Data Min Knowl Disc
DOI 10.1007/s10618-017-0521-2

BSig: evaluating the statistical significance of biclustering solutions

Rui Henriques[1] · **Sara C. Madeira**[2,3]

Received: 8 February 2016 / Accepted: 12 June 2017

Abstract Statistical evaluation of biclustering solutions is essential to guarantee the absence of spurious relations and to validate the high number of scientific statements inferred from unsupervised data analysis without a proper statistical ground. Most biclustering methods rely on merit functions to discover biclusters with specific homogeneity criteria. However, strong homogeneity does not guarantee the statistical significance of biclustering solutions. Furthermore, although some biclustering methods test the statistical significance of specific types of biclusters, there are no methods to assess the significance of flexible biclustering models. This work proposes a method to evaluate the statistical significance of biclustering solutions. It integrates state-of-the-art statistical views on the significance of local patterns and extends them with new principles to assess the significance of biclusters with additive, multiplicative, symmetric, order-preserving and plaid coherencies. The proposed statistical tests provide the unprecedented possibility to minimize the number of false positive biclusters without incurring on false negatives, and to compare state-of-the-art biclustering algorithms according to the statistical significance of their outputs. Results on synthetic and real data support the soundness and relevance of the proposed contributions, and stress the need to combine significance and homogeneity criteria to guide the search for biclusters.

Responsible editor: Ian Davidson.

✉ Rui Henriques
rmch@tecnico.ulisboa.pt

Sara C. Madeira
sacmadeira@ciencias.ulisboa.pt

1 INESC-ID and DEI, Instituto Superior Técnico, Universidade de Lisboa, Lisbon, Portugal

2 LASIGE, Faculdade de Ciências, Universidade de Lisboa, Lisbon, Portugal

3 INESC-ID, Lisbon, Portugal

Published online: 28 June 2017

Keywords Biclustering · Statistical significance · Pattern mining

1 Introduction

Given a real-valued or symbolic matrix, biclustering seeks to find subsets of rows with specific homogeneity across a subset of columns. Biclustering has been applied for the analysis of gene expression data, biological and social networks, collaborative filtering data, growth phenotype data, genomic structural variations, text data, chemical data, among other applications (Henriques et al. 2015; Gnatyshak et al. 2012; Henriques and Madeira 2016b; Alzahrani et al. 2017; Madeira and Oliveira 2004). Despite the relevance of the biclustering task for several biomedical and social applications, 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 of the 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 insufficient since good levels of homogeneity can appear by chance in the sample data (commonly observed for small biclusters).

A few statistical views are available to assess the significance of specific types of constant biclusters (Califano et al. 2000; Bellay et al. 2011; Ramon et al. 2013), such as dense biclusters (Tanay et al. 2002), low-variance biclusters (Lee et al. 2015), and constant biclusters with sequential constraints (Madeira and Oliveira 2007). However, these views are not generalizable towards more flexible coherencies, including biclusters with varying coherency strength and additive, multiplicative, plaid or order-preserving assumptions. Furthermore, the statistical evaluation of biclusters is also challenged by the need to guarantee that the retrieved biclusters deviate from expectations. As a result, statistical assessments should be able to minimize the risk of false positive discoveries (biclusters that appear by chance on the sample data) without increasing the risk of excluding relevant biclusters (false negatives).

This work explores why existing efforts in the field of biclustering are not yet able to address the enumerated problems (insufficiency of homogeneity criteria and challenges associated with the statistical assessment of biclusters with flexible coherency); surveys the limitations and contributions from biclustering and related research streams (including pattern mining and inferential statistics/estimation theory); and proposes a statistical method to efficiently test the significance of biclustering solutions with a strict upper limit on the risk of false discoveries. This assessment can be used to either filter biclusters or as a sound heuristic to narrow the search space of biclustering algorithms. In this context, this work provides five major contributions:

1. sound statistical tests able to evaluate the significance of (real-valued) biclusters with constant coherency;
2. first statistical tests to robustly assess: 1) additive models, 2) multiplicative models, 3) plaid models, 4) order-preserving models, 5) symmetric models;
3. theoretical and empirical analysis on how coherency impacts significance;
4. a new multi-hypothesis correction to test deviation from expectations, thus addressing computational bottlenecks of non-conservative corrections while minimizing the risk of false negatives of conservative corrections;

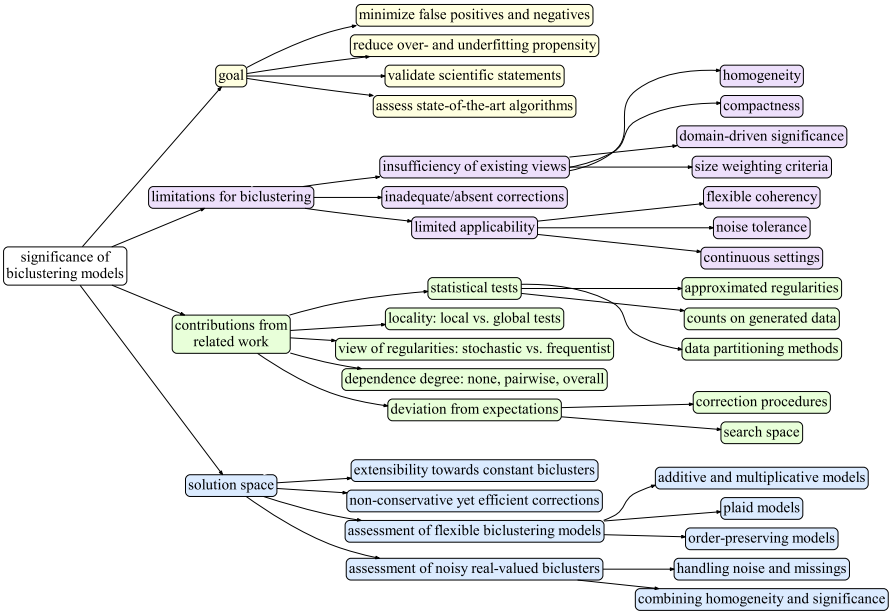

Fig. 1 Challenges and proposed contributions to evaluate biclustering solutions

5. principles to guarantee the assessment of biclustering solutions with arbitrary-high levels of noise and missing values, and to consistently combine significance and homogeneity views for a complete evaluation.

These contributions are critical to:

- validate the increasing number of implications that are derived from real data without statistically sound guarantees;
- evaluate and compare state-of-the-art biclustering algorithms with regards to the significance of their solutions;
- output solutions without spurious biclusters (false positives/negatives);
- offer a sound criteria to reduce the high number of biclusters outputted by exhaustive approaches;
- guide the biclustering tasks, promoting the efficiency of searches.

These contributions are consistently integrated within an evaluation method, implemented in the BSig (**B**iclustering **Sig**nificance) toolbox.[1] Figure 1 summarizes the enumerated problems and contributions.

Accordingly, this paper is organized as follows. Section 2 provides background concepts. Section 3 surveys the current challenges and relevant contributions to address the target problem. Section 4 proposes the statistical principles to robustly evaluate biclustering solutions with varying coherency and quality; and introduces the BSig method to consistently combine these principles. Section 5 provides empirical evidence of the relevance and soundness of the proposed method. It further compares

[1] https://web.ist.utl.pt/rmch/software/bsig/.

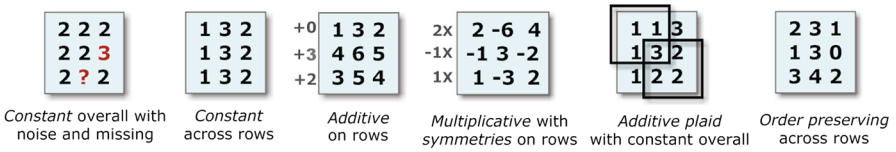

Fig. 2 Illustrative discrete biclusters with varying coherency assumption and quality

state-of-the-art biclustering algorithms regarding their statistical significance. Finally, the contributions and implications of this work are synthesized.

2 Background

Definition 1 Given a matrix, $A = (X, Y)$, with a set of rows $X = \{x_1, .., x_N\}$, set of columns $Y = \{y_1, .., y_M\}$, and elements $a_{ij} \in \mathbb{R}$ relating row i and column j:

- A **bicluster** $B = (I, J)$ is a $n \times m$ submatrix of A, where $I = (i_1, .., i_n) \subset X$ is a subset of rows and $J = (j_1, .., j_m) \subset Y$ is a subset of columns;
- The **biclustering task** aims to identify a set of biclusters $\mathcal{B} = \{B_1, .., B_s\}$ such that each bicluster $B_k = (I_k, J_k)$ satisfies specific *homogeneity criteria*.

The homogeneity criteria are commonly guaranteed through the use of a merit function, such as the variance of the values in a bicluster (Madeira and Oliveira 2004). In stochastic approaches, a set of parameters that describe the biclustering solution are learned by optimizing the merit (or likelihood) function (Hochreiter et al. 2010). Alternatively, merit functions can be defined to locally maximize greedy iterative searches, to combine row- and column-based clusters, to exploit matrices recursively, or to guide the space exploration in exhaustive searches (Madeira and Oliveira 2004).

The merit function determines the coherency and quality of biclusters. The *coherency* of a bicluster is defined by the observed correlation of values (coherency assumption) and by the allowed deviation from expectations (coherency strength). A bicluster can have coherency of values across its rows, columns or overall elements, with values typically following constant, additive, multiplicative, symmetric, order-preserving and plaid assumptions (Henriques et al. 2015). The *quality* of a set of biclusters is defined by the type and amount of accommodated noise. Definitions 2-4 formalize these concepts, and Fig. 2 illustrates biclusters with varying coherency assumptions. Table 6 in *appendix* motivates the relevance of assessing biclusters with flexible coherency by listing biological, clinical and social data contexts where such biclusters are commonly found.

Definition 2 Let the elements in a bicluster $a_{ij} \in (I, J)$ have **coherency** across rows $a_{ij} = c_j + \gamma_i + \eta_{ij}$ (or columns $a_{ij} = c_i + \gamma_j + \eta_{ij}$), where c_j (or c_i) is the value of column j (or row i), γ_i (or γ_j) is the adjustment for row i (or column j), and η_{ij} is the noise factor of a_{ij}.

Definition 3 Let $\bar{A}$ be the amplitude of the range of values in a matrix A. Given a real-valued matrix A, the **coherency strength** is a range $\delta \in [0, \bar{A}]$, such that

$a_{ij} = c_j + \gamma_i + \eta_{ij}$ where $\eta_{ij} \in [-\delta/2, \delta/2]$. Given a symbolic matrix, the coherency strength δ is defined by the number of symbols $\mathcal{L}$, $\delta = \frac{1}{|\mathcal{L}|}$.

Definition 4 The γ factors define the **coherency assumption**: **constant** when $\gamma = 0$, **multiplicative** if a_{ij} is better described by $c_i\gamma_j + \eta_{ij}$ (or $c_j\gamma_i + \eta_{ij}$), and **additive** otherwise. **Symmetries** can be accommodated on rows, $a_{ij} \times k_i$ where $k_i \in \{1, -1\}$. **Order**-preserving assumption is verified when the values of rows induce the same linear ordering across columns. A **plaid** assumption considers the cumulative effect of the contributions from multiple biclusters on areas where their rows and columns overlap.

Definition 5 The bicluster **pattern** φ_B is the ordered set of values in the absence of adjustment and noise factors: $\varphi_B = \{c_j \mid y_j \in J\}$ for coherency across rows (or $\varphi_B = \{c_i \mid x_i \in I\}$ for column-based coherency). The bicluster **support** $\sup_B$ is the number of rows $n = |I|$ (or columns $m = |J|$) respecting φ_B.

Given the illustrative additive bicluster with coherency on rows $B = (\{x_1, x_2, x_3\}, \{y_1, y_2, y_3\})$ from Fig. 2 where $a_{ij} \in \mathbb{N}^+$. This bicluster can be described by $a_{ij} = c_j + \gamma_i$ with pattern $\varphi_B = \{c_1 = 1, c_2 = 3, c_3 = 2\}$, supported by three rows with additive factors $\gamma_1 = 0$, $\gamma_2 = 3$ and $\gamma_3 = 2$.

To our knowledge, there are no contributions to assess the significance of biclustering solutions with flexible coherency (Definition 2-4). Nevertheless, mappings between pattern mining and biclustering have been recently proposed by Henriques et al. (2015) (see Definition 6), opening a new direction for robust assessments since sound statistical views have been largely researched in the context of pattern mining (Gionis et al. 2007; Kirsch et al. 2012).

Definition 6 Let $\mathcal{L}$ be a finite set of items, and P a pattern be a composition of items (itemset I, rule $I_1 \rightarrow I_2$ or sequence $I_1..I_n$, where $I_i \subseteq \mathcal{L}$). Given a set of transactions $D = \{P_1, .., P_n\}$, let the *coverage* Φ_P of a pattern P be the set of transactions in D in which P occurs ($\{i \mid P \subseteq P_i\}$), its *support* sup_P to be the coverage size ($|\Phi_P|$), and its *length* to be the number of items ($|P|$). Given D and a minimum support and length thresholds, θ_1 and θ_2, *pattern mining* aims to compute: $\{(P, \Phi_P) \mid sup_P \geq \theta_1 \wedge |P| \geq \theta_2\}$.

