Les Cahiers du GERAD ISSN: 0711–2440

Metabolic Data Learning: Forestogram Using Spike-and-Slab Models

V. Partovi Nia
M. Ghannad-Rezaie

G–2014–18

March 2014

Metabolic Data Learning: Forestogram Using Spike-and-Slab Models

Vahid Partovi Nia

GERAD & Department of Mathematical and Industrial Engineering
Polytechnique Montréal
Montréal (Québec) Canada, H3C 3A7

vahid.partovinia@polymtl.ca

Mostafa Ghannad-Rezaie

Department of Electrical Engineering
Massachusetts Institute of Technology
Cambridge, MA 02139-4307, U.S.A.

mrezaie@mit.edu

March 2014

Les Cahiers du GERAD

G–2014–18

Abstract: In many applications, such as metabolomics, data are composed of several continuous measurements of subjects (tissues) over multiple variables (metabolites). Measurement values are put in a matrix with subjects in rows and variables in columns. The analysis of such data requires grouping subjects and variables to provide a primitive guide toward data modelling. A common approach is to group subjects and variables separately, and construct a clustering tree once on rows and another time on columns. This simple approach provides a grouping visualization through two separate trees, which is difficult to interpret jointly. Another approach is to partition the matrix to provide a joint clustering, but this method looses the visualization tool being attractive for biologists. We propose a binary tree built on the matrix directly, thus providing a collection of three-dimensional trees that we call forestogram. We propose a hierarchical spike-and-slab model to provide a robust clustering in the presence of noise. Furthermore, we suggest an extension of the model that quantifies discriminant rows and columns. The log posterior is encouraged to be used as the similarity measure for comparing groupings and building the forestogram. As a consequence, the biclustering algorithm becomes fully automated. We apply our proposed method on real metabolomic measurements.

Key Words: Agglomerative clustering, Bayesian clustering, Dendrogram, Metabolic data, Spike-and-slab model.

Acknowledgments: This research is simulated by collaboration with the Swiss NCCR Plant Survival team in particular Prof. Samuel Zeeman and Dr. Gaëlle Messerli . We thank Professor Art Owen for his helpful comments on the early draft of the paper. We also appreciate the Department of Statistics of Stanford's hospitality. This work was supported by the Swiss SNF Fellowship PBELP2-125531S and the Natural Sciences and Engineering Research Council of Canada Grant 418034-2012.

1 Introduction

The scope of this paper is two fold. First, to advertise the advantages of incorporating spike-and-slab models in matabolomics. Second, to develop an agglomerative method for biclustering. The introduction section, therefore, is divided into three subsections. In Section 1.1 metabolic analysis is reviewed. Section 1.2 gives an overview on cluster analysis. We briefly introduce biclustering methods in Section 1.3.

1.1 Metabolic Analysis

Study of metabolism reveals deep underlying connections between the gene expression profile and the cell physiology (Fiehn et al., 2000). Furthermore, it provides an important tool for the study of metabolic disorders. Metabolic data are frequently collected using analytical chemistry methods, time-of-flight mass spectrometry (Vaidyanathan et al., 2001), infra-red spectrometry (Thomas et al., 2000), or gas chromatography-mass spectrometry (Gohlke and McLafferty, 1993), the latter being one of the low-cost tools. Gas chromatography mass spectrometry produce a continuous measurement of metabolic content of a tissue, usually tissues with different genetic backgrounds, within multiple samples. The information in the metabolic data could be regarded as a functional signature of physiological status of an organism, which is regulated through genetic background and environmental clues. Understanding such data can uncover the missing link between genotype and phenotype in presence of environmental factors. This understanding may help devising new generations of biomarkers to study biological disorders.

Metabolic datasets are relatively small compared to genetic datasets, but their analysis is still an important issue. Such studies help to reveal physiological fingerprint of a cell. Physiological patterns have a complex relationship with the genetic composition of cells. Studying such relationships still is an important challenge in cell biology with a lot of room for further methodological developments. A careful study of metabolite patterns helps to understand the metabolism pathway. These pathways clarify the interaction of certain metabolites with one or several genes. Consequently, a cell with similar genomic background may give different metabolic response, resulting the heterogeneity in data.

Despite the power of entirely graphical approaches such as looking at projected data on principal components axes (Yeung and Ruzzo, 2001), such methods disregard the assumption that a gene or a metabolite may play different roles. Furthermore, these methods in analysis of multivariate data, are concentrated on capturing the second order dependence between variables. Higher order dependency inherited in the non-linear nature of metabolic processes cannot be captured using these classical techniques.

Another highly interpretable class of methods applied to metabolic data is clustering. Clustering is a learning method whereby similar subjects are placed into disjoint groups on the basis of measurement of several variables. This technique is flexible to capture non-linear patterns (Redestig et al., 2007). A proper analysis of metabolic data should not require any prior information about the number of clusters since the prior condition and the significance of conditions to metabolism of the cell is unknown. Traditional clustering methods, such as hierarchical clustering (Everitt et al., 2011) and the k-means clustering (Hartigan and Wong, 1979), that have been used commonly in these analyses require the number of clusters to be known, in order to provide a proper grouping. Biologists prefer hierarchical clustering, since a visual tool is produced through a binary tree, called dendrogram. This tree helps applied researchers to decide what is the proper number of clusters. A certain grouping is produced by cutting this tree at a specific height. We keep this visualization tool in our new method. Moreover, our approach helps practitioners further, by providing a reference about an appropriate height to cut the tree. However, there is a price to pay while we develop a mathematical reference for the number of clusters. We suppose a statistical model for data and a prior distribution for cluster configurations. Throughout this section, we argue that a statistical model, e.g. a spike-and-slab model, works like a distance or a clustering linkage, also chosen implicitly by practitioners in algorithmic techniques. Algorithmic approaches may or may not produce the same fitting procedure, for a more detailed discussion about the two views in classification see Boulesteix and Schmid (2014).

Any grid partitioning of a data matrix corresponds to a joint grouping of rows and columns. Partitioning data vectors was focus of research for many years, but partitioning of a matrix, called biclustering, remained

unattended until recently. Biclustering has become a topic of more interest due to its modern applications in metabolomics, proteomics, and genetics (Zhang, 2010). Similar to clustering, most of the biclustering methods require a prior knowledge about the number of partitions (Lazzeroni and Owen, 2002). We relax this prior knowledge by taking a Bayesian approach and optimizing the resulting posterior over different number of partitions. The new method produces an easy-to-understand three-dimensional visualization of different matrix partitions, we call *forestogram*. This extended collection of the row and column trees provides a deeper insight into the clustering process. The classical row dendrogram and the column dendrogram can be extracted from the resulting forest by projection on rows and columns.

1.2 Clustering Analysis

Grouping, clustering, partitioning, or sometimes called unsupervised learning is a difficult problem, because search for an optimal data partitioning over the space of all possible groupings is awkward. This space is discrete, unordered, and have a large cardinality. Therefore, it is difficult to optimize any criterion over the grouping space. Optimality criterion for a grouping may have different definitions in different contexts. Briefly, optimality of a grouping is defined using a similarity measure or sometimes called linkage. It is supposed that data belonging to the same cluster are more similar, or have less distance from each other, on average. This approach is called average linkage in hierarchical clustering context. We propose a model-based view, a different approach, but still related to the linkage paradigm.

The basis for suggesting a model for clustering is the following. It appears that if groups with less average distance are merged, these clusters must have close measures of central tendency or equivalently less dispersion. More formally, suppose three singleton clusters each being a univariate measurement say y_1, y_2, and y_3. Merging y_1 with y_2 is meaningful if these two values are close to each other, i.e. $(y_1 - y_2)^2$ is the smallest distance among the others. This means $(y_1 - y_2)^2$ must be smaller than $(y_1 - y_3)^2$ and also smaller than $(y_2 - y_3)^2$. Equivalently, this means the variance in the merged cluster

$$\sum_{i=1}^{2}(y_i - \overline{y})^2 \tag{1}$$

must be the smallest, where $\overline{y} = \frac{1}{2}(y_1 + y_2)$, see Figure 1 (left panel) for the geometrical insight. When more than two clusters, each with different number of observations is available, the notation requires more indices. One way to generalize the latter variance (1), is to sum over all observations and clusters. Mathematically, denote the data in cluster I, by y_{Ii}, where I varies between 1 and total number of clusters, say k. If cluster I include n_I objects, the aim is to group the data in a way that the sum of within-cluster variances

$$\sum_{I=1}^{k}\sum_{i=1}^{n_I}(y_{Ii} - \overline{y}_I)^2 \tag{2}$$

is minimized, see Figure 1 (right panel).

