Model-based approaches to synthesize microarray data: a unifying review using mixture of SEMs

F. Martella[*] J.K. Vermunt [†]

Abstract

Several statistical methods are nowadays available for the analysis of gene expression data recorded through microarray technology. In this paper, we take a closer look at several Gaussian mixture models which have recently been proposed to model gene expression data. It can be shown that these are special cases of a more general model, called mixture of structural equation models (mixture of SEMs), which has been developed in psychometrics. This model combines mixture modeling and SEMs by assuming that component-specific means and variances are subject to a structural equation model. The connection with SEM is useful for at least two reasons: 1) it shows more explicitly the basic assumptions of existing methods, and 2) it helps in straightforward development of alternative mixture models for gene expression data with

[*] corresponding author: Francesca Martella, Dipartimento di Scienze Statistiche, Sapienza University of Rome, P.le Aldo Moro, 5 - I 00185 Rome (Italy). E-mail: francesca.martella@uniroma1.it

[†] Department of Methodology and Statistic, Tilburg University, Tilburg, The Netherlands. E-mail: j.k.vermunt@uvt.nl

alternative mean/covariance structures. Different specifications of mixture of SEMs for clustering gene expression data are illustrated using two benchmark datasets.

Keywords: Mixture of SEMs, microarray data, biclustering, simultaneous clustering and dimensional reduction, correlated data.

1 Introduction

DNA microarray techniques are used to measure the expression levels of thousands of genes simultaneously. From a statistical point of view, microarray data are a collection of real numbers (representing expression levels) arranged in a $(n \times J)$ data matrix $\mathbf{Y}$, where, usually, rows and columns represent genes and experimental conditions. The latter may refer to different time points, environmental conditions, cellular states, organs, treatments, tissue samples, and so on. In most cases, the generic element of this matrix, y_{ij}, represents the log-ratio between the j-th experimental condition and a reference condition for the i-th gene ($i = 1, ..., n$; $j = 1, ..., J$). Summarizing biologically relevant information from such high dimensional data is a challenging task. Among others, several clustering techniques have been proposed for this purpose; from a biological point of view, in fact, it is often meaningful to cluster genes and/or experimental conditions (that is, rows and columns of the data matrix). In a clustering context, genes (rows) with similar expression patterns (across conditions) are assumed to belong to the same cluster of genes; conversely, experimental conditions (columns) with similar expression patterns (across genes) are assumed to belong to the same cluster of conditions.

Conditional on the biological question, attention is focus on one of the following three types of microarray experimental designs: 1. genes are measured on different tissue samples (or different treatments); 2. genes are measured at different time points (time-course series); 3. gene measurements are replicated on each tissue sample (technical replicates) at different time points. In the first design, it may be useful to cluster genes with similar expression levels to predict gene functions or identify sets of genes that are regulated by the same organism. In this context, it may also be interesting to define clusters of tissue samples reflecting biological categories such as cell tumor progression, mutational status of a disease, and so on. In particular, if tissue samples belong to known groups of tissues (for example, normal and abnormal tissues or known subtypes of a disease), it may be interesting to find genes which are associated with disease status (clinical markers), that is, genes that are differentially expressed in the observed experimental conditions. A more complex situation is related to discovering local expression patterns; in this case, clusters of genes may show similar activation patterns under a class of experimental conditions, which is not defined a priori. The task is here to joint clusters experimental conditions and genes, where cellular processes are active only in a subset of the experiments and groups of genes participate in the cellular process of interest. The second design arises since biological systems are dynamic; thus, microarray experiments performed over time give us a way to observe cellular mechanisms in action. Analyses in this area have focused on identifying genes with differential expressions over time, to identify clusters of genes with similar time profiles (co-expressed genes) or causal networks generating the observed data. The last experimental design is the most complex one; here, multiple measurements are taken over time usually for a different purpose than time-course data. More specifically, mRNA from a tissue sample is hybridized several times to reduce the uncertainty about gene expression levels and the effect of measurement error. It can be expected that the repeated measurements will also increase robustness of clustering results. In terms of the possible purpose of the analysis, this design may be considered as more similar to the first than to the second one.

This paper focuses on cluster analysis via mixture models. Model-based clustering dates back at least to [1], [2], [3], [4], [5],[6], [7], [8] and [9] and became popular in microarray data analysis thanks to the works of [10], [11] and [12]. Model-based clustering approaches are based on the assumption that the data come from an underlying finite mixture model, where each mixture component corresponds to a cluster. Finite mixtures of Gaussian densities are by far the most commonly used structure in model-based clustering. The goal of model-based clustering is to provide a partition of the data into clusters of homogeneous observations; to achieve this, after model fitting model-based clustering requires an additional step to assign each observation to a different cluster according to some pre-specified rule.

Whereas this approach offers various advantages compared to classical clustering techniques [13], its use in gene expression data applications may face specific questions. Problems occur when the aim is to cluster tissue samples, since the number of objects to be clustered (tissue samples) is usually much smaller than the number of variables (genes). In this case, the standard mixture model should be adapted to prevent singularity of component-specific covariance matrices. A further problem may arise when clustering genes in time series or repeated measurement contexts, due to the potential dependence among gene profiles which may also vary across tissue samples. Recent examples of model-based clustering for these kinds of microarray data are discussed in [14] and [15]. A third limitation of standard model-based clustering methods is that they are not designed to cluster simultaneously genes and conditions, i.e. to provide a biclustering structure. Extensions of standard mixture models have been recently developed for this purpose by [16] and [17].

The aim of this paper is to show that most model-based clustering methods developed in the microarray data context can be seen as special cases of a finite mixture of structural equation models (Mixture of SEMs). This framework has been developed in the field of psychometrics, quite independently from other fields of applied statistics. We take the mixture of SEMs as the starting point, and show in the next sections how to unify and extend a wide variety of statistical methods proposed for model-based clustering of microarray data. Additionally, we show how these are connected with the more general mixture of SEMs model to discover interesting relationships among previously separate methods with possibly different biological purposes becoming visible. The mixture of SEMs framework may help statisticians develop more flexible and realistic models, as well as make analyses easier, since biologists/genetists can focus on the applied rather than on the theoretical side.

The rest of the paper is structured as follows. In section 2, we introduce the mixture of SEMs framework. Section 3 proposes different specifications of mixture of SEMs to cluster tissue samples and/or to apply dimensional reduction on genes, including mixture of factor analyzers (MFA), its extensions (PGMM1-PGMM8, EPGMM1-EPGMM4), mixture of probabilistic principal component analyzers (MPCA), and two-way models for simultaneous reduction and classification introduced by

[18] (RV). Section 4 shows how mixtures of SEMs are related to two-way mixtures of factor analyzers proposed by [19] for the simultaneous clustering of genes and tissue samples (Biclustering). In section 5, we highlight connections between mixture of SEMs and mixture models for clustering genes where repeated time-course data have to be handled, focusing on ideas developed by [20] (LMM1). A more refined specification of mixture of SEMs is presented in section 6, where the linear mixed models proposed by [21] (LMM2, LMM3 and LMM4) to deal with correlated genes (in different experimental designs) is shown to be a special case of a mixture of SEMs. The mixture of SEMs framework is illustrated by discussing the analysis of two benckmark datasets from the microarray literature: the yeast galactose data of [22] and the colorectal carcinoma data of [23]. Last section presents concluding remarks.

2 Modeling framework: The mixture of SEMs

Let us take as starting point the structural equation model (SEM). Structural equation modeling is mainly used to assess relations among manifest and latent variables; given its generality, it may be considered as nesting many multivariate techniques such as multiple regression, path analysis, confirmatory factor analysis, etc. Readers unfamiliar with common factor modelling, are referred to the recent overview by [24].

A SEM is composed of two parts, a measurement part relating the observed outcomes to the latent variables and a structural part that specifies the relationships among the latent variables. Even though it is very popular in some statistical areas, SEM is not commonly used in mainstream statistics, since it often lacks a strict formalization (because of its widespread flexibility) and a standardized terminology (because of the different application fields). About ten years ago, several authors proposed modelling unobserved heterogeneity through a general model by unifying SEM and finite mixture model for clustering ([25], [26], [27], [28], [29] and [30]). This hybrid approach is called mixture of SEMs. In detail, it is defined as a finite mixture model with a structural equation model describing component-specific parameters. In particular, when a Gaussian distribution is considered, a SEM can be used to constrain component-specific mean vectors and/or covariance matrices. This approach may lead to parsimonious Gaussian mixture models that are more stable with respect to models with unrestricted parameter sets. The aim of linking mixture models and SEMs is not only to avoid singularities in the estimation phase, but also to test hypotheses concerning the relationship among variables within component-specific distributions. Below, we first introduce standard mixture of SEMs. Then, we show how the mixture SEM framework can be used to analyze gene expression data (by imposing specific constraints); that is, how it can be linked to already known models for model-based clustering of gene expression data.

