The penalized biclustering model and related algorithms

Thierry Chekouo and Alejandro Murua*

Département de Mathématiques et de Statistique, Université de Montréal, CP 6128, Succ. Centre-ville, Montréal, Québec, Canada H3C 3J7

(Received 18 May 2012; accepted 14 December 2014)

Biclustering is the simultaneous clustering of two related dimensions, for example, of individuals and features, or genes and experimental conditions. Very few statistical models for biclustering have been proposed in the literature. Instead, most of the research has focused on algorithms to find biclusters. The models underlying them have not received much attention. Hence, very little is known about the adequacy and limitations of the models and the efficiency of the algorithms. In this work, we shed light on associated statistical models behind the algorithms. This allows us to generalize most of the known popular biclustering techniques, and to justify, and many times improve on, the algorithms used to find the biclusters. It turns out that most of the known techniques have a hidden Bayesian flavor. Therefore, we adopt a Bayesian framework to model biclustering. We propose a measure of biclustering complexity (number of biclusters and overlapping) through a penalized plaid model, and present a suitable version of the deviance information criterion to choose the number of biclusters, a problem that has not been adequately addressed yet. Our ideas are motivated by the analysis of gene expression data.

Keywords: clustering; deviance information criterion; gene expression; mixture; model selection; plaid model

JEL Classification: Y800

1. Introduction

The term biclustering seems to have been first introduced by Cheng and Church [9] (although, Hartigan [16] seems to have been the first to have developed similar techniques). It refers to the simultaneous clustering of the individuals of a population and the features defining the individuals. In general, this is better understood if one thinks of a data matrix where the individuals correspond to the rows and the features to the columns. Biclustering is clustering done simultaneously in two dimensions (rows and columns). A bicluster is a sub-matrix of the data matrix where the rows exhibit a similar pattern across the columns, and the columns exhibit a
similar pattern across the rows. Biclustering techniques have important applications in bioinformatics [32], marketing [12], and text mining [6], to mention a few fields. For example, in bioinformatics, they are usually applied to gene expression data. These are matrices of gene expression levels (rows) obtained under different experimental conditions (columns). Clustering methods (as opposed to biclustering ones) such as hierarchical clustering [29] or k-means [34] will group only the genes (or only the conditions) into subsets that convey biological significance. Consider the gene expression data associated with the yeast cell cycle data [13]. These data set was obtained by combining 10 separate time-courses from different biological and cellular processes. These are the so-called diauxic shift process; the mitotic cell division cycle; the sporulation process; a shock by high temperature, reducing agents and low temperature; and the centrifugal elutriation. The data show the fluctuation of the log-expression levels of 2467 genes over 79 time-points. These data have been analyzed by several researchers [10,13,19,22]. The goal is to find genes that present similar patterns within and across processes. The biological and or cellular functions associated with several genes are unknown. Having these genes share the same cluster with known genes gives analysts an indication about their functions. We note here that contrary to a bicluster, a cluster is a subset of genes of the original data matrix. In other words, a cluster is a sub-matrix with the same number of columns as the original data matrix. Since a gene may belong to only one cluster, all genes in the same cluster must present similar co-regulation patterns across all experimental conditions (i.e. time-courses). This is unrealistic, since a gene may participate in multiple biological pathways (functions) that could or not be co-active under all experimental conditions. This is the reason why biclustering is so relevant in the context of gene expression data. Genes in a bicluster could belong to other biclusters (e.g. signaling diverse biological pathways), and could be co-regulated (i.e. present a similar pattern) only in a subset of experimental conditions. For the particular yeast data, one would certainly expect to observe co-regulation within each process. Although much more interesting, it would be rare to find co-regulated genes across many processes. Forcing co-regulation across all processes, as in clustering, will hinder a possible rich data structure. Biclustering does not force this co-regulation, but it does not hinder it either. Figure 1 shows the log-expressed data together with the permutation of the rows (genes) and columns (experimental conditions) associated with some of the biclusters found by applying the biclustering method introduced later in this work. Note that as expected these biclusters involve few time-courses

Figure 1. Yeast cycle data. Biclusters 3 and 9 found by our penalized plaid model: within the context of the whole data matrix (left) and zoomed in (right). For details see Section 6.
and several genes. They also overlap, which means that these overlapping genes participate in diverse biological pathways. These are two properties of biclustering that clustering does not possess.

Biclustering has not received much attention from the statistical community. Many of the algorithms proposed in the literature have no probabilistic or statistical foundation (for a good survey on the topic see for example [27]). Hence, little is known about the adequacy and limitations of the models and the efficiency of the algorithms. Very few models for biclustering have been proposed. One of the most popular is the plaid model of Lazzeroni and Owen [22]. In this model, the expectation of each entry of the data matrix is written as a sum of overlapping biclusters (see Equation (1)). The parameters are estimated by least squares. In order to facilitate their estimation, the labels are relaxed and assumed continuous. Based on the same model, Turner et al. [33] proposed a constrained estimation of the labels by binary least squares. Gu and Liu [15] generalized the plaid model by introducing bicluster dependent variances within a Bayesian framework. However, their model constraints the overlapping structure of the biclusters to one dimension. That is, the overlapping consists of either only rows or only columns. Caldas and Kaski [7] also extended the plaid model within a Bayesian framework. A drawback of the estimation algorithms presented in [7,15] is that their Gibbs samplers require either the inversion of potentially high-dimensional matrices (depending on the dimension of the data) and/or the computation of products of potentially high-dimensional matrices. These require huge computational costs and may be intractable in practice when faced with high-dimensional data sets. Zhang [35] tried to overcome these limitations also within a Bayesian framework, by estimating the number of biclusters and the bicluster parameters sequentially, that is, one bicluster at the time, using an ICM-type (iterated conditional modes) algorithm [4,23].

In this work, we shed light on the actual statistical models behind the algorithms. This allows us to generalize most of the known popular biclustering techniques, and to justify, and many times improve on, the algorithms used to find the biclusters. It turns out that most of the known techniques have a hidden Bayesian flavor. Therefore, we proposed a Bayesian framework to model biclusters. We show that algorithms such as the one of Lazzeroni and Owen [22] and Cheng and Church [9] can be justified as applications of ICM [4,23], or can be embedded into a Metropolis–Hastings paradigm. We introduce a modified extended version of the Bayesian plaid model, the penalized plaid model, that aims at controlling the amount of bicluster overlapping in the fitting. Our model fully accounts for a general overlapping structure as opposed to just one-dimensional overlapping as in the model of Gu and Liu [15]. The parameters are determined all at once by a dedicated Markov chain Monte Carlo (MCMC) sampler as opposed to the sequential algorithm of Zhang [35]. Our model also takes into account the problem of identifiability of the row and column effects. Inspired by the ANOVA model, it assumes that the sum of these effects vanishes within each bicluster. In addition, the penalized plaid model may be seen as a continuous extension of the non-overlapping model of Cheng and Church [9] to the plaid model. We also introduce a criterion to choose the number of biclusters, a problem that has not yet been addressed properly in the literature. Our criterion is a modified version of the deviance information criterion (DIC) [30] based on the marginal or conditional distribution over the labels, and on maximum a posteriori (MAP) estimates of the model parameters.

The paper is organized as follows. Biclustering as a mixture model is treated in Section 2. The expectation maximization (EM), hard-EM, and ICM algorithms are described in Section 3. In Section 4, we describe a Bayesian framework for biclustering and introduce our penalized plaid model. Our experiments through simulations are shown in Section 5. An application of our penalized plaid model methodology to elucidate the biclustering structure of the gene expression data associated with the yeast data [13] is described in Section 6. We end up our exposition with some conclusions in Section 7.
2. Biclusters are mixtures

Let $Y = (y_{ij})$ be the data matrix, with $p$ cases (the rows) and $q$ conditions (the columns). In practice, in bioinformatics applications, we think of the cases as genes, and of the conditions as different experimental conditions imposed on the genes. A bicluster is given by the equation $E(y_{ij}|\text{cell } (i,j) \in k) = \mu_k + \alpha_{ik} + \beta_{jk}$, where $k$ denotes the bicluster, $E(\cdot| \in k)$ is the conditional expectation given that the cell is in bicluster $k$, $\mu_k$ is the overall mean of the objects in the bicluster, $\alpha_{ik}$ is the effect of the $i$th case in bicluster $k$, and $\beta_{jk}$ denotes the effect of the $j$th condition on bicluster $k$. For identifiability purposes, we need to impose the constraints $\sum_{i \in k} \alpha_{ik} = \sum_{i \in k} \beta_{jk} = 0$, where $i \in k$ refers to all the cases in bicluster $k$ and $j \in k$ refers to all the conditions in bicluster $k$.

We note that the most known biclustering techniques only give explicit assumptions on the mean of the observations. The implicit assumptions of the distribution of the observations need to be derived from the algorithms used to find the biclusters. However, it is easy to see that in all of them the errors may be assumed independent identically distributed Gaussian random variables. The variance is either common to all biclusters or bicluster dependent. Also note that the observed mean is modeled conditionally on knowing the bicluster membership.

Let $\rho_{ik} = 1$ if the $i$th case is in bicluster $k$, and let it be zero otherwise. Similarly, let $\kappa_{jk} = 1$ if the $j$th condition is in bicluster $k$, and let it be zero otherwise. The bicluster problem consists of estimating the number of biclusters $K$ and the labels $(\rho, \kappa) = \{ (\rho_{ik}, \kappa_{jk}) \}$ for $i = 1, \ldots, p$, $j = 1, \ldots, q$, and $k = 1, \ldots, K$. We will also use the notation $\rho_i = (\rho_{ik})$ and $\kappa_j = (\kappa_{jk})$, and $\rho_i^* \kappa_j = (\rho_{ik}\kappa_{jk} : k = 1, 2, \ldots, K)$. We will refer to all the labels $\rho_i^* \kappa_j$ as the membership labels. Most of the techniques suggested in the literature assume that $K$ is known. This is usually a sufficiently large number $[9,22,33,35]$. However, many algorithms are sequential in the sense that they uncover one bicluster at a time. In this case, $K$ is determined sequentially according to some stopping criterion. The sequential search for $K$ is somewhat preferred since it apparently helps to discover large biclusters.

