Bayesian Analysis (2022) **17**, Number 3, pp. 969–995

Biclustering via Semiparametric Bayesian Inference[†*]

Alejandro Murua[‡] and Fernando Andrés Quintana[§]

Abstract. Motivated by classes of problems frequently found in the analysis of gene expression data, we propose a semiparametric Bayesian model to detect biclusters, that is, subsets of individuals sharing similar patterns over a set of conditions. Our approach is based on the well-known plaid model by Lazzeroni and Owen (2002). By assuming a truncated stick-breaking prior we also find the number of biclusters present in the data as part of the inference. Evidence from a simulation study shows that the model is capable of correctly detecting biclusters and performs well compared to some competing approaches. The flexibility of the proposed prior is demonstrated with applications to the analysis of gene expression data (continuous responses) and histone modifications data (count responses).

Keywords: clustering, CAR model, stick-breaking prior, gene expression.

1 Introduction

Assume we record measurements $\{y_{ij}\}$ corresponding to a sample of $i = 1, \ldots, n$ individuals (e.g. genes) on each of $j = 1, \ldots, J$ conditions. The data structure can thus be summarized in matrix form. Assume also the main interest is in identifying subsets of individuals sharing consistent patterns over a subset of conditions. In other words, we are concerned with the detection of biclusters (Pontes et al., 2015). Such problems arise frequently in the analysis of gene expression data (Tanay et al., 2002).

Several methods have been proposed in the literature to tackle the problem of finding biclusters in data coming in this matrix format. Biclustering was first discussed by Hartigan (1972) in the context of creating a method for simultaneously grouping rows and columns of a matrix (coining for this the term *direct clustering*). A generic form of approaching this problem consists of clustering analysis, in which individuals and conditions are each separately partitioned and the combined subsets later assessed. This idea could be criticized because of the potentially large number of clusters being created, including possibly meaningless subsets, and also because no overlapping is

[*]AM was partially funded by a Discovery grant 2019-05444 from the Natural Sciences and Engineering Research Council of Canada, and by a Fundamental Research Projects grant PRF-2017-20 from the Institute for Data Valorization (IVADO) of Quebec, Canada. FQ was partially supported by grant FONDECYT 1180034 and by ANID - Millennium Science Initiative Program - NCN17_059.

[†]We thank Peter Müller and Yanxun Xu for facilitating us the locations used in their analysis of the HM dataset discussed in their work and reconsidered here in Section 4.2.

[‡]Département de Mathématiques et de Statistique, Université de Montréal, alejandro.murua@umontreal.ca

[§]Departamento de Estadística, Facultad de Matemáticas, Pontificia Universidad Católica de Chile, and ANID - Millennium Science Initiative Program - Millennium Nucleus Center for the Discovery of Structures in Complex Data, quintana@mat.uc.cl

 https://doi.org/10.1214/21-BA1284

allowed among selected subsets. Nevertheless, this has been traditionally a common way to tackle the problem of biclustering. Approaches based on this idea can be found in Getz et al. (2000) and in Tang and Zhang (2005). The case of latent block models (LBM, see, e.g. Govaert and Nadif, 2014) is one such instance where independent partitions are considered for rows and columns of the data matrix. These partitions are inferred from the data in a model-based fashion. For a review of computational approaches to LBM see also Bouveyron et al. (2019). A different approach was considered by Cheng and Church (2000), who proposed an algorithm based on identifying submatrices with similar entries, as measured by a mean squared residue. Tanay et al. (2002) proposed *SAMBA* (Statistical-algorithmic method for bicluster analysis), which is based on viewing the data structure as a bipartite graph with edge weights assigned according to a certain probability model in such a way that heavy sub-graphs coincide with biclusters with high probability. The Factor analysis for bicluster acquisition (FABIA), developed by Kasim et al. (2010) considers instead a class of multiplicative models that exploits a sparse factorization of the data matrix that allows for heavy-tailed data, paired with a model selection approach to detect biclusters under a Laplacian prior to enforce sparsity. For a recent review on these and other bicluster methods for gene expression data, see Pontes et al. (2015), and for the case of biclustering discrete multivariate data, see Fernández et al. (2019).

Other model-based approaches devised for detection of biclusters are available in the literature. One traditional alternative (on which the approach to be discussed later is based) is the plaid model (Lazzeroni and Owen, 2002). In the plaid model, the expected value for each y_{ij} is computed by summing contributions from each bicluster, and biclusters need not be disjoint, i.e. they may have nonempty overlapping. An improvement of their original algorithm for finding biclusters is described in Turner et al. (2005a). Extensions of this approach were considered, with different forms of prior construction, in Zhang (2010), Chekouo and Murua (2015a), and Chekouo et al. (2015). Under a parametric prior framework for the plaid model, Caldas and Kaski (2008) discuss an efficient implementation using a collapsed Gibbs sampler. Gu and Liu (2008) present an approach that, based on identifiability considerations, restricts biclusters to overlap on either rows or columns. Many of these approaches involve specifying a fixed number of biclusters. A different approach that does not involve this restriction was discussed in Xu et al. (2013). They proposed a nonparametric Bayesian Poisson model for histone modifications (HMs) by means of a zero-enriched Pólya urn model (Sivaganesan et al., 2011) that allows for clustering of HMs and also accounting for the fact that many HMs are idle and play no role for clustering. They also allow for each subset of HMs to define their own clustering of genomic locations, thus constructing a nested structure of biclusters. We stress here that by construction, the nested biclustering structure does not allow for overlapping. Ni et al. (2020) proposed a model for feature allocation that can also produce overlapping biclusters of patient-disease and symptom-disease, but using a completely different approach than ours, based on matrix factorizations. Their model is in reality specifically designed for the case of categorical entries in the data matrix. Li et al. (2020) described a mixed effects model for periodontal data that takes into account the spatial configuration of teeth, using a non-overlapping bicluster construction that features repulsion to induce sparsity by way of a determinantal point

process. Ren et al. (2020) propose a method that identifies biclusters of patients sharing similar patterns over time, with applications to monitoring data on 24-hr ambulatory blood pressure. They construct the biclusters by means of a Dirichlet process prior at the level of patients and a baseline distribution with time change points. And recently, Zhou et al. (2021) consider Dirichlet-multinomial mixtures and matrix factorizations for the analysis of taxon abundance data.

In this work, we revisit the plaid model with a nonparametric prior that allows us to detect biclusters that may have nonempty intersections, while at the same time lifting the restriction of fixing a priori the number of biclusters. At the heart of the plaid model definition is the introduction of two binary matrices (details will be given later in Section 2) that indicate whether genes and conditions form part of a certain bicluster. These matrices also imply a *null* cluster that has the purpose of concentrating the background noise. Our hierarchical model construction specifies the prior distribution of these matrices by way of two separate stick-breaking processes on latent variables that define these binary quantities via thresholding. Furthermore, we model the collection of latent variables defining the binary matrix for genes by way of a conditional autoregressive specification. This allows us to introduce information on which genes are a priori more likely to be part of a given bicluster. A byproduct of the double stick-breaking construction is that we can infer the number of biclusters. In practice, we assume a certain maximum K of biclusters, which can be large enough to pose no practical restriction. At the same time, we introduce a penalization term that discourages the formation of very large biclusters that may otherwise be detected. Large biclusters are usually hard to interpret and, many times, meaningless. The prior construction is general and can be coupled with specific sampling models such as a Gaussian, unimodal or Poisson likelihood, depending on particular needs and/or data features. We illustrate and implement all these instances, and compare results with other competitor approaches that were designed for data structures matching our inferential target. The comparisons are carried out in the context of extensive simulations studies and also with datasets already analyzed elsewhere.

The main novelties of this paper can then be summarized as follows: (1) we free the traditional plaid model from the restriction of a pre-specified number of biclusters, while allowing for a wide range of possible sampling models to accommodate for various data formats; (2) we define the binary indicator matrices by way of a novel approach based on thresholding a double stick-breaking prior that allows us to provide inference on the number and conformation (i.e. genes and conditions) of possibly overlapping biclusters; and (3) we introduce a penalty prior that controls the size of biclusters.

This paper is organized as follows. Section 2 describes the proposed hierarchical model and the corresponding prior construction. The model has several components, and we describe each of these, discussing their role and impact on the resulting inference. A detailed posterior simulation strategy is described in an accompanying Supplementary Materials (Murua and Quintana, 2021) file, and for convenience, a summary of some of its less standard aspects is presented in Section 2.3. A simulation study aimed at evaluating and comparing model performance is described in Section 3. Data illustrations for both continuous and count outcomes are described in Section 4. The article concludes with a discussion in Section 5.

2 The modeling approach

The proposed model is of hierarchical type and builds on the plaid model. As discussed below, the sampling model adopts either an additive or multiplicative form depending on the type of data available. The prior involves a double stick-breaking construction that is used to overcome the restrictions posed by fixing a priori de number of biclusters. In what follows, we specify all of the model components, discussing their role for the desired inference problem, namely, uncovering meaningful biclusters.

2.1 Sampling model

We start our discussion with the definition of the mean response, and then proceed to the specification of sampling distribution. We deal here mainly with the continuous and count response cases, which are the most frequently found in practical applications from fields such as genetics.

Plaid model for signals in sampling model

Let y_{ij} denote the score collected for individual (say, gene) $1 \leq i \leq n$ and condition $1 \leq j \leq J$. The data structure is then of matrix form, with individuals typically represented as rows and conditions as columns. We consider below the case of continuous and count scores, which may naturally arise in various applications. Our approach for the detection of biclusters, i.e., subsets of individuals exhibiting similar behavior across a subset of conditions or columns, is based on the plaid model introduced by Lazzeroni and Owen (2002). In the continuous response case, this model expresses each y_{ij} as the sum of signals coming from potentially several biclusters plus a noise term. To explain the main idea, we start by introducing the bicluster indicators. Let $\boldsymbol{\rho} = \{\rho_{i\ell}\}$ and $\boldsymbol{\kappa} = \{\kappa_{j\ell}\}$ be matrices with binary entries defining the biclusters in the following way: gene i and condition j are part of bicluster $1 \leq \ell \leq K$ provided that $\rho_{i\ell} = \kappa_{j\ell} = 1$. Note that a maximum number K of possible biclusters is assumed (this number can be assumed to be large enough to pose no practical restriction). Also define $\gamma_{ij} = \prod_{\ell=1}^{K}(1 - \rho_{i\ell}\kappa_{j\ell})$, and note that $\gamma_{ij} = 1$ only when gene i and condition j are not part of any bicluster. We refer to this cluster as the *null cluster*, representing what we term as the noise present in the data. The null cluster was first introduced by Chekouo and Murua (2015a), and was also used in Chekouo et al. (2015). Xu et al. (2013) also used a similar null cluster, but they defined it as a collection of subclusters nested inside larger biclusters. The plaid model assumes a sampling model where responses are conditionally independent given the bicluster matrices $\boldsymbol{\rho}$ and $\boldsymbol{\kappa}$. In our work, depending on the nature of the data, the plaid model is either an additive or a multiplicative model, the latter case arising as an additive specification in the log-scale. Details follow.