2.1 Specification of a mixture of SEMs

Let $\mathbf{y}_i$ $(i = 1, ..., n)$ be a J-dimensional data vector from a finite mixture model. The marginal density function is defined by:

$$f(\mathbf{y}_i; \boldsymbol{\phi}) = \sum_{k=1}^{K} \pi_k f(\mathbf{y}_i; \boldsymbol{\theta}_k), \tag{1}$$

where π_k represent the k-th component weight, $\pi_k \geq 0$, for $k = 1, ..., K$, with $\sum_{k=1}^{K} \pi_k = 1$, $f(\mathbf{y}_i; \boldsymbol{\theta}_k)$ represents the component-specific density indexed by the parameter vector $\boldsymbol{\theta}_k$ ($k = 1, ..., K$), while $\boldsymbol{\phi} = \{\pi_k, \boldsymbol{\theta}_k\}_{k=1,...,K}$ is the "global"parameter vector. We will assume that the component-specific density is Gaussian, where a given hierarchical structure will be used to define the parameters $\boldsymbol{\theta}_k$. As shown in [28], conditional on the i-th unit belonging to the k-th component, the SEM can be defined as follows:

$$\mathbf{y}_i = \boldsymbol{\upsilon}_k + \mathbf{V}_k \mathbf{u}_{ik} + \mathbf{e}_{ik}, \tag{2}$$

$$\mathbf{u}_{ik} = \boldsymbol{\alpha}_k + \mathbf{B}_k \mathbf{u}_{ik} + \boldsymbol{\xi}_{ik}. \tag{3}$$

The first equation represents the *measurement part* of the model, where $\boldsymbol{\upsilon}_k$ represents a J-dimensional component-specific vector of intercepts, $\mathbf{V}_k$ is a $J \times Q_k$ component-specific factor loadings matrix, $\mathbf{u}_{ik}$ is a Q_k-dimensional, component-specific, latent variable vector, $\mathbf{e}_{ik}$ is a J-dimensional, component-specific, measurement error vector ($k = 1, ..., K$, $i = 1, ..., n$). The second equation defines the relationships among the latent variables (referred to as the *structural part* of the model); here, $\boldsymbol{\alpha}_k$ is a Q_k-dimensional component-specific vector of intercepts, $\mathbf{B}_k$ is a $Q_k \times Q_k$ component-specific regression coefficient matrix (with null diagonal elements), while $\boldsymbol{\xi}_{ik}$ is a Q_k-dimensional component-specific residual vector ($k = 1, ..., K$). It is worth noticing that $\mathbf{B}_k$ should be defined as such that $(\mathbf{I} - \mathbf{B}_k)$ is non singular. Moreover, $\mathbf{e}_{ik}$ and $\boldsymbol{\xi}_{ik}$ are assumed to be mutually independent. The most general model is based on the following additional assumptions:

- $\mathrm{E}(\mathbf{e}_{ik}) = \mathbf{0}$, $\mathrm{cov}(\mathbf{e}_{ik}) = \mathbf{D}_k$ and $\mathbf{e}_{ik} \sim N(\mathbf{0}, \mathbf{D}_k)$;

- $\mathrm{E}(\boldsymbol{\xi}_{ik}) = \mathbf{0}$, $\mathrm{cov}(\boldsymbol{\xi}_{ik}) = \boldsymbol{\Phi}_k$ and $\boldsymbol{\xi}_{ik} \sim N(\mathbf{0}, \boldsymbol{\Phi}_k)$;

where $\mathbf{D}_k$ and $\mathbf{\Phi}_k$ represent component-specific covariance matrices; these matrices may be non-diagonal. It follows that the mixture component densities in (1) are J-variate Normal density functions with parameters $\boldsymbol{\theta}_k = \{\boldsymbol{\mu}_k, \mathbf{\Sigma}_k\}$ $(k = 1, ..., K)$. To define the corresponding component-specific means and covariances, let us consider a set of K indicator variables z_{ik} of component membership: $z_{ik} = 1$ if the i-th unit belongs to the k-th cluster, 0 otherwise. Thus, component-specific mean vectors and covariance matrices can now be expressed as follows

$$\mathrm{E}(\mathbf{y}_i|z_{ik} = 1) = \boldsymbol{\mu}_k = \boldsymbol{\upsilon}_k + \mathbf{V}_k\mathrm{E}(\mathbf{u}_{ik}) + \mathrm{E}(\mathbf{e}_{ik}) = \boldsymbol{\upsilon}_k + \mathbf{V}_k(\mathbf{I} - \mathbf{B}_k)^{-1}\boldsymbol{\alpha}_k,$$

$$\mathrm{cov}(\mathbf{y}_i|z_{ik} = 1) = \mathbf{\Sigma}_k = \mathbf{V}_k\mathrm{cov}(\mathbf{u}_{ik})\mathbf{V}_k' + \mathbf{D}_k = \mathbf{V}_k(\mathbf{I} - \mathbf{B}_k)^{-1}\mathbf{\Phi}_k(\mathbf{I} - \mathbf{B}_k')^{-1}\mathbf{V}_k' + \mathbf{D}_k.$$

It should be noted that the structural parameters in the structure for the component-specific covariance matrix in models (1)-(3) are not identifiable. A common method to identify the component-specific covariance matrix is to fix the elements in $\mathbf{V}_k$ and/or $\mathbf{B}_k$ to constant values. In this paper, we restrict ourselves to situations where $\mathbf{B}_k = \mathbf{0}$, since, as far as we know, models with alternative specifications for $\mathbf{B}_k$ have not been proposed in the microarray context yet. However, by taking the more general formulation as a starting point, it is possible to define several useful extensions; for example, in the clustering of time-course data, it may be relevant to impose a Markov-type structure on time-specific latent factors.

2.2 Parameter estimation and model selection

even though this paper does not focus on estimation, we believe that a brief description of parameter estimation and model selection is mandatory. The most general mixture of SEMs (1)-(3) has a total of

$JK + JQ_kK + \frac{J(J+1)K}{2} + Q_kK + (Q_k^2 - Q_k) + \frac{Q_k(Q_k+1)K}{2} + (K-1)$ parameters; in particular, terms $\frac{J(J+1)K}{2}$ and $\frac{Q_k(Q_k+1)K}{2}$ refer to the component-specific covariance matrices $\mathbf{D}_k$ and $\mathbf{\Phi}_k$, JK and Q_kK to the component-specific intercept vectors $\boldsymbol{\upsilon}_k$ and $\boldsymbol{\alpha}_k$, JQ_kK and $Q_k^2 - Q_k$ to the component-specific factor loadings and regression parameter matrices $\mathbf{V}_k$ and $\mathbf{B}_k$, and $(K-1)$ to the mixing weights π_k. For parameter estimation, the same approaches used for standard finite mixture models may be used. As discussed in [13], these include graphical methods, methods of moments, minimum-distance methods, maximum likelihood and Bayes methods. However, the maximum likelihood (ML) framework is the most commonly used approach to fit a mixture of SEMs. The optimization of the likelihood function wrt model parameters is mostly based on the Expectation-Maximization (EM) algorithm [31] with the Newton-Raphson algorithm [32].

In the clustering context, the interest is not only on model parameter estimates but also on posterior probabilities of component membership, w_{ik}, defined by:

$$\Pr(z_{ik} = 1|\mathbf{y}; \boldsymbol{\phi}) = w_{ik} = \frac{\hat{\pi}_k \varphi(\mathbf{y}_i; \hat{\boldsymbol{\mu}}_k, \hat{\mathbf{\Sigma}}_k)}{\sum_{k=1}^{K} \hat{\pi}_k \varphi(\mathbf{y}_i; \hat{\boldsymbol{\mu}}_k, \hat{\mathbf{\Sigma}}_k)} \tag{4}$$

where $\hat{\pi}_k, \hat{\boldsymbol{\mu}}_k, \hat{\mathbf{\Sigma}}_k$ represent component-specific parameters estimated at the current iteration of the EM algorithm. To allocate each unit in the sample a MAP (Maximum A Posteriori) rule could be used, where each observation is assigned to the cluster corresponding to the highest w_{ik} $(k = 1, ..., K)$. In the following, we will assume that the number of mixture components K and latent variables Q_k $(k = 1, ..., K)$ is fixed. However, in practice, these are unknown quantities which must be estimated from observed data. One of the advantages of model-based clustering compared to standard clustering algorithms is that it provides a more formal basis to choose the number of clusters; in this case, the choice of K or Q_k involves comparison of possible models. This procedure is common to standard finite mixture models, where the number of components is usually determined by penalized likelihood methods (AIC, BIC, AWE, etc.; see [33] for details). Although there is a considerable amount of literature concerning this issue, an optimal method to determine the number of components can not be found (for recent work, see [34], [35] and [36]). In particular, [34] argued that BIC tends to overestimate the number of mixture components and propose the ICL criterion as a solution; however, [35] showed that the ICL criterion tends to select models with too few components; therefore, they propose a method combining BIC and ICL to join the best of both criteria. A recent proposal is discussed in [36], who emphasized that there in practical situations the concept of cluster has not a unique definition. For example, we may look for a cluster characterized by a high variance but notthat different from another (more homogeneous) cluster when the average is concerned; otherwise, we may look for clusters with low variance but far from each other on the mean scale. Therefore, [36] propose different methods according to different definitions, based on a ridgeline analysis of modalities in Gaussian mixtures, dip test, Bhattacharya dissimilarity, a direct estimator of misclassification rate and the strength of predicting pairwise cluster memberships.