Applied researchers feel more comfortable in using Euclidean distance instead of adopting a statistical model. However, as we describe in the following, a model-based view can produce the same clustering algorithm. Moreover, the model-based approach, helps us answering questions related to statistical issues. Issues like stability of an estimated grouping, estimating the number of unknown clusters, and ranking or selecting important clustering rows and columns. In the following we describe the equivalent model-based method to the average linkage clustering.

Suppose y_1 and y_2 are in the same cluster. Therefore, they follow the same probabilistic pattern, say independent Gaussian distribution with unit variance and mean θ. The joint density, then, is

$$f(y_1, y_2 \mid \theta) = f(y_1 \mid \theta) f(y_2 \mid \theta) \propto \exp\left\{-\frac{1}{2}\sum_{i=1}^{2}(y_i - \theta)^2\right\}. \tag{3}$$

When θ is unknown, the plug-in principle is applied and the maximum likelihood estimation of $\hat{\theta} = \overline{y} = \frac{1}{2}(y_1 + y_2)$ replaces θ. It is evident by comparing (3) with (1) that minimizing the within-cluster variance,

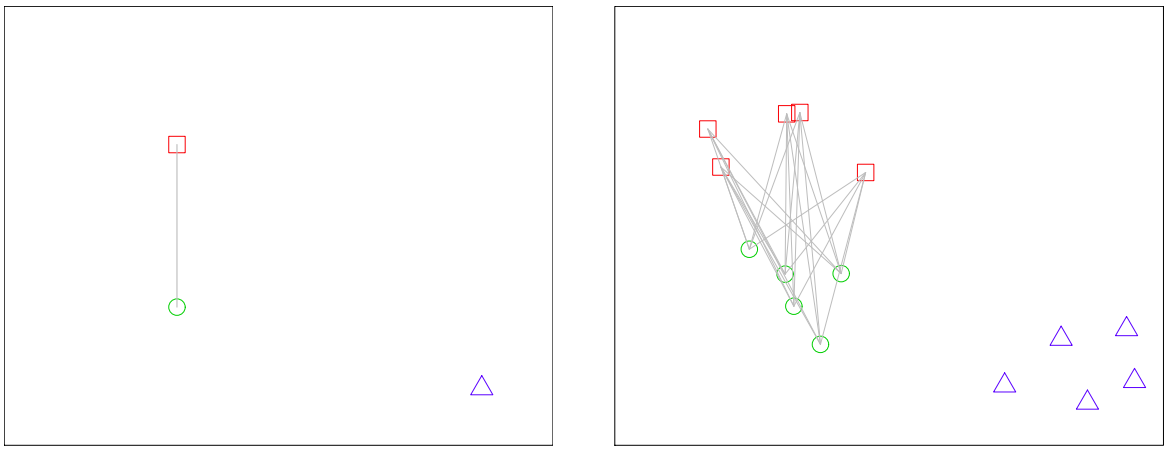

Figure 1: Three clusters of a bivariate data, each cluster is shown with a different symbol and color. The distance between a pair of clusters is denoted by a straight line. Left panel: three singleton clusters. Right panel: three clusters each of size five, visualizing the average linkage.

which is argued to be the average linkage clustering, actually mimics maximization of the joint data density under independent Gaussian assumption.

The generalization for more than one clusters is then straightforward. Assume data belong to the cluster I are independently distributed according to the Gaussian distribution with mean θ_I and unit variance. Suppose clusters are independent, then the joint data density is

$$f(\mathbf{y} \mid \boldsymbol{\theta}) \propto \exp\left\{-\frac{1}{2}\sum_{I=1}^{k}\sum_{i=1}^{n_I}(y_{Ii}-\theta_I)^2\right\}. \tag{4}$$

The double sum inside the exponent function is the within-variance cluster variance, except the maximum likelihood estimator $\hat{\theta}_I = \overline{y}_I = \frac{1}{n_I}\sum_{i=1}^{n_I} y_{Ii}$ replaces the unknown centre θ_I. It is trivial that maximizing the plugged-in version of (4) is equivalent to minimizing (2).

Unlike the frequentist approach which aims to find the grouping that maximizes the best possible scenario, i.e. the maximum likelihood plugged-in joint density, the Bayesian approach optimizes the averaged density. This averaging requires assuming a prior distribution for the unknown parameter θ_I, for instance a Gaussian distribution with mean μ and known variance σ_θ^2

$$\begin{aligned} y_{Ii} \mid \theta_I &\overset{\text{iid}}{\sim} \text{Gaussian}(\theta_I, 1) \\ \theta_I &\overset{\text{iid}}{\sim} \text{Gaussian}(\mu, \sigma_\theta^2). \end{aligned}$$

The objective, then, is to find a grouping that maximizes the averaged density

$$\begin{aligned} f(\mathbf{y};\mu,\sigma_\theta^2) &\propto \prod_{I=1}^{k}\int_{-\infty}^{\infty}\prod_{i=1}^{n_I}\exp\left\{-\frac{1}{2}(y_{Ii}-\theta_I)^2\right\}\exp\left\{-\frac{1}{2\sigma_\theta^2}(\theta_I-\mu)^2\right\}d\theta_I \\ &= \prod_{I=1}^{k}\int_{-\infty}^{\infty}\exp\left\{-\frac{1}{2}\sum_{i=1}^{n_I}(y_{Ii}-\theta_I)^2\right\}\exp\left\{-\frac{1}{2\sigma_\theta^2}(\theta_I-\mu)^2\right\}d\theta_I. \end{aligned} \tag{5}$$

A dendrogram which is the end-product of hierarchical clustering is a tree which organizes clusters. The traditional method for hierarchical clustering is a bottom-up (agglomerative) or a top-down (divisive) algorithm (Kaufman and Rousseeuw, 1990). Divisive methods start with all data in one cluster and consecutively divide clusters until ending with each observation as a singleton. Agglomerative methods start with each observation as a single cluster and successively merge the two closest clusters until one cluster containing

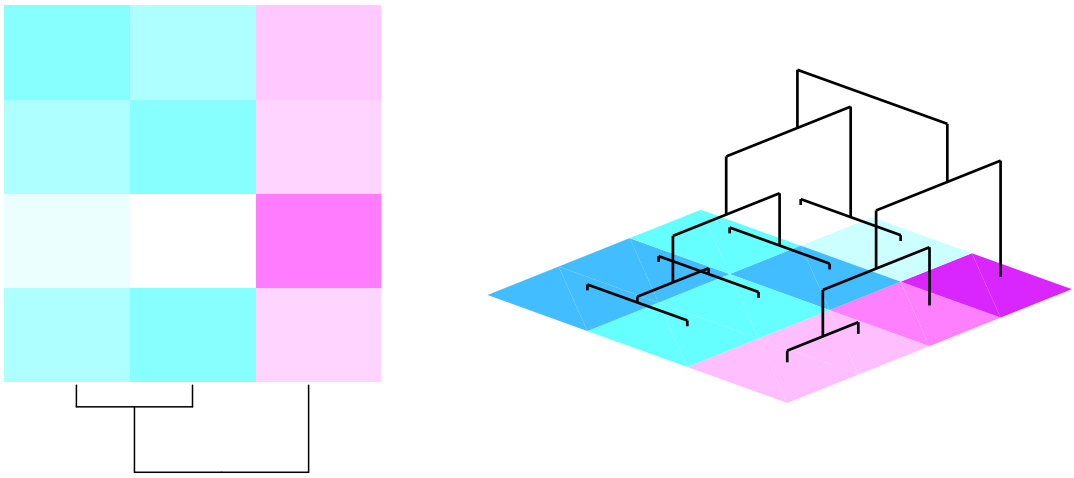

Figure 2: Dendrogram (left panel) is tree that organizes a collection of vectors. Forestogram (right panel) organizes a matrix and is composed of collection of row and column trees. A dendrogram can be extracted from a forestogram through projection on the row or the column side.

all observations is achieved. The nearest clusters are merged based on a given linkage also called similarity/dissimilarity measure. We therefore suggest to replace the linkage with the averaged density to produce a hierarchical Bayesian algorithm. This allows to take advantage of inferential machinery that supports a model-based view. We focus on developing an agglomerative partitioning method, but a divisive clustering algorithm can be constructed as well.

