The International Journal of Biostatistics

Volume 4, Issue 1

2008

Article 3

Biclustering of Gene Expression Data by an Extension of Mixtures of Factor Analyzers

Francesca Martella, *Sapienza, Università di Roma*

Marco Alfò, *Sapienza, Università di Roma*

Maurizio Vichi, *Sapienza, Università di Roma*

Recommended Citation:

Martella, Francesca; Alfò, Marco; and Vichi, Maurizio (2008) "Biclustering of Gene Expression Data by an Extension of Mixtures of Factor Analyzers," *The International Journal of Biostatistics*: Vol. 4: Iss. 1, Article 3.

DOI: 10.2202/1557-4679.1078

Biclustering of Gene Expression Data by an Extension of Mixtures of Factor Analyzers

Francesca Martella, Marco Alfò, and Maurizio Vichi

Abstract

A challenge in microarray data analysis concerns discovering local structures composed by sets of genes that show homogeneous expression patterns across subsets of conditions. We present an extension of the mixture of factor analyzers model (MFA) allowing for simultaneous clustering of genes and conditions. The proposed model is rather flexible since it models the density of high-dimensional data assuming a mixture of Gaussian distributions with a particular component-specific covariance structure. Specifically, a binary and row stochastic matrix representing tissue membership is used to cluster tissues (experimental conditions), whereas the traditional mixture approach is used to define the gene clustering. An alternating expectation conditional maximization (AECM) algorithm is proposed for parameter estimation; experiments on simulated and real data show the efficiency of our method as a general approach to biclustering. The Matlab code of the algorithm is available upon request from authors.

KEYWORDS: mixture of factor analyzers, biclustering, microarray data

1 Introduction

In the microarray literature, analyzed data are usually stored in a $n \times J$ gene expression matrix, $\mathbf{Y} = \{Y_{ij}\}$, where n (rows) and J (columns) represent, respectively, the n genes observed on J experimental conditions (sample tissues, *etc.*). Depending on the goal of the cluster analysis, the columns may be regarded as variables and the rows as observations, which is common in standard multivariate statistical approaches; otherwise, the roles of rows and columns may be reversed. In fact, if the aim is to identify clusters of genes that show similar regulatory mechanisms, the columns may be considered as variables while rows represent the observations to be classified. On the other hand, if the aim is to find homogeneous groups of experimental conditions, the rows (genes) are regarded as the variables and columns (samples) as the observations. In the latter case, unlike traditional multivariate statistical analysis, the number of variables, n , may exceed the number of observations, J .

In any case, assumptions of conventional clustering methods, that are based on similarity between observations across all variables, might be too restrictive. Observations may in fact be homogeneous only under a limited set of conditions while showing little similarities outside these conditions. In this case, a group of observations may be represented by a row cluster and by an associated subset of the columns which characterizes the cluster. For example, if samples represent patients with a certain disease, it may not be appropriate to require every sample to belong to a cluster, since this does not allow for misdiagnosis. Such limitations of standard clustering techniques have motivated the development of *biclustering* techniques, where rows and columns are clustered *simultaneously*. In general, given an observed matrix, the biclustering method generates *biclusters* (or blocks) - *i.e.* subsets of rows that exhibit similar behavior across subset of columns, or *viceversa*. The term was originally introduced by Hartigan (1972, 1975) and only recently discussed by Cheng and Church (2000) in the microarray data analysis. Since then, many biclustering methods have appeared in the context of bioinformatic area; in the last ten years, two structured taxonomies of biclustering methods have been proposed by Van Mechelen, Bock and De Boeck (2004) and Madeira and Oliveira (2004). In the former paper, the overview is built starting from a traditional statistical/data analytic perspective while in the latter paper, the topic is discussed on the bioinformatics side.

Examples of biclustering methods in a mixture context have been developed by Govaert and Nadif (2003), who propose a block mixture model by using Bernoulli mixtures, and by Li J. and Zha H. (2006), who propose a two-way Poisson mixture model in the the analysis context. Finally, Prelic et al. (2006)

provide a methodology to compare and validate biclustering methods using a fast divide-and-conquer algorithm (Bimax).

In this paper, we review the *mixtures of factor analyzers* model introduced by Ghahramani and Hinton (1996), and successively developed by McLachlan and Peel (2000) and McLachlan et al. (2002). Martella (2006) compared the mixture of factor analyzers with the proposal of Rocci and Vichi (2002). Mixtures of factor analyzers have originally been used for model-based density estimation when facing high dimensional data and for local dimensional reduction. Here, we propose to modify the model in a two-dimensional context in order to cluster simultaneously genes and experimental conditions, *i.e.* for biclustering purposes.

The paper is organized as follows. In the next section, we describe the mixture of factor analyzers, the extension of MFA to the biclustering context and underline some peculiar features, such as the identifiability issue, the estimation algorithm and the model choice. Furthermore, to assess the behavior of the proposed approach we discuss the analysis of simulated as well as real gene expression data. Concluding remarks are given in section 4.

