V

Significance Guarantees of Local Descriptive Models

Overview

Assessing the statistical significance of the local (descriptive) models data is required to guarantee the discovery of relevant regions from high-dimensional data, as well as to filter, validate or weight the increasing number of implications in literature derived from the analysis of biomedical and social data. As largely motivated, guaranteeing the statistical significance of the modeled regions of interest is required to minimize the propensity of a given learning function to over/underfit the observed data. This ensures the adequacy of the capacity term of a learning function, and thus its ability to generalize.

Despite the relevance of guaranteeing the statistical significance of local descriptive models, there is not yet a ground truth on how to assess the significance of flexible biclustering models from tabular data neither of cascade and event-set models from structured data. As such, this book proposes principles for the robust assessment of regions with varying criteria of homogeneity (including varying coherency and tolerance to noise) learned from tabular and structured data contexts.

Furthermore, this book integrates the proposed significance views with homogeneity views to affect the learning of local models. To illustrate this requirement, consider the problems associated with two major learning options. A first option (with propensity to overfit data) is to model regions with high homogeneity only. However, optimizing homogeneity levels is of limited use since good levels can appear by chance for small regions. A second option (with propensity to underfit data) is to model large regions (satisfying some homegeneity criteria) to minimize the chance of delivering non-significant biclusters. For this aim, some loose form of coherency is assumed and high levels of noise are tolerated. Understandably, these classes can be seen as distinct poles of the performance axis. The first option neglects the significance component of performance, while the second option prioritizes significance in detriment of homogeneity (coherency and quality).

The need to guarantee the significance of local descriptive models can be decomposed according to seven major requirements:

- robust statistical tests for biclustering, with an efficient assessment of deviation from expectations **[R4.1]**;
- significance assessment of additive, multiplicative, symmetric, order-preserving and plaid models **[R4.2]**;
- significance assessment of biclusters with arbitrary-high levels of noise **[R4.3]**;
- significance assessment of biclusters discovered in real-valued data contexts and of biclusters with continuous ranges of shifting and scaling factors **[R4.4]**;
- significance assessment of cascade models from three-way time series and of arrangements of events from multi-sets of events **[R4.5]**;
- inference of global constraints to enforce bounded guarantees on the significance of regions, and their effective incorporation within the learning process **[R4.6]**;
- combined homogeneity-significance views for adequately assessing the relevance of regions **[R4.7]**.

Chapters 1-3 propose methods to test and ensure the statistical significance of regions in tabular data contexts.

Chapter 1 motivates this need and surveys the current limitations that prevent the assessment of flexible biclustering models. To address these limitations, we propose a robust statistical framework to: 1) assess biclusters using stochastic and frequentist views able to adequately model the impact of data dimensionality; 2) minimize the number of false positives (outputted non-significant biclusters) and false negatives (non-retrieved significant biclusters) by effectively and efficiently testing deviations from expectations; and 3) infer global constraints that guarantee the significance of a given bicluster when they are satisfied.

Chapter 2 provides the first attempt for the statistical assessment of biclusters with flexible coherency and quality by extending the previous contributions towards biclustering models with non-constant coherency assumptions and arbitrary-high levels of noise, and consistently integrating quality and significance views.

Chapter 3 extends the previous contributions towards real-valued data contexts where the coherency of a given bicluster may not be known apriori. Principles from integral calculus are used to enlarge the proposed statistical assessments to biclusters characterized by continuous ajustement factors. Finally, we guarantee the applicability of the proposed contributions for data domains with non-identically distributed features.

To guarantee the consistency of the contents in this book with the standard notation in statistics (instead of machine learning), we revised the notation of two concepts. Given a tabular dataset $\mathbf{A}$ with $\mathbf{X}$ observations and $\mathbf{Y}$ features, and a bicluster $\mathbf{B}=(\mathbf{I}\subseteq\mathbf{X},\mathbf{J}\subseteq\mathbf{Y})$, then $N=|\mathbf{X}|$ is the data size, $M=|\mathbf{Y}|$ is the data dimensionality, $n=|\mathbf{I}|$ is the number of observations in $\mathbf{B}$ and $m=|\mathbf{J}|$ is the number of features in $\mathbf{B}$.

Finally, *Chapter 4* provides principles to test and ensure the statistical significance of regions in structured data contexts. For this aim, we enrich existing statistical views of sequential data to statistically assess cascade models learned from three-way time series and arrangements of events from multi-sets of events.

Index of Requirements and Contributions

Tables 1-4 exhaustively list the proposed contributions throughout this book.

Table 1: Major contributions to assess the statistical significance of biclustering models (*Chapter* V-1).

R4: Guarantee the significance of flexible local models;
R4.1: Robust assessment of the statistical significance of (discrete) biclusters;
- **C4.1a**: Structured view on the contributions/limitations to statistically model biclusters from PM, biclustering and inferential statistics;
- **C4.1b**: Systemic empirical analysis on how bicluster's properties (support, length, pattern, coherency, quality) and data properties (size, dimensionality, regularities) affect statistical significance;

R4.1.1: Robust statistical tests;
- **C4.1.1.1**: Statistical tests to assess significance of biclusters based on bionomial tails estimated from: 1) the observed regularities (using either stochastic or frequentist views), 2) generated data, and 3) hold-out partitions;
- **C4.1.1.2**: Extension of the proposed tests to guarantee their applicability to biclusters with different forms of constant coherency;
- **C4.1.1.3a**: Extension of bionomial calculus to adequately measure the impact of data dimensionality on the significance;
- **C4.1.1.3b**: Removal of efficiency bottlenecks when assessing bicluster's patterns with non-indexed columns;
- **C4.1.1.4a**: Principles to compute sound statistical views when assuming varying forms of dependency (inc. pairwise and overall) based on: 1) joint probability calculus, and 2) generated data based on n-wised dependencies;
- **C4.1.1.4b**: Dynamic programming principles to tackle efficiency bottlenecks when testing biclusters with n-wised dependencies;
- **C4.1.1.5**: Integrative assessment method where statistical decisions (tests, stochastic/frequentist views, dependency form) are dynamically selected from the input data properties or user expectations;

R4.1.2: Non-conservative yet efficient assessment of deviation from expectations;
- **C4.1.2.1**: New multi-hypotheses correction based on a variant of Hochbert procedures able to effectively test deviations and minimize the risk of false negatives of conservative peers (type-II errors) and false positives (type-I errors);
- **C4.1.2.2**: Binary space partitioning algorithm for efficient non-conservative corrections from an arbitrary-high number of hypotheses;

R4.1.3: Guarantee the applicability to tabular data contexts with complex domains (mixtures of nominal, ordinal and numeric features);
R4.2: Robust assessment of additive, multiplicative, symmetric, order-preserving and plaid coherencies;
R4.3: Robust assessment of regions with arbitrary-high levels of noisy and missing elements;
R4.4: Robust assessment in real-valued data contexts;
R4.5: Robust assessment of regions from structured data;
R4.6: Inference of global constraints;
R4.6.1: Expectations on the properties of biclusters assuring their statistical significance of biclusters;
- **C4.6.1.1**: Global statistical tests (founded on the Poisson distribution to characterized deviations on the number of occurring biclusters) to infer minimum number of rows and columns in a bicluster that guarantees statistical significance;
- **C4.6.1.2**: Extensions to guarantee an adequate balance between type-I or type-II errors (address problems of peer contributions);
- **C4.6.1.3**: Extension of the proposed global tests to guarantee their adequate applicability over non-constant models;

R4.6.2: Effective incorporation within the learning process;
- **C4.6.2**: Principles to adequately prune of the search space of biclustering tasks in the presence of global and/or local constraints;

R4.7: Integrating homogeneity and significance views for a complete assessment of biclustering models;

Table 2: Proposed contributions to assess the statistical significance of non-constant and noisy biclustering models (*Chapter* V-2).

R4.2: Robust assessment of additive, multiplicative, symmetric, order-preserving and plaid coherencies;
C4.2.1.1: Assessment of additive (and symmetric) models based on the examination of patterns with different amplitudes;
C4.2.1.2: Assessment of multiplicative models based on the examination of patterns who share greatest common divisors;
C4.2.1.3: Assessment of order-preserving models based on the allowed linear orderings (permutations);
C4.2.1.4: Assessment of plaid models by recovering the original coherence in the absence of overlapping layers;
C4.2.2: Theoretically inferred equations to characterize the impact of flexible coherency assumptions on the search space;
C4.2.3: Depth-first searches with dynamic programming principles to avoid redundant computations;
R4.3: Robust assessment of regions with arbitrary-high levels of noisy and missing elements;
R4.3.1: Assessment of noisy biclusters;
C4.3.1a: Feature-based mode (or median for numeric/ordinal data) calculus to compute pattern expectations of constant models;
C4.3.1b: Extended mode calculus with removal of adjustments to assess additive, multiplicative, symmetric and plaid models;
C4.3.1c: Extension to compute mode permutations to model expectations of order-preserving models;
R4.3.2: Assessment of sparse biclusters;
C4.3.2: Retrieval of pattern expectations for: 1) structurally sparse data (such as network data) where elements may not have values assigned, and 2) data with arbitrary number of missings where elements may have multi-item imputations;
R4.7: Integrating homogeneity and significance views for a complete assessment of biclustering models;
C4.7.1: Effective combination of the p-value given by the probability of a given bicluster to deviate from expectations affected with the p-value associated with the probability of a bicluster to have unexpectedly low levels of noise;
C4.7.2: Alternative scores for combining views: 1) adjustments by quality based on deviation from expectations; and 2) adjustments based on the area to benefit low probable patterns;

Table 3: Contributions to assess the significance of real-valued biclustering models with continuous factors (*Chapter* V-3).

R4.4: Robust assessment in real-valued data contexts;
R4.4.1: Robust assessment of real-valued biclusters with (possibly unknown) coherency strength;
C4.4.1.1: Principles to estimate the original coherency strength and recover the underlying coherency assumption;
C4.4.1.2a: Non-biased estimators of the true significance of real-valued biclusters with lower and upper bounds (bar envelope) on the expected statistical significance;
C4.4.1.2b: Extended estimators to assess (real-valued) biclusters with non-constant coherencies;
C4.4.1.3: Comparison of the proposed estimators with the assessment of discretized biclusters (with multi-item assignments);
R4.4.2: Assessment of additive and multiplicative biclusters with continuous adjustment factors;
C4.4.2a: Estimate-driven retrieval of the allowed ranges of shifting and scaling factors;
C4.4.2b: Effective assessment of additive models based on the integral of the product of slided density functions;
C4.4.2c: Effective assessment of multiplicative models based on the integral of the product of size-adjusted (scaled) density functions;
C4.4.2d: Interpolation principles to guarantee the efficiency of the integral calculus;
R4.1.3: Guarantee the applicability to tabular data contexts with complex domains (mixtures of nominal, ordinal and numeric features);
C4.1.3.1: Three-step extension of the proposed assessments for complex domains (tests assuming dedicated distributions per feature);
C4.1.3.2: Principles to model the impact of dimensionality on significance for non-identically distributed features;
C4.1.3.3: Principles to approximate the properties of the search space, and thus the applied correction;
C4.1.3.4: Support for biclusters with mixtures of coherency assumptions;

Table 4: Major contributions to assess the statistical significance of local descriptive models from structured data (*Chapter* V-4).

R4.5: Robust assessment of the statistical significance of local descriptive models from structured data;
R4.5.1: Significance assessment of cascade models and arrangements of events;
C4.5.1.1: Related work on how to infer null models, non-parametric equations and hypothesis tests to test temporal patterns;
C4.5.1.2a: Statistical view on the expected support of sequential patterns by combining sequence- and itemset-wise views;
C4.5.1.2b: Hypothesis testing from this view with guarantees of deviations from expectations based on new efficient procedure to compute Hochbert corrections for large outputs and to parametrically control the rate of false positives and negatives;
C4.5.1.3a: Alternative statistical view on how to assess sequential patterns from probabilistic models, by testing: 1) the weight unexpectedness associated with the probability paths of a region, or 2) the region coverage and items' dependence;
C4.5.1.3b: Extensibility of statistical tests on probabilistic finite automata towards complex hidden Markov models;
C4.5.1.4: Shown compliance of these contributions to assess temporal patterns and, ultimately, cascades and arrangements of events;
R4.5.2: Learning guidance from statistical significance criteria;
C4.5.2a: Local statistical tests to monotonically prune the search space for deterministic methods;
C4.5.2b: Principles to guide the learning of Markov-based models and their decoding step;
C4.5.2c: Inference of global constraints from deviations on the Poisson distribution of expected occurrences;

1

Assessing the Statistical Significance of Biclustering Solutions

Despite the relevance of the biclustering task for several biomedical and social applications (Table I-1.1), there is not yet an accepted ground truth on how to guarantee the statistical significance of biclustering solutions. This is due to the fact that most existing approaches are guided by merit functions to guarantee the homogeneity of biclusters, but commonly do not subject them to sound statistical evaluation. Understandably, optimizing homogeneity levels is of limited use since good levels of homogeneity can appear by chance in the sample data since commonly observed for small biclusters. Although robust statistical principles are available to assess very specific types of constant biclusters (such as dense biclusters [616, 376] and biclusters with sequential constraints [427]), these principles are not generalizable for more flexible types of biclusters, including biclusters with varying coherency strength and assumptions.

In addition to these observations, the statistical assessment of biclusters is further challenged by two problems. First, the need to guarantee that the retrieved biclusters deviate from expectations to minimize the number of false positive biclusters (outputted biclusters that appear by chance on the sample data). Second, the need to infer significance criteria (such as expectations on the minimum size of biclusters) to guide the biclustering task without increasing the risk of excluding relevant biclusters (false negative biclusters). These problems are respectively associated with type-I and type-II errors.

This chapter explores why existing efforts in the field of biclustering are not yet able to address these problems, surveys the available contributions from related research streams and defines robust principles to efficiently assess the significance of biclustering solutions with a strict upper limit on the risk of false positives. This assessment can be used to either filter biclusters or as a sound heuristic to narrow the search space of biclustering algorithms, thus increasing their effectiveness and efficiency. In this context, five major contributions are proposed:

- structured view on the contributions and limitations of different research streams, including pattern mining and inferential statistics (estimation theory), for the effective assessment of biclustering models;
- extension of these contributions to: 1) guarantee their applicability to biclustering solutions with different forms of constant coherency, 2) adequately model the impact of dimensionality on the significance of biclusters, and 3) address efficiency bottlenecks associated with frequentist counts and assessments of regions without a non-fixed subset of i.i.d. features;
- integrative assessment method where statistical decisions (statistical tests, stochastic/frequentist views, dependency form) are dynamically selected from the properties of input data or from user expectations;
- a new multi-hypothesis correction to effectively and efficiently test deviation from expectations, thus addressing computational bottlenecks of non-conservative corrections while minimizing the risk of false negatives of conservative corrections;
- global statistical tests to infer constraints centered on expectations on biclusters that guarantee their statistical significance, surpassing the propensity towards type-II errors of existing global tests.

These contributions are critical to: 1) validate the increasing number of implications that are derived from real data without statistically sound guarantees, 2) evaluate and compare state-of-the-art biclustering algorithms with

regards to the significance of their solutions, and 3) guide the biclustering tasks, promoting the significance of their outputs and efficiency of the searches. The gathered results confirm the soundness and relevance of the proposed contributions to both assess and guarantee the significance of biclustering outputs, and stress the need to combine significance and homogeneity to affect the discovery of biclusters.

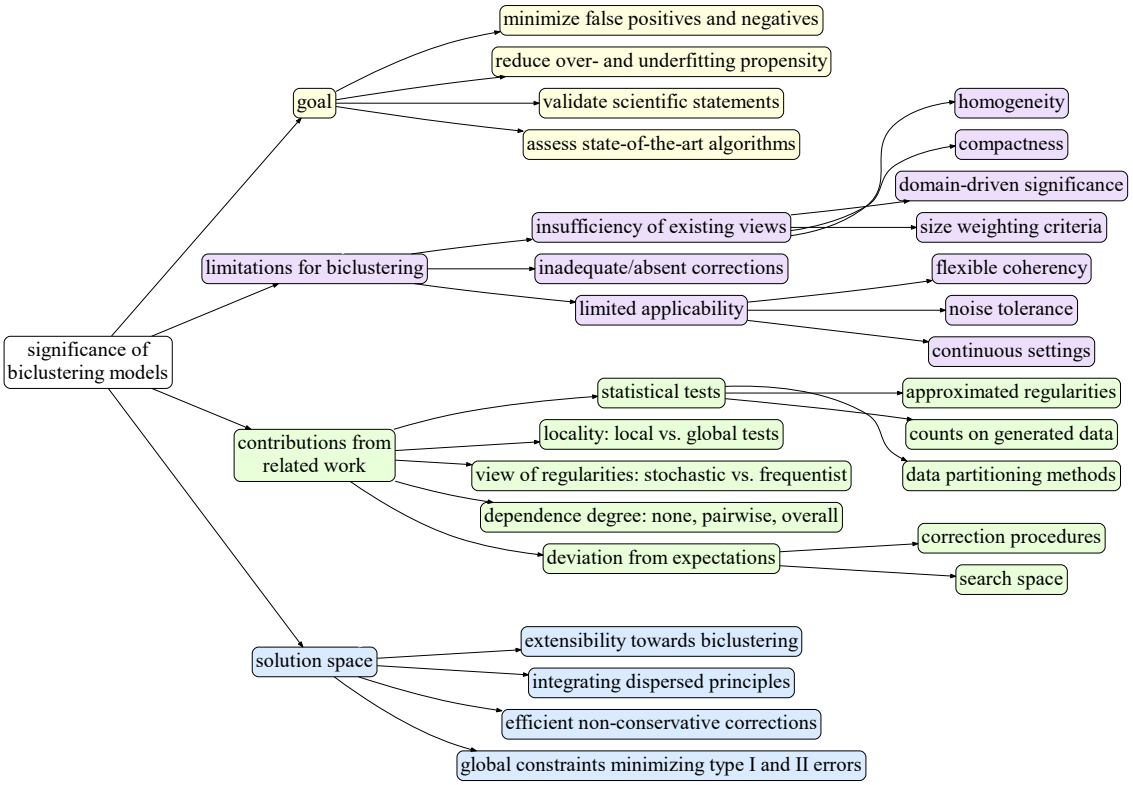

Figure 1.1: Challenges, limitations and proposed contributions for the significance assessment of biclustering models.

The enumerated problems and contributions are visually depicted in Figure 1.1. Accordingly, this chapter is organized as follows. *Section 1.1* describes the current challenges of statistically assessing biclustering models, and *Section 1.2* surveys related work streams with relevant contributions and limitations for the target problem. *Section 1.3* proposes a methodology to robustly assess biclustering solutions. *Section 1.4* provides initial empirical evidence of its relevance. Finally, the contributions and implications of this work are summarized.

Revised Notation: N, M, n and m. To guarantee the consistency of the contents in this book with the standard notation in statistics (instead of machine learning), we revised the notation of two concepts. Given a tabular dataset with $\mathbf{X}$ observations and $\mathbf{Y}$ features, and a bicluster $\mathbf{B}=(\mathbf{I}\subseteq\mathbf{X},\mathbf{J}\subseteq\mathbf{Y})$, then $N=|\mathbf{X}|$ is the data size, $M=|\mathbf{Y}|$ is the data dimensionality, $n=|\mathbf{I}|$ is the number of observations in $\mathbf{B}$ and $m=|\mathbf{J}|$ is the number of features in $\mathbf{B}$.

1.1 Background

In *Book III*, we provided an extensive background on the biclustering task according from the point of view of its homogeneity. According to this perspective, biclustering seeks to find subsets of observations (rows/columns) in tabular data with specific form and strength of coherency across a subset of features (columns/rows). In its simplistic form, the elements of a bicluster $a_{ij} \in (I, J)$ can be described by $a_{ij} = c_j + \gamma_i + \eta_{ij}$ for coherency on rows (or $a_{ij} = c_i + \gamma_j + \eta_{ij}$ for coherency on columns). In addition to the target coherency, biclustering may show specific forms and degree of tolerance to noise (modeled by the η_{ij} component). The coherency and quality components of the homogeneity criteria are commonly guaranteed through the use of merit functions. Illustrative merit functions for stochastic, greedy, recursive and exhaustive approaches to biclustering were surveyed in Table III-1.3.

To formalize the universe of discourse (using the revised definition of n, m, N and M), we need to revisit the notion of bicluster pattern and support. Given a bicluster B, the bicluster *pattern* $\varphi_{\mathbf{B}}$ is the ordered set of expected values in the absence of adjustment and noise factors: $\varphi_{\mathbf{B}}=\cup_{j=1}^{m}\{c_j\}$ for coherency on rows (or $\varphi_{\mathbf{B}}=\cup_{i=1}^{n}\{c_i\}$ for coherency on columns), and bicluster *support* $\sup_{\mathbf{B}}$ is, respectively, the number of observations $n=|I|$ (or $m=|J|$) respecting $\varphi_{\mathbf{B}}$. Given a bicluster with $\varphi_{\mathbf{B}}$ pattern and $\sup_{\mathbf{B}}$ support, its statistical significance is given by the deviation of its support against expectations. *Basics 1.1* provides an illustrative instantiation of the target task.

Basics 1.1 Motivation: assessing the significance of a constant bicluster

Consider a discrete tabular dataset with values in $\mathcal{L}$, N=1000 rows, M=200 columns, and $|\mathcal{L}|$=5 items uniformly distributed. Assume that we observe 3 columns with constant items across 25 rows. Is this bicluster statistically significant? The probability of the pattern occurrence is $p_{\varphi_{\mathbf{B}}}=1/5^3$. A simple binomial calculus shows that the probability to have at least 25 supporting rows is approximately $p_B=\binom{M}{m}\sum_{x=n}^{N}\binom{N}{x}p_{\varphi_{\mathbf{B}}}{}^{x}(1-p_{\varphi_{\mathbf{B}}})^{N-x}$=5.0E-3 where $n=|\mathbf{I}|$ and $m=|\mathbf{J}|$. Although $p_{\mathbf{B}}$ appears to be considerably low, we need to consider the space of all similar biclusters: $s=5^3$. Assuming the conservative Bonferroni correction under α=0.05 significance, $p_{\mathbf{B}}$ is assessed against α/s=4.E-4. Under these assumptions, $\mathbf{B}$ is rejected (false positive discovery).