Given a matrix A, a set of transactions D can be derived by: 1) concatenating a_{ij} discretized elements with their column indexes (to learn constant patterns according to Henriques and Madeira 2014b) or 2) ordering columns according to their values per row (to learn order-preserving patterns according to Henriques and Madeira 2014a). Let Ψ_P of a pattern P in D be its columns, and Υ_P be its items in $\mathcal{L}$. Given A, **pattern-based biclustering** aims to learn a set of *biclusters* $\cup_k B_k$ from patterns $\cup_k P_k$ (discovered on transactions D derived from A) by mapping $I_k = \Phi_{P_k}$, $J_k = \Psi_{P_k}$ and $\varphi_{B_k} = \Upsilon_P$.

Figure 3 maps a (symbolic) matrix into two distinct transactional databases (given by index concatenations and orderings) for the subsequent discovery of constant and order-preserving biclusters derived from frequent patterns.

To motivate the target problem, consider a discrete matrix with 1000 rows, 200 identically distributed columns and 5 symbols uniformly distributed. Assume that we

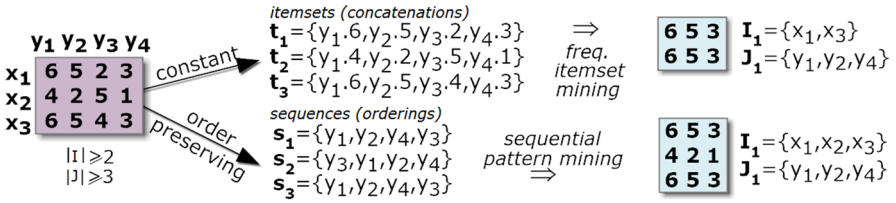

Fig. 3 Pattern-based biclustering (Henriques et al. 2015): discovery of two illustrative biclusters with constant and order-preserving assumptions based on frequent itemsets and frequent subsequences from transactional data mapped from the input data matrix

observe a pattern with constant symbols on 3 columns across 25 rows. Is the associated (pattern-based) bicluster, B, statistically significant? The probability of the pattern occurrence is $p_{\varphi_B} = 1/5^3$. A binomial calculus shows that the probability to have at least 25 supporting rows is $p_B = 5.0\text{E-}3$. Although p_B is considerably low, we need to consider the space of all similar biclusters, $s = 5^3$, to guarantee it deviates from expectations. Assuming the conservative Bonferroni correction at $\alpha = 0.05$, p_B is assessed against $\alpha/s = 4\text{E-}4$. Under these assumptions, B is rejected. Understandably, this illustrative assessment needs to be revised to control false negatives and enable the evaluation of real-valued biclusters with noise and non-constant coherencies.

3 Related work

3.1 Limitations of the state-of-the-art evaluation of biclustering solutions

Despite the rapidly increasing number of contributions on the field of biclustering, assessing the statistical significance of biclustering solutions has been poorly explored (Henriques 2016). We discuss below why this is the case and identify the major limitations of existing efforts.

Optimization and scoring schema. Stochastic approaches for biclustering rely on multivariate distributions to approximate data (Hochreiter et al. 2010). However, the learned distributions are not considered for further testing the significance of biclusters. Instead, they just derive biclusters from the learned parameters as soon as specific convergence criteria are satisfied. Alternative approaches for biclustering (Madeira and Oliveira 2004) commonly define an objective metric of the homogeneity of the biclusters to either guide the discovery, sort the found biclusters and/or filter the discovered bicluster. Understandably, these functions do not guarantee that a bicluster is not found by chance in the sample data. Small biclusters can have high levels of homogeneity by chance. Some of the available merit functions compensate this undesirable effect by weighting the size of biclusters to benefit larger biclusters (Mitra and Banka 2006). However, this 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 have been applied to guarantee that homogeneity is below a residual error with a particular significance and statistical power

(Wang et al. 2002). Serin and Vingron (2011) proposed tests 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 prevent a bicluster from occurring by chance.

Domain-driven significance. Statistical tests can be also used to guarantee the domain relevance of each bicluster by computing the enriched terms of a bicluster against a knowledge base (Huang et al. 2009). Noureen et al. (2009) compared the performance of biclustering algorithms on gene expression data. Pio et al. (2012) proposed a t Test to evaluate the hypothesis that rows/columns within the same biclusters are more functionally similar according to enriched terms in Gene Ontology (GO) than rows/columns belonging to different biclusters. Although domain-driven indicators can be considered to be the ultimate criteria to assess biclustering solutions, they suffer from three major drawbacks. First, knowledge bases are incomplete and prone to errors. Second, domain-driven evaluation only targets one dimension of the bicluster at a time (either the group of rows or columns). Third, domain significance and statistical significance are not always in agreement. In this context, domain-driven views should be complemented with statistical significance views to better assess or promote the relevance of biclusters.

Testing specific types of biclusters. A few biclustering methods perform robust tests to guarantee the significance of their solutions (Califano et al. 2000; Ramon et al. 2013; Tanay et al. 2002; Bellay et al. 2011). However, these tests cannot be easily extended to evaluate flexible and noisy biclustering solutions (whose relevance is highlighted in Table 6 in *appendix*).

Koyuturk et al. (2004) proposed an objective statistical function to test unusually dense biclusters in binarized matrices by assuming that values are binomially distributed. Tanay et al. (2002) proposed a generalized assessment of dense biclusters by mapping the matrix into a weighted graph (weights are assumed to be normally distributed) and computing a 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. Lee et al. (2015), Balakrishnan et al. (2011) and Chen and Xu (2016) proposed alternative and rigorous statistical frameworks to test dense biclusters using finite distributions. Despite their relevance, the provided formulations are meant to test biclusters with low variance (as these works tackle the problem of submatrix localization) and assume that the observed data is well described by an univariate distribution. eCCC-Biclustering (Madeira and Oliveira 2007) assesses noise-tolerant constant biclusters B with contiguous columns (given by a string pattern φ_B) by computing the probability of a bicluster with the same size of B to occur by chance in a matrix with the same frequency of contiguous pairs of symbols or ranges of values (first-order Markov assumption). Finally, Ramon et al. (2013), Bellay et al. (2011) and Califano et al. (2000) propose alternative statistical tests to assess the probability of a constant bicluster to deviate from different forms of background noise.

Despite the relevance of the surveyed works, they can only handle specific forms of constant coherency, and therefore their approaches are not applicable to assess biclustering solutions outputted by alternative algorithms.

3.2 Contributions from related streams of research

A substantial portion of the contributions for the statistical assessment of local regularities has been developed in the context of pattern mining (PM) (Gionis et al. 2007; Kirsch et al. 2012). In PM, and according to Definition 6, a pattern can be mapped into a bicluster with coherency across rows and no noise ($\eta_{ij} = 0$). The relevance of a pattern is defined by its support (number of rows) and length (number of columns). The statistical significance of a pattern is then given by the probability of its support and/or length to deviate from expectations. Accordingly, we now survey studies that aim to guarantee the statistical significance of pattern-based solutions and minimize their risk towards false positives and negatives. These contributions provide the structural principles to address the target problem since, to our knowledge, no contributions to evaluate the statistical significance of biclusters with non-trivial forms of coherency and arbitrary tolerance to noise have been proposed to date.

To test the significance of a (pattern-based) bicluster B, the regularities of the input matrix A, Θ, need to be adequately modeled to assess the probability of occurrence of bicluster B, p_B. p_B can be computed by testing B against: *1)* approximated distributions, *2)* randomized synthetic datasets, and *3)* hold-out data partitions. Multiple studies approximate the distributions underlying data, Θ, by either fitting distributions for each row $x_i \sim \Theta_i$, column $y_j \sim \Theta_j$, or for the overall matrix $a_{ij} \sim \Theta$ (Karian and Dudewicz 2010). Then, p_B is often estimated from the joint probability of a specific pattern φ_B to occur, p_{φ_B}, for a minimum number of rows by computing Binomial tails (Madeira et al. 2010). Instead of relying on the observed data, some approaches generate multiple synthetic datasets from the underlying data regularities Θ (Kirsch et al. 2012). Here, p_B can be estimated by testing the φ_B support in the observed data against the h generated datasets.[2] Gionis et al. (2007) 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 (1998) relaxed this criteria, by assuming independence of items among transactions while preserving their frequencies. Monte Carlo sampling is an additional option (Kirsch et al. 2012). Webb (2007) proposed a hold-out method to control the false discovery rate by testing the support of patterns found in the exploratory and holdout partitions using paired t tests.

In pattern discovery, density functions that compare the observed support sup_B against the expected support $\hat{sup}_B$ (with $\hat{\sigma}(sup_B)$ deviation) have been proposed by using one of the previously surveyed assessment options. Significance ratios include: $sup_B/\hat{sup}_B$ and $|sup_B\text{-}\hat{sup}_B|/\hat{\sigma}(sup_B)$, among others[3] (DuMouchel 1999; Kirsch et al. 2012). Scores to identify spurious pattern discoveries have been also proposed by Bolton et al. (2002). Alternative scores have been proposed in the context of Bayesian analyzes by Silberschatz and Tuzhilin (1996) and association rule mining by Scheffer (2005), Hämälinen and Nykänen (2008) and Zhang et al. (2004). The problem with

[2] An illustrative statistical test is to rely on the percentage of synthetic 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.

[3] $((1 - v(B))/(1 - E[v(B)])) \cdot (E[v(B)]/v(B))$, where $v(B)$ is the fraction of transactions with some but not all φ_B items, and $E[v(B)]$ is the expectation of $v(B)$ in a random dataset (Aggarwal and Yu 1998).

the use of ratios is that, instead of measuring the probability of a pattern's occurrence, ratios are subjective indicators of its relevance.

To address this problem, statistical tests have been proposed. χ^2 tests were proposed by Silverstein et al. (1998) (and revised by DuMouchel and Pregibon 2001) to assess a pattern based on the degree of dependence among its constituent items using synthetic data. DuMouchel (1999) and DuMouchel and Pregibon (2001) proposed Bayesian assessments with shrinkage estimates to provide a conservative true probability of support's significance and hence minimize the number of false discoveries. The statistical significance has been also estimated from a Bayesian network with parameters derived from Θ (Jaroszewicz and Scheffer 2005). However, these Bayesian estimates neither provide a general mechanism for applying hypothesis tests nor assess deviations from expectations. Kirsch et al. (2012) 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. Although global parameters can be inferred efficiently and used right-away as an heuristic for biclustering searches, small yet significant biclusters (very low φ_B) are incorrectly seen as false positive discoveries.

Finally, the works of Ramon et al. (2013), Bellay et al. (2011) and Califano et al. (2000) test the probability of constant biclusters to occur against randomized data with different forms of background noise. However, these tests are deemed to assess solutions of a specific algorithm, not being generalizable.

Deviation from expectations. To guarantee that the probability of occurrence of a bicluster deviates from the expected probability, its significance needs to be corrected against the space of similar biclusters. The space of similar biclusters depends essentially on the allowed coherencies, placed similarity criteria (e.g. biclusters with the same area or pattern length) and applicable relaxations (Bay and Pazzani 2001). For a given space, 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). To avoid a large increase in the number of neglected relevant biclusters (false negatives), non-conservative options, such as Holm (1979) and Hochbert procedures, can be used and still verify the FWER constraint. Alternatively, non-FWER procedures for multi-hypotheses can be placed to minimize false negatives, while still providing adequate guarantees on the risk of false positives (Benjamini and Yekutieli 2001; Benjamini and Hochberg 1995).

4 Statistical evaluation of biclustering solutions

The solution space is organized as follows. First, we extend the surveyed statistical views in the context of pattern mining to assess constant biclustering solutions Sect. 4.1). In this context, new statistical tests are proposed together with a new correction procedure able to simultaneously minimize type-I errors (false positives) and type-II errors (false negatives). These contributions are then further extended to: (1) assess additive, multiplicative, symmetric and order-preserving models Sect. 4.2); (2) assess bicusters with arbitrary levels of noise and missings (Sect. 4.3); and (3) address

intrinsic challenges related with the assessment of real-valued biclusters (Sect. 4.4). Finally, BSig (**B**iclustering **Sig**nificance) method is introduced (Sect. 4.5).

4.1 Significance of constant biclusters

Given a set of items $\mathcal{L}$, a discrete bicluster B is referred as *perfect* if it does not contain noisy elements, $a_{ij} \in \mathcal{L} \wedge \eta_{ij} = 0$. Under the mapping between pattern mining (PM) and biclustering proposed by Henriques et al. (2015), the surveyed PM-based statistical views (Sect. 3.2) are applicable to assess discrete, perfect constant biclusters. To this aim, and according to related work, local and global statistical tests can be defined (against the approximated data regularities, permuted/randomized data or hold-out partitions) using either stochastic or frequentist views with different degrees of dependence. Fixing these decisions essentially depends on: *1)* the properties of the input matrix, and *2)* the selected biclustering approach (as it determines the type and noise-tolerance of 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 first provide a structured view on how these contributions can be consistently combined, and then propose a variant of the Hochbert procedure to tackle the efficiency bottlenecks of non-conservative FWER corrections.