2 Methods

2.1 Mixture of factor analyzers

Let $\mathbf{y}_i$ be a J -dimensional data vector. We assume that $\mathbf{y}_i$ comes from a population with K clusters defined through the following component labels:

$$z_{ik} = \begin{cases} 1 & \text{if } i \text{ belongs to the } k\text{-th cluster} \\ 0 & \text{otherwise} \end{cases}.$$

Under the general mixture of factor analyzers model, $\mathbf{y}_i$ is modeled as

$$\mathbf{y}_i = \boldsymbol{\mu}_k + \mathbf{V}_k \mathbf{u}_{ik} + \mathbf{e}_{ik},$$

with probability π_k ($k = 1, \dots, K$), $i = 1, \dots, n$, where $\boldsymbol{\mu}_k$ is a component-specific mean vector, $\mathbf{V}_k$ is a $J \times Q_k$ component-specific factor loadings matrix, $\mathbf{u}_{ik}$ is a Q_k -dimensional ($Q_k < J$) vector of component-specific latent variables (factors), which are assumed to be i.i.d. draws from $N(\mathbf{0}, \mathbf{I}_{Q_k})$, where $\mathbf{I}_{Q_k}$ denotes the $Q_k \times Q_k$ identity matrix. Furthermore, $\mathbf{e}_{ik}$ are i.i.d. Gaussian component-specific random variables with mean $\mathbf{0}$ and covariance matrix $\mathbf{D}_k = \text{diag}(\sigma_{1k}^2, \dots, \sigma_{Jk}^2)$, that are assumed to be independent of $\mathbf{u}_{ik}$.

Thus, conditional on component k we have:

$$f(\mathbf{y}_i|\mathbf{z}_{ik}) = N(\boldsymbol{\mu}_k, \mathbf{V}_k \mathbf{V}_k' + \mathbf{D}_k),$$

with probability π_k ; therefore the marginal density of $\mathbf{y}_i$ is given by:

$$f(\mathbf{y}_i) = \sum_{k=1}^K \pi_k N(\boldsymbol{\mu}_k, \mathbf{V}_k \mathbf{V}_k' + \mathbf{D}_k). \quad (1)$$

For estimation purposes it is useful to define the MFA model in terms of conditional densities. Using previous assumptions, we obtain

$$f(\mathbf{y}_i|\mathbf{z}_{ik}, \mathbf{u}_{ik}) = N(\boldsymbol{\mu}_k + \mathbf{V}_k \mathbf{u}_{ik}, \mathbf{D}_k).$$

Thus, the set of model parameters is $\boldsymbol{\phi}_k = \{\boldsymbol{\mu}_k, \mathbf{V}_k, \boldsymbol{\pi}_1, \dots, \boldsymbol{\pi}_{K-1}, \mathbf{D}_k\}$, $k = 1, \dots, K$. This model has been introduced by Ghahramani and Hinton (1996) and successively extended by McLachlan and Peel (2000) and McLachlan et al. (2002), for density estimation in high dimensional data with “local” dimension reduction. Tipping and Bishop (1999) develop a mixture of probabilistic principal components model which is closely related to model (1) with unequal, isotropic error component-specific matrices $\mathbf{D}_k = \psi_k \mathbf{I}_J$ with $Q_k = J$, $k = 1, \dots, K$.

2.2 Extension of MFA model for biclustering

If the aim is to cluster not only the units (genes) but also the variables (experimental conditions) of the observed data matrix, one must modify the mixture of factor analyzers model (1).

To introduce variable clustering, we replace the component-specific factor loadings matrix, $\mathbf{V}_k$, with a *binary row stochastic* matrix, $\mathbf{B}_k = \{b_{jl}\}$ ($j = 1, \dots, J$, $l = 1, \dots, Q_k$ and $k = 1, \dots, K$) representing column cluster membership, *i.e.* such that $b_{jl} = 1$ if and only if the j -th condition belongs to the l -th cluster, and 0 otherwise. Furthermore, factors $\mathbf{u}_{ik}$ are replaced with specific-block mean vectors $\mathbf{u}_k = (u_{k1}, \dots, u_{kQ_k})$, which are assumed to be Gaussian variates with mean $\mathbf{0}$ and covariance matrix $\mathbf{I}_{Q_k}$ (*i.e.*, the $\mathbf{u}_k$ are independent and homoscedastic, $k = 1, \dots, K$).

Note the proposed model may assume different specifications whether the $\mathbf{D}_k$ and $\mathbf{B}_k$ matrix are constrained to be equal across gene clusters or not. The full range of possible constraints provides a class of 4 different models (see Table 1).

Model	Variable Member. Matrix	Error Matrix	Covariance Par.	Total Par.
CC	Constr.	Constr.	$2J$	$(K-1) + J(K+2)$
CU	Constr.	Unconstr.	$J(K+1)$	$(K-1) + J(2K+1)$
UC	Unconstr.	Constr.	$J(K+1)$	$(K-1) + J(2K+1)$
UU	Unconstr.	Unconstr.	$2KJ$	$(K-1) + 3KJ$

