BAYESIAN BICLUSTERING WITH THE PLAID MODEL

José Caldas, Samuel Kaski

Helsinki Institute for Information Technology
Department of Information and Computer Science
Helsinki University of Technology
P.O. Box 5400, FI-02015 TKK, Finland

ABSTRACT

Biclustering is an active and promising research topic in unsupervised learning. With the aim of uncovering condition-specific similarities between objects, it may be applied in areas such as collaborative filtering and bioinformatics. The plaid model is amongst the most flexible biclustering models. However, its potential has not yet been fully explored. In this paper we extend the plaid model with a Bayesian framework and a collapsed Gibbs sampler. We show that the new method is useful in a gene expression study both in finding gene-specific associations between microarrays and condition-specific associations between genes.

1. INTRODUCTION

We consider the common setup of when the input data set has the form of a matrix $\boldsymbol{Y}$, where entry Y_{ij} refers to an observation of object i under condition j. Two common examples of this type of data are user-movie ratings or gene mRNA measurements under various clinical conditions. Such data sets are called *dyadic* [1].

Classical clustering algorithms such as k-means or hierarchical clustering (see [2]) attempt to find relations between objects that hold under all conditions. However, this may sometimes be an overly restrictive constraint: For instance, a set of users might have a similar opinion about movies of a given type, but may disagree when it comes to other types of movies; similarly, mRNA expression may be homogeneous in a given set of genes, but only under particular biological conditions. The task of searching for associations between objects in subsets of the conditions is most frequently known as *biclustering*. Two main areas of application of biclustering algorithms have been collaborative filtering and gene expression analysis. In this article we mostly discuss algorithms for the latter application.

Using DNA microarrays, it is possible to measure mRNA expression levels for a large set of genes under different experimental conditions. Clustering of gene expression data allows one to both confirm existing biological knowledge and formulate new hypotheses on the functional role of uncharacterized genes (first studies in [3]). However, similarity in expression between genes is not necessarily condition-independent. It has been observed that co-expression of genes is often the result of the activation of condition-specific cellular processes [4, 5].

Many biclustering algorithms and models have already been proposed (for a review and taxonomy see [6]). Among them is the plaid model [7], which is arguably one of the most flexible biclustering models up to now. Although it has a high applicability potential, it has not yet been used to a large extent. Its original authors provide a greedy, partly heuristic inference algorithm [7], which leaves open the question of whether a more principled probabilistic approach would produce better results.

In this paper we provide a Bayesian framework for the plaid model and introduce a collapsed Gibbs sampler for inferring the posterior distribution of bicluster memberships, more specifically the binary membership variables that indicate which genes and which microarrays belong to each bicluster. We run the new inference mechanism on a subset of a human microarray data set [8] to show that the method yields better associations between genes and conditions than would be expected either by chance of by running a hierarchical clustering algorithm. Finally, we also relate the plaid model to two other models [9, 10].

The structure of this article is the following: In section 2 we describe the original formulation of the plaid model and our new Bayesian extension. In section 3 we compare the plaid model with two other models [9, 10]. In section 4 we conduct two tests on our method and show that it provides consistent results. Finally, in section 5 we conclude on the value of the new formulation and provide directions for future work.

The authors are with the Adaptive Informatics Research Centre. This work was supported in part by the PASCAL2 Network of Excellence of the European Community and Tekes (Multibio project). We would like to thank J. Sinkkonen, L. Lahti, and A. Klami for helpful feedback and comments. We further thank L. Lahti for his ideas and suggestions regarding the experimental protocol.

2. THE PLAID MODEL

In this section we review the original formulation for the plaid model and introduce a Bayesian extension for it.

2.1. Original Formulation

Consider a data matrix $\boldsymbol{Y}$ of dimensions $N \times M$. Each entry Y_{ij} refers to the observation of object i under condition j. The plaid model consists of a bias plus a sum of K layers, where each layer (or bicluster) is an ANOVA model [7],

$$Y_{ij} = \mu_0 + \sum_{k=1}^{K} (\mu_k + \alpha_{ik} + \beta_{jk})\, \rho_{ik}\, \kappa_{jk}. \qquad (1)$$

The authors implicitly assume that there is Gaussian noise, as the parameter inference algorithm is based on the minimization of a quadratic error function.