Basics 1.2 Coming to terms with the terms

Given a tabular dataset $\mathbf{A}$, a bicluster is said to be *observed* if it appears in $\mathbf{A}$ (i.e. $\mathbf{I}\subseteq\mathbf{X}\wedge\mathbf{J}\subseteq\mathbf{Y}$), and it is said to be *unobserved* otherwise. The *statistical significance* of a bicluster (either observed or unobserved) is the probability of rejecting the following hypothesis: the bicluster is likely to appear in data sampled from the observed regularities $P_{\mathbf{A}}$. In an abbreviated form, we refer to this hypothesis as the bicluster's *probability of occurrence*, $p_{\mathbf{B}}$. In this context, given a dataset $\mathbf{A}$, the probability of an observed bicluster to occur is non-necessarily 100% since this probability is computed against the underlying data regularities $P_{\mathbf{A}}$ (assuming a preserved data size and dimensionality).

According to the terms discussed in *Basics 1.2*, let the probability of a bicluster to occur in a (n,m)-space sampled from $P_{\mathbf{A}}$ be $p_{\mathbf{B}}$. Given a dataset $\mathbf{A}$, the statistical significance of a bicluster $\mathbf{B}$ in $\mathbf{A}$ can be alternatively seen as the confidence level associated with the rejection of $p_{\mathbf{B}}$ (null hypothesis).

1.1.1 Limitations

Despite the increasing number of contributions in the field of biclustering, assessing the significance of biclustering solutions has been poorly explored. Below, we cover the major limitations of existing efforts.

Optimization and Scoring Schema. Although stochastic methods for biclustering rely on multivariate distributions to approximate the data, these methods do not use the learned density functions to test the significance of biclusters. Instead, they derive biclusters from the parameters of the learned models as soon as specific convergence criteria is satisfied.

Alternative approaches to biclustering define an objective metric of the relevance of the biclusters to narrow the search space and/or sort and filter the discovered biclusters. Typically, this scoring scheme commonly measures the homogeneity of the biclusters and, optionally, their overlapping degree using similarity penalizations (such as penalizations based on the Jaccard index). Understandably, these functions do not guarantee the likelihood that a bicluster is not found by chance in the sample data. Small biclusters can have high levels of homogeneity by chance. This propensity towards false discoveries is aggravated in high-dimensional spaces.

Some proposed merit functions in literature compensate this undesirable effect by weighting the size of biclusters in order to benefit the discovery of larger biclusters [503]. However, this strategy is insufficient to guarantee the significance of biclusters and often promotes a weak homogeneity, increasing the risk of false positive discoveries.

Testing Homogeneity Levels. Statistical tests can be applied to guarantee that the homogeneity is below a residual error with a particular confidence and statistical power. In DeBi [578], statistical tests are proposed to guarantee the compactness of biclusters by verifying if the exclusion or inclusion of specific rows or columns improves their homogeneity. However, these tests also suffer from the previous problem: homogeneity does not guarantee that a

bicluster is not discovered by chance and therefore it is not an adequate indicator of significance.

Additionally, tests can be used to guarantee the domain relevance of each bicluster by computing enrichment p-values against knowledge bases [67, 672]. Again, this is a poor metric of significance due to three major reasons. First, this evaluation only targets one dimension of the bicluster at a time (either the group of rows or columns). Second, knowledge bases are incomplete and prone to errors. Third, domain significance and statistical significance are not always in agreement.

Testing Specific Types of Biclusters. A few biclustering methods perform robust statistical tests to guarantee the significance of specific biclustering solutions. However, they cannot be easily extended to evaluate flexible biclustering solutions. Koyuturk et al. [376] defines an objective statistical function to find unusually dense biclusters in binarized matrices (high proportion of 1s). This is performed by assuming a memoryless dataset where the number of 1s, k, is binomially distributed. Chernoff's bound is used to compute the minimum density deviation from the mean that guarantees the significance of the bicluster, p-value $\propto \sqrt{\frac{k \times |\mathbf{I}| \times |\mathbf{J}|}{|\mathbf{X}| \times |\mathbf{Y}|}}$.

A generalized assessment of dense biclusters for real-valued data is done by SAMBA [616]. This algorithm maps the matrix into a weighted graph (whose weights are assumed to be normally distributed) and computes the p-value for each bicluster (subgraph) based on the probability of finding a bicluster with at least the same weight. However, even in the presence of corrections, this assessment is optimistically biased and only applicable to biclusters with differential values (i.e. biclusters with δ=1/2 strength of coherency).

CCC-Biclustering [427] measures the significance of a bicluster $\mathbf{B}$ with constant values across rows given by a string pattern $\varphi_{\mathbf{B}}$ (adjacent conditions) by computing the tail of the binomial distribution from the probability $p_{\mathbf{B}}$ of a bicluster respecting $\varphi_{\mathbf{B}}$ to occur by chance in a matrix with the same frequency of contiguous pairs of symbols (first-order Markov assumption).

Other statistical tests have been proposed to study the probability of a bicluster (following a strict coherency assumption) to occur against background noise [57, 116]. Understandably, these assessments rely on assumptions only applicable for specific types of constant biclusters, and therefore are not extensible towards biclusters following more flexible coherency criteria.

1.2 Related Work

A substantial portion of the contributions to assess the statistical significance of local regions has been developed in the context of pattern mining (PM) [367, 190, 252, 587]. Pattern miners that apply statistical significance tests before outputting patterns include [97, 410, 671, 704]. As previously seen, a pattern can be mapped as a bicluster with: *1)* constant values across rows (a_{ij}=c_j), and *2)* no noise (η_{ij}=0). In this context, the relevance of a pattern is defined by its support (number of rows) and pattern length (number of columns). This mapping (formalized in Defs.III-1.6 and III-1.7) allowed the definition of several approaches for pattern-based biclustering in the last years [500, 578, 509]. Unless background knowledge is available, the minimum support is typically low in pattern-based biclustering [5], leading to an increase risk of false positive biclusters (outputting non-significant biclusters). To minimize this risk, *Section 1.2.1* surveys available statistical tests to guarantee the significance of constant biclusters derived from patterns, and *Section 1.2.2* describes correction procedures to test their deviation from expectations.

1.2.1 Significance of Constant Biclusters

Computing Expectations. To test the significance of a bicluster $\mathbf{B}$, the regularities of the input matrix $\mathbf{A}$, Θ, need to be adequately modeled to assess the probability of $\mathbf{B}$ occurrence, $p_{\mathbf{B}}$. There are three major directions to make this assessment. $p_{\mathbf{B}}$ can be computed by testing $\mathbf{B}$ against: *1)* approximated distributions underlying the original matrix, *2)* randomized synthetic datasets (assuming similar distributions or permutations of the observed values), and *3)* hold-out data partitions.

First, some studies approximate the distributions underlying data, Θ, by either fitting distributions for each row $\mathbf{x}_i \sim \Theta_i$, column $\mathbf{y}_j \sim \Theta_j$, or for the overall matrix $a_{ij} \sim \Theta$ [358]. Then, $p_{\mathbf{B}}$ is often estimated from the joint probability of a specific pattern $\varphi_{\mathbf{B}}$ to occur, $p_{\varphi_{\mathbf{B}}}$, for a minimum number of rows by computing Binomial tails [427].

Second, instead of relying on the observed data, some approaches generate multiple synthetic datasets from the underlying data regularities Θ [252, 367]. Here, $p_{\mathbf{B}}$ can be estimated by testing the observed $\varphi_{\mathbf{B}}$ support against the number of observations that support $\varphi_{\mathbf{B}}$ pattern in each one of the h generated datasets, $\{sup_1,..,sup_h\}$[1]. Gionis et al. [252] generated datasets based on all arrangements of transactions that satisfy the exact item frequencies and average transaction lengths of the original dataset. Megiddo and Srikant [450] relaxed this criteria, by assuming independence of items among transactions while preserving their frequencies. Monte Carlo sampling is an additional option [367].

Finally, assessments based on the creation of hold-out data partitions [670] have been additionally proposed to control the false discovery rate [63]. This option is accomplished by testing the support of patterns found in the exploratory against the holdout partitions using paired t-tests.

Statistical Tests against Expectations. In pattern discovery, density functions have been proposed based on the deviation of the observed support against the expected support (derived from one of the introduced strategies). Let $s\hat{u}p_{\mathbf{B}}$ to be the expected support of a pattern $\varphi_{\mathbf{B}}$ and $\hat{\sigma}(sup_{\mathbf{B}})$ to be the expect support deviation, then some significance ratios include: $sup_{\mathbf{B}}/s\hat{u}p_{\mathbf{B}}$, $|sup_{\mathbf{B}}\text{-}s\hat{u}p_{\mathbf{B}}|$, and $(sup_{\mathbf{B}}\text{-}s\hat{u}p_{\mathbf{B}})/\hat{\sigma}(sup_{\mathbf{B}})$ [190, 189, 597]. Similar scores have been also proposed in the context of association rule mining [569, 288, 704]. Aggarwal and Yu [5] relate the significance of an itemset with the quantity $((1\text{-}v(\mathbf{B}))/(1\text{-}E[v(\mathbf{B})])) \cdot (E[v(\mathbf{B})]/v(\mathbf{B}))$, where $v(\mathbf{B})$ is the fraction of transactions containing some but not all the items of $\varphi_{\mathbf{B}}$ and $E[v(\mathbf{B})]$ is the expectation of $v(\mathbf{B})$ in a random dataset where items occur in transactions independently. Bayesian analysis is employed in [586] to derive alternative scores. The problem with the use of the introduces ratios is that, instead of expressing the probability of occurrence, they are simple indicative of the relevance of a pattern.

To address this problem, statistical tests have been proposed. χ^2 tests were proposed by Silverstein et al. [587] (and revised by DuMouchel and Pregibon [189]) to assess a pattern based on the degree of dependence among its constituent items using synthetic data. In the absence of generated background datasets, the computation of Binomial tails has been also applied to compute the probability of string patterns to be observed for a minimum set of rows [427]. Bayesian assessments with shrinkage estimates [190, 189] were proposed to provide a conservative true probability of support's significance and hence minimize the number of false discoveries. However, these estimates do not provide a general mechanism for applying hypothesis. The significance has been also estimated from a Bayesian network with parameters derived from Θ [351].

Kirsch et al. [367] proposed a novel methodology to identify a global and meaningful support threshold ρ that yields a substantial deviation from what would be expected in a random dataset with the same item frequencies (following [587]). Despite the presence of other attempts to infer global parameters of significance (including Bayesian and Markov-based views), they were conducted in a purely qualitative fashion. Although global parameters can be inferred efficiently and used right-away as an heuristic for biclustering methods, small yet significant biclusters (very low $p_{\varphi_{\mathbf{B}}}$) are incorrectly seen as false discoveries.

1.2.2 Deviation from the Expectations

To guarantee that the probability of a bicluster to occur deviates from the expected probability of occurrence, its significance needs to be corrected against the search space of similar biclusters. The search space depends essentially on the: 1) allowed coherencies, 2) considered similarity criterion (e.g. biclusters with the same area or pattern length), and 3) applicable relaxations. Illustrating, Bay and Pazzani propose a division of the search space

[1] An illustrative statistical test is the percentage of datasets with support higher than $\hat{\theta}$: $p(x) = \frac{1}{h}\Sigma_{i=1}^{h} f(x-\hat{\theta})$, where $f(z)$=1 if $z \leq 0$ and 0 otherwise.

of similar association rules in multiple batches according to the number of items in the rule antecedent [48]. To our knowledge, there are not yet principles to efficiently characterize the search space of flexible biclusters.

For a given search space, robust correction procedures commonly rely on the family-wise error rate (FWER) – the probability of accepting at least one false positive (flagging a non-significant subspace as significant) [64]. A conservative correction is the Bonferroni adjustment [580], which bounds the α risk of false discoveries when assessing s tests by the corrected α/s level for each test. To avoid a large increase in the number of neglected relevant biclusters (false negatives), non-conservative options should be adopted in order to trade-off the risks of type-I errors (false positives) and type-II errors (false negatives). The Holm and Hochbert procedures are less conservative, while still verify the FWER constraint [325]. Alternative correction procedures have been proposed in the context of multiple comparisons [353] and oversearch [541]. Finally, non-FWER procedures for multi-hypotheses tests can be considered to minimize false negatives, while still providing adequate guarantees on the risk of false positives [64, 63].

1.3 Solution

Having covered structural contributions from related work to guarantee the statistical significance of pattern mining outputs, *Section 1.3.1* proposes key extensions to guarantee their applicability for constant biclusters and integrates them within a consistent method. *Section 1.3.2* proposes new procedures to test multi-hypothesis with guarantees on type-I and type-II errors. For this aim, a variant of the Hochbert procedures is proposed to surpass the efficiency bottlenecks of non-conservative FWER corrections in the presence of large search spaces. Finally, *Section 1.3.3* proposes a method to infer global constraints (and guide biclustering searches) based on the minimum expected size that guarantees the statistical significance of a given bicluster.

1.3.1 Assessing Perfect Constant Biclusters: Integrating Statistical Principles

A taxonomy with the decisions impacting the statistical assessment of biclusters was provided in Figure 1.1. Accordingly, the surveyed contributions can be used to define local and global statistical tests (against the observed matrix, generated datasets or hold-out partitions) using either stochastic or frequentist views with different degrees of dependence among items. Fixing these decisions essentially depends on: *1)* the properties of the input matrix (size, dimensionality and regularities), and *2)* the behavior of the selected biclustering method (as it determines the homogeneity of biclusters).

Given a set of items $\mathcal{L}$, a discrete bicluster $\mathbf{B}$ is referred as *perfect* if it does not contain noisy elements, $a_{ij} \in \mathcal{L} \wedge \eta_{ij}{=}0$. Under the proposed mapping, existing contributions are applicable to the assessment of discrete, perfect and constant biclusters. Despite the relevance of existing contributions, the majority of them still suffer from a lack of robust statistical views and the absence of adequate corrections. In this context, we provide a structured view on how these contributions can be consistently combined.

Statistical Tests

Assuming coherency is observed across rows, statistical tests based on binomial tails are robust as they surpass the need to rely on approximations from inequalities associated with tail bounds. In this context, a binomial tail defines the probability of a bicluster with pattern $\varphi_{\mathbf{B}}$ to occur across a set of rows, $p'_{\mathbf{B}} = P(Z \geq m)$ with $Z \sim Bin(p_{\varphi_{\mathbf{B}}}, N)$, where $m{=}|\mathbf{I}|$, $N{=}|\mathbf{X}|$ and $p_{\varphi_{\mathbf{B}}}$ is the probability of the $\varphi_{\mathbf{B}}$ items to occur. Contrasting with the research on pattern mining (where patterns are derived from transactional databases), the significance of a bicluster from a tabular data context needs to be adjusted by the probability of the bicluster to occur for any combination of columns, which is approximately given by $p_{\mathbf{B}} = 1 - P(Z < m)^{\binom{M}{m}} = \binom{M}{m} P(Z \geq m)$ when assuming that columns have a similar distribution of items. This proposed probability, given by (1.1), essentially depends on $p_{\varphi_{\mathbf{B}}}$ and on the size and dimensionality of both the bicluster and the input matrix. Although the previous calculus allows for $p_B{>}1$ when

there is a high likelihood that multiple biclusters with φ_B pattern and more than n rows occur, this probability can be bounded by 1 or, alternatively, used for subsequent $p_B \approx 0$ hypothesis testing.

Figure 1.2 provides an illustrative application of the proposed statistical principles to test the significance of a perfect constant bicluster with coherency across rows in a discrete space ($a_{ij} \in \{0..7\}$). The input coherency as defined by the number of items affects both of the sides of the testing equation: the probability of $\varphi_{\mathbf{B}}$ occurrence (and consequently $p_{\mathbf{B}}$) and the search space (and consequently the significance level of the applied correction).

$$p_{\mathbf{B}} = \binom{M}{m} \sum_{x=n}^{N} \binom{N}{x} {p_{\varphi_{\mathbf{B}}}}^{x} (1 - p_{\varphi_{\mathbf{B}}})^{N-x} \tag{1.1}$$

Similarly, a bicluster with coherency across columns has $p_{\mathbf{B}} = \binom{N}{n} P(Z \geq |J|)$ with $Z \sim Bin(p_{\varphi_{\mathbf{B}}}, M)$.

For assessing biclusters with a constant value ($\sigma \in \mathcal{L}$) on both rows and columns, statistical tests can be defined by assuming a memoryless dataset where the number of σ occurrences is binomially distributed. Under this assumption, the statistical tests proposed in [376, 57] can be used to test unusually dense biclusters in binarized matrices (high proportion of σ items against the remaining $\mathcal{L} \backslash \sigma$ items) based on the σ frequency.

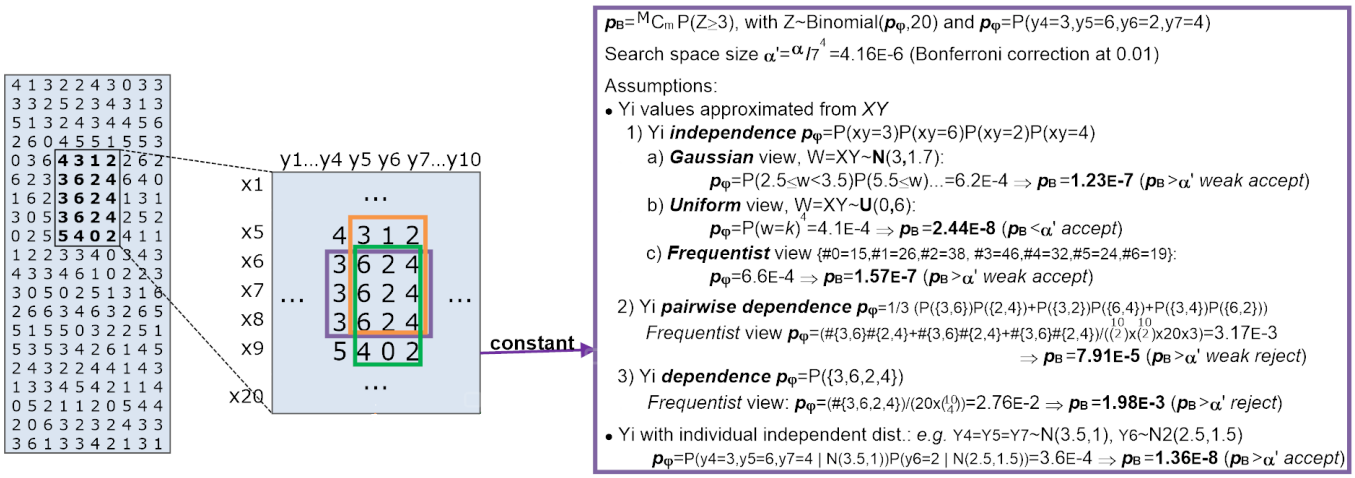

Figure 1.2: Illustrative statistical assessment of constant (non-noisy) biclusters in discrete settings.

The $p_{\varphi_{\mathbf{B}}}$ Calculus

Given the common biclustering context, where coherency is verified across observations, the $p_{\varphi_{\mathbf{B}}}$ calculus essentially depends on the approximated regularities underlying data, $P_{\mathbf{A}} \sim \Theta$. Θ can be given by a univariate distribution, or, when independence is assumed among columns (or rows), by a N-order (or M-order) multivariate distribution.

Assuming column independence, $p_{\varphi_{\mathbf{B}}}$ is either the joint probability of the observed items to occur on the corresponding columns (e.g. $P(y_4{=}3)P(y_5{=}6)P(y_6{=}2)P(y_7{=}4)$) or the sum of the Cartesian product when columns are not fixed (e.g. $\Sigma_{j=4}^{7} \Pi_{a \in \{3,6,2,4\}} P(y_j{=}a)$). For this case, we propose the use of dynamic programming principles to avoid the potential efficiency bottlenecks associated with the repeated computation of subsets of products.

Figure 1.2 assumes items to be *ordinal* to show how the different Θ approximations impact results. Under this assumption, continuous distributions can be applied. However, whenever these distributions show a large error (given, for instance, by the normalized mean square error between the approximated and observed models), the use of fitting tests to find alternative distributions (without overfitting propensity) is of critical importance. Illustrating, when the mode item differs from the median item, multimodal distributions can be dynamically selected.

Alternatively, frequentist distributions can be approximate to compute p_{φ_B}. In fact, they show the best fit and are not prone to problems associated with complex distributions of frequencies per items. However, in the presence of small sized matrices, they are susceptible to overfit the observed data.

Whenever this is the cases, a more robust strategy is to generate h background datasets and replace the Binomial calculus based on $p_{\varphi_{\mathbf{B}}}$ by a test, such as an unilateral t-Student, based on the h support estimates of $\varphi_{\mathbf{B}}$.

Dependency Between Items

If a form of dependence is assumed between the elements in a bicluster, either probabilistic or frequentist views can be considered. Frequentist views are the default option when pairwise or overall dependency among the items of a bicluster pattern $\varphi_{\mathbf{B}}$ is assumed. In this context, the possible combinations of dependent items from a pattern $\varphi_{\mathbf{B}}$ are thus either counted in the original dataset or used to generate the background datasets.

When coherency across rows is considered, the mean estimator for the counts of subsets of items per row can be used. In addition to the counting of subsets of items, there is the need to define principles for combining their influence to compute the probability of the overall pattern, $p_{\varphi_{\mathbf{B}}}$. According to the illustrative matrix in Figure 1.2, $p_{\varphi_{\mathbf{B}}} = P(\{y_4, y_5, y_6, y_7\}=\{3,6,2,4\}) = (\#\{3,6\}\#\{2,4\}+\#\{3,2\}\ \#\{6,4\}+\#\{3,4\}\#\{6,2\})/({}^{10}C_2\times{}^{10}C_2\times 20\times 3)$, or, in the presence of an odd number of items, $p_B = P(\{3,6,6\})=(\#\{3,6\}\#\{3\}\times 2 + \#\{6,6\}\#\{3\})/({}^{10}C_2\times 10\times 20\times 3)$. Different strategies can be used to compute the pairwise counts, such as the average counts per row (if $\{x_1=\{4,1,3,2,2,4,3,0,3,3\},..,x_{20}\}$, then $\#\{3,2\} = \frac{1}{20}({}^3C_1 \times^2 C_1 + ..)$ and $\#\{3,3\} = \frac{1}{20}({}^3C_2 + ..)$). Similar principles are used when assuming alternative forms of dependence among the items of a bicluster pattern $\varphi_{\mathbf{B}}$. We propose the use of dynamic programming to compute counts and permutations apriori, thus avoiding redundant computational effort associated with the $p_{\varphi_{\mathbf{B}}}$ calculus.