Using the labels, the bicluster model becomes

$$E(y_{ij}|\rho, \kappa) = \sum_k \rho_{ik}\kappa_{jk}(\mu_k + \alpha_{ik} + \beta_{jk}).$$

(1)

It is useful to also consider the special constant mean bicluster $E(y_{ij}|\rho_0\kappa_0 = 1) = \mu_0$. This is what we called the 0-bicluster component. This is used to model data cells that do not belong to any other bicluster. The 0-bicluster is a special cluster in the model in that we always assume that the event $\{\rho_0\kappa_0 = 1\}$ occurs if and only if the event $\{\sum_{k=1}^K \rho_{ik}\kappa_{jk} = 0\}$ occurs. This event is better expressed mathematically by the indicator function $\gamma_{ij} = \prod_{k=1}^K (1 - \rho_{ik}\kappa_{jk})$.

We assume that the conditional variance $\text{Var}(y_{ij}|\rho, \kappa) = \sigma^2(\rho_i, \kappa_j) = \sigma^2(\rho_i^* \kappa_j)$ depends only on the membership labels. We also assume that the distribution of $y_{ij}$ given the membership labels is Gaussian. This is an implicit assumption in most of the known biclustering algorithms. These assumptions imply that the marginal distribution of $y_{ij}$ is a Gaussian mixture with $2^K$ components

$$\frac{1}{\sigma_0} \phi \left( \frac{y_{ij} - \mu_0}{\sigma_0} \right) p(\gamma_{ij} = 1) + \sum_{\rho_i^* \kappa_j} \frac{1}{\sigma(\rho_i^* \kappa_j)} \phi \left( \frac{y_{ij} - \mu(\rho_i^* \kappa_j)}{\sigma(\rho_i^* \kappa_j)} \right) p(\rho_i^* \kappa_j),$$

(2)

where $\mu(\rho_i^* \kappa_j) = \sum_{k=1}^K \rho_{ik}\kappa_{jk}(\mu_k + \alpha_{ik} + \beta_{jk})$, $\sigma_0^2$ is the variance associated with the 0-bicluster, and $p(\rho_i^* \kappa_j)$ is the prior probability associated with the mixture component generated by the set of membership labels $\rho_i^* \kappa_j$. Note that each possible combination of biclusters forms a mixture component. We will refer to the components of the mixture involving only one

T. Chekouo and A. Murua

1258
bicluster as biclusters; that is, a bicluster is a component with $\sum_k \rho_{ik} \kappa_{jk} = 1$. The other components will be referred to as plaids or combination biclusters; that is, a plaid is a component with $\sum_k \rho_{ik} \kappa_{jk} > 1$. In practice, most of the combination biclusters are empty. Hence, the actual number of components is much smaller than $2^K$ when $K$ is large. Intuitively, the component variances should be kept constant (as opposed to increase) independently of how many biclusters form the plaid. This is what the plaid model assumes [22], that is, $\sigma(\rho_i \ast \kappa_j) = \sigma$ for all $(\rho_i, \kappa_j)$. Although the plaid model is a regression model, the marginal of $y_{ij}$ is a Gaussian mixture with $2^K$ components. Also note that the original plaid model does not explicitly include the 0-bicluster component. However, it is implicitly given by a normal distribution with mean $\mu_0$ and variance $\sigma^2$ equal to the regression variance. That is, the mixture model given by Equation (2) is still valid with $\sigma_0 = \sigma(\rho_i \ast \kappa_j) = \sigma$, and $\mu_k = \mu_0 + \mu_k'$, where $\{\mu_k'\}$ are the deviations from the overall mean $\mu_0$. If we suppose that every data cell $y_{ij}$ in the data matrix may belong to only one bicluster, then the labels must satisfy the constraints $\sum_{k=0}^{K} \rho_{ik} \kappa_{jk} = 1$, for all $(i, j)$. It is easily seen that this case corresponds to a $(K + 1)$-component Gaussian mixture for the marginal of $y_{ij}$

$$p(y_{ij}) = \sum_{k=0}^{K} \frac{1}{\sigma_k} \phi \left( \frac{y_{ij} - \mu_k - \alpha_{ik} - \beta_{jk}}{\sigma_k} \right) p(\rho_{ik} \kappa_{jk} = 1),$$

with $\alpha_{i0} = \beta_{j0} = 0$ for all $(i, j)$. This corresponds to the model introduced by Cheng and Church [9]. In this model, there is no overlapping between biclusters, that is, each cell belongs to only one bicluster, though a line or a column may belong to more than one bicluster. (We note that the notion of overlapping is not always the same in the literature. Some authors say that there is overlapping if a line or column belongs to more than one bicluster [15,22,33].)

In what follows, we will work with the general mixture model given by Equation (2). This model is a special case of the so-called mixture of experts model in the machine learning literature [18]. However, note that in the usual mixture of experts model $p(\rho_i \ast \kappa_j)$ is replaced by $\pi_{ijk}$. In general, the biclustering parametrization of the labels will yield a more parsimonious model, unless these probabilities are modeled as a function of the observed variables (e.g. through logistic regression). This is difficult to do in the biclustering framework since usually there is no further information on the data cell, that is, there is no covariable information. To our knowledge, this type of model has not been proposed in the biclustering literature or, at least, it has not been widely applied. Instead, the Lazzeroni and Owen plaid model has become more popular. In this model, a data cell can be seen as having contributions from several biclusters. Many researchers think of this property as if the data cell were a member of several biclusters. However, as we can see from the mixture formulation of the model, this is not true. Membership in several biclusters is better thought of as a data cell that has high probabilities of belonging to several biclusters. Instead, a data cell with contributions from several biclusters in the plaid model corresponds to a cell lying in a combination bicluster of the mixture.

3. Parameter estimation

Since biclustering corresponds to finite mixture modeling, a straightforward application of the EM algorithm [11] appears as a good procedure to find estimates of the parameters and bicluster labels. This is the case for at least the Cheng and Church model where only $K + 1$ sets of parameters need to be estimated. The general bicluster mixture model involves a number of parameters that increase exponentially with the number of biclusters $K$. It is perhaps for this reason that the literature does not favor EM as a valuable alternative. However, we will see here that instead hard-EM has been preferred. We remark that there is no mention of hard-EM in the biclustering literature. The community appears not to have realized yet that this is the algorithm most researchers are using to fit the models. Let $\theta$ denote the set of all parameters in the model. We will
write $\mu(\rho_i \ast \kappa_j)$ as $\mu(\rho_i \ast \kappa_j, \theta)$ to make explicit the dependency of this quantity on $\theta$. Also, let $p(\rho, \kappa | Y, \theta)$ be the conditional probability of the labels given the data and the parameters. In the EM algorithm, the expectation step corresponds to the computation of

$$Q(\theta | \hat{\theta}) = -\frac{1}{2} \sum_{(\rho, \kappa), (i, j)} p(\rho, \kappa | Y, \hat{\theta}) \left\{ \log \sigma^2(\rho_i \ast \kappa_j) + \frac{(y_{ij} - \mu(\rho_i \ast \kappa_j, \theta))^2}{\sigma^2(\rho_i \ast \kappa_j)} - 2 \log p(\rho_i \ast \kappa_j) \right\},$$

where $\hat{\theta}$ is the current estimate of the parameters. The goal of computing this quantity is the M-step, that is, the maximization of $Q(\theta | \hat{\theta})$. Note, however, that this quantity is intractable for large $K$. A possible solution is to approximate the sum over all possible values of the labels $(\rho, \kappa)$ by the largest term in the sum. But this term will depend on the value of $\theta$. Another popular solution, the so-called hard-EM algorithm, is to replace the sum by the term associated with the largest weight $p(\hat{\rho}, \hat{\kappa} | Y, \hat{\theta})$. Once the labels have been fixed, the M-step for the plaid model (recall that in this model $\sigma(\rho_i \ast \kappa_j) = \sigma$) corresponds to maximize over $\theta$

$$\hat{Q}(\theta | \hat{\theta}) = -\frac{n}{2} \log \sigma^2 - \frac{1}{2\sigma^2} \sum_{i,j} (y_{ij} - \mu(\hat{\rho}_i \ast \hat{\kappa}_j, \theta))^2 + \sum_{i,j} \log p(\hat{\rho}_i \ast \hat{\kappa}_j).$$

Note that the hard-EM algorithm in this setup coincides with a simple version of the ICM algorithm [4,23]. That is, for the current value of the parameters $\hat{\theta}$, the labels are estimated by the mode of the full conditional distribution of the labels $p(\rho, \kappa | Y, \hat{\theta})$. And once the labels have been chosen, the parameters are updated by the mode of the ‘full conditional’ of $\theta$, which is proportional to the likelihood. We remark that a full version of the ICM algorithm would maximize over $\theta$ by performing a series of consecutive maximizations of the full conditionals of the lower dimensional parameters that make up $\theta$. Therefore, a truly hard-EM iteration would be achieved by performing an ICM within ICM in order to maximize over $\theta$.