The variables are defined as follows:

- $\boldsymbol{\rho}$ and $\boldsymbol{\kappa}$ are matrices of size $N \times K$ and $M \times K$, respectively, containing binary membership variables. Here $\rho_{ik} = 1$ iff object i belongs to bicluster k, and $\kappa_{jk} = 1$ iff condition j belongs to bicluster k.
- $\boldsymbol{\mu} \in \mathbb{R}^K$, $\boldsymbol{\alpha} \in \mathbb{R}^{N \times K}$, and $\boldsymbol{\beta} \in \mathbb{R}^{M \times K}$ have the same semantics as in standard ANOVA models. For each bicluster k, $\boldsymbol{\alpha}_{\cdot k}$ and $\boldsymbol{\beta}_{\cdot k}$ are defined as departures from the mean μ_k, so that

$$\sum_{i=1}^{N} \alpha_{ik} \rho_{ik} = 0 \qquad (2)$$

and

$$\sum_{j=1}^{M} \beta_{jk} \kappa_{jk} = 0. \qquad (3)$$

- The variable $\mu_0 \in \mathbb{R}$ indicates a bicluster to which all the points belong.

Each bicluster k is specified by which of the variables $\boldsymbol{\rho}_{\cdot k}$ and $\boldsymbol{\kappa}_{\cdot k}$ are equal to 1. It corresponds to a submatrix of $\boldsymbol{Y}$. The plaid model also allows each gene and condition to belong to more than one bicluster.

The authors of the original plaid model devised a partly heuristic, iterative algorithm that attempts to minimize the quadratic error between the data matrix supplied as input and the model given in (1) [7].

2.2. Bayesian Extension

We first specify probability distributions for the model variables in a standard way. We then describe a compact way to represent the relationship between $\boldsymbol{Y}$ and the remaining variables.

Each point Y_{ij}, conditioned on the parameters μ_0, $\boldsymbol{\alpha}_{i\cdot}$, $\boldsymbol{\beta}_{j\cdot}$, $\boldsymbol{\rho}_{i\cdot}$, $\boldsymbol{\kappa}_{j\cdot}$, and σ^2, is assumed to follow a Gaussian distribution,

$$Y_{ij} \;\sim\; N\left(\mu_0 + \sum_{k=1}^{K} (\alpha_{ik} + \beta_{jk})\, \rho_{ik}\, \kappa_{jk}, \sigma^2\right) \qquad (4)$$

(for succinctness, conditioning on the relevant parameters is omitted from the above formula). We have adapted the model for gene expression in a bicluster from (1). We removed the variable μ_k and relaxed the constraints (2) and (3), as it allows for a simpler and more direct handling of the model, without losing generality; notice that there is unidentifiability in the original model. It is assumed, as normally in modelling, that the data points in $\boldsymbol{Y}$, given the relevant parameters, are uncorrelated and share the same scalar variance parameter σ^2.

As for the parameters μ_0, $\boldsymbol{\alpha}$, and $\boldsymbol{\beta}$, we also assign Gaussian distributions to them,

$$\mu_0 \;\sim\; N(0, \sigma_\mu^2 \sigma^2), \qquad (5)$$
$$\alpha_{ik} \;\sim\; N(0, \sigma_\alpha^2 \sigma^2), \qquad (6)$$
$$\beta_{jk} \;\sim\; N(0, \sigma_\beta^2 \sigma^2), \qquad (7)$$

where σ_μ^2, σ_α^2, and σ_β^2 are scalar hyper-parameters specified by the user. In the remaining text we will refer to μ_0, $\boldsymbol{\alpha}$, and $\boldsymbol{\beta}$ collectively as $\boldsymbol{\Theta}$,

$$\boldsymbol{\Theta} = [\mu_0\, \boldsymbol{\alpha}_{\cdot 1}\, \boldsymbol{\beta}_{\cdot 1}\, \ldots\, \boldsymbol{\alpha}_{\cdot K}\, \boldsymbol{\beta}_{\cdot K}]^T.$$

Thus, $\boldsymbol{\Theta}$ follows a Gaussian distribution,

$$\boldsymbol{\Theta} | \sigma^2, \sigma_\mu^2, \sigma_\alpha^2, \sigma_\beta^2 \;\sim\; N(\mathbf{0}, D), \qquad (8)$$

where D is a diagonal covariance matrix.