High-order form of dependency among items is only robust for small patterns or large matrices since missing a single item from a lengthy pattern does not contribute to its counts. This leads to overly pessimistic views of the statistical significance of a bicluster, increasing the propensity of the assessment towards type-II errors. To avoid this, frequentist views can be replaced by probabilistic views to model different forms of dependency between based on conditional probability calculus.

Integrative View

In this section, we extended the statistical principles proposed in the context of pattern mining to: 1) guarantee their applicability to biclusters with different forms of constant coherency, 2) avoid efficiency bottlenecks associated with the $p_{\varphi_{\mathbf{B}}}$ calculus for bicluster's patterns with non-indexed columns, 3) allow an accurate $p_{\varphi_{\mathbf{B}}}$ calculus when considering different forms of dependency among $\varphi_{\mathbf{B}}$ items, and 4) avoid the computationally expensive task of performing an arbitrary-high number of counts. To guarantee the correct application of these principles, we propose an integrative assessment of constant biclusters grounded on the calculus of binomial tails based on the regularities of data Θ (1.1). When independence is assumed (default setting), Θ is directly approximated from data by fitting multivariate distributions (according to tests and fitting measures proposed in [358]). Otherwise, row-based counts are performed on the original dataset when $N>200 \wedge n\sqrt{M}>5000$ (see experimental evidence) and on $h=30$ background datasets (generated using the randomization principles proposed by Ojala et al. [499]) for the remaining cases.

1.3.2 Robust and Efficient Corrections

Given $|\mathcal{L}|=7$, $\varphi_{\mathbf{B}}=\{3,6,2,4\}$ and assuming columns to be fixed, the search space is defined by the set of all patterns with the same size (7^4 biclusters). If columns are not fixed, larger spaces need to be considered ($7^4\binom{10}{4}$ biclusters), thus increasing both the probability of a pattern to occur $p_{\varphi_{\mathbf{B}}}$ as well as the correction effect (from testing multiple hypothesis). When considering the conservative Bonferroni correction, the considered α confidence to test the null hypothesis is simply divided by the search space size s. However, this correction often leads to pessimistic views of the true significance and it is therefore associated with a high susceptibility towards false negatives (rejecting biclusters that are statistically significant). In fact, the majority of available assessments either ignore the application of corrections (propensity to type-I errors) or apply the Bonferroni adjustments (propensity to type-II errors).

To address this problem, the Hochbert correction can be applied. The p-values from the s biclusters from the generated search space are sorted according to their probability of occurrence, $\{p_{\mathbf{B}_1},..,p_{\mathbf{B}s}\}$, and the p-value

$max_{p_{\mathbf{B}j}} : \forall_{1 \leq j \leq s} p_{\mathbf{B}j} \leq \alpha/(s - j + 1)$ is outputted as the corrected level. Despite offering a good compromise between type-I and type-II errors, this strategy is impracticable in presence of a large number of biclusters as it implies a high number of Binomial tail calculus.

To tackle this problem, we propose the use of a binary space partitioning (BSP) method that recursively subdivides the space based on the frequency of each item. For the introduced example, assuming that the descendant order of items by frequency is {3,2,4,5,1,6,0}, BSP starts by computing the probability for bicluster with $\varphi_{\mathbf{B}}$={5,5,5,5} to divide the space, and compares its probability with the corrected significance $\alpha/(7^4/2)$. In case this probability is lower, then BSP compares $p_{\mathbf{B}}$ with $\varphi_{\mathbf{B}}$={6,6,6,6} against $\alpha/(7^4/2 + 7^4/4)$, and, if it is still lower, BSP tests $p_{\mathbf{B}}$ with $\varphi_{\mathbf{B}}$={6,6,0,0} and so forth. A total of 10 tests provides already a highly accurate approximation of the correct significance with heightened efficiency.

1.3.3 Global Constraints

Global constraints are expectations that can be used to characterize the significance of biclustering solutions without relying on individual tests. Let $Q_{m,n}$ be the number of coherent biclusters with m features (columns/rows) with at least n supporting observations (rows/columns). There is a specific number of observations θ, such that for all $n \geq \theta$ the distribution of $\hat{Q}_{m,n}$ is well approximated by a Poisson distribution [367]. Based on this approximation, efficient parametric tests can be employed to estimate the null hypothesis space, which corresponds to $Q_{m,n}$ not deviating from the expected number of biclusters with $|I| \geq n \wedge |J|$=m (that is, $Q_{m,n}$ coming from a Poisson distribution of suitable expectation). The number of observations θ that guarantee the significance of a bicluster with m features can, thus, be estimated by iteratively testing the Poisson distribution fitting for an incrementally large number of observations. Relying on the work of Kirsch et al. [367], these computations can be efficiently performed recurring to approximated analytical bounds b_1 and b_2 that satisfy: $min_\theta : P(b_1(m,n) + b_2(m,n) \leq \epsilon) \geq 1\text{-}\delta$ with 1-δ statistical power. Since the estimation of these bounds were presented in the context of pattern mining and require the search for all frequent itemsets for each new support, we propose the use of an exhaustive pattern-based biclustering approach, such as the proposed BicPAM algorithm, to exhaustively search for all biclusters either in the original dataset or in background datasets under a tight noise-allowance criteria.

This assessment can be used to tackle the computational cost of individually testing a large number of bicluster (particularly critical under non-conservative corrections procedures), as well as to define an heuristics that can be easily plugged in learning process to constrain the search space.

1.3.3.1 Addressing Criticisms: Minimizing type-I and type-II Errors

A criticism of global strategies is related with the fact that despite m-length biclusters with at least θ support are considered to be statistically significant, this does not necessarily imply that every m-length bicluster is significant. Illustrating, non-significant biclusters with patterns associated with items with frequency above average (high $p_{\varphi_{\mathbf{B}}}$) often have more than ρ supporting observations and thus are incorrectly seen as significant (false positives). Additionally, m-length biclusters with $p_{\varphi_{\mathbf{B}}}$ below average yet less than θ supporting observations may be incorrectly seen as non-significant (false negatives). In this context, even when the inference of the θ support arguably follows principles to minimize the rate of false discoveries [367], the fact that only a single support threshold is outputted is unavoidably associated with the enumerated problems.

In order to overcome these problems, we propose the computation of a lower and upper bound for the minimum number of rows that guarantee a deviation from expectations $\theta \in [\theta_{min}, \theta_{max}]$. In this context, the lower bound can be used to minimize the risk of false negatives, while the upper bound to minimize the risk of true negatives. Under this range of values, a biclustering can made decisions regarding the output of a bicluster with n=$|I|$ observations based on its φ_B pattern. Whenever a $\varphi_{\mathbf{B}}$ pattern is primarily composed by items with low frequency, then the reference support θ should decrease (closer to θ_{min}); while when composed by frequent items, θ should increase

towards a threshold closer to θ_{max}. These decisions can be dynamically performed based on the observed $\varphi_{\mathbf{B}}$ pattern. Illustrating, given a dataset where the least and most frequent items have f_{min} and f_{max} frequencies, then for a m-pattern length (where $p_{\varphi_{\mathbf{B}}} \in [(f_{min})^m, (f_{max})^m]$ when assuming independence among items): $\theta = \frac{(\theta_{max}-\theta_{min})(p_{\varphi_{\mathbf{B}}}-(f_{min})^m)}{(f_{max})^m-(f_{min})^m} + \theta_{min}$ (an alternative calculus of $p_{\varphi_{\mathbf{B}}}$ bounds can be applied). For computing θ_{max}, the exhaustive biclustering search (BicPAM) applied within the previously described strategy can be parameterized with individual statistical tests on certain biclusters (biclusters with $p_{\varphi_{\mathbf{B}}}$ above average) in order to exclude biclusters with support above θ yet non-significant from the outputs. An approximate estimation of the θ_{min} (with adequacy shown from empirical evidence) results from applying of the exhaustive biclustering search to not only return all found biclusters with m length but additionally all statistically significant biclusters with m-1 length.

An alternative and more simplistic procedure to compute the $[\theta_{min}, \theta_{max}]$ thresholds is to derive them from the expected support for low and high probable biclusters' patterns (given by the 20-percentile and 80-percentile of the $p_{\varphi_{\mathbf{B}}}$ range) that turn them statistically significant. These computations (see *Section 1.4*) can be efficiently performed by reversing the binomial calculus and estimating n using principles from a binary space partitioning.

1.4 Results and Discussion

In this section, we provide empirical evidence for the relevance of the proposed principles. First, we use local statistical tests to study the minimum number of observations and features that guarantee the significance of a constant bicluster for data with varying size and dimensionality. Second, we assess the effectiveness and efficiency of the proposed correction procedure. Third, we show how different statistical decisions (such as distribution and dependency assumptions) impact the significance of biclusters observed in real data. Finally, we analyze inferred global constraints derived from real data. The proposed statistical methods[2] were coded in Java (JVM v1.6.0-24), and run in an Intel Core i3 1.80GHz with 6GB of RAM.

Impact of N, M, n, m, $|\mathcal{L}|$ and $\varphi_{\mathbf{B}}$ in Statistical Significance. Tables 1.1 and 1.2 describe how the statistical significance of a constant bicluster varies: with its support, length, coherency strength; with the size, dimensionality and regularities of the original matrix; and with the probability associated with the $\varphi_{\mathbf{B}}$ pattern. In these analyzes, we assess the expected minimum support required to guarantee the significance of a bicluster. We assume $p_{\varphi_{\mathbf{B}}}$ to have items with average, above-average and below-average probability of occurrence. Illustrating, a bicluster with pattern length m=$|\mathbf{J}|$=5 and coherency strength given by $|\mathcal{L}|$=5, observed in a dataset with N=20000 observations and M=100 features, should have a minimum support of n_{min}=20, $\hat{n}$=33 and n_{max}=55 observations for below-average, average and above-average $p_{\varphi_{\mathbf{B}}}$ (or in other words $n \in [20, 53]$ with $\hat{n}$=33) in order to be significant. Similarly to global constraints, the inferred expectations from these tables can be used to guide the learning.

Table 1.1 shows how this minimum number of observations in the bicluster varies with the considered coherency strength (number of items), $|\mathcal{L}|$, and its pattern length (number of features), m. The *coherency strength* (number of items) largely impacts significance: fewer items are associated with a looser homogeneity and therefore larger biclusters are required to preserve statistical significance. Assuming m=5, N=20000 and M=100 and a constant coherency, for differential values ($|\mathcal{L}|$=2): $n \in [70, 309]$ with $\hat{n}$=153, while when considering $|\mathcal{L}|$=5, the expected minimum number of rows decreases significantly, $n \in [20, 53]$ with $\hat{n}$=33. Understandably, the pattern *length* m also strongly affects p_{φ_B} and thus the expected minimum number of supporting observations. Assuming $|\mathcal{L}|$=5, N=20000 and M=100, then $\hat{n}$=93, $\hat{n}$=33 and $\hat{n}$=12 are respectively expected for m=3, m=5 and m=7 lengths.

Table 1.2 shows how the minimum support of a bicluster to guarantee its statistical significance varies with the number of observations, N, and features, M, of the input dataset. The *data size*, N, impacts significance as it shapes the binomial tails. Assuming $|\mathcal{L}|$=5, m=5 and M=100, the expected minimum number of rows to guarantee significance is $\hat{n}$=8 for a constant bicluster when N=2000, and largely increases to $\hat{n}$=44 when N=50000

[2] Available in web.ist.utl.pt/rmch/software/bsig/

(magnitude of human genome). Although the data dimensionality, M, also visibly affects $\hat{n}$, it is not so critical as previous variables since when considering coherency across rows. If coherency across columns is targeted, we would observe the inverse effect: increased sensitivity to the number of data features and less sensitivity to the number observations.

	m	5	5	5	5	**m**	**2**	**3**	**5**	**7**
	N	20000	20000	20000	20000	**N**	20000	20000	20000	20000
	M	100	100	100	100	**M**	100	100	100	100
	$\|\mathcal{L}\|$	**3**	**4**	**5**	**10**	$\|\mathcal{L}\|$	5	5	5	5
$p_{\varphi_B}=\frac{1}{\|L\|}^m$	$\mathbf{n}_{min}$	**153**	**59**	**33**	**11**	$\mathbf{n}_{min}$	**930**	**93**	**33**	**12**
$p_{\varphi_B}\in[\frac{0.8}{\|L\|}^m, \frac{1.2}{\|L\|}^m]$	$\mathbf{n}_{min}$	[70,309]	[32,106]	[20,53]	[7,14]	$\mathbf{n}_{min}$	[619,1304]	[139,375]	[20,53]	[9,18]
$\alpha'=\frac{\alpha}{\|L\|^m}$	α'	2.1E-4	4.9E-5	1.6E-5	5.0E-7	α'	2.0E-3	4.0E-4	1.6E-5	6.4E-7

Table 1.1: Impact of the number of items, $|\mathcal{L}|$, and pattern length, $|J|$, on the expected minimum number of observations in the bicluster that guarantees its statistical significance (for fixed number of data observations n and features m).

	m	5	5	5	5	5	**m**	5	5	5
	N	**200**	**500**	**2000**	**10000**	**50000**	**N**	20000	20000	20000
	M	100	100	100	100	100	**M**	**50**	**200**	**1000**
	$\|\mathcal{L}\|$	5	5	5	5	5	$\|\mathcal{L}\|$	5	5	5
$p_{\varphi_B}=\frac{1}{\|L\|}^m$	$\mathbf{n}_{min}$	**8**	**9**	**14**	**24**	**44**	$\mathbf{n}_{min}$	**31**	**35**	**40**
$p_{\varphi_B}\in[\frac{0.8}{\|L\|}^m, \frac{1.2}{\|L\|}^m]$	$\mathbf{n}_{min}$	[5,9]	[7,12]	[10,18]	[16,37]	[30,94]	$\mathbf{n}_{min}$	[19,50]	[23,56]	[27,63]
$\alpha'=\frac{\alpha}{\|L\|^m}$	α'	1.6E-5	1.6E-5	1.6E-5	1.6E-5	1.6E-5	α'	1.6E-5	1.6E-5	1.6E-5

Table 1.2: Impact of data size and dimensionality on the expected minimum number of observations in a constant bicluster that guarantees its statistical significance (for a fixed number of items and pattern length).

Multi-Hypotheses Correction. Figure 1.3 compares the impact of considering different correction procedures that respect the family-wise error. We considered two correction procedures: Bonferroni and the proposed variant of the Hochbert procedure (using 20 iterations/tests to compute the adjusted significance) against a baseline significance level of α=5%. For this analysis we generated a discrete dataset with N=5000 rows and M=100 columns with a distribution of the items that can be approximated by a Gaussian distribution $N(\frac{|\mathcal{L}|}{2}, \frac{|\mathcal{L}|}{5})$. The provided assessment is made for a planted bicluster with fixed number of features ($|\mathbf{J}|$=4) and observations ($|\mathbf{I}|$=20), and varying coherency strength ($|\mathcal{L}| \in \{4\}$). Understandably, the observed differences between the correction procedures reveal that a non-conservative procedure that still preserves the family-wise error rate is critical to reduce the risk of false negative biclusters (type-II error). Furthermore, under the introduced binary partitioning principles for the application of the Hochbert correction, the statistical significance can be computed efficiently, even in the presence of very large search spaces. Finally, the analysis provided in Figure 1.3 also underlines the importance of using correction procedures that take into account different coherency assumptions (see *Chapter V-2*) for an adequate identification of the cut-off thresholds that guarantee the deviation from the expected probability of occurrence.

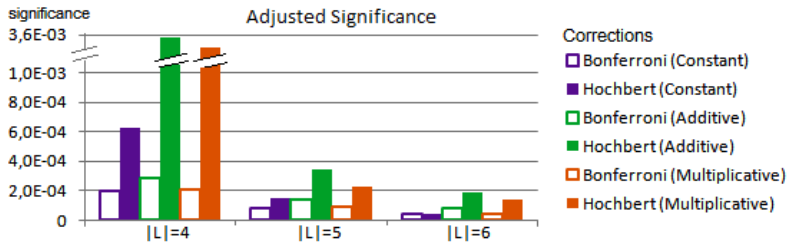

Figure 1.3: Significance needed to guarantee that the probability of occurrence of a bicluster deviates from expectations (using Bonferroni and Hochbert procedures with α=5%).

Modeling Assumptions. Figure 1.4 assesses the impact of considering different assumptions for modeling the regularities of real discrete data, including varying distributions and dependence degree between items. For this aim, we used the Yeast cell cycle dataset [619] and extracted the largest perfect biclusters with different coherency assumptions from each one of these data settings (exhaustively mined with BicPAM [310]) for a coherency

strength given by $|\mathcal{L}|$=5. In this context, the proposed integrative assessment method was parameterized with stochastic views and frequentist views (with pairwise and overall dependence between items). Understandably, the largest biclusters are significant. The average number of rows of these biclusters increases under non-constant assumptions. This effect is compensated by their higher p_{φ_B} (see *Chapter V-2*), thus leading to comparable levels of significance with constant biclusters. Two major observations can be retrieved. First, assuming pairwise and overall conditional dependence between φ_B items can largely impact the significance analysis. Conditional dependence, and in particular overall items' dependence, should be only made available with high-dimensional datasets ($M > 100$). Otherwise, the counts of co-occurrences can be associated with pessimistic probabilities since missing a single item from a lengthy pattern does not contribute to its support. Second, statistical significance levels under (item-independent) frequentist and Uniform approximations are similar since the inputted coherency strength is associated with equiprobable Gaussian cut-off points and therefore items in $\mathcal{L}$ have identical frequency.

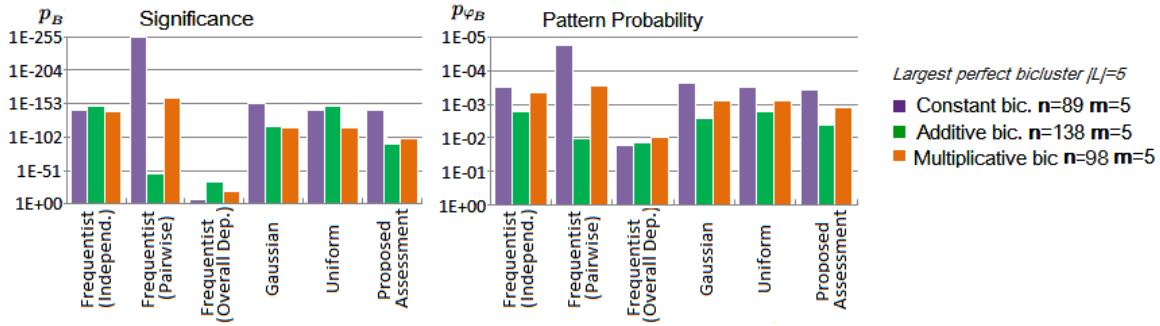

Figure 1.4: Impact of stochastic and frequentist views to assess large biclusters discovered from expression data [619].

Global Constraints. Figure 1.5 illustrates the expected versus observed number of exhaustively mined biclusters for the inference of global constraints for the Yeast cycle dataset [619] when considering two degrees of homogeneity given by a loose coherency strength ($|\mathcal{L}|$=3) and stricter coherency strength ($|\mathcal{L}|$=5). Two major observations can be retrieved from this analysis. First, the inference of size constraints are very dependent on the chosen dataset. Illustrative constraints inferred from this analysis include: m=3 ⇒ n>40 and m=4 ⇒ n>15 (assuming $|\mathcal{L}|$=5). As we relax the homogeneity criterion by decreasing the number of items, the size constraints visibly change. Illustrating, under $|\mathcal{L}|$=3, m=4 ⇒ n>70 and m=5 ⇒ n>40 are inferred rules.

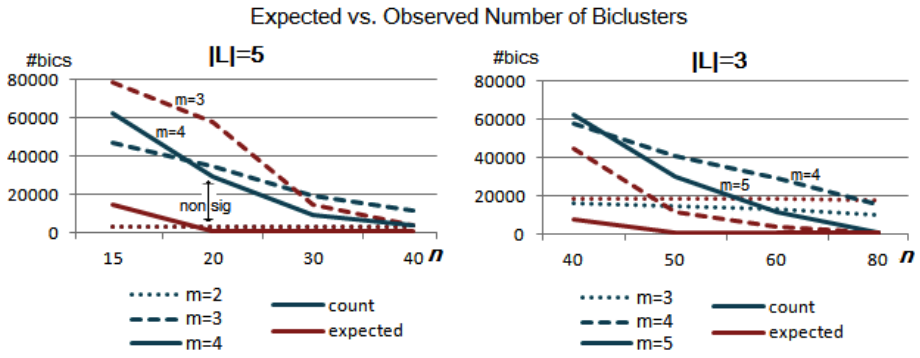

Figure 1.5: Expected vs. observed number of constant biclusters with varying numbers of observations and features for the Yeast cycle dataset with varying coherency strength ($|\mathcal{L}|$=3 and $|\mathcal{L}|$=5).

Soundness. The analysis provided in Table 1.3 shows the relevance of incorporating statistical significance criteria in order to recover only true positive biclusters from synthetic data. For this end, we selected the 1000×100 data setting described in Table II-3.1. However, instead of planting 10 biclusters with the provided distributions for the number of rows and columns, we planted 20 biclusters: 10 biclusters with a pattern, size and length that guarantee their significance and 10 biclusters without statistical significance. In particular, three of the planted significant biclusters are small yet their pattern φ_B highly improbable. Similarly, three of the planted non-significant biclusters

are reasonably large yet the probability of pattern φ_B to occur is above average. In this context, Table 1.3 measures how BicPAM (with default behavior) is able to recover the true positive biclusters only. For this end, BicPAM was applied in the absence of statistical views and in the presence of local and global statistical tests. The results confirm the relevance of using the proposed statistical framework to guide the learning. Two major observations are retrieved. First, the incorporation of the proposed statistical views is critical to guarantee that BicPAM is able to recover the true biclusters. Second, local statistical tests are more robust since global tests in general fail to correctly assess small (yet improbable) and large (yet probable) biclusters. Understandably, this option comes with an additional efficiency cost. The provided results also highlight the relevance of the proposed variant for the application of global tests (*Section 1.3.3.1*) since it minimizes the risk of false positive/negative discoveries.