Probability of occurrence p_B. When coherency is observed across rows, binomial tails can be used to robustly compute the probability of a bicluster $B = (I, J)$ with pattern φ_B to occur across a set of rows, $p'_B = P(Z \geq |I|)$ with $Z \sim \text{Bin}(p_{\varphi_B}, |X|)$, where p_{φ_B} is the probability of the φ_B items to occur. Consider $n = |I|$, $m = |J|$, $N = |X|$ and $M = |Y|$ to be, respectively, the number of rows and columns in a given bicluster and the number of rows and columns of the inputted dataset. This probability, given by (1), essentially depends on p_{φ_B} and on the size of both the bicluster and the input matrix. The significance needs to be adjusted by the probability of the bicluster to occur for any combination of columns, which is approximately given by $p_B = 1 - P(Z < |\ I\ |)^{\binom{M}{m}} = \binom{M}{m}\text{P}(\text{Z} \geq |\ I\ |)$ when assuming that columns have a similar distribution of items.

$$p_B = \binom{M}{m} \sum_{x=n}^{N} \binom{N}{x} {p_{\varphi_B}}^x (1 - p_{\varphi_B})^{N-x} \tag{1}$$

Although the previous calculus allows for $p_B > 1$ when it is likely to observe the occurrence of more than one bicluster with φ_B pattern and at least n rows occur for a given matrix, this probability can be bounded by 1 or, alternatively, used for subsequent $p_B \approx 0$ hypothesis testing.

Similarly, a bicluster with coherency across columns has $p_B = \binom{N}{n} P(Z \geq |\ J\ |)$ (2) with $Z \sim \text{Bin}(p_{\varphi_B}, M = |\ Y\ |)$. Variants can be defined to further assess biclusters with a constant value ($\sigma \in \mathcal{L}$) on both rows and columns. For this case, 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

by Koyuturk et al. (2004) and Bellay et al. (2011) 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.

The p_{φ_B} **calculus** essentially depends on the regularities underlying data, Θ, which 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_{φ_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 considering multiple orderings of φ_B values (e.g. $\Sigma_{j=4}^{7}\Pi_{a\in\{3,6,2,4\}}P(y_j = a)$). For this case, we suggest the use of dynamic programming to avoid the redundant computation of subsets of products. Figure 4 assumes items to be ordinal ($a_{ij} \in \{0..6\}$) to show how the different Θ approximations impact the assessment of a perfect constant bicluster with coherency across rows. Both univariate and multivariate distributions are applied, as well as varying forms of dependency, to stress the relevance of adequately defining a null data model. The coherency strength as implicitly defined by the number of symbols ($|\mathcal{L}| = 7$) affects both sides of the testing equation: the probability of φ_B occurrence (and consequently p_B) and the space of similar biclusters (and consequently the significance level of the applied correction).

As illustrated, frequentist distributions can be considered to compute p_{φ_B}. They are the default option when either a pairwise or overall form of *dependency* among items in φ_B is assumed. In this context, the possible combinations of dependent items in φ_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,

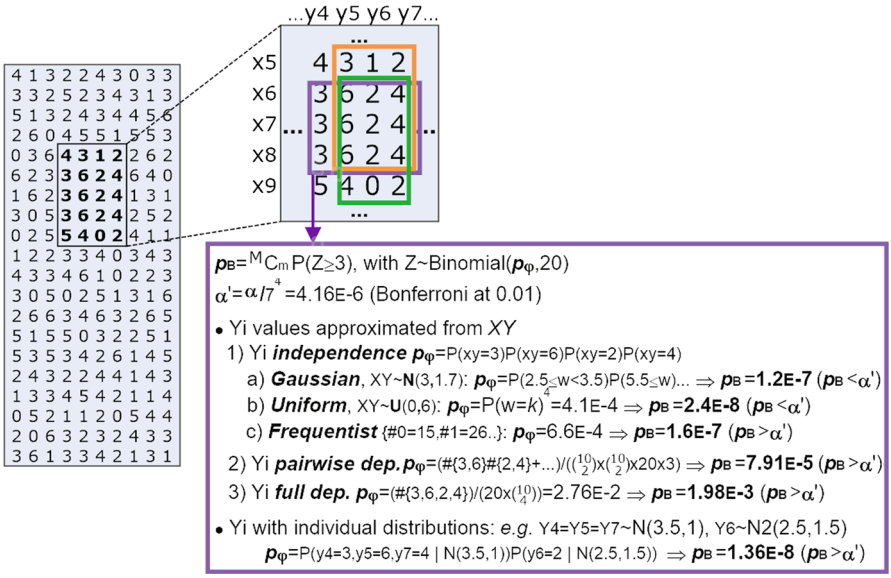

Fig. 4 Illustrative assessment of a (non-noisy) constant bicluster in discrete settings

there is the need to define principles for combining their influence to compute the probability of the overall pattern, p_{φ_B} (see the pairwise and overall dependence scenarios provided in Fig. 4). Note that a high-order 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.

In the presence of small matrices, frequentist views are susceptible to overfit the observed data. For these cases, a more robust strategy is to generate h background datasets and replace the Binomial calculus by a test, such as an unilateral t-Student, based on the h support estimates of φ_B.

This section extended the statistical principles proposed in the context of pattern mining to: 1) allow an accurate p_{φ_B} calculus (when considering different distributions and forms of dependency), 2) guarantee their applicability to biclusters with different forms of constant coherency, 3) avoid efficiency bottlenecks associated with the p_{φ_B} calculus for bicluster's patterns with non-indexed columns, and 4) avoid the computationally expensive task of performing an arbitrary-high number of counts.

Robust and efficient corrections. Given $\varphi_B = \{3,\ 6, 2, 4\}$, when assuming columns to be fixed, the space of similar biclusters is defined by the set of all patterns with the same size (7^4 biclusters). Otherwise, larger spaces need to be considered ($7^4\binom{10}{4}$ biclusters), thus increasing both the probability of a pattern to occur p_{φ_B} and the correction effect (from testing multiple hypothesis). When considering the Bonferroni correction, the considered α confidence is simply divided by the space size s. Since this correction assumes the space of other similar biclusters to be more significant, $p_B > \alpha/s$ does not imply that the occurrence of B is not significant. Thus, to avoid a high number of false negatives (rejected biclusters that are significant), the Hochbert correction can be applied. The p values from the s biclusters are sorted, $\{p_{B_1}, .., p_{Bs}\}$, and the p value $max_{p_{Bj}} : \forall_{1\leq j\leq s}\ p_{Bj} \leq \alpha/(s - j + 1)$ is outputted as the corrected level. Understandably, 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 make 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 order of items by frequency is {3,2,4,5,1,6,0}, BSP starts by computing the probability for bicluster with $\varphi_B = \{5, 5, 5, 5\}$, and compares its probability with the corrected significance $\alpha/(7^4/2)$. If it is lower, then BSP compares p_B with $\varphi_B = \{6, 6, 6, 6\}$ against $\alpha/(7^4/2 + 7^4/4)$, and, if lower, BSP tests p_B with $\varphi_B = \{6, 6, 0, 0\}$ and so forth. A total of 10 tests provides already a good approximation of the correct significance with heightened efficiency.

4.2 Significance of biclusters with non-constant coherency

Despite the relevance of the contributions from previous section, they cannot be applied to assess biclusters with more flexible forms of coherency. In order to extend the pro-

posed principles for non-constant biclusters, two aspects need to be carefully revised: *1)* the p_{φ_B} calculus (affecting the Binomial tail) and, *2)* the space of similar biclusters (affecting the deviation analysis).

Additive biclusters. Consider $\mathcal{R}$ to be a finite set of integers with bijective correspondence to the set of items $\mathcal{L}$. Contrasting with constant biclusters, the set of items per row (column) can vary for an additive coherency across rows (columns). Given $\mathcal{L} = \{0..3\}$, if the items $\{2, 0, 1\}$ are observed as a row of an additive bicluster, $\{3, 1, 2\}$ ($\gamma = 1$) can also be observed as an alternative row. The pattern φ_B is given by the underlying items in the absence of shifting factors ($\gamma = 0$). p_{φ_B} thus differs from the constant assumption since new combinations of items are allowed due to the presence of shifting factors. The allowed number of shifts of an additive pattern φ_B is determined by the difference between the maximum allowed value, $max_{a_{ij}}(A)$, and the highest value in φ_B, $max_{a_{ij}}(\varphi_B)$, as well as by $min_{a_{ij}}(\varphi_B) - min_{a_{ij}}(A)$. These ranges are used to generate the shifting factors required to compute p_{φ_B}.

Although the probability of an additive pattern to occur is higher than that of a constant pattern respecting the same φ_B, the size of the space of additive biclusters is smaller leading to a subtler correction and higher testing significance level. Both sides of the equation are affected. The space size, s, of an additive bicluster is defined by:

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

The intuition behind this calculus is to count the combination of patterns with different amplitudes. When the amplitude is $\Delta = 0$, 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 $\Delta = 1$, 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 φ_B values. Consider $m = 4$ and the amplitude to be 3, then the last parcel of (2) is given by $\binom{4}{3}\binom{1}{1}2^0+\binom{4}{2}(\binom{2}{2}2^0+\binom{2}{1}2^1)+...$, capturing the possible combinations of values with this amplitude.

Multiplicative biclusters. Similarly, multiplicative coherency impacts both p_{φ_B} and the applied correction. Given $\mathcal{L} = \{-6..6\}$, if the pattern $\varphi_B = \{3, -1, 2\}$, 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 space of similar biclusters, which essentially depends on the amplitude of values, the multiplicative 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 $\gamma \in \mathbb{N} \wedge i \in \mathbb{N}$, until all the range of items are covered. The number of scaling factors (space of similar multiplicative biclusters) for a specific φ_B with m columns is typically less than the number of shifting factors (space of similar additive biclusters) and is given by: $\sum_{\varphi \in \mathcal{L}^m}(gcd(\varphi) = 1)$ (3), where $gcd(\varphi)$ is the greatest common divisor of the m values of φ pattern. The idea behind this calculus is to remove the

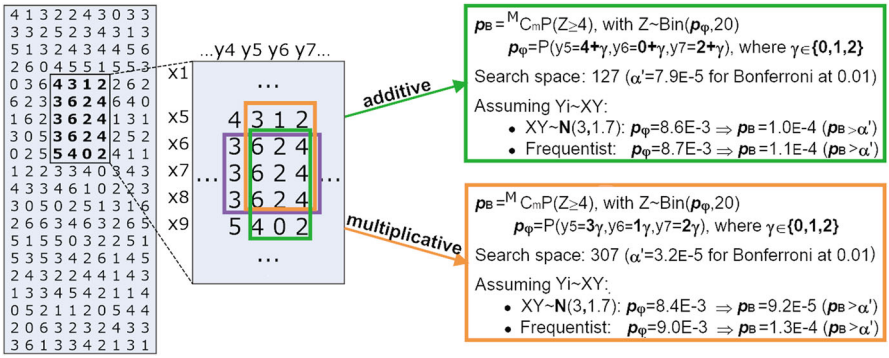

Fig. 5 Illustrative tests of non-noisy and discrete additive and multiplicative biclusters

patterns with scaling factors $\gamma \neq 1$. If $gcd(\varphi)$ differs from 1, this means that the pattern can be derived from a simpler φ pattern with $gcd(\varphi) = 1$. To avoid the generation-and-test of all possible combinations of items, corrections can be performed using a depth-first search and 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.

Figure 5 provides an illustrative assessment of an additive and multiplicative bicluster. It shows how the search space size and the enhanced $p_{\varphi_\mathbf{B}}$ calculus affect the corrected significance threshold and the observed p value $p_\mathbf{B}$.

Order-preserving biclusters. Order-preserving biclusters are known for their flexibility, embedding constant, additive and multiplicative models (Henriques and Madeira 2014a). Their columns (rows) define a linear ordering (monotonically increasing) respected across rows (columns). Order-preserving biclusters can simultaneously capture constant, additive and multiplicative coherencies. A bicluster with m columns is described by one of the $m!$ possible linear orderings. Thus, the probability of occurrence and the size of the space of similar 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 number of similar biclusters, $m!$. The optimum significance level is thus $\alpha/m!$. Based on these observations, global properties, such as the minimum number of rows of a bicluster with m columns, can be directly inferred for a $N \times M$ matrix by satisfying the binomial calculus: $P(Z \geq n) < \alpha/m!$, with $Z \sim \text{Bin}(1/m!, N)$.

Symmetric biclusters. A bicluster following a symmetric assumption can consider symmetries on rows (or columns). Illustrating, the patterns {2,4,3} and {-1,-3,-2} cannot be described by a single additive or multiplicative model, but can be described by a symmetric additive model. Symmetries can be also accommodated with orderings (Henriques and Madeira 2014a). Understandably, the presence of symmetries will also cause the probability of occurrences to be larger as well as the corrected testing significance level (as a result of a smaller space of similar biclusters). Similarly to the previous coherencies, p_{φ_B} is computed by adding the probabilities with the symmetries

of the allowed patterns associated with φ_B. For non-constant models, this means to include the possible symmetries on each scaling or shifting factor associated with φ_B. The space of a symmetric bicluster is approximately half of the original search space. For instance, when considering symmetries over constant biclusters, the number of comparable biclusters is given by $\mid \mathcal{L} \mid^m /2 - 1$.

Biclusters with plaid effects. 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 (I_k, J_k) for the a_{ij} element (0 if $x_i \notin I_k \vee y_j \notin J_k$) and $a_{ij} \in \mathcal{R}$. In accordance with the original definition of the plaid model (Lazzeroni and Owen 2002), θ_{ijk} is able 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_{φ_B} calculus and applied correction assume that the underlying pattern is $\varphi_B = \{0, 2, 1\}$.

4.3 Significance of noisy biclusters

Despite the relevance of the proposed statistical tests, they cannot be applied as-is to 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. A noisy discrete bicluster is a bicluster where some of its elements $a_{ij} \in B$ do not respect the overall coherency criteria, $\eta_{ij} \neq 0$.