However, for standard model-based clustering methods only a single choice (the number of components) has to be made. In the mixture of SEMs context, model choice can be more complex when, for example, both dimensional reduction of observations (clustering) and variables (clustering or factorial techniques) are involved. In this case, once the number K of clusters has been specified, the number of latent variables Q_k within each cluster ($k = 1, ..., K$) has to be estimated. Since automatic model selection is usually desirable, most of the studies discussed in the next sections use for this purpose penalized likelihood criteria, such as AIC and BIC.

3 Case 1a: Clustering tissue samples and gene-reduction

Let us start considering the case of tissue sample clustering. In this context, the original matrix $\mathbf{Y}$ has to be transposed to let $\mathbf{y}_i$ represent the i-th tissue sample $(i = 1, ..., n)$. As mentioned before, the classification of few tissue samples with information on a very large number of genes represents a non standard problem in statistics, where observations are usually considered as n independent realizations of a J-dimensional random variable, $n > J$ to avoid near-singular estimates of (component-specific) covariance matrices. If the aim is to cluster few tissue samples characterized by a very large number of genes, we may face some problems in parameter estimation. A first raw solution could be to constrain on these quantities to reduce the number of parameters; we may turn to:

- local independence: component-specific covariance matrices are diagonal matrices (i.e. independent genes);
- covariance matrices are equal across the mixture components;
- component-specific covariance matrices are reparameterized by an eigenvalue decomposition as proposed in [37].

A more refined solution, to solve overparameterization problems getting a better trade-off between the use of simple and full covariance matrices, is the so-called *mixture of factor analyzers* (MFA, [38]). It can be considered as an extension of standard factor analysis models to deal with heterogeneous populations; in detail, data within each mixture component are generated according a standard factor model with substantial dimension reduction. Starting with the pioneering work of [38] in the Neural Computation community, MFA has received, thanks to its flexible structure, considerable interest in many other research communities. In particular, in the statistics community, [13] and [39] developed the MFA as a solution to simultaneous clustering of tissue samples and local dimensionality reduction for the feature space of genes (see [40] for an application to real data). It could be interesting to show that MFA is a particular mixture of SEMs by making the following assumptions in equations (1)-(3): $\boldsymbol{\alpha}_k = \mathbf{0}$, $\mathbf{B}_k = \mathbf{0}$ (that is $\boldsymbol{\xi}_{ik} = \mathbf{u}_{ik}$), $\mathbf{D}_k = diag(\sigma^2_{1k}, ..., \sigma^2_{Jk})$ and $\boldsymbol{\Phi}_k = \mathbf{I}$ where $\mathbf{I}$ denotes a Q_k-dimensional identity matrix, $i = 1, ..., n$ and $k = 1, ..., K$. As it can be noticed, MFA constrains the latent variables to be uncorrelated. The component-specific mean vector and covariance matrix are defined by: $\mathrm{E}(\mathbf{y}_i|z_{ik} = 1) = \boldsymbol{\upsilon}_k$, $\mathrm{cov}(\mathbf{y}_i|z_{ik} = 1) = \mathbf{V}_k\mathbf{V}'_k + \mathbf{D}_k$. Conditionally on $\mathbf{u}_{ik}$, they can be expressed as follows: $\mathrm{E}(\mathbf{y}_i|z_{ik} = 1, \mathbf{u}_{ik}) = \boldsymbol{\upsilon}_k + \mathbf{V}_k\mathbf{u}_{ik}$, $\mathrm{cov}(\mathbf{y}_i|z_{ik} = 1, \mathbf{u}_{ik}) = \mathbf{D}_k$. As it can be seen from the latter expression, the elements in $\mathbf{y}_i$ are conditionally independent given the latent factors $\mathbf{u}_{ik}$ ($\mathbf{D}_k$ is diagonal). Thus, dependence among the observed genes in the i-th tissue sample within the k-th cluster ($i = 1, ..., n$; $k = 1, ..., K$) are due to the latent factor only. In other words, MFA controls the number of parameters by modelling the component-specific covariance matrices $\boldsymbol{\Sigma}_k = \mathbf{V}_k\mathbf{V}'_k + \mathbf{D}_k$ and explains the correlations between genes through latent factors $\mathbf{u}_{ik}$.

As opposed to the previous case, let us assume that we are interested in identifying the variables (genes) that best discriminate between two or more groups. This can be achieved by assuming that the component-specific mean vectors lie in a common subspace identified by latent factors that best explain the between group variability. With this purpose, let us consider the mixture of SEMs (1)-(3) with the following constraints: $\boldsymbol{\upsilon}_k = \mathbf{0}$, $\mathbf{B}_k = \mathbf{0}$, $\mathbf{V}_k = \mathbf{V}$, $\boldsymbol{\xi}_{ik} = \mathbf{0}$ and unconstrained $\mathbf{D}_k$. These assumptions imply $\mathbf{u}_{\boldsymbol{ik}} = \boldsymbol{\alpha}_k$ ($i = 1, ..., n$ and $k = 1, ..., K$). Therefore, the component-specific mean vector and covariance matrix have the following reparameterization: $\mathrm{E}(\mathbf{y}_i|z_{ik} = 1) = \mathbf{V}\boldsymbol{\alpha}_k$ and $\mathrm{cov}(\mathbf{y}_i|z_{ik} = 1) = \mathbf{D}_k$. This model was introduced by [18] as a two-way model for simultaneous dimensional reduction and clustering of units (and will be referred to as RV). It could also be considered as a MFA model with non-random factors, $\boldsymbol{\upsilon}_k = \mathbf{0}$, $\mathbf{V}_k = \mathbf{V}$ and free component-specific covariance matrices, $\mathbf{D}_k$ ($k = 1, ..., K$). In [40] MFA and RV, are shown to give good clustering results when applied to data described by [41]. It is worth noticing that, while MFA solves the over-parametrization problem by modelling the component-specific covariance matrices, RV can be used when the number of tissue samples is smaller than the number of genes. As observed in [40], when a common spherical covariance matrix for the analyzed data can be assumed, RV is more efficient in terms of computational complexity.

Another model which is closely related to MFA is the mixture of probabilistic principal component analyzers (MPCA) proposed by [42] to model high dimensional data with relatively few parameters and applied as model-based clustering method in the context of microarray data by [43]. The principal component approach is motivated by the aim of projecting observed data onto an optimal subspace. In detail, MPCA resembles MFA with unequal, isotropic, error component-specific matrices $\mathbf{D}_k = \psi_k \mathbf{I}_J$ with $Q_k = J$, $k = 1, ...K$. It has to be noted that the isotropic model is highly constrained as the covariance structure in J dimensions is based on a single parameter (ψ_k).

More recently, [44] proposed a family of eight mixture models (Parsimonious Gaussian mixture models) where the following three constraints can be imposed $\mathbf{V}_k = \mathbf{V}$, $\mathbf{D}_k = \mathbf{D}$ and $\mathbf{D}_k = \psi_k \mathbf{I}_J$. The four special cases in which $\mathbf{V}_k = \mathbf{V}$, thanks to the reduced number of parameters involved, can be useful for clustering tissue samples and gene-reduction (PGMM1, PGMM2, PGMM3, PGMM4). In [45] further the Parsimonious Gaussian mixture model is extended leading to twelve mixture models (expanded PGMM) by reparametrizing $\mathbf{D}_k = \omega_k \mathbf{H}_k$, where $\omega_k \in \Re$ and $\mathbf{H}_k = diag\{\vartheta_1, ..., \vartheta_J\}$ such that $|\mathbf{H}_k| = 1$ for $k = 1, ..., K$. This expanded family is applied to two well known gene expression data sets.

