Statistical Modelling 2005; **5**: 1–25

Mixture of linear mixed models for clustering gene expression profiles from repeated microarray experiments

Gilles Celeux[1], Olivier Martin[2] and Christian Lavergne[3]
[1]Department of Mathematics, University Paris-Sud, Paris, France
[2]INRA, Unité Protéomique, Montpellier, France
[3]Institut de Mathématiques et de Modélisation de Montpellier, Montpellier, France

Abstract: Data variability can be important in microarray data analysis. Thus, when clustering gene expression profiles, it could be judicious to make use of repeated data. In this paper, the problem of analysing repeated data in the model-based cluster analysis context is considered. Linear mixed models are chosen to take into account data variability and mixture of these models are considered. This leads to a large range of possible models depending on the assumptions made on both the covariance structure of the observations and the mixture model. The maximum likelihood estimation