Table 1: Covariance structures derived from different constraints

The structure of $\mathbf{B}_k$ leads to a peculiar form of the component-specific covariance matrix of data $\mathbf{y}_i$ ($i = 1, \dots, n$). In fact, the covariance model adopted implies: (i) a block diagonal correlation structure; (ii) the correlation between variables depends on the variances only. In fact, the smaller the variance of one variable, the higher the correlation is among the other variables in the block. Assumption (i) allows a partition of the variables which is useful in many applications. For example in gene expression data this structure is called the clumpy dependence; assumption (ii) is specific to biclustering where we assume observations are homogeneous (with small within variance) only under limited block of conditions that therefore are highly correlated. To briefly show an example of this concept, suppose that the variable membership and error matrices of the generic k -th component of the extended mixture of factor analyzers model are as follows:

$$\mathbf{B}_k = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{pmatrix}$$

while error specific covariance matrix is:

$$\mathbf{D}_k = \begin{pmatrix} 0.1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0.01 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0.2 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 2 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 10 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 8 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 13 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 70 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 4 \end{pmatrix}.$$

The resulting component-specific covariance matrix $\mathbf{B}_k \mathbf{B}_k' + \mathbf{D}_k$ is a *block diagonal matrix*, i.e. a block matrix having on its main diagonal Q_k blocks formed by square matrices of size J_l ($l = 1, \dots, Q_k$ with $\sum_{l=1}^{Q_k} J_l = J$) such that the off-diagonal blocks are null matrices. In particular, all variables have a different variance, while the correlation between variables is equal to 1 if the variables are within the same variable cluster, otherwise is equal to 0.

In other words, the covariance matrix is a block-diagonal matrix of the form:

$$\mathbf{B}_k \mathbf{B}_k' + \mathbf{D}_k = \begin{pmatrix} 1.1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1.01 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1.2 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 3 & 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 11 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 9 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & 14 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 71 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 5 \end{pmatrix},$$

while the corresponding correlation matrix is:

$$\mathbf{C}_k = \begin{pmatrix} 1 & 0.95 & 0.87 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0.95 & 1 & 0.91 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0.87 & 0.91 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0.17 & 0.19 & 0.15 & 0 & 0 \\ 0 & 0 & 0 & 0.17 & 1 & 0.1005 & 0.08 & 0 & 0 \\ 0 & 0 & 0 & 0.19 & 0.10 & 1 & 0.09 & 0 & 0 \\ 0 & 0 & 0 & 0.15 & 0.08 & 0.09 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0.05 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.05 & 1 \end{pmatrix}.$$

Note this feature of the component-specific covariance matrix keeps with

the microarray biclustering problems, where we expected to have a huge amount of gene expression values co-regulated in small blocks.

Using the proposed model, we avoid having to compute the inverses of the estimated $J \times J$ specific covariance matrices $\mathbf{B}_k \mathbf{B}_k' + \mathbf{D}_k$ that may be singular for J large relative to n . This is because the inverse of the current $J \times J$ matrix $\mathbf{B}_k \mathbf{B}_k' + \mathbf{D}_k$ can be split up as:

$$(\mathbf{B}_k \mathbf{B}_k' + \mathbf{D}_k)^{-1} = \mathbf{D}_k^{-1} - \mathbf{D}_k^{-1} \mathbf{B}_k (\mathbf{I} + \mathbf{B}_k \mathbf{D}_k^{-1} \mathbf{B}_k)^{-1} \mathbf{B}_k' \mathbf{D}_k^{-1},$$

with the advantage to involve only the inverses of $Q_k \times Q_k$ matrices, since $\mathbf{D}_k$ is a diagonal matrix. This feature allows to fix either genes or tissues as observations (rows).

The main difference between the Ghahramani and Hinton (1996) and the proposed reparametrizations is that the former assumes the matrix $\mathbf{V}_k$ is a factor loading matrix, and therefore the resulting model is not strictly identifiable since $\mathbf{V}_k$ can be rotated without affecting the results, provided that the latent vectors $\mathbf{u}_k$ are counter-rotated. On the other hand, we assume that the matrix $\mathbf{B}_k$ is binary and row stochastic and this helps making the model identifiable. In fact, let $\mathbf{T}$ be an arbitrary $Q_k \times Q_k$ orthogonal matrix such that $\mathbf{T} \mathbf{T}' = \mathbf{T}' \mathbf{T} = \mathbf{I}_{Q_k}$. Then, consider the orthogonal transformation $\mathbf{B}_k^* = \mathbf{B}_k \mathbf{T}$. The transformed class-conditional covariance matrix is equal $\mathbf{B}_k \mathbf{B}_k' + \mathbf{D}_k$. In fact, the covariance structure of the transformed matrix has the form:

$$\mathbf{B}_k^* \mathbf{B}_k^{*'} + \mathbf{D}_k = \mathbf{B}_k \mathbf{T} \mathbf{T}' \mathbf{B}_k' + \mathbf{D}_k = \mathbf{B}_k \mathbf{B}_k' + \mathbf{D}_k.$$

Furthermore, matrix $\mathbf{B}_k$ is unique, but for labels permutation. In fact, $\mathbf{B}_k^*$ for definition must be binary and row stochastic, *i.e.*, $\mathbf{B}_k^* \mathbf{1}_{Q_k} = \mathbf{1}_J$ and $b_{jlk}^* \in (0, 1)$, but this is verified only if $\mathbf{T} = \mathbf{I}_{Q_k}$ or any permutation matrix which simply affects column clusters labels.