Sampling model for continuous responses: In the case of continuous data, the mean response under the plaid model, $\mu_{ij} \doteq E(y_{ij})$, is assumed to be additive

in the row and column effects. Specifically, it is expressed as

$$\mu_{ij} = \sum_{\ell=1}^{K} (\mu_\ell + \alpha_{i\ell} + \beta_{j\ell}) \rho_{i\ell} \kappa_{j\ell} + \gamma_{ij} \mu_N, \quad i = 1, \ldots, n, \quad j = 1, \ldots, J. \tag{2.1}$$

Sampling model for count responses: In the case of count responses (e.g., Poisson data), the mean response in the plaid model is assumed to have multiplicative row and column effects (or equivalently, additive in a logarithmic scale). It is expressed as

$$\mu_{ij} = \left[\prod_{\ell=1}^{K} \exp\{(\mu_\ell + \alpha_{i\ell} + \beta_{j\ell}) \rho_{i\ell} \kappa_{j\ell}\} \right] \exp\{\gamma_{ij} \mu_N\}, \quad i = 1, \ldots, n, \quad j = 1, \ldots, J. \tag{2.2}$$

The rightmost term in (2.1) or (2.2) represents the background noise or null cluster, i.e., everything that does not belong to a bicluster. Under either case, the sampling level parameters $\alpha_{i1}, \ldots, \alpha_{iK}$ and $\beta_{j1}, \ldots, \beta_{jK}$ are not identified and need to be constrained. A standard procedure for doing so consists of imposing the restriction $\sum_{i=1}^{n} \alpha_{i\ell} \rho_{i\ell} = \sum_{j=1}^{J} \beta_{j\ell} \kappa_{j\ell} = 0$ for all $1 \le \ell \le K$. This constraint can be easily implemented with a change of variables. Concretely, write $\boldsymbol{\alpha}_\ell = (\alpha_{1\ell}, \alpha_{2\ell}, \ldots, \alpha_{n\ell})$, and denote the number of rows in bicluster ℓ as $r_\ell = \sum_{i=1}^{n} \rho_{i\ell}$. The constraint on $\boldsymbol{\alpha}_\ell$ can be expressed as $\boldsymbol{\alpha}_\ell = \boldsymbol{V}_\ell \boldsymbol{a}_\ell$, for an n-dimensional vector $\boldsymbol{a}_\ell$, and an $n \times n$ matrix $\boldsymbol{V}_\ell$ whose components are given by $(\boldsymbol{V}_\ell)_{ii} = \rho_{i\ell}(1 - (1/r_\ell))$, and $(\boldsymbol{V}_\ell)_{ij} = -\rho_{i\ell}/r_\ell$ if $i \neq j$. In other words,

$$\alpha_{i\ell} = \rho_{i\ell} a_{i\ell} - \frac{1}{r_\ell} \sum_{k=1}^{n} \rho_{k\ell} a_{k\ell} = \rho_{i\ell} a_{i\ell} - \frac{1}{r_\ell} \boldsymbol{\rho}_\ell' \boldsymbol{a}_\ell,$$

where $\boldsymbol{\rho}_\ell = (\rho_{1\ell}, \ldots, \rho_{n\ell})$. It is easily seen that with these transformations, $\sum_{i=1}^{n} \alpha_{i\ell} \rho_{i\ell} = \boldsymbol{\rho}_\ell' \boldsymbol{\alpha}_\ell = 0$, for all ℓ. So, instead of working with $\boldsymbol{\alpha}_\ell$, we can work with $\boldsymbol{a}_\ell$. In a similar manner, we can replace the vector $\boldsymbol{\beta}_\ell$ with a vector $\boldsymbol{b}_\ell$, so that $\boldsymbol{\beta}_\ell = \boldsymbol{W}_\ell \boldsymbol{b}_\ell$, where $(\boldsymbol{W}_\ell)_{ii} = \kappa_{i\ell}(1 - (1/c_\ell))$, and $(\boldsymbol{W}_\ell)_{ij} = -\kappa_{i\ell}/c_\ell$ if $i \neq j$, and where $c_\ell = \sum_{j=1}^{J} \kappa_{j\ell}$ is the number of columns in bicluster ℓ. Prior choices for these parameters are discussed below.

Error distribution for sampling model

So far we have defined the mean of the signals through (2.1) or (2.2), depending on the data type under consideration. To properly define a sampling model, we consider two versions of likelihood function: Gaussian, and Poisson. The first is appropriate for signals on a continuous scale, while the second is suitable for count data such as in the histone modification (HMs) example discussed in Section 4. In defining the sampling model we recall that the distribution for observed responses should combine the signals from potentially overlapping biclusters and also those coming from the background noise. The likelihood models we use are thus:

(a) Gaussian likelihood model: In this case we combine the Gaussian signals with three different specifications for the noise term that we have found useful in practice. These choices result in the sampling models $p(y_{ij}|\boldsymbol{\rho},\boldsymbol{\kappa})$ that are specified in Table 1 below:

Noise	$p(y_{ij}\|\boldsymbol{\rho},\boldsymbol{\kappa})$
Gaussian	$\exp\left\{-\frac{1}{2\sigma_e^2}(y_{ij}-\mu_{ij})^2(1-\gamma_{ij})-\frac{1}{2\sigma_{0e}^2}(y_{ij}-\mu_N)^2\gamma_{ij}\right\}$
Uniform	$\exp\left\{-\frac{1}{2\sigma_e^2}(y_{ij}-\mu_{ij})^2(1-\gamma_{ij})-\gamma_{ij}\log(2\xi_N)\mathbf{1}_{\{y_{ij}\in[\mu_N-\xi_N,\mu_N+\xi_N]\}}\right\}$
Unimodal	$\exp\Big\{-\frac{1}{2\sigma_e^2}(y_{ij}-\mu_{ij})^2(1-\gamma_{ij})$ $+\left(\frac{1}{2}\log(\pi/[2\sigma_{0e}^2])+\log\big(1-\Phi(\lvert\mu_N-y_{ij}\rvert/\sigma_{0e})\big)\right)\gamma_{ij}\Big\}$

Table 1: Likelihood models in the Gaussian case with different distributional assumptions for the background noise.

These choices are motivated to provide a sensitivity assessment to specific distributional assumptions for the background noise. The third choice consists of noise modeled as a unimodal distribution based on the Rayleigh distribution. More precisely, by employing the unimodal distributions representation derived by Khintchine (1938), and following the work of Paez and Walker (2018), we find that any unimodal density $f_Y(y)$ can be expressed as

$$f_Y(y)=\int(2s)^{-1}\mathbf{1}_{s\in(\mu-s,\mu+s)}f_S(s)f_\mu(\mu)\,dsd\mu.$$

Even though the Khintchine representation just described is general, we are mainly interested in assessing how departures from normality assumptions for the background noise component affect the likelihood model. We thus consider a parametric mixture specification in the above unimodal representation and leave the nonparametric aspects of the model for the definition of prior distribution of the binary matrices $\boldsymbol{\rho}$ and $\boldsymbol{\kappa}$, as detailed below. Under this view, we have found in practice the Rayleigh distribution to be a useful choice for mixing distribution. This is because by taking $f_S(s)=(s/\sigma_{0e}^2)\exp(-s^2/[2\sigma_{0e}^2])$ we get a closed-form expression:

$$\begin{aligned}
f_Y(y)&=\int(2\sigma_{0e}^2)^{-1}\mathbf{1}_{y\in(\mu-s,\mu+s)}\exp(-s^2/[2\sigma_{0e}^2])f_\mu(\mu)\,dsd\mu\\
&=\int f_\mu(\mu)d\mu\int_{|\mu-y|}^{+\infty}(2\sigma_{0e}^2)^{-1}\exp(-s^2/[2\sigma_{0e}^2])\,ds\\
&=\frac{\sqrt{2\pi}}{2\sigma_{0e}}\int\int_{|\mu-y|}^{+\infty}\frac{1}{\sqrt{2\pi}\sigma_{0e}}\exp\left(-\frac{s^2}{2\sigma_{0e}^2}\right)f_\mu(\mu)\,dsd\mu\\
&=\frac{\sqrt{2\pi}}{2\sigma_{0e}}\int\left\{1-\Phi\left(\frac{|\mu-y|}{\sigma_{0e}}\right)\right\}f_\mu(\mu)d\mu,
\end{aligned}$$

which simplifies to what is indicated in Table 1. The prior specification of this part of the model proceeds by assuming standard distributions for mean and variance parameters:

$$\{\mu_\ell : \ell \geq 1\} \overset{\text{iid}}{\sim} N(0, \sigma_\mu^2), \qquad \mu_N \sim N(0, \sigma_{\mu,N}^2) \tag{2.3}$$

and

$$\sigma_\alpha^{-2} \sim \text{Gamma}(a_\alpha, b_\alpha), \quad \sigma_\beta^{-2} \sim \text{Gamma}(a_\beta, b_\beta), \quad \sigma_\mu^{-2}, \sigma_{\mu,N}^{-2} \overset{\text{iid}}{\sim} \text{Gamma}(a_\mu, b_\mu), \tag{2.4}$$

$$\sigma_e^{-2} \sim \text{Gamma}(a_e, b_e). \tag{2.5}$$

In addition we assume

$$\begin{aligned} &\sigma_{0e}^{-2} \sim \text{Gamma}(a_{0e}, b_{0e}) \qquad \text{under Gaussian noise, or} \\ &\xi_N^{-2} \sim \text{Gamma}(a_{0e}, b_{0e}) \qquad \text{under uniform noise.} \end{aligned} \tag{2.6}$$

Finally the specification of this part of the model is completed by assuming that

$$\{a_{i\ell} : i \geq 1, \ell \geq 1\} \overset{\text{iid}}{\sim} N(0, \sigma_\alpha^2) \qquad \text{and} \qquad \{b_{j\ell} : j \geq 1, \ell \geq 1\} \overset{\text{iid}}{\sim} N(0, \sigma_\beta^2). \tag{2.7}$$

(b) Poisson likelihood model: In this case, we have found it useful to assume that a realization in the null cluster comes from a Beta negative binomial distribution with parameters (r, α_N, β_N). Under this assumption, the density of an observation y in the null cluster is given by

$$\begin{aligned} p_N(y|r, \alpha_N, \beta_N) &= \frac{\Gamma(r+y)}{\Gamma(y+1)\Gamma(r)} \frac{\Gamma(\alpha_N + r)\Gamma(\beta_N + y)}{\Gamma(\alpha_N + \beta_N + r + y)} \frac{\Gamma(\alpha_N + \beta_N)}{\Gamma(\alpha_N)\Gamma(\beta_N)} \\ &= \frac{B(\alpha_N + r,\, \beta_N + y)}{yB(r, y)B(\alpha_N,\, \beta_N)}, \end{aligned}$$

where $B(\cdot, \cdot)$ stands for the beta function. We suppose that β_N is fixed (for example, we can fix β_N to a small positive value). Therefore, the likelihood model for count data can be expressed as

$$\begin{aligned} p(y_{ij}|\boldsymbol{\rho}, \boldsymbol{\kappa}) = \exp\Big\{ &(1 - \gamma_{ij})[y_{ij} \log \mu_{ij} - \mu_{ij} - \log(y_{ij}!)] \\ &+ \gamma_{ij}\big[\log B(\alpha_N + r,\, \beta_N + y_{ij}) - \log\big(y_{ij} B(r, y_{ij}) B(\alpha_N,\, \beta_N)\big)\big] \Big\}. \end{aligned}$$

Unlike the previous Gaussian case, we find it convenient to assume the priors on the bicluster means $\{\mu_\ell\}$, and unconstrained row and column effect coefficients $\{a_{i\ell}\}$ and $\{b_{j\ell}\}$ to be given by i.i.d. log-gamma distributions, that is, $\pi(\mu_\ell) \propto \exp(\beta_G \mu_\ell - e^{\psi(\beta_G) + \mu_\ell})$,

$$\pi(a_{i\ell}) \propto \exp(\beta_G a_{i\ell} - e^{\psi(\beta_G) + a_{i\ell}}), \qquad \pi(b_{j\ell}) \propto \exp(\beta_G b_{j\ell} - e^{\psi(\beta_G) + b_{j\ell}}),$$

where β_G is a constant, and $\psi(\cdot)$ denotes the digamma function. Note that this choice gives a zero-mean prior to each of the $\{\mu_\ell\}$, $\{a_{i\ell}\}$ and $\{b_{j\ell}\}$ parameters inside the exponential expression in (2.2).