Option	MS($\mathcal{B},\mathcal{H}$) (coverage of significant bics)	MS($\mathcal{H},\mathcal{B}$) (exclusion of non-significant bics)	Fraction of discovered significant bics	Average number of discovered significant bics	Average number of discovered non-significant bics
No statistical tests	0,97	0,61	52%	10,0	9,3
Local Statistical tests	0,96	0,95	100%	10,0	0,0
Global statistical tests	0,84	0,81	72%	7,3	2,8
Revised global statistics	0,89	0,87	86%	8,9	1,4

Table 1.3: Ability of BicPAM to recover only statistically significant biclusters from a diverse set of 20 planted biclusters (10 significant and 10 non-significant with varying properties) in the presence and absence of local and global statistical tests for 30 data instances of the 1000×100 setting (Table II-3.1).

1.5 Summary of Contributions and Implications

This chapter proposes a consistent set of principles to assess the significance of constant biclustering solutions in discrete settings. To motivate the relevance of this task, we provided a structured overview on the limitations of existing assessments of biclustering models, as well as on the potential contributions from related research on pattern mining and statistics' foundations. To answer this task, three major contributions were proposed.

First, relying on the established mapping between pattern mining and biclustering, we consistently integrate these dispersed statistical principles. In particular, local and global tests with either stochastic and frequentist views with varying dependence degree are dynamically selected based on the properties of the input data. Furthermore, the impact of the data dimensionality on the probability calculus is soundly modeled for biclustering models with either coherency on rows or columns.

Second, a new variant of Huchbert search space correction is proposed to: 1) surpass the efficiency bottlenecks of non-conservation corrections, and 2) minimize the risk of false positive and false negative discoveries.

Finally, global statistical tests are revised to guarantee an adequate inference of constraints to set robust expectations on the minimum number of rows and columns of biclusters according to their observed pattern $\varphi_{\mathbf{B}}$. This extension to minimize the propensity towards type-I and II errors when the frequency of items from a discrete dataset is non-uniform.

Results from synthetic and real data reveal the adequacy of the proposed principles as a robust, simplistic and flexible way to flag false positive biclusters with varying coherency, and to study size constraints that may turn a bicluster significant.

Future Work. This work opens new directions for future work. The proposed principles can be used to define heuristics to guide learning tasks, thus promoting the efficiency of exhaustive searches. Both efficient local tests and the inference of global constraints are critical to narrow the search space. Another relevant direction, tackled by the following sections, is to extend the assessment in order to adequately test biclusters with flexible coherency assumptions, arbitrary levels of noise and continuous ranges of shift and scale factors. A final direction is to revise the proposed statistical tests to provide guarantees on the rate of false negative discoveries.

2

Significance of Biclusters with Flexible Coherencies

Despite the relevance of the contributions proposed in previous chapter to assess the statistical significance of biclustering models, they suffer from three major challenges. First, they cannot be directly applied to assess non-constant biclusters. To our knowledge, there are not yet solid contributions towards this end. Modeling shifts, scales, plaid effects and orders impacts not only the probability of a specific pattern to occur but also the properties of the search space and therefore the applied correction. In this context, there is the need for new statistical tests that guarantee the robust assessment of biclustering solutions with flexible coherency assumptions.

Second, the proposed statistical tests are not prepared to adequately assess biclusters in the presence of arbitrary-high levels of noise and missings. Contrasting with the pattern mining task (from which the previously proposed statistical tests were inspired from), biclustering is by definition prepared to tolerate noise according to the inputted homogeneity criteria. As such, most state-of-the-art biclustering algorithms cannot be assessed using the proposed tests due to variations on the observed pattern of a given bicluster across its observations.

Finally, many biclustering methods output large biclusters with high tolerance to noise in the attempt of bypassing the significance assessment. In fact, such methods commonly guarantee the absence of false positive biclusters with regards to their significance. However, this comes at the cost of a weak homogeneity. Understandably, both homogeneity and significance views are relevant [475], and therefore their jointly optimization is preferred over a strict preference towards one view (causing a detrimental impact on the remaining view).

To address these observations, this chapter extends the statistical assessment framework proposed in *Chapter V-1*. As a result, five major contributions are provided:

- local statistical tests to robustly assess: 1) additive models, 2) multiplicative models, 3) plaid models, 4) order-preserving models and 5) symmetric models;
- principles to guarantee the assessment of biclustering models with arbitrary-high levels of noisy and missing values, including dedicated principles to extend each coherency assumption and detect the underlying assumption when unknown;
- extension of the strategies to infer global constraints towards non-constant models;
- theoretically inferred equations to characterize the impact of flexible coherency assumptions on the properties of the search space;
- integration of significance and homogeneity views for a complete evaluation of biclustering models;

These contributions, visually depicted in Figure 2.1, are integrated within BSig (Biclustering Significance) toolbox, a toolbox for statistical analysis of regions derived from tabular data contexts. The collected results provide empirical evidence for the relevance of these contributions and stress the need to use the underlying principles to affect the discovery of biclusters. Furthermore, we provide the first attempt to compare biclustering algorithms with regards to not only the homogeneity but also the significance of their outputs.

As the background for learning of non-constant biclustering models was provided throughout *Book III*, and the background on the statistical assessment of biclustering models given by *Chapter V-1*, this chapter directly

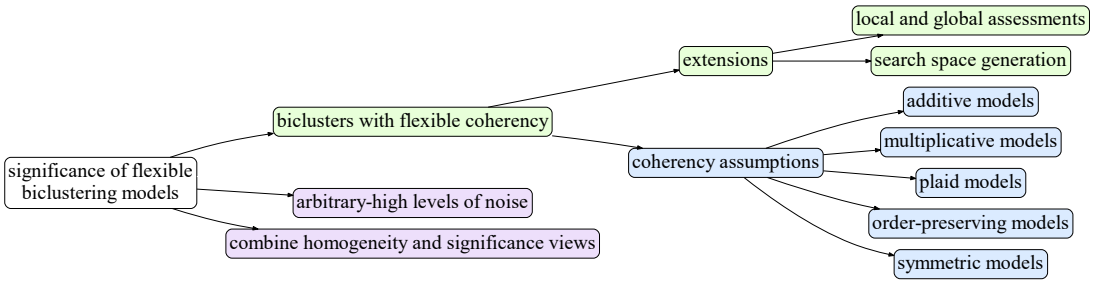

Figure 2.1: Contributions for the assessment of biclustering models with flexible coherency and quality.

proceeds with the solution space[1]. *Section 2.1* extends the proposed statistical tests towards biclusters with flexible coherencies. *Section 2.2* revises the extended tests to adequately assess biclusters with arbitrary levels of noise and missing values, and to provide integrative views of the performance of biclustering algorithms. *Section 6.2* provides the results from the application of the extended tests on synthetic and real data. Finally, we summarize the major contributions of this chapter and pinpoint their implications.

2.1 Solution: Significance of Biclusters with Flexible Coherency Assumptions

In order to extend these principles for non-constant biclusters, two aspects need to be carefully revised: *1)* the $p_{\varphi_\mathbf{B}}$ calculus (affecting the Binomial tail) and, *2)* the new search space (affecting the deviation analysis). Below, we entail these concerns for locally testing biclusters with additive (*Section 2.1.1*), multiplicative (*Section 2.1.2*), plaid (*Section 2.1.3*), order-preserving (*Section 2.1.4*) and symmetric (*Section 2.1.5*) coherency assumptions. Finally, we show how to infer minimum size constraints from globally testing non-constant biclusters (*Section 2.1.6*).

2.1.1 Additive Model

Consider $\mathcal{R}$ to be a finite set of integers with bijective correspondence to the set of items $\mathcal{L}$. A discrete bicluster under an additive assumption is $(\mathbf{I}, \mathbf{J})$ with $a_{ij} = c_j + \gamma_i$, where $c_j \in \mathcal{R}$ is the expected value for feature $\mathbf{y}_j$ and $\beta_j \in \mathcal{R}$ is the adjustment for observation $\mathbf{x}_i$ (shifting factor). Contrasting with constant biclusters, the set of items per observation (row/column) can vary for an additive coherency across features (columns/rows). Given $\mathcal{L}$=$\{0..3\}$, if the items $\{2, 0, 1\}$ are observed as an observation of an additive bicluster, $\{3, 1, 2\}$ (γ=1) can also be observed as an alternative row. The pattern φ_B is given by the underlying items in the absence of shifting factors (γ=0).

In this context, $p_{\varphi_\mathbf{B}}$ thus differs from the constant assumption since new combinations of items are allowed due to the presence of shifting factors. Assuming a correspondence of items to a set of contiguous integers, the allowed number of shifts of an additive pattern $\varphi_\mathbf{B}$ is determined by the difference between the amplitude of values allowed in the input dataset, $\widehat{\mathbf{A}}$, and the amplitude/range of values observed on the bicluster pattern $\widehat{\varphi_\mathbf{B}}$= $max_{a_{ij}}(\varphi_\mathbf{B}) - min_{a_{ij}}(\varphi_\mathbf{B})$. As such, given $\varphi_\mathbf{B}$, the number of allowed shifts, $\widehat{\mathbf{A}} - \widehat{\varphi_\mathbf{B}}$, are used to generate the different combinations of items to compute $p_{\varphi_\mathbf{B}}$:

$$p_{\varphi_B} = \sum_{\gamma=0}^{\widehat{\mathbf{A}}-\widehat{\varphi_\mathbf{B}}} P(\cap_{c_j \in \varphi_B}\{y_j = c_j + \gamma\}) \tag{2.1}$$

where y_j is the random variable following a distribution from the observed feature values $\mathbf{y}_j$. Although the probability of an additive bicluster to occur is greater than a constant bicluster respecting the same pattern φ_B, the search space size of additive biclusters is smaller due to the large number of combinations of items per pattern $\varphi_\mathbf{B}$, and thus the applied correction is subtler and the testing significance level is higher. Both sides of the equation are affected. The search space size, s, of an additive bicluster is defined by:

[1] According to the revised notation, given $\mathbf{X}$ observations, $\mathbf{Y}$ features and a bicluster $\mathbf{B}$=($\mathbf{I}\subseteq\mathbf{X}$,$\mathbf{J}\subseteq\mathbf{Y}$), then N=|$\mathbf{X}$|, M=|$\mathbf{Y}$|, n=|$\mathbf{I}$| and m=|$\mathbf{J}$|.

$$1 + \sum_{i=1}^{m-1} \binom{m}{m-i} + \sum_{d=2}^{|\mathcal{L}|} \left(\sum_{i=1}^{m-1} \binom{m}{m-i} \left(\sum_{j=1}^{i} \binom{i}{j} \times (d-1)^{i-j} \right) \right) \tag{2.2}$$

where m=$|\mathbf{J}|$. The intuition behind this calculus is to count the combination of patterns with different amplitudes. When the amplitude is $AMP(\varphi_{\mathbf{B}})$=0, the m columns have the same item, and only one count is considered. Consider m=4, the patterns {2,2,2,2} and {3,3,3,3} can co-occur as rows of a single additive bicluster. When the amplitude is $\widehat{\varphi_{\mathbf{B}}}$=1, only two items are considered and thus $\Sigma_{i=1}^{n-1}\binom{n}{n-i}$ defines the number of possible arrangements. For m=4, $\binom{4}{3}+\binom{4}{2}+\binom{4}{1}$=14. Finally, when the amplitude is higher, new arrangements are counted assuming that up to m−2 columns can be filled with any item between the maximum and minimum $\varphi_{\mathbf{B}}$ values. Consider m=4 and the amplitude to be 3, then the last parcel of (2.2) is given by $\binom{4}{3}\binom{1}{1}2^0 + \binom{4}{2}(\binom{1}{2}2^0 + \binom{2}{1}2^1)$ + ..., capturing the possible combinations of values with this amplitude. Figure V-2.2 provides an illustrative assessment of an additive bicluster.

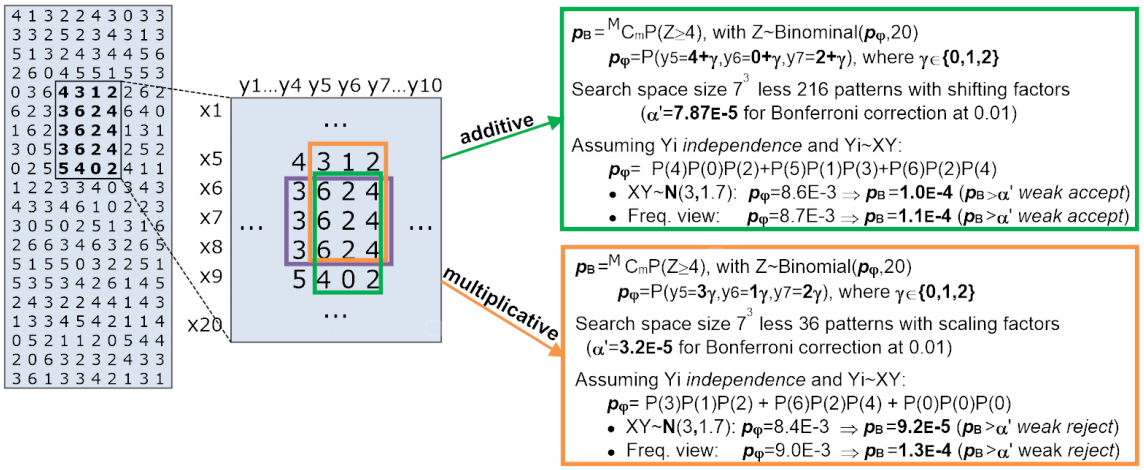

Figure 2.2: Illustrative statistical assessment of (non-noisy) additive and multiplicative biclusters in discrete settings.

2.1.2 Multiplicative Model

Similarly, the multiplicative coherency assumption impacts both the $p_{\varphi_{\mathbf{B}}}$ calculus and the corrected level of significance. Considering a bijective correspondence between a set of integers $\mathcal{R}$ and items $\mathcal{L}$, a discrete multiplicative bicluster is $(\mathbf{I},\mathbf{J})$ with elements $a_{ij} = c_j\gamma_i$, where $a_{ij} \in \mathcal{R}$, $c_j \in \mathcal{R}$ is the expected value for feature $\mathbf{y}_j$ and $\beta_j \in \mathcal{R}$ is the adjustment for observation $\mathbf{x}_i$ (scaling factor). Given $\mathcal{L}$={-6..6}, if the pattern $\varphi_{\mathbf{B}} = \{3,-1,2\}$, is observed within a row of a multiplicative bicluster, three additional combinations of items can be observed ($\gamma \in \{-2,-1,2\}$): {-6,2,-4}, {-3,1,-2} and {6,-2,4}. Contrasting with the additive search space, which essentially depends on the amplitude of values, the multiplicative search space depends on the ratios between all values within a pattern. Scaling factors can be defined by incrementally exploring $\gamma = i$ and $\gamma = 1/i$, where $\gamma \in \mathbb{N} \wedge i \in \mathbb{N} \wedge a_{ij} \in \mathcal{R}$, until all the valid patterns are covered. This generated set patterns are used to compute $p_{\varphi_{\mathbf{B}}}$. Figure 1.2 provides an illustrative assessment of a multiplicative bicluster.

The number of scaling factors (determining the search space of multiplicative biclusters) for a bicluster with m columns is typically less than the number of shifting factors (additive search space) and is given by:

$$\sum_{\varphi \in \mathcal{L}^m} \left(gcd(\varphi) = 1 \right) \tag{2.3}$$

where $gcd(\varphi)$ is the greatest common divisor of the m values of φ pattern. (2.3) provides a naïve calculus based on the generation-and-test of all possible combinations of items. The idea behind this calculus is to remove the patterns with scaling factors $\gamma \neq 1$ (counting the patterns with no scaling factors). If $gcd(\varphi)$ differs from 1, this means that the pattern can be derived from a simpler φ pattern with $gcd(\varphi)$=1 (e.g. $\varphi\prime$={2,-4,4} has a double scale of φ={1,-2,2}). To avoid the combinatorial computational complexity of generating-and-testing all possible combinations of items, corrections can be performed using: 1) a depth-first search to minimize memory usage,

and 2) dynamic programming to avoid redundant computations of the greatest common divisor. This is carried by storing intermediary calculus on the respective nodes of the tree structure.

2.1.3 Plaid Model

Given a finite set of integers $\mathcal{R}$, a plaid model is a composition of biclusters, $a_{ij} = \sum_{k=0}^{K} \theta_{ijk}$ (simplified equation), where $\theta_{ijk} \in \mathcal{R}$ specifies the contribution of bicluster $(\mathbf{I}_k, \mathbf{J}_k)$ for the a_{ij} element (0 if $\mathbf{x}_i \notin \mathbf{I}_k \vee \mathbf{y}_j \notin \mathbf{J}_k$) and $a_{ij} \in \mathcal{R}$. In accordance with Def.III-4.1, θ_{ijk} is sufficiently expressive to model biclusters with constant, additive and multiplicative coherency assumptions. In this context, in order to statistically test a bicluster from a plaid model, two steps are required. First, the original coherency, θ_{ijk}, associated with an observed bicluster needs to be recovered by removing the plaid effects associated with contributions associated with overlapping biclusters. Second, either the statistical tests proposed in previous chapter or the extended tests proposed in previous sections are applied depending on whether the observed bicluster is constant, additive or multiplicative. Illustrating, consider an observation from an additive bicluster with $\{a_{i2}{=}3, a_{i3}{=}5, a_{i5}{=}2\}$ values and contributions $\{a_{i2}{=}2, a_{i3}{=}2, a_{i3}{=}0\}$ from other biclusters, then the $p_{varphi_{\mathbf{B}}}$ calculus and applied correction assume that the underlying pattern is $\varphi_{\mathbf{B}} = \{0, 2, 1\}$.

2.1.4 Order-Preserving Model

Order-preserving biclusters are known for their inherent flexibility (embedding constant, additive and multiplicative models) and relevance across domains (Table III-6.1). The values of the features (columns/rows) included in an order-preserving bicluster define a π linear ordering (monotonically increasing) that is respected across a subset of observations (rows/columns). In this context, a bicluster with m features is described by one of the $m!$ possible linear orderings. As such, and contrasting with previous coherency assumptions, the probability of occurrence and the size of the search space of order-preserving biclusters is well-defined.

The probability of occurrence of a m-length pattern φ_B is $p_{\varphi_B}{=}1/m!$. Interestingly, since every m-length pattern has equal probability of occurrence, the applied correction procedure simply needs to adjust the significance level according to the size of the search space, $m!$. The corrected significance level is thus $\alpha/m!$, the exact (rather than a conservative) estimation.

Based on these observations, global properties, such as the minimum number of θ rows for a bicluster with m columns, can be directly defined. These properties are extracted by checking the ranges of key parameters – the number of rows N of the original matrix and the (n, m)-size of a bicluster – by satisfying the binomial calculus: $P(Z \geq n) < \alpha/m!$, with $Z{\sim}\text{Bin}(1/m!, N)$.

2.1.5 Symmetric Models

A discrete bicluster $(\mathbf{I}, \mathbf{J})$ following a symmetric assumption has symmetries on observations $a_{ij} = c_i \times a'_{ij}$ where $c_i \in \{\text{-}1, 1\}$ is the symmetry factor for each row (or column) and $a'_{ij} \in \mathcal{R}$ is a bicluster element whose value belongs to a finite set of integers and is determined by the underlying coherency assumption (either constant, additive, multiplicative or order-preserving). Illustrating, the patterns $\{2,4,3\}$ and $\{\text{-}1,\text{-}3,\text{-}2\}$ cannot be described by any of the previous assumptions, yet can be described by a symmetric additive model with $\varphi_{\mathbf{B}} = \{0, 2, 1\}$ (with observations respectively modeled by $\gamma_i{=}2 \wedge c_i{=}1$ and $\gamma_i{=}1 \wedge c_i{=}\text{-}1$). Understandably, this symmetric relaxation will also cause the probability of occurrences to be larger, as well as the corrected testing significance level (as a result of a smaller search space size). Similarly to the previous coherencies, $p_{\varphi_{\mathbf{B}}}$ is computed by verifying if an observed combination of items has a valid symmetric and, in this case, by adding the probabilities associated with the symmetric patterns. For non-constant models, this is performed on each scaling or shifting factor associated with $\varphi_{\mathbf{B}}$. In the context of an order-preserving assumption, the binomial calculus underlying the target statistical test is well-defined: $P(Z \geq n) < \alpha/2m!$, with $Z{\sim}\text{Bin}(1/2m!, N)$. The search space of a symmetric bicluster is approximately half of the original search space. For instance, when considering symmetries over constant biclusters, the size of the search space is given by $|\mathcal{L}|^m/2 - 1$.

2.1.6 Global Constraints

In *Section V-1.3.3*, we introduced the possibility to infer global constraints given by expectations on minimum number of rows and columns in a bicluster that approximately guarantees its statistical significance. In the context of a constant model, this was done by estimating the minimum number of rows/columns that guarantees a deviation from expectations (assuming a Poisson test). For this aim, expectations were retrieved from the application of an exhaustive search for constant biclusters on a null/randomized/permuted dataset was proposed. In order to guarantee the extensibility of these tests towards non-constant biclusters, we simply need to guarantee that expectations are given by the exhaustive discovery of non-constant biclusters on a null dataset. This can be easily accomplished using BicPAMS (natively prepared to exhaustively discover biclusters with parameterizable coherency assumption). This enables the application of the Poisson test for a given coherency assumption at a time and thus the retrieval of the minimum size expectations for biclusters with that coherency. Furthermore, the principles proposed in *Section V-1.3.3.1* can be further incorporated to minimize the risks of making a false positive and false negative assessment.

2.2 Solution: Significance Assessment of Biclusters from Noisy and Sparse Matrices

The proposed statistical tests are below extended to enable the assessment of biclusters from noisy data (*Section 2.2.1*) and sparse data (*Section 2.2.2*). Finally, *Section 2.2.3* provides integrative performance views to surpass delineate discrepancies between significance and homogeneity views.

2.2.1 Biclustering Models with Arbitrary-High Levels of Noise

A noisy discrete bicluster is a bicluster where some of its elements $a_{ij} \in \mathbf{B}$ do not respect the overall coherency criteria, $\eta_{ij} \neq 0$. Understandably, when this is the case, the true values (values in the absence of noise) associated with a bicluster become blurry, which hampers the retrieval of the bicluster pattern $\varphi_{\mathbf{B}}$ and can lead to large errors on the assessment of its true statistical significance.