Since hard-EM is a form of ICM with improper priors (i.e. uniform priors on unbounded domains of the parameters such as the means and the variances), one should wonder if the use of proper priors would improve the results. Also, the use of priors allows the use of the Gibbs sampler and more generally, Metropolis–Hastings techniques, to obtain not only point estimates of the parameters but also proper posterior distributions of the parameters. These are discussed in Section 4.

### 3.1 The EM estimates of the labels

Most observations will fall in a single bicluster (as opposed to a combination bicluster). Hence, in general, the number of non-void combination biclusters (components) is much smaller than $2^K$. The expectation step can only be carried over if $K$ is not large. In some problems, $K$ may be moderate, so that a solution through EM may be possible. Also, the EM algorithm appears to be a sensible procedure for the important case where each observation is supposed to be in a single bicluster. In this case, the bicluster mixture consists of only $K + 1$ components (including the 0-bicluster). The EM updating equations for the general case are shown in Section B of the Supplemental Online Material. The derived formulas are easily adapted to the $K$-component mixture case. Unfortunately, despite the simplicity of the updating formulas, it does not seem possible to get reliable estimates of $\rho_{ik}$ and $\kappa_{jk}$ from the EM algorithms for biclusters described here. In fact, note that for any $j = 1, \ldots, q$

$$P(\rho_{ik} = 1 | Y) = P(\rho_{ik} = 1, \kappa_{jk} = 1 | Y) + P(\rho_{ik} = 1, \kappa_{jk} = 0 | Y) \geq P(\rho_{ik} \kappa_{jk} = 1 | Y).$$ (3)

Thus, if there is a single column $j$ such that $p(\rho_{ik} \kappa_{jk} = 1 | Y)$ is large, then the probability that the row $i$ belongs to bicluster $k$ will be high as well. To illustrate this point, we used an artificial
data set with two non-overlapping biclusters. Figure 2 shows the two biclusters (leftmost panel) and the estimated probabilities of bicluster membership. The middle panel image shows the EM-estimated posterior membership probabilities $P(\rho_1 = 1, \kappa_j = 1 | Y)$, whilst the right panel image shows the ones for $P(\rho_2 = 1, \kappa_j = 1 | Y)$. Observe, for example, that in the middle panel image (estimated Bicluster 1) there are many spots of large (black) probabilities around the black rectangle (Bicluster 1), surrounded by very small (white) probabilities. Therefore, by Equation (3), almost all the rows belong to Bicluster 1. The same can be concluded from the right panel image; that is, almost all the rows belong to Bicluster 2. Unfortunately, this property makes biclustering different from clustering and renders biclustering as a mixture model an impractical solution. Perhaps, it is exactly to avoid this problem that the literature has favored a hard-EM type solution.

4. A Bayesian biclustering framework

In this section, we consider a fully Bayesian model for biclustering. The Bayesian framework will allow us to employ the Gibbs sampler and more complex techniques derived from the Metropolis–Hastings algorithm. We note that ICM may be seen as a deterministic greedy Gibbs, where instead of generating stochastic samples, one simply proposes the mode of the conditional distribution.

Given the bicluster labels $(\rho, \kappa)$, we define $I_k$ to be the set of rows making up bicluster $k$, and $J_k$ to be the set of columns in bicluster $k$, $k = 1, \ldots, K$. Bicluster $k$ is given by $B_k = I_k \times J_k$. The number of elements in bicluster $k$ will be denoted as $n_k$. The number of rows and columns in this bicluster will be denoted by $r_k$ and $c_k$, respectively. Note that $n_k = r_k \times c_k$.

We suppose that given the bicluster labels, the prior of the row effects $\{\alpha_{ik}\}$ (note: throughout the manuscript the notation $\{a_i\}$ will denote the set with elements $a_i$ for all possible admissible values of the index $i$, unless otherwise stated) is a multivariate normal distribution with mean zero and variance–covariance matrix $\text{Cov}(\alpha_{ik}, \alpha_{i'k}) = (1 - 1/r_k) \sigma^2_\alpha$, if $i' = i \in I_k$, and is equal to $-\sigma^2_\alpha / r_k$, otherwise. We also suppose that the prior for $\{\beta_{jk}\}$ given the bicluster labels is a similarly defined zero-mean multivariate normal distribution. These prior distributions ensure that the row and column effects add up to zero on each bicluster [20]. The distributions are degenerate since the variance–covariance matrices are singular. This fact is used in the derivation of the full conditional distribution of these parameters (see the Supplemental Online Material).

The priors for the means $\mu_0$ and $\{\mu_k\}$ are given by independent zero-mean normal distributions with variances $\sigma^2_{\mu_0}$ and $\sigma^2_{\mu_k}$, respectively. The prior of the bicluster variances $\sigma^2_k$ is independent inverse-$\chi^2(s^2, v)$ (i.e. a scaled-inverse-chi-squared distribution) with scaling parameter $s^2$ and $v$ degrees of freedom. The prior for the variance $\sigma^2_0$ associated with the zero-bicluster is also an inverse-$\chi^2(s^2_0, v_0)$.
The penalized plaid model. In general, the biclustering may contain overlapping biclusters (producing plaid or combination biclusters). The amount of overlapping may be controlled by imposing a prior that restricts it. We assume that the prior for the bicluster labels is proportional to \(\prod_{ij} \exp\{-\lambda [1 - y_{ij} - \sum_{k=1}^K \rho_{ik} k_{jk}]\}\). The parameter \(\lambda \geq 0\) controls the amount of biclustering overlapping. The larger the \(\lambda\), the less the overlapping. We will refer to the plaid model with this prior as the penalized plaid model. Note that for very large values of \(\lambda\) the model is a non-overlapping bicluster model. If \(K = 1\) or \(\lambda = 0\), the prior on the labels becomes a uniform prior.

4.1 The full conditionals

The Gibbs sampler and the ICM algorithm rely on the knowledge of the full conditionals of the parameters. In this section, we only deal with important properties of the full conditional of the labels. Further details on the full conditional of the labels and the other parameters of the model are considered in Section C of the Supplemental Online Material.

4.1.1 The labels

As before, \(\theta\) denotes the set of parameters of the model. Let \(z_{ijk} = y_{ij} - \sum_{k' \neq k} \rho_{ik} k_{jk} (\mu_{k'} - \alpha_{ik'} - \beta_{jk'})\). Note that \(y_{ij} = \prod_{k=1}^K (1 - \rho_{ik} k_{jk})\). For a given \(k\), we will write \(y_{jk} = \prod_{k'=1}^K (1 - \rho_{ik} k_{jk})\).

Consider the quantities

\[
A_{ik} = \exp \left\{ -\frac{1}{2} \sum_{j \in J_k} \frac{(z_{ijk} - \mu_k - \alpha_{ik} - \beta_{jk})^2}{\sigma^2(\rho_i * k_j)} \right\} \left( \prod_{j \in J_k} \sigma^2(\rho_i * k_j) \right)^{-1/2},
\]

\[
B_{ik} = \exp \left\{ -\frac{1}{2\sigma_0^2} \sum_{j \in J_k} (y_{ijk} - \mu_0)^2 \right\} \left( \sigma_0^2 \right)^{-1} \left( \frac{\sigma_0}{\sum_{j \in J_k} y_{ijk} \sigma_0} \right)^{1/2},
\]

\[
C_{ik} = \exp \left\{ -\frac{1}{2} \sum_{j \in J_k} (1 - y_{ijk}) \left( \frac{z_{ijk}}{\sigma^2(\rho_i * k_j)} + \log \sigma^2(\rho_i * k_j) \right) \right\},
\]

\[
D_{ik,\rho_{ik}} = \exp \left\{ \frac{1}{2} \sum_{j \in J_k} (1 - y_{ijk}) \left( \frac{z_{ijk}^2}{\sigma^2(\rho_i * k_j)} + \log \sigma^2(\rho_i * k_j) \right) \right\}.
\]

In what follows, \(\rho_{(-ik)}\) will denote the set of all row labels except \(\rho_{ik}\). One can show (see Section C of the Supplemental Online Material) that the full conditionals of \(\rho_{ik}\) satisfy

\[p(\rho_{ik} | y_{ij}, (\rho_{(-ik)}, k), \theta) \propto A_{ik}^{\rho_{ik}} B_{ik}^{1-\rho_{ik}} C_{ik}^{1-\rho_{ik}} D_{ik,\rho_{ik}} \pi(\rho_{ik}),\]

where \(\pi(\rho_{ik}) = \exp\{-\lambda \sum_{j \sum_{k'=1}^K \rho_{ik} k_{jk} + y_{ijk} + (1 - y_{ijk}) k_{jk} (\rho_{ik} - 1)\}\). In particular, the ratio \(p(\rho_{ik} = 1 | y_{ij}, (\rho_{(-ik)}, k), \theta) / p(\rho_{ik} = 0 | y_{ij}, (\rho_{(-ik)}, k), \theta)\) is given by

\[A_{ik} B_{ik} C_{ik} D_{ik,1}^{-1} D_{ik,0}^{-1} \exp\{-\lambda \sum_j (1 - y_{ijk}) k_{jk}\}.
\]