2.2 Hierarchical prior structure

An essential component of the plaid model is the pair of binary matrices indicating bicluster membership. Recall that gene i and condition j form part of bicluster $1 \leq \ell \leq K$ provided that $\rho_{i\ell} = \kappa_{j\ell} = 1$. The prior distribution we construct, which is explained in detail below, has three main distinct goals: (1) to allow for overlapping biclusters by adopting the plaid model, but without a pre-specified number of biclusters; (2) to express prior beliefs on how genes or rows interact with each other when forming biclusters; and (3) we use information on the expected proportion of data that will form the biclusters, without restricting the number of detected biclusters in any way. The third goal is specifically designed to help avoiding the detection of too many (potentially spurious or hard to interpret) biclusters. We describe all these elements next.

Prior for matrix $\boldsymbol{\kappa}$

We assume first a maximum number K of biclusters, taken to be large enough so that it does not restrict in any practical way the desired inference. Next, the probabilities of bicluster membership are defined by way of a stick-breaking prior (Ishwaran and James, 2001) truncated at K, which frees us from the restriction just described. The main idea here is to use an ordered set of latent thresholds (the ζ_ℓ's below), computed by deterministically transforming the probability mass (the t_ℓ's below) induced by a stick-breaking process. This way, elements are assigned to biclusters according to a thresholding process that employs a collection of latent scores (the $T_{j\ell}$'s below). Specifically, for the case of the column membership variables $\boldsymbol{\kappa}$ we assume a truncated process construction of the form

$$t_1 = U_1 \quad \text{and} \quad t_\ell = U_\ell \prod_{k=1}^{\ell-1} (1 - U_k) \quad \text{for} \quad \ell = 2, \ldots, K-1, \tag{2.8}$$

where

$$U_1, U_2, \ldots, U_{K-1} \overset{\text{iid}}{\sim} \text{Beta}(1, M_c) \quad \text{and} \quad M_c \sim \text{Gamma}(a_{mc}, b_{mc}).$$

Setting $U_K = 1$ we have the same weight distribution as a truncated version of the Dirichlet process (Sethuraman, 1994) conditional on M_c, and in particular, $P(\sum_{\ell=1}^{K} t_\ell = 1 \mid M_c) = 1$. Based on the previously constructed stick-breaking process we specify the latent thresholds as $\zeta_1 = 0 = \Phi^{-1}(1/2)$, and

$$\zeta_\ell = \Phi^{-1}\big(\{1 + t_1 + \cdots + t_{\ell-1}\}/2\big), \qquad 2 \leq \ell \leq K-1, \tag{2.9}$$

where $\Phi^{-1}(\cdot)$ denote the inverse of the standard normal CDF. The binary matrix $\boldsymbol{\kappa}$ is then defined as

$$\kappa_{j\ell} = I\{T_{j\ell} > \zeta_\ell\}, \quad \text{where} \quad T_{j\ell} \overset{\text{iid}}{\sim} \text{Normal}(0, 1), \tag{2.10}$$

and $I\{A\}$ denotes the indicator function of the set A. This defines a categorical assignment of bicluster memberships with values ranging from 1 to K, and with a stochastic reduction in the underlying probabilities, to discourage the formation of small spurious biclusters.

Prior for matrix $\boldsymbol{\rho}$

For the gene selection matrix $\boldsymbol{\rho}$ we set up a prior that uses available information describing which genes are more likely to form part of the same group. This information is commonly available in biclustering for genetic problems. See the discussion below. We translate this information into a prior distribution in terms of a neighboring structure, which is a natural way of encoding known relations among genes. We construct the joint prior distribution for $\boldsymbol{\rho}$ by resorting to a sequence of latent multivariate normal random vectors that follow a conditionally autoregressive (CAR) specification, and another stick-breaking process similar to the one describing the prior on $\boldsymbol{\kappa}$. These elements are described next.

We start with the first column of $\boldsymbol{\rho}$, i.e. $(\rho_{11}, \ldots, \rho_{n1})'$. Let $\boldsymbol{Z}_1' = (Z_{11}, \ldots, Z_{n1})$ be a corresponding random vector of latent scores. We assume that

$$\rho_{i1} = I\{Z_{i1} > 0\}, \quad i = 1, \ldots, n, \tag{2.11}$$

which is a common way to define a set of binary outcomes in terms of a deterministic link involving the latent scores. See, e.g. Albert and Chib (1993). To build dependence in the joint distribution of the binary variables, we consider a multivariate normal distribution, expressed conditionally as

$$Z_{i1} \mid \boldsymbol{Z}_1^{-i} \sim N\left(\sum_{r=1}^{n} q_{ir} Z_{r1}, d_i^{-1}\right), \tag{2.12}$$

where $\boldsymbol{Z}_1^{-i}$ is $\boldsymbol{Z}_1$ with the ith component removed, $q_{i_1 r}$ is non-zero if and only if $i_1 \sim r$ and $r \neq i_1$, that is, if the distinct genes i_1 and r are *a priori* believed to belong to the same bicluster (more on this below), and d_i is a precision (reciprocal of the variance) parameter. Note then that $q_{ii} = 0$ by definition. Let $\boldsymbol{Q}$ be the $n \times n$ matrix with $Q_{ii} = d_i$ and $Q_{ir} = -d_i q_{ir}$. Besag (1974) proved that the CAR assumption (2.12) defines a valid joint multivariate normal distribution provided that $\boldsymbol{Q}$ is symmetric and positive definite, in which case we have that

$$\boldsymbol{Z}_1 \sim N_n\left(\boldsymbol{0}, \boldsymbol{Q}^{-1}\right). \tag{2.13}$$

An additional consideration is that (2.11) implies lack of identifiability under changes of scale of Z_{i1}. To overcome this deficiency, we choose $d_i = 1$ for all $1 \leq i \leq n$ in the definition of $\boldsymbol{Q}$. The relation given by the prior belief membership "$\sim$" induces a graph with vertices given by $1, \ldots, n$ and edges connecting i_1 and i_2 if and only if $i_1 \sim i_2$. The degree ν_i of any vertex i in the graph is given by its neighborhood size, that is, $\nu_i = \text{Card}(i_r : i \sim i_r)$. The matrix $A = (A_{ij})$ given by $A_{ij} = 1$ if $i \sim j, i \neq j$, and zero otherwise, is known as the adjacency matrix. The matrix $L = \text{diag}(\nu_1, \ldots, \nu_n) - A$ is the so-called Laplacian of the graph. When $d_i = 1$ for all $1 \leq i \leq n$, and $q_{ir} = \nu_i^{-1}$, the matrix $\boldsymbol{Q}$ coincides with the random-walk normalized Laplacian matrix. We work with a slight modification of this matrix by setting $q_{ir} = \lambda/\nu_i$, for $\lambda \in \mathbb{R}$. Thus, our model for $\boldsymbol{Q}$ corresponds to a generalized Laplacian. Moreover, we assume that all degrees ν_i are the same (that is, that all points have the same number of neighbors in the graph).

Our generalized Laplacian is diagonally dominant provided that $|\lambda| < 1$. This latter condition ensures the invertibility of $\boldsymbol{Q}$. Therefore, we assume a prior distribution

$$\lambda \sim \text{Beta}(a_\lambda, b_\lambda) \tag{2.14}$$

for this parameter that so defines the prior mean for the latent variables $\{Z_{1\ell}\}$.

The remaining columns of $\boldsymbol{\rho}$ are defined through additional latent scores. Denote these columns by $\boldsymbol{Z}_2, \boldsymbol{Z}_3, \ldots$. We assume these to be i.i.d with the same distribution as $\boldsymbol{Z}_1$, given in (2.13). Our prior construction aims at encouraging a change in the number of genes participating in subsequent biclusters, and that row sizes of these are stochastically sorted in decreasing order, with largest biclusters (in number of rows) appearing first. To this effect, we move the cutoff in (2.11) and consider a stick-breaking construction of the form

$$w_1 = V_1 \quad \text{and} \quad w_\ell = V_\ell \prod_{k=1}^{\ell-1}(1 - V_k), \quad \ell \geq 2, \tag{2.15}$$

where

$$V_1, V_2, \ldots \overset{\text{iid}}{\sim} \text{Beta}(1, M_r) \quad \text{and} \quad M_r \sim \text{Gamma}(a_{mr}, b_{mr}).$$

As in the case of the column variables $\boldsymbol{\kappa}$, this prior structure implies $P(\sum_{\ell=1}^{\infty} w_\ell = 1 \mid M_r) = 1$. With this, we assume the additional columns of $\boldsymbol{\rho}$ to still be defined via thresholding, but change (2.11) to

$$\rho_{i\ell} = I\{Z_{i\ell} > \theta_\ell\}, \quad 1 \leq i \leq n, \quad \ell \geq 1, \tag{2.16}$$

where $\theta_1 = 0 = \Phi^{-1}(1/2)$, and

$$\theta_\ell = \Phi^{-1}\left(\frac{1 + w_1 + \cdots + w_{\ell-1}}{2}\right), \quad \ell \geq 2. \tag{2.17}$$

Note that, by construction, $0 = \theta_1 < \theta_2 < \theta_3 < \cdots$ and $\theta_\ell \uparrow \infty$ as $\ell \to \infty$ with probability 1.

