrap procedure. The methods are demonstrated using real data on critically endangered hammerhead sharks and evaluated with a simulation study. Our approach is more generally applicable than the hurdle model and so can be easily extended to GLMM, for example.

C1628: **Maximum empirical likelihood estimation of mixing distributions for overdispersed count data modelling**
Presenter: **William Aeberhard**, University of Geneva, Switzerland
Co-authors: Eva Cantoni, Stephane Heritier

For fitting purposes and consistent inference, the correct specification of the relation between the variance of a response variable and its mean (known as the variance function) is crucial. For a response consisting of counts, typically involving overdisperion, many models already exist and offer a wide variety of variance functions. This variety is appealing, yet the data analyst is still faced with the somewhat arbitrary choice of a specification over another, without many means to check the relevance of such a choice. Most of these variance function specifications are in fact the result of the specification of an unobservable mixing distribution, whose realizations can be seen as random intercepts (one per observation) appearing in the mean of Poisson distributions, themselves giving rise to the observed counts. A well-known case is a gamma underlying mixing distribution yielding negative binomial outcomes, implying a quadratic variance function. We propose here a semi-parametric generalized linear model for count data where the mixing distribution is estimated by maximum empirical likelihood along with the regression coefficients. By doing so, the shape of the variance function is somehow estimated from the data and the data analyst needs not to worry about its specification. Theoretical results as well as simulations under different settings and an application will be presented.

C1633: **Count data regression with excess zeros - a flexible framework using the GLM toolbox**
Presenter: **Christian Kleiber**, Universität Basel, Switzerland
Co-authors: Achim Zeileis

Many data sets encountered in count data modeling exhibit a substantial number of zeros. Often, this number is so large that moving from a Poisson to a negative binomial model cannot fix the problem and the zeros require special treatment. The hurdle model is a two-part model for counting data with excess zeros, comprising a binary response part for zeros vs. non-zeros and a zero-truncated count distribution for the positive counts. We show how not only the binary part but also the count component can be conveniently analyzed within the GLM framework. We present several applications: First, the GLM approach helps to identify computational pitfalls such as the nonexistence of the MLE under certain circumstances. Next, it paves the way for flexible extensions along the lines of additive models. We present an additive version of the negative binomial hurdle model using GAM machinery. Furthermore, regularization via boosting algorithms is available more or less like for classical GLMs. An R package named countreg with the relevant functionality is currently under development, first versions are already available from http://r-forge.r-project.org/projects/countreg/.

Authors Index

next meeting . . .

COMPSTAT 2016

Palacio Calatrava
Oviedo, Spain

August 23–26, 2016