The term \(D_{ik,\rho_{ik}}\) may be ignored for models whose variances do not depend on \((i,j)\). In particular, for the plaid model, the logarithm of this ratio is \(-1/(2\sigma^2) \sum_{j \in J_k} \{ (z_{ijk} - \mu_k - \alpha_{ik} - \beta_{jk})^2 + (1 - y_{ijk}) z_{ijk}^2 + y_{ijk} (y_{ij} - \mu_0)^2 \} - \lambda \sum_j (1 - y_{ijk}) k_{jk}\). The full conditional for \(k_{jk}\)'s are found in a similar way by symmetry.
Next, consider the non-overlapping bicluster model. Note that in this case the term $C_{ik}$ may be conveniently written as

$$
\exp \left\{-\frac{1}{2} \sum_{k' \geq 0} \rho_{ik'} \sum_{j \in J_k \cap J_{k'}} \frac{z_{ikj}}{\sigma_{k'j}^2} \right\} \prod_{k' \geq 0 \setminus k} \sigma_{k'j}^{-c_{ik'k}},
$$

where $c_{ik'k}$ is the number of cells in the set $\{(i, j) \in B_k : j \in J_k\}$. Also note that the term $B_{ik}$ may be written as $\exp(-\rho'_{ik}/2\sigma_0^2) \sum_{j \in J_{ik} \cap J_k} \sigma_0^{-c_{ik0}}$, for $z_{ikj} = y_{ij} - \mu_0$, and where $J_0$ and $c_{i0k}$ are defined as before but for the zero-bicluster. We have

$$
p(\rho_{ik} = 1 | \{y_{ij}\}, (\rho_{(-ik)}, \kappa), \theta) = \frac{p(\rho_{ik} = 1 | \{y_{ij}\}, (\rho_{(-ik)}, \kappa), \theta)}{p(\rho_{ik} = 0 | \{y_{ij}\}, (\rho_{(-ik)}, \kappa), \theta)} = \sigma_k^{-c_k} \prod_{k' \geq 0 \setminus k} \sigma_{k'j}^{-c_{ik'k}} \exp \left\{-\frac{1}{2} \sum_{j \in J_k} (z_{ikj} - \mu_k - \alpha_{ik} - \beta_{jk})^2 + \frac{1}{2} \sum_{k' \geq 0 \setminus k} \rho_{ik'} \sum_{j \in J_{ik} \cap J_{k'}} \frac{z_{ikj}}{\sigma_{k'j}^2} \right\}.
$$

Let

$$
E_{ik} = \prod_{k' \geq 0 \setminus k} \sigma_{k'j}^{-c_{ik'k}} \exp \left\{\frac{1}{2} \sum_{k' \geq 0 \setminus k} \rho_{ik'} \sum_{j \in J_{ik} \cap J_{k'}} \frac{z_{ikj}^2}{\sigma_{k'j}^2} \right\}.
$$

Performing the Gibbs sampler for the row labels in this model would favor inclusion, that is, $\rho_{ik} = 1$, if $A_{ik} E_{ik} > 1$. That is, inclusion is favored when the likelihood associated with having the row $i$ in bicluster $k$ is larger than that associated with having row $i$ elsewhere. Although this is very reasonable, it is not the only argument to include row $i$ in bicluster $k$. Cheng and Church [9] thought of favoring inclusion when the average square error $c_{ik}^{-1} \sum_{j \in J_k} (y_{ij} - \mu_k - \alpha_{ik} - \beta_{jk})^2$ is smaller than the current estimate of the bicluster variance $\sigma_k^2$. Also, a row is a candidate to be eliminated from bicluster $k$ if the average square error is larger than the current estimate of the bicluster variance $\sigma_k^2$. We may incorporate these otherwise ad hoc ideas into a Metropolis–Hastings sampling procedure for sampling the labels as follows. Given the current values of the labels $\kappa_j$, our proposal $q_{ik}(\rho'_{ik} | \kappa_j)$ proposed $\rho'_{ik}$ with probability proportional to $\sigma_k^{-\rho'_{ik} c_k} e^{-\rho'_{ik} c_k/2E_{ik}'_{ik}}$. Note that an inclusion, that is, $\rho'_{ik} = 1$, may only be proposed if $\rho_{ik} < \kappa_j = 0$ for all $1 \leq k' \neq k$, and $j \in J_k$. Hence, only the rows $i$ for which all cells in $\{i\} \times J_k$ are in the zero-bicluster may be proposed for inclusion in the $kth$ bicluster. For these admissible rows, $E_{ik} = \sigma_0^2 \exp((1/2\sigma_0^2) \sum_{j \in J_k} z_{ikj}^2)$. Therefore, the proposal will favor inclusion if $\sigma_k^2 / \sigma_0^2 < \exp((s_{k0}^2 - \sigma_0^2) / \sigma_0^2)$, where $s_{k0}^2 = \sum_{j \in J_k} z_{ikj}^2 / c_k$. That is, inclusion is favored when the current 'sample' estimate of the variance in the zero-bicluster of the columns in bicluster $k$, $s_{k0}^2$, is relatively large in comparison to the overall variance of the zero-bicluster, $\sigma_0^2$. The Metropolis–Hastings acceptance ratio is

$$
\alpha(\rho'_{ik}, \rho_{ik}) = \min \left\{1, \exp \left(-\frac{\rho'_{ik} - \rho_{ik}}{2} \left[ \sum_{j \in J_k} \frac{(z_{ikj} - \mu_k - \alpha_{ik} - \beta_{jk})^2}{\sigma_k^2} - c_k \right] \right) \right\}.
$$

Note that the proposal $\rho'_{ik} = \rho_{ik}$ is always accepted. Also always accepted are the proposals $\rho'_{ik} = 1$, that is, inclusion of row $i$ in bicluster $k$, if $(c_k)^{-1} \sum_{j \in J_k} (z_{ikj} - \mu_k - \alpha_{ik} - \beta_{jk})^2 \leq \sigma_k^2$, and
\( \rho_{ik} = 0 \), that is, removal of row \( i \) from bicluster \( k \), if \((c_k)^{-1} \sum_{j \in J_k} (e_{ijk} - \mu_k - \alpha_{ik} - \beta_{jk})^2 \geq \sigma_k^2 \).

The procedure is similar for the plaid model: it suffices to set \( \sigma_k = \sigma_0 = \sigma \) in the earlier formulas.

Note that these moves do not consider moving or swapping rows between arbitrary bicluster. Only swaps between one fixed bicluster and the zero-bicluster are allowed. Also note that these moves are sufficient to move around all biclusters. However, this procedure might be inefficient in the sense that it may take several moves to swap rows between biclusters. The root of this problem lies in the fact that most biclustering algorithms work with only one bicluster at the time. Once a bicluster and its parameters have been estimated, the labels already estimated are no longer touched. Instead, a search for a new bicluster starts. Once the search for new biclusters is finished, the labels may be re-estimated. However, most of the time this latter step is not done. We note that this type of procedure resembles ICM over the biclusters, that is, only one bicluster is estimated at the time by letting the others fixed. To allow a move between biclusters in the non-overlapping model, we might proceed as follows. We have to propose two biclusters: \( k' \) the ‘targeted’ bicluster and \( k \) the ‘selected’ bicluster. In order to be able to propose a move of row \( i \) from the targeted bicluster \( k' \) to the selected bicluster \( k \), all columns in \( J_k \) need to be included in the \( k' \)th bicluster, that is, \( J_k \cap J_{k'} = J_k \). This is the condition similar to the one already encountered when considering a move from the zero-bicluster to the \( k \)th one. That is, we must have \( E_{ik} = \sigma_{k'}^2 \exp\{(1/2\sigma_{k'}^2) \sum_{j \in J_k} (y_{ij} - \mu_{k'} - \alpha_{ik} - \beta_{jk})^2\} \). Consider the same proposal as before. However, this time \( k' \) needs to be proposed as well. This may be done uniformly among the admissible biclusters. Note that in this case, the ‘exclusion’ move corresponds to the reversal move, that is, of moving the \( i \)th row in the \( k \)th bicluster to the \( k' \)th bicluster. This is possible since the \( k \)th bicluster is admissible for the reversal move. The move will be favored if \( \sigma_k^2/\sigma_{k'}^2 < \exp\{(s_{kk}^2 - \sigma_k^2)/\sigma_{k'}^2\} \), where \( s_{kk}^2 = \sum_{j \in J_k} (y_{ij} - \mu_k - \alpha_{ik} - \beta_{jk})^2/c_k \). That is, the move is favored when the current ‘sample’ estimate of the variance in the \( k' \)th bicluster of the columns in bicluster \( k \), \( s_{kk'}^2 \), is relatively large in comparison to the overall variance of the \( k' \)th bicluster, \( \sigma_{k'}^2 \). Not surprisingly, the Metropolis–Hastings acceptance ratio and the general conclusions about when the moves are always accepted are exactly the same as before. This is due to the symmetry concerning all the biclusters in the formulas. We would like to stress that although these moves are always reasonable for the selected bicluster \( k \), they are not necessarily good moves for the targeted bicluster \( k' \). The best balance move may be obtained by performing the Gibbs sampler on the labels (see the comment aforementioned).

4.1.2 The penalty parameter

The penalty \( \lambda \) may be fixed a priori to a suitable value. However, in the absence of information about its value, one may choose to estimate it. In this latter case, \( \lambda \) is the penalty parameter of the model. We assume a gamma\((a, b)\) prior to it. The full conditional of \( \lambda \), \( p(\lambda | (\rho, \kappa), [y_{ij}]) \), is proportional to

\[
\pi(\rho, \kappa | \lambda) \pi(\lambda) \propto \prod_{ij} [Z_{ij}(\lambda)]^{-1} \lambda^{a-1} \exp\{-\lambda(b + |1 - \gamma_{ij} - \sum_k \rho_{ik}\kappa_{jk}|)\},
\]

where \( Z_{ij}(\lambda) = \sum_k \sum_{\rho_{ik}, \kappa_{jk}} \exp\{-\lambda|1 - \gamma_{ij} - \sum_k \rho_{ik}\kappa_{jk}|\} = 1 + e^{\lambda}([1 + e^{-\lambda}]^K - 1) \). To generate \( \lambda \) in the MCMC sampling, we use a Metropolis–Hastings step. The proposal for \( \lambda' \) is again a gamma distribution

\[
g(\lambda' | \lambda) \propto \lambda'(a-1) \exp \left\{ -\lambda'(b + \sum_{i,j} |1 - \gamma_{ij} - \sum_k \rho_{ik}\kappa_{jk}|) \right\}.
\]
Thus, the acceptance probability ratio in the Metropolis–Hastings step is given by

$$
\min \left( 1, \prod_{i,j} \left( \left[ Z_{ij}(\lambda') \right]^{-1} Z_{ij}(\lambda) \right) \right) = \min \left( 1, e^{\rho(\lambda-\lambda')} \frac{X(\lambda)}{X(\lambda')} \right),
$$

where $X(\lambda) = (e^{-\lambda} + (1 + e^{-\lambda})K - 1)^{pq}$ for all $\lambda > 0$. In our simulations below, we note that $\lambda$ may be seen as a measure of complexity of the data structure: its value decreases with the number of biclusters and the amount of bicluster overlapping in the data.