The above definition does not enforce a maximum number of biclusters as described earlier. To set this maximum number to K, we simply truncate the stick-breaking construction as was done when defining the prior for $\boldsymbol{\kappa}$. We achieve this by letting $V_K = 1$, so that $w_1 + \cdots + w_K = 1$ surely, and furthermore set $\theta_{K+1} = \infty$ in (2.16). Of course we can set a large enough value of K so as to represent no practical limitation in our capacity to detect biclusters.

The neighborhood structure of genes

There are several ways to incorporate prior beliefs in the gene graph edges. A simple solution consists of considering the distances between genes (e.g., correlation or Euclidean distances), and placing an edge between the k_{nn}-nearest-neighbors of every gene. A more involved distance corresponds to using Lin's pairwise similarities (Lin,

1998; Resnik, 1995) between genes obtained from gene ontologies (Ashburner et al., 2000). These latter distances were used in Chekouo et al. (2015). We adopt here a symmetrized version of the k_{nn} graph. In practice k_{nn} depends on the size of the graph. In our applications and simulations we used the Euclidean distances, and set k_{nn} to 30 as recommended by authors using similar graphs (Stanberry et al., 2008; Chekouo et al., 2015).

A penalty on the number of elements in biclusters

We have found it useful to incorporate in the model a prior belief on the proportion of data that actually forms biclusters under each of the likelihood models described earlier. This is done so as to avoid the formation of biclusters that are too large when there is little evidence that they are present in the data. In principle, the prior belief is passed as a geometric distribution, so that

$$p\Big(\sum_{\ell=1}^{K}\sum_{i=1}^{n}\sum_{j=1}^{J}\rho_{i\ell}\kappa_{j\ell}=m\Big)\propto\exp\{-\lambda_p m\}.$$

However, note that

$$\exp\Big\{-\lambda_p\sum_{\ell=1}^{K}\sum_{i=1}^{n}\sum_{j=1}^{J}\rho_{i\ell}\kappa_{j\ell}\Big\}=\prod_{i=1}^{n}\prod_{j=1}^{J}\exp\Big\{-\lambda_p\sum_{\ell=1}^{K}\rho_{i\ell}\kappa_{j\ell}\Big\}.$$

Thus, we may see the penalty as a product of nJ distributions

$$p_{ij}\Big(\sum_{\ell=1}^{K}\rho_{i\ell}\kappa_{j\ell}\Big)\propto\exp\Big\{-\lambda_p\sum_{\ell=1}^{K}\rho_{i\ell}\kappa_{j\ell}\Big\}.$$

Taking into consideration the normalizing constants of these probability mass functions, we get nJ identical Binomial$(K, e^{-\lambda_p}/(3+e^{-\lambda_p}))$ distributions

$$p_{ij}\Big(\sum_{\ell=1}^{K}\rho_{i\ell}\kappa_{j\ell}=m\Big)=\binom{K}{m}e^{-\lambda_p m}\,\frac{3^{K-m}}{(3+e^{-\lambda_p})^K}.$$

This result comes from the fact that for each pair (i,j)

$$\sum_{\rho,\kappa}\exp\Big\{-\lambda_p\sum_{\ell=1}^{K}\rho_{i\ell}\kappa_{j\ell}\Big\}=\sum_{m=0}^{K}\sum_{(\rho,\kappa)\in R_m}e^{-\lambda_p m},$$

where $R_m=\{(\rho,\kappa):\sum_{\ell=1}^{K}\rho_{i\ell}\kappa_{j\ell}=m\}$. The cardinalities of the sets R_m, for $m=0,\ldots,K$ are computed next. To have exactly m biclusters for which $\rho_{i\ell}\kappa_{j\ell}=1$, we need to have $\rho_{i\ell}=1$ and $\kappa_{j\ell}=1$ simultaneously exactly m times. So for each $\ell=1,\ldots,K$, $\rho_{i\ell}\kappa_{j\ell}=1$ only $2^{J-1}2^{n-1}$ times, and $\rho_{i\ell}\kappa_{j\ell}=0$, the remaining $(2^J2^n-2^{J-1}2^{n-1})=3\times2^{J-1}2^{n-1}$ times. So, $|R_m|=\binom{K}{m}3^{K-m}\big(2^{J-1}2^{n-1}\big)^K$. Therefore

$$\sum_{m=0}^{K}\sum_{(\rho,\kappa)\in R_m}e^{-\lambda_p m}=\big(2^{J-1}2^{n-1}\big)^K\sum_{m=0}^{K}\binom{K}{m}3^{K-m}e^{-\lambda_p m}=\big(2^{J-1}2^{n-1}\big)^K(3+e^{-\lambda_p})^K.$$

The desired penalty is now introduced by assuming independent identically distributed Binomial distributions for each point (i, j). With this, the prior on the number of observations in biclusters is specified through a Binomial distribution with parameters nJK, and $e^{-\lambda_p}/(3 + e^{-\lambda_p})$, that is

$$p\left(\sum_{\ell=1}^{K}\sum_{i=1}^{n}\sum_{j=1}^{p}\rho_{i\ell}\kappa_{j\ell} = m\right) = e^{-\lambda_p m}\binom{nJK}{m}3^{nJK-m}/(3 + e^{-\lambda_p})^{nJK}.$$

The mean and variance of this distribution depend on the maximum number of clusters K. To lessen this dependency, we replace λ_p by $\log K + \lambda_p$. This change results in the expected number of observations in biclusters equal to $nJe^{-\lambda_p}/(3 + e^{-\lambda_p}/K) \approx e^{-\lambda_p}nJ/3$, and a variance of $nJ3e^{-\lambda_p}/(3+e^{-\lambda_p}/K)^2 \approx e^{-\lambda_p}nJ/3$ for moderate to large values of K. This can also be derived from the Poisson approximation to the binomial, noting that

$$nJK \times \tfrac{e^{-\log K+\lambda_p}}{3+e^{-\log K+\lambda_p}} \overset{K\to\infty}{\longrightarrow} nJe^{-\lambda_p}/3,$$

which is independent of K. That is, by replacing λ_p by $\log K+\lambda_p$ we obtain a distribution on the number of observations in biclusters very close to a Poisson with mean $e^{-\lambda_p}nJ/3$. To choose a reasonable value for λ_p, we can set the mean of the Binomial distribution to a value close to the prior belief on the number of points that form clusters: $\hat{\lambda}_p = -\log(3EB/nJ)$, where EB is the number of observations that are expected to form part of biclusters. Alternatively, one can also set a prior on λ_p, for example, $\lambda_p \sim \text{Exp}(\theta_p)$, with $\theta_p^{-1} = \hat{\lambda}_p$. In our experiments, this prior works well.

2.3 Implementation

To implement inference for the models discussed earlier we use a hybrid MCMC posterior simulation scheme. A complete description of all the required steps is given in the Supplementary Materials file. This also includes the specific changes needed for each of the likelihood models adopted here. Nevertheless, we offer in this section a general discussion and summary of the main steps involved in the implementation.

The full model considers likelihood location-level parameters $\{\mu_\ell\}$, $\{a_{i\ell}\}$, $\{b_{j\ell}\}$, $\{\rho_{i\ell}\}$, $\{\kappa_{j\ell}\}$ and μ_N. In addition to these, the model involves various variance parameters (some of which change with the assumptions on background noise distribution), the CAR matrix parameter λ, the penalty parameter λ_p, and the stick-breaking parameters M_c and M_r. To begin this summary, recall that the bicluster row membership indicators $\{\rho_{i\ell}\}$ are by construction deterministic functions of $\{Z_{i\ell}\}$, and $\{w_\ell\}$, according to (2.16) and (2.17). Moreover, the $\{w_\ell\}$ parameters are also deterministic functions of the stick-breaking quantities $\{V_\ell\}$, as seen from (2.15). Similarly, the bicluster column membership indicators $\{\kappa_{j\ell}\}$ are deterministic functions of the $\{T_{j\ell}\}$ and $\{U_\ell\}$ parameters, as seen from (2.10), (2.8) and (2.9). Thus, in the sampler, the $\{\rho_{i\ell}\}$ and $\{\kappa_{j\ell}\}$ parameters are replaced by $\{Z_{i\ell}\}$, $\{V_\ell\}$, $\{T_{j\ell}\}$ and $\{U_\ell\}$. In the Gaussian case the full conditionals for mean and fixed effects are available in closed form, while in the Poisson case, Metropolis-Hastings (MH) moves are proposed. In addition, the full conditionals

for variance parameters in the Gaussian case also have closed-form expressions, and so is the case of the stick-breaking parameters M_r and M_c.

We discuss next some of the special moves required to sample from those parameters defining the binary matrices $\boldsymbol{\rho}$ and $\boldsymbol{\kappa}$, which as described earlier, are constructed via thresholding. For the row labels, $\rho_{i\ell} = I\{Z_{i\ell} > \theta_\ell\}$ together with (2.17) imply that the full conditional distribution of $Z_{i\ell}$ can be written as the mixture of two truncated normal distributions. The full conditionals of the stick breaking variables V_ℓ are more difficult to handle. Let $\eta_{i,\ell+1}$ denote the variable

$$\Big(2\Phi(Z_{i\,\ell+1}) - (1 + w_1 + \cdots + w_{\ell-1})\Big) \Big/ \prod_{k=1}^{\ell-1}(1 - V_k),$$

where $\Phi(\cdot)$ stands for the standard normal cdf. Set $\mu^{(0)}_{ij,\ell} = \sum_{k\neq\ell}^{K}(\mu_k + \alpha_{ik} + \beta_{jk})\rho_{ik}\kappa_{jk} + \gamma_{ij}\mu_0$, and $\mu^{(1)}_{ij,\ell} = \mu^{(0)}_{ij,\ell} + (\mu_\ell + \alpha_{i\ell} + \beta_{j\ell})\kappa_{j\ell}$, and define $g_0^{(ij,\ell)} = g\big(y_{ij}|\mu^{(0)}_{ij,\ell}, \text{else}\big)$, $g_1^{(ij,\ell)} = g\big(y_{ij}|\mu^{(1)}_{ij,\ell}, \text{else}\big)$, where $g(\cdot|\mu^{(r)}_{ij,\ell}, \text{else})$ stands for the likelihood of the model (i.e., either, Gaussian or Poisson), evaluated at $\mu_{ij,\ell} = \mu^{(r)}_{ij,\ell}$, $r = 0, 1$, and "else" refers to all other model parameters fixed at their currently imputed values. In Section A.4.1 of the Supplementary Materials file we show that

$$\begin{aligned} &P(V_\ell|V_{-\ell}, y, \boldsymbol{\Theta}) \\ &\quad\propto (1 - V_\ell)^{M-1} \prod_{j=1}^{J} \left(\prod_{i:\eta_{i,\ell+1}>V_\ell} g_1^{ij,\,\ell+1}\right)\left(\prod_{i:\eta_{i,\ell+1}\leq V_\ell} g_0^{ij,\,\ell+1}\right) \frac{e^{-[\lambda_p+\log(3)]m}\Gamma(nJK+1)}{\Gamma(m+1)\Gamma(nJK-m+1)}, \end{aligned} \tag{2.18}$$

with $V_{-\ell} = \{V_1, \ldots, V_{K-1}\} \setminus \{V_\ell\}$, where the rightmost term comes from the penalty prior, and $m = \sum_{i=1}^{n}\sum_{j=1}^{p}\sum_{\ell}^{K}\rho_{i\ell}\kappa_{j\ell}$ is the total size of the biclusters. Note that there are about 2^{nJ} possible combinations in the above product term (depending on the values of the $\boldsymbol{\kappa}$-labels). Therefore, it is more efficient to consider a MH sampling for the stick-breaking variables. For example, we could sample proposal updates from the associated Beta prior distribution, or consider transformations to near-normality of the stick-breaking variables followed by MH moves that mimic the suggestion of Roberts and Rosenthal (2009). More details are provided in Section A.4 of the Supplementary Materials file.