The relation between $\boldsymbol{\Theta}$ and the data matrix $\boldsymbol{Y}$ can be expressed compactly. Consider a vectorized representation of $\boldsymbol{Y}$, obtained by vertically juxtaposing all of its columns (column-major order). We can express (4) as

$$\boldsymbol{Y} | \boldsymbol{\Theta}, \boldsymbol{\rho}, \boldsymbol{\kappa}, \sigma^2 \;\sim\; N(\boldsymbol{A}\boldsymbol{\Theta}, \sigma^2 I), \qquad (9)$$

where the matrix $\boldsymbol{A} = [A_0 A_1 \dots A_K]$ is defined as

$$A_0 = \mathbf{1}_{NM\times 1}, \tag{10}$$

$$A_{k\geq 1} = [A_{k1} A_{k2}], \tag{11}$$

$$A_{k1} = \begin{bmatrix} \rho_{1k}\kappa_{1k} & & \\ & \ddots & \\ & & \rho_{Nk}\kappa_{1k} \\ \dots & \dots & \dots \\ \rho_{1k}\kappa_{Mk} & & \\ & \ddots & \\ & & \rho_{Nk}\kappa_{Mk} \end{bmatrix}, \tag{12}$$

$$A_{k2} = \begin{bmatrix} \rho_{1k}\kappa_{1k} & & \\ \vdots & & \\ \rho_{Nk}\kappa_{1k} & & \\ \dots & \dots & \dots \\ & & \rho_{1k}\kappa_{Mk} \\ & & \vdots \\ & & \rho_{Nk}\kappa_{Mk} \end{bmatrix}. \tag{13}$$

Structures of the same sort as in $\boldsymbol{A}$ are common in the context of ANOVA models [11].

For a fully Bayesian approach we assign probability distributions to the binary membership variables $\boldsymbol{\rho}$ and $\boldsymbol{\kappa}$. For each bicluster k,

$$\boldsymbol{\rho}_{\cdot k}|\pi_k \sim \text{Binomial}(N, \pi_k), \tag{14}$$

$$\boldsymbol{\kappa}_{\cdot k}|\lambda_k \sim \text{Binomial}(M, \lambda_k). \tag{15}$$

The parameters π_k and λ_k are respectively the probability of a gene belonging to a bicluster k and the probability of a condition belonging to bicluster k. The variables $\boldsymbol{\rho}$ and $\boldsymbol{\kappa}$ are assumed to be independent.

As is standard in Bayesian models, we also assign a prior to π_k and λ_k. We consider each of them to follow a Beta distribution, which is the conjugate prior of the Binomial distribution,

$$\pi_k \sim \text{Beta}(\delta_\rho^{(k)}, \gamma_\rho^{(k)}), \tag{16}$$

$$\lambda_k \sim \text{Beta}(\delta_\kappa^{(k)}, \gamma_\kappa^{(k)}). \tag{17}$$

Given the above specification, we can calculate the probability distribution of $\boldsymbol{\rho}_{\cdot k}$ and $\boldsymbol{\kappa}_{\cdot k}$ after integrating out π_k and λ_k,

$$P(\boldsymbol{\rho}_{\cdot,k}) = \frac{B(n_1 + \delta_\rho^{(k)}, N - n_1 + \gamma_\rho^{(k)})}{B(\delta_\rho^{(k)}, \gamma_\rho^{(k)})}, \tag{18}$$

$$P(\boldsymbol{\kappa}_{\cdot,k}) = \frac{B(m_1 + \delta_\kappa^{(k)}, M - m_1 + \gamma_\kappa^{(k)})}{B(\delta_\kappa^{(k)}, \gamma_\kappa^{(k)})}, \tag{19}$$

where n_1 and m_1 are respectively the number of objects and conditions in bicluster k, and $B(x, y)$ is the beta function,

$$B(x, y) = \int_0^1 t^{x-1}(1-t)^{y-1} dt. \tag{20}$$

The probability distributions of $\boldsymbol{\rho}$ and $\boldsymbol{\kappa}$ are in turn given by

$$P(\boldsymbol{\rho}) = \prod_{k=1}^{K} P(\boldsymbol{\rho}_{\cdot,k}), \tag{21}$$

$$P(\boldsymbol{\kappa}) = \prod_{k=1}^{K} P(\boldsymbol{\kappa}_{\cdot,k}). \tag{22}$$