### 4.2 Estimating the number of biclusters

The problem of estimating the number of biclusters is very rarely treated in the literature. For the most part, the number of biclusters is either fixed a priori, or estimated sequentially by some ad hoc stopping rule such as stopping when no more biclusters of a minimum size are found. For example, the algorithm of Cheng and Church [9] fixes the number of biclusters $K$ a priori. Then, it discovers one bicluster at a time. At each one of the $K$ iterations, the algorithm starts with an initial bicluster that contains all rows and columns. The previously discovered biclusters are masked with uniform random numbers. The process is repeated until the $K$ biclusters are found. A similar ad hoc criterion is applied in the algorithms of Lazzeroni and Owen [22], Turner et al. [33], and Zhang [35]. Therein, the number of maximum biclusters $K$ is fixed a priori. The optimal number of biclusters is chosen according to a measure of relevance of the last bicluster found. This measure compares a candidate bicluster with a pure noise bicluster. See the aforementioned references for further details.

In this work, we proposed instead two modified versions of the DIC [30] that are suited for biclustering. As pointed out by Celeux et al. [8], in order to properly formulate the DIC, a model needs a focus parameter. Unfortunately, this focus parameter is not obvious for mixture models. Since in the clustering setup, the labels may be seen as latent variables, we first suggest considering the marginals $E_{p,\kappa}\pi(y|\theta, \rho, \kappa)$ in the computation of DIC. This corresponds to choosing $\theta$ as the focus parameter; that is, our DIC, which we will refer to as $\text{DIC}_m$, is given by

$$
\text{DIC}_m = -2E_\theta[\log (E_{p,\kappa}\pi(y|\theta, \rho, \kappa))|y] + p_m(\tilde{\theta}_m),
$$

where $\tilde{\theta}_m$ denotes the MAP estimator of $\theta$, and

$$
p_m(\tilde{\theta}_m) = -2E_\theta[\log (E_{p,\kappa}\pi(y|\theta, \rho, \kappa))|y] + 2\log (E_{p,\kappa}\pi(y|\tilde{\theta}_m, \rho, \kappa))
$$

is the so-called effective dimension. As suggested by Celeux et al. [8], we use the MAP estimator so as to have a positive effective dimension. This measure works well in our experiments. However, it is computationally expensive, since the labels’ marginals have to be computed. An alternative measure to the $\text{DIC}_m$ is the conditional DIC:

$$
\text{DIC}_c = -2E_{\tilde{\theta},\rho,\kappa}[\log p(y|\theta, \rho, \kappa)|y] + p_c(\tilde{\theta}, \tilde{\rho}, \tilde{\kappa}),
$$

where $(\tilde{\theta}, \tilde{\rho}, \tilde{\kappa})$ is the MAP estimator of $(\theta, \rho, \kappa)$, and

$$
p_c(\tilde{\theta}, \tilde{\rho}, \tilde{\kappa}) = -2E_{\tilde{\theta},\rho,\kappa}[\log p(y|\theta, \rho, \kappa)|y] + 2\log p(y|\tilde{\theta}, \tilde{\rho}, \tilde{\kappa})
$$

is the corresponding effective dimension. Note that $\text{DIC}_c$ is similar to the original DIC proposed by Spiegelhalter et al. [30]. The difference is that we use the MAP estimator $(\tilde{\theta}, \tilde{\rho}, \tilde{\kappa})$ to compute the effective dimension, whereas Spiegelhalter et al. use the expected value of $(\theta, \rho, \kappa)$. In our
experiments, DICc works as well as DICm, but its computation is much faster. This finding may be a bit surprising at first, since after all, in the clustering setup, the labels are not thought of as parameters but as latent variables. However, a closer look at the biclustering model reveals that the labels \((\rho, \kappa)\) are not equivalent to the labels in the clustering setup. In fact, our analysis of biclustering as a mixture model in Section 3.1 clearly shows that it is the product \(\rho_{\kappa j k}\) that is equivalent to the cluster labels. The two sets of labels \((\rho, \kappa)\) are more similar to parameters than latent variables. This is especially true for the overlapping bicluster model. Note that in the clustering setup, this latter model does not and cannot exist.

In our experiments, we have also compared these two measures with more classical ones, such as the Akaike information criterion (AIC) [1] and the Bayesian information criterion (BIC) [28]. In order to compute them, we have used the value of the parameters \((\hat{\Theta}, \hat{\rho}, \hat{\kappa})\) that maximizes the likelihood among the values generated by our MCMC sampler as the estimator of the maximum likelihood estimator.

4.3 Measuring MCMC convergence

As a stopping rule for exploring the support of the posterior distribution, we used the Kolmogorov–Smirnov (KS) test as suggested by Robert and Casella [25, Ch. 8, pp. 370–372]. We compare the ‘empirical’ distribution of the log-likelihood from MCMC sub-samples taken \(G\) samples apart. The gap \(G\) between the sub-samples is used to run the KS test with two (quasi-) independent populations. To monitor the MCMC convergence to the posterior distribution, we plotted the KS test \(p\)-values against the number of iterations.

5. Experiments with artificial data

In this section, we show the results of applying the Bayesian biclustering models to diverse simulated data sets. The goal is to study the behavior of the models under two real complexities in the data: the number of biclusters and the amount of bicluster overlapping. We will see that augmenting the number of biclusters deteriorates the performance of all the algorithms studied. The same is true if the amount of overlapping between biclusters increases, though the performance of many of the algorithms is not affected as much as when the number of biclusters is increased. We also study the robustness of our algorithms to departures from normality in the errors.

The data were simulated for a number of biclusters \(K\in\{2, 4, 6, 8, 10\}\). The biclusters were allowed to overlap. For each \(K\), we generated three scenarios of data with \(p\in\{100, 400, 1000\}\) rows and \(q=50\) columns. The first scenario corresponded to non-overlapping biclusters. These data were generated according to the non-overlapping model. The second and third scenarios allowed for a small and sizable amount of overlapping biclusters, respectively. These two scenarios were simulated according to the Bayesian plaid model described in this paper. Figure 3 shows some examples of our simulated data sets. To study the robustness of the algorithms to lighter tail distributions of the errors, we also generated data sets as before but with Student-\(T\) distributed-errors with 5, 15, and 30 degrees of freedom instead of normal-distributed ones.

More specifically, for \(K=2\), \(\mu_k\sim N(2k, 0.05)\), \(k=1, 2\), and \(\mu_0\sim N(0, 0.05)\). For \(K=4\) and \(K=6\), the data were generated with \(\mu_k\sim N(-2(k+1), 0.05)\), \(k=1, \ldots, K\). For \(K=8\) and \(K=10\), \(\mu_k\sim N((4k^2 + 8)/(k+1), 0.05)\), \(k=1, \ldots, K\). For \(K\in\{4, 8, 10\}\), \(\mu_0\sim N(0, 0.05)\), and for \(K=6\), \(\mu_0\sim N(1, 0.05)\). The row effects \(\alpha_{ik}\) were generated from normal distributions with means set to \(\mu_{\alpha_{ik}}=2(1+e^{-k})^{-1} - \sum_r 2(1+e^{-r})^{-1}/r_k\), and variances set equal to their prior variances, \(k=1, \ldots, K\). The column effects \(\beta_{jk}\) were generated similarly. For the non-overlapping bicluster data, we generated the variances as draws from the distributions \(\sigma_0^2\sim\text{inverse-}\chi^2(0.01, 3)\), and \(\sigma_k^2\sim\text{inverse-}\chi^2(s_k^2, 3)\), where \(s_k^2=k/100\), \(k=1, \ldots, K\). For the overlapping bicluster data, we draw the variance \(\sigma^2\) from an inverse-\(\chi^2(0.01, 3)\). The
hyper-parameters were set to the values: $\nu = \nu_0 = 1$, $s^2 = s_0^2 = 0.05$, and $\sigma^2_\alpha = \sigma^2_\beta = \sigma^2_\mu = \sigma^2_{\mu_0} = 0.5$. For the penalized plaid model, the prior of $\lambda$ was a gamma distribution with shape $a = 16$ and scale $b = 8$ (that is, the mean prior was set to 2, and the variance to 0.25). We create four independent data sets for each of the 180 set of parameters.

5.1 Model comparison

In this section, we compare several biclustering models and algorithms for a known biclustering comprising $K$ biclusters. The case of unknown $K$ is dealt with in the next section. The models compared were (A) the non-overlapping bicluster model, (B) the plaid model, and (C) our penalized plaid model with double exponential prior on the labels. The estimation methods compared were (I) the Gibbs sampler and (II) the Metropolis–Hastings algorithm. The Gibbs and Metropolis–Hastings samplers were run with 10,000 burn in iterations. Only 10,000 iterations were kept after the burn in to do the comparison analysis between the models. To initialize our algorithms, we combined the $K$-means algorithm [34] with a technique similar to that of Turner et al. [33]. Further details can be found in Section D of the Supplemental Online Material. As explained in Section 4.3, we monitored the convergence of the algorithm by looking at the KS test.