2.3 Parameter estimation

To estimate model parameters, we use an Alternating Expectation Conditional Maximization (AECM) algorithm derivated from the general form discussed in Meng and VanDyk (1997). This algorithm is an extension of the EM algorithm using different definitions of missing data at different stages. Let us start by partitioning the vector of unknown parameters ϕ in $(\phi_1; \phi_2)$; ϕ_1 contains the mixing proportions π_k , the elements of the component means μ_k . The subvector ϕ_2 contains the elements of $\mathbf{B}_k$, $\mathbf{D}_k$ ($k = 1, \dots, K$) and the specific-block means $\mathbf{u}_k$. One iteration of the AECM algorithm consists of two cycles:

one E-step (in the standard form for Gaussian mixture models) and one CM-step (one for each sub-vector of $\boldsymbol{\phi}$) for each cycle.

In the first cycle, we define $\mathbf{z}_i = (z_{i1}, \dots, z_{iK})$ as the missing data ($i = 1, \dots, n$; $k = 1, \dots, K$). The complete data log-likelihood function has the following form:

$$\begin{aligned} l_{C_1}(\mathbf{y}, \mathbf{z}) &= \sum_{i=1}^n \sum_{k=1}^K z_{ik} [\log \pi_k + \log N(\boldsymbol{\mu}_k, \mathbf{B}_k \mathbf{B}_k' + \mathbf{D}_k)] = \\ &= \sum_{i=1}^n \sum_{k=1}^K z_{ik} \left[\log \pi_k - \frac{J}{2} \log 2\pi - \frac{1}{2} \log |\mathbf{B}_k \mathbf{B}_k' + \mathbf{D}_k| - \right. \\ &\quad \left. - \frac{1}{2} \text{tr} \{ (\mathbf{y}_i - \boldsymbol{\mu}_k)(\mathbf{y}_i - \boldsymbol{\mu}_k)' (\mathbf{B}_k \mathbf{B}_k' + \mathbf{D}_k)^{-1} \} \right]. \end{aligned}$$

The expected complete-data log-likelihood is therefore

$$\begin{aligned} H_1(\boldsymbol{\mu}_k, \pi_k) &= \sum_{k=1}^K n_k \log \pi_k - \frac{nJ}{2} \log 2\pi - \\ &\quad - \frac{1}{2} \sum_{k=1}^K n_k \log |\mathbf{B}_k \mathbf{B}_k' + \mathbf{D}_k| - \\ &\quad - \sum_{k=1}^K n_k \text{tr} \{ \mathbf{S}_k (\mathbf{B}_k \mathbf{B}_k' + \mathbf{D}_k)^{-1} \}; \end{aligned}$$

where $n_k = \sum_{i=1}^n w_{ik}$ and w_{ik} is the expected value of z_{ik} given by

$$w_{ik} = \frac{\hat{\pi}_k N(\hat{\boldsymbol{\mu}}_k, \hat{\mathbf{B}}_k \hat{\mathbf{B}}_k' + \hat{\mathbf{D}}_k)}{\sum_{k=1}^K \hat{\pi}_k N(\hat{\boldsymbol{\mu}}_k, \hat{\mathbf{B}}_k \hat{\mathbf{B}}_k' + \hat{\mathbf{D}}_k)},$$

while

$$\mathbf{S}_k = \frac{\sum_{i=1}^n w_{ik} (\mathbf{y}_i - \boldsymbol{\mu}_k)(\mathbf{y}_i - \boldsymbol{\mu}_k)'}{n_k}$$

is the component-specific sample covariance matrix, $k = 1, \dots, K$.

Thus, at the first CM-step, maximizing the expected complete-data log-likelihood with respect to π_k and $\boldsymbol{\mu}_k$, we obtain:

$$\hat{\pi}_k = \frac{n_k}{n}$$

and

$$\hat{\boldsymbol{\mu}}_k = \frac{\sum_{i=1}^n w_{ik} \mathbf{y}_i}{n_k}.$$

In the second CM-step, we consider $\mathbf{z}_i = (z_{i1}, \dots, z_{iK})$ and $\mathbf{u}_k = (u_{k1}, \dots, u_{kQ_k})$ to be the missing data ($k = 1, \dots, K$). Therefore, the complete-data log-likelihood has the following form:

$$l_{C_2}(\mathbf{y}, \mathbf{z}, \mathbf{u}) = \sum_{i=1}^n \sum_{k=1}^K z_{ik} [\log \pi_k + \log N(\boldsymbol{\mu}_k + \mathbf{V}_k \mathbf{u}_k, \mathbf{D}_k) + \log f(\mathbf{u}_k)].$$

The expected complete-data log-likelihood is

$$\begin{aligned} H_2(\mathbf{B}_k, \mathbf{D}_k, \mathbf{u}_k) = C + \sum_{k=1}^K \left[\frac{n_k}{2} \log |\mathbf{D}_k^{-1}| - \frac{n_k}{2} \text{tr}\{\mathbf{D}_k^{-1} \mathbf{S}_k\} + \right. \\ \left. + \sum_{i=1}^n w_{ik} (\mathbf{y}_i - \boldsymbol{\mu}_k) \mathbf{D}_k^{-1} \mathbf{B}_k E(\mathbf{u}_k | \mathbf{y}_i, \boldsymbol{\mu}_k, \mathbf{D}_k, \mathbf{B}_k) - \right. \\ \left. - \frac{1}{2} \text{tr}\left\{ \mathbf{B}_k' \mathbf{D}_k^{-1} \mathbf{B}_k \sum_{i=1}^n w_{ik} E(\mathbf{u}_k \mathbf{u}_k' | \mathbf{y}_i, \boldsymbol{\mu}_k, \mathbf{D}_k, \mathbf{B}_k) \right\} \right], \end{aligned} \quad (2)$$

where C is a normalizing constant independent of $\mathbf{u}_k$, $\mathbf{B}_k$ and $\mathbf{D}_k$.