Finally, we assume σ^2 to follow a scaled inverse-chi-square distribution,

$$\sigma^2 \sim \text{Inv}{-}\chi^2(\upsilon, \sigma_0^2). \tag{23}$$

2.3. Collapsed Gibbs Sampler

We are interested in sampling from the posterior distribution of the membership variables $\boldsymbol{\rho}$ and $\boldsymbol{\kappa}$. Applying Bayes' law, we obtain

$$P(\boldsymbol{\rho}, \boldsymbol{\kappa}|\boldsymbol{Y}) = \frac{P(\boldsymbol{\rho}, \boldsymbol{\kappa})P(\boldsymbol{Y}|\boldsymbol{\rho}, \boldsymbol{\kappa})}{P(\boldsymbol{Y})}, \tag{24}$$

where $P(\boldsymbol{Y})$ is a normalizing factor that depends neither on $\boldsymbol{\rho}$ nor on $\boldsymbol{\kappa}$, and $P(\boldsymbol{Y}|\boldsymbol{\rho}, \boldsymbol{\kappa})$ is the conditional probability density function of $\boldsymbol{Y}$ after integrating $\boldsymbol{\Theta}$ and σ^2 out. That integral is given by

$$P(\boldsymbol{Y}|\boldsymbol{\rho}, \boldsymbol{\kappa}) = \int P(\boldsymbol{Y}|\boldsymbol{\rho}, \boldsymbol{\kappa}, \boldsymbol{\Theta}, \sigma^2) P(\boldsymbol{\Theta}) P(\sigma^2) d\boldsymbol{\Theta} d\sigma^2.$$

Solving the integral, one obtains a multivariate Student-t distribution,

$$\boldsymbol{Y}|\boldsymbol{\rho}, \boldsymbol{\kappa} \sim t_\upsilon \left(0, \left(\frac{I - AQA^T}{\sigma_0^2}\right)^{-1}\right), \tag{25}$$

where $Q = (D^{-1} + A^T A)^{-1}$ [12].

A collapsed Gibbs sampler over $\boldsymbol{\rho}$ and $\boldsymbol{\kappa}$ samples iteratively from the conditional probability distribution of each variable ρ_{ik} or κ_{jk} in turn, conditioned on all the other variables and $\boldsymbol{Y}$. The difference to a standard Gibbs sampler is that $\boldsymbol{\Theta}$ has been integrated out. We refer to the conditional probability distributions as $P(\rho_{i,k}|\boldsymbol{\rho}_{-(i,k)}, \boldsymbol{\kappa}, \boldsymbol{Y})$ and $P(\kappa_{j,k}|\boldsymbol{\kappa}_{-(j,k)}, \boldsymbol{\rho}, \boldsymbol{Y})$, where $\boldsymbol{\rho}_{-(i,k)}$ is obtained from $\boldsymbol{\rho}$ by discarding ρ_{ik}, and $\boldsymbol{\kappa}_{-(j,k)}$ is obtained from $\boldsymbol{\kappa}$ by discarding κ_{jk}. The derivation of the sampler is similar for variables in $\boldsymbol{\rho}$ and variables in $\boldsymbol{\kappa}$. For succinctness, we provide the derivation only for $\boldsymbol{\kappa}$.

The conditional distribution of the (j, k)-th component of $\boldsymbol{\kappa}$ is given by

$$P(\kappa_{j,k}|\boldsymbol{\kappa}_{-(j,k)}, \boldsymbol{\rho}, \boldsymbol{Y}) = \frac{P(\kappa_{j,k}, \boldsymbol{\kappa}_{-(j,k)}, \boldsymbol{\rho}, \boldsymbol{Y})}{P(\boldsymbol{\kappa}_{-(j,k)}, \boldsymbol{\rho}, \boldsymbol{Y})}. \tag{26}$$

The above probability distribution can be more easily computed by first calculating the odds, which after some simplifications are expressible as

$$\frac{P(\kappa_{j,k}=1|\boldsymbol{\kappa}_{-(j,k)},\boldsymbol{\rho},\boldsymbol{Y})}{P(\kappa_{j,k}=0|\boldsymbol{\kappa}_{-(j,k)},\boldsymbol{\rho},\boldsymbol{Y})} = \frac{P(\kappa_{j,k}=1,\boldsymbol{\kappa}_{\cdot,k})}{P(\kappa_{j,k}=0,\boldsymbol{\kappa}_{\cdot,k})} \times \frac{P(\boldsymbol{Y}|\boldsymbol{\rho},\kappa_{j,k}=1,\boldsymbol{\kappa}_{-(j,k)})}{P(\boldsymbol{Y}|\boldsymbol{\rho},\kappa_{j,k}=0,\boldsymbol{\kappa}_{-(j,k)})}.$$