In order to compute the correct probability $p_{\mathbf{B}}$ in these cases, we propose a strategy that aims to identify the true bicluster pattern $\varphi_{\mathbf{B}}$. Figure 2.3 illustrates this strategy.

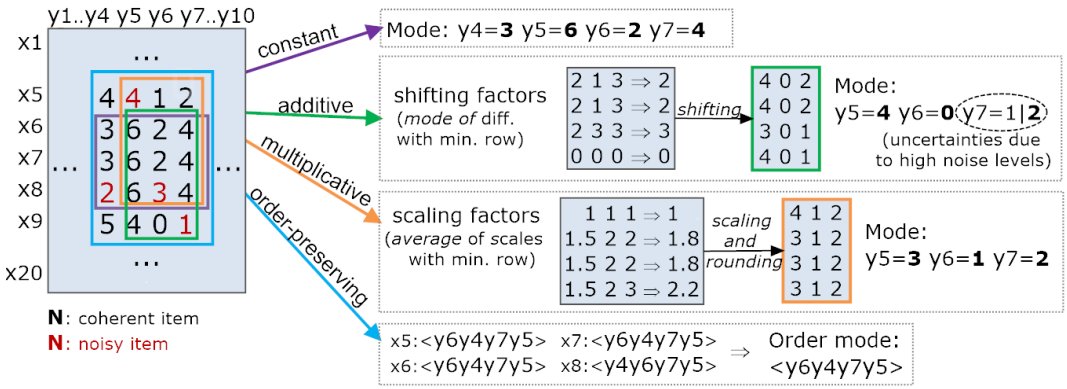

Figure 2.3: Illustrative retrieval of $\varphi_{\mathbf{B}}$ coherence of biclusters in noisy contexts.

The idea behind this strategy is that the levels of noise can be significantly high, yet not sufficient to corrupt expectations on the observed values. For this aim, we propose the mode calculus to retrieve the true value. In the context of a constant bicluster, the true pattern $\varphi_{\mathbf{B}}$ is given by the mode of items per feature $vecy_j \in \mathbf{J}$.

For additive and multiplicative biclusters, this calculus is performed in three steps. First, the shifting or scaling factor associated with each observation in $\mathbf{I}$ are computed by assuming that the observation-conditional values do not have noise. Illustrating, given an additive bicluster with the observed set of items $\{2,3,3\}$ for an observation $\mathbf{x}_i$, a shifting factor γ=2 is assigned to $\mathbf{x}_i$. Second, the computed factors are applied for each observation in $\mathbf{I}$ to retrieve a set of possible bicluster patterns. For the illustrative bicluster, $\varphi_{\mathbf{B}}$=$\{0,1,1\}$ would be the computed pattern for $\mathbf{x}_i$.

Finally, the set of expected patterns form a (noisy) constant bicluster and the mode is applied on each feature to retrieve the true $\varphi_{\mathbf{B}}$ for the $p_{\varphi_{\mathbf{B}}}$ calculus. Under this assumption, the proposed tests in previous section are applied from the inputted bicluster's support, length and expected pattern.

Similarly, symmetric models rely on the identification and removal of symmetric factors (possibly in combination with shifting and scaling factors) in order to identify the true bicluster pattern.

Given a plaid model, three steps are performed: 1) the plaid effects are removed to test separately the contributions of each bicluster, 2) the coherency assumption underlying a given bicluster is identified, and 3) the previous principles to retrieve the true bicluster pattern are applied.

Finally, although assessing the significance of order-preserving biclusters does not depend on the retrieval of the underlying true pattern, *Pointer 2.1* explores how to recover the true permutations of an order-preserving bicluster in noisy data contexts.

Pointers 2.1 Recovering true permutations from noisy order-preserving models

Order-preserving biclusters with arbitrary amounts of noise are associated with a variable number of mismatches on the observed orderings of features per observation. In this context, the true permutations of features can be retrieved by analyzing the most frequent subsets of ordered features (given by maximal sequential patterns) across observations. Consider an order-preserving bicluster with $\{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3\}$ observations associated with $\{(y_2y_3)y_1y_4\}$, $\{y_2(y_1y_3)y_4\}$ and $\{y_3(y_1y_2)y_4\}$ permutations. The maximal ordering (sequential pattern) $\{y_2y_1y_4\}$ is respected in $\mathbf{x}_1$ and $\mathbf{x}_2$ observations and the maximal ordering $\{y_3y_1y_4\}$ satisfied in $\mathbf{x}_1$ and $\mathbf{x}_3$ observations. In this context, $\mathbf{x}_1$ is the observation that supports most of the maximal orderings. Therefore, it is selected as the (mode) observation to determine the properties associated with the true pattern $\varphi_B \vdash$.

2.2.2 Biclustering Models from Sparse Data

Most of existing biclustering algorithms tolerate different forms and amount of noise depending on the input homogeneity criteria. Contrasting, only few biclustering algorithms are able to robustly accommodate missing elements in the outputted biclusters. The problems of imputation methods in the context of the biclustering task were explored in *Chapter III-2*, while the need to analyze sparse data derived from network data motivated in *Chapter III-8*. These contributions open new possibilities for the effective discovery of biclusters from sparse data settings, where a bicluster may be associated with an arbitrary-high number of missing elements. As discussed in *Chapter III-9*, two major factors contribute the sparsity levels of the observed biclusters: 1) structural sparsity either associated with non-imputed missing values (tabular data) or disconnected nodes (network data), and 2) induced sparsity from removing uninformative regions based on background knowledge or user expectations. In this context, the statistical assessment of biclusters with missing elements $a_{ij} = ?$ is increasingly critical.

For this aim, we similarly propose the use of the mode calculus to retrieve the true pattern under the assumption that at least one non-missing value is observed per feature. In this context, missing values are removed from the mode calculus. Illustrating, given an additive bicluster $(\mathbf{I}, \mathbf{J})$, the shifting factors are computed based on non-missing elements for each row $\mathbf{x}_i \in \mathbf{I}$, then the mode calculus is applied for each feature $\mathbf{y}_j \in \mathbf{J}$ to retrieve the expected $\varphi_{\mathbf{B}}$. In the presence of biclusters with both noisy and missing values, this strategy can be consistently combined with principles in previous section.

2.2.3 Combining Significance and Homogeneity Views

Biclustering methods that tolerate high levels of noise tend to deliver large biclusters, often highly significant. In these contexts, significance comes at a cost of the tolerated noise, instead of being associated with a non-noisy deviation from expectations. Understandably, this situation is undesirable since it can mask the true statistical significance of the bicluster. In this context, to guarantee more fair assessments, three strategies can be followed.

First, the analysis of significance should be complemented with the analysis of homogeneity levels.

Second, significance scores can be adjusted by the amount of noise computed using the overall difference from the mode calculus proposed in *Section 2.2.1*. Illustrating, consider a bicluster with 5 observations and 3 features

and a total of 5 elements deviated from the expected true pattern. In this context, $\frac{1}{3}$ of overall elements are noisy and therefore this fraction can be used to weight the significance score. The fraction of identified elements as noisy can be more effectively used to reflect the (noise-sensitive) statistical significance of a given bicluster. For this aim, statistical tests can be proposed based on the conditional analysis of the significance $P(n_{\varphi_\mathbf{B}} \leq X | (\Sigma_{a_{ij} \in \mathbf{B}} \eta_{ij}) < \frac{1}{3})$.

Due to the structural complexity of this statistical test, we propose its simplification based on the Kolmogorov axiom. As such, the p-value given by the probability of a bicluster to deviate from expectations (based on its support, length and expected pattern) can be divided by the p-value associated with the probability of a bicluster has unexpectedly low levels of noise (using contributions from related work [57]).

In this context, if a bicluster has unexpectedly low levels of noise, the noise p-value will be close to 1 and therefore the significance p-value is not affected. Contrasting, if the bicluster tolerates a large amount of noise, an adjustment over the original significance p-value (increasing its value) is observed.

Finally, an alternative analysis is to adjust significance scores by the area of the bicluster in order to benefit (smaller) biclusters with low probable patterns. This score allows the differentiation between approaches that discover smaller (significant) biclusters whose deviation is essentially due to the low p_{φ_B} and approaches that discover larger (significant) biclusters whose deviation is essentially due to the accommodation of noise.

2.3 Results and Discussion

The gathered results are organized in three parts. First, we briefly show how fundamental properties of the search space vary with the coherency assumption and strength. Second, we undertake an in-depth analysis of how the properties of biclustering models and of the input data affect significance. Finally, we provide an inital comparison of state-of-the-art biclustering algorithms with varying coherency assumption according to both the significance and homogeneity of their outputs.

In addition to these results, in *Section V-1.4* we provided initial views on how the significance of non-constant biclustering solutions varied with the applied correction (Figure V-1.3) and selected statistical tests (Figure V-1.4).

Search Space. The impact of the coherency assumption and strength in the search space is illustrated in Figure 2.4. Search space is here given by biclusters with the same pattern length (m=4). Understandably, the order-preserving search space is the most flexible and therefore the most compact. The additive search space increases at a significant lower rate than constant and multiplicative search spaces due to the higher chance of a pattern $\varphi_\mathbf{B}$ to be described by multiple shifting factors. The multiplicative search space is comparable with constant search space due to the low probability of a pattern $\varphi_\mathbf{B}$ being described by a scaling factor $\gamma \neq 1$. In the presence of symmetries, both constant and multiplicative search spaces reduce visibly.

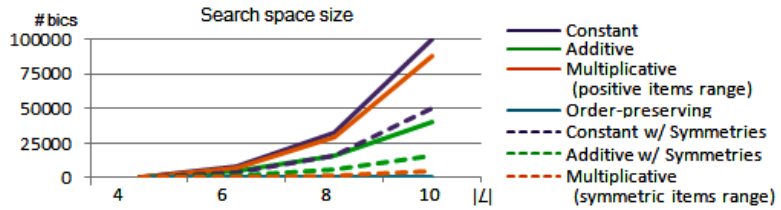

Figure 2.4: Impact of coherency type and strength on the search space size.

Impact of Coherency Assumption, N, M, n, m, $|\mathcal{L}|$ and $\varphi_\mathbf{B}$ in Significance. Tables 2.1, 2.2 and 2.3 extend the analysis provided in Tables V-1.1 and V-1.2 towards non-constant assumptions. In this context, we show how the significance of a bicluster varies: with its support, length, pattern, coherency assumption and strength; and with the size, dimensionality and regularities of the original matrix. For this aim, we show how these different variables affect the expected minimum support of a bicluster required to guarantee its statistical significance.

Table 2.1 explores how different properties affect the significance of order-preserving biclusters assuming that

the search space is given by all the biclusters with same number of columns. As previously discussed, the significance assessment in this context does not directly depends on the data regularities, Θ. The minimum size of biclusters, $n \times m$, can be approximately inferred from the size of the input data, $N \times M$. In particular, the minimum support of a bicluster to guarantee its statistical significance is highly dependent on: 1) the pattern length m – e.g. n=1012 rows for m=4 and n=15 rows for m=8 when N=20000 (and M=100) – and, 2) the number of data observations N – e.g. (n=14,m=6) for N=500 rows and (n=50,m=6) when N=10000 (and M=100).

m	**4**	**5**	**6**	**8**	**10**
N	20000	20000	20000	20000	20000
M	100	100	100	100	100
α'	2.1E-3	4.2E-4	6.9E-5	1.2E-6	1.4E-8
$\mathbf{n}_{min}$	**1012**	**262**	**76**	**15**	**7**
p_B	1.9E-3	3.4E-4	4.0E-5	7.4E-8	4.3E-11

m	6	6	6	6	6
N	**200**	**500**	**2000**	**10000**	**50000**
M	100	100	100	100	100
α'	6.9E-5	6.9E-5	6.9E-5	6.9E-5	6.9E-5
$\mathbf{n}_{min}$	**11**	**14**	**23**	**50**	**141**
p_B	1.2E-5	3.4E-5	4.6E-5	5.9E-5	3.3E-5

m	6	6	6
N	20000	20000	20000
M	**50**	**200**	**1000**
α'	6.9E-5	6.9E-5	6.9E-5
$\mathbf{n}_{min}$	**72**	**79**	**88**
p_B	3.0E-5	4.4E-5	2.9E-5

Table 2.1: Minimum order-preserving observations to guarantee significance across m features for N data size.

The analyzes provided in Tables 2.2 and 2.3 consider two φ_B settings for the additive coherency – one associated with $|\mathcal{L}|$ shifting factors ($\widehat{\varphi_{\mathbf{B}}}$=0) and another with $|\mathcal{L}|$-3 shifting factors ($\widehat{\varphi_{\mathbf{B}}}$=3) – and two scaling factors for $\varphi_{\mathbf{B}}$ under a multiplicative coherency. Generally, we observe that when moving from constant to more flexible coherency assumptions, larger biclusters are required to preserve significance. This is due to the strong effect of these assumptions on the $p_{\varphi_{\mathbf{B}}}$ calculus (even when considering already their larger impact on the applied correction). In particular, additive biclusters tend to be larger than multiplicative peers due to the higher chance of a given pattern $\varphi_{\mathbf{B}}$ to be described by a shifting factor. Assuming m=5, $|\mathcal{L}|$=5, N=20000 and M=100, the expected minimum number of rows $n \in [20, 53]$ ($\hat{n}$=33) for a constant bicluster increases to $n \in [27, 80]$ ($\hat{n}$=47) for multiplicative biclusters with a pattern described by two shifting patterns and up-most to $n \in [42, 151]$ ($\hat{n}$=81) for additive biclusters.

Table 2.2 shows how the minimum number of observations to guarantee significance varies with coherency strength and pattern length. An increase in coherency strength (fewer items/looser homogeneity) requires larger biclusters to guarantee deviations from expectations. Assuming m=5, N=20000 and M=100 and the multiplicative assumption, the expected minimum support $n \in [111, 549]$ ($\hat{n}$=259) for a differential strength ($|L|$=3) decreases visibly to $n \in [27, 80]$ ($\hat{n}$=47) when considering $|\mathcal{L}|$=5. A decrease in pattern length m requires a large increase in the number of observations. Assuming $|\mathcal{L}|$=5, N=20000 and M=100, $\hat{n}$=427, $\hat{n}$=47 and $\hat{n}$=15 are respectively expected for m=3, m=5 and m=7 under a multiplicative coherency, while $\hat{n}$=610, $\hat{n}$=59 and $\hat{n}$=18 are respectively expected for m=3, m=5 and m=7 under an additive coherency with $\widehat{\varphi_{\mathbf{B}}}$=2.

		m	5	5	5	5	**m**	**2**	**3**	**5**	**7**
		N	20000	20000	20000	20000	**N**	20000	20000	20000	20000
		M	100	100	100	100	**M**	100	100	100	100
		$\lvert\mathcal{L}\rvert$	**3**	**4**	**5**	**10**	$\lvert\mathcal{L}\rvert$	5	5	5	5
Additive with $\widehat{\varphi_{\mathbf{B}}}$=0	$p_{\varphi_B}=\frac{1}{\lvert L\rvert}^m \lvert L\rvert$	$\mathbf{n}_{min}$	**362**	**149**	**81**	**21**	$\mathbf{n}_{min}$	**4270**	**982**	**81**	**20**
	$p_{\varphi_B} \in [\frac{0.8}{\lvert L\rvert}^m \lvert L\rvert, \frac{1.2}{\lvert L\rvert}^m \lvert L\rvert]$	$\mathbf{n}_{min}$	[150,784]	[70,299]	[42,151]	[14,31]	$\mathbf{n}_{min}$	[2762,6027]	[525,1580]	[42,151]	[13,32]
	$\alpha'=\alpha/Eq.(2.2)$	α'	2.4E-4	6.4E-5	2.4E-5	1.2E-6	α'	5.6E-3	8.2E-4	2.4E-5	8.1E-7
Additive with $\widehat{\varphi_{\mathbf{B}}}$=2	$p_{\varphi_B}=\frac{1}{\lvert L\rvert}^m (\lvert L\rvert\text{-}3)$	$\mathbf{n}_{min}$	**153**	**92**	**59**	**19**	$\mathbf{n}_{min}$	**2621**	**610**	**59**	**18**
	$p_{\varphi_B} \in [\frac{0.8}{\lvert L\rvert}^m (\lvert L\rvert\text{-}3), \frac{1.2}{\lvert L\rvert}^m (\lvert L\rvert\text{-}3)]$	$\mathbf{n}_{min}$	[70,308]	[46,174]	[32,105]	[14,28]	$\mathbf{n}_{min}$	[1700,3682]	[337,988]	[32,105]	[11,26]
	$\alpha'=\alpha/Eq.(2.2)$	α'	2.4E-4	6.4E-5	2.4E-5	1.2E-6	α'	5.6E-3	8.2E-4	2.4E-5	8.1E-7
Multiplicative with (γ_1,γ_2)	$p_{\varphi_B}=2\frac{1}{\lvert L\rvert}^m$	$\mathbf{n}_{min}$	**259**	**91**	**47**	**12**	$\mathbf{n}_{min}$	**1786**	**427**	**47**	**15**
	$p_{\varphi_B} \in [2\frac{0.8}{\lvert L\rvert}^m, 2\frac{1.2}{\lvert L\rvert}^m]$	$\mathbf{n}_{min}$	[111,549]	[46,173]	[27,80]	[10,16]	$\mathbf{n}_{min}$	[1162,2499]	[240,684]	[27,80]	[10,22]
	$\alpha'=\alpha/Eq.(2.3)$	α'	4.1E-4	1.0E-4	3.5E-5	1.8E-6	α'	5.0E-3	9.8E-4	3.5E-5	1.3E-6
Constant	$p_{\varphi_B}=\frac{1}{\lvert L\rvert}^m$	$\mathbf{n}_{min}$	**153**	**59**	**33**	**11**	$\mathbf{n}_{min}$	**930**	**93**	**33**	**12**
	$p_{\varphi_B} \in [\frac{0.8}{\lvert L\rvert}^m, \frac{1.2}{\lvert L\rvert}^m]$	$\mathbf{n}_{min}$	[70,309]	[32,106]	[20,53]	[7,14]	$\mathbf{n}_{min}$	[619,1304]	[139,375]	[20,53]	[9,18]

Table 2.2: Impact of the number of items, $|\mathcal{L}|$, and pattern length, m, on the expected minimum support that guarantees the significance of constant, additive and multiplicative biclusters for a fixed data size (N and M).

Table 2.3 explores how data size and dimensionality impact the minimum support of a bicluster to guarantee its statistical significance. Assuming $|\mathcal{L}|$=5, m=5 and M=100, the expected minimum number of observations

to guarantee significance is $\hat{n}$=17 and $\hat{n}$=24 for respectively a multiplicative and additive ($\widehat{\varphi_B}$=0) bicluster when N=2000, and largely increases to $\hat{n}$=81 and $\hat{n}$=153 for the same biclusters when N=50000 (magnitude of human genome). This analysis shows that data size largely impacts significance analysis. Similarly, data dimensionality also affects $\hat{n}$. However, the effect of an increase dimensionality on the increase of minimum support is softer than previous variable when targeting coherency across observations. To facilitate the analysis of Tables 2.1-2.3, Figure 2.5 provides a graphical representation of the most relevant results.

		m	5	5	5	5	5	**m**	5	5	5
		N	**200**	**500**	**2000**	**10000**	**50000**	**N**	20000	20000	20000
		M	100	100	100	100	100	**M**	**50**	**200**	**1000**
		$\lvert\mathcal{L}\rvert$	5	5	5	5	5	$\lvert\mathcal{L}\rvert$	5	5	5
Additive with $\widehat{\varphi_B}$=0	$p_{\varphi_B}=\frac{1}{\lvert L\rvert}^m \lvert L\rvert$ $p_{\varphi_B}\in[\frac{0.8}{\lvert L\rvert}^m \lvert L\rvert, \frac{1.2}{\lvert L\rvert}^m \lvert L\rvert]$ $\alpha'=\alpha/Eq.(2.2)$	$\mathbf{n}_{min}$	**11**	**14**	**24**	**53**	**153**	$\mathbf{n}_{min}$	**77**	**85**	**94**
		$\mathbf{n}_{min}$	[8,14]	[10,19]	[16,36]	[30,93]	[71,307]	$\mathbf{n}_{min}$	[39,146]	[44,157]	[50,170]
		α'	2.4E-5	2.4E-5	2.4E-5	2.4E-5	2.4E-5	α'	2.4E-5	2.4E-5	2.4E-5
Additive with $\widehat{\varphi_B}$=2	$p_{\varphi_B}=\frac{1}{\lvert L\rvert}^m (\lvert L\rvert\text{-}3)$ $p_{\varphi_B}\in[\frac{0.8}{\lvert L\rvert}^m (\lvert L\rvert\text{-}3), \frac{1.2}{\lvert L\rvert}^m (\lvert L\rvert\text{-}3)]$ $\alpha'=\alpha/Eq.(2.2)$	$\mathbf{n}_{min}$	**10**	**12**	**19**	**40**	**107**	$\mathbf{n}_{min}$	**56**	**62**	**69**
		$\mathbf{n}_{min}$	[8,12]	[9,16]	[13,28]	[24,67]	[52,105]	$\mathbf{n}_{min}$	[30,101]	[34,110]	[40,120]
		α'	2.4E-5	2.4E-5	2.4E-5	2.4E-5	2.4E-5	α'	2.4E-5	2.4E-5	2.4E-5
Multiplicative with (γ_1,γ_2)	$p_{\varphi_B}=2\frac{1}{\lvert L\rvert}^m$ $p_{\varphi_B}\in[2\frac{0.8}{\lvert L\rvert}^m, 2\frac{1.2}{\lvert L\rvert}^m]$ $\alpha'=\alpha/Eq.(2.3)$	$\mathbf{n}_{min}$	**9**	**11**	**17**	**33**	**81**	$\mathbf{n}_{min}$	**44**	**49**	**54**
		$\mathbf{n}_{min}$	[6,11]	[8,14]	[12,24]	[20,52]	[42,151]	$\mathbf{n}_{min}$	[25,77]	[29,84]	[33,90]
		α'	3.5E-5	3.5E-5	3.5E-5	3.5E-5	3.5E-5	α'	3.5E-5	3.5E-5	3.5E-5
Constant	$p_{\varphi_B}=\frac{1}{\lvert L\rvert}^m$ $p_{\varphi_B}\in[\frac{0.8}{\lvert L\rvert}^m, \frac{1.2}{\lvert L\rvert}^m]$	$\mathbf{n}_{min}$	**8**	**9**	**14**	**24**	**44**	$\mathbf{n}_{min}$	**31**	**35**	**40**
		$\mathbf{n}_{min}$	[5,9]	[7,12]	[10,18]	[16,37]	[30,94]	$\mathbf{n}_{min}$	[19,50]	[23,56]	[27,63]

Table 2.3: Impact of dataset size on the expected minimum support that guarantees the statistical significance of constant, additive and multiplicative biclusters (for a fixed number of items and pattern length).