4 Case 1b: Biclustering

The models discussed so far represent approaches that treat microarray data matrix as *asymmetric*; that is, rows and columns play rather different roles. The rows contain the objects which have to be clustered, whereas the columns contain the variables which are projected using factor techniques. As highlighted before, when analyzing DNA microarray experiments discovering clusters of genes with similar biological features and clusters of tissue samples with similar gene expression profiles may be a key point to detect meaningful biological functions. In such situations, it could be more useful to apply a *symmetric* approach, where the two modes of the data matrix have similar roles and both are summarized by clustering techniques. In the bioinformatic area, this approach is often referred to as *biclustering* and has become very popular in the last decade. In this section, we show how mixtures of SEMs can be useful to describe approaches to simultaneous clustering of genes and tissue samples.

Biclustering was originally introduced by [46],[47] and only later discussed by [48] in the microarray data analysis; for a review of biclustering techniques proposed in this area, see [49] and [50]. In particular, examples of biclustering methods in a finite mixture context have been developed by [51], who propose a block mixture model by using Bernoulli mixtures, and by [52], who propose a two-way Poisson mixture model for text analysis context. However, as far as we know, the only biclustering method for gene expression data based on finite mixture models has been proposed by [19]. Here, we focus on the latter model and show that it represents a particular case of mixture of SEMs. As before, let $\mathbf{y}_i$ be the J-dimensional observed vector for the i-th gene and assume the MFA structure holds except for the factor loading matrix $\mathbf{V}_k$. To introduce tissue samples clustering, $\mathbf{V}_k$ is specified, as a *binary row stochastic* matrix representing tissues cluster membership. Specifically, $\mathbf{V}_k$ is used to cluster tissue samples, whereas a traditional finite mixture approach is used to define the gene clustering (for details on the update of $\mathbf{V}_k$ see [19]). It has to be noted that this model leads to a component-specific covariance matrix, $\mathbf{V}_k\mathbf{V}_k' + \mathbf{D}_k$, having a block diagonal structure, *i.e.* a block matrix having on its main diagonal Q_k blocks formed by square matrices of size J_l $(l = 1, ..., Q_k$ with $\sum_{l=1}^{Q_k} J_l = J)$ such that the off-diagonal blocks are null matrices. In particular, the smaller the variance of a tissue, the larger the correlation among other tissues within the block. While the tissues have unequal variances, the covariance between tissues is equal to 1 if the tissues are within the same cluster and 0 otherwise. This biclustering model was introduced by [19] and applied to real data described by [53]. It has to be observed that this can be also fitted, without facing problems of singular estimates, to the transposed data matrix (rows are tissue samples and columns are genes), due to the possibility to split the component-specific covariance matrix, $\mathbf{V}_k\mathbf{V}_k' + \mathbf{D}_k$ (see [19] for details). Further extensions of this approach could be defined, such as allowing $\mathbf{\Phi}_k$ and $\mathbf{D}_k$ to be unconstrained; however, one should be careful because increasing the number of free parameters may yield identifiability problems.

5 Case 2: Clustering genes with technical replicates

Recently, some authors have underlined the necessity to develop models taking into account technical variability (replicates from each tissue sample) to improve the quality of analysis (for a discussion, see [54] among others). This kind of variability is usually modelled through *linear mixed-effects models* (LMMs), in a context similar to *multilevel models*, defined to handle repeated measures corresponding to the same individual. In the microarray field, [55] and [56] gave some examples of LMM studies, and, in particular, [57] and [58] propose ad hoc procedures to deal with repeated measurements when clustering microarray data. As far as the latter is concerned, a more refined solution has been developed by [20], where a finite mixture of LMMs is proposed to account for measurement variability where clustering is performed (LMM1). Here, we show how this model can be embedded in the mixture of SEMs framework. Assuming that $\mathbf{y}_i$ is a J-dimensional observed vector for the i-th gene containing R replicates for each of T tissue samples ($J = RT$). To account for replicates in a finite mixture framework, [20] simply constrain all measurements of a gene to belong to the same mixture component. They explicitly model the covariance structure between the r-th and r'-th technical replicates on the i-th gene and the t-th tissue sample, $cov(y_{itr}, y_{itr'})$. Thus, considering models (1)-(3), we assume $\boldsymbol{\upsilon}_k = \mathbf{X}\boldsymbol{\beta}_k$ where $\mathbf{X}$ is a known $(RT \times T)$ design matrix defined by

$$\mathbf{X} = \begin{pmatrix} \mathbf{1}_R & \mathbf{0}_R & ... & \mathbf{0}_R \\ \mathbf{0}_R & \mathbf{1}_R & ... & \mathbf{0}_R \\ ... & ... & ... & ... \\ \mathbf{0}_R & ... & \mathbf{0}_R & \mathbf{1}_R \end{pmatrix}$$

with $\mathbf{1}_R$ and $\mathbf{0}_R$ representing unit and null vectors of size R while $\boldsymbol{\beta}_k$ is a T-dimensional vector of fixed effects. Furthermore, $\mathbf{V}_k = \mathbf{X}$, $\mathbf{B}_k = \mathbf{0}$ and $\boldsymbol{\alpha}_k = \mathbf{0}$. These last two constraints imply $\boldsymbol{\xi}_{ik} = \mathbf{u}_{ik}$, where the terms $\boldsymbol{\xi}_{ik}$ represent zero-mean T-dimensional vectors of random effects capturing the common variability between gene expression and cluster center profiles. Finally, $\boldsymbol{\Phi}_k = \tau_k^2\mathbf{I}$ and $\mathbf{D}_k = \sigma_k^2\mathbf{I}$. Given these assumptions, the component-specific mean vector and covariance matrix of a mixture of SEMs become: $\mathrm{E}(\mathbf{y}_i|z_{ik} = 1) = \mathbf{X}\boldsymbol{\beta}_k$ and $\mathrm{cov}(\mathbf{y}_i|z_{ik} = 1) = \tau_k^2\mathbf{X}\mathbf{X}' + \sigma_k^2\mathbf{I}$. This model assumes that measurements corresponding to different genes are independent, while the covariance between two technical replicates (on any gene from the same tissue sample) is equal to τ_k^2, while, the variance of a gene is given by $\tau_k^2 + \sigma_k^2$. As mentioned in [20], this model can be extended accounting also for common variation in different tissue samples; that is, by modeling the covariance between two different tissues within the same gene through the assumption of another gene-specific random effect not depending on a given tissue. For further details, see [20].

6 Case 3: Clustering correlated genes

In sections 4 and 5, we have discussed particular specifications for a mixture of SEMs to handle clustering of genes in different experimental conditions. In particular, when the biclustering model is fitted by considering genes as rows and tissue samples as columns of a data matrix, potential dependence between genes is often ignored. Similarly, in LMM1 measurements corresponding to different genes are assumed to be independent, since attention is focused on the variability due to repeated measurements on any couple (gene, tissue sample). In both cases, independence may not hold since genes within a given tissue sample may be associated; even though, in practice, we can just proceed by ignoring this correlation. In this section, we focus on a family of models introduced by [21] to provide a unified approach to cluster genes by taking into account dependence in a wide variety of experimental situations. Also the models discussed in [21] can be placed within the finite mixture of SEMs framework. In particular, in the next subsections, we will focus on empirical situations considered in [21], namely: clustering of time-course data (LMM2), clustering of genes with replicates (LMM3) and clustering of genes with a known tissue sample partition (LMM4).