1.3 Biclustering

Biclustering or co-clustering refers to partitioning a matrix. Like clustering, biclustering methods are devided into distance-based, or model-based methods. Early biclustering was based on a distance, and proposed by Hartigan (1972). However, after three decades was applied in practice by Cheng and Church (2000) due to lack of computational power. Model-based techniques can be divided into two categories: (i) the frequentist approach where the statistical parameters of the model are treated as fixed unknowns like in (4) (Lazzeroni and Owen, 2002); and (ii) the Bayesian approach where a prior distribution is associated to the model parameters like in (5) (Gu and Liu, 2008; Zhang, 2010). Indeed most biclustering techniques apply model-based techniques nowadays, for a comprehensive review see Madeira and Oliveira (2004) and Tanay et al. (2005). In model-based biclustering, observations in each bicluster are supposed to be drawn independently from a parametric form (Sheng et al., 2003; Gan et al., 2008; van Uitert et al., 2008; Hochreiter et al., 2010).

The difficulty of biclustering methods, like clustering, is to optimize an objective function over all the possible submatrices. One way to reduce search space is to search only over the agglomerative path. A visual bi-product of agglomerative method over a matrix is a collection of binary trees, that we call forestogram, see Figure 2.

In Section 2 we introduce the notation and a spike-and-slab model for biclustering of continuous data. In Section 3 we discuss the agglomerative Bayesian biclustering and Bayesian forestogram. Section 4 applies the suggested method on replicated metabolomics data.

2 Spike-and-Slab Model

Let $\mathbf{Y}_{n\times p}$ denote the matrix entries, i.e. n subjects measured over p variables. Denote the data clustering by $\mathcal{C}$, a discrete random variable with probability mass function $f(\mathcal{C})$ and cardinality $||\mathcal{C}|| = k$, where $1 < k < np$

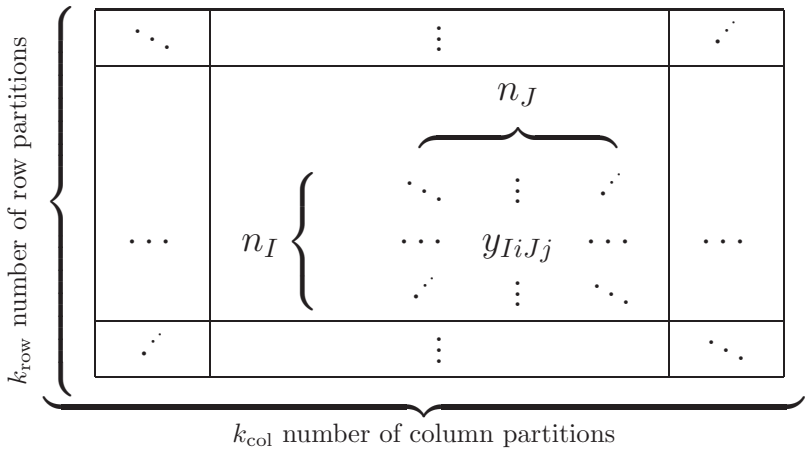

Figure 3: Visual illustration of grid partitioning of a matrix.

is the number of disjoint clusters. In Bayesian clustering, the data grouping is the parameter of interest to be estimated from data. Therefore, in this approach, a model is adopted for the data given a grouping, also called likelihood, and a prior distribution is assumed for grouping. This section focuses on presenting a useful model when a grouping is given, i.e. $f(\mathbf{Y} \mid \mathcal{C})$. The goal, after, is to optimize the grouping posterior $f(\mathcal{C} \mid \mathbf{Y}) \propto f(\mathbf{Y} \mid \mathcal{C}) f(\mathcal{C})$. This is equivalent to clustering np observations. Optimizing over any possible grouping of data entries ignores the matrix structure of data. We are particularly interested in non-overlapping grid clustering of the matrix $\mathbf{Y}$, mainly because grid clustering of a matrix is equivalent to mutual grouping of rows and columns. As a consequence of this grid restriction, the grouping $\mathcal{C}$ now can be decomposed into row-cluster $\mathcal{C}_{\text{row}}$ with $||\mathcal{C}_{\text{row}}|| = k_{\text{row}}$ and column-cluster $\mathcal{C}_{\text{col}}$ with $||\mathcal{C}_{\text{col}}|| = k_{\text{col}}$, evidently the number of biclusters $k = k_{\text{row}} k_{\text{col}}$. Suppose that the entries of $\mathbf{Y}$, y_{IiJj}, is univariate, where $I = 1, \ldots, k_{\text{row}}$ indexes row-clusters, $J = 1, \ldots, k_{\text{col}}$ indexes column-clusters, $i = 1, \ldots, n_I$ and $j = 1, \ldots, n_J$ denote the ith observation in row-cluster I and jth observation in column-cluster J, see Figure 3. If y_{IiJj} is a univariate random variable, data are called unreplicated, otherwise are called replicated. The integers n_I and n_J denote the number of observations in row-cluster I and column-cluster J, accordingly.

Hence, the total number of rows $n = \sum_{I=1}^{k_{\text{row}}} n_I$, and the total number of columns $p = \sum_{J=1}^{k_{\text{col}}} n_J$. If $\mathbf{y}_{IJ}$ is the data in bicluster IJ, the joint density can be written as $f(\mathbf{Y} \mid \mathcal{C}_{\text{row}}, \mathcal{C}_{\text{col}}) = \prod_{I=1}^{k_{\text{row}}} \prod_{J=1}^{k_{\text{col}}} f(\mathbf{y}_{IJ} \mid \mathcal{C}_{\text{row}}, \mathcal{C}_{\text{col}})$, in which

$$f(\mathbf{y}_{IJ} \mid \mathcal{C}_{\text{row}}, \mathcal{C}_{\text{col}}) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \prod_{i=1}^{n_I} \prod_{j=1}^{n_J} f(y_{IiJj} \mid \boldsymbol{\psi}, \mathcal{C}_{\text{row}}, \mathcal{C}_{\text{col}}) dF_{\boldsymbol{\psi} \mid \mathcal{C}_{\text{row}}, \mathcal{C}_{\text{col}}}, \tag{6}$$

for some real valued parameter vector $\boldsymbol{\psi}$. We suggest models with analytically tractable marginals (6) since its fast evaluation gives two important advantages. First, it is possible to estimate the hyperparameters of $F_{\boldsymbol{\psi} \mid \boldsymbol{\mathcal{C}}_{\mathbf{row}}, \boldsymbol{\mathcal{C}}_{\mathbf{col}}}$ through empirical Bayes. Second, we can compare different pair of groups efficiently and build a binary tree. If the marginal density (6) is intractable an analytical approximation can be used instead.

A very simple parametric model for biclustering continuous data is the Gaussian mean model

$$\begin{aligned} y_{IiJj} \mid \theta_{IJ} &\overset{\text{iid}}{\sim} \text{Gaussian}(\theta_{IJ}, \sigma^2), \\ \theta_{IJ} &\overset{\text{iid}}{\sim} \text{Gaussian}(\mu, \sigma_\theta^2), \\ &\theta_{IJ} \in \mathbb{R}, \qquad \sigma^2, \sigma_\theta^2 > 0. \end{aligned} \tag{7}$$

The marginal density (6) is tractable, taking $\psi = \theta_{IJ}$. The other hyperparameters σ^2, σ_θ^2, and μ we consider given at the moment, but we suggest to estimate them using empirical Bayes in practice. Here σ^2 is the variance of biclusters being equal for all submatrices, σ_θ^2 is the variance of the mean signal that separates such biclusters, and μ is the centre of data. In other words, μ reflects the data average, and the magnitude of

the signal to noise ratio σ_θ^2/σ^2 reflects the difficulty of the biclustering problem. One may suppose different bicluster variances through indexing σ^2, e.g. σ_{IJ}^2. In order to keep the analytical tractability of (6) we may assume that σ_{IJ}^2 is distributed according to the inverse Gamma law. Smith et al. (2008) show that clustering with varying cluster variances produce junk clusters using agglomerative method. Therefore, we focus on models with varying mean, but common variance σ^2.