At the CM-step of the second cycle, we derive the estimates of $\mathbf{D}_k$, $\mathbf{u}_k$ and $\mathbf{B}_k$ maximizing expression (2) considering one of the possible models in Table 1.

Here, we focus on the most general model, UU, where no constraints are imposed on $\mathbf{D}_k$ and $\mathbf{B}_k$. Differentiating $H_2(\mathbf{B}_k, \mathbf{D}_k, \mathbf{u}_k)$ with respect to $\mathbf{D}_k^{-1}$ and solving the corresponding likelihood equation, we obtain:

$$\hat{\mathbf{D}}_k = \text{diag}\{\mathbf{S}_k - \mathbf{B}_k \mathbf{L}_k \mathbf{S}_k\},$$

where $\mathbf{L}_k = \mathbf{B}_k' (\mathbf{B}_k \mathbf{B}_k' + \mathbf{D}_k)^{-1}$.

As far as the estimates of $\mathbf{u}_k$ are concerned ($k = 1, \dots, K$), let us consider the joint Gaussian distribution of $(\mathbf{y}_i, \mathbf{u}_k)$ given the membership vector z_{ik} :

$$\begin{bmatrix} \mathbf{y}_i \\ \mathbf{u}_k \end{bmatrix} | z_{ik} \sim MVN \left[\begin{pmatrix} \boldsymbol{\mu}_k \\ \mathbf{0} \end{pmatrix}, \begin{pmatrix} \mathbf{B}_k \mathbf{B}_k' + \mathbf{D}_k & \mathbf{B}_k \\ \mathbf{B}_k' & \mathbf{I}_{Q_k} \end{pmatrix} \right].$$

To find a feasible estimator for $\mathbf{u}_k$, we use the posterior mean of $\mathbf{u}_k$ given $\mathbf{y}_i$ and $z_{ik} = 1$ ($k = 1, \dots, K$; $i = 1, \dots, n$) which is equal to

$$\hat{\mathbf{u}}_k = E(\mathbf{u}_k | \mathbf{y}_i, z_{ik} = 1) = \frac{\mathbf{B}_k' (\mathbf{B}_k \mathbf{B}_k' + \mathbf{D}_k)^{-1} \sum_{i=1}^n w_{ik} (\mathbf{y}_i - \boldsymbol{\mu}_k)}{n_k}.$$

This estimator closely resembles the Thomson estimator (1951) of factor scores.

As far as the estimation of $\mathbf{B}_k$ is concerned, we choose the unit value in each column as follows

$$b_{jl} = \begin{cases} 1 & \text{if } H_2(\cdot, b_{jl} = 1) = \max_h H_2(\cdot, b_{jh} = 1) \\ 0 & \text{otherwise} \end{cases}$$

with $j = 1, \dots, J$, $l, h = 1, \dots, Q_k$ and $l \neq h$.

Thus, the AECM algorithm consists of the following steps:

1. Choose initial values for the parameter vector.
2. First cycle
 - (a) E-step: Update w_{ik} ($i = 1, \dots, n$ and $k = 1, \dots, K$);
 - (b) CM-step: Update π_k and $\boldsymbol{\mu}_k$ ($k = 1, \dots, K$).
3. Second cycle
 - (a) E-step: Update w_{ik} ($i = 1, \dots, n$ and $k = 1, \dots, K$);
 - (b) CM-step: Update $\mathbf{S}_k$, $\mathbf{L}_k$, $\mathbf{B}_k$, $\mathbf{D}_k$ and $\mathbf{u}_k$ ($k = 1, \dots, K$) in turn.
4. Compute the log-likelihood function for the current parameter values. If the function increase is larger than a fixed threshold, iterate once more according to 2. Otherwise, the process has converged.

3 Applications

3.1 Simulation study

In this section we provide a simulation study to show the finite sample properties of the proposed model. It is evaluated in terms of recovering row and column partitions by using the Modified Rand Index (Hubert and Arabie, 1985) that measures concordance between the true and the estimated partitions. Model parameters are estimated by using the AECM algorithm described in the previous section. In particular, the performance of the algorithm are evaluated using the following measures:

1. $\text{Mrand}(\mathbf{Z}, \mathbf{W}^*)$, i.e. the Modified Rand Index between $\mathbf{Z} = \{z_{ik}\}$, the true matrix of row cluster membership and the crispy estimated matrix $\mathbf{W}^* = \{w_{ik}^*\}$ ($k = 1, \dots, K$), where each element w_{ik}^* is defined as:

$$w_{ik}^* = \begin{cases} 1 & \text{if } k = \operatorname{argmax}_{k=1, \dots, K} w_{ik}, \\ 0 & \text{otherwise} \end{cases};$$

2. $\text{Mrand}(\mathbf{B}_k^*, \mathbf{B}_k)$, i.e. the Modified Rand Index between the true matrix of column cluster membership $\mathbf{B}_k$ and the estimated matrix $\hat{\mathbf{B}}_k$.