We compare our six model-method combinations with five other alternative biclustering algorithms. These are, the Cheng and Church’s biclustering algorithm, the plaid model of Turner et al. [33], and the FABIA [17], SAMBA [31], and Spectral [21] methods. We decided to use the plaid model implementation of Turner et al. because it seems to perform better than the original Lazzeroni and Owen’s method. FABIA (factor analysis for bicluster acquisition) is a multiplicative model inspired by factor analysis with $K$ factors. The biclusters are found by detecting sparsity on the row and column effects. The parameters are estimated by a variational EM algorithm. SAMBA (statistical-algorithmic method for bicluster analysis) is based on graph theory and probabilistic models of the data. The gene expression matrix is represented as a bipartite graph where the two sets of nodes are the genes and the experimental conditions. An edge from a gene to a condition means that the expression level of that gene changes significantly in that condition. A bicluster is a weighted sub-graph where the subset of genes are co-regulated under the subset of conditions. The weights in the graph are found according to a probabilistic model. The number of biclusters is fixed in advance. A heuristic algorithm is used to look for the biclusters, which are allowed to overlap. The Spectral algorithm uses a singular value decomposition to identify bicluster structures in the data. As in FABIA, Spectral also assumes a multiplicative model. The expression level of a specific gene $i$ in a condition $j$ is expressed as a
product of three independent factors. The first factor is called the hidden base expression level which is constant. The second factor represents the gene expression tendencies across different conditions. The last factor represents the role of particular conditions over the gene expression. Biclustering corresponds to finding the underlying block structure of the first factor. Therefore, the biclustering has a checkerboard-like structure. As in SAMBA, the number of biclusters is fixed in advance.

5.1.1 The F1 measure

In order to compare the performance of the models and algorithms, we used the so-called F1 measure [2]. The F1 measure has been extensively used in the text mining literature and recently introduced to the biclustering literature [26,33]. It results from the harmonic mean of precision and recall. These indices are defined as follows. Let \( B \) be a bicluster, \( r_B \) be the number of genes in \( B \), \( c_B \) be the number of conditions in \( B \), and \( n_B = r_B c_B \) be the number of elements in \( B \). Suppose that we wish to compare a target bicluster \( A \) and a known bicluster \( B \). Then

\[
\text{recall} = \frac{(r_A \cap B)(c_A \cap B)}{n_B}, \quad \text{precision} = \frac{(r_A \cap B)(c_A \cap B)}{n_A}.
\]

Recall measures the proportion of elements in \( B \) that belong to \( A \) and precision measures the proportion of elements in \( A \) captured in \( B \). Turner et al. [33] used these indices to measure quality of the biclusters but they mistakenly named precision by specificity. The F1 measure is defined as

\[
F_1(A, B) = \frac{2}{\frac{1}{n_A} + \frac{1}{n_B}}.
\]

When several biclusters are to be compared, we can use an F1-type average. Let \( M_1 = \{A_1, \ldots, A_k\} \) be the set of estimated (target) biclusters, and \( M_2 = \{B_1, \ldots, B_\ell\} \) be the set of true (known) biclusters. We measure the similarity of the estimate \( M_1 \) to the true biclustering \( M_2 \) by

\[
S(M_1, M_2) = \left(\frac{1}{k}\right) \sum_{i=1}^{k} \max_j F_1(A_i, B_j).
\]

Note that \( S(M_1, M_2) \leq 1 \), and it is equal to 1.0 if \( M_1 = M_2 \).

A visual summary of the results can be seen in Figures 4 and 5. It is clear from these plots that our six model-method combination algorithms do much better than the other five competing ones, regardless of the degree of overlapping and the amount of departure from normality. In particular, the fitting methods suggested by Cheng and Church [9] for the mixture model, and by Turner et al. [33] for the plaid model are not very competitive in comparison with the Bayesian models proposed in this paper. We carried out an analysis of variance of the square-root of the F1 measure (as suggested by the Box and Cox family of power transformations, [5]) for the 11 models compared. Note that the boxplots in Figure 5 reveal no significant effect on the \( F_1 \)-measure from the amount of overlapping in the biclustering, and no noticeable effect from the amount of departure from normality (measured by the degrees of freedom) on the error distribution. In fact, 91% of the variance observed in \( \sqrt{F_1} \) is explained only by the type of algorithm (58%), the interaction effect between the number of clusters and the type of algorithm (19%), the number of clusters (8.8%), and the interaction between the size of the data and the type of algorithm (5.7%). Therefore, only the type of algorithm, the number of clusters, and the size of the data are influential in the performance of the 11 algorithms compared (these three variables together with their corresponding second- and third-order interactions explain 97% of the variance observed in the data). We also performed an analysis of variance for the three models (penalized plaid, plaid, and Mixture model), fitted using the two methods (the Gibbs sampler or the Metropolis–Hastings sampler) presented in Section 4, for the three amounts of overlapping (no overlapping, moderate overlapping, and heavy overlapping). The ANOVA revealed that a model with all second-order interactions fitted well the results. The number of clusters explains 69% of the variation observed in the response, followed by the type of model (almost 9%) and the interaction between the size of the data and the number of clusters (almost 9%), and the size of the data (6.3%). The Gibbs sampler performs better than the Metropolis–Hastings algorithm.
Figure 4. The $F_1$-measure mean performance of the six model–method combinations and the five alternative algorithms as a function of the number of clusters $K$. ‘Gibbs plaid’ and ‘Metropolis plaid’ stand for the plaid models fitted with the Gibbs sampler and the Metropolis–Hastings procedure, respectively. ‘Gibbs plaid penalized’ and ‘Metropolis plaid penalized’ stand for the penalized plaid models fitted with the Gibbs sampler and the Metropolis–Hastings procedure, respectively. ‘Gibbs CC’ and ‘Metropolis CC’ refer to the Cheng and Church’s algorithm fitted with the Gibbs sampler and the Metropolis–Hastings procedure, respectively. ‘CC’ stands for the original Cheng and Church’s algorithm and ‘Turner’ stands for the plaid algorithm of Turner et al. of Section 4 for the data sets with eight and ten biclusters. Furthermore, a multiple comparison analysis showed that the plaid and penalized plaid models performed similarly and were the best models. We note that the value of the penalty parameter $\lambda$ depends a posteriori on the complexity of the data. Its logarithm decreases with the number of biclusters and the amount of overlapping (see Figure 6) and seems to reach a plateau when the number of biclusters is larger than six. As such the logarithm of $\lambda$ may be used as a measure of data complexity. Note that the complexity is dominated by the number of biclusters when this is large relatively to the data size. The amount of overlapping does not appear to influence it much in this case.

5.2 Choice of model

We applied the criteria described in Section 4.2 for selecting the number of biclusters. We only used the Gibbs sampler on the penalized plaid model to evaluate the criteria, since the above results have shown that this is one of the best combinations for biclustering estimation. Figure 7 shows averages $\text{DIC}_c$, $\text{DIC}_m$, AIC, and BIC as a function of the number of biclusters for $K = 2$ and 10. The averages for each model selection criterion were computed over five different data set replicates. For this simulation, we chose a scenario with moderate overlapping with a number of rows $p = 400$, and a number of columns $q = 50$. The data were generated with Student-$T$ distributed error terms with five and a hundred degrees of freedom (e.g. nearly a normal distribution). The penalized plaid model algorithm was run with a burn-in period of 5000 iterations. The next 5000 iterations were used to compute the four model selection criteria. In order to compute the AIC and BIC criteria, all the model parameters were counted in the penalty term, except for the latent variables $\rho$ and $\kappa$. We had then a total of $(p + q - 1) \times K + 2$ parameters. We note that for $K = 2$, whatever the degrees of freedom, there is a large decrease in the value of all four criteria when passing from a model with one bicluster to a model with two biclusters. In addition,
Figure 5. Boxplots of the $F_1$-measure means for the 11 different models compared. The boxplots are ordered as follows (from left to right) in each panel. GCC, GPL, and GPE stand for the non-overlapping mixture model, the plaid model, and the penalized plaid model, respectively, fitted with the Gibbs sampler; MCC, MPL, and MPE stand for the non-overlapping mixture model, the plaid model, and the penalized plaid model, respectively, fitted with the Metropolis–Hastings procedure; CC stands for the original method suggested by Cheng and Church to fit a mixture model; FAB stands for the FABIA algorithm, SAM, for the SAMBA algorithm, SPE, for the Spectral method, and TUR, for the plaid model implementation of Turner et al. The first row displays the results for each scenario. The second row displays the results for different error term distributions (all scenarios combined).

Figure 6. The profile means of the logarithm of the $\lambda$ parameter in the penalized plaid model.

the minima of BIC and AIC are at the true number of biclusters ($K = 2$), whereas the minima of DIC$_c$ and DIC$_m$ are simultaneously at two and three biclusters. The valley is deeper for BIC than the other criteria. When $K = 10$, DIC$_c$, DIC$_m$, and AIC indicate an appropriate number of biclusters, whereas BIC suggests about four biclusters, that is nearly half the number of true
biclusters. These observations together with other simulations (not shown here) lead us to only consider DIC$_c$, DIC$_m$, and AIC as reliable criteria to choose the number of biclusters in the data.