6.1 Case 3a: Clustering time-course data

Our aim here is to cluster n genes whose expression levels are repeatedly measured at J different time points. With respect to expressions (1)-(3), the model (referred to as LMM2) defines $\boldsymbol{\upsilon}_k$ to be a J-dimensional vector of fixed effects describing the conditional mean of the i-th unit (gene) values in the k-th cluster; we have $\mathbf{B}_k = \mathbf{0}$ and $\boldsymbol{\alpha}_k = \mathbf{0}$ so that $\mathbf{u}_{ik} = \boldsymbol{\xi}_{ik}$. The term $\mathbf{V}_k \boldsymbol{\xi}_{ik}$ is specified as

$\mathbf{V}_k\boldsymbol{\xi}_{ik} = b_{ik}\mathbf{1}_J + \mathbf{c}_k$, where b_{ik} is a cluster-specific gene-effect which is constant across time points with $\text{var}(b_{ik}|z_{ik} = 1) = \theta_{b_k}$, and $\mathbf{c}_k = (c_{k1}, ..., ckJ)'$ represents time and cluster-specific effect associated to the J time points with $\text{cov}(\mathbf{c}_k|z_{ik} = 1) = \theta_{c_k}\mathbf{I}$, so that $\boldsymbol{\Phi}_k = \theta_{b_k}\mathbf{1}_J\mathbf{1}'_J + \theta_{c_k}\mathbf{I}$. It has to be noticed that, within the same cluster, the gene specific-effect b_{ik} is used to model the covariance between time points corresponding to the same gene, while, c_k allows the covariance between expression levels of different genes measured at the same time point to be non-zero. Moreover, we assume $\mathbf{D}_k = \sigma_k^2\mathbf{I}$, i.e. the cluster-specific error covariance matrix is a diagonal matrix with constant variance across J time. Given these assumptions, the component-specific mean vector and covariance matrix of the LMM2 are defined by $\text{E}(\mathbf{y}_i|z_{ik} = 1) = \boldsymbol{\upsilon}_k$ and $\text{cov}(\mathbf{y}_i|z_{ik} = 1) = \sigma_k^2\mathbf{I} + \theta_{b_k}\mathbf{1}_J\mathbf{1}'_J + \theta_{c_k}\mathbf{I}$. Within the k-th cluster, the covariance structure between the gene expression levels can be summarized as: $\text{cov}(y_{ij}, y_{i'j'}) = \theta_{b_k}\delta_i^{i'} + \theta_{c_k}\delta_j^{j'} + \sigma_k^2\delta_i^{i'}\delta_j^{j'}$ (where $\delta_i^{i'} = 1$ if $i = i'$, 0 otherwise). Thus, within the same cluster the covariance between expression levels of different genes measured at the same time point is θ_{c_k}, the covariance between expression levels corresponding to different time points but to the same gene is equal to θ_{b_k} and the variance of a gene is $\theta_{b_k} + \theta_{c_k} + \sigma_k^2$.

6.2 Case 3b: Clustering genes with replicates

In this section, we discuss the technical situation dealt with in section 5; that is, there are R replicates on T tissue samples; here, however, we are not only interested in modelling the covariance between replicates corresponding to the same gene, but also the potential dependence of the expression levels entailing any pair of genes from the same tissue, which is commonly encountered in practice. This can be dealt with by adding another random effect (latent factor) to the finite mixture of SEMs described in section 5 (LMM1). This factor is assumed to be shared by genes from the same tissue at the same replicate. Specifically, on the basis of expressions (1)-(3) LMM3 is based on $\boldsymbol{\upsilon}_k = \mathbf{X}\boldsymbol{\beta}_k$, where $\mathbf{X}$ and $\boldsymbol{\beta}_k$ are defined as in section 5, $\mathbf{V}_k\boldsymbol{\xi}_{ik} = \mathbf{X}\mathbf{b}_{ik}+\mathbf{c}_k$, where $\mathbf{b}_{ik}$ represents the cluster-gene specific effect shared by replicates from the same gene and the same tissue sample with $\text{cov}(\mathbf{b}_k|z_{ik} = 1) = \theta_{b_k}\mathbf{I}$, while $\mathbf{c}_k$ is a RT-dimensional vector accounting for correlations between genes with $\text{cov}(\mathbf{c}_k|z_{ik} = 1) = \theta_{c_k}\mathbf{I}$, so that $\boldsymbol{\Phi}_k = \theta_{b_k}\mathbf{X}\mathbf{X}' + \theta_{c_k}\mathbf{I}$. Moreover, $\mathbf{D}_k = \text{diag}(\mathbf{X}\boldsymbol{\vartheta}_k)$, where $\boldsymbol{\vartheta}_k = (\sigma^2_{1k}, ..., \sigma^2_{Tk})$, that is $\mathbf{D}_k$ is a diagonal matrix with different variances across the T tissues and constant across the replicates. Therefore, the component-specific mean vector and covariance matrix from a LMM3 are, respectively, $\text{E}(\mathbf{y}_i|z_{ik} = 1) = \mathbf{X}\boldsymbol{\beta}_k$ and $\text{cov}(\mathbf{y}_i|z_{ik} = 1) = \theta_{b_k}\mathbf{X}\mathbf{X}' + \theta_{c_k}\mathbf{I} + \text{diag}(\mathbf{X}\boldsymbol{\vartheta}_k)$. Thus, conditional on the k-th cluster, the covariance structure between expression levels has the following form: $\text{cov}(y_{itr}, y_{i't'r'}) = \theta_{b_k}\delta_i^{i'}\delta_t^{t'} + \theta_{c_k}\delta_t^{t'}\delta_r^{r'} + \sigma^2_{tk}\delta_i^{i'}\delta_t^{t'}\delta_r^{r'}$.That is, within the same cluster the covariance between replicates on any genes measured on the same tissue sample is θ_{b_k}, the covariance between genes (from the same tissue and replicate) is θ_{c_k}, and the variance of a gene measured on the t-th tissue sample is $\theta_{b_k} + \theta_{c_k} + \sigma^2_{tk}$. Other types of covariance structures for gene expression data with replicates are described in [59].

6.3 Case 3c: Clustering genes with a known tissue sample partition

This section deals with situations where we aimed to find (eventually correlated) genes that best discriminate between two groups of tissue samples; such genes may be called marker genes. The problem is similar to biclustering with the main difference being that information about the partition of tissue samples is known. Let J_1 and J_2 the size of the two groups, which can be thought of "abnormal" and "normal" tissue samples, respectively, with $J_1 + J_2 = J$. Expressions (1)-(3), according to LMM4 can be specified as $\boldsymbol{\upsilon}_k = \mathbf{X}\boldsymbol{\beta}_k$, where $\boldsymbol{\beta}_k$ is a 2-dimensional vector of fixed effects with $\mathbf{X}$ being a $(J \times 2)$ binary matrix where, after proper rearranging, the first J_1 rows are (1 0) and the next J_2 rows are (0 1). This can easily be generalized to situations where more than two known tissue groups are available; including one extra column in per tissue group. In LMM4, $\mathbf{V}_k\boldsymbol{\xi}_{ik} = \mathbf{X}\mathbf{b}_{ik}+\mathbf{c}_k$, where $\mathbf{b}_{ik}$ is a 2-dimensional random effect explaining association between tissue samples in the same group, with $\text{cov}(\mathbf{b}_{ik}|z_{ik} = 1) = \theta_{b_k}\mathbf{I}$, and $\mathbf{c}_k$ is a J-dimensional vector modelling correlations between genes, within the same partition, with $\text{cov}(\mathbf{c}_k|z_{ik} = 1) = \theta_{c_k}\mathbf{I}$, so that $\boldsymbol{\Phi}_k = \theta_{b_k}\mathbf{X}\mathbf{X}' + \theta_{c_k}\mathbf{I}$. The component-specific error variance can be different between groups of tissue samples, $\mathbf{D}_k = \text{diag}(\mathbf{X}\boldsymbol{\vartheta}_k)$, where $\boldsymbol{\vartheta}_k = (\sigma^2_{1k}, \sigma^2_{2k})$. The component-specific mean vector and covariance matrix of the LMM4 have the same expression for those in the previous subsection even though the interpretation is different. In fact, conditional on belonging to the k-th cluster and the g-th group of tissues, the covariance between tissue samples is θ_{b_k}, the covariance between genes (measured on the same tissue sample) is θ_{c_k}, and the variance of a gene measured on the g-th tissue sample group $(g = 1, 2)$ is $\theta_{b_k} + \theta_{c_k} + \sigma^2_{gk}$. As it can be noticed, expression values corresponding to tissue samples are correlated only if the tissues are in the same group.

7 Real examples

In this section, results obtained on real data sets are discussed, to show the use of different specifications of finite mixtures of SEMs discussed in several empirical examples. In particular, since the second microarray experiment design (see section 1) can be easily related to the others, we decided to focus on the other two types: genes measured on different tissue samples as in [22] and genes with technical replicates as in [23].