A similar type of full conditional is obtained for the latent $T_{j\ell}$ scores, related to the column labels. Here, we again obtain that the full conditionals are mixtures of truncated normal distributions and MH moves are considered when updating the stick-breaking variables U_ℓ (see Sections A.1 and A.4.2 in the Supplementary Materials file).

Finally, the full conditional distribution of λ implied by assumptions (2.12) through (2.14) has a complicated form, but we are able to derive a good approximation as

$$p(\lambda \mid \text{else}) \approx C\,\lambda^{a_\lambda - 1}(1-\lambda)^{b_\lambda + \frac{nK}{2\nu} - 1}\,(1 + \lambda/\nu)^{\frac{nK}{2}}\,\exp\big\{(\lambda/2\nu)\,\text{trace}\big(\boldsymbol{A}\hat{\boldsymbol{V}}\big)\big\},$$

where $\hat{\boldsymbol{V}} = \sum_{\ell=1}^{K} \boldsymbol{Z}_\ell \boldsymbol{Z}_\ell' / K$, C is the corresponding normalizing constant, which we evaluate numerically, ν is the common degree of the graph, and the matrix $\boldsymbol{A}$ is the adjacency matrix associated with the data graph (see Section 2.2).

3 Simulation study

We next describe a simulation study designed to evaluate the performance of the proposed model, and to compare it against some alternative methods designed to uncover biclusters in the same setting we have described. These methods are (1) the algorithm by Cheng and Church (2000); (2) the plaid method of Lazzeroni and Owen (2002), as improved by Turner et al. (2005a); and (3) the penalized biclustering method described in Chekouo and Murua (2015a). The first two methods are implemented in the `R` package `biclust` (Kaiser and Leisch, 2008), while the third has been implemented in `Java` and is publicly available on the web-sites of the authors (Chekouo and Murua, 2015b). While other methods are available, as described in Kasim et al. (2017), we have limited this comparison to some of the model-based approaches, discarding purely algorithmic alternatives.

Data were generated according to a number of different scenarios based on the number of biclusters $K \in \{2, 4, 8, 16\}$, the type of overlap between biclusters, either conditions-only overlap (C, or columns-only overlap), or conditions and genes overlap, (C&G, or row-and-columns overlap), and the type of noise, either normal (Gaussian), uniform, or unimodal noise. Three data sizes were considered. The smaller datasets generated consist of 355 rows and 17 columns, a choice that mimics the well-known yeast cycle dataset that recorded the gene expressions of yeast over seventeen different conditions (Cho et al., 1998; Mewes et al., 1999; Tavazoie et al., 1999; Yeung et al., 2001). Two large datasets were also considered. They consist respectively of 2000 rows and 38 columns, and 2000 rows and 76 columns. These mimic the retina detachment dataset GSE28133 found at NCBI/GEO (Edgar et al., 2002) and used by several researchers (Delyfer et al., 2011; Chekouo et al., 2015). The synthetic data were generated following the setup used by Chekouo et al. (2015). Sizes and positions of biclusters were visually generated so as to have sufficient but not extensive overlap. Some datasets are illustrated in Figure 1.

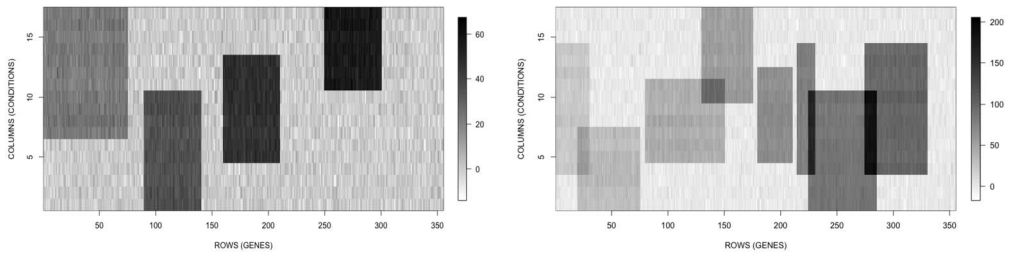


Figure 1: Some examples of datasets used in the comparison study. The data matrices depicted are transposed only for convenience of graphical illustration.

For each bicluster $k \in \{1, \ldots, K\}$, the overall bicluster mean was set to a draw from a normal distribution with mean $k + 1$ and unit variance. Row effects were generated as realizations of a multivariate normal with mean vector given by $\alpha_{ki} = 2/(1+\exp\{-i\}) - \bar{\alpha}_k$, where $\bar{\alpha}_k = \text{mean}\{2/(1+\exp\{-i\})\ i = 1, \ldots, |r_k|\}$ (recall that r_k and c_k denote, respectively, the number of rows and the number of columns in bicluster k). The column effects were generated similarly by replacing r_k by c_k. We devised scenarios corresponding to low and to high variance. The low variance scenario has the row and column effects variances as well as the bicluster mean variances set to 2.0. The error variance was drawn from a gamma distribution with parameters 2.0 and 5.0. The high variance scenario has the row and column effects variances as well as the bicluster mean variances set to 10.0. The error variance was drawn from a gamma distribution with parameters 1.5 and 15.0. For the larger datasets (with 76 columns), we also considered a moderate variance scenario. This has the row and column effects variances as well as the bicluster mean variances set to 6.0. The error variance was drawn from a gamma distribution with parameters 2.0 and 12.0. For the three scenarios, the noise parameters were set as follows:

Normal Noise: in this case the data were generated from a zero-mean normal distribution with variance defined as twice the error variance;

Uniform Noise: in this case each data point in the null cluster is generated as a draw from a uniform distribution in $(-5|u_1|, 5|u_2|)$, where for each data point, the pair (u_1, u_2) is a draw from a standard bivariate normal distribution;

Unimodal Noise: in this case each point is a draw from a Rayleigh-standard-Normal unimodal distribution with Rayleigh parameter $\sigma_{0e}^2 = 20$.

We considered five replications for each data scenario described above, except for scenarios associated with the larger dataset (76 columns) which were replicated according to the Graeco-Latin square design displayed in Table 2. Three factors were simultaneously varied: variance scenario {Low, Moderate (Mod), High}, noise distribution used to generate the data {normal, uniform, unimodal}, and number of true biclusters $\{4, 8, 16\}$. A fourth factor, the noise model assumed by our model, denoted here by *SbCAR* for the stick-breaking-CAR prior, was also varied. Because the larger datasets are considerably more computational demanding, this design was conceived so as to measure the performance of the different algorithms on several factors, but with less runs.

The biclustering results were evaluated using the F_1-measure (Santamaria et al., 2007; Turner et al., 2005b; Chekouo et al., 2015) of agreement between the true biclustering and the estimated biclustering yielded by the different methods that were compared. The F_1-measure is the harmonic mean between the so-called "recall" and "precision." For given biclusters B_1 and B_2, these are defined as follows: recall $= |B_1 \cap B_2|/|B_2|$, and precision $= |B_1 \cap B_2|/|B_1|$, so that $F_1(B_1, B_2) = 2(1/\,\text{recall} + 1/\,\text{precision})^{-1}$. Given two biclustering collections $\mathcal{A} = \{A_1, \ldots, A_G\}$ and $\mathcal{B} = \{B_1, \ldots, B_K\}$, we define for each $A_i \in \mathcal{A}$, $F_1(A_i, \mathcal{B}) = \max_{B \in \mathcal{B}} F_1(A_i, B)$, and similarly, for each $B_j \in \mathcal{B}$, $F_1(B_j, \mathcal{A}) = \max_{A \in \mathcal{A}} F_1(B_j, A)$. We measure the similarity between the two biclustering

Noise	Variance Low	Mod	High
normal	Biclusters = 4 SbCAR = normal	Biclusters = 8 SbCAR = unimodal	Biclusters = 16 SbCAR = uniform
unimodal	Biclusters = 8 SbCAR = uniform	Biclusters = 16 SbCAR = normal	Biclusters = 4 SbCAR = unimodal
uniform	Biclusters = 16 SbCAR = unimodal	Biclusters = 4 SbCAR = uniform	Biclusters = 8 SbCAR = normal

Table 2: Simulation designed to vary three factors simultaneously: Variance, Noise, and Number of true biclusters. The notation SbCAR = `xyz`, means that the model SbCAR was run assuming a noise distribution equal to `xyz`.

collections using a symmetrized version of the F_1-measure defined as

$$F_1(\mathcal{A}, \mathcal{B}) = \frac{1}{2}\left(\frac{1}{G}\sum_{i=1}^{G} F_1(A_i, \mathcal{B}) + \frac{1}{K}\sum_{j=1}^{K} F_1(B_j, \mathcal{A})\right).$$

Choosing the biclusters To choose the biclusters and their number we look at the posterior means of the labels $p(\rho_{i\ell} = 1 \mid y)$, $p(\kappa_{j\ell} = 1 \mid y)$, and also at $p(\rho_{i\ell}\kappa_{j\ell} = 1 \mid y)$. We set the ℓ-th bicluster to be the submatrix formed by the ensemble of rows i and columns j for which $p(\rho_{i\ell} = 1|y)$ and $p(\kappa_{j\ell} = 1 \mid y)$ are large enough. That is, if they are larger than a predetermined threshold such as 0.5, which is the threshold used in our simulations. The same procedure can also be done by only looking at $p(\rho_{i\ell}\kappa_{j\ell} = 1 \mid y)$. Our simulation results yielded the same or similar results with either choice of posterior means. Besides empty biclusters, we also discarded biclusters containing only a handful of cells (less than five).

Since, as discussed earlier, our normal-likelihood model allows three types of noise in the data, namely, normal, uniform and unimodal noise distributions, our model was run assuming normal, uniform and unimodal noise distributions on each dataset, regardless of the actual noise used to generate the data. For the cases $K \in \{2, 4, 8\}$, we also compare the results with our Poisson-likelihood model. For this, we applied the exponential transform to the data (that is, $x \mapsto \exp(x)$), and scaled the data so as to avoid numerical problems with the exponentiation. In our experiments, we transform a data point x to $y = \exp(\text{scale } x/\max_x)$, where $\max_x$ denotes the maximum over the absolute values of the data, and scale is chosen so that $\log(y) \leq 30$. The hyperparameters M_r and M_c were given Gamma priors with shape and rates equal to $(2, 20)$, and $(2, 100)$, respectively. The inverse-variance hyperparameters were set as follows: $a_\mu = a_\alpha = a_\beta = 2.1$, and $b_\mu = b_\alpha = b_\beta = 1.1$. Finally, a symmetrized 10-nearest-neighbor graph structure with Euclidean distance was employed when running these simulations.