Assessing noisy biclusters. In order to compute the correct probability p_B in these cases, we propose a strategy that aims to identify the original φ_B pattern. 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 φ_B is given by the mode of items per column $y_j \in J$. For additive and multiplicative biclusters, this calculus is performed in three steps. First, the shifting or scaling factor associated with each row in I are computed by assuming that the row-conditional values do not have noise. Illustrating, given an additive bicluster with the $\{2, 3, 3\}$ items on row x_i, a shifting factor $\gamma = 2$ is assigned to x_i. Second, the computed factors are applied for each row in I to retrieve the set of possible patterns (including $\{0,1,1\}$ from row x_i). The set of possible patterns form a (noisy) constant bicluster. In this context, the mode is applied on each column to retrieve the true φ_B for the p_{φ_B} calculus. Similarly, symmetric models rely on the identification and removal of

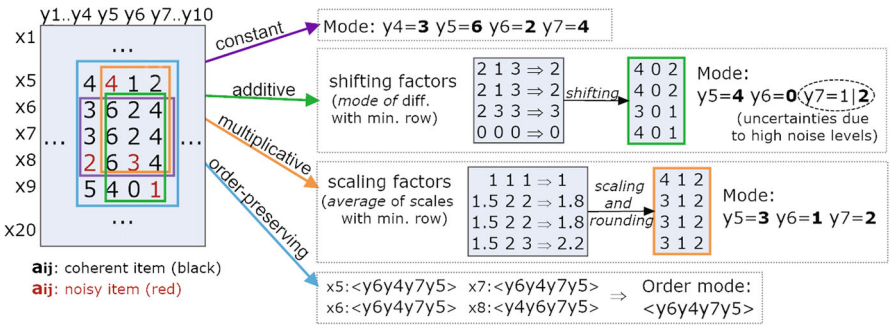

Fig. 6 Retrieval of the true pattern φ_B of noisy biclusters with varying coherency

symmetric factors in order to identify the true 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 the assessment of noisy order-preserving biclusters does not depend on the retrieval of the underlying true pattern, the mode of permutations can be computed to recover the underlying orderings. Figure 6 illustrates these strategies. Given the true pattern, the proposed tests in previous sections can be directly applied to assess the significance of noisy biclusters.

Assessing biclusters with missing values. An increasing number of biclustering algorithms is able to accommodate missings in the outputted biclusters (Henriques et al. 2015). These contributions open the possibility to discover biclusters from sparse matrices, where a bicluster may be associated with an arbitrary-high number of missing elements. To enable the assessment of biclusters with missing elements, 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 column. In this context, missings are removed from the mode calculus. In the presence of biclusters with both noisy and missings, this strategy can be consistently combined with the previous assessment of noisy biclusters.

Combining significance and homogeneity. 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 can 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. Illustrating, consider a bicluster with 5 rows and 3 columns and a total of 5 elements deviated from the expected true pattern. In this context, $\epsilon = \frac{1}{3}$ 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 $P(n \leq x \mid (\Sigma_{a_{ij} \in B} \eta_{ij}) < \epsilon)$. Due to the inherent 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 (by testing the amount of value deviations). 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.

4.4 Significance of real-valued biclusters

The applicability of the previous statistical views towards real-valued biclusters is dependent on the application of discretization procedures, which can introduce uncertainty, and does not account for the possibility of assessing bicluster with continuous adjustment factors, $\gamma \in \mathbb{R}$ (where γ was introduced in Definition 2). Below we describe how can discrete assessments be applied for this end without incurring in undesirable drawbacks associated with the item-boundaries problem, and propose an alternative strategy able to robustly bound the significance of real-valued biclusters.

"Appendix A2" extends these procedures with new 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.

Noise-free discretization. The first option towards the analysis of real-valued biclusters is to discretize data, thus mapping the discovered biclusters into discrete biclusters. To fix an adequate alphabet length, the coherency strength, δ (Definition 3), needs to be either given or estimated. The majority of available biclustering methods explicitly define a coherency strength. For the few cases where coherency strength might not be available, it can be inferred by analyzing the deviations from the expected values. When the inputted data is associated with ranges of coherent values δ that differ for different biclusters, a discretization procedure can be directly applied for each bicluster (bypassing the need to estimate an overall coherency strength). The pros and cons of alternative discretization methods, such as equal-depth, bin and distribution-centered methods, have been largely discussed in literature (Carmona-Saez et al. 2006; Mahfouz and Ismail 2009; Okada et al. 2007).

The mode calculus (Sect. 4.2) is then applied to deal with both structural noise and the noise introduced from the applied discretization (associated with the item-boundaries problem). In this context, after retrieving the discretized φ_B, the calculus of p_B should rely on continuous distributions approximated from the original values. The

use of a continuous probabilistic view is preferable over a frequentist view in order to minimize the errors associated with the applied discretization. Figure 7a illustrates this strategy using statistical break-points of a Gaussian distribution for data discretization. However, real values near break-point boundaries can be assigned to different items, leading to more uncertainty when retrieving the mode pattern.

To tackle this problem, elements with values near break-points can be assigned to more than one item. The underlying idea is to reduce the influence of noisy elements during the mode calculus. As such, 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 $y_2 \in J$: $\{a_{1,2} = \{1\}, a_{3,2} = \{0, 1\}, a_{5,2} = \{1\}, a_{6,2} = \{0, 1\}\}$. 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. This strategy is also illustrated in Fig. 7a.

Range-based assessment. An alternative strategy, and the default option in BSig, is to rely on multiple probability estimates, one estimate per row (or column) in the bicluster, with the coherency range δ applied around the observed value. For a constant bicluster with coherency across rows:

$$\hat{p}_{\varphi_B} = T\left(\bigcup_{x_i \in I}\left\{p^i_{\varphi_\mathbf{B}}\right\}\right), \quad where \quad p^i_{\varphi_\mathbf{B}} = \prod_{y_j \in J} P(a_{ij} - \delta/2 \leq Y_j \leq a_{ij} + \delta/2) \tag{3}$$

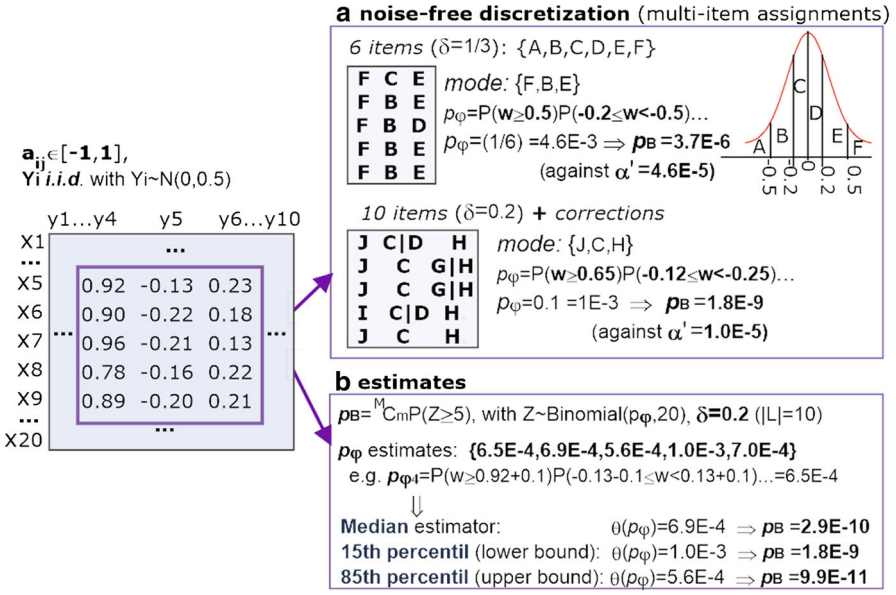

Fig. 7 Strategies to extend the significance assessment to real-valued biclusters

where $p^i_{\varphi_B}$ is the probability of the set of observed items in ith row to occur, Y_j is a random variable (with distribution drawn from values in y_j column), and T is the estimator of the true probability from the inputted n estimates. This estimator $\hat{p}_{\varphi_B}$ is then used to compute the Binomial tails in order to calculate the estimator of the true probability, $p_B = \binom{M}{m} \Sigma_{x=n}^{N} \binom{N}{x} (\hat{p}_{\varphi_B})^x (1 - \hat{p}_{\varphi_B})^{N-x}$ (where $N = |X|$, $M = |Y|$, $n = |I|$ and $m = |J|$). Alternatively, given a set of estimates p^i_B based on $p^i_{\varphi_\mathbf{B}}$ estimates, the estimator of the true probability can be directly given by $\hat{p}_B = \mathrm{T}(\{p^1_B, p^2_B, .., p^n_B\})$. For both options, the estimator T needs to be adequately defined. We propose the median estimate and percentiles, such as the 15th and 85th percentiles, to model the true probability of occurrence (p_{φ_B} and p_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 by Brown (1947)). Figure 7b illustrates this strategy.

Equation (3) is natively prepared to assess constant biclusters, yet it can be consistently extended towards additive, multiplicative, plaid and symmetric models by applying the principles from two previous sections.

4.5 BSig method

To guarantee the correct application of the principles proposed throughout Sects. 4.1–4.4, we propose BSig (**B**iclustering **Sig**nifcance). BSig is described in *Algorithm* 1. Four major steps are considered. *First*, the null data model, Θ, is determined. To this end, when independence between elements is assumed (default setting), Θ is directly approximated from data by fitting multivariate distributions (according to tests and fitting measures surveyed by Karian and Dudewicz 2010). 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. 2008) for the remaining cases.

Second, the coherency assumption and coherency strength of a given bicluster are identified. To this end, the adjustment factors are approximated to check whether the target bicluster is well described by a constant, additive or multiplicative coherency; if there are localized forms noise explained by plaid effects; or if there are orderings skewing previous checks. In the context of real-valued data, adjustment factors are then removed and value deviations from expectancies computed to estimate the coherency strength required for the proposed statistical tests.

Third, p_{φ_B} is estimated using Θ regularities and the principles from Sects. 4.2–4.4 to deal with non-constant, noisy and (possibly) real-valued aspects. Only then, statistical tests grounded on the calculus of binomial tails (1) are applied to compute p_B.

Fourth, the corrected significance is efficiently computed using the proposed Hochbert correction by generating similarly-sized biclusters with varying φ_B pattern according to a binary space partitioning procedure. Finally, the gathered p value is tested against this corrected significance threshold.

BSig is applied under three major assumptions: 1) considers that the coherency of a given bicluster is well-defined; 2) when coherency strength is not known, it is empirically estimated; and 3) relies on data fitting tests to guarantee an adequate null data model. To our knowledge, these assumptions (estimation of the underlying coherency and data regularities) should pertain to any method aiming to statistically assess flexible biclustering solutions.

Algorithm 1: BSig core steps

Input: A data and $\mathcal{B}$ biclusters
$\Theta \leftarrow$ fitDist(A), $\Theta^T \leftarrow$ fitDist(A^T) *//continuous vs. count-based*
foreach $B \in \mathcal{B}$ **do**
  if orientation(B)=*column* **then** $B \leftarrow B^T$, $A \leftarrow A^T$, exchange(Θ,Θ^T)
  factors $\leftarrow$ retrieveAdjustmentFactors(B,A)
  if $a_{ij} \in \mathbb{R}$ **then** $\delta \leftarrow$ estimateStrength(B,A,factors)
  else $\delta \leftarrow 1$ *//assuming* $\mathcal{L} \rightarrowtail \mathbb{N}$ *(injective)*
  if *hasPlaidEffects(*B,A,*factors)* **then**
    plaids $\leftarrow$ removePlaidEffects(B,A,factors,$\mathcal{B}$)
  coherency $\leftarrow$ estimateCoherencyAssumption(B,A,factors)
  if coherency=*orderPreserving* **then** $p_B \leftarrow$ permutationTest(B) *//Section 4.2*
  else if $a_{ij} \in \mathbb{R}$ **then**
    if coherency=*constant* **then** $p_{\varphi_B} \leftarrow$ Eq.3 (with Θ, δ) *//Section 4.4*
    else if coherency=*additive* **then**
      $p_{\varphi_B} \leftarrow$ Eq.A2 (with Θ, δ, factors) *//Appendix A2*
    else $p_{\varphi_B} \leftarrow$ Eq.A3 (with Θ, δ, factors) *//Appendix A2*
    $p_B \leftarrow$ Eq.1 (with p_{φ_B}) *//Section 4.4*
  else
    $\varphi_B \leftarrow$ noiseTolerantPatternEstimation(B,coherency,factors) *//Section 4.3*
    $\Phi_B \leftarrow$ validPatterns(φ_B,coherency,factors) *//Section 4.2*
    if *small(*A*)* **then** $p_B \leftarrow$ noiseTolerantCountsOnBackgroundData(Θ,h,Φ_B)
    else
      $p_{\varphi_B} \leftarrow$ computeProb(Φ_B) *//Section 4.2*
      $p_B \leftarrow$ Eq.1 *//Section 4.1*
  $p_B \leftarrow$ noiseProbCalibration(p_B,B,factors) *//Section 4.3*
  $\alpha \leftarrow$ 1E-3; $\varphi_S \leftarrow \emptyset$
  if $a_{ij} \in \mathcal{L}$ **then**
    foreach $i \in \{1..10\}$ **do**
      $\varphi_S \leftarrow$ binaryPartitioning(φ_B,φ_S)
      if *HolmVerification(*φ_S,α*)* **then** break
    $\alpha' \leftarrow$ HolmCorrection(φ_S,Θ)
  if orientation(B)=*column* **then** $A \leftarrow A^T$, exchange(Θ,Θ^T)
  B.decision $\leftarrow$ t-Test(p_B,α')

5 Results

The results were collected and analyzed in five steps. First, we undertake an in-depth analysis of how the properties of biclusters and input data determine significance. Second, we provide evidence for the soundness of the proposed statistical tests and measure the impact of varying statistical decisions on the observed significance. Third, we motivate the relevance of assessing the statistical significance of biclustering models. Fourth, we show how the coherency of biclusters affects the space of similar biclusters and, consequently, the applied correction. Finally, we provide an initial comparison of state-of-the-art biclustering algorithms with varying coherency assumption according to both the significance and homogeneity of their outputs.