In the simulation setting, the algorithm starts with random matrices $\mathbf{Z}$ and $\mathbf{B}_k$. To avoid to be trapped into local maxima, we fixed the number of random starts equal to 10, retaining the solution with the best fit in term of maximized log-likelihood. This number of random starts is sufficient to significantly reduce the probability to obtain a local maximum, although when a very large error is present in the data, the number of starting points should be increased. Then, we generate the specific-block mean vectors $\mathbf{u}_k$ to ensure specific-block mean vectors within a row cluster is sufficiently far from the average of specific-block mean vectors corresponding to other clusters. Three error levels (Low, Medium, High) have been considered in order to have different levels of homogeneity within biclusters, by multiplying the matrix $\mathbf{D}_k$, by 5, 10 and 50, respectively ($k = 1, \dots, K$). Finally, we generate $R = 100$ samples according to the proposed model in a $J = 80$ dimensional space with a number of units equal to $n = 1000$ considering two different situations:

1. the data matrix is formed by 5 biclusters defined by a partition of units with $K=2$ clusters and $Q_1=3$, $Q_2 = 2$ variable clusters;
2. the data matrix is formed by 3 biclusters defined by a partition of units with $K=2$ clusters and $Q_1=2$, $Q_2 = 1$ variable clusters.

For each data set, biclusters have been randomly placed within the data matrix by permuting rows and columns and the AEEM algorithm has been applied to recover biclusters and, specifically, partitions of units and variables generating the data (Figure 1 and 2). Tables 2 and 3 show the average Mrand indices for every experiment.

As it can be observed, the algorithm performs well in recovering the true partitions of units and variables under all levels of error. Obviously, the algorithm performs better in recovering both partitions when the error level is low (see row 1, Tables 2-3); by increasing the error level, the average Mrand index for both unit and variable partitions is still significantly newsworthy.

Error level	Average Mrand for unit partitions	Average Mrand for variable partitions		Average number of iterations
		$\mathbf{B}_1$	$\mathbf{B}_2$	
Low	1.00	1.00	0.98	2.78
Medium	0.94	0.96	0.93	3.50
High	0.80	0.89	0.70	5.05

Table 2: Simulation results for the experiment with $K=2$, $Q_1 = 3$, $Q_2 = 2$

Error level	Average Mrand for unit partitions	Average Mrand for variable partitions	Average number of iterations
Low	1.00	1.00	2.69
Medium	0.90	0.98	3.00
High	0.78	0.60	7.03

Table 3: Simulation results for the experiment with $K=2$, $Q_1 = 2$, $Q_2 = 1$

As far as the computational times in both experiments are concerned, we can observe that the algorithm converges quite fastly even if it start with random matrices $\mathbf{Z}$ and $\mathbf{B}_k$. In order to further speed up its convergence, these matrices could be chosen in a rationale way (*e.g.*, using some clustering method, *i.e.* the k -means algorithm).

3.2 Real data example

We analyze the benchmark data set on cutaneous melanoma described by Bittner et al. (2000). The original aim of this study is to determine whether or not molecular profiles generated by cDNA microarrays could be used to identify distinct subtypes of cutaneous melanoma, a malignant neoplasm of the skin. The data consists of 38 samples from tissue biopsies and tumor cell lines, with 31 cutaneous melanomas and 7 controls; samples come from male and female patients aged 29 to 75, with 3 patients of unknown age. The mRNA has been extracted and Cy5-labelled cDNA is created for the 31 cutaneous melanoma and the 7 control samples; a single reference probe, labelled Cy3, is used for all the 38 samples. The Cy5 and Cy3-labelled cDNAs

are mixed for each sample and hybridized to a separate melanoma microarray. The hybridized array has been scanned using both red and green lasers, and the resulting image has been analyzed. 3613 over 8150 cDNAs have been identified as adequately measured, and gene expression ratios of Cy5/Cy3 are calculated. We directly downloaded the ratio expression levels of the just filtered 3613 genes from the web site associated with the Bittner et al. paper. To avoid distortions of the data resulting from ratios where the signal in one channel (Cy5 or Cy3) is large, and the signal in the other channel is undetectable, ratios greater than 50 and lower than 0.02 are truncated to 50 and 0.02, respectively; the ratios are transformed to a logarithm scale (base 2) and normalized by subtracting the median log-ratio for that tissues, so that the median log-ratio within a tissue is zero. No normalization is performed across tissues, since a single reference probe was used for all of them. Bittner et al. (2000) discuss the analysis of 31 samples, excluding the 7 control samples; they are considering average linkage clustering on these 31 samples by employing one minus the Pearson correlation coefficient between log-ratios as a dissimilarity measure between two tissues. In this way, they obtained two clusters of 12 and 19 samples, which have been further validated by multidimensional scaling (MDS) and through CAST, a non-hierarchical clustering algorithm (Ben-Dor, Shamir and Yakhini, 1999). Both MDS and CAST identify the same major cluster of 19 samples found by the average linkage.