Figures 2 and 3 display barplots with means and standard deviations of the F_1 measure comparing the true collections of biclusters to what is estimated by each of the methods studied, for the case of sixteen biclusters, and for the case of eight biclusters,

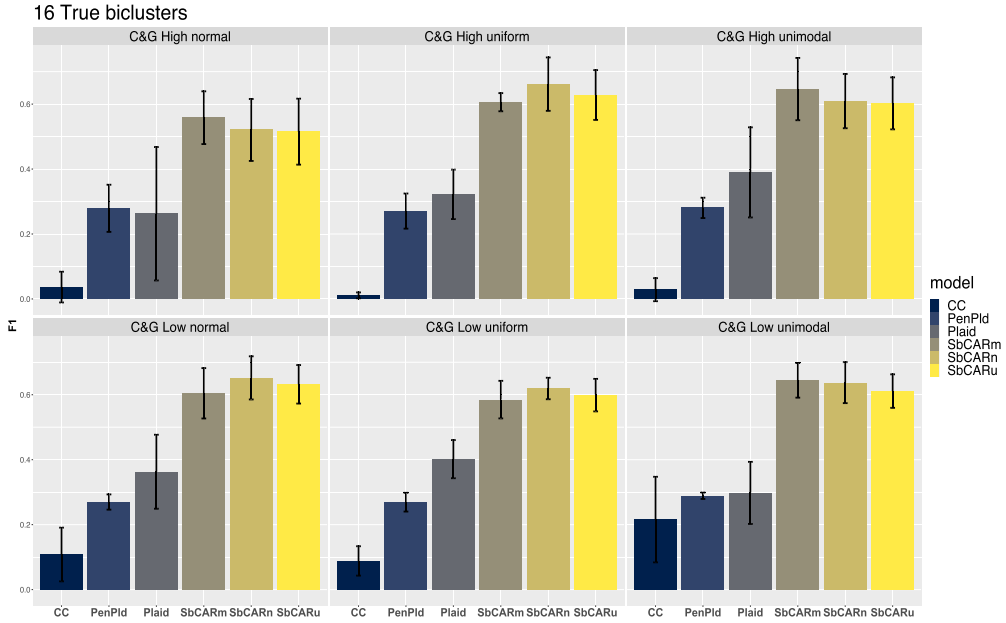


Figure 2: F_1 comparison results for the large datasets (with 36 conditions/columns). "Low" and "High" refer to the low and high variance scenarios, respectively. "C" and "C&G" refer to the conditions-only and conditions and genes overlap scenarios. The barplots titles "normal", "uniform" and "unimodal" refer to the type of noise used to generate the data. The methods displayed under the bars in the figure are, from left to right, Cheng & Church (CC), penalized-plaid (PenPld), plaid (Plaid), and the proposed normal-likelihood model with normal (SbCARn), uniform (SbCARu) and unimodal (SbCARm) noise fits (sampling models). Error bars represent variability over five repeated simulations; see the text for further explanation.

respectively (the results for two and four biclusters for the smaller datasets are similar and are shown in Section B of the Supplementary Materials file). Our proposed model is denoted here as *SbCAR* for the stick-breaking-CAR prior. In Figure 3 only the normal-noise fit (SbCARn) results are displayed. For the small datasets, the unimodal noise model's performance is very similar to that of the normal noise; the performance of the uniform noise model is also similar but not as good as in the case of the normal and unimodal noise models. This is in contrast with the performance of the uniform fit in the large datasets (see Figure 2).

Overall, the results for the small datasets show that the best performing methods are the proposed normal-likelihood model with either normal or unimodal (not shown) noise fit, the proposed Poisson-likelihood model, and the Penalized-Plaid model of Chekouo et al. (2015). However, for four biclusters (not shown here), the best performer is the Penalized-Plaid model. But, its performance is closely followed by those of the Poisson-likelihood and normal-likelihood models. For both type of datasets, our proposed models

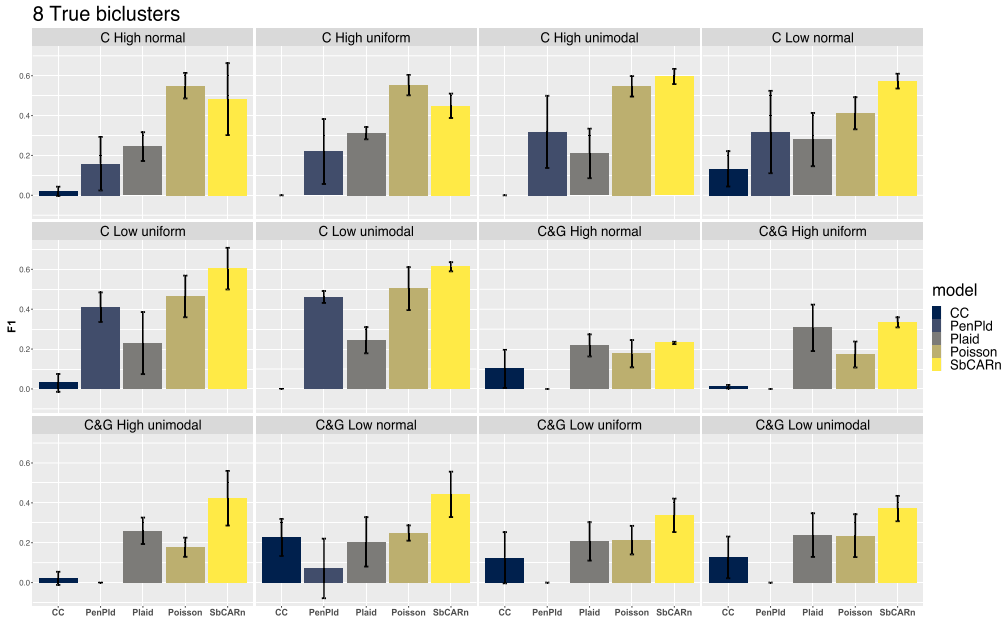


Figure 3: F_1 comparison results for the small datasets. "Low" and "High" refer to the low and high variance scenarios, respectively. "C" and "C&G" refer to the conditions-only and condition-and-genes overlap scenarios. The barplots titles "normal", "uniform" and "unimodal" refer to the type of noise used to generate the data. The methods displayed under the bars in the figure are, from left to right, Cheng & Church (CC), penalized-plaid (PenPld), plaid (Plaid), the proposed Poisson-likelihood model with negative-binomial noise (Poisson), and the proposed normal-likelihood model with normal noise fit (SbCARn). Error bars represent variability over five repeated simulations; see the text for further explanation.

are clearly the best performers when the number of biclusters is large (eight or sixteen). For the large datasets, the performance of our models is by far the best. The specific type of noise (i.e. sampling model) used when fitting our model does not appear to have much impact on inference and generally speaking, model performance. Nevertheless, the model with uniform noise has a slightly better performance, but this finding is shadowed by the typically longer computational time it requires to carry out posterior inference when compared to the other two alternatives we tried. In the accompanying Supplementary Materials file, Section C, we report on summary statistics obtained from some of the runs we have carried out. Taking all of this into account, our practical recommendation is to choose either the unimodal or normal sampling models.

The results from the simulation with the larger datasets (with 76 conditions/columns) are displayed in Table 3. All statistics were computed from three runs of each model as specified in the Graeco-Latin design of the simulation shown in Table 2. For this simulation we added the method BBC for Bayesian biclustering described in Gu

and Liu (2008). This method adopts a parametric prior for the plaid model, and based on identifiability considerations, imposes a certain restriction on biclusters, namely, that they can overlap either on rows or columns but not both. However, this method did not yield any results when the number of biclusters were set to 8 or to 16. Note that in all the runs of our model SbCAR, we have not imposed any constraints on the biclusters found except for the minimum number of cells which, as explained earlier, was set to five cells. Furthermore all competing methods were run by letting them know the true number of biclusters to be found. So in this simulation we have experimented with setting a minimum number of genes and conditions in each bicluster in the results of SbCAR. It seems reasonable to set the number of genes and the number of conditions forming a bicluster in proportion with their dimension in the data. So for the larger dataset with 76 conditions, we asked for 5% of genes and 10% of conditions as minimum sizes to form a bicluster. These translate to be a minimum of 100 genes and a minimum of 8 conditions to form an "interesting" bicluster. We remark here that these constraints are considered only as part of the posterior processing required to actually determine the biclusters, and are not at all related to other aspects of the method such as the definition of K or the penalty term discussed earlier. The results with these constraints are denoted by SbCARx in the corresponding table. The last row of Table 3 gives the overall performance of each method over all nine scenarios of Table 2. It is clear that SbCAR performs best overall, specially with the variant SbCARx.

		Model					
Factor	Levels	BBC	CC	PenPlaid	Plaid	SbCAR	SbCARx
Biclusters	4	0.632 (0.080)	0.019 (0.015)	0.788 (0.197)	0.707 (0.044)	0.586 (0.116)	0.810 (0.141)
	8	–	0.069 (0.110)	0.448 (0.021)	0.296 (0.154)	0.594 (0.034)	0.682 (0.038)
	16	–	0.024 (0.004)	0.270 (0.029)	0.224 (0.114)	0.395 (0.038)	0.445 (0.030)
Variance	Low	–	0.077 (0.103)	0.491 (0.268)	0.454 (0.269)	0.563 (0.141)	0.714 (0.232)
	Mod	–	0.015 (0.011)	0.444 (0.159)	0.450 (0.180)	0.548 (0.110)	0.641 (0.204)
	High	–	0.019 (0.017)	0.570 (0.372)	0.322 (0.347)	0.464 (0.116)	0.581 (0.126)
Noise	normal	–	0.014 (0.006)	0.504 (0.244)	0.402 (0.314)	0.525 (0.155)	0.686 (0.254)
	uniform	–	0.011 (0.010)	0.421 (0.182)	0.329 (0.288)	0.545 (0.129)	0.649 (0.172)
	unimodal	–	0.087 (0.095)	0.581 (0.369)	0.495 (0.200)	0.505 (0.110)	0.601 (0.159)
Overall	performance	0.211 (0.319)	0.037 (0.061)	0.502 (0.249)	0.409 (0.246)	0.525 (0.116)	0.645 (0.177)

Table 3: F_1 means and standard deviations (within parenthesis) for the large dataset (2000 rows and 76 columns) for the three different factors of the simulation. All statistics are taken from three runs of each model as specified in the Graeco-latin design for the simulation. The model BBC did not yield any results when the number of biclusters were set to 8 or 16, which is why the corresponding entries are empty. The results for SbCARx are obtained by filtering out biclusters with less than 100 rows or 8 columns. The results for the other models were obtained by asking the corresponding algorithms to find exactly 4, 8 or 16 biclusters, depending on the true number of biclusters in the data.