Many high-dimensional applications involve a lot of noise variables. It is more convenient to assume that data are generated by a mixture of two distributions, one spike distribution that represents the noise and is concentrated around a constant, often zero, and another distribution with diffuse tails representing a significant signal (George and McCulloch, 1997). In high-dimensional clustering it is more meaningful to assume that large chunk of data are noise perhaps because of inclusion of a large number of non-discriminating variables. We propose to add another hierarchy to the model (7) in order to build a spike-and-slab model (Mitchell and Beauchamp, 1988) and produce a biclustering method with less sensitivity to noise variables

$$\begin{aligned} y_{IiJj} \mid \theta_{IJ} &\overset{\text{iid}}{\sim} \text{Gaussian}(\theta_{IJ}, \sigma^2), \\ \theta_{IJ} \mid \gamma_{IJ} &\overset{\text{iid}}{\sim} \text{Gaussian}(\mu, \gamma_{IJ}\sigma_\theta^2), \\ \gamma_{IJ} &\overset{\text{iid}}{\sim} \text{Bernoulli}(q), \\ \theta_{IJ}, \mu \in \mathbb{R}, \quad \sigma^2, \sigma_\theta^2 > 0, &\qquad \gamma_{IJ} \in \{0,1\}, \quad 0 < q < 1. \end{aligned} \tag{8}$$

Note that in (8), Gaussian$(\mu, 0)$ denotes a degenerate distribution at μ. Model (8) defines a mixture of two densities for the cluster centre; one when $\gamma_{IJ} = 0$, giving a distribution concentrated about μ (spike), and another when $\gamma_{IJ} = 1$, giving a distribution with diffuse tails (slab). As a consequence, the marginal distribution for data is mixture of Gaussian(μ, σ^2) and Gaussian$(\mu, \sigma^2 + \sigma_\theta^2)$. Model (8) is a spike-and-slab model at μ. If the average of data is subtracted $\mu = 0$ it reduces to a spike-and-slab model at zero. Still the marginal density (6) is tractable by taking $\boldsymbol{\psi} = (\theta_{IJ}, \gamma_{IJ})$, see Appendix for calculations. Noise biclusters are submatrices with $\gamma_{IJ} = 0$. Therefore, using the spike-and-slab model (8) one can judge about the important biclusters through the Bayes factor of $\gamma_{IJ} = 1$ versus $\gamma_{IJ} = 0$. Though in real applications discovering noise variables (columns) and noise subjects (rows) is more of interest.

We propose to use (8) with slight modification for row or column noise detection

$$\begin{aligned} y_{IiJj} \mid \theta_{IJ} &\overset{\text{iid}}{\sim} \text{Gaussian}(\theta_{IJ}, \sigma^2), \\ \theta_{IJ} \mid \gamma_{IJ} &\overset{\text{iid}}{\sim} \text{Gaussian}(\mu, \gamma_{IJ}\sigma_\theta^2), \\ \gamma_{IJ} \mid \delta_{Ii}, \delta_{Jj} &\overset{\text{iid}}{\sim} \text{Bernoulli}(\delta_{Ii}\delta_{Jj}q), \\ \delta_{Ii} &\overset{\text{iid}}{\sim} \text{Bernoulli}(0.5), \\ \delta_{Jj} &\overset{\text{iid}}{\sim} \text{Bernoulli}(0.5), \\ \theta_{IJ}, \mu \in \mathbb{R}, \quad \sigma^2, \sigma_\theta^2 > 0, &\qquad \gamma_{IJ}, \delta_{Ii}, \delta_{Jj} \in \{0,1\}, \quad 0 < q < 1. \end{aligned} \tag{9}$$

The row i of the row-cluster I separates columns only if $\delta_{Ii} = 1$. Similarly, column j of column-cluster J partitions the rows only if $\delta_{Jj} = 1$. Therefore, the Bayes factor of $\delta_{Ii} = 1$ versus $\delta_{Ii} = 0$ can be used to judge about discriminant rows. Likewise, the Bayes factor of $\delta_{Jj} = 1$ versus $\delta_{Jj} = 0$ gives a clue about discriminant columns. We propose a Bernoulli distribution with succes probability 0.5 for both indicator variables. Often, there is no information apriori about the proportion of active rows, or active columns, so a fair choice looks reasonable. On the other hand, we found that estimating these values through maximizing the marginal likelihood is numerically inefficient.

In many metabolomic studies data are replicated, i.e. the same tissue is analyzed several times. Therefore, a generalization of (9) is required to account for subject replication.

Let R_{Ii} denote replications of subject i in row-cluster I. In unreplicated case $R_{Ii} = 1, \forall i, I$. The total number of samples, i.e. the number of rows of $\mathbf{Y}$, is $n = \sum_{I=1}^{k_{\text{row}}} \sum_{i=1}^{n_I} R_{Ii}$. A straightforward generalization of (9) is

$$
\begin{aligned}
y_{IiJjr} \mid \varepsilon_{IiJj} &\overset{\text{iid}}{\sim} \text{Gaussian}(\varepsilon_{IiJj}, \sigma^2),\\
\varepsilon_{IiJj} \mid \theta_{IJ} &\overset{\text{iid}}{\sim} \text{Gaussian}(\theta_{IJ}, \sigma_\varepsilon^2),\\
\theta_{IJ} \mid \gamma_{IJ} &\overset{\text{iid}}{\sim} \text{Gaussian}(\mu, \gamma_{IJ}\sigma_\theta^2),\\
\gamma_{IJ} \mid \delta_{Ii}, \delta_{Jj} &\overset{\text{iid}}{\sim} \text{Bernoulli}(\delta_{Ii}\delta_{Jj}q),\\
\delta_{Ii} &\overset{\text{iid}}{\sim} \text{Bernoulli}(0.5),\\
\delta_{Jj} &\overset{\text{iid}}{\sim} \text{Bernoulli}(0.5),\\
\theta_{IJ}, \mu \in \mathbb{R}, \quad \sigma^2, \sigma_\theta^2 > 0, &\qquad \gamma_{IJ}, \delta_{Ii}, \delta_{Jj} \in \{0,1\}, \quad 0 < q < 1.
\end{aligned} \tag{10}
$$

The extra level of variation ε_{IiJj} mimics the variation between mean of replicated data of size R_{Ii} both generated with the same center. One may run the unreplicated model (9) on the average of replicated tissues instead, but this approach ignores the uncertainty of this average and treats the averaged tissues as a single realization. This issue might be crucial if replications are unequal. The model (10) takes each variable as a column-clustering object but a block of rows as a row-clustering object. The marginal posterior derived from (10) is calculated in the Appendix.

3 Agglomerative Biclustering

In Bayesian clustering a prior distribution is assumed for grouping, and a likelihood is supposed for the data given grouping. Bayesian agglomerative method uses the log posterior as the natural similarity measure to build the binary tree. Consider the row-cluster $\mathcal{C}_{\text{row}}$ is defined over n rows, with $||\mathcal{C}_{\text{row}}|| = k_{\text{row}}$ being the number of blocks of the row-cluster. Similarly suppose the column-cluster $\mathcal{C}_{\text{col}}$ is defined over p columns, with $||\mathcal{C}_{\text{col}}|| = k_{\text{col}}$ being the number of blocks of the column-cluster. We assume groupings are exchangeable a priori , and hence only need to specify a prior distribution for the number of blocks in the partition, and the sizes of the blocks, see Booth et al. (2008) for discussion. Heard et al. (2006) suggest a uniform discrete prior for the number of distinct clusters, and the uniform Multinomial-Dirichlet prior for the cluster sizes given the number of groups. This clustering prior favours small number of clusters and is fast to evaluate. Further we suppose that row-clusters and column-clusters are independent apriori, hence $f(\mathcal{C}_{\text{row}}, \mathcal{C}_{\text{col}}) = f(\mathcal{C}_{\text{row}})f(\mathcal{C}_{\text{col}})$. Following Heard et al. (2006)

$$
f(\mathcal{C}_{\text{row}}) \propto \frac{(k_{\text{row}}-1)\prod_{I=1}^{k_{\text{row}}}(n_I!)}{n(n+k_{\text{row}}-1)!}, \qquad f(\mathcal{C}_{\text{col}}) \propto \frac{(k_{\text{col}}-1)\prod_{J=1}^{k_{\text{col}}}(n_J!)}{p(p+k_{\text{col}}-1)!}. \tag{11}
$$

Initially every matrix entry is regarded as a separate bicluster, so the initial row-cluster has $k_{\text{row}} = n$ blocks and the initial column-cluster has $k_{\text{col}} = p$ blocks. At each stage, every possible merger of pair of row blocks and pair of column blocks is considered, and the merger that maximizes the posterior is applied, either a row merger or a column merger. This gives a symmetric algorithm with respect to the rows and the columns, i.e. the same result is produced over the transpose of the data matrix.