The proposed statistical methods are included in BSig toolbox.[4] and implemented in Java (JVM v1.6.0-24) The experiments were run using an Intel Core i3 1.80GHz with 6GB of RAM.

For the experiments, we generated synthetic data and considered five gene expression datasets[5,6]: *dlbcl* to study responses to chemotherapy (Rosenwald et al. 2002), *hughes* to characterize nucleosome occupancy (Lee et al. 2007), *gasch* to measure yeast responses to varying environmental stimuli (Gasch et al. 2000), *ycycle* to study yeast cell cycle (Tavazoie et al. 1999) and *ystress* to study yeast gene expression in response to stress (Eisen et al. 1998).

Impact of n, m, N, M and p_{φ_B} in Significance Tables 1, 2 and 3 describe how the significance of a bicluster varies with its size and coherency, with the size and regularities of the original matrix, and with the probability of the φ_B pattern. These analyzes determine the expected minimum size of a bicluster that guarantees its significance. These size expectations can be used to guide biclustering algorithms and classifiers reliant on discriminative biclusters.

Tables 1 and 2 show how the expected minimum number of rows in a bicluster varies with the considered coherency strength, number of columns in the bicluster m, and data size (number of rows, N, and columns, M). We assume p_{φ_B} to have items with average, above-average and below-average probability of occurrence. We observe that when moving from constant to more flexible coherencies (or looser coherency strength), larger biclusters are required to preserve significance. The number of data rows, N, impacts significance as it shapes the binomial tails. Given $|\mathcal{L}| = 5$, $m = 5$ and $M = 100$, the expected minimum number of rows is $\hat{n} = \text{x}$, $\hat{n} = 17$ and $\hat{n} = 24$ for respectively a constant, multiplicative and additive ($\Delta = 0$) bicluster when $N = 2000$, and $\hat{n} = \text{x}$, $\hat{n} = 81$ and $\hat{n} = 153$ for the same biclusters when $N = 50000$. The number of data columns, M, affects $\hat{n}$ in a less accentuated way when assuming coherency across rows. If coherency across columns is targeted, we would observe the inverse effect: increased sensitivity to the number of data columns. Coherency strength (number of items) also largely impacts significance. Given $m = 5$, $N = 20000$ and $M = 100$ and constant coherency, for differential expression ($|\mathcal{L}| = 3$): $n \in [70, 309]$ with $\hat{n} = 153$, while when considering $|\mathcal{L}| = 5$: $n \in [20, 53]$ with $\hat{n} = 33$. Finally, the pattern length m strongly affects p_{φ_B} and thus the expected number of rows. 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 $m = 7$ under a constant coherency ($\hat{n} = 610$, $\hat{n} = 59$ and $\hat{n} = 18$ under an additive coherency with $\Delta = 2$).

Table 3 explores structural properties of order-preserving biclusters. Order-preserving significance analysis does not directly depend on data regularities, Θ. Here, the minimum number of rows that guarantees the significance of an order-preserving bicluster is highly sensitive to the pattern length m – e.g. $n = 1012$ rows for $m = 4$ and $n = 15$ rows for $m = 8$ (given $N = 20000$ and $M = 100$) – and to the number

[4] Available in http://web.ist.utl.pt/rmch/software/bsig/.

[5] http://www.bioinf.jku.at/software/fabia/gene_expression.html.

[6] http://chemogenomics.stanford.edu/supplements/03nuc/datasets.html.

Table 1 Impact of the number of items, $|\mathcal{L}|$, and pattern length, m, on the expected minimum number of rows ($\mathbf{n}_{min}$) in the bicluster that guarantees the statistical significance (in accordance with Algorithm 1) of constant, additive and multiplicative biclusters for a fixed data size (N and M)

		m	5	5	5	5	**2**	**3**	**5**	**7**
		N	20000	20000	20000	20000	20000	20000	20000	20000
		M	100	100	100	100	100	100	100	100
		$\mid \mathcal{L} \mid$	**3**	**4**	**5**	**10**	5	5	5	5
Constant	$p_{\varphi_B} = \frac{1}{\mid L \mid}^m$	$\mathbf{n}_{min}$	**153**	**59**	**33**	**11**	**930**	**93**	**33**	**12**
	$p_{\varphi_B} \in [\frac{0.8}{\mid L \mid}^m, \frac{1.2}{\mid L \mid}^m]$	$\mathbf{n}_{min}$	[70,309]	[32,106]	[20,53]	[7,14]	[619,1304]	[139,375]	[20,53]	[9,18]
	$\alpha' = \frac{\alpha}{\mid L \mid^m}$	α'	2.1E-4	4.9E-5	1.6E-5	5.0E-7	2.0E-3	4.0E-4	1.6E-5	6.4E-7
Additive with $\Delta = 0$	$p_{\varphi_B} = \frac{1}{\mid L \mid}^m \mid L \mid$	$\mathbf{n}_{min}$	**362**	**149**	**81**	**21**	**4270**	**982**	**81**	**20**
	$p_{\varphi_B} \in [\frac{0.8}{\mid L \mid}^m \mid L \mid, \frac{1.2}{\mid L \mid}^m \mid L \mid]$	$\mathbf{n}_{min}$	[150,784]	[70,299]	[42,151]	[14,31]	[2762,6027]	[525,1580]	[42,151]	[13,32]
	$\alpha' = \alpha / Eq.(3.2)$	α'	2.4E-4	6.4E-5	2.4E-5	1.2E-6	5.6E-3	8.2E-4	2.4E-5	8.1E-7
Additive with $\Delta = 2$	$p_{\varphi_B} = \frac{1}{\mid L \mid}^m (\mid L \mid -2)$	$\mathbf{n}_{min}$	**153**	**92**	**59**	**19**	**2621**	**610**	**59**	**18**
	$p_{\varphi_B} \in [\frac{0.8}{\mid L \mid}^m (\mid L \mid -2), \frac{1.2}{\mid L \mid}^m (\mid L \mid -2)]$	$\mathbf{n}_{min}$	[70,308]	[46,174]	[32,105]	[14,28]	[1700,3682]	[337,988]	[32,105]	[11,26]
	$\alpha' = \alpha / Eq.(3.2)$	α'	2.4E-4	6.4E-5	2.4E-5	1.2E-6	5.6E-3	8.2E-4	2.4E-5	8.1E-7
Multiplicative with (γ_1, γ_2)	$p_{\varphi_B} = 2\frac{1}{\mid L \mid}^m$	$\mathbf{n}_{min}$	**259**	**91**	**47**	**12**	**1786**	**427**	**47**	**15**
	$p_{\varphi_B} \in [2\frac{0.8}{\mid L \mid}^m, 2\frac{1.2}{\mid L \mid}^m]$	$\mathbf{n}_{min}$	[111,549]	[46,173]	[27,80]	[10,16]	[1162,2499]	[240,684]	[27,80]	[10,22]
	$\alpha' = \alpha / Eq.(3.3)$	α'	4.1E-4	1.0E-4	3.5E-5	1.8E-6	5.0E-3	9.8E-4	3.5E-5	1.3E-6

Table 2 Impact of data size on the expected minimum number of rows ($\mathbf{n}_{min}$) in the bicluster that guarantees the statistical significance (in accordance with Algorithm 1) of constant, additive and multiplicative biclusters (for a fixed number of items and pattern length)

		m **N** **M** $\mid \mathcal{L} \mid$	5 **200** 100 5	5 **500** 100 5	5 **2000** 100 5	5 **10000** 100 5	5 **50000** 100 5	5 20000 **50** 5	5 20000 **200** 5	5 20000 **1000** 5
Constant	$p_{\varphi_B} = \frac{1}{\|L\|}^m$	$\mathbf{n}_{min}$	**8**	**9**	**14**	**24**	**44**	**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]	[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	1.6E-5	1.6E-5	1.6E-5
Additive with $\Delta = 0$	$p_{\varphi_B} = \frac{1}{\|L\|}^m \mid L \mid$	$\mathbf{n}_{min}$	**11**	**14**	**24**	**53**	**153**	**77**	**85**	**94**
	$p_{\varphi_B} \in [\frac{0.8}{\|L\|}^m \mid L \mid, \frac{1.2}{\|L\|}^m \mid L \mid]$	$\mathbf{n}_{min}$	[8,14]	[10,19]	[16,36]	[30,93]	[71,307]	[39,146]	[44,157]	[50,170]
	$\alpha' = \alpha / Eq.(3.2)$	α'	2.4E-5	2.4E-5	2.4E-5	2.4E-5	2.4E-5	2.4E-5	2.4E-5	2.4E-5
Additive with $\Delta = 2$	$p_{\varphi_B} = \frac{1}{\|L\|}^m (\mid L \mid -2)$	$\mathbf{n}_{min}$	**10**	**12**	**19**	**40**	**107**	**56**	**62**	**69**
	$p_{\varphi_B} \in [\frac{0.8}{\|L\|}^m (\mid L \mid -2), \frac{1.2}{\|L\|}^m (\mid L \mid -2)]$	$\mathbf{n}_{min}$	[8,12]	[9,16]	[13,28]	[24,67]	[52,105]	[30,101]	[34,110]	[40,120]
	$\alpha' = \alpha / Eq.(3.2)$	α'	2.4E-5	2.4E-5	2.4E-5	2.4E-5	2.4E-5	2.4E-5	2.4E-5	2.4E-5
Multiplicative with (γ_1, γ_2)	$p_{\varphi_B} = 2\frac{1}{\|L\|}^m$	$\mathbf{n}_{min}$	**9**	**11**	**17**	**33**	**81**	**44**	**49**	**54**
	$p_{\varphi_B} \in [2\frac{0.8}{\|L\|}^m, 2\frac{1.2}{\|L\|}^m]$	$\mathbf{n}_{min}$	[6,11]	[8,14]	[12,24]	[20,52]	[42,151]	[25,77]	[29,84]	[33,90]
	$\alpha' = \alpha / Eq.(3.3)$	α'	3.5E-5	3.5E-5	3.5E-5	3.5E-5	3.5E-5	3.5E-5	3.5E-5	3.5E-5

Table 3 Minimum number of order-preserving rows ($\mathbf{n}_{min}$) that guarantees significance across m columns for data with N rows (in accordance with Algorithm 1)

m	**4**	**5**	**6**	**8**	**10**	6	6	6	6	6	6	6	6
N	20000	20000	20000	20000	20000	**200**	**500**	**2000**	**10000**	**50000**	20000	20000	20000
M	100	100	100	100	100	100	100	100	100	100	**50**	**200**	**1000**
α'	2.1E-3	4.2E-4	6.9E-5	1.2E-6	1.4E-8	6.9E-5	6.9E-5	6.9E-5	6.9E-5	6.9E-5	6.9E-5	6.9E-5	6.9E-5
$\mathbf{n}_{min}$	**1012**	**262**	**76**	**15**	**7**	**11**	**14**	**23**	**50**	**141**	**72**	**79**	**88**
p_B	1.9E-3	3.4E-4	4.0E-5	7.4E-8	4.3E-11	1.2E-5	3.4E-5	4.6E-5	5.9E-5	3.3E-5	3.0E-5	4.4E-5	2.9E-5

of data rows N – e.g. ($n = 14, m = 6$) for $N = 500$ rows and ($n = 50, m = 6$) for $N = 10000$ (given $M = 100$).

Soundness. The analysis provided in Table 4 tests the relevance of the proposed statistical tests to recover only true positive biclusters from synthetic data. For this end, we relied on available generation procedures for biclustering data (Henriques and Madeira 2015) to generate 30 synthetic datasets ($N = 1000$, $M = 100$) with 20 planted biclusters (10 significant biclusters and 10 non-significant biclusters). In particular, three of the planted significant biclusters are small yet their pattern φ_B is 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, we assessed the ability of exhaustive biclustering algorithms in BicPAMS (parameterized with default behavior (Henriques et al. 2017)) to recover the true positive biclusters only when the proposed statistical tests are used to post-filter the outputs. The results confirm the relevance of using the proposed statistical tests to recover the true biclusters. Alternative global tests fail to correctly assess small (yet improbable) and large (yet probable) biclusters.