7.1 The yeast galactose data

The yeast galactose data discussed in [22] consider $n = 205$ genes containing $R = 4$ replicates on $T = 20$ tissue samples. As in [57] and [59], we imputed all the missing values using a k-nearest neighbor method. The expression levels of the genes reflect four functional categories in the Gene Ontology (GO) listing [60]; thus, we expect they are clustered together. To cluster genes, we applied the specifications of the finite Mixture of SEMs discussed in section 5 (LMM1) and 6.2 (LMM3). All the results have been obtained using R EMMIX-WIRE library developed by [59] to fit LMMs via the EM algorithm (see `http://www.maths.uq.edu.au/~gjm/mix_soft/EMMIX-WIRE/index.html`). As highlighted before, if we use LMM1 we ignore the potential dependence between expression levels from any pair of genes on the same tissue, which is accounted for by LMM3. In detail, the former model assumes that replicates from the same gene and the same tissue share some random effects ($\mathbf{u}_{ik}$), while in the latter specific random effects are also shared by expressions measured on the same tissue but corresponding to different genes ($\mathbf{c}_k$). While, LMM1 assumes that the k-th component variance is

Cluster	Estimated	Estimated	True
	LLM1 membership	LLM3 membership	Membership
1	80	80	83
2	15	21	15
3	96	90	93
4	14	14	14

Table 1: LMM1 and LMM3 model results compared to the known membership

constant across tissues, it could be different in LMM3. Thus, we applied LMM1 and LMM3 to group genes into $K = 4$ clusters; we do not discuss fitting for varying numbers of components K, since we aim at comparing their performance in reproducing the known functional categories. We measured the degree of agreement between the known categories and the estimated cluster memberships by using the Modified Rand Index ([61]), whose value equals 1 in the case of perfect agreement. In our study, the values are 0.976 for LMM1 and 0.978 for LMM3, which are good results compared with several model-based and hierarchical clustering algorithms considered by [57]. Table 1 shows that the number of genes assigned to each group is similar for the two mixture models, and similar to the known functional membership. Additional information on the correlation among expression levels estimated by LMM3 shows that this correlation is significantly higher in cluster 1 (0.6973), than in cluster 4 (0.3667), cluster 3 (0.2166), or cluster 2 (0.1425). This could be a useful information to proceed with further biological research.

7.2 The human colorectal carcinoma data

This dataset has been previously analyzed by [23] and [59] using parametric clustering method and LMM4, respectively. Gene expression levels have been measured using a PCR (polymerase chain reaction). However, as in [59], the data set i a good example of association between genes and malignancy in human colorectal carcinoma. To allow for direct comparability with previous analyses on the same dataset, we consider 1536 gene expression levels over 100 tumoral samples and 11 normal samples, pre-processed using the same method as in [59]. Previous analyses, see e.g. [23] and [59], provided a partition of genes based on their expression levels in three gene clusters; more specifically, [23] found a cluster (27 genes) containing 17 genes linked to human colorectal carcinoma. [59], using the information on the known tissue partition as proposed in section 6.3, found a cluster with 15 genes consisting of a subset of the 17 genes listed by [23], which reduces the discriminating "important" genes list. Here, we use this dataset for two analyses:

1. Clustering of tissue samples and gene-reduction using MFA ([38]) and RV ([18]) models;
2. Clustering of genes and tissue samples to recover both gene and tissue sample partitions by using the biclustering model introduced in [19].

Results have been obtained using MATLAB routines based on the EM algorithm (and extensions) which have been developed by the corresponding author.

For the first analysis, we aim at recovering the "true" partition of tissue samples, and for this reason no model selection criteria have been used. We fitted the RV model to the transposed data matrix (n=111, J = 1536) with K=2, Q=2, and a common, spherical, covariance matrix $\mathbf{\Sigma}$, and the MFA model with K=2, Q=2, and a common covariance matrix. Both models have difficulties to correctly separated the 100 tumoral samples from the 11 normal ones (the Modified Rand Index is below 0.8). Further, we fitted the RV model with K=3, Q=2, and a common, spherical, covariance matrix $\mathbf{\Sigma}$

and the MFA model with K=3, Q=2, and common covariance matrix to further classify the 100 tumoral tissues in two subgroups according to having distant metastases or not (we know that 29 and 71 are with/without distant metastases). Also in this case, both models have difficulties to correctly separate the tissue samples in three groups even though the Modified Rand index values are slightly higher (but still below 0.8) and the BIC criteria values are lower than for the models with $K = 2$. A possible reason for the inability to recover the true partition of tissue samples is the low number of controls (11 normal tissues), together with the presence of many missing data.

For the second analysis, we take the original data matrix (n=1536, $J = 111$) and fit the biclustering model with three gene clusters and several number of tissue clusters, with the purpose: 1) compare the gene partition with the ones obtained by [23] and [59]; 2) find a more meaningful tissue partition. We run the algorithm several times to avoid local maxima, choosing the best solution through BIC among 100 starting points. This has been reached with $Q_1 = Q_2 = 1$, $Q_3 = 2$; that is, two gene clusters constant across tissues and one gene cluster containing discriminant genes between two tissue groups (the third cluster). The last bicluster includes 312 genes with two clusters of 82 and 29 tissue samples, and contains most of the 27 genes selected by [23]; 56 tissue samples out of 82 are a subset of 71 tissues without distant metastases and 11 out of 29 are a subset of 29 tissues with distant metastases. Although this biclustering model ignores the possible dependence between genes, it is able to select a potential bicluster as relevant to tumor progression process.

8 Comments and future research

We have presented a general framework that provides a richer description of model-based clustering developed in microarray analysis, and discussed the empirical evidence from two real datasets. The full range of models and the corresponding model parameter specifications are summarized in Table 2.

The first fifteen models (MFA [38], MPCA [42], RV [18], PGMM1-PGMM8 [44], EPGMM1-EPGMM4 [45]) in Table 2 are examples of approaches to provide partition of observations under dimensional reduction of the feature space, and for this reason they may be useful to clustering tissue samples avoiding singularity of component-specific covariance matrices. The Biclustering model ([19]) is defined to discover local expression patterns in microarray data. It can be fitted to the original or transposed data matrix (with tissue samples as rows and genes as columns) taking into account that, in the former case, we are ignoring correlation between genes. The last four models (LMM1 [20], LMM2 [21], LMM3 [21], LMM4 [21]) deal with gene clustering. LMM1 [20] and LMM3 [21] models take into account technical variability due to replicates on each tissue sample, but the latter is tailored to model also the dependence between any pair of gene values. LMM2 [21] is an important tool for clustering time-course data; it takes into account the correlations between time points corresponding to a single gene and between any pair of genes for the same time point. Finally, LMM4 [21] is a powerful approach to discover genes which are associated with disease status (clinical markers): it takes into account the correlation between tissues in the same group and between pair of genes of the same tissue.