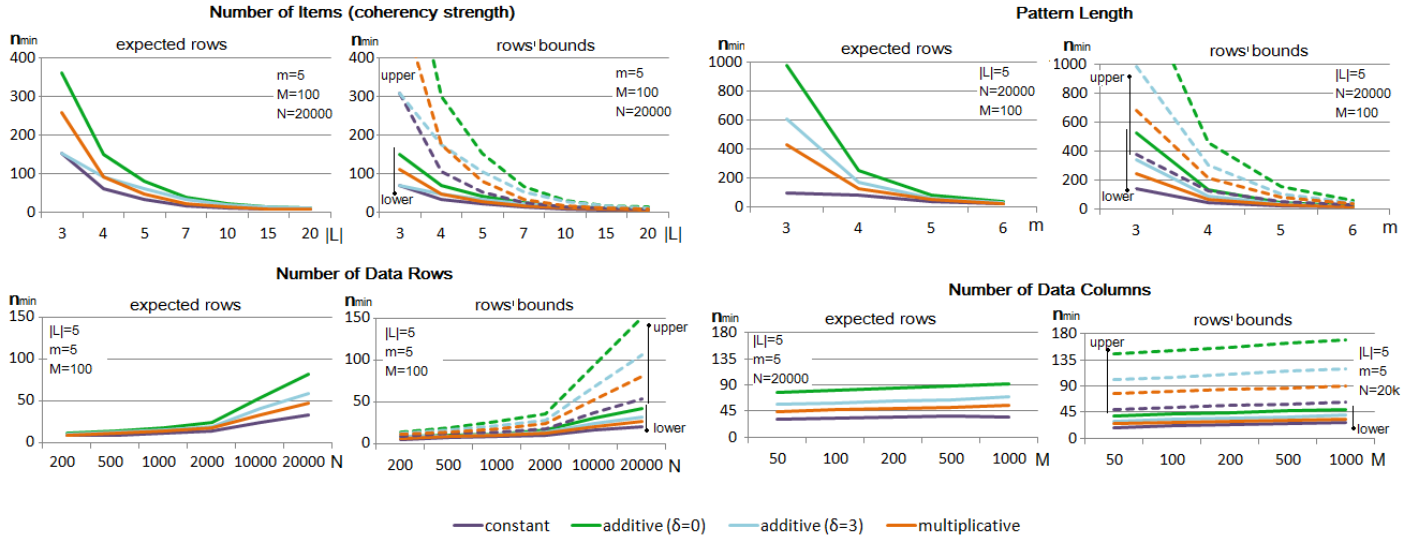

Figure 2.5: Impact of coherency strength, coherency assumption, data size and dimensionality, p_{φ_B} and pattern length on the expected minimum support of a bicluster that guarantees its statistical significance.

Comparing Biclustering Solutions. To test the statistical significance of biclusters discovered by different biclustering methods in real data settings, we selected five state-of-the-art algorithms[2]: FABIA with sparse prior [324] (optimally tuned to discover multiplicative biclusters), ISA [342] (able to discover additive biclusters), CC [134] (for discovering biclusters with flexible yet approximately constant coherence), OPSM [59] (to discover order-preserving biclusters) and BicPAM [310, 311] (able to discover all the previous coherencies). The applied statistical tests were selected according to the proposed integrative assessment and the coherency assumption of each method. In the context of BicPAMS, biclusters are annotated with the underlying coherency assumption for this end. The number of seeds for FABIA, CC and ISA was 10, the number of iterations for OPSM was varied from

[2]To run the experiments, we used: fabia package [324] from R, BicAT software [45] and BicPAMS software [310].

10 to 100. BicPAMS was used with a default parameterization (closed pattern representations, iterative searches with decreasing coherence strength, and a simple merging procedure (70% overlap) and filtering of biclusters overlapping with a larger bicluster on more than 40% of its elements).

Figure 2.6 provides an initial view on their performance with regards to both a significance and homogeneity view of their outputs for two datasets (gene expression of Yeast along a cell cycle [619] and under stress conditions [198]) under a 5-item discretization. Homogeneity was derived from the differences from the expected non-noisy pattern using a simple loss function (normalized mean squared error) between each row of the bicluster and the mode $\varphi_{\mathbf{B}}$ pattern. We can observed that the proposed statistical tests provide a simplistic yet robust tool to study the significance of biclustering algorithms. Three observations can be retrieved. First, CC and Fabia discover biclusters with stronger significance than ISA and BicPAMS (without post-processing options), but that accommodate arbitrary high-levels of noise (looser homogeneity levels using merit functions sensitive to non-constant coherencies [690]). Second, although OPSM implements principles to guarantee the significance of order-preserving solutions [59], some outputted biclusters are still below the adjusted significance threshold. Finally, the use of post-processing options in BicPAMS is associated with a good balance between the significance and homogeneity of the discovered solutions.

This analysis further supports the importance of integrating significance and homogeneity views, and of using the proposed statistical tests (or comparable minimum size expectations) to guide these methods to minimize the risk towards false positive biclusters.

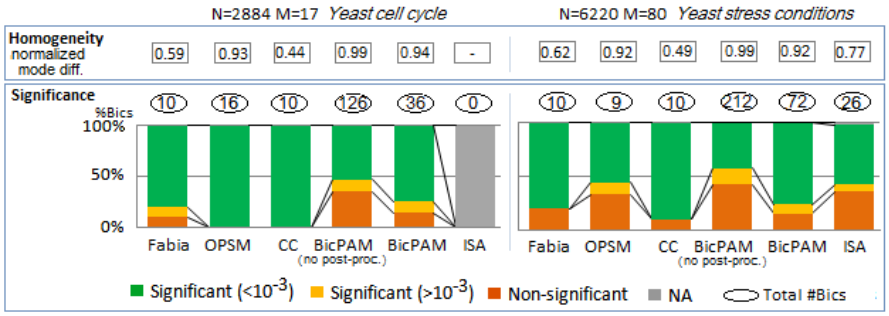

Figure 2.6: Comparison of the statistical significance and homogeneity of biclusters delivered from Fabia, ISA, OPSM, CC and BicPAM over gene expression data.

2.4 Summary of Contributions and Implications

This chapter extended the previous statistical tests for biclusters with varying coherency assumptions and quality. First, we modeled the impact that additive, multiplicative, plaid, order-preserving and symmetric models have on the probability calculus and search space. Second, we extended these principles to enable assessments in noisy contexts where the underlying coherency of a bicluster becomes blurry. Finally, to correctly assess large biclusters discovered at a cost of tolerating large amounts of noise, a new statistical view produced by weighting significance by the amount of noise in a bicluster is proposed.

An experimental analysis was conducted to understand: 1) the effects of coherency assumption and strength on the search space and 2) how significance varies with the support, length, coherency strength, coherency assumption and pattern of a bicluster in relation to the size, dimensionality and regularities of the input data. Furthermore, we conducted an initial comparison of biclustering algorithms that stress both the relevance of revealing their significance and of combining this view with homogeneity.

The contributions of this chapter open a key door for the robust and integrative assessment of flexible biclustering models in discrete data contexts. When analyzing high-dimensional data, this guarantees a decreased propensity towards overfitting by minimizing false positives (removal of non-significant biclusters from the outputs), as well as propensity towards underfitting by minimizing false negatives (recovery of significant biclusters when there is evidence that the outputted set of biclusters is incomplete).

3

Assessing the Significance of Real-valued Biclusters

The previous chapters incrementally described statistical tests to robustly assess the significance of discrete biclusters with flexible coherency and quality. Nevertheless, three prominent problems in state-of-the-art biclustering research remain unaddressed. First, the applicability of these tests towards the analysis of real-valued biclusters is undesirably dependent on the use of discretization procedures, which introduce an additional source of noise that can mask their true significance.

Second, the learning of real-valued biclusters opens new challenges related with the possibility of the underlying adjustment factors, γ (according to Def.II-3.1), having values on a continuous range ($\gamma \in \mathbb{R}$). This possibility contrasts with previous assumptions made in the context of discrete data settings where, for instance, additive or multiplicative models have a limited number of factors.

Finally, up to this point, the proposed statistical tests relied on the assumption that the inputted dataset is given by identically distributed features. As such, these statistical tests are also applicable to matrix and network data. However, they cannot be applied as-is over tabular datasets with non-identically distributed numeric features or, harder, with mixtures of numeric and categoric features. The need for learning biclustering models from these data contexts (and thus to adequately assess their significance) was motivated in *Chapter III-3*.

To address these observations, this chapter extends the statistical assessment framework proposed along the two previous chapters. As a result, five major contributions are provided:

- new estimator of the true significance of real-valued biclusters with (optional) lower and upper bounds on the expected significance;
- effective and efficient assessment of biclusters with continuous shifting factors based on the integral of the product of *slided* density probability functions;
- effective and efficient assessment of biclusters with continuous scaling factors based on the integral of the product of *scaled* density probability functions;
- principles to recover the original coherency strength and assumption of the observed biclusters;
- extension of the assessment for tabular data contexts with non-identically distributed values;

These contributions are synthesized in Figure 3.1. The collected results provide initial empirical evidence for the effectiveness of these contributions and stress the need to use the underlying principles to shape the behavior of biclustering algorithms. We also confront biological significance and statistical, showing that, although correlated, are not always in agreement.

As the background on the statistical assessment of flexible biclustering models was given in *Chapters V-1* and *V-2*, this chapter directly proceeds with the solution space[1]. *Section 3.1* extends the proposed statistical tests towards real-valued biclustering models with continuous adjustment factors. *Section 3.2* summarizes major principles to guarantee the applicability of these tests towards tabular data with varying properties. *Section 3.3* provide results from the application of the extended statistical tests. Finally, the major contributions and implications of this chapter are synthesized.

[1] According to the revised notation, given $\mathbf{X}$ observations, $\mathbf{Y}$ features and a bicluster $\mathbf{B}=(\mathbf{I}\subseteq\mathbf{X},\mathbf{J}\subseteq\mathbf{Y})$, then $N=|\mathbf{X}|$, $M=|\mathbf{Y}|$, $n=|\mathbf{I}|$ and $m=|\mathbf{J}|$.

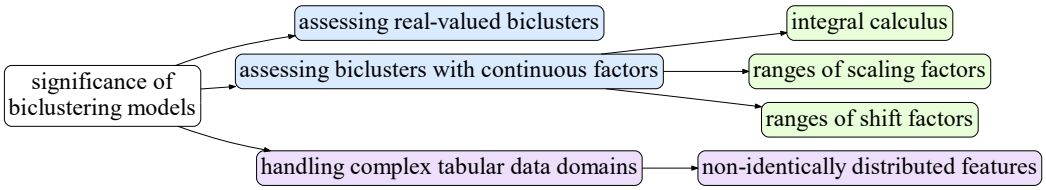

Figure 3.1: Contributions for the assessment of real-valued biclustering models with (possibly) continuous adjustment factors from tabular data.

3.1 Solution: Significance of Real-valued Biclusters

The previous contributions are extensible for real-value data contexts under the identification of the underlying coherency strength, δ, followed by the application of a discretization procedure. Below we describe how this option can be adequately applied in order to minimize the propensity towards discretization drawbacks. In order to address its problems, *Section 3.1.1* proposes an alternative strategy to assess real-valued biclusters. Finally, *Sections 3.1.2* and *3.1.3* extend these procedures with principles from integral calculus in order to guarantee their applicability to the assessment of real-valued biclusters with continuous ranges of shifting and scaling factors.

Baseline Option. A simplistic option towards the analysis of real-valued biclusters is to discretize data and map the discovered regions into discrete biclusters. The mode calculus (*Section V-2.2*) is then applied to deal with both structural noise and the introduced noise from the applied discretization (associated with the item-boundaries problem). The pros and cons of alternative discretization methods, such as equal-depth, bin and distribution-centered methods, have been largely discussed in *Chapters III-1* and *III-2* and throughout literature [120, 430, 500].

To fix an adequate alphabet length, the coherency strength is either given or estimated. The majority of available biclustering methods explicitly define a coherency strength (whether fixed or parameterizable). For the few cases where coherency strength might not be available, it can be approximately inferred from the values of one or more biclusters against the range of values from the input matrix. This is done by analyzing the deviations from the expected values, where the expected values within a bicluster are given by the mean of real-values in the absence of adjustment factors. When the inputted data is associated with ranges of coherent values δ that differ for different biclusters, one of two strategies should be considered. First, a normalization procedure can be applied before the retrieval of a fixed coherency strength. Second, a discretization procedure can be directly applied respecting the identifying ranges (bypassing the need to estimate an adequate coherency strength).

In this context, after retrieving the discretized $\varphi_{\mathbf{B}}$, the calculus of $p_{\mathbf{B}}$ should rely on continuous distributions approximated from the original real-values. The use of a continuous probabilistic view is preferable over a frequentist view to minimize the errors associated with the applied discretization. Figure 3.2a illustrates this strategy using statistical break-points of a Gaussian distribution for data discretization.

However, this strategy suffers from two major problems. First, real values near break-point boundaries can be assigned to different items, leading to more uncertainty when retrieving the mode pattern. Second, shifting and scaling factors can be verified on continuous range of values, and a discretization prevents their adequate assessment.

3.1.1 Assessing Real-Valued Biclusters

To address the first problem associated with this baseline procedure two strategies can be considered. A first strategy is to identically rely on a discretization step, yet to assign the elements with values near break-points to more than one item. The underlying idea is to reduce the influence of these elements during the mode calculus. Accordingly, the proposed mode calculus can be easily revised to either equally weight co-occurring items or to ignore these elements in order to reduce the bias of the mode calculus. Illustrating, consider a bicluster (discretized

using $\mathcal{L}$={-1,0,1}) with the four following observations for feature $\mathbf{y}_2 \in \mathbf{J}$: $\{a_{1,2}=\{1\}, a_{3,2}=\{0,1\}, a_{5,2}=\{1\}, a_{6,2}=\{0,1\}\}$. In this scenario, the mode for this feature can be given by $mode(1,0,1,0)$ in the absence of multi-item assignments and either by $mode(1,1)$ or $mode(1,0,1,1,0,1)$ in the presence of multiple items. Figure 3.2a illustrates this strategy.

A second strategy, the default option, is to rely on multiple probability estimates, one estimate per observation in the bicluster, with the coherency range δ applied around the observed value. For a constant bicluster:

$$\hat{p}_{\varphi_\mathbf{B}} = T(\bigcup_{\mathbf{x}_i \in \mathbf{I}} \{p^i_{\varphi_\mathbf{B}}\}), \quad where \quad p^i_{\varphi_\mathbf{B}} = \prod_{\mathbf{y}_j \in \mathbf{J}} P(a_{ij} - \delta/2 \leq y_j \leq a_{ij} + \delta/2) \tag{3.1}$$

where $p^i_{\varphi_\mathbf{B}}$ is the probability of the set of observed items for i^{th} observation to occur, y_j is the random variable (with distribution drawn from $\mathbf{y}_j$ feature values), and T is the estimator of the true probability from the inputted n estimates. This estimator $\hat{p}_{\varphi_\mathbf{B}}$ s then used for the computation of the Binomial tails in order to compute the estimator of the true probability, $p_\mathbf{B} = \binom{M}{m} \Sigma_{x=n}^{N} \binom{N}{x} (\hat{p}_{\varphi_\mathbf{B}})^x (1 - \hat{p}_{\varphi_\mathbf{B}})^{N-x}$ (where $N=|\mathbf{X}|$, $M=|\mathbf{Y}|$, $n=|\mathbf{I}|$ and $m=|\mathbf{J}|$).

The (3.1) equation is natively prepared to assess constant biclusters, yet it can be consistently extended towards additive, multiplicative, plaid and symmetric models by applying the principles introduced in *Chapter V-2*. In this context, the inputted coherency strength and the observed pattern is used to determine the allowed shifting and scaling factors, which are used to compute $p_{\varphi_\mathbf{B}}$. Note that the previously proposed assessment for order-preserving biclusters remains valid whether permutations are derived from discrete or real-valued data.

Given a set of estimates, the estimator of the true probability, $\hat{p}_B = T(\{p_{\mathbf{B}_1}, p_{\mathbf{B}_2}, .., p_{\mathbf{B}n}\}$ where $n=|\mathbf{I}|$, needs to be adequately defined. We propose the median estimate and percentiles, such as the 15^{th} and 85^{th} percentiles, to model the true probability of occurrence ($p_{\varphi_\mathbf{B}}$ and $p_\mathbf{B}$) with an error bar envelope, providing lower and upper bounds on the significance of a bicluster. The lower and upper bounds can be alternatively seen as conservative and optimistic estimations of the true significance. These estimates are non-biased estimators of the true significance (proof in [102]). Figure 3.2b illustrates this strategy.

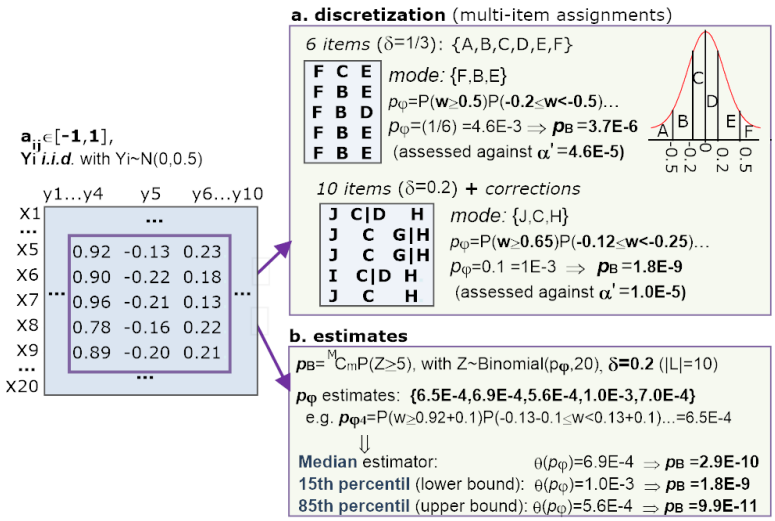

Figure 3.2: Strategies for the extension of the significance assessment for biclusters in real-value settings.

3.1.2 Dealing with Continuous Ranges of Shifting Factors

The drawback of the previous strategies is their inability to model as-is continuous adjustment factors. In previous chapter, a finite number of factors could be generated, $\gamma \in \mathcal{R}$. However, in the context of real-valued data, adjustment factors belong to a real interval, $\gamma \in \mathbb{R}$. Illustrating, consider the $\{1.3, 2.2, 1.7\}$ combination of values, a continuous range of coherent values under additive or multiplicative assumptions can be generated based on the exploration of γ factors (e.g. shifting γ∈[-1.3, 1.8] or scaling γ∈[0, 1.8] factors for values $a_{ij} \in [0,4]$).

In order to compute an adequate level for the p_{φ_B} probability of additive and multiplicative models, and subsequently of symmetric and plaid models (with underlying additive/multiplicative assumptions), we propose a

technique based on the integration calculus of intervals and on the multiplication of density mass functions.

Consider the additive coherency assumption. Let the maximum and minimum observed values for a particular observation $\mathbf{x}_i \in \mathbf{I}$ of a bicluster to be, respectively, $max_{\mathbf{J}|\mathbf{x}_i}$ and $min_{\mathbf{J}|\mathbf{x}_i}$. Also consider the range of real-values of the matrix $\mathbf{A}$ to be $[min_{\mathbf{A}}, max_{\mathbf{A}}]$. Then for a particular pattern $\mathbf{J}|\mathbf{x}_i$ the shifting factors are defined by the interval $\gamma \in [\gamma_1 = -(min_{\mathbf{A}} - min_{\mathbf{J}|\mathbf{x}_i}), \gamma_2 = max_{\mathbf{A}} - max_{\mathbf{J}|\mathbf{x}_i}]$. The probability of a particular value a_{ij} to occur under this shifting interval is:

$$\int_{a_{ij}+\gamma_1}^{a_{ij}+\gamma_2} f(x) = \int_{\gamma_1}^{\gamma_2} f(x + a_{ij}) \tag{3.2}$$

where $f(x)$ is the distribution function that approximates a_{ij} values. This calculus assumes that the range of observed values $\widehat{\mathbf{A}}$ are linearly adjusted to guarantee an unitary coherency strength $\delta \approx 1$.

The probability of two values a_{ij} (a_1) and $a_{i(j+1)}$ (a_2) to occur under this shifting interval is not simply the product of their individual probabilities since a simple product would allow for non-coherent values (e.g. $\{a_1+\gamma_1, a_2+\gamma_2/2\}$).

In order to correctly account for the combination of values with continuous shifting ranges, the distribution functions need to be aligned by the target column value and multiplied. The resulting function delivers the product of the individual probabilities. Finally, the area behind this curve between γ_1 and γ_2 values is computed in order to retrieve a estimate of the probability p_{φ_B} for the Binomial tail calculus. This strategy is illustrated in Figure 3.3, under the assumption that the values of the $\mathbf{A}$ matrix are either described by a single Uniform or Gaussian distribution.

Given $\varphi_{\mathbf{B}(i)} = \{a_{i1}, .., a_{im}\}$ combination of values for i row, $p_{\varphi_{\mathbf{B}(i)}}$ can be approximated by:

$$\int_{\gamma_1}^{\gamma_2} \Pi_{j=1}^{m} f(x + a_{ij}) \tag{3.3}$$

In order to compute this probability efficiently we propose the calculus of its approximate area by interpolating 100 points between γ_1 and γ_2.

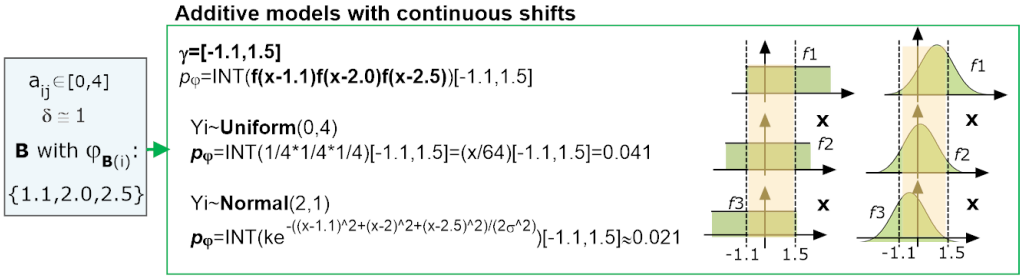

Figure 3.3: Illustrative integral of the product of slided density functions to assess biclusters with continuous ranges of shifting factors.