Note that the DIC$_c$ and DIC$_m$ curves are almost indistinguishable. This indicates that the expected marginal likelihood is similar to the expected conditional likelihood in these examples. This is also seen in the gene expression application described in the next section. This is partly due to a relatively small effective dimension compared to the expected log-likelihood. Also, after thousands of iterations of the Gibbs sampler, the distribution of the bicluster labels becomes very asymmetric, strongly signaling the final bicluster memberships. As a consequence, the expected log-likelihood is very similar to the log of the expected likelihood.

6. Applications to gene expression arrays

We have applied our penalized plaid model to elucidate the biclustering structure of the gene expression data associated with the yeast cell cycle data [13] described in the introduction. The original data contained some missing values (1.9% of the data), which we imputed as in [22] by using the sum of the row and column means less the overall mean.

For comparison purposes, we have also applied the other five biclustering algorithms described in Section 5, that is, the Cheng and Church’s biclustering algorithm, the plaid model of Turner et al. [33], and the FABIA [17], SAMBA [31], and Spectral [21] methods. In order to compare the results of these methods and our penalized plaid biclustering method on these data, we followed the procedure suggested by Prelic et al. [24]. This consists of evaluating the biological relevance of the obtained biclusters with respect to the enrichment given by the GO annotations. This is explained in the following.
6.1 Biological interpretation of the biclusters

We use the gene ontology (GO) database [3] to investigate which terms are under- or over-represented on each of the estimated bicluster and on each ontology. The GO project is a major bioinformatics initiative with the aim of standardizing the representation of gene attributes across species and databases. The GO project consists of three structured controlled vocabularies or ontologies that describe gene products in terms of their associated biological processes, cellular components, and molecular functions in a species-independent manner.

If a high fraction of the genes forming a bicluster, say Bicluster $k$, are assigned to a GO term for a given ontology, then we can say that this GO term is over-represented in Bicluster $k$ for that ontology. In general, this is difficult to assess, since the large number of genes makes it likely to find spurious but significant relations between GO terms and biclusters just by chance. The Bioconductor project [14] uses the Fisher exact test to measure such relations. In order to control the rate of false positive relations, we employed a Bonferroni correction by adjusting the number of GO terms present in the data. Even after this adjustment, we found several relations that are significant within each estimated bicluster.

In what follows, as in the work of Lazzeroni and Owen [22], we will say that gene $i$ is up-regulated (respectively, down-regulated) within the $k$th bicluster, if $\mu_k + \alpha_{ik}$ is positive (respectively, negative). We will say that the effect of the $j$th condition is positive (respectively, negative) within the $k$th bicluster if $\mu_k + \beta_{jk}$ is positive (respectively, negative).

6.2 Results

Figure 8 shows the proportion of biclusters for which GO terms are over-represented at different levels of significance for each type of GO function for each of the five alternative methods described earlier and our penalized plaid biclustering method. We note that FABIA and the Cheng and Church’s algorithms did not find any bicluster in the data. In general, the best performing method in terms of GO enrichment seems to be the penalized plaid biclustering. This method is clearly superior to the other five in terms of GO biological processes and GO component cellular categories. However, the Spectral algorithm appears to perform better than the other methods in terms of GO molecular function categories. A closer look at the Spectral results reveals that this method found only four biclusters. Moreover, all four biclusters contain exactly the same genes (only seven genes) which are highly enriched with GO molecular functions.

Figure 8. GO enrichment results associated with all five biclustering methods described in Section 5 and the penalized plaid model. From left to right, the plots correspond to the GO biological process, GO cellular component and GO molecular function enrichments. From darker to lighter color, the bars correspond to the penalized plaid model, Spectral, the plaid model of Turner et al., and SAMBA. The significance levels considered were (from left to right in each plot): 0.01, $10^{-4}$, $10^{-5}$, and $10^{-6}$. 
6.3 **The findings of the penalized plaid model**

Next, we describe the findings from the application of the penalized plaid model to these data. Figure 9 shows that the DIC\textsubscript{m} and DIC\textsubscript{c} criteria select thirteen biclusters for the yeast cell cycle data, while the AIC selects five biclusters. The AIC criterion differs from the DIC because this time the data size is large enough to make the penalty term much more relevant. Recall that the number of parameters in our model is proportional to the data size. The five-bicluster solution aggregates too much of the data and hides interesting small biclusters, and consequently, the amount of enrichment of the larger biclusters selected by AIC is smaller than the enrichment found in the thirteen-bicluster solution favored by DIC\textsubscript{m}. In addition, the mean squared error (MSE) associated with the five-bicluster solution is 0.24, whereas the MSE associated with the thirteen-bicluster solution is 0.20. That is, the DIC-favored solution offers a 20% reduction in the MSE. We also note that the F1 measure between the five-bicluster and the thirteen-bicluster solution is 0.51, which indicates a good agreement between these two solutions. In fact, Bicluster 3 of the thirteen-bicluster solution, which is very interesting biologically (see below), and Bicluster 3 of the five-bicluster solution share the same experimental conditions and approximately 88% of their genes. Moreover, almost all the experimental conditions of Bicluster 9 in the thirteen-bicluster solution (seven conditions out of eight) are included in Bicluster 5 of the five-bicluster solution and they share approximately 76% of their genes. Therefore, since smaller biclusters are more appealing to biologists (they are easier to analyze and more likely to be enriched), we decided to continue the analysis of the biclustering chosen by the DIC\textsubscript{m}. The number of genes and the number of conditions associated with each bicluster are displayed in Table 1. About 8% of the genes and 14% of conditions are in a single bicluster and only 3% of the genes are in the zero-bicluster. 86% of the genes and 80% of the conditions are found in overlapping biclusters. Biclusters 4 and 7 are the two largest, with 1388 genes and 1381 genes, respectively. Bicluster 5 is the smallest one. It is composed of five genes. We note that most of the genes are annotated by GO terms, that is, they have a known biological function within GO.

The smaller biclusters, in terms of the number of genes or experimental conditions included, are Biclusters 3, 5, 6, 9, 10, 11, and 13. In general, the smaller biclusters are the most interesting to analyze since their genes are more likely to share common functions. They are shown in Figure 10 along with the gene ($\mu_k + \alpha_{ik}$) and experimental condition ($\mu_k + \beta_{jk}$) effects.

Bicluster 3 has 908 genes, all down-regulated, and consists of five experimental conditions of sporulation. All the conditions act negatively. The main molecular functions that characterize this bicluster are associated with structural constituents of ribosomes and rRNA binding. The biological processes are associated with cytoplasmic translation, bio-synthetic and catabolic processes. The cellular components correspond to ribosomes and cytoplasm. This bicluster is similar to the third bicluster found in Lazzeroni and Owen [22]. Bicluster 5 contains only five genes. This
Table 1. Summary of the results for the yeast cell cycle data.

| Bicluster | Number of conditions | Number of genes | Annotated genes | Bicluster | Number of conditions | Number of genes | Annotated genes |
|-----------|----------------------|-----------------|-----------------|-----------|----------------------|-----------------|-----------------|
| 0         | 5                    | 71              | 71              | 7         | 31                   | 1381            | 1378            |
| 1         | 41                   | 415             | 414             | 8         | 22                   | 680             | 679             |
| 2         | 53                   | 488             | 487             | 9         | 8                    | 711             | 711             |
| 3         | 5                    | 908             | 907             | 10        | 3                    | 884             | 882             |
| 4         | 28                   | 1388            | 1383            | 11        | 4                    | 789             | 788             |
| 5         | 36                   | 5               | 5               | 12        | 10                   | 953             | 948             |
| 6         | 7                    | 492             | 491             | 13        | 9                    | 1105            | 1103            |

Figure 10. Biclusters 3, 5, 6, 10, 11, and 13 of the Yeast cell cycle data. The upper subplots correspond to the gene (column) effects, and the right subplots correspond to the experimental conditions (row) effects.

bicluster is characterized by hydrolase activity (gene YLR286C), cell division (genes YNL327 and YNL066W), and cellular budding (genes YKL185W, YNL327W, and YNL066W). Bicluster 6 consists of 492 genes, six time-points of the sporulation and one time-point of the diauxic shift. In this bicluster, the genes are up-regulated and the experimental conditions present positive effects. The over-represented GO-term functions are related to the threonine-type endopeptidase activity, the DNA binding, the cell cycle, and the DNA replication processes. This bicluster is similar to the first bicluster found in [22]. Bicluster 10 consists of 884 genes and 3 time-points of the sporulation. The over-represented GO-term functions are related to catalytic and oxidoreductase activities, and the catabolic process. Bicluster 11 consists of three times of the sporulation. Its associated GO-term significant functions are related to the structural constituent of ribosomes and rRNA binding, the cytoplasmic translation, the metabolic process, and the gene expression. All the genes were down-regulated in Biclusters 10 and 11. Moreover, the
experimental conditions presented negative effects in these biclusters. Bicluster 13 consists of 1105 genes and nine cdc15 experimental conditions. These conditions from the mitotic cell cycle show negative effects. Almost all the genes were down-regulated (95%). The associated GO-term significant functions are related to proteolysis, the cellular protein catabolic process, and the modification-dependent protein catabolic process.

The zero-bicluster has 71 genes and five time-points from the alpha-factor (alpha.77), the sporulation (spo5.2), the heat shock (heat.10), and the diauxic shift (diau.a, diau.b). No gene is over-represented and no condition has an effect in this zero-bicluster.

Although most of the small biclusters consist primarily of experimental conditions from sporulation, these biclusters are very different. They play different biological roles in terms of cellular components, biological processes, and molecular functions. Lazzeroni and Owen [22] detected similar biclusters. However, the ones found with our model are also related to other functions that have not been reported before.