The first term is easily and efficiently computable. The second term relies on obtaining the product AQA^T, where the matrix Q is obtained by inverting $(D^{-1}+A^TA)$ (see (25)). Obtaining that inverse from scratch is very costly due to its size. However, every time the value of a binary membership variable is switched, the update to Q is of a low rank. This allows us to make use of the Sherman-Morrison-Woodbury (SMW) identity [13]. A particular case of the SMW identity states that, for a given invertible matrix M, its inverse M^{-1}, and an update UV, we have

$$(M+UV)^{-1} = M^{-1} - M^{-1}U(I+VM^{-1}U)^{-1}VM^{-1}.$$

The advantage of using the above formula is that when M is $n \times n$, and both U and V are of a low rank (that is, U is $n \times m$, V is $m \times n$, and $m \ll n$), it is significantly cheaper to compute $(I+VM^{-1}U)^{-1}$ than to directly calculate $(M+UV)^{-1}$. We follow that approach for updating Q (the actual structure of U and V are omitted for brevity). Due to the accumulation of numerical errors after each update, Q must effectively be computed from scratch after a few iterations.

3. RELATED WORK

Segal *et al.* have proposed a probabilistic model for decomposing gene expression into partially overlapping cellular processes [9]. They define a set of genes $N = \{g_1, \ldots, g_n\}$, a set of microarrays $A = \{a_1, \ldots, a_k\}$, and a set of expression objects $E = \{e_{1,1}, \ldots, e_{n,k}\}$, which relate genes to microarrays. Given the existence of j biological processes, the binary attributes $g.M_1, \ldots, g.M_j$ specify, for each gene g, the processes to which it belongs. The continuous attributes $a.C_1, \ldots, a.C_j$ describe, for each microarray a, the level of activation of each process. The expression level of a gene g in a microarray a (represented as an attribute of the corresponding expression object $e_{g,a}$) follows a Gaussian distribution,

$$e_{g,a}.Level \sim N\left(\sum_{p=1}^{j} g.M_p \cdot a.C_p, \sigma_a^2\right). \quad (27)$$

Gene expression is therefore modelled as an additive combination of cellular processes, where each process contributes with a continuous value $a.C_p$, which varies from microarray to microarray. Each gene has the option of belonging to each bicluster or not, by use of the binary membership attribute $g.M_p$.

By considering the following constraints in the Bayesian plaid model, we can obtain Segal's model (apart from the variance parameters σ_a^2):

$$\mu_0 = 0,\ \boldsymbol{\alpha} = \mathbf{0},\ \boldsymbol{\kappa} = \mathbf{1}. \quad (28)$$

Both models are equivalent if we further restrict the variance parameters σ_a^2 to be the same. The attributes $g.M_p$ have the same semantics as the variables ρ_{ik} in the Bayesian plaid model, and the attributes $a.C_p$ are equivalent to the variables β_{jk}. Notice that enforcing the constraints in (28) (apart from $\boldsymbol{\kappa} = \mathbf{1}$) amounts to forcing the parameters σ_μ^2 and σ_α^2 to be zero.

Another model which is related to the plaid model is the one by Meeds *et al.* [10]. The authors model a dyadic data matrix $\boldsymbol{X} \in \mathbb{R}^{n \times m}$ as

$$\boldsymbol{X}|\boldsymbol{U},\boldsymbol{V},\boldsymbol{W} \sim f(\boldsymbol{U}\boldsymbol{W}\boldsymbol{V}^T,\Theta), \quad (29)$$

where $\boldsymbol{U} \in \{0,1\}^{n \times K}$ and $\boldsymbol{V} \in \{0,1\}^{m \times K}$ are binary matrices, $\boldsymbol{W} \in \mathbb{R}^{K \times K}$ is a weight matrix, f is a probability density function (we consider the case when f is a Gaussian distribution), and Θ is a given parameterization. The matrices $\boldsymbol{U}$ and $\boldsymbol{V}$ are equivalent to the matrices $\boldsymbol{\rho}$ and $\boldsymbol{\kappa}$. Consider an object i under condition j, corresponding to the data point X_{ij}. According to the above model, its probability density function is