The posterior function typically includes some hyperparameters. We suggest to estimate these hyperparameters using empirical Bayes at the earliest stage of the agglomerative algorithm, providing an automatic choice of hyperparameters.

Assuming the uniform Multinomial-Dirichlet clustering prior (11), given a grouping of variables $\mathcal{C}_{\text{col}}$, the ratio of posteriors for merging subject block I with block I' is

$$
\frac{f(\mathcal{C}_{\text{row}}^{I,I'}, \mathcal{C}_{\text{col}} \mid \mathbf{Y})}{f(\mathcal{C}_{\text{row}}, \mathcal{C}_{\text{col}} \mid \mathbf{Y})} = \frac{f(\mathbf{y}_{\text{row}}^{I,I'} \mid \mathcal{C}_{\text{col}})}{f(\mathbf{y}_{\text{row}}^{I} \mid \mathcal{C}_{\text{col}})f(\mathbf{y}_{\text{row}}^{I'} \mid \mathcal{C}_{\text{col}})} \times \frac{(n+k_{\text{row}}-1)(n_I+n_{I'})!}{(k_{\text{row}}-1)n_I!n_{I'}!}, \tag{12}
$$

where $\mathcal{C}_{\text{row}}^{I,I'}$ is the row grouping after merging row block I with row block I', $\mathbf{y}_{\text{row}}^{I,I'}$ is the data in row block I and I', $\mathbf{y}_{\text{row}}^{I}$ is the data in row block I and $\mathbf{y}_{\text{row}}^{I'}$ is the data in row block I', and n_I and $n_{I'}$ are the number of subjects in row block I and I', respectively. Similarly, the ratio of posteriors given a fixed grouping of

subjects $\mathcal{C}_{\text{row}}$ for merging column blocks J with J' is

$$\frac{f(\mathcal{C}_{\text{row}}, \mathcal{C}_{\text{col}}^{J,J'} \mid \mathbf{Y})}{f(\mathcal{C}_{\text{row}}, \mathcal{C}_{\text{col}} \mid \mathbf{Y})} = \frac{f(\mathbf{y}_{\text{col}}^{J,J'} \mid \mathcal{C}_{\text{row}})}{f(\mathbf{y}_{\text{col}}^{J} \mid \mathcal{C}_{\text{row}}) f(\mathbf{y}_{\text{col}}^{J'} \mid \mathcal{C}_{\text{row}})} \times \frac{(p + k_{\text{col}} - 1)(n_J + n_{J'})!}{(k_{\text{col}} - 1) n_J! n_{J'}!}. \tag{13}$$

Here $\mathcal{C}_{\text{row}}$ and $\mathcal{C}_{\text{col}}$ are the row and the column groupings in the previous step of the agglomerative algorithm. Merging the clusters continues until a single block containing all rows and columns is achieved. The absolute logarithm of the posterior ratio produces the length of the arms of the binary tree. This length corresponds to the log Bayes factor for comparing two successive groupings. Having the log Bayes factor as the arms of the tree provides immediate and statistically-sensible comparison of different groupings, directly on the tree. The corresponding pseudo code for this algorithm is

1. Initialize $||\mathcal{C}_{\text{row}}|| = k_{\text{row}} = n$, $||\mathcal{C}_{\text{col}}|| = k_{\text{col}} = p$; this initialization means $\mathcal{C}_{\text{row}}$ and $\mathcal{C}_{\text{col}}$ are both collection of singleton clusters. Set the model hyperparameters to some values or estimate them using empirical Bayes.
2. Compute $\mathcal{C}_{\text{row}}^{\max} = \text{argmax}_{I,I'} f(\mathcal{C}_{\text{row}}^{I,I'}, \mathcal{C}_{\text{col}} \mid \mathbf{Y})$ and $\mathcal{C}_{\text{col}}^{\max} = \text{argmax}_{J,J'} f(\mathcal{C}_{\text{row}}, \mathcal{C}_{\text{col}}^{J,J'} \mid \mathbf{Y})$.
3. If $f(\mathcal{C}_{\text{row}}^{\max}, \mathcal{C}_{\text{col}} \mid \mathbf{Y}) > f(\mathcal{C}_{\text{row}}, \mathcal{C}_{\text{col}}^{\max} \mid \mathbf{Y})$, merge the best two row blocks $\mathcal{C}_{\text{row}}^{\max}$, otherwise merge the best two column blocks $\mathcal{C}_{\text{col}}^{\max}$.
4. Stop if $k_{\text{row}} = 1$ and $k_{\text{col}} = 1$, otherwise return to 2.

The posterior $f(\mathcal{C}_{\text{row}}, \mathcal{C}_{\text{col}} \mid \mathbf{Y})$ maximizes for some $(\mathcal{C}_{\text{row}}, \mathcal{C}_{\text{col}})$ on this agglomerative path, providing the best biclustering in the posterior sense on this path; see the top panel of Figure 6. Estimating hyperparameters through empirical Bayes and choosing the cutting point by maximum a posteriori principle on the agglomerative path provides a fully automatic biclustering aglorithm. Despite the well-posed statistical properties of this approach, we still keep the visual guide for other possible grouping, through forestogram.

4 Data Analysis

The metabolomic data set (Messerli et al., 2007) consists of 14 genetic mutants of *Arabidopsis thaliana* measured over 43 metabolites. The measurements were obtained through gas chromatography mass spectrometry, with three replicates for one of the mutants, *ColWT*, and four replicates for the rest. These mutant strains are described as follows: two mutants are defective in starch biosynthesis (*pgm* and *isa2*); four are defective in starch degradation (*sex1*, *sex4*, *mex1* and *dpe2*)–a comparative mutant accumulates starch as a pleiotropic effect (*tpt*)–four are uncharacterized mutants (*deg172*, *deg263*, *ke103* and *sex3*); and finally, three are wild-type plants (*WsWT*, *RLDWT*, and *ColWT*). The idea was to regroup the mutants and indicate what avenues should be explored first when seeking to characterize the plants. Simultaneously, it is of interest to discover which metabolites have a similar functionality over these samples, and which plants are informative to discover such functionality. The logarithm of spectra of the raw data were first preprocessed, a subset of 43 reliably detected metabolites has been selected from the many measured metabolites. We apply our developed method on the preprocessed data to figure out which replicated tissues are similar and which metabolites have a similar pattern over the samples. Our spike-and-slab model, further, finds out noise subjects and metabolites.

Figure 4 shows the forestogram using model (10) while hyperparameters are estimated at the first stage of hierarchical clustering, t.e. treating each data entry as a separate block. The estimated hyperparameters and their standard errors are $\hat{\mu} = 0.083(0.028), \hat{\sigma}^2 = 0.159(0.005), \hat{\sigma_\varepsilon}^2 = 0.373(0.032), \hat{\sigma_\theta}^2 = 5.155(2.773)$, and $\hat{q} = 0.034(0.019)$. Figure 5 shows the projected forestogram, estimated bicluster, important metabolites, and important plants. The projected tree of the forest helps to study the row and the column trees marginally. A finer understanding of relation between these two marginal trees is inherent in the forestogram of Figure 4.

The agglomerative log posterior is shown in Figure 6 top panel. The projected path of the top panel on row and column sides is illustrated in the middle panel. The importance of variables and subjects quantified through the Bayes factor is reported in the bottom panel; bottom left panel shows the row log Bayes factor $\delta_{Ii} = 1$ versus $\delta_{Ii} = 0$, and the bottom right panel shows the column log Bayes factor $\delta_{Jj} = 1$ versus $\delta_{Jj} = 0$.