3.1.3 Dealing with Continuous Ranges of Scaling Factors

The probability of occurrence of a combination of real values $\varphi_{\mathbf{B}(i)}$ on the i^{th} observation of a bicluster under a multiplicative coherency across observations can be approximated using similar principles to the ones proposed in previous section. Considering max_i and min_i to be the maximum and minimum real values for a given observation I_i and r to be the range real values in $\mathbf{A}$. When only positive values are allowed, the scaling range is $[\gamma_1 = 0, \gamma_2 = r/max_i]$. When negatives values are allowed the scaling range is given by $[\gamma_1 = -d, \gamma_2 = d]$ where $d = max(max_i, min_i)/(r/2)$.

The probability of multiple values to occur is given by the integral of the product of the size-adjusted density functions for the $[\gamma_1, \gamma_2]$ interval.

Why the size adjustment is necessary? Consider the pair of observed values $\{a_1 = 1, a_2 = 2.5\}$ and the scaling range to be $\gamma \in [0, 1]$. This means that the density function to estimate the a_1 value is considered for the interval

[0,1], while the density function to estimate a_2 is considered over [0,2.5]. Therefore, the density functions need to be normalized with regards to their size: $f(x/a_1)$ and $f(x/a_2)$.

Given $\varphi_{\mathbf{B}(i)}=\{a_{i1},..,a_{im}\}$ combination of values for i row, $p_{\varphi_{\mathbf{B}(i)}}$ can be approximated by:

$$\int_{c_1}^{c_2} \Pi_{i=1}^{n} f(x/a_i) \tag{3.4}$$

Similarly, an efficient computation of the (3.4) integral calculus is made available recurring to interpolation whenever the multiplication of the inputted density functions is complex.

This strategy is illustrated in Figure 3.4, under the assumption that the values of the A matrix are either described by a single Uniform or Gaussian distribution.

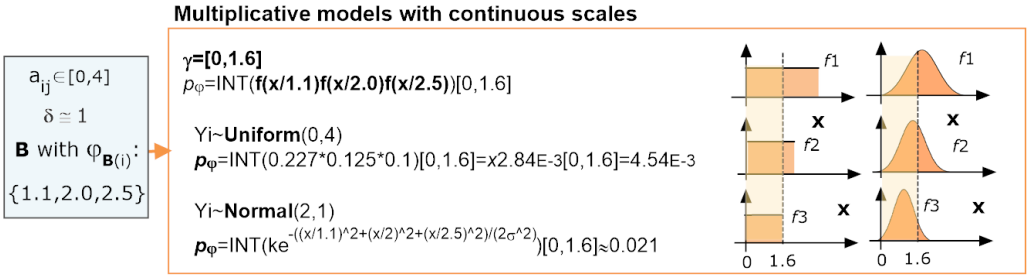

Figure 3.4: Illustrative integral of the product of scaled density functions to assess biclusters with continuous ranges of scaling factors.

3.2 Solution: Tabular Data with Non-Identically Distributed Features

The proposed assessments throughout this and previous chapters are applicable towards tabular data with identically distributed attributes. In particular, the proposed contributions are applicable to an either symbolic or real-valued network dataset under the mapping proposed in *Chapter III-8* (where two matrices are derived from the corresponding minimal bipartite graph) and using the principles to assess biclusters from sparse data contexts proposed in *Section III-9.2*. Nevertheless, high-dimensional tabular data can have non-identically distributed features with possibly distinct domains. In this context, *Section III-3.2.3* extended biclustering to learn from these tabular data contexts. To be able to assess these models, we propose the extension of the statistical assessment framework towards non-trivial tabular datasets. For this aim, we introduce a method based on three steps. First, the distribution of values per feature are approximated using either stochastic or frequentist views depending on whether the feature domain is numeric or categoric. Second, the median/mean calculus is used to retrieve the expected true pattern $\varphi_{\mathbf{B}}$. Third, the probability of this pattern to occur, $p_{\varphi_{\mathbf{B}}}$, is directly inferred from joint probability. That is, $p_{\varphi_{\mathbf{B}}}=P(\cup_{\mathbf{y}_j\in\mathbf{B}}\{c_j \mid \mathcal{Y}_j\})$. Finally, the binomial tails are computed, $p'_{\mathbf{B}}=\Sigma_{x=n}^{N}\binom{N}{x}(\hat{p}_{\varphi_{\mathbf{B}}})^x(1-\hat{p}_{\varphi_{\mathbf{B}}})^{N-x}$.

Despite the consistency of these calculus, two key challenges remain unanswered. First, the impact of dimensionality is not modeled by the previous process. Understandably, the probability of a given bicluster with specific size to occur in a dataset with few features differs from its probability to occur in a dataset with additional features. However, determining the exact effect of dimensionality is computationally complex for tabular data. Illustrating, consider a bicluster $\mathbf{B}$ with two categoric features $\{\mathbf{y}_2,\mathbf{y_4}\}\subseteq\mathbf{Y}$ (where $|\mathcal{Y}_2|=2$, $|\mathcal{Y}_4|=5$) following a discrete Uniform distribution from a dataset with $M=20$ features. In order to model the probability of a similar bicluster with same pattern length ($|\mathbf{J}|=2$) to be discovered for any other pair of features, the conditional probability associated with each pair of features needs to be adequately weighted against $p_{\varphi_{\mathbf{B}}}=\frac{1}{2}\frac{1}{5}$. In the presence of numeric features, the underlying coherency strength needs to be estimated (see *Section 3.1*). In these contexts, modeling the impact of dimensionality is only computationally feasible for datasets with a small set of features. As such, for high-dimensional data, approximations can be considered, being $p'_{\varphi_{\mathbf{B}}}\leq p_{\mathbf{B}}\leq\binom{m}{M}p'_{\varphi_{\mathbf{B}}}$ looser bounds of the dimensionality impact.

Second, the properties of the search space (determining the applied correction) are not defined. In the presence of a constant bicluster, the product of the cardinality of the features in the bicluster can be used to estimate the search space size. The cardinality of a feature is either given by the number of ordinal/nominal items when categoric or by the range of accepted values divided by coherency strength when numeric. Considering the illustrative bicluster from previous paragraph, the space of similar bicluster is approximately given by $|\mathcal{Y}_2| \times |\mathcal{Y}_2| = 4$.

In the presence of an order-preserving bicluster, the search space is well-defined and its size given by $m!$.

Considering now the remaining coherencies (additive, multiplicative, symmetric and plaid), these are only applicable for (possibly non-identically distributed) numeric features. In these contexts, the statistical tests provided in this chapter can be used.

Finally, *Section III-3.2.3* introduced the notion of a mixture of coherencies when a bicluster is composed by a subset of numeric features and a subset of categoric features. Consider $\varphi_{\mathbf{B}|\mathbb{R}}$ and $\varphi_{\mathbf{B}|\mathcal{L}}$ to be a division of the pattern according to the observed values in the subset of numeric and categoric features. Illustrating, $\varphi_{\mathbf{B}} = \{2.1, 3.0, \mathrm{B}, \mathrm{A}\}$, where $\varphi_{\mathbf{B}|\mathbb{R}} = \{2.1+\gamma, 3.0+\gamma\}$ and $\varphi_{\mathbf{B}|\mathcal{L}} = \{\mathrm{B}, \mathrm{A}\}$. In this context, the probability of the pattern to occur, $p_{\varphi_{\mathbf{B}}}$, can be simply given by the product $p_{\varphi_{\mathbf{B}|\mathbb{R}}} \times p_{\varphi_{\mathbf{B}|\mathcal{L}}}$ where $\varphi_{\mathbf{B}|\mathbb{R}}$ follows a (possibly) additive/multiplicative/symmetric/plaid model (given by principles introduced earlier in this chapter) and $\varphi_{\mathbf{B}|\mathcal{L}}$ necessarily follows a constant model.

In order to better approximate the effects of dimensionality and search space on the computed statistical significance, additional strategies can be consider. Illustrative strategies include balancing the cardinality of categoric features and normalizing the distribution of values of numeric features (see *Section III-3.2.3*). In a tabular data context where the distribution and coherency strength of categoric and numeric features is identical, the suggested effects to model the impact of dimensionality, $\binom{m}{M}$, and search space size, $|\mathcal{Y}|^m$, are in fact exact estimations.

3.3 Results and Discussion

Results are organized as follows. First, we assess the relevance of modeling continuous ranges of factors when modeling non-constant biclustering models. Second, we provide brief evidence of the relevance of bounding significance of real-valued biclusters derived from noisy data contexts. Finally, we present significance views collected for tabular data with non-identically distributed features. The statistical tests proposed in this and previous chapters are included in BSig (Biclusters Significance) toolbox, implemented in Java (JVM v1.6.0-24). Experiments were computed using an Intel Core i5 2.30GHz with 6GB of RAM.

Continuous Ranges of Shifting and Scaling Factors. Table 3.1 shows how the required minimum support to guarantee the statistical significance of a real-valued bicluster with continuous shifts/scales varies with the number of observations, N, and features, M, of the input dataset. For this analysis we applied the introduced integral analysis to collect a robust statistical estimation of $p_{\varphi_{\mathbf{B}}}$ (Table 3.2). Two major observations can be retrieved. Both the size and dimensionality of data affect the significance levels, being the effect of varying the size of data clearly more accentuated since the assessment was applied over biclusters with coherency across rows. Second, the observed pattern also largely determines the computed significance levels as it determines the range of allowed shifts and scales. Understandably, the larger the allowed range, the higher is the probability of a bicluster pattern to occur and thus the higher is the number of minimum rows in the bicluster to guarantee its significance.

Noisy Real-valued Biclusters. Two strategies were proposed for assessing the significance of noisy biclusters in real-valued matrices: *1)* retrieving the $\varphi_{\mathbf{B}}$ pattern by using the mode (discretized data) or median (real-valued data) of the bicluster's values on columns to compute $p_{\mathbf{B}}$, or *2)* rely on a estimator of row estimates. We undertook an experimental analysis where we applied some of the state-of-the-art biclustering methods over gene expression data to produce noisy biclusters with varying properties. Figure 3.5 illustrates the median, 15-percentile and 85-percentile estimators of the true significance of nine illustrative biclusters gathered from the application of three

			N	**200**	**500**	**2000**	**10000**	10000	10000	10000
			M	100	100	100	100	**50**	**400**	**1000**
m=3	Multiplicative	$\gamma \in [0, 0.2]$	n_{min}	7	10	18	42	39	45	48
		$\gamma \in [0, 0.5]$	n_{min}	9	12	23	58	55	62	66
		$\gamma \in [0, 1]$	n_{min}	14	21	44	137	132	144	150
	Additive	$\gamma \in [0, 0.2]$	n_{min}	8	11	19	44	41	48	51
		$\gamma \in [0, 0.5]$	n_{min}	11	15	31	82	78	87	91
		$\gamma \in [0, 1]$	n_{min}	14	21	44	137	132	144	150
m=5	Multiplicative	$\gamma \in [0, 0.2]$	n_{min}	4	5	6	8	7	10	11
		$\gamma \in [0, 0.5]$	n_{min}	5	6	8	12	10	14	16
		$\gamma \in [0, 1]$	n_{min}	7	9	13	23	21	27	30
	Additive	$\gamma \in [0, 0.2]$	n_{min}	6	7	9	13	11	15	17
		$\gamma \in [0, 0.5]$	n_{min}	7	8	11	18	16	20	23
		$\gamma \in [0, 1]$	n_{min}	7	9	13	23	21	27	30

Table 3.1: Impact of data size and dimensionality on the expected minimum number of observations in biclusters with continuous adjustment factors to guarantee their statistical significance (assuming a δ=0.2 coherency strength, uniform background values, and additive and multiplicative coherencies with varying ranges of allowed shifts/scales).

		m=3			m=5		
		$\gamma \in [0, 1]$ φ_B={0.4,0.4,0.4}	$\gamma \in [0, 0.5]$ φ_B={0.5,1,0.6}	$\gamma \in [0, 0.2]$ φ_B={0.2,0.9,1}	$\gamma \in [0, 1]$ φ_B={0.4,0.4,0.4,0.4,0.4}	$\gamma \in [0, 0.5]$ φ_B={0.5,1,0.6,0.8,0.7}	$\gamma \in [0, 0.2]$ φ_B={0.2,0.9,1,0.3,0.6}
Multiplicative p_{φ_B}	δ=0.2 $\delta \in [0.1, 0.3]$	8.0E-3 [1.0E-3,2.7E-2]	2.4E-3 [3.0E-4,8.1E-3]	1.4E-3 [1.8E-4,4.9E-3]	3.2E-4 [1.0E-5,2.4E-3]	5.4E-5 [1.7E-6,4.1E-4]	1.0E-5 [3.2E-7,7.9E-5]
Additive p_{φ_B}	δ=0.2 $\delta \in [0.1, 0.3]$	8.0E-3 [1.0E-3,2.7E-2]	4.0E-3 [5.0E-4,1.4E-2]	1.6E-3 [2.0E-4,5.4E-3]	3.2E-4 [1.0E-5,2.4E-3]	1.6E-4 [5.0E-6,1.2E-3]	6.4E-5 [2.0E-6,4.9E-4]

Table 3.2: Expected probability of different patterns to occur in biclusters with continuous shifts and scales from data with approximately uniform distribution of values (a_{ij}∈[0,1]).

biclustering methods – CC [134], ISA [342] and Fabia [324] – over Yeast dataset [619] under a coherency strength $\delta = \widehat{\mathbf{A}} /5$. CC, ISA and Fabia were selected as they are prepared to discover biclusters with flexible coherency assumptions. For simplicity sake, we excluded the significance levels obtained under the first strategy from this analysis as they were close to the median estimator. This analysis reveals the importance of bounding significance whenever possible. Interestingly, although the absolute significance of the ISA biclusters is worse than the remaining methods, the variance of ISA estimates is approximately null. This is because ISA guarantees a strong homogeneity that penalizes the discovery of large biclusters. Contrasting, Fabia and CC methods provide more significant biclusters at a cost of allowing increased levels of noise.

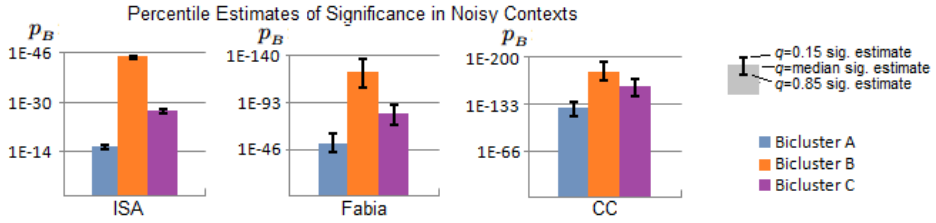

Figure 3.5: Median, 15-percentile and 85-percentile *significance* of an illustrative set of biclusters derived from the application of three different biclustering methods over the Yeast cycle dataset.

Statistical vs. Biological Significance. To understand whether the statistical significance of a biclustering model is correlated with its biological significance, we conducted two analyzes provided in Table 3.3 and Figure 3.6. Table 3.3 shows how the percentage of biclusters with enriched gene ontology terms[2] varies for non-significant versus significant biclusters (using the statistical tests proposed in this chapter). For this analysis, we applied BicPAM with default behavior on three expression datasets (description in *Section III-2.7.3*): *dlblc* dataset (660 genes, 180 conditions) [560], *hughes* dataset (6300 genes, 300 conditions) [395], and *gasch* dataset (6152 genes, 176 conditions) [239]. From the gathered results, we observe that both the percentage of enriched biclusters and their average number of terms increases for biclustering models with guarantees of statistical significance. Although it appears that statistical and biological significance are correlated, their correlation can be spurious if it is primarily explained by a third variable: the size of biclusters. The results show that solutions with guarantees of

[2]According to the domain-driven analysis described in *Section II-2.2*.

statistical significance tend to deliver larger biclusters. Also, a commonly well-know drawback of term enrichment analysis is its bias to favor large (bi)clusters [458, 578]. However, we can observed that no biases have been incurred since there were no significant differences in the number of genes/rows between statistically significant and non-significant (the differences in size were primarily explained by the number of conditions/columns).

To complement this analysis, Figure 3.6 plots all the biclusters discovered by BicPAM in the *gasch* dataset according to their statistical significance and number of enriched terms (biological significance). Similarly to the previous analysis, this analysis suggests that, although statistical and biological significance are not strongly correlated, there is a soft emerging trend. These analyzes pinpoint the importance of considering both views to evaluate and (possibly) guide biclustering algorithms.

Dataset	Filtering criteria	♯Bics	Average $\|I\| \times \|J\|$	%Enriched bics
Gasch	Significant	93	409×9	85%
	Non-significant	56	430×6	79%
	All	149	411×8	83%
Dlblc	Significant	29	89×7	56%
	Non-significant	27	65×5	44%
	All	56	83×7	50%
Hughes	Significant	31	343×8	81%
	Non-significant	16	389×6	69%
	All	47	360×7	77%

Table 3.3: Characterizing the biclusters found by BicPAM (default behavior) over gasch, dlblc and hughes data according to their: statistical significance (fraction of significant biclusters), size, and biological significance (enriched biclusters).

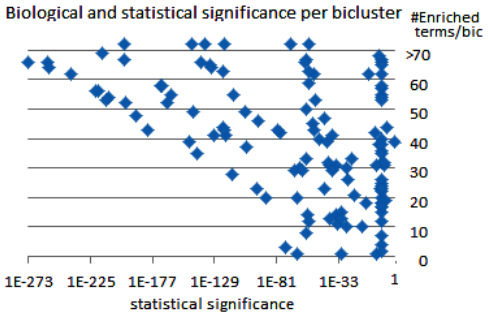

Figure 3.6: Correlation between biological and statistical significance: number of enriched terms per bicluster versus statistical significance for the biclustering output from applying BicPAM over the *gasch* dataset [238].

3.4 Summary of Contributions and Implications

This chapter extended the statistical assessment of biclustering models in real-valued data settings, proposing estimators to bound the significance of biclusters. For the adequate modeling of continuous ranges of shifting and scaling factors, new tests based on the integration of the product of the shifted/scaled distributions on these ranges are proposed. These statistical tests enable the assessment of real-valued additive and multiplicative models and, consequently, of plaid models. Finally, BSig was extended to tabular data contexts possibly characterized by a combination of nominal, ordinal and numeric features non-identically distributed.

Results support these contributions, and confront statistical significance with biological significance, showing that they are not always in agreement (stressing the relevance of the proposed statistical views).

Built upon the contributions from previous chapters, the proposed integrative assessments provide the unique opportunity to: 1) effectively and efficiently assess the significance of flexible biclustering models learned from high-dimensional data contexts (minimizing the risk of false positives/negatives and thus data over/underfitting propensity), 2) handle advanced learning aspects including non-trivial coherency assumptions from (possibly) sparse, noisy and non-identically distributed features, and 3) validate the increasing number of scientific implications made from biomedical and social local data analysis without a proper statistical ground truth.

4

Significance of Local Descriptive Models Learned from Structured Data

The statistical views proposed in previous chapters can be used to ensure the significance of local descriptive models from tabular data. Yet, these views are not generalizable for structured data due to the inherent temporal nature of its regions. These regions are given by cascades and arrangements of events, whose relevance was throughout *Chapters IV-1* and *IV-2* seen as a measure of their support (number of observations). However, the support model has no mechanism to eliminate regions that occur simply by chance. When observations are high-dimensional, short cascades or compact arrangements can occur frequently by chance. Consequently, support alone cannot distinguish between statistically significant and spurious regions.

To address this problem, this chapter proposes a sound and efficient approach to mine statistically significant regions against a null model to describe the input data. This is done by first seeing these regions as temporal patterns and by seeing their support as a random variable, whose distribution under the null model is approximated. Then, the statistical significance of these patterns is assessed by testing the observed frequency against the expected frequency. In this context, four major contributions result from this chapter:

- statistical tests to assess the significance of sequential patterns from itemset sequences and evidence of their extensibility to assess temporal patterns and, ultimately, cascades and arrangements of events;
- principles to guarantee an adequate deviation from expectations (principles to guarantee the applicability of Hochbert corrections for large outputs and to simultaneously control the rate of false positives and negatives);
- statistical tests to assess regions encoded within probabilistic models, including complex Markov models;
- incorporation of the previous statistical views to guide the learning methods: 1) use of local statistical tests to monotonically prune the search space, 2) (optional) inference of global constraints from deviations on the number of expected occurrences, and 3) use of statistical tests on probabilistic models to condition the learning and guide the decoding step.

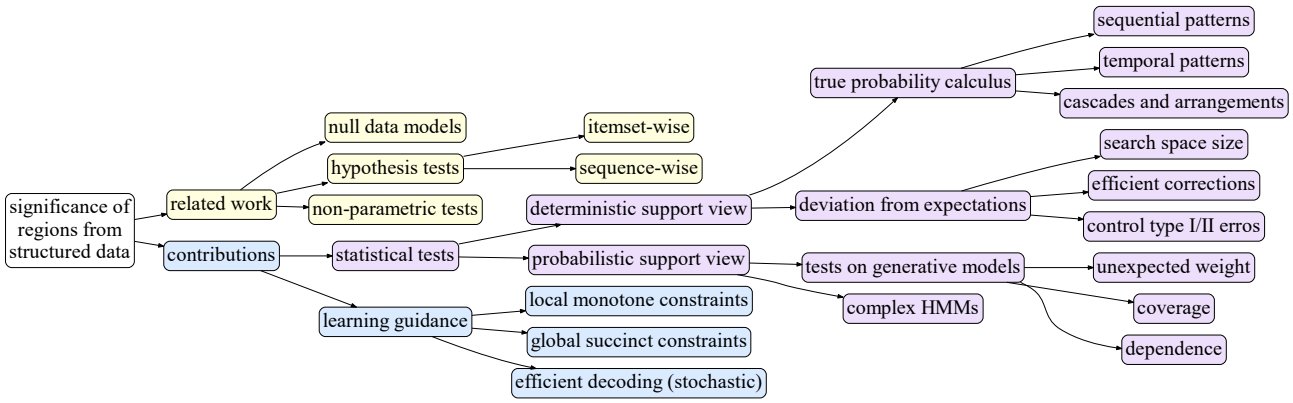

Figure 4.1: Contributions to assess the statistical significance of regions from structured data.