4 Data illustrations

We next consider applications in the case of continuous and count data, using the models developed earlier in Section 2. Recall that, as explained in Section 2.2 we adopt here the procedure of defining the graph structure in terms of Euclidean distances obtained from gene ontologies, specifically by resorting to the structure induced by the symmetrized 30-nearest-neighbor of every gene.

4.1 Yeast cell data

We consider data on gene expression for yeast cell expression data discussed and available in Eisen et al. (1998), and analyzed by several authors. See Chekouo and Murua (2015a) and references therein. The data are the result of combining information on yeast expression levels from microarray data coming from shock experiments and some previously published microarray data originating in time-courses from various cell processes, including the mitotic cell division cycle, diauxic shift, and sporulation. Specifically, we have the fluctuation of the log (base 2) expression levels of 2467 genes over 79 time-points (conditions). A small fraction of the data (about 1.9%) are missing. To make our results comparable with previous analyses of these data, we imputed the missing data following the procedure applied in Lazzeroni and Owen (2002) and Chekouo and Murua (2015a). This consists of replacing missing values by the sum of the corresponding row (gene) mean plus the column mean (time point) minus the overall mean. The main goal of the analysis is to find genes that exhibit consistent patterns over a subset of time-points. A close study of the noise cluster yielded by the penalized plaids model of Chekouo and Murua (2015a), hinted that a unimodal Rayleigh distribution such as the one describe in Section 2.1 would fit better the noise in the data than a normal distribution. For more details, see the Supplementary Materials.

Previous published work dealing with these data report a fair number of large biclusters with a large proportion of data forming part of at least one bicluster. Based on these studies and our observation on the cluster noise, we fit a unimodal-noise model with $K = 16$, and use a log-Normal proposal for the penalty parameter λ_p with mean 0.05 when executing the Metropolis step associated to running the MCMC procedure described in the Supplementary Materials file. The hyperparameters were set as in the simulation studies of Section 3.

We found 14 biclusters. In particular, at least three of the most important biclusters found by Lazzeroni and Owen (2002) and Chekouo and Murua (2015a) are also found by the proposed model, but one of the biclusters was in our case split into two. These are displayed in Figure 4, with genes shown in the rows and time points in the columns. Note that these biclusters overlap with each other, which implies that a model that imposes a nested structure would be inadequate to properly capture this behavior. The similarity between the conditions selected in these biclusters and those found in Chekouo and Murua (2015a) is $F_1 = 0.72$, while the similarity between the genes selected is $F_1 = 0.62$. The overall similarity is $F_1 = 0.48$. The drop in overall similarity is mostly due to the splitting of the Penalized-Plaid bicluster 3 into two biclusters, and the fact

that, in general, our model selected less genes in the biclusters. This latter fact surely entails better enrichment in the biclusters found by the proposed model.

4.2 Histone modifications data

We consider here a dataset on histone modifications (HMs), previously analyzed in Xu et al. (2013). Histones are proteins that package and order the DNA into structural units called nucleosomes. Histones have been found to play a role in gene regulation, and combinations of HMs have also been linked to cancer prognosis, and DNA repair, among other things. See the discussion in Xu et al. (2013). The data come from a ChiP-Seq experiment for CD4+ T lymphocytes (Barski et al., 2007; Wang et al., 2008) where 39 HMs are reported, with a total of 300 genomic locations including both, promoter and insular regions. The main goal of the application is to identify subsets of genomic locations with similar patterns of HMs.

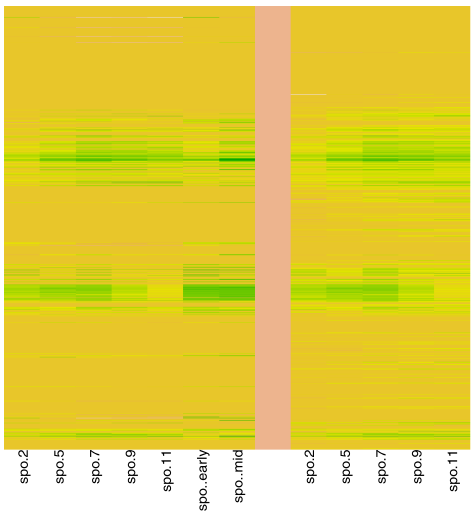
As in the paper of Xu et al. (2013), we reduced the data to a hundred genomic locations: fifty from promoter regions and 50 from insulator regions. These genomic locations are exactly the same ones selected by Xu et al. (2013). The data consist of counts of HMs over 100 genomic locations and 39 types of HMs. We applied our Poisson-likelihood model with row and column multiplicative effects and Negative Binomial noise to these ChIP-Seq data. We set $K = 16$, and use an exponential prior of intensity 10 for the penalty parameter λ_p. The hyperparameters were set as in the simulation studies of Section 3.

The model found three biclusters. These are displayed in Figure 5. They are very similar to those reported by Xu et al. (2013). In fact, the F_1 similarity measure between the two associated clusterings formed by the genomic locations is 0.61. Basically, bicluster 1 corresponds to active-HM set 1 of (Xu et al., 2013), bicluster 2 to active-HM set 3, and bicluster 3, to active-HM set 2. Thus, the conclusions from our analysis are very similar to the ones in (Xu et al., 2013). In particular, they corroborate the findings on the study of histone patterns in the human genome of Wang et al. (2008).

5 Discussion

We have proposed a model to detect biclusters from a sample in the form of a matrix of data, typically with subjects in the rows and conditions in the columns. Our proposal can be used with continuous or count responses and may be regarded as a nonparametric extension of the plaid model by Lazzeroni and Owen (2002). The hierarchical prior structure features a CAR model on a latent scale to incorporate prior information on which genes are more likely to form part of the same group, teamed with a stick-breaking prior for encouraging changes in the number of genes that constitute subsequent biclusters. A suitable MCMC posterior simulation procedure was devised to make inference under this model, particularly in what refers to detecting biclusters. Extensive simulation studies were also carried out to test the model and to compare its performance against other competitors. The results showed that our proposal performs well for our bicluster detection goals.

Biclusters UML1 (left) and PP1 (right)



Biclusters UML2 (left) and PP2 (right)

Biclusters UML3 (left), UML4 (center), and PP3 (right)

Figure 4: Yeast cell data: Three important biclusters found by the Unimodal-likelihood model (UML) and the Penalized-Plaid model (PP). Solid vertical bars are shown to separate biclusters.

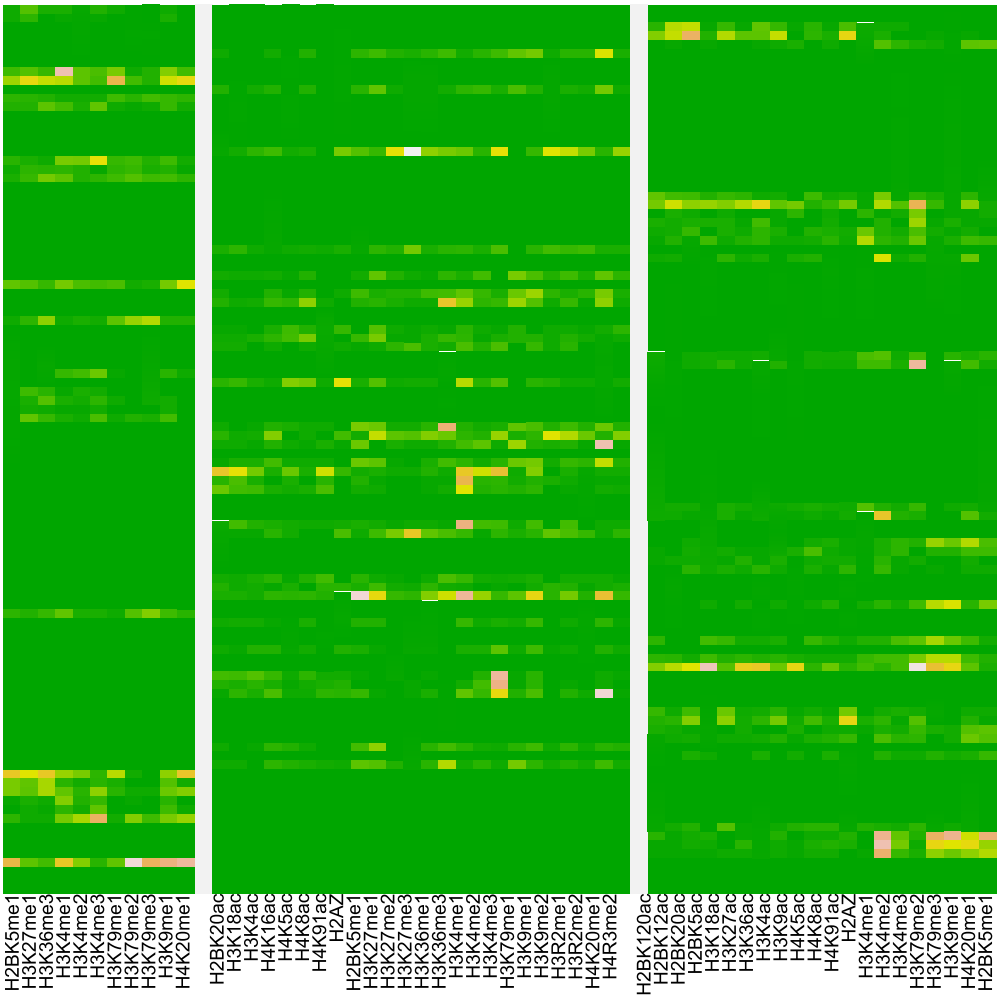


Figure 5: Histone Modifications Data: The three biclusters found by the Poisson-likelihood model with multiplicative effects for the ChIP-Seq data.

Our model included a particular definition of the columns in the gene selection matrix $\boldsymbol{\rho}$ based on a stick-breaking construction. An alternative way to define the joint prior model $p(\boldsymbol{\rho})$ could consider an Ising model with a neighboring structure constructed from prior information on gene interactions, just as in the approach described in Section 2. Developing further this idea is part of work to be carried out in the future.

Supplementary Material

Supplementary Materials to manuscript "Biclustering via Semiparametric Bayesian Inference" (DOI: 10.1214/21-BA1284SUPP; .pdf).