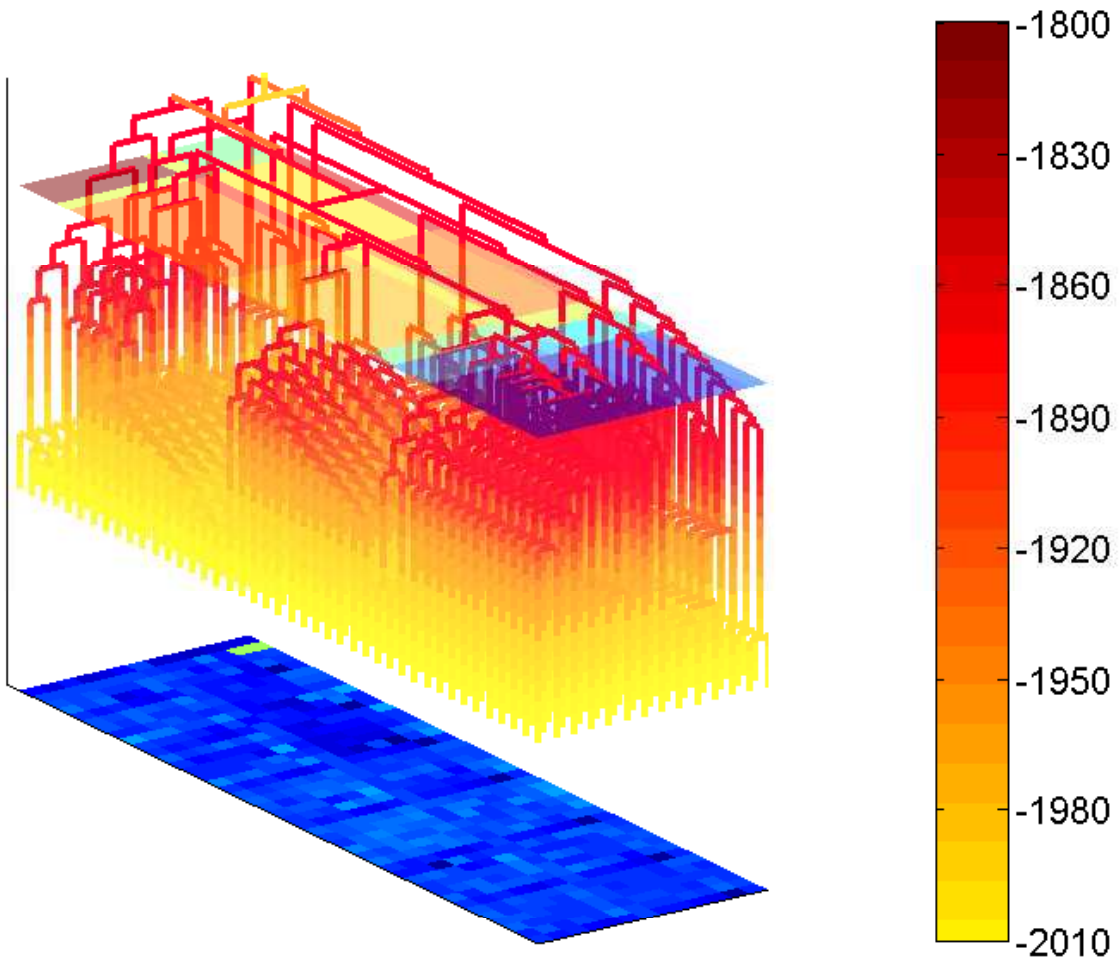

Figure 4: Forestogram demonstrated over the image plot of the metabolic data. The projection of the forest over the row and the column sides is demonstrated in Figure 5. More details about the construction of this forest is visualized in Figure 6.

5 Discussion

We introduced a fully automatic hierarchical biclustering using a suitable spike-and-slab model with tractable marginals for metabolic data. The model allows to incorporate data replication and identifies noise tissues and metabolites. Grouping with a tractable log posterior provides a visualization facility through forestogram. Some other Bayesian models allow several choices for the distribution of centers with tractable marginals; for instance, in model (8) an asymmetric Laplace distribution for θ_{IJ} (Bhowmick et al., 2006) also produces tractable marginals and hence provides another similarity measure for biclustering. Biclustering other types of data such as discrete data is feasible using the introduced approach through other marginally tractable models, e.g. conjugate models. Grouping with log posterior as the similarity measure provides a visualization facility through forestogram. In this work we focused on parametric models, but conjugate Dirichlet mixtures also can be used for a similar analysis (Savage et al., 2009).

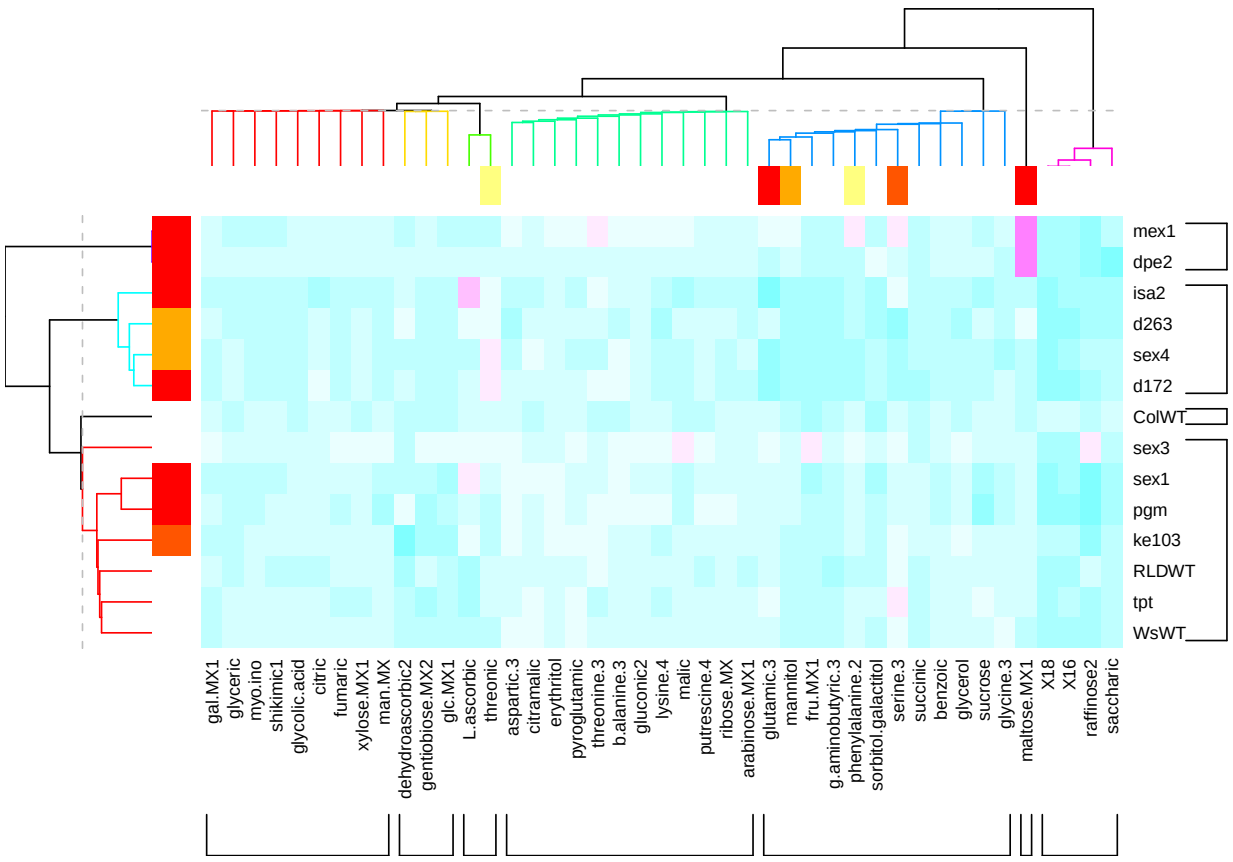

Figure 5: Projected forestogram of Figure 4 over row and column sides. Important metabolites and plants shown on the margin of the image plot of the metabolomic data. The forestogram is cut at the maximum a posteriori grouping over the agglomerative path and the estimated grouping is shown at the opposite side of each marginal dendrogram. Metabolite and plant importance is calculated for the maximum a posteriori grouping using model (10), a heat color is shown for an important metabolite or plant, and blank otherwise, see also Figure 6.

Forestogram is a connection between biclustering and three-dimensional binary tree. This approach is different with running hierarchical clustering on rows and columns separately. The introduced method is computationally more expensive, but visually finer, and statistically more interpretable. While posterior is used as the arms of the tree, the arms of the forestogram have a probabilistic interpretation, since the length of each arm corresponds to a log Bayes factor. This facilitates the comparison of different groupings directly on the forestogram if needed. However, visualization through forestogram is infeasible for large matrices since a trivial implementation of the suggested algorithm is of $O(n^3p^3)$ apart from the one time optimization needed for estimation of hyperparameters. If the postrior has Lance–Williams property (Lance and Williams, 1967) the computation can be improved to $O(n^2p^2 \log n \log p)$.