Figure provides a structured view on the proposed contributions. This chapter is structured as follows. *Section 4.1* provides the background on the target task. *Section 4.2* surveys related work for the statistical assessment of

patterns from temporal data. *Section 4.3* describes the solution space. Finally, this chapter provides a discussion on the soundness of the proposed contributions and points out relevant directions for future work.

4.1 Background

From a statistical point of view, the discovery of regions (either from tabular or structured data) is always achieved on a finite sample of observations drawn from an unknown target distribution $P_{\mathcal{X}}$. The ability to approximate $P_{\mathcal{X}}$ is particularly challenge for high-dimensional data with a small sample of observations. The smaller the sample, the higher the risk to have a large statistical bias. Yet, the existing algorithms for region discovery from structured data [424, 309, 21] are almost always applied without any consideration of the underlying statistical data distribution. In other words, these algorithms commonly do not assess the likelihood that an extracted region is a spurious byproduct of the sampling rather than a consistent pattern in the target distribution.

Let us explore why this is so. The learning of local descriptive models from structured data has been mainly driven by the discovery of temporal patterns (including chords, sequential patterns, episodes [467, 470, 321]) according to frequency criteria. The use of a minimum support threshold to identify frequent temporal patterns (regions) is in fact driven by the desire to distinguish true patterns from those appearing by chance. Yet many regions can occur frequently by chance, particularly if: the pattern is short, the number of observations is high, and the pattern values appear frequently in the database. When the minimum support is set low, the support of some patterns can satisfy this threshold simply by chance (false positives). Contrasting, the use of high support thresholds may result in missing statistically significant patterns (false negatives). In other words, although false regions can be extracted by chance (referred as false positives) and true regions can be missed due to the sampling (referred as false negatives), existing learning algorithms do little to control the risk of false discoveries.

To simplify the statistical assessment of the target regions – cascades and arrangements –, both can be seen as temporal patterns from temporally-enriched itemset sequences (according to Def.IV-2.3). In this context, we hypothesize that their statistical assessment is similar to existing statistical views of sequential patterns from itemset sequences. Nevertheless, despite the support model has been the basis for sequential pattern mining [424, 700, 523], two limitations remain. First, there is a lack of contributions on how to test the observed support against expectations. As previously motivated, support alone cannot distinguish between statistically significant patterns and random occurrences. Second, these methods mine sequential patterns with exact matching, which are vulnerable to noise in the data. Contrasting, cascades and arrangements of events are founded on noise-tolerant temporal patterns. In this context, the underlying problem targeted by this chapter is the formulation of a statistical view of SPM and its extension to guarantee its applicability towards approximate temporal patterns in general, and cascades and arrangements of events in particular.

4.2 Related Work

A temporal pattern with high frequency may not be interesting if it is statistically expectable, while a temporal pattern with low frequency may be interesting if it is statistically unexpected. The first key-point for this assessment is to choose an appropriate null model to describe the input data from which expectations can be assessed. According to the surveyed research in *Section V-1.2*, there are three major directions for this end: *1)* approximation of adequate distributions of the data $P_{\mathcal{X}}$ [347, 322], *2)* generation of randomized data (either using permutations or distributions) [253, 367, 450], and *3)* creation of hold-out data partitions to test the support of patterns [63].

Based on the expectations computed from these null models, two options can be considered for the final statistical assessment: 1) non-parametric equations or 2) hypothesis testing.

An illustrative *non-parametric equation* is given by the application of Chernoff bounds on the difference between the observed and expected value of a random variable (the pattern support). Given a temporal pattern s, the

observed value θ is the number of observations supporting s, while the expected value $E(\theta)$ is the pattern's support according to the computed null model. Assuming the error to be lower bounded by ϵ, then $\forall_{\epsilon\in[0,1[}, P(|\theta - E(\theta)| \geq \epsilon) < e^{-2n\epsilon^2}$, where n=|**X**| is the number of observations. This inequality can be either solved for ϵ or, by fixing ϵ, can be solved for n to provide a lower bound of the required minimum number of observations (sample data size) to satisfy the given error ϵ. Despite its utility, the use of the Chernoff bounds do not allow the distinction between false positives and negatives. To address this problem, Laur et al. [388] extended this assessment to either control false negative rate or false positive rate. Yet, a choice still needs to be placed.

An alternative option is to statistically test the difference between the observed and expected pattern support. When a *hypothesis test* is performed, there are two possible errors: 1) the risk of rejecting a correct null hypothesis (false positive rate) α, and 2) the risk of not rejecting a false null hypothesis (false negative rate) β. Existing research are mainly focused on reducing false positives at a risk of increasing false negatives, not providing a bound for both risks α and β [347, 450]. Furthermore, a large part of existing work do not place correction procedures to adequately assess the deviation from expectations [379, 347].

To our knowledge, we identify four major works on how to statistically assess sequential patterns derived from itemset sequences. First, Jacquemont et al. [347] rely on global tests to infer minimum support thresholds in the form of statistical constraints to guarantee the extraction of significant patterns from a compact and generalized representation of itemset sequences in the form of a probabilistic automaton. Second, Gwadera and Crestani [284] ranks sequential patterns with respect to their significance given a null model by combining itemset-wise and sequence-wise views. For this aim, they use a mixture model obtained by conditioning the length of the sequence to test unexpectedly high number of precedences, and a maximal entropy model to test unexpectedly high dependency between a set of items. This approach, while theoretically sound, may become inefficient with a large set of items, and cannot be used to infer global constraints. Third, Low-Kam et al. [423] test the observed against the expected support by combinatorially enumerating the possible positions for the itemset of each sequential pattern and by considering that the probability of an item appearing at a particular position in the itemset sequence is independent of both its position in the sequence and the appearance of other items in the sequence. An upper bound on the associated p-value is provided for greater efficiency and principles are place to integrate this assessment within SPM methods. Finally, Kum et al. [379] proposes an alternative statistical view of sequential patterns to guarantee deviations from random patterns (referred as spurious patterns), and extends this assessment with three additional criteria: recoverability (how well a planted region in synthetic data is detected), the number of redundant patterns, and the degree of extraneous items in the patterns. In this light, a good model should not only minimize the number of spurious patterns, but also achieve a high recoverability with a low level of extraneous items and low redundancy (tackled concerns throughout *Book IV*).

There is a complementary rich body of literature on how to guarantee the statistical significance of patterns retrieved from temporal data, including (possibly noisy) strings and motifs [279, 537, 692], multiple alignments [265, 279], and dependent event patterns [403]. However, these methods assume that input data is given by time series, and thus are hardly extensible towards the assessment of alternative patterns from itemset sequences.

4.3 Solution

Below, we derive from existing contributions from related work, a statistical framework to assess the statistical significance of regions from structured data. For this aim, *Section 4.3.1* defines statistical views to test sequential patterns from itemset sequences. *Section 4.3.2* extends this assessment to guarantee a reduced number of false positives and negatives. *Section 4.3.3* provides an alternative view on how to assess sequential patterns given by high-probable paths from probabilistic Markov-based models. *Section V-4.3.4* extends these contributions to enable the assessment of cascades and arrangements of events from temporal data. Finally, *Section V-4.3.5* provides a synthetic view on how the proposed tests can be used to guide the learning of local descriptive models.

4.3.1 Assessing the Significance of Sequential Patterns

Given a sequential pattern s with observed support $\sup_s = |\Phi_s|$, $p_s = E(\frac{sup_s}{n})$ is a estimation of its true probability. Since the true probability is unknown, one can formulate a hypothesis on its real value. The null hypothesis is that p_s is not high enough to consider s as statistically significant ($H_0 : p_s \leq \alpha$). In this context, rather than directly comparing the observed and expected frequency, this difference is tested using an adequate corrected significance level that guarantees an adequate deviation from expectations.

According to the work of Kum et al. [379], the expected support for a pattern s can be informally formulated as $p_s = E(\frac{sup_s}{n}) = P(s \text{ appears at least once}) = 1 - P(s \text{ never appears})$. Given a set of symbols $\mathcal{L}$, the expected support of a sequential pattern with two sequential items (one preceding the other) in $\mathcal{L}$ is given by:

$$E\Big(\frac{sup(i_1 \rightarrow i_2)}{n}\Big) = 1 - P(\neg(i_1 \rightarrow i_2)) = 1 - \Big(p(i_1)\sum_{j=1}^{n}\Big((1-p(i_2))^{n-j}(1-p(i_1))^{j-1}) + (1-p(i_1))\Big)^n\Big) \tag{4.1}$$

where i_1 and i_2 are two items in $\mathcal{L}$ (not necessarily distinct from each other), m is the number of itemsets $P(i_1 \rightarrow i_2)$ denotes the probability that item i_1 is followed by item i_2 at least one time in a sequence, and $p(i_1)$ is the probability of item i_1 appearing in the sequence at any particular position. The probability of observing longer sequential patterns is calculated recursively. Illustrating, the calculation of $P(i_1 \rightarrow i_2 \rightarrow i_3)$, is facilitated by knowing the probability of the shorter pattern $i_2 \rightarrow i_3$.

Considering now the possibility to consider co-occurring items in a sequential pattern. Given an itemset I with d items, $I=\{i_1, .., i_d\}$, then its probability is given by $p(I)=\Pi_{i_k \in I} p(i_k)$. As such, by replacing the probability of an arbitrary element, $p(i_k)$, by $p(i_1)p(i_2)..p(i_d)$ in Eq.(4.1), we can derive the expected support of a sequential pattern.

$$E\Big(\frac{sup(I_1 \rightarrow (I_2 \rightarrow .. \rightarrow I_3))}{n}\Big) = 1 - \Big(\prod_{i_k \in I_1} p(i_k)\sum_{j=1}^{n}\Big((1-p(I_2 \rightarrow .. \rightarrow I_3))^{n-j}(1-\prod_{i_k \in I_1} p(i_k))^{j-1}) + (1-\prod_{i_k \in I_1} p(i_k))\Big)^n\Big) \tag{4.2}$$

As the quantity $p(i_k)$ for an item in $\mathcal{L}$ is not known, it must be estimated from the input data and necessarily exhibits some sampling error. Consequently, $E(\frac{\sup(s)}{n})$, also exhibits a form of sampling error. For the estimation of $p(i_k)$, we propose the use of either probabilistic distributions or frequentist countings depending on the size and dimensionality of the input data (according to the principles proposed in *Section V-1.3.1*). For this end, the dimensionality of an itemset sequence is computed according to the calculus introduced in *Basics I-1.6*.

4.3.2 Controlling False Positive and Negative Rate

Search Space. To guarantee that the computed probability adequately deviates from expectations, the confidence threshold to assess the statistical significance of pattern needs to be corrected according to the properties of the space of similar patterns. However, and contrasting with the statistical assessment of patterns from tabular data, there is a higher complexity associated with the inference of the properties the space of similar patterns from structured data. In this context, the few studies that apply a correction consider the search space to be approximately given by the retrieved/frequent patterns under a closed pattern representation [423]. Since the characterization of the search space of temporal patterns is out of the scope of our work, we rely on this approximation.

Minimizing False Negatives. Under this approximation, different corrections can be applied. When considering the Bonferroni correction, the confidence to test the null hypothesis is simply divided by the search space size. However, the given search space size can be arbitrarily voluminous (high number of frequent closed sequential patterns), leading to pessimistic views of the true significance and to a greater susceptibility towards false negatives (rejecting biclusters that are statistically significant). To address this problem, the Hochbert correction can be applied. The p-values from the retrieved patterns are sorted according to their expected probability of occurrence, $\{p_{s_1}, p_{s_2}, .., p_{s_d}\}$, and the p-value $max_{p_{s_j}} : \forall_{1 \leq j \leq d} p_{s_j} \leq \alpha/(d - j + 1)$ is outputted as the corrected level.

Despite offering a good compromise between type-I and type-II errors, this strategy is computationally expensive for large outputs. To reduce the computationally complexity, we propose the sequential patterns to be ordered according to the number of items (which approximately determines the order of their significance level), and the average significance of the 5 sequential patterns around the computed index j (according to the Hochbert equation) to be used to adjust the confidence level α.

Parameterizable Control of the Risks. The previously proposed correction aims to minimize the false positive rate, α (the probability of outputting a non-significant or false frequent pattern), without compromising the false negative rate. As such, in order to be able to take a well-founded decision, α can be fixed (usually 5%, although it depends on the application domain) to compute a bound of rejection. More formally, $\alpha = P(p(w) > k'|H_0 true)$. Yet, we can also orient the statistical testing according to the risk of false negatives, β (probability of rejecting a significant or true frequent pattern). More formally, $\beta = P(p(w) < k'|H_1 true)$. The work of Jacquemont et al. [347] explores way of affecting both risks simultaneously, and goes further on defining a lower bound on how many observations $|\mathbf{X}|$ are needed to not exceed a priori fixed α and β risks, showing under certain conditions how the number of observations affects the guarantees of precision (1-α) and recall (1-β).

4.3.3 Assessing the Significance of Highly-Probable Paths in Graphs

Although temporal patterns were introduced as the baseline option to model regions from structured data, throughout *Chapters IV-3* and *IV-4* we introduce the alternative option of stochastically modeling the input data under a Markov assumption. In this context, regions are given by highly probable paths from the underlying lattices. Under this assumption, there are two major options to assess the statistical significance of these regions.

First, sequential patterns can be decoded from the learned lattices (according to Alg.IV-9) and the previously discussed statistical tests applied to test their significance. In this context, the proposed decoding process[1] should consider relaxed probability thresholds to minimize the number of false negatives (non-significant patterns).

Second, new statistical tests can be defined to assess the significance of a region based on the deviation of its generative likelihood against expectations. In other words, the probabilities of the paths associated with a region should be unexpectedly high. This option is essential to assess noise-tolerant regions. The previous support model requires a pattern to be supported by an observation if and only if it is fully satisfied by that observation. However, noise can cause the exact matching methods to miss relevant regions and affect the significance calculus.

Yet, a critical concern with regards to this second option is related with the need to use a probabilistic model that enables the correct estimation of the true probability of a pattern. Hingston [322] proposed a method to compute such estimates from a probabilistic deterministic finite state automaton. However, *Chapters IV-3* and *IV-4* shown the relevance of using hidden Markov models to address the challenges associated with the modeling of cascades and arrangements of events. Relevantly, Dupont et al. [193] showed that not only finite state automata (PDFA), but also Hidden Markov Models, are probabilistic models enabling adequate estimations of the true probability of a pattern. In fact, it can be theoretically shown that under certain conditions the two models are equivalent [193]. Under a Markov model, statistical tests of significance have been proposed for string patterns [537, 568], generating functions (with variable-length gaps) [220], and structured motifs [556, 705]. Yet, there is a lack of literature on how to adequately assess temporal patterns from Markov-based models with non-trivial architectures (such as the architectures proposed in *Chapter IV-3*).

In this context, we propose two groups of statistical tests to test regions directly in the probabilistic model, where each group offers unique specificities of interest. The first option is to test the unexpectedness of a given region by seeing the probabilistic model as a weighted graph and by assessing how the observed density of the subgraph associated with the region deviates from its weight expectations. For this aim, the statistical tests proposed by Tanay

[1] Borges and Levene [88] proposed alternative decoding principles for the extraction of frequent patterns from weighted graph structures and Dupont et al. [192] present a method to extract relevant subgraphs between nodes of interest with Hidden Markov Models using random walks.

et al. [616] assume the weights of the edges in the graph (corresponding to transmission and emission probabilities of a hidden Markov model) are normally distributed, and compute the p-value for each region (subgraph) based on the probability of finding another similar region with at least the same weight. The second option is based on the statistical tests proposed by Jacquemont et al. [347]. Accordingly, the significance of decodable sequential patterns in the probabilistic model is assessed based on two conditions: 1) pattern coverage (a pattern s=<I_1..I_d> must cover a significant part of the probability density of all sequences), and 2) statistical dependence between itemsets (is the majority of sequences containing <I_1..I_{d-1}> also contain <I_1..$I_{d-1}I_d$>). The first condition is assessed using a proportion test aiming to verify if the estimate of p(s) on the probabilistic model is high enough, while the second is verified under a Fisher exact test. In this context, a statistically significant pattern satisfies both the proportion and dependence constraints in a given probabilistic model.

4.3.4 Assessing the Significance of Cascade Models and Arrangements of Events

The previous sections provide robust statistical views to assess the significance of deterministic and stochastic models of sequential patterns. As discussed throughout *Book IV*, regions of interest from structured data can be given by cascades (in the presence of three-way time series) or arrangements of events (in the presence of multi-sets of events), and seen as a specific variant of sequential patterns, where item occurrences are annotated with a time frame with an expectation for their occurrence (according to Def.IV-2.4). An illustrative instantiation of such pattern is <$(\{a,c\},\varphi_2),(\{d\},\varphi_4)$>, where $\{a,c\}$ items are expected to co-occur in the second time partion (φ_2), preceed by d item, which is expected to occur in the fourth time partition (φ_4). By ignoring the time annotations, these temporal patterns become sequential patterns and the statistical tests proposed in *Section 4.3.1* can be directly applied to assess their significance. For this aim, whenever multiple time granularities are used to compose different datasets (according to Def.IV-2.3), the expected probability of a particular item to occur, co-occur and preceed another item should be assessed in the context of a specific time granularity. Furthermore, the emptyset symbol (associated with time partitions without item occurrences) should be ignored for this end.

Under these assumptions, the statistical significance of cascades and arrangements of events can be soundly assessed. Nevertheless, an important concern is whether the time annotations can be use to gain further knowledge and correct the expected true probability of a temporal pattern. Although one might think that such knowledge can be used to minimize the possible optimistic biases incurred from focusing uniquely on orderings, the time annotations are simply expectations (there can be arbitrary-high temporal misalignments between observations). In this context, the allowed degree of misalignment between observations is dependent on the applied algorithms and therefore these annotations cannot be feasibly incorporated as part of a statistical test. Illustrating, assume that the previous temporal pattern, <$(\{a,c\},\varphi_2),(\{d\},\varphi_4)$>, is supported by $\{x_1,x_2,x_3\}$ observations. The $\{a,c\}$ items may co-occur in the first time partition of $\{x_1,x_3\}$ observations and in the fourth time partition for x_2 observation, yet the expected time of occurrence is the second partition. As such, by assuming that temporal misalignments can occur, the statistical tests based on the estimator of the true probability (V-4.2) are not optimistic. As a result, the significance of cascades and arrangements of events is robustly assessed.

For the purpose of assessing these regions directly on the customized Markov-based models (*Chapter IV-3*), the statistical tests provided in the previous section can be soundly applied. Despite the inherent specificities of the proposed architectures, the principles to test the unexpectedness of the weights of the probability paths associated with a region are preserved. For instance, the use of dedicated states to emit delimiters of time partitions (*Section IV-3.2.6*) does not affect the statistical tests since the probability of finding similar regions with at least the same weight is equally affected by these customizations. In this context, the learned transmissions and emissions of the customized HMMs can be used to offer an alternative and noise-tolerant statistical view of the significance of cascades and arrangements of events.

4.3.5 Using Significance Criteria to Guide the Learning

The previous statistical tests can be used as a filtering procedure to remove non-significant cascades and arrangements. However, this option prevents the exploration of possible efficiency gains from narrowing the search space. For this end, three strategies can be used. First, in the presence of deterministic methods, the statistical tests can be pushed into mining step in the form of monotonic constraints: if a region is statistical significant, a region with more items/events or observations is necessarily statistically significant.

Second, in the presence of stochastic methods, the statistical view of the significance of regions from probabilistic models can be used to guide the efficient decoding of statistically significant regions. For this purpose, the work by Jacquemont et al. [347] shows how to incorporate constraints to accelerate decoding procedures on probabilistic models. Expectations on the properties of a region that guarantee its significance can be also used to shape the architectures (e.g. increasing the minimum number of precedences).

Finally, global constraints can be alternatively inferred without the need to rely on individual tests by considering that the number of expected regions with a fixed number of items or events per observation is well-approximated by a Poisson distribution. Under this assumption, the principles proposed in *Section V-1.3.3* can be applied to detect the minimum support for regions with a given number of items or events per observation by employing parametric tests to deviations on the observed number of regions against expectations. Similarly to the proposed strategy, enhancements can be consider to simultaneously minimize false positives and false negatives.

4.4 Summary of Contributions and Implications

This chapter proposes a statistical view of deterministic and probabilistic models of structured data to assess the statistical significance of regions in these data contexts. For this aim, we first surveyed relevant work for the statistical assessment of temporal patterns from structured data. Then, we extended statistical tests to assess the significance of sequential patterns from itemset sequences under a null model in order to guarantee their applicability towards temporal patterns. In particular, we tackled the problem of guaranteeing their adequate deviation from expectations, proposing a new way to efficiently compute non-conservative corrections by reducing the need to apply a large number of statistical tests. Furthermore, we further studied the possibility of assessing the statistical significance of regions encoded in probabilistic models based on the unexpectedness of their weight and coverage. This assessment was shown to be compliant with the previously proposed hidden Markov models. Finally, we proposed new principles to guide the previously proposed algorithms to learn from structured in order to focus their search on regions with promising statistical significance.

Empirical Evidence. Unlike remaining chapters, the contributions of this chapter were mainly grounded on (statistical) principles already found in literature. In particular, our contributions are seen an increment on top of four previous works [379, 347, 423, 284]. The soundness of these statistical views for retrieving significant patterns, as well as initial empirical evidence on synthetic and real data, can be found in the referenced articles. This chapter goes further on tackling some of their limitations and extending the statistical framework towards cascades and arrangements of events. Since the compliance of the statistical views with these new classes of local descriptive models is shown, we consider the empirical validation to be less critical and thus seen as part of future work.

Future Work. Three major directions are identified for future work. First, we expect to draw an in-depth analysis on how the properties of a given region (such as their support, number of items/events and tolerated noise) and of the input dataset (size, dimensionality and global regularities) determine its statistical significance. Second, we aim to compare the inherent properties associated with the statistical tests proposed for deterministic and probabilistic representations of temporal patterns. Finally, we expect to extend these statistical views, currently prepared to deal with symbolic temporal data, towards real-valued data, enabling a (possibly) more accurate analysis of regions retrieved from numeric three-way time series and of multi-sets of event with numeric attributes.