7. Conclusions

In this work, we introduced an extension of the plaid model, the penalized plaid model. This model incorporates a penalty parameter, \( \lambda \), that controls (or measures) the amount of bicluster overlapping. Within this model, we find the original plaid model of Lazzeroni and Owen [22] when \( \lambda = 0 \). At the other extreme (\( \lambda \to +\infty \)), we find a homogeneous-variance version of the non-overlapping model of Cheng and Church [9]. We have proposed both the Gibbs sampler and the Metropolis–Hastings algorithm to estimate the parameters. We note that our Metropolis–Hastings sampler applied to the mixture model with bicluster-dependent variances justifies the optimality of the otherwise ad hoc algorithm proposed by Cheng and Church [9].

We have shown that although the biclustering problem may be studied as a mixture model, the commonly used (soft) EM algorithm for mixtures does not seem appropriate. Instead, an ICM-like or hard-EM algorithm appears to be more suitable. In fact, we note that most of the underlying algorithms for biclustering reported in the literature may be justified using hard-EM or ICM. However, we have shown through our simulations that the results derived from our MCMC implementation of the models are better than the original Cheng and Church and plaid model algorithms. Our results are also better than the results from other alternative algorithms such as FABIA, SAMBA, and Spectral.

We defined the DIC that seems specially suitable for the biclustering problem. Our DIC for biclustering was inspired by the work of Celeux et al. [8] who noticed that the original DIC was not well defined for mixture models. In our experiments, our marginal and conditional DIC criteria performed very well. Although we believe that, in principle, the marginal DIC should be preferred to the conditional DIC, we admit that this latter criterion is easier and faster to compute.

We applied our penalized plaid model to the yeast cell cycle data of Eisen et al. [13]. We found thirteen biclusters in the data as indicated by the conditional DIC model selection criterion. Among the biclusters, we obtained the main biclusters found in Lazzeroni and Owen [22]. We showed that these biclusters are very different as indicated by their diverse biological roles obtained using the GO ontology annotations [14].

A JAVA implementation of the penalized plaid model may be found in the website of the second author www.dms.umontreal.ca/~murua/.

Acknowledgements

The authors would like to thank the editor and the anonymous reviewers for their comments and suggestions to improve the exposition of the paper.
Disclosure statement
No potential conflict of interest was reported by the authors.

Funding
This research was supported by NSERC [grant number 327689-06].

Supplemental data
The supplemental material provides further details on the EM and Bayesian algorithms described in this work. It also provides more information on the procedure to initialize the parameters and labels for the Markov chain Monte Carlo implementation of the penalized plaid model. The URL addresses of the biclustering packages used in this study are also provided in the supplemental material.

Supplemental data for this article can be accessed at 10.1080/02664763.2014.999647.

References
[1] H. Akaike, A new look at the statistical model identification, IEEE Trans. Automat. Control 19 (1974), pp. 716–723.
[2] J. Allan, J. Carbonell, G. Doddington, J. Yamron, and Y. Yang, Topic detection and tracking pilot study: Final Report, Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop, Landsdowne Conference Resort, Landsdowne, Virginia, USA, 1998, pp. 194–218.
[3] M. Ashburner, C.A. Ball, J.A. Blake, D. Botstein, H. Butler, J.M. Cherry, A.P. Davis, K. Dolinski, S.S. Dwight, J.T. Eppig, M.A. Harris, D.P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J.C. Matise, J.E. Richardson, M. Ringwald, G.M. Rubin, and G. Sherlock, Gene ontology: tool for the unification of biology, Nat. Genet. 25 (2000), pp. 25–29.
[4] J. Besag, On the statistical analysis of dirty pictures, J. R. Stat. Soc. Ser. B 48 (1986), pp. 259–302.
[5] G.E.P. Box and D.R. Cox, An analysis of transformations, J. R. Stat. Soc. Ser. B 26 (1964), pp. 211–243.
[6] S. Busygin, O. Prokopyev, and P.M. Pardalos, Biclustering in data mining, Comput. Oper. Res. 35 (2008), pp. 2964–2987.
[7] J. Caldas and S. Kaski, Bayesian biclustering with the plaid model, Proceedings of the IEEE International Workshop on Machine Learning for Signal Processing XVIII, Cancun, Mexico, 2008, pp. 291–296.
[8] G. Celeux, F. Forbes, C.P. Robert, and D.M. Titterington, Deviance information criteria for missing data models, Bayesian Anal. 1 (2006), pp. 651–674.
[9] Y. Cheng and G. Church, Biclustering of expression data, International Conference on Intelligent Systems for Molecular Biology, Vol. 12, La Jolla, California, USA, 2000, pp. 61–86.
[10] S. Chu, J. DeRisi, M. Eisen, J. Mulholland, D. Botstein, P.O. Brown, and I. Herskowitz, The transcriptional program of sporulation in budding yeast, Science 282 (1998), pp. 699–705.
[11] A. Dempster, N. Laird, and D. Rubin, Maximum likelihood from incomplete data via the EM algorithm, J. R. Stat. Soc. Ser. B 39 (1977), pp. 1–38.
[12] S. Dolnicar, S. Kaiser, K. Lazarevski, and F. Leisch, Biclustering, J. Travel Res. 51 (2012), pp. 41–49.
[13] M. Eisen, P. Spellman, P. Brown, and D. Botstein, Cluster analysis and display of genome-wide expression patterns, Genetics 95 (1998), pp. 14863–14868.
[14] S. Falcon and R. Gentleman, Using GOSstats to test gene lists for GO term association, Bioinformatics 23 (2007), pp. 257–258.
[15] J. Gu and S. Liu, Bayesian biclustering of gene expression data, The International Conference on Bioinformatics & Computational Biology, BMC Genomics, Vol. 9, Las Vegas, Nevada, USA, 2008, pp. 113–120.
[16] J.A. Hartigan, Direct clustering of a data matrix, J. Amer. Statist. Assoc. 67 (1972), pp. 123–129.
[17] S. Hochreiter, U. Bodenhofer, M. Heusel, A. Mayr, A. Mitterecker, A. Kasim, T. Khramtsova, S.V. Sanden, D. Lin, W. Talloen, L. Bijnens, H.W.H. Gohlmann, Z. Shkedy, and D.A. Clevert, FABIA: Factor analysis for bicluster acquisition, Bioinformatics 26 (2010), pp. 1520–1527.
[18] M.I. Jordan and R.A. Jacobs, Hierarchical mixtures of experts and the EM algorithm, Neural Comput. 6 (1994), pp. 181–214.
[19] H. Katsuhisa and T. Hiroyuki, Statistical estimation of cluster boundaries in gene expression profile data, Bioinformatics 17 (2001), pp. 1143–1151.
[20] C. Kaufmann and R. Sain, Bayesian functional ANOVA modeling using Gaussian process prior distributions, Int. Soc. Bayesian Anal. 5 (2010), pp. 123–150.
[21] Y. Kluger, R. Basri, J.T. Chang, and M. Gerstein, Spectral biclustering of microarray data: co-clustering genes and conditions, Genome Res. 13 (2003), pp. 703–716.

[22] L. Lazzeroni and A. Owen, Plaid models for gene expression data, Statist. Sinica 12 (2002), pp. 61–86.

[23] D.V. Lindley and A.F.M. Smith, Bayes estimates for the linear model, J. R. Stat. Soc. Ser. B 34 (1972), pp. 1–41.

[24] A. Prelić, S. Bleuler, P. Zimmermann, A. Wille, P. Bühlmann, W. Gruissem, L. Hennig, L. Thiele, and E. Zitzler, A systematic comparison and evaluation of biclustering methods for gene expression data, Bioinformatics 22 (2006), pp. 1122–1129.

[25] C.P. Robert and G. Casella, Monte Carlo Statistical Methods, Springer, New York, 1999.

[26] R. Santamaria, L. Quintales, and R. Theron, Methods to bicluster validation and comparison in microarray data, IDEAL 07 Proceedings of the Eighth International Conference on Intelligent Data Engineering and Automated Learning, Birmingham, UK, 2007, pp. 780–789.

[27] C.M. Sara and A.L. Oliveira, Biclustering algorithms for biological data analysis: A survey, IEEE Trans. Comput. Biol. Bioinform. 1 (2004), pp. 24–45.

[28] G.E. Schwarz, Estimating the dimension of a model, Ann. Statist. 6 (1978), pp. 461–464.

[29] R. Sokal and C. Michener, A statistical method for evaluating systematic relationships, Univ. Kansas Sci. Bull. 38 (1958), pp. 1409–1438.

[30] D.J. Spiegelhalter, N.G. Best, B.P. Carlin, and A. van der Linde, Bayesian measures of model complexity and fit, J. R. Stat. Soc. Ser. B 64 (2002), pp. 583–640.

[31] A. Tanay, R. Sharan, and R. Shamir, Discovering statistically significant biclusters in gene expression data, Bioinformatics 18 (2002), pp. S136–S144.

[32] A. Tanay, R. Sharan, and R. Shamir, Biclustering algorithms: A survey, in Handbook of Computational Molecular Biology, S. Aluru, ed., Computer and Information Science Series, Chapman & Hall/CRC Press, Boca Raton, FL, 2005.

[33] H. Turner, T. Bailey, and W. Krzanowski, Improved biclustering of microarray data demonstrated through systematic performance tests, Comput. Statist. Data Anal. 48 (2005), pp. 235–254.

[34] J.H. Ward, Hierarchical groupings to optimize an objective function, J. Amer. Statist. Assoc. 58 (1963), pp. 234–244.

[35] J. Zhang, A Bayesian model for biclustering with applications, J. R. Stat. Soc. B 59 (2010), pp. 635–656.