Importance values for ordering effective variable and subjects can be evaluated rapidly once a sensible grouping is given. If one is interested only in disjoint grid biclustering but not the three-dimensional tree, Markov chain Monte Carlo methods can be used. Adopting Markov chain Monte Carlo sampling enables averaging the variable and subject importance over sampled groupings. In our suggested method the importance is calculated only based on one grouping: the maximum a posteriori grouping found on the agglomerative path. Posterior computations and biclustering calculations for this research has been implemented in R statistical software (R Core Team, 2014). The R package `baybi` is under development on R-Forge and will be released on R-CRAN in the near future. The forestogram graph has been implemented in MATLAB (2014).

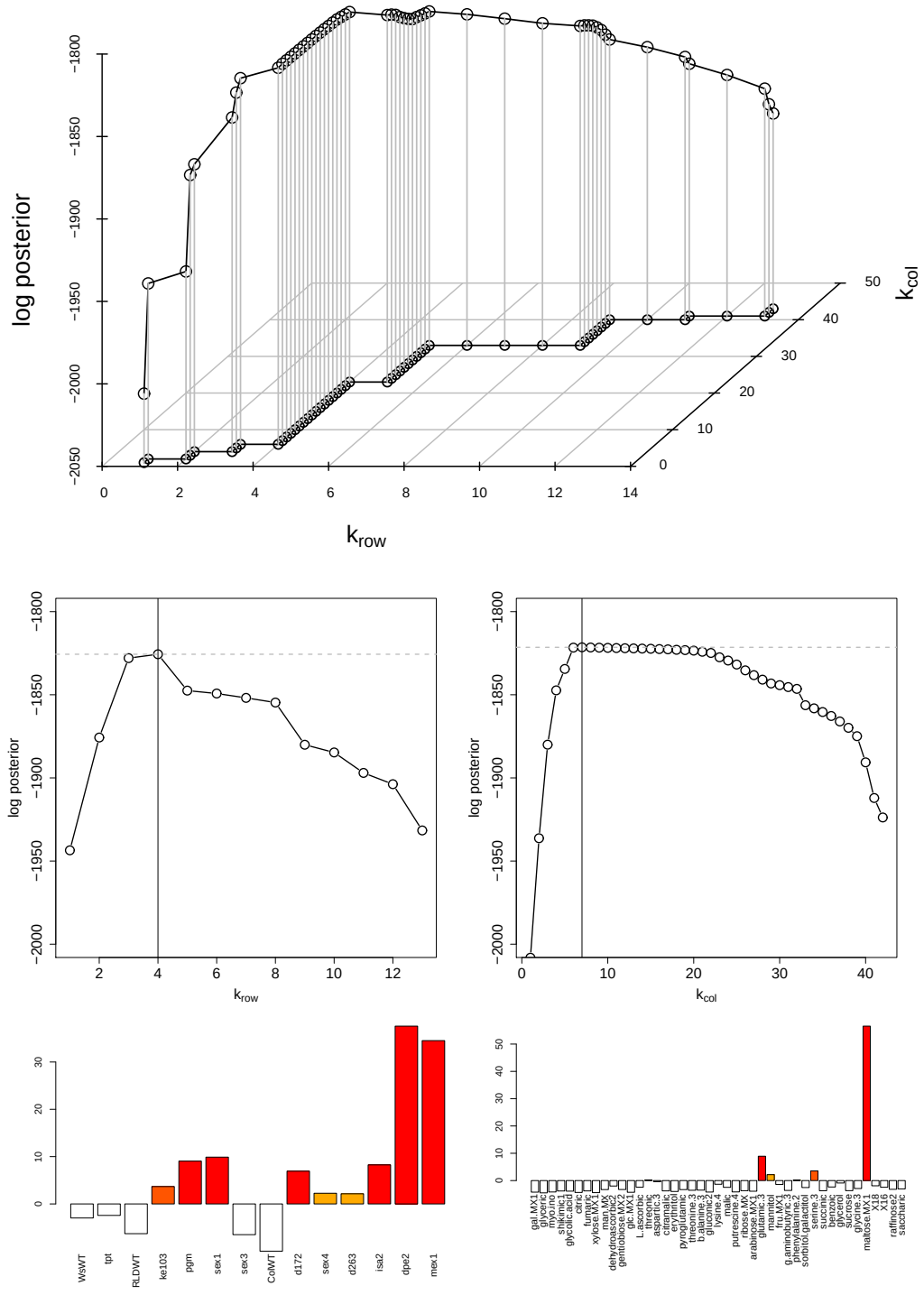

Figure 6: Top: agglomerative log posterior values over the agglomerative path for the metabolomic data shown on the number of metabolite clusters (k_{col}) and the number of plant clusters (k_{row}). Middle: the projected log posterior path of the top panel, for plants (left panel), and for metabolites (right panel). The horizontal dashed line refers to the maximum a posteriori cutting point on the log posterior path of the top panel and the vertical solid line to the optimal number of groups of plants (left panel) and metabolites (right panel). Bottom: the log Bayes factor as the importance measure for plants (left panel) and for metabolites (right panel); a heat color (important) if the log Bayes factor is positive and blank (unimportant) otherwise, see also Figure 5.

Appendix

In this section the marginal density for a given row-cluster $\mathcal{C}_{\text{row}}$ and column-cluster $\mathcal{C}_{\text{col}}$, $f(. \mid \mathcal{C}_{\text{row}}, \mathcal{C}_{\text{col}})$, is denoted by $f(.)$ for the sake of simplicity in notation. In building the tree we suppose $\delta_{Ii} = 1$ and $\delta_{Jj} = 1$. These parameters are only useful for testing noise rows or columns after a forestogram is built.

The marginal density for the model (10) is simply a random effect model with disappearing random component θ_{IJ} whose appearance is controlled by the Bernoulli random variable γ_{IJ}. We marginalize first over γ_{IJ},

$$f(\mathbf{Y}) = \prod_{I=1}^{k_{\text{row}}} \prod_{J=1}^{k_{\text{col}}} \{q f(\mathbf{y}_{IJ} \mid \gamma_{IJ} = 1) + (1-q) f(\mathbf{y}_{IJ} \mid \gamma_{IJ} = 0)\}. \tag{14}$$

The density for a given $\gamma_{IJ} = 0$ is

$$f(\mathbf{y}_{IJ} \mid \gamma_{IJ} = 0) = \prod_{j=1}^{n_J} \prod_{i=1}^{n_I} \int_{-\infty}^{\infty} \prod_{r=1}^{R_{Ii}} f(y_{IiJjr} \mid \varepsilon_{IiJj}) f(\varepsilon_{IiJj} \mid \gamma_{IJ} = 0) d\varepsilon_{IiJj}, \tag{15}$$

in which $f(y_{IiJjr} \mid \varepsilon_{IiJj}) = (2\pi\sigma^2)^{-1/2} \exp\left\{-\frac{1}{2\sigma^2}(y_{IiJjr} - \varepsilon_{IiJj})^2\right\}$, and $f(\varepsilon_{IiJj} \mid \gamma_{IJ} = 0) = (2\pi\sigma_\varepsilon)^{-1/2}$ $\exp\{-\frac{1}{2\sigma_\varepsilon^2}(\varepsilon_{IiJj} - \mu)^2\}$. After making (15) a complete square in terms of ε_{IiJj},

$$\begin{aligned} f(\mathbf{y}_{IJ} \mid \gamma_{IJ} = 0) = \prod_{i=1}^{n_I} \prod_{j=1}^{n_J} \quad & (2\pi)^{-RIi/2} \sigma^{1-R_{Ii}} (R_{Ii}\sigma_\varepsilon^2 + \sigma^2)^{-1/2} \\ \times \quad & \exp\left\{ -\frac{1}{2\sigma^2} \left(\sum_{r=1}^{R_{Ii}} y_{IiJjr}^2 - R_{Ii} \overline{y}_{IiJj} \right) - \frac{(\overline{y}_{IiJj} - \mu)^2}{2(\sigma_\varepsilon^2 + \sigma^2/R_{Ii})} \right\}, \end{aligned} \tag{16}$$

in which $\overline{y}_{IiJj} = R_{Ii}^{-1} \sum_{r=1}^{R_{Ii}} y_{IiJjr}$.