References

Albert, J. H. and Chib, S. (1993). "Bayesian analysis of binary and polychotomous response data." *Journal of the American Statistical Association*, 88(422): 669–679. MR1224394. 977

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., and Sherlock, G. (2000). "Gene ontology: tool for the unification of biology." *Nature Genetics*, 25: 25–29. 979

Barski, A., Cuddapah, S., Cui, K., Roh, T., Schones, D., Wang, Z., Wei, G., Chepelev, I., and Zhao, K. (2007). "High-resolution profiling of histone methylations in the human genome." *Cell*, 129(4): 823–837. 989

Besag, J. (1974). "Spatial interaction and the statistical analysis of lattice systems." *Journal of the Royal Statistical Society. Series B. Methodological*, 36: 192–236. With discussion by D. R. Cox, A. G. Hawkes, P. Clifford, P. Whittle, K. Ord, R. Mead, J. M. Hammersley, and M. S. Bartlett and with a reply by the author. MR0373208. 977

Bouveyron, C., Celeux, G., Murphy, T. B., and Raftery, A. E. (2019). *Model-based clustering and classification for data science*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge. With applications in R. MR3967046. doi: https://doi.org/10.1017/9781108644181. 970

Caldas, J. and Kaski, S. (2008). "Bayesian biclustering with the plaid model." In *2008 IEEE Workshop on Machine Learning for Signal Processing*, 291–296. 970

Chekouo, T. and Murua, A. (2015a). "The penalized biclustering model and related algorithms." *Journal of Applied Statistics*, 42(6): 1255–1277. MR3317943. doi: https://doi.org/10.1080/02664763.2014.999647. 970, 972, 982, 988

Chekouo, T. and Murua, A. (2015b). "The penalized biclustering plaid model." http://www.dms.umontreal.ca/~murua/software/penalizedplaid.zip. Software. 982

Chekouo, T., Murua, A., and Raffelsberger, W. (2015). "The Gibbs-plaid biclustering model." *The Annals of Applied Statistics*, 9(3): 1643–1670. MR3418739. doi: https://doi.org/10.1214/15-AOAS854. 970, 972, 979, 982, 983, 985

Cheng, Y. and Church, G. (2000). "Biclustering of expression data." In *Proceedings of the Eighth International Conference on Intelligent Systems for Molecular Biology*, volume 1, 93–103. 970, 982

Cho, R. J., Campbell, M. J., Winzeler, E. A., L., S., Conway, A., Wodicka, L., Wolfsberg,

T. G., Gabrielian, A. E., Landsman, D., Lockhart, D. J., and Davis, R. W. (1998). "A genome-wide transcriptional analysis of the mitolic cell cycle." *Molecular Cell*, 2(1): 65–73. 982

Delyfer, M. N., Raffelsberger, W., Mercier, D., Korobelnik, J. F., Gaudric, A., Charteris, D. G., Tadayoni, R., Metge, F., Caputo, G., Barale, P. O., Ripp, R., Muller, J. D., Poch, O., Sahel, J. A., and Léveillard, T. (2011). "Transcriptomic analysis of human retinal detachment reveals both inflammatory response and photoreceptor death." *PLoS One*, 6(12): e28791. 982

Edgar, R., Domrachev, M., and Lash, A. E. (2002). "Gene Expression Omnibus: NCBI gene expression and hybridization array data repository." *Nucleic Acids Research*, 30(1): 207–210. http://www.ncbi.nlm.nih.gov/geo. 982

Eisen, M., Spellman, P., Brown, P., and Botstein, D. (1998). "Cluster analysis and display of genome-wide expression patterns." *Genetics*, 95: 14863–14868. 988

Fernández, D., Arnold, R., Pledger, S., Liu, I., and Costilla, R. (2019). "Finite mixture biclustering of discrete type multivariate data." *Adv. Data Anal. Classif.*, 13(1): 117–143. MR3935193. doi: https://doi.org/10.1007/s11634-018-0324-3. 970

Getz, G., Levine, E., and Domany, E. (2000). "Coupled two-way clustering analysis of gene microarray data." *Proceedings of the National Academy of Sciences*, 97(22): 12079–12084. 970

Govaert, G. and Nadif, M. (2014). *Co-Clustering: Models, Algorithms and Applications*. Wiley-ISTE. 970

Gu, J. and Liu, J. S. (2008). "Bayesian biclustering of gene expression data." *BMC Genomics*, 9(Suppl I):S4(1): 1–10. 970, 986

Hartigan, J. A. (1972). "Direct clustering of a data matrix." *Journal of the American Statistical Association*, 67(337): 123–129. 969

Ishwaran, H. and James, L. F. (2001). "Gibbs sampling methods for stick-breaking priors." *Journal of the American Statistical Association*, 96(453): 161–173. MR1952729. doi: https://doi.org/10.1198/016214501750332758. 976

Kaiser, S. and Leisch, F. (2008). "A toolbox for bicluster analysis in R." In *COMPSTAT 2008—Proceedings in Computational Statistics*, 201–208, CD–ROM. Physica-Verlag/Springer, Heidelberg. MR2509605. 982

Kasim, A., Mayr, A., Mitterecker, A., Lin, D., Clevert, D.-A., Göhlmann, H. W. H., Bijnens, L., He usel, M., Hochreiter, S., Van Sanden, S., Khamiakova, T., Bodenhofer, U., Talloen, W., and Shkedy, Z. (2010). "FABIA: factor analysis for bicluster acquisition." *Bioinformatics*, 26(12): 1520–1527. 970

Kasim, A., Shkedy, Z., Kaiser, S., Hochreiter, S., and Talloen, W. (2017). *Applied Biclustering Methods for Big and High-Dimensional Data*. CRC Press, Boca Raton, FL. 982

Khintchine, A. (1938). "On unimodal distributions." *Izvestiya Nauchnolssledovatel'skoyo Instituta Matematiki i Mekka*, 2: 1. 974

Lazzeroni, L. and Owen, A. (2002). "Plaid models for gene expression data." *Statistica Sinica*, 12(1): 61–86. Special issue on bioinformatics. MR1894189. 969, 970, 972, 982, 988, 989

Li, Y., Bandyopadhyay, D., Xie, F., and Xu, Y. (2020). "BAREB: A Bayesian repulsive biclustering model for periodontal data." *Statistics in Medicine*, 39(16): 2139–2151. MR4108756. doi: https://doi.org/10.1002/sim.8536. 970

Lin, D. (1998). "An information-theoretic definition of similarity." In *Proceedings of the 15th International Conference on Machine Learning*, Morgan Kaufmann, San Francisco, CA, 296–304. 978

Mewes, H. W., Heumann, K., Kaps, A., Mayer, K., Pfeiffer, F., Stocker, S., and Frishman, D. (1999). "MIPS: a database for genomes and protein sequences." *Nucleic Acids Research*, 27(1): 44–48. 982

Murua, A., and Quintana, F. A. (2021). "Supplementary material for: Biclustering via Semiparametric Bayesian Inference." *Bayesian Analysis*. doi: https://doi.org/10.1214/21-BA1284SUPP. 971

Ni, Y., Müller, P., and Ji, Y. (2020). "Bayesian double feature allocation for phenotyping with electronic health records." *Journal of the American Statistical Association*, 115(532): 1620–1634. MR4189742. doi: https://doi.org/10.1080/01621459.2019.1686985. 970

Paez, M. S. and Walker, S. G. (2018). "Modeling with a large class of unimodal multivariate distributions." *Journal of Applied Statistics*, 45(10): 1823–1845. MR3811846. doi: https://doi.org/10.1080/02664763.2017.1396296. 974

Pontes, B., Giráldez, R., and Aguilar-Ruiz, J. S. (2015). "Biclustering on expression data: A review." *Journal of Biomedical Informatics*, 57: 163–180. 969, 970

Ren, Y., Sivaganesan, S., Altaye, M., Amin, R. S., and Szczesniak, R. D. (2020). "Biclustering of medical monitoring data using a nonparametric hierarchical Bayesian model." *Stat*, 9(1): e279. MR4104225. doi: https://doi.org/10.1002/sta4.279. 971

Resnik, P. (1995). "Using information content to evaluate semantic similarity in a taxonomy." In *Proceedings of the 14th International Joint Conference on Artificial Intelligence*, 448–453. 978

Roberts, G. O. and Rosenthal, J. S. (2009). "Examples of adaptive MCMC." *Journal of Computational and Graphical Statistics*, 18(2): 349–367. MR2749836. doi: https://doi.org/10.1198/jcgs.2009.06134. 981

Santamaria, R., Quintales, L., and Theron, R. (2007). "Methods to Bicluster Validation and Comparison in Microarray Data." *Springer Verlag Berlin Heidelberg*. 983

Sethuraman, J. (1994). "A constructive definition of Dirichlet priors." *Statistica Sinica*, 4(2): 639–650. MR1309433. 976

Sivaganesan, S., Laud, P. W., and Müller, P. (2011). "A Bayesian subgroup analy-

sis with a zero-enriched Polya urn scheme." *Statistics in Medicine*, 30(4): 312–323. MR2758864. doi: https://doi.org/10.1002/sim.4108. 970

Stanberry, L., Murua, A., and Cordes, D. (2008). "Functional connectivity mapping using the ferromagnetic Potts spin model." *Human Brain Mapping*, 422–440. 979

Tanay, A., Sharan, R., and Shamir, R. (2002). "Discovering statistically significant biclusters in gene expression data." *Bioinformatics*, 18(suppl 1): S136–S144. 969, 970

Tang, C. and Zhang, A. (2005). "Interrelated two-way clustering and its application on gene expression data." *International Journal on Artificial Intelligence Tools*, 14(04): 577–597. 970

Tavazoie, S., Hughes, J. D., Campbell, M. J., Cho, R. J., and Church, G. M. (1999). "Systematic determination of genetic network architecture." *Nature Genetics*, 22: 281–285. 982

Turner, H., Bailey, T., and Krzanowski, W. (2005a). "Improved biclustering of microarray data demonstrated through systematic performance tests." *Computational Statistics & Data Analysis*, 48(2): 235–254. MR2133586. doi: https://doi.org/10.1016/j.csda.2004.02.003. 970, 982

Turner, H., Bailey, T., and Krzanowski, W. (2005b). "Improved biclustering of microarray data demonstrated through systematic performance tests." *Computational Statistics and Data Analysis*, 48: 235–254. 983

Wang, Z., Zang, C., Rosenfeld, J., Schones, D., Barski, A., Cuddapah, S., Cui, K., Roh, T., Peng, W., Zhang, M., and Zhao, K. (2008). "Combinatorial patterns of histone acetylations and methylations in the human genome." *Nature Genetics*, 40(7): 897–903. 989

Xu, Y., Lee, J., Yuan, Y., Mitra, R., Liang, S., Müller, P., and Ji, Y. (2013). "Nonparametric Bayesian bi-clustering for next generation sequencing count data." *Bayesian Analysis*, 8(4): 759–780. MR3150468. doi: https://doi.org/10.1214/13-BA822. 970, 972, 989

Yeung, K. Y., Fraley, C., Murua, A., Raftery, A. E., and Ruzzo, W. L. (2001). "Model-based clustering and data transformations for gene expression data." *Bioinformatics*, 17(10): 977–987. 982

Zhang, J. (2010). "A Bayesian model for biclustering with applications." *Journal of the Royal Statistical Society. Series C. Applied Statistics*, 59(4): 635–656. MR2758627. doi: https://doi.org/10.1111/j.1467-9876.2010.00716.x. 970

Zhou, F., He, K., Li, Q., Chapkin, R. S., and Ni, Y. (2021). "Bayesian biclustering for microbial metagenomic sequencing data via multinomial matrix factorization." *Biostatistics*. Kxab002. URL https://doi.org/10.1093/biostatistics/kxab002. 971