Figure 8 assesses the impact of assuming different distributions and dependence degree when modeling the regularities of gene expression data (*ycycle*). Considering a coherency strength given by $| \mathcal{L} | = 5$, this analysis compares the significance of the largest perfect biclusters (exhaustively mined with BicPAMS (Henriques et al. 2017)) using stochastic views, frequentist views (with pairwise and overall dependence between items), and the proposed assessment method based on dynamically selected tests as a function of the observed data size (N and M) and regularities (Θ). Understandably, the average number of rows for the largest non-constant biclusters is higher than the number of rows for the constant peers. However, this effect is compensated by their higher p_{φ_B}, leading to comparable levels of significance against constant biclusters. We observe that the conditional dependence between φ_B items can largely impact the significance analysis (counts are associated with pessimistic views since missing a single item from a lengthy pattern does not contribute to its support) and, therefore, should be only considered for high-dimensional datasets ($M > 100$).

Table 4 Ability of exhaustive biclustering algorithms to recover only true biclusters from a set of 20 planted biclusters (10 significant and 10 non-significant) in the presence and absence of local and global statistical tests for 30 data instances (1000×100 setting and Jaccard-based match scores MS($\mathcal{B}, \mathcal{H}$) and MS($\mathcal{H}, \mathcal{B}$) in accordance with Henriques and Madeira 2015)

Option	MS($\mathcal{B}, \mathcal{H}$) (coverage of significant biclusters)	MS($\mathcal{H}, \mathcal{B}$) (exclusion of non-significant biclusters)	Fraction of found significant biclusters (%)	Avg. number of found significant biclusters	Avg. number of found non-significant biclusters
No statistical tests	0.97	0.61	52	10.0	9.4
Global statistical tests (Henriques 2016)	0.84	0.81	74	8.1	2.8
Proposed statistical tests	0.97	0.95	100	10.0	0.0

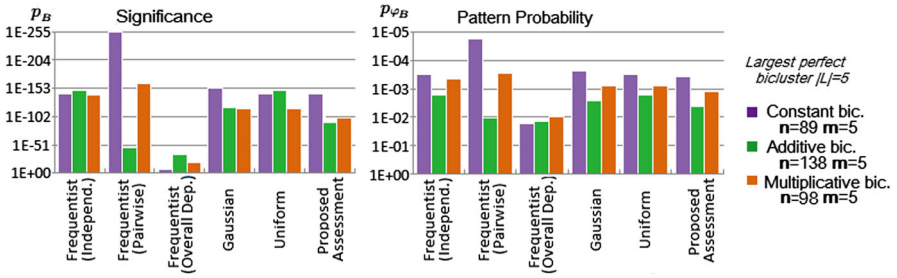

Fig. 8 Impact of stochastic and frequentist views when assessing large biclusters discovered from *ycycle* dataset

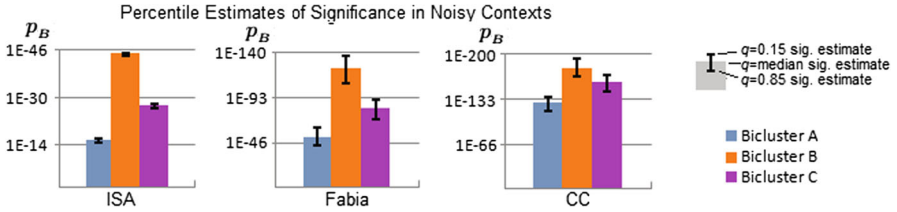

Fig. 9 Percentiles of *significance* of an illustrative set of biclusters selected from the outputs of ISA, Fabia and CC algorithms collected for the *ycycle* dataset

Finally, we applied state-of-the-art biclustering methods over *ycycle* dataset to identify noisy biclusters with varying properties, and then evaluated their significance using the proposed strategies to assess real-valued biclusters. Figure 9 illustrates the median, 15-percentile and 85-percentile estimators of the true significance of three illustrative biclusters (corresponding to the 50-, 90- and 70-percentiles of the discovered biclusters with regards to their statistical significance). gathered from the application of three biclustering methods – CC (Cheng and Church 2000), ISA (Ihmels et al. 2004) and Fabia (Hochreiter et al. 2010) – respectively prepared to discover biclusters with constant, scaling and shifting factors. This analysis reveals the importance of bounding significance whenever possible. Interestingly, although the absolute significance of ISA's biclusters is worse that of their peers, their variance is approximately zero (since ISA guarantees a strong homogeneity that penalizes large biclusters). Contrasting, Fabia and CC provide more significant biclusters at a cost of tolerating high levels of noise.

Statistical versus 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 5 and Fig. 10. Table 5 shows how the average number of enriched gene ontology terms (p value of hypergeometric test after correction[7] below 1E-3 using) varies for significant and non-significant biclusters. For this analysis, we applied BicPAMS with default behavior on *dlbcl*, *hughes* and *gasch* datasets. From the gathered results, although it appears that statistical and biological significance are correlated, this correlation is spurious as it is primarily explained by a third variable: the number of rows within a bicluster. A commonly well-know draw-

[7] Using *Yeastract* http://yeastract.com and *Enrichr* http://amp.pharm.mssm.edu/Enrichr.

Table 5 Characterization of biclustering models (found by BicPAMS with default behavior) in gasch, dlbcl and hughes datasets according to their: statistical significance (fraction of significant biclusters), size, and biological significance (enriched biclusters)

Dataset	Filtering criteria	#Bics	Average $\|I\| \times \|J\|$	Average #terms enriched per bic
Gasch	Significant	94	807×9	74
	Non-significant	13	101×4	51
	All	107	721×8	71
Dlbcl	Significant	32	102×7	39
	Non-significant	9	34×4	23
	All	41	87×6	36
Hughes	Significant	57	1431×8	69
	Non-significant	11	128×4	47
	All	68	1201×7	65

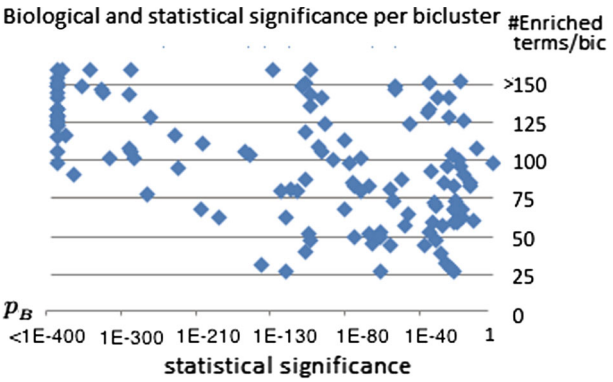

Fig. 10 Correlation between biological and statistical significance: number of enriched terms per bicluster versus statistical significance of the exhaustive biclustering exploration of the *gasch* dataset with BicPAMS (Henriques et al. 2017)

back of term enrichment analysis is its bias towards large (bi)clusters with a large number of rows (columns) independently of φ_B pattern (Huang et al. 2009).

To complement this analysis, Fig. 10 plots all the biclusters discovered by BicPAMS in the *gasch* dataset according to their statistical significance and number of enriched terms (biological significance). This analysis suggests that statistical significance and biological significance are not necessarily correlated. These analyzes pinpoint the relevance of using both views to evaluate and (possibly) guide biclustering algorithms.

Applied correction. The impact of the coherency assumption and strength in the space of similar biclusters is illustrated in Fig. 11. The space is here given by biclusters with the same pattern length ($m = 4$). Understandably, the order-preserving space of similar biclusters is the most flexible and therefore the most compact. The space of similar additive biclusters increases at a significant lower rate than constant and multiplicative spaces due to the higher chance of a pattern φ_B to be described by

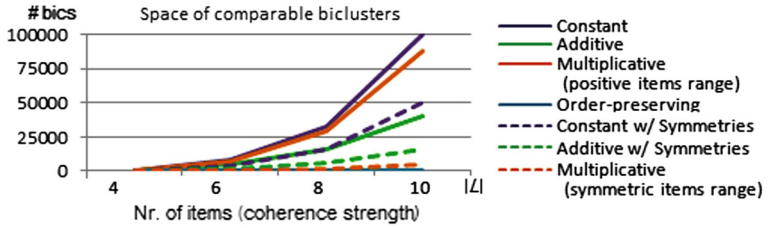

Fig. 11 Impact of coherency assumption/strength on the space of comparable biclusters

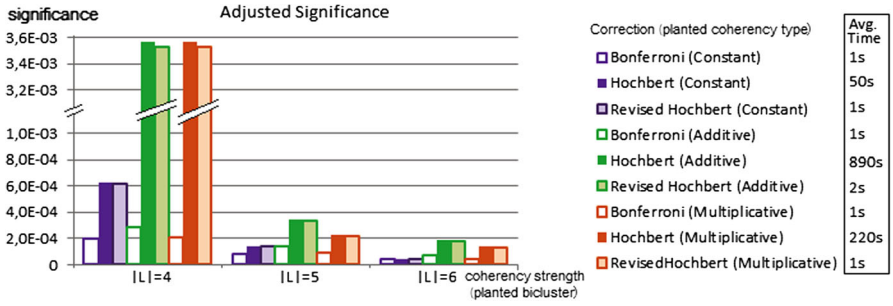

Fig. 12 Minimum probability p_B to guarantee that a bicluster B deviates from expectations (using Bonferroni and Hochbert procedures with $\alpha = 5\%$)

multiple shifting factors. The multiplicative space is comparable with constant space due to the low probability of a pattern φ_B being described by a scaling factor $\gamma \neq 1$. In the presence of symmetries, the size of both constant and multiplicative spaces reduce visibly.

Figure 12 compares the impact of different correction procedures – Bonferroni and the revised Hochbert (parameterized with 20 iterations to compute the adjusted significance). For this analysis we assessed the significance of a planted bicluster ($n = 20 \wedge m = 4$) against a discrete dataset with $N = 5000$ rows and $M = 100$ columns with a varying number of items ($\mid \mathcal{L} \mid \in \{4\}$) distributed according to $N(\frac{|\mathcal{L}|}{2}, \frac{|\mathcal{L}|}{5})$. The large differences observed between the correction procedures support the relevance of using non-conservative corrections that preserve the family-wise error in order to reduce the risk of false negatives. This analysis also underlines the importance of using correction procedures that take into account the different coherency assumptions for an adequate identification of the probabilistic levels that guarantee a significant deviation from the expected probability of occurrence. Additionally, the residual computational complexity of the revised Hochbert procedure together with its comparable levels of significance (against the original Hochbert procedure) support the efficiency and effectiveness of the proposed correction.

Comparing biclustering solutions. To test the significance of biclusters discovered in real settings, we selected five state-of-the-art biclustering algorithms[8]: FABIA with

[8] To run experiments, we used: fabia package from R, BicAT (Barkow et al. 2006) and BicPAMS (Henriques et al. 2017) software.

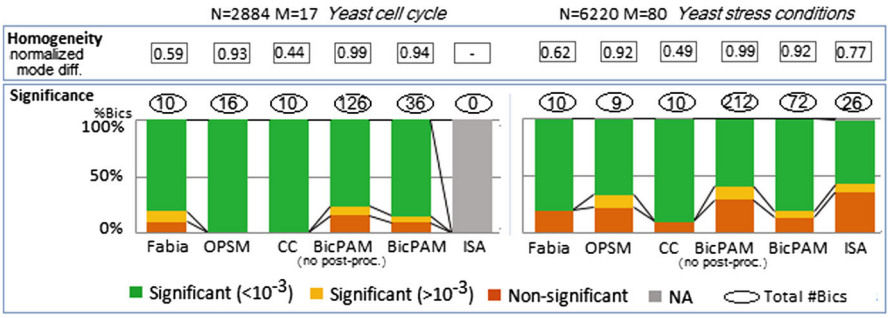

Fig. 13 Comparison of significance and homogeneity of biclusters delivered from Fabia, ISA, OPSM, CC and BicPAM over gene expression data

sparse prior (Hochreiter et al. 2010) (able to discover multiplicative biclusters), ISA (Ihmels et al. 2004) (able to discover additive biclusters), CC (Cheng and Church 2000) (able to discover biclusters with flexible coherence that can accommodate shifts and scales), OPSM (Ben-Dor et al. 2003) (able to discover order-preserving biclusters) and BicPAM (Henriques and Madeira 2014b) (able to discover all the previous coherencies). The number of seeds for FABIA and ISA was set to 10, the number of iterations for OPSM was varied from 10 to 100. BicPAM was parameterized with closed pattern representations, iterative searches with decreasing coherence strength, and merging (70% overlap) and filtering (30% of overlap) procedures. The remaining parameters of the selected algorithms were set by default. Figure 13 provides an initial view on their performance with regards to the significance and homogeneity of their outputs for two datasets (gene expression of Yeast along a cell cycle (*ycycle*) and under stress conditions (*ystress*) 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 φ_B pattern. We can observe that the proposed tests provide a simplistic yet robust tool to study the significance of biclustering algorithms. CC and Fabia discover biclusters with higher significance than ISA and BicPAM (without post-processing options), but accommodate high amounts of noise (looser homogeneity levels according to merit functions sensitive to non-constant coherencies (Yang et al. 2002)). The use of post-processing options in BicPAM is associated with a good balance between the significance and homogeneity of the discovered solutions. Although OPSM implements principles to guarantee the significance of order-preserving solutions, some outputted biclusters are still below the adjusted significance threshold.