$$X_{ij}|\boldsymbol{U}_i,\boldsymbol{V}_j,\boldsymbol{W} \sim f(\boldsymbol{U}_i\boldsymbol{W}\boldsymbol{V}_j^T,\Theta), \quad (30)$$

where $\boldsymbol{U}_i$ is the i-th line of $\boldsymbol{U}$ and $\boldsymbol{V}_j$ is the j-th line of $\boldsymbol{V}$. The quadratic form $\boldsymbol{U}_i\boldsymbol{W}\boldsymbol{V}_i^T$ can be also expressed as $\sum_{k_1=1}^{K}\sum_{k_2=1}^{K} W_{k_1,k_2}U_{ik_1}V_{jk_2}$. The case when $\boldsymbol{W}$ is a diagonal matrix is equivalent to a particular version of the original plaid model, namely the one where $\boldsymbol{\alpha} = \mathbf{0}$ and $\boldsymbol{\beta} = \mathbf{0}$ [10].

4. EXPERIMENTS

We made a brief proof-of-concept study by applying the Bayesian plaid model to 79 preprocessed microarrays portraying mRNA gene expression in several human tissues [8]. We analyzed four gene ontology groups, one at at time, to find out whether the biclustering finds subgroups of the groups. The groups were rhythmic processes, regulation of biosynthetic processes, growth regulation, and cell division; expression within each group varies clearly throughout the tissues.

An indirect indication of the meaningfulness of a bicluster is that it finds a pattern over the conditions (here tissues). That is, restricted to the genes belonging to the bicluster, the arrays belonging to the bicluster should be more related than by chance. In order to quantify how well the method

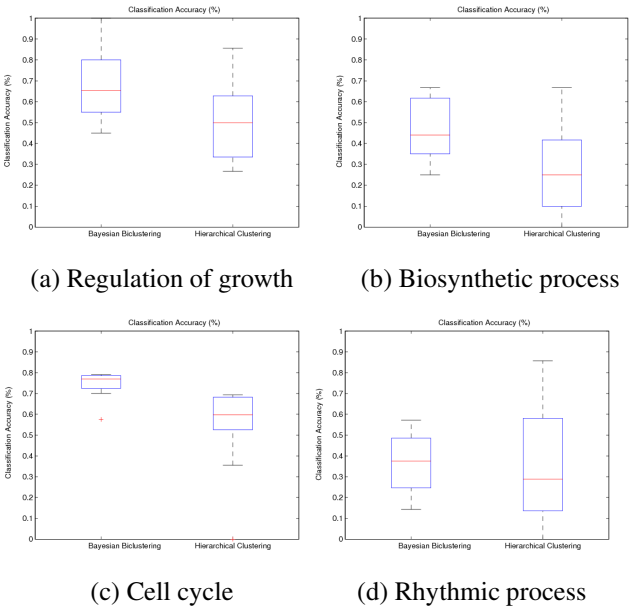
Gene Set	Average Corr. Gain	P-value
Regulation of Growth	0.1367	< 0.001
Biosynthetic process	0.2261	< 0.001
Cell cycle	0.1255	0.007
Rhythmic process	0.2396	0.004

Table 1. Average gain in correlation between pairs of microarrays in the same bicluster, when restricted to the subset of genes in that bicluster.

performs that task, we computed the Pearson correlation coefficient between pairs of microarrays in the same bicluster, restricted to the set of genes in that same bicluster. We then computed the coefficient between the same pairs of microarrays, but this time using all genes. We measured the average gain in correlation when going from the full set of genes to the set of genes in the same bicluster, and calculated a p-value obtained as the empirical probability of obtaining a higher gain by selecting random sets of genes of the same size as the set of genes in a bicluster. Table 1 shows the results for the above experiment. The Bayesian biclustering method appears to consistently choose subsets of genes that significantly increase the correlation between microarrays associated with those subsets.

It is at least as important that the genes belonging to the same bicluster are functionally related. We tested the "purity" of the found biclusters in terms of the Biological Process ontology subclasses of the chosen "main" class (rhythmic processes, regulation of biosynthetic processes, growth regulation, or cell division). We assigned each bicluster to one biological process with a majority voting system, where each gene votes for the gene ontology terms it is associated with. When classifying a gene then, it is first removed from its bicluster, the bicluster is assigned a biological process without taking the gene into account, and the gene becomes classified correctly if its class equals the class of the bicluster. We repeated the experiment for all genes (cross-validation) and also did the same with a standard hierarchical clustering algorithm [3].