Here, we focus on biclustering of genes and tissue samples, by using the proposed model. We denote the observed gene expression levels by $\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_n$, where $\mathbf{y}_i$ is a 31-dimensional vector representing the expression level of the i -th gene on 31 cutaneous melanoma samples. Thus, we have $J = 31$ and $n = 3613$.

We fit the proposed model for different numbers of genes and tissues clusters. For each pair, we run the AEEM several times to avoid local maxima, choosing the best solution through BIC and AIC criteria. In details, we obtain that the best choice of number of gene clusters is equal to 4: up-regulated (1164 genes), down-regulated (781 genes), discriminating (238 genes) and no-regulated (1430 genes) gene clusters; where we define as discriminating the genes which are differentially expressed in the two tissue clusters. In fact, only for these discriminating genes we obtain that the best choice is 2 tissue clusters, while any partition for the other gene clusters. In other words, the best solution corresponds to 4 gene clusters in the data matrix, *e.g.* 3 clusters represent not-discriminating genes, and the latter (the smallest) corresponds to discriminating genes (Figure 3). The resulting bicluster formed by 238 genes within two clusters of 12 and 19 tissue samples, coincides with the class distinction that was detected and identified as biologically meaningful in Bittner

et al. (2000).

We may notice that the same data set has been investigated by other authors using different approaches. For instance, Goldestein et al. (2002) use several hierarchical clustering methods, Heydebreck et al. (2001) propose the ISIS method (Identifying Splits with Clear Separation) and Rocci and Vichi (2007) propose an extension of the standard K -means algorithm to two dimensions. Rocci and Vichi (2007) and Goldestein et al. (2002) get a partition with 21 and 10 tissue samples. As far as gene clusters are concerned, while the former obtain 4 gene clusters with only one (formed by 98 genes) containing differentially expressed genes, the latter obtain 182 discriminant genes. On other hand, Heydebreck et al. (2001) focus only on tissues clustering and obtain a partition of tissues coinciding with our and Bitter et al class distinction, except that their method suggests to reassign one sample to the cluster of 19 samples.

Moreover, we performed comparative experiments to some existing biclustering methods making advantages from the freely available BicAT toolbox (Barkow et al., 2006), a software platform that integrates various biclustering techniques. In details, we choose to run these methods that do not need a priori number of gene and tissue clusters, but it is defined within the algorithm itself. In particular, the Order-preserving Submatrix Algorithm, proposed by Ber-Dor et al. (2003), yields 16 gene clusters of size 4-5-15-17-18-32-88-92-95-98-101-125-185-388-952-2075 corresponding to subsets of tissues given by 17-16-15-14-13-12-11-10-9-8-7-6-5-4-3-2, respectively. The Iterative Signature Algorithm (Ihmels et al., 2002), instead, reaches 9 genes clusters (100-87-74-71-67-64-64-61-51) corresponding to small tissue clusters of size 2 (one for each of them). We can observe that tissue samples and genes may be grouped in different ways, according to different biological factors and on the molecular point of view these different groupings may correspond to expression patterns in different, relatively small biclusters. Therefore, it is usually difficult to see how one may be capable of inferring a biological basis for the observed biclusters. However, by using the proposed model, we achieve a partition of tissue samples in 2 groups which previous investigations identify with differing invasive qualities of the cutaneous melanomas, and also obtain that only few up-regulated genes seem to be highly significant in explaining the tissue partition.

3.3 Model choice

In the standard non-hierarchical clustering methods a single choice of the number of clusters for the row (objects) has to be taken. Here, the model choice is

more complex: after the number K of row clusters is specified, it is necessary to choose the number of column clusters Q_k , $k = 1, \dots, K$. Since automatic model choice are in general desirable, we used AIC and BIC information criteria; they penalize the log-likelihood function with the number of free parameters involved in the model. AIC is the Akaike Information Criterion defined as $-2\log(L) + 2p$, where L is the maximized likelihood and p is the number of model parameters. BIC has a different penalization, *i.e.*, $-2\log(L) + \log(n)p$, where n is the number of observations; BIC in general gives a stronger penalization (for $n \geq 8$) than AIC and tends to choose more parsimonious models.

4 Concluding remarks

In recent years, a growing attention has been paid to discover local structures composed by sets of genes that show coherent expression patterns across subsets of experimental conditions. Several biclustering techniques for analyzing microarray data have been proposed (see Madeira and Oliveria, 2004 for a review). Most frequently used techniques for biclustering are heuristically motivated. The contribution of this paper is to extend the mixtures of factor analyzers introduced by Ghahramani and Hinton (1996) to allow for simultaneous clustering of genes and tissue samples. Specifically, we introduce variable clustering into the mixture of factor analyzers, replacing the standard (component-specific) factor loadings matrix with a binary and row stochastic matrix representing variable membership; the cluster of variables is defined with row clusters identified by a standard finite mixture. This leads to a peculiar shape of the component-specific covariance matrix which keeps with the microarray biclustering problem. To estimate model parameters, we use the Alternating Expectation Conditional Maximization (AECM) algorithm (Meng and VanDyk, 1997). Applications on real data have shown that our proposal helps significantly the biological understanding of resulting biclusters. Finally, it can be highlighted that although we focus on a specialized application area which is microarray expression level modeling, the proposed model may be useful in many application areas, as for instance, text web mining, customer satisfaction or, in general, for all contexts, where the aim is to identify blocks (or biclusters), which satisfy some specific characteristics of homogeneity.