6 Conclusions and future work

This work proposes a robust set of statistical principles, implemented within a method termed BSig, to assess the significance of biclustering solutions with varying coherency and quality. We covered the limitations of the existing statistical assessments of biclustering solutions and potential contributions from related research on pattern mining and

statistics' foundations. To answer this task, five major contributions were proposed. First, relying on the established mapping between pattern mining and biclustering, we consistently integrated dispersed statistical principles to assess constant biclusters in discrete settings. Second, a new variant of Huchbert correction was proposed to surpass the efficiency bottlenecks of non-conservation corrections, and minimize the risk of false positive and false negative discoveries. Third, new principles were proposed to extend this assessment towards flexible coherencies: additive, multiplicative, symmetric, plaid and order-preserving assumptions. Fourth, we extended these principles to enable assessments in noisy contexts where the underlying coherency of a bicluster becomes blurry. In this context, in order to correctly assess large biclusters discovered at a cost of tolerating large amounts of noise, we proposed new statistical tests consistently integrating significance and homogeneity views. Finally, new principles were proposed to guarantee the applicability of previous tests towards real-valued biclusters, including multiple estimators to bound the significance of biclusters and discretization methods without susceptibility to boundary problems. In this context, we also rely on integral calculus to enable the assessment of biclusters with continuous ranges of shifting and scaling factors.

Results from synthetic and real data show the soundness of the proposed method to flag false positive biclusters with varying coherency. This evidence is essential to validate the increasing number of implications derived from the analysis of local relations within biomedical and social data. We also conducted an experimental analysis to show 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. We further confronted statistical and domain-driven significance, showing that they are not always in agreement. Finally, we conducted an initial comparison of biclustering algorithms that stress both the relevance of revealing their statistical significance and of combining this view with homogeneity views.

This work opens new directions for future work. The proposed principles can be used to define heuristics to guide learning tasks. Both efficient local tests and the inference of global constraints are critical to narrow the search space. Another relevant direction is to extend the proposed statistical tests to minimize false negatives, thus balancing type-I and type-II errors. This is particularly relevant for the unsupervised analysis of high-dimensional data in order to reduce the overfitting risk of the learned biclustering models by minimizing false positives (exclusion of non-significant biclusters), as well as the underfitting risk by minimizing false negatives (recovery of significant biclusters when there is evidence that the outputted set is incomplete).

Acknowledgements This work was supported by *FCT* under the Neuroclinomics2 Project PTDC/EEI-SII/1937/2014, Research Grant SFRH/BD/75924/2011 to RH, INESC-ID plurianual Ref. UID/CEC/50021/2013, and LASIGE Research Unit Ref. UID/CEC/00408/2013.

Compliance with Ethical Standards

Conflicts of interest The authors declare that they have no conflict of interest.

Appendix 1: Relevance of biclustering with flexible coherency

High-dimensional biomedical and social data is characterized by the presence of biclusters with flexible coherency assumptions (Table 6). Table 6 motivates the relevance of such biclusters, highlighting data contexts where their discovery is relevant for different purposes.

Table 6 Relevance of non-constant biclusters for biomedical and social data analysis

Coherency	Illustrative biclusters across biomedical and social domains
Additive and Multiplicative	Coherencies used to allow the occurrence of shifting and scaling factors across observations. Illustrating, two genes may be regulated in the same subset of conditions (features) but show different expression levels explained by a shifting or scaling factor associated with their distinct responsiveness, or the bias introduced by the applied measurement and preprocessing (Henriques and Madeira 2014b). These factors are also critical to analyze physiological and clinical data to handle the structural differences across individuals (Henriques et al. 2015). In social domains, these factors are relevant to model social interactions with coherent behavior but differing in the extent of frequency and popularity of actions, and to group subjects with identical variation of preferences during browsing and collaborative filtering (Gnatyshak et al. 2012)
Order Preserving	Order-preserving biclusters were originally proposed to find genes co-expressed within a temporal progression (such as stages of a disease or drug response) (Ben-Dor et al. 2003). Yet, they have been also largely applied in static biological contexts where gene expression or molecular concentrations coherently vary across samples (Henriques and Madeira 2014a). This coherence can be also applied to: find sets of nodes in (social and biological) networks with an order-preserving degree of influence across another set of nodes; to support task planning and scheduling; and to discover order-preserving preferences from collaborative filtering data (Henriques 2016). Order-preserving biclusters can emulate constant, additive and multiplicative coherencies, leading to more inclusive solutions with larger and less noise-susceptible regions
Symmetric	In biological contexts, symmetries are key to simultaneously capture activation and repression mechanisms within biological processes associated with biclusters in transcriptomic, proteomic or metabolic data (Henriques and Madeira 2016a). In social contexts, symmetries are used to capture opposed (yet correlated) regularities associated with trading, tweeting, browsing and (e-)commerce activity (Henriques 2016). Symmetries can be combined with the previous coherencies
Plaid	Plaid models are essential to describe overlapping regulatory influence in biological contexts and cumulative effects in the interactions between nodes in social networks (Lazzeroni and Owen 2002; Mankad and Michailidis 2014). Illustrating, consider a gene activate in a set of biological processes, a plaid coherence can consider their cumulative effect on the expression of a gene when more than one of these processes is active at a time. The plaid model can be also applied to study regulatory cascades, user behavior and trading operations, as these data contexts are also characterized by mutual influences between biclusters (Henriques 2016)

Appendix 2: Continuous adjustment factors

The proposed statistical tests can be extended to support biclusters with continuous adjustment factors. 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 $\gamma \in [-1.3, 1.8]$ or scaling $\gamma \in [0, 1.8]$ factors for values $a_{ij} \in [0, 4]$). In order to robustly compute 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 based on the integral of the product of (either *slided* or *scaled*) density probability functions.

Continuous ranges of shifting factors

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

$$\int_{a_{ij}+\gamma_1}^{a_{ij}+\gamma_2} f(x) = \int_{\gamma_1}^{\gamma_2} f(x + a_{ij}) \tag{A1}$$

where $f(x)$ is the distribution function that approximates a_{ij} values. This calculus assumes that the range of observed values $\hat{A}$ 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 Fig. 14, under the assumption that the values in A are either described by an Uniform or Gaussian distribution. Given $\varphi_B^i = \{a_{i1}, .., a_{im}\}$ combination of values, $p_{\varphi_B^i}$ can be approximated by:

$$\int_{\gamma_1}^{\gamma_2} \Pi_{j=1}^{m} f(x + a_{ij}) \tag{A2}$$

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.

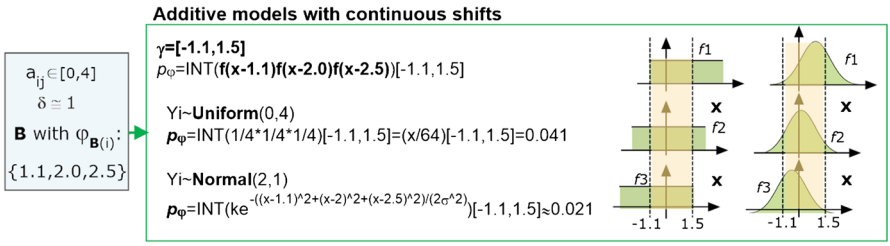

Fig. 14 Illustrative integral of the product of slided density functions to assess biclusters with continuous ranges of shifting factors

Continuous ranges of scaling factors

The probability of occurrence of a combination of real values φ_B^i on the ith row of a bicluster under a multiplicative coherency across rows 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 values of a given row x_i and $\bar{A}$ to be the range real values in A. When only positive values are allowed, the scaling range is $[\gamma_1 = 0, \gamma_2 = \bar{A}/max_i]$. When negatives values are allowed the scaling range is given by $[\gamma_1 = -d, \gamma_2 = d]$ where $d = max(max_i, -min_i)/(\bar{A}/2)$.

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 $\varphi_B^i = \{a_{i1}, .., a_{im}\}$ combination of values for i row, $p_{\varphi_B^i}$ can be approximated by:

$$\int_{c_1}^{c_2} \Pi_{i=1}^{n} f(x/a_i) \tag{A3}$$

Similarly, an efficient computation of the (A3) integral calculus is made available recurring to interpolation whenever the multiplication of the inputted density functions is complex. This strategy is illustrated in Fig. 15, under the assumption that the values of the A matrix are either described by a single Uniform or Gaussian distribution.

Appendix 3: Complementary results

Table 7 shows how the required minimum of rows that guarantee the statistical significance of a real-valued bicluster with continuous shifts/scales varies with the number rows and columns of the input dataset. Two major observations can be retrieved. Both the size and dimensionality of data affect the significance levels, being the effect of

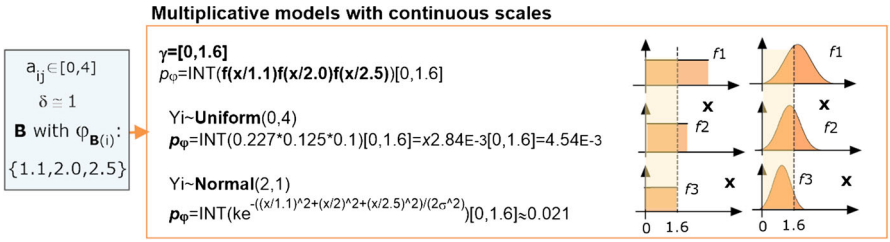

Fig. 15 Illustrative integral of the product of scaled density functions to assess biclusters with continuous ranges of scaling factors

Table 7 Impact of data size and dimensionality on the expected minimum number of observations ($\mathbf{n}_{min}$) in biclusters with continuous adjustment factors to guarantee their statistical significance (assuming a $\delta = 0.2$ coherency strength, uniform background values, and additive and multiplicative coherencies with varying ranges of allowed shifts/scales). Algorithm 1 was applied to compute statistical significance

			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

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 (Table 8). 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.

Figure 16 provides the graphical representation of the results gatheres throughout Tables 1, 2 and 3, thus showing the expected minimum number of rows in a bicluster that guarantees its significance for varying: coherency assumption, pattern expectations φ_B, coherency strength $|\mathcal{L}|$, pattern length m, data size N, and data dimensionality M.

Table 8 Expected probability of different patterns to occur in biclusters with continuous shifts and scales from data with approximately uniform distribution of values ($a_{ij} \in [0,1]$)

m=3	$\gamma \in [0, 1]$ 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}
Multip. 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]
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]
m=5	$\gamma \in [0, 1]$ 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}
Multip. p_{φ_B} δ=0.2 $\delta \in [0.1, 0.3]$	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 $\delta \in [0.1, 0.3]$	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]

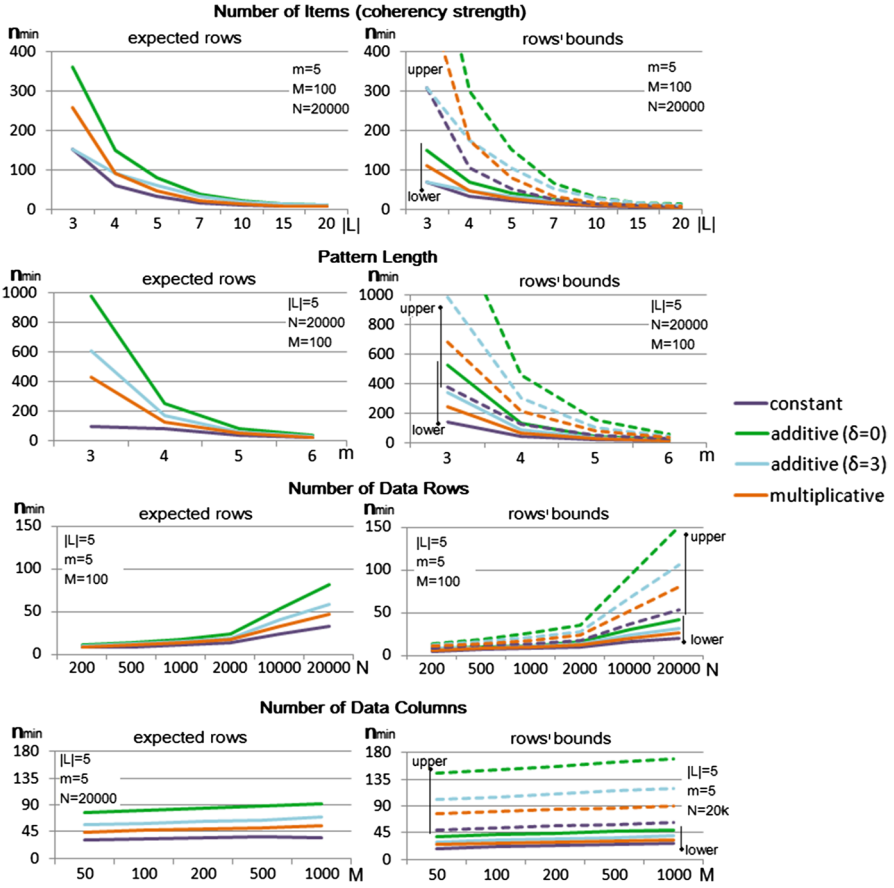

Fig. 16 Impact of coherency strength, pattern length, data size/dimensionality on the expected minimum number of bicluster's rows that guarantee its statistical significance

References

Aggarwal CC, Yu PS (1998) A new framework for itemset generation. In: Proceedings of the seventeenth ACM SIGACT-SIGMOD-SIGART symposium on principles of database systems, ACM, New York, NY, USA, PODS '98, pp 18–24, doi:10.1145/275487.275490

Alzahrani M, Kuwahara H, Wang W, Gao X (2017) Gracob: a novel graph-based constant-column biclustering method for mining growth phenotype data. Bioinformatics. doi:10.1093/bioinformatics/btx199

Balakrishnan S, Kolar M, Rinaldo A, Singh A, Wasserman L (2011) Statistical and computational tradeoffs in biclustering. In: NIPS 2011 workshop on computational trade-offs in statistical learning, vol 4