Model	$\boldsymbol{\upsilon}_k$	$\mathbf{V}_k\mathbf{u}_{ik}$	$\boldsymbol{\alpha}_k$	$\mathbf{B}_k$	$\mathbf{D}_k$	$\boldsymbol{\Phi}_k$	$\boldsymbol{\mu}_k$	$\boldsymbol{\Sigma}_k$
MFA [38]	free	free	$\mathbf{0}$	$\mathbf{0}$	$diag(\sigma^2_{1k}, ..., \sigma^2_{Jk})$	$\mathbf{I}$	$\boldsymbol{\upsilon}_k$	$\mathbf{V}_k\mathbf{V}'_k + \mathbf{D}_k$
MPCA [42] $\equiv$ PGMM5 [44]	free	free	$\mathbf{0}$	$\mathbf{0}$	$\psi_k\mathbf{I}$	$\mathbf{I}$	$\boldsymbol{\upsilon}_k$	$\mathbf{V}_k\mathbf{V}'_k + \psi_k\mathbf{I}$
RV [18]	$\mathbf{0}$	$\mathbf{V}\mathbf{u}_{ik}$	$\mathbf{u}_{ik}$	$\mathbf{0}$	free	$-$	$\mathbf{V}\boldsymbol{\alpha}_k$	$\mathbf{D}_k$
PGMM1 [44]	free	$\mathbf{V}_k = \mathbf{V}$	$\mathbf{0}$	$\mathbf{0}$	free	$\mathbf{I}$	$\boldsymbol{\upsilon}_k$	$\mathbf{V}\mathbf{V}' + \mathbf{D}_k$
PGMM2 [44]	free	$\mathbf{V}_k = \mathbf{V}$	$\mathbf{0}$	$\mathbf{0}$	$\mathbf{D}$	$\mathbf{I}$	$\boldsymbol{\upsilon}_k$	$\mathbf{V}\mathbf{V}' + \mathbf{D}$
PGMM3 [44]	free	$\mathbf{V}_k = \mathbf{V}$	$\mathbf{0}$	$\mathbf{0}$	$\psi_k\mathbf{I}$	$\mathbf{I}$	$\boldsymbol{\upsilon}_k$	$\mathbf{V}\mathbf{V}' + \psi_k\mathbf{I}$
PGMM4 [44]	free	$\mathbf{V}_k = \mathbf{V}$	$\mathbf{0}$	$\mathbf{0}$	$\psi\mathbf{I}$	$\mathbf{I}$	$\boldsymbol{\upsilon}_k$	$\mathbf{V}\mathbf{V}' + \psi\mathbf{I}$
PGMM6 [44]	free	free	$\mathbf{0}$	$\mathbf{0}$	$\mathbf{D}$	$\mathbf{I}$	$\boldsymbol{\upsilon}_k$	$\mathbf{V}_k\mathbf{V}'_k + \mathbf{D}$
PGMM7 [44]	free	free	$\mathbf{0}$	$\mathbf{0}$	$\psi\mathbf{I}$	$\mathbf{I}$	$\boldsymbol{\upsilon}_k$	$\mathbf{V}_k\mathbf{V}'_k + \psi\mathbf{I}$
$PGMM8$ [44]	free	free	$\mathbf{0}$	$\mathbf{0}$	free	$\mathbf{I}$	$\boldsymbol{\upsilon}_k$	$\mathbf{V}_k\mathbf{V}'_k + \mathbf{D}_k$
EPGMM1 [45]	free	free	$\mathbf{0}$	$\mathbf{0}$	$\omega\mathbf{H}$ $\omega \in \Re$ $\mathbf{H}_k = diag\{\vartheta_{1k}, ..., \vartheta_{Jk}\}$	$\mathbf{I}$	$\boldsymbol{\upsilon}_k$	$\mathbf{V}_k\mathbf{V}'_k + \omega\mathbf{H}_k$
EPGMM2 [45]	free	free	$\mathbf{0}$	$\mathbf{0}$	$\omega\mathbf{H}_k$ $\omega \in \Re$ $\mathbf{H} = diag\{\vartheta_1, ..., \vartheta_J\}$	$\mathbf{I}$	$\boldsymbol{\upsilon}_k$	$\mathbf{V}_k\mathbf{V}'_k + \omega\mathbf{H}$
EPGMM3 [45]	free	$\mathbf{V}_k = \mathbf{V}$	$\mathbf{0}$	$\mathbf{0}$	$\omega\mathbf{H}_k$ $\omega \in \Re$ $\mathbf{H} = diag\{\vartheta_{1k}, ..., \vartheta_{Jk}\}$	$\mathbf{I}$	$\boldsymbol{\upsilon}_k$	$\mathbf{V}\mathbf{V}' + \omega\mathbf{H}_k$
EPGMM4 [45]	free	$\mathbf{V}_k = \mathbf{V}$	$\mathbf{0}$	$\mathbf{0}$	$\omega_k\mathbf{H}_k$ $\omega_k \in \Re$ $\mathbf{H} = diag\{\vartheta_{1k}, ..., \vartheta_{Jk}\}$	$\mathbf{I}$	$\boldsymbol{\upsilon}_k$	$\mathbf{V}\mathbf{V}' + \omega_k\mathbf{H}_k$
Biclustering [19]	binary row stochastic	free	$\mathbf{0}$	$\mathbf{0}$	$diag(\sigma^2_{1k}, ..., \sigma^2_{Jk})$	$\mathbf{I}$	$\boldsymbol{\upsilon}_k$	$\mathbf{V}_k\mathbf{V}'_k + \mathbf{D}_k$
LMM1 [20]	$\mathbf{X}\boldsymbol{\beta}_k$	$\mathbf{X}\mathbf{u}_{ik}$	$\mathbf{0}$	$\mathbf{0}$	$\sigma^2_k\mathbf{I}$	$\tau^2_k\mathbf{I}$	$\mathbf{X}\boldsymbol{\beta}_k$	$\tau^2_k\mathbf{X}\mathbf{X}' + \sigma^2_k\mathbf{I}$
LMM2 [21]	$\boldsymbol{\upsilon}_k$	$\mathbf{X}b_{ik}$+$\mathbf{c}_k$	$\mathbf{0}$	$\mathbf{0}$	$\sigma^2_k\mathbf{I}$	$\theta_{b_k}\mathbf{1}\mathbf{1}' + \theta_{c_k}\mathbf{I}$	$\boldsymbol{\upsilon}_k$	$\sigma^2_k\mathbf{I} + \theta_{b_k}\mathbf{1}\mathbf{1}' + \theta_{c_k}\mathbf{I}$
LMM3 [21]	$\mathbf{X}\boldsymbol{\beta}_k$	$\mathbf{X}b_{ik}$+$\mathbf{c}_k$	$\mathbf{0}$	$\mathbf{0}$	$\mathrm{diag}(\mathbf{X}(\boldsymbol{\vartheta}_k))$ $\boldsymbol{\vartheta}_k = (\sigma^2_{1k}, ..., \sigma^2_{Tk})$	$\theta_{b_k}\mathbf{X}\mathbf{X}' + \theta_{c_k}\mathbf{I}$	$\mathbf{X}\boldsymbol{\beta}_k$	$\theta_{b_k}\mathbf{X}\mathbf{X}' + \theta_{c_k}\mathbf{I}+$ $\mathrm{diag}(\mathbf{X}(\boldsymbol{\vartheta}_k))$
LMM4 [21]	$\mathbf{X}\boldsymbol{\beta}_k$	$\mathbf{X}b_{ik}$+$\mathbf{c}_k$	$\mathbf{0}$	$\mathbf{0}$	$\mathrm{diag}(\mathbf{X}(\boldsymbol{\vartheta}_k))$ $\boldsymbol{\vartheta}_k = (\sigma^2_{1k}, \sigma^2_{2k})$	$\theta_{b_k}\mathbf{X}\mathbf{X}' + \theta_{c_k}\mathbf{I}$	$\mathbf{X}\boldsymbol{\beta}_k$	$\theta_{b_k}\mathbf{X}\mathbf{X}' + \theta_{c_k}\mathbf{I}+$ $\mathrm{diag}(\mathbf{X}(\boldsymbol{\vartheta}_k))$

Table 2: Different model specifications derived from the mixture of SEMs

There are several advantages to define a common approach to the methods we have discussed in this paper. First, the proposed approach emphasizes that all the techniques are essentially particular specifications of the same model since they correspond to specific restrictions imposed to a general on mixture of SEMs model; this unifying model would encompass a broad range of existing models as specific cases. Essentially this formal framework helps visualize connections and differences between related approaches, it highlights specific restrictions implied by each particular model, and it potentially leads to a powerful and flexible path for development of new modelling approaches in microarray data analysis. The models discussed in these pages are by no means exhaustive, but we believe that, this systematic organization can be used by interested readers as a good starting point to learn and apply the most important models proposed in the last few years in the context of model-based clustering of microarray data. One promising direction for future research involves new specifications of mixture of SEMs corresponding to $\mathbf{B}_k \neq \mathbf{0}$. For example, a Markov structure could be imposed on the latent factors to cluster genes with repeated measurements (as mentioned in section 2.1). Another example could be the inclusion of higher-level factors for biclustering of repeated measure data; that is, by using one or more latent factors for all tissues at one time, and a set of higher-level factors connecting the factors at different time points. Finally, the unification in a single software package of specific cases of mixture of SEMs for microarray data analyses could be investigated.

9 Acknowledgements

We are happy to acknowledge, with sincere thanks, the referees for feedback and reflections which helped refine many of the ideas discussed in a previous version, and Dr. Marco Alfò for his valuable discussions and comments on a previous the draft of this paper.

References

[1] Wolfe JH. Object cluster analysis of social areas. Berkeley ed. University of California: Master's thesis; 1963.

[2] Wolfe JH. Pattern clustering by multivariate mixture analysis. Multivariate Behavioral Research. 1970;5:329–350.

[3] Day NE. Estimating the components of a mixture of normal distributions. Biometrika. 1969;56:463–474.

[4] Scott AJ, Symons MJ. Clustering methods based on likelihood ratio criteria. Biometrics. 1971;27:387–397.

[5] Binder DA. Bayesian Cluster Analysis. Biometrika. 1978;65:31–38.

[6] McLachlan GJ. The classification and mixture maximum likelihood approaches to cluster analysis. Handbook of Statistics, Amsterdam: North-Holland. 1982;2:199–208.

[7] McLachlan GJ, Basford KE. Mixture Models: Inference and applications to clustering. New York: Marcel Dekker Inc; 1982.

[8] Banfield JD, Raftery AE. Model-based Gaussian and non-Gaussian clustering. Biometrics. 1993;49(3):803–821.

[9] Celeux G, Govaert G. Gaussian parsimonious clustering models. Pattern Recognition. 1995;28:781–793.

[10] Holmes I, Bruno WJ. Finding regulatory elements using joint likelihoods for sequence and expression profile data. In: 8, editor. Proc Int Conf Intell Syst Mol Biol; 2000. p. 202–10.