The density $f(\mathbf{y}_{IJ} \mid \gamma_{IJ} = 1)$ corresponds to the hierarchical model

$$\begin{aligned} y_{IiJjr} \mid \varepsilon_{IiJj} &\overset{\text{iid}}{\sim} \text{Gaussian}(\varepsilon_{IiJj}, \sigma^2), \\ \varepsilon_{IiJj} \mid \theta_{IJ} &\overset{\text{iid}}{\sim} \text{Gaussian}(\theta_{IJ}, \sigma_\varepsilon^2), \\ \theta_{IJ} &\overset{\text{iid}}{\sim} \text{Gaussian}(\mu, \sigma_\theta^2), \\ & \varepsilon_{IiJj}, \theta_{IJ}, \mu \in \mathbb{R}, \quad \sigma^2, \sigma_\varepsilon^2, \sigma_\theta^2 > 0. \end{aligned} \tag{17}$$

Evaluation of $f(\mathbf{y}_{IJ} \mid \gamma_{IJ} = 1)$ is straightforward using the standard mixed effect model matrix notation. Suppose an appropriate design matrix $\mathbf{Z}$ with $\sum_{i=1}^{n_I} R_{Ii} + n_J$ rows and $n_I + n_J$ columns and vector $\boldsymbol{\varepsilon}$ is a vector of length $n_I + n_J$ with entries ε_{IiJj}. One may re-write the hierarchical model (17) as

$$\begin{aligned} \mathbf{y}_{IJ} \mid \boldsymbol{\varepsilon} &\sim \mathcal{N}(\mu + \mathbf{Z}\boldsymbol{\varepsilon}, \sigma^2 \mathbf{I}), \\ \boldsymbol{\varepsilon} &\sim \mathcal{N}(\mathbf{0}, \mathbf{W}), \end{aligned}$$

in which $\mathcal{N}$ denotes the multivariate Gaussian distribution, $\mathbf{I}$ is the identity matrix, and $\mathbf{W}$ is an $(n_I + n_J) \times (n_I + n_J)$ uniform matrix with main diagonals $\sigma_\varepsilon^2 + \sigma_\theta^2$ and off-diagonals σ_θ^2 obtained after integration over a univariate θ_{IJ}. Using standard mixed effect calculations we have

$$\mathbf{y}_{IJ} \sim \mathcal{N}(\mu \mathbf{1}, \boldsymbol{\Sigma}), \tag{18}$$

where $\mathbf{1}$ denotes the unit vector and the covariance matrix $\boldsymbol{\Sigma} = \sigma^2 \mathbf{I} + \mathbf{Z}\mathbf{W}\mathbf{Z}'$. The covariance matrix $\boldsymbol{\Sigma}$ corresponds to an $(n_J + \sum_{i=1}^{n_I} R_{Ii}) \times (n_J + \sum_{i=1}^{n_I} R_{Ii})$ symmetric matrix with main diagonals $\sigma^2 + \sigma_\varepsilon^2 + \sigma_\theta^2$ and off-diagonals $\sigma_\varepsilon^2 + \sigma_\theta^2$ or σ_θ^2.

References

Bhowmick, D., Davison, A. C., Goldstein, D. R. and Ruffieux, Y. (2006) A Laplace mixture model for identification of differential expression in microarray experiments. *Biostatistics* **7**, 630–641.

Booth, J. G., Casella, G. and Hobert, J. P. (2008) Clustering using objective functions and stochastic search. *Journal of the Royal Statistical Society, Series B* **70**, 119–139.

Boulesteix, A.-L. and Schmid, M. (2014) Machine learning versus statistical modeling. Biometrical Journal (to appear).

Cheng, Y. and Church, G. M. (2000) Biclustering of expression data. In *International Conference on Intelligent Systems for Molecular Biology*, volume 8, pp. 93–103.

Everitt, B., Landau, S., Leese, M. and Stahl, D. (2011) *Cluster Analysis*. New York: Wiley.

Fiehn, O., Kopka, J., Dormann, P., Altmann, T., Trethewey, N. R. and Willmizer, L. (2000) Metabolic profiling represents a novel and powerful approach for plant functional genomics. *Nature Biotechnology* **18**, 1157–1161.

Gan, X., Liew, A. and Yan, H. (2008) Discovering biclusters in gene expression data based on high-dimensional linear geometries. *BMC Bioinformatics* **9**, 209.

George, E. I. and McCulloch, R. E. (1997) Approaches for Bayesian variable selection. *Statistica Sinica* **7**, 339–373.

Gohlke, R. S. and McLafferty, F. W. (1993) Early gas chromatography/mass spectrometry. *Journal of the American Society for Mass Spectrometry* **4**, 367–371.

Gu, J. and Liu, J. S. (2008) Bayesian biclustering of gene expression data. *BMC Genomics* **9**, S4+.

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

Hartigan, J. A. and Wong, M. A. (1979) A k-means clustering algorithm. *Applied Statistics* **28**, 100–108.

Heard, N. A., Holmes, C. C. and Stephens, D. A. (2006) A quantitative study of gene regulation involved in the immune response of *Anopheline* mosquitoes: An application of Bayesian hierarchical clustering of curves. *Journal of the American Statistical Association* **101**(473), 18–29.

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

Kaufman, L. and Rousseeuw, P. J. (1990) *Finding Groups in Data: An Introduction to Cluster Analysis*. New York: Wiley.

Lance, G. N. and Williams, W. T. (1967) A general theory of classifactory sorting strategies, II. clustering systems. *Computer Journal* **10**, 271–277.

Lazzeroni, L. and Owen, A. (2002) Plaid models for gene expression data. *Statistica Scinica* **12**, 61–68.

Madeira, S. and Oliveira, A. (2004) Biclustering algorithms for biological data analysis: a survey. *IEEE Transactions on Computational Biology and Bioinformatics* pp. 24–45.

MATLAB (2014) *version 8.3 (R2014a)*. Natick, Massachusetts: The MathWorks Inc.

Messerli, G., Partovi Nia, V., Trevisan, M., Kolbe, A., Schauer, N., Geigenberger, P., Chen, J., Davison, A. C., Fernie, A. R. and Zeeman, S. C. (2007) Rapid classification of phenotypic mutants of Arabidopsis via metabolite fingerprinting. *Plant Physiology* **143**, 1481–1492.

Mitchell, T. J. and Beauchamp, J. J. (1988) Bayesian variable selection in linear regression (with discussion). *Journal of the American Statistical Association* **83**, 1023–1036.

R Core Team (2014) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Redestig, H., Repsilber, D., Sohler, F. and Selbig, J. (2007) Integrating functional knowledge during sample clustering for microarray data using unsupervised decision trees. *Biometrical Journal* **49**, 214–229.

Savage, R. S., Heller, K., Xu, Y., Ghahramani, Z., Truman, W. M., Grant, M., Denby, K. J. and Wild, D. L. (2009) R/BHC: fast Bayesian hierarchical clustering for microarray data. *BMC Bioinformatics* **10**, 242.

Sheng, Q., Moreau, Y. and De Moor, B. (2003) Biclustering microarray data by Gibbs sampling. *Bioinformatics* **19**, 196–205.

Smith, J., Anderson, P. and Liverani, S. (2008) Separation measures and the geometry of Bayes factor selection for classification. *Journal of the Royal Statistical Society, Series B* **70**, 957–980.

Tanay, A., Sharan, R. and Shamir, R. (2005) Biclustering algorithms: A survey. In *Handbook of Computational Molecular Biology*, ed. S. Aluru. CRC Press.

Thomas, N., Goodacre, R., Timmins, E., Gaudoin, M. and Fleming, R. (2000) Fourier transform infrared spectroscopy of follicular fluids from large and small antral follicles. *Human Reproduction* **15**, 1667.

van Uitert, M., Meuleman, W. and Wessels, L. (2008) Biclustering sparse binary genomic data. *Journal of Computational Biology* **15**, 1329–1345.

Vaidyanathan, S., Rowland, J., Kell, D. and Goodacre, R. (2001) Discrimination of aerobic endospore-forming bacteria via electrospray-ionization mass spectrometry of whole cell suspensions. *Analytical Chemistry* **73**, 4134–4144.

Yeung, K. and Ruzzo, W. (2001) Principal component analysis for clustering gene expression data. *Bioinformatics* **17**, 763–774.

Zhang, J. (2010) A Bayesian model for biclustering with applications. *Journal of the Royal Statistical Society, Series C* **59**, 635–656.