Barkow S, Bleuler S, Prelić A, Zimmermann P, Zitzler E (2006) Bicat: a biclustering analysis toolbox. Bioinformatics 22(10):1282. doi:10.1093/bioinformatics/btl099

Bay SD, Pazzani MJ (2001) Detecting group differences: mining contrast sets. Data Min Knowl Discov 5(3):213–246. doi:10.1023/A:1011429418057

Bellay J, Atluri G, Sing TL, Toufighi K, Costanzo M, Ribeiro PSM, Pandey G, Baller J, VanderSluis B, Michaut M, Han S, Kim P, Brown GW, Andrews BJ, Boone C, Kumar V, Myers CL (2011) Putting genetic interactions in context through a global modular decomposition. Genome Res 21(8):1375–1387. doi:10.1101/gr.117176.110

Ben-Dor A, Chor B, Karp R, Yakhini Z (2003) Discovering local structure in gene expression data: the order-preserving submatrix problem. J Comput Biol 10(3–4):373–384. doi:10.1089/10665270360688075

Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the royal statistical society Series B (Methodological), pp 289–300, doi:10.2307/2346101

Benjamini Y, Yekutieli D (2001) The control of the false discovery rate in multiple testing under dependency. Ann Stat 1165–1188. doi:10.1214/aos/1013699998

Bolton RJ, Hand DJ, Adams NM (2002) Determining hit rate in pattern search. Springer, Berlin, pp 36–48. doi:10.1007/3-540-45728-3_4

Brown GW (1947) On small-sample estimation. Ann Math Stat 18(4):582–585

Califano A, Stolovitzky G, Tu Y (2000) Analysis of gene expression microarrays for phenotype classification. Int Conf Intell Syst Mol Biol 8:75–85

Carmona-Saez P, Chagoyen M, Rodriguez A, Trelles O, Carazo JM, Pascual-Montano A (2006) Integrated analysis of gene expression by association rules discovery. BMC Bioinform 7(1):54. doi:10.1186/1471-2105-7-54

Chen Y, Xu J (2016) Statistical-computational tradeoffs in planted problems and submatrix localization with a growing number of clusters and submatrices. J Mach Learn Res 17(1):882–938

Cheng Y, Church GM (2000) Biclustering of expression data. Intell Syst Mol BiolPress 8:93–103

DuMouchel W (1999) Bayesian data mining in large frequency tables, with an application to the fda spontaneous reporting system. Am Stat 53(3):177–190. doi:10.2307/2686093

DuMouchel W, Pregibon D (2001) Empirical bayes screening for multi-item associations. In: Proceedings of the seventh ACM SIGKDD international conference on knowledge discovery and data mining, ACM, New York, NY, USA, KDD '01, pp 67–76, doi:10.1145/502512.502526

Eisen MB, Spellman PT, Brown PO, Botstein D (1998) Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci 95(25):14,863–14,868

Gasch AP, Spellman PT, Kao CM, Carmel-Harel O, Eisen MB, Storz G, Botstein D, Brown PO (2000) Genomic expression programs in the response of yeast cells to environmental changes. Mol Biol Cell 11(12):4241–4257. doi:10.1091/mbc.11.12.4241

Gionis A, Mannila H, Mielikäinen T, Tsaparas P (2007) Assessing data mining results via swap randomization. ACM Trans Knowl Discov Data 1(3). doi:10.1145/1297332.1297338

Gnatyshak D, Ignatov D, Semenov A, Poelmans J (2012) Gaining insight in social networks with biclustering and triclustering. In: Perspectives in business informatics research, LNBIP, vol 128. Springer, Berlin Heidelberg, pp 162–171, doi:10.1007/978-3-642-33281-4_13

Hämälinen W, Nykänen M (2008) Efficient discovery of statistically significant association rules. In: 2008 Eighth IEEE international conference on data mining (ICDM), pp 203–212. doi:10.1109/ICDM.2008.144

Henriques R (2016) Learning from high-dimensional data using local descriptive models. PhD thesis, Instituto Superior Tecnico, Universidade de Lisboa, Lisboa

Henriques R, Madeira S (2014a) Bicspam: flexible biclustering using sequential patterns. BMC Bioinform 15(1):130. doi:10.1186/1471-2105-15-130

Henriques R, Madeira SC (2014b) Bicpam: pattern-based biclustering for biomedical data analysis. Algorithms Mol Biol 9(1):27. doi:10.1186/s13015-014-0027-z

Henriques R, Madeira SC (2015) Biclustering with flexible plaid models to unravel interactions between biological processes. IEEE/ACM Trans Comput Biol Bioinform (TCBB) 12(4):738–752. doi:10.1109/TCBB.2014.2388206

Henriques R, Madeira SC (2016a) Bic2pam: constraint-guided biclustering for biological data analysis with domain knowledge. Algorithms Mol Biol 11(1):23. doi:10.1186/s13015-016-0085-5

Henriques R, Madeira SC (2016b) Bicnet: flexible module discovery in large-scale biological networks using biclustering. Algorithms Mol Biol 11(1):1–30. doi:10.1186/s13015-016-0074-8

Henriques R, Antunes C, Madeira SC (2015) A structured view on pattern mining-based biclustering. Pattern Recognit 48(12):3941–3958. doi:10.1016/j.patcog.2015.06.018

Henriques R, Ferreira FL, Madeira SC (2017) Bicpams: software for biological data analysis with pattern-based biclustering. BMC Bioinform 18(1):82. doi:10.1186/s12859-017-1493-3

Hochreiter S, Bodenhofer U, Heusel M et al (2010) Fabia: factor analysis for bicluster acquisition. Bioinformatics 26(12):1520–1527. doi:10.1093/bioinformatics/btq227

Holm S (1979) A simple sequentially rejective multiple test procedure. Scand J Stat 6:65–70

Huang DW, Sherman BT, Lempicki RA (2009) Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Res 37(1):1. doi:10.1093/nar/gkn923

Ihmels J, Bergmann S, Barkai N (2004) Defining transcription modules using large-scale gene expression data. Bioinformatics 20(13):1993. doi:10.1093/bioinformatics/bth166

Jaroszewicz S, Scheffer T (2005) Fast discovery of unexpected patterns in data, relative to a bayesian network. In: Proceedings of the eleventh ACM SIGKDD international conference on knowledge discovery in data mining, ACM, New York, NY, USA, KDD '05, pp 118–127. doi:10.1145/1081870.1081887

Karian Z, Dudewicz E (2010) Handbook of fitting statistical distributions with R. Taylor & Francis, Milton Park

Kirsch A, Mitzenmacher M, Pietracaprina A, Pucci G, Upfal E, Vandin F (2012) An efficient rigorous approach for identifying statistically significant frequent itemsets. J ACM 59(3):12:1–12:22. doi:10.1145/2220357.2220359

Koyuturk M, Szpankowski W, Grama A (2004) Biclustering gene-feature matrices for statistically significant dense patterns. In: Proceedings. 2004 IEEE computational systems bioinformatics conference (CSB), pp 480–484. doi:10.1109/CSB.2004.1332467

Lazzeroni L, Owen A (2002) Plaid models for gene expression data. Statistica Sinica 12(1):61–86. http://www.jstor.org/stable/24307036

Lee JD, Sun Y, Taylor JE (2015) Evaluating the statistical significance of biclusters. In: Advances in neural information processing systems 28 (NIPS), Curran Associates, Inc., pp 1324–1332

Lee W, Tillo D, Bray N, Morse RH, Davis RW, Hughes TR, Nislow C (2007) A high-resolution atlas of nucleosome occupancy in yeast. Nat Genet 39(10):1235–1244. doi:10.1038/ng2117

Madeira SC, Oliveira AL (2004) Biclustering algorithms for biological data analysis: a survey. IEEE/ACM Trans Comput Biol Bioinform (TCBB) 1(1):24–45. doi:10.1109/TCBB.2004.2

Madeira SC, Oliveira AL (2007) An efficient biclustering algorithm for finding genes with similar patterns in time-series expression data. In: Asia Pacific bioinformatics conference, pp 67–80

Madeira SC, Teixeira MC, Sa-Correia I, Oliveira AL (2010) Identification of regulatory modules in time series gene expression data using a linear time biclustering algorithm. IEEE/ACM Trans Comput Biol Bioinform (TCBB) 7(1):153–165. doi:10.1109/TCBB.2008.34

Mahfouz MA, Ismail MA (2009) Bidens: Iterative density based biclustering algorithm with application to gene expression analysis. Int J Comput Electr Automa Control Inf Eng 3(1):40–46

Mankad S, Michailidis G (2014) Biclustering three-dimensional data arrays with plaid models. J Comput Graph Stat 23(4):943–965. doi:10.1080/10618600.2013.851608

Megiddo N, Srikant R (1998) Discovering predictive association rules. In: Proceedings of the fourth international conference on knowledge discovery and data mining, AAAI Press, KDD'98, pp 274–278

Mitra S, Banka H (2006) Multi-objective evolutionary biclustering of gene expression data. Pattern Recognit 39(12):2464–2477. doi:10.1016/j.patcog.2006.03.003

Noureen N, Kulsoom N, de la Fuente A, Fazal S, Malik SI (2009) Functional and promoter enrichment based analysis of biclustering algorithms using gene expression data of yeast. In: 2009 IEEE 13th international multitopic conference (INMIC), IEEE, pp 1–6, doi:10.1109/INMIC.2009.5383144

Ojala M, Vuokko N, Kallio A, Haiminen N, Mannila H (2008) Randomization of real-valued matrices for assessing the significance of data mining results. In: Proceedings of the 2008 SIAM international conference on data mining (SDM), SIAM, vol 8, pp 494–505. doi:10.1137/1.9781611972788.45

Okada Y, Fujibuchi W, Horton P (2007) A biclustering method for gene expression module discovery using closed itemset enumeration algorithm. IPSJ Trans Bioinform 3(SIG5):183–192. doi:10.2197/ipsjdc.3.183

Pio G, Ceci M, D'Elia D, Loglisci C, Malerba D (2012) A novel biclustering algorithm for the discovery of meaningful biological correlations between mirnas and mrnas. EMBnetjournal 18(A). doi:10.14806/ej.18.A.375

Ramon J, Miettinen P, Vreeken J (2013) Detecting bicliques in gf[q]. In: Proceedings of the European conference on machine learning and knowledge discovery in databases, vol 8188, Springer New York, Inc., New York, NY, USA, ECML PKDD, pp 509–524. doi:10.1007/978-3-642-40988-2_33

Rosenwald A, Wright G, Chan WC, Connors JM, Campo E, Fisher RI, Gascoyne RD, Muller-Hermelink HK, Smeland EB, Giltnane JM, Hurt EM, Zhao H, Averett L, Yang L, Wilson WH, Jaffe ES, Simon R, Klausner RD, Powell J, Duffey PL, Longo DL, Greiner TC, Weisenburger DD, Sanger WG, Dave BJ, Lynch JC, Vose J, Armitage JO, Montserrat E, López-Guillermo A, Grogan TM, Miller TP, LeBlanc M, Ott G, Kvaloy S, Delabie J, Holte H, Krajci P, Stokke T, Staudt LM (2002) The use of molecular profiling to predict survival after chemotherapy for diffuse large-b-cell lymphoma. N Engl J Med 346(25):1937–1947. doi:10.1056/NEJMoa012914

Scheffer T (2005) Finding association rules that trade support optimally against confidence. Intell Data Anal 9(4):381–395. doi:10.1007/3-540-44794-6_35

Serin A, Vingron M (2011) Debi: Discovering differentially expressed biclusters using a frequent itemset approach. Algorithms Mol Biol 6:1–12. doi:10.1186/1748-7188-6-18

Silberschatz A, Tuzhilin A (1996) What makes patterns interesting in knowledge discovery systems. IEEE Trans Knowl Data Eng 8(6):970–974. doi:10.1109/69.553165

Silverstein C, Brin S, Motwani R (1998) Beyond market baskets: generalizing association rules to dependence rules. Data Min Knowl Discov 2(1):39–68. doi:10.1023/A:1009713703947

Tanay A, Sharan R, Shamir R (2002) Discovering statistically significant biclusters in gene expression data. Bioinformatics 18(suppl1):S136. doi:10.1093/bioinformatics/18.suppl_1.S136

Tavazoie S, Hughes J, Campbell M, Cho R, Church G (1999) Systematic determination of genetic network architecture. Nature Genet 22(3):281–285. doi:10.1038/10343

Wang H, Wang W, Yang J, Yu PS (2002) Clustering by pattern similarity in large data sets. In: Proceedings of the 2002 ACM SIGMOD international conference on management of data, ACM, New York, NY, USA, SIGMOD '02, pp 394–405. doi:10.1145/564691.564737

Webb GI (2007) Discovering significant patterns. Mach Learn 68(1):1–33. doi:10.1007/s10994-007-5006-x

Yang J, Wang W, Wang H, Yu P (2002) delta-clusters: capturing subspace correlation in a large data set. In: Proceedings 18th international conference on data engineering (ICDE), IEEE, pp 517–528. doi:10.1109/ICDE.2002.994771

Zhang H, Padmanabhan B, Tuzhilin A (2004) On the discovery of significant statistical quantitative rules. In: Proceedings of the tenth ACM SIGKDD international conference on knowledge discovery and data mining, ACM, New York, NY, USA, KDD '04, pp 374–383. doi:10.1145/1014052.1014094