[11] Barash Y, Friedman N. Context-specific Bayesian clustering for gene expression data. Journal of Computational Biology. 2002;9:169–191.

[12] Yeung KY, Fraley C, Murua A, Raftery AE, Ruzzo WL. Model-based clustering and data transformations for gene expression data. Bioinformatics. 2001;17(10):977–987.

[13] McLachlan GJ, Peel D. Finite Mixture Models. New York: Wiley Series in Probability and Statistics; 2000.

[14] Baek J, McLachlan GJ, Flack L. Mixtures of factor analyzers with common factor loadings: applications to the clustering and visualisation of high-dimensional data. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2010;(32):1298–1309.

[15] McNicholas PD, Murphy TB. Model-based clustering of longitudinal data. The Canadian Journal of Statistics. 2010;38(1):153–168.

[16] Martella F, Alfò M, Vichi M. Hierarchical mixture models for biclustering in microarray. Statistical Modelling. 2010;To appear.

[17] Zhang J. A Bayesian model for biclustering with applications. Journal of the Royal Statistical Society: Series C (Applied Statistics). 2010;54(4):635–656.

[18] Rocci R, Vichi M. A two-way model for simultaneous reduction and classification; 2002. Atti della XLI Riunione Scientifica, 5th-7th June.

[19] Martella F, Alfò M, Vichi M. Biclustering of Gene Expression Data by an Extension of Mixtures of Factor Analyzers. The International Journal of Biostatistics. 2008;4(1):3.

[20] Celeux G, Martin O, Lavergne C. Mixture of linear mixed models for clustering gene expression profiles from repeated microarray experiments. Statistical Modelling. 2005;5(3):243–267.

[21] Ng SK, McLachlan GJ, Wang K, Ben-Tovim Jones L, Ng SW. A Mixture model with random-effects components for clustering correlated gene-expression profiles. Bioinformatics. 2006;22(14):1745–1752.

[22] Ideker T, Thorsson V, Ranish JA, Christmas R, Buhler J, Eng JK, et al. Integrated Genomic and Proteomic Analyses of a Systematically Perturbed Metabolic Network. Science. 2001;92:929–934.

[23] Muro S, Takemasa I, Oba S, Matoba R, Ueno N, Maruyama C, et al. Identification of expressed genes linked to malignancy of human colorectal carcinoma by parametric clustering of quantitative expression data. Genome Biol. 2003;4(3):1–10.

[24] Rabe-Hesketh S, A S. Classical latent variable models for medical research. Statistical Methods in Medical Research. 2008;17(1):5–32.

[25] Armingerm G, Steinm P. Finite mixtures of covariance structure models with regressors. Sociological Method & Research. 1997;26(2):148–182.

[26] Jedidi K, Jagpal HS, DeSarbo WS. Finite-mixture structural equation models for response-based segmentation and unobserved heterogeneity. Marketing Science. 1997;16(1):39–59.

[27] Yung YF. Finite mixtures in confirmatory factor-analysis models. Psychometrika. 1997;62(3):297–330.

[28] Dolan CV, van der Maas HLJ. Fitting multivariate normal mixtures subject to structural equation modeling. Psychometrika. 1998;63(3):227–253.

[29] Arminger G, Stein P, Wittenberg J. Mixtures of conditional mean and covariance-structure models. Psychometrika. 1998;64(4):475–494.

[30] Zhu HT, Lee SY. A Bayesian analysis of finite mixtures in the LISREL model. Psychometrika. 2001;66(1):133–152.

[31] Dempster AP, Laird NM, Rubin DA. Maximum likelihood from incomplete datavia the EM Algorithm. Journal of the Royal Statistical Society B. 1977;39(1):1–38.

[32] Titterington DM, Smith AFM, Makov UE. Statistical Analysis of Finite Mixture Distributions. New York: John Wiley; 1985.

[33] Fraley C, Raftery AE. Model-based clustering, discriminant analysis, and density estimation. J Am Stat Assoc. 2002;97:611–631.

[34] Biernacki C, Celeux G, Govaert G. Assessing a mixture model for clustering with the integrated completed likelihood. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2000;22:719–725.

[35] Baudry JK, Raftery AE, Celeux G, Lo K, Gottardo R. Combining Mixture Components for Clustering; 2008. Technical Report n. 6644.

[36] Henning C. Methods for merging Gaussian mixture components. Adv Data Anal Classif. 2010;4(3):3–34.

[37] Banfield JD, Raftery AE. Model-based Gaussian and non-Gaussian clustering. Biometrics. 1993;49:803–821.

[38] Ghahramani G, Hinton GE. The EM algorithm for mixtures of factor analyzers. Toronto (Canada), M5S 1A4: Department of Computer Science, University of Toronto; 1996. Report No.: CRG-TR-96-1.

[39] McLachlan GJ, Bean RW, Peel D. A mixture model-based approach to the clustering of microarray expression data. Bioinformatics. 2002;18(3):413–422.

[40] Martella F. Classification of microarray data with factor mixture models. Bioinformatics. 2006;22(2):202–208.

[41] Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science. 1999;286:531–537.

[42] Bishop CM, Tipping ME. Mixtures of probabilistic principal component analyzers. Neural Computation. 1999;11(2):443–482.

[43] Yoshioka T, Morioka R, Kobayashi K, Oba S, Ogawsawara N, S I. Clustering of Gene Expression Data by Mixture of PCA Models. Archive Proceedings of the International Conference on Artificial Neural Networks table of contents. 2002;2415:796.

[44] McNicholas PD, Murphy TB. Parsimonious Gaussian mixture models. Statistics and Computing. 2008;18:285–296.

[45] McNicholas PD, Murphy TB. Model-Based Clustering of Microarray Expression Data via Latent Gaussian Mixture Models. Bioinformatics. 2010;26(21):2705–2712.

[46] Hartigan JA. Direct clustering of a data matrix. J Am Stat Assoc. 1972;67:123–129.

[47] Hartigan JA. Clustering algorithms. New York: York Wiley; 1975.

[48] Cheng Y, Church GM. Biclustering of expression data. Proc Int Conf Intell Syst Mol Biol. 2000;p. 93–103.

[49] Madeira SC, Oliveira AL. Biclustering Algorithms for Biological Data Analysis: A Survey. IEEE/ACM Transactions on Computational Biology and Bioinformatics. 2004;1(1):24–45.

[50] Prelic A, Bleuler S, Zimmermann P, Wille A, Bhlmann P, Gruissem W, et al. A systematic comparison and evaluation of biclustering methods for gene expression data. Bioinformatics. 2006;22(9):1122–1129.

[51] Govaert G, Nadif M. Clustering with block mixture models. Pattern Recognition. 2003;36(2):463–473.

[52] Li J, Zha H. Two-way Poisson mixture models for simultaneous document classification and word clustering. Computational Statistics and Data Analysis. 2006;50(1):163–180.

[53] Bittner M, Meltzer P, Khan J, Chen Y, Jiang Y, Seftor E, et al. Molecular classification of cutaneous malignant by gene expression profiling. Nature. 2000;406(6795):536–540.

[54] Lee MLT, Kuo FC, Whitmorei GA, Sklar J. Importance of replication in microarray gene expression studies: statistical methods and evidence from repetitive cDNA hybridizations. Proc Natl Acad Sci USA. 2000;97:9834–9838.

[55] Wolfinger RD, Gibson G, Wolfinger ED, Bennett L, Hamadeh H, Bushel P, et al. Assessing gene significance from cDNA microarray expression data via mixed models. J Comput Biol. 2001;8:625–637.

[56] Efron B, Tibshirani R, Goss V, Chu G. Microarrays and their use in a comparative experiment. Dept Statistics, Stanford Univ.; 2000. No: 37B/213.

[57] Yeung KY, Medvedovic M, Bumgarner RE. Clustering gene-expression data with repeated measurements. Genome Biol. 2003;4(R34).

[58] Luan Y, Li H. Clustering of time-course gene expression data using a mixed-effects model with B-splines. Bioinformatics. 2003;19(4):474–482.

[59] Ng SK, McLachlan GJ, Bean RW, Ng SW. Clustering replicated microarray data via mixtures of random effects models for various covariance structures. In: Boden M, Bailey TL, editors. In Conferences in Research and Practice in Information Technology, Vol. 73. Sydney: The Australian Computer Society; 2006. p. 29–33.

[60] Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene Ontology: tool for the unification of biology. Nat Genet. 2000;25:2529.

[61] Milligan GW, Cooper MC. A study of the comparability of external criteria for hierarchical cluster analysis. Multivariate Behavioral Research. 1986;21:441–458.