Figure 1 shows box plots of classification accuracy in each of the four groups of genes. With the exception of the rhythmic process-related genes, Bayesian biclustering consistently finds sets of genes that are more functionally homogeneous than the ones found by hierarchical clustering. In order to evaluate the significance of the improvements, we applied McNemar's test to each data set. The test takes as input a 2×2 contingency table. In the current context, we assign each gene to each cell in the table, depending on whether it was correctly or incorrectly classified by one method (line) and on whether it was correctly or incorrectly classified by the other method (column). The null hypothesis is that the marginal frequencies in the lines are the same as the marginal frequencies in the columns. This is equivalent to stating that both classification algorithms are performing similarly. The p-values are shown in table 2. The results confirm that the improvement over the hierarchical clustering algorithm is significant in all data sets except the one pertaining to rhythmic process genes.



(a) Regulation of growth (b) Biosynthetic process

(c) Cell cycle (d) Rhythmic process

Fig. 1. Box plots of classification accuracy in biclusters found by the new method and in clusters found by a standard hierarchical clustering algorithm.

Gene Set	P-value
Regulation of Growth	3×10^{-5}
Biosynthetic process	3.1×10^{-4}
Cell cycle	10^{-6}
Rhythmic process	0.015

Table 2. Applying McNemar's test to each set of genes.

5. DISCUSSION

We have introduced a Bayesian formulation for the well-known plaid model [7]. We derived a collapsed Gibbs sampler for inferring the posterior distribution of the (binary) bicluster membership variables, and showed how to efficiently obtain the samples. The model is intrinsically related to an earlier method that searches for overlapping biological processes as well as to a recent binary latent factor model. By applying the resulting method to a gene expression data set, we have demonstrated that it finds both meaningful sets of genes and gene-specific associations between microarrays. The application was a very small-scale proof-of-concept study which we will next extend to larger case

studies. The goal is to find which cellular processes and parts of pathways are activated in each biological condition.

6. REFERENCES

[1] T. Hofmann, J. Puzicha, and M. I. Jordan, "Learning from dyadic data," in *Advances in Neural Information Processing Systems 11*, S. A. Solla, T. K. Leen, and K-R Müller, Eds., pp. 466–472. MIT Press, Cambridge, MA, 1999.

[2] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, Wiley-Interscience, New York, second edition, 2001.

[3] M. B. Eisen, P. T. Spellman, P. O. Brown, and D. Botstein, "Cluster analysis and display of genome-wide expression patterns," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 95, pp. 14863–14868, 1998.

[4] E. Segal, N. Friedman, D. Koller, and A. Regev, "A module map showing conditional activity of expression modules in cancer," *Nature Genetics*, vol. 36, no. 10, pp. 1090–1098, 2004.

[5] A. Battle, E. Segal, and D. Koller, "Probabilistic discovery of overlapping cellular processes and their regulation," *Journal of Computational Biology*, vol. 12, no. 7, pp. 909–927, 2005.

[6] S. Madeira and O. Oliveira, "Biclustering algorithms for biological data analysis: A survey," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 1, no. 1, pp. 24–45, 2004.

[7] L. Lazzeroni and A. Owen, "Plaid models for gene expression data," *Statistica Sinica*, vol. 12, no. 1, pp. 61–86, 2002.

[8] A. I. Su et al., "A gene atlas of the mouse and human protein-encoding transcriptomes," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101, no. 16, pp. 6062–6067, 2004.

[9] E. Segal, A. Battle, and D. Koller, "Decomposing gene expression into cellular processes," in *Proceedings of the 8th Pacific Symposium on Biocomputing, Lihue, Hawaii, USA*, 2003, pp. 89–100.

[10] E. Meeds, Z. Ghahramani, R. M. Neal, and S. T. Roweis, "Modeling dyadic data with binary latent factors," in *Advances in Neural Information Processing Systems 19*, B. Schölkopf, J. Platt, and T. Hoffman, Eds., pp. 977–984. MIT Press, Cambridge, MA, 2007.

[11] H. Scheffé, *The Analysis of Variance*, John Wiley and Sons, Inc., New York, 1999.

[12] A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin, *Bayesian Data Analysis*, Chapman & Hall/CRC, second edition, 2003.

[13] G. Golub and C. Van Loan, *Matrix Computations*, The John Hopkins University Press, Baltimore, third edition, 1996.