Figure 1: (1000×80) data matrices with $K=2$, $Q_1 = 3$, $Q_2 = 2$. Three error levels a), b) and c) are considered. Entries are randomly permuted

a) matrix in 5 biclusters, low error

a2) permuted matrix a)

b) matrix in 5 biclusters, medium error

b2) permuted matrix b)

c) matrix in 5 biclusters, high error

c2) permuted matrix c)

Figure 2: (1000×80) data matrices with $K=2$, $Q_1 = 2$, $Q_2 = 1$. Three error levels a), b) and c) are considered. Entries are randomly permuted

a) matrix in 3 biclusters, low error

a2) permuted matrix a)

b) matrix in 3 biclusters, medium error

b2) permuted matrix b)

c) matrix in 3 biclusters, high error

c2) permuted matrix c)

Figure 3: Best solution for Bittner et al. (2000) data: $K = 4$, $Q_1 = Q_3 = Q_4 = 1$, $Q_2 = 2$

References

- Barkow S., Bleuler S., Prelic A., Zimmermann P. and Zitzler E. (2006) BicAT: a biclustering analysis toolbox. *Bioinformatics*, 22(10):1282-1283.
- Ben-Dor A., Shamir R. and Yakhini Z. (1999) Clustering gene expression patterns. *Journal Comput. Biol.*, 6(3-4) 281-97.
- Ben-Dor A. et al. (2003) Discovering local structure in gene expression data: the order-preserving submatrix problem. *J. Comput. Biol.*, 10 373384.
- Bittner M. et al. (2000) Molecular classification of cutaneous malignant by gene expression profiling. *Nature*, 406 (6795) 536-540.
- Cheng Y. and Church G.M. (2000) Biclustering of expression data. *Proceedings of the Eighth International Conference on Intelligent Systems for Molecular Biology (ISMB)*, 8:93 103.
- Ghahramani Z. and Hinton G.E. (1996) The EM algorithm for mixture of factor analyzers. *Technical Report*, CRG-TR-96-1, 8, University of Toronto.
- Goldstein D., Ghosh D. and Conlon E. (2002) Statistical issues in the clustering of gene expression data. *Stat. Sinica*, 12 219241.

- Govaert G. and Nadif M. (2003) Clustering with block mixture models. *Pattern Recognition*, 36(2): 463-473.
- Hartigan J.A. (1972) Direct clustering of a data matrix. *Journal Am. Stat. Assoc*, 67 123-129.
- Hartigan J.A. (1975) Clustering algorithms. New York: York Wiley.
- Heydebreck A.V., Huber W., Poustka A., Vingron M. (2001) Identifying splits with clear separation: a new class discovery method for gene expression data. *Bioinformatics*, 17 107-114.
- Hubert L. and Arabie P. (1985) Comparing partitions. *Journal of Classification*, 2 193-218.
- Ihmels J. et al. (2002) Revealing modular organization in the yeast transcriptional network. *Nat. Genet.*, 31 370377.
- Li J. and Zha H. (2006) Two-way Poisson mixture models for simultaneous document classification and word clustering. *Computational Statistics and Data Analysis*, 50(1) 163-180.
- Madeira Sara C. and Oliveira Arlindo L. (2004) Biclustering Algorithms for Biological Data Analysis: A Survey. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 1 24-45.
- Martella F. (2006) Classification of microarray data with factor mixture models. *Bioinformatics*, 22(2): 202-208.
- McLachlan G.J. and Peel D. (2000) Mixtures of factor analyzers. In: Langley, P. (ed) Proceedings of the Seventeenth International Conference on Machine Learning. Morgan Kaufmann, San Francisco.
- McLachlan G.J. et al (2002) A mixture model-based approach to the clustering of microarray expression data. *Bioinformatics*, 18: 413-422.
- Meng X.L. and Van Dyk D.A. (1997) The EM algorithm -an old folk song sung to a fast new tune. *Journal R. Statist. Soc.*, B 59 511-567.
- Prelic A. et al. (2006) A systematic comparison and evaluation of biclustering methods for gene expression data. *Bioinformatics*, 22, 11221129.
- Rocchi R. and Vichi M. (2002) A two-way model for simultaneous reduction and classification. *Atti della XLI Riunione Scientifica, Universit di Milano-Bicocca* (5th-7th June).

Rocci R. and Vichi M. (2007) Two-mode multi-partitioning. *Computational Statistics and Data Analysis*, in press.

Tipping M.E. and Bishop C.M. (1999) Mixtures of probabilistic principal component analysers. *Neural Computation*, 11(2) 443-482.

Thomson G.H. (1951) The factor analysis of human ability. (5th ed.) Boston: Houghton Mifflin Company.

Van Mechelen I., Bock H.H., De Boeck P. (2004) Two-mode clustering methods: A structured overview. *Statistical Methods in medical Research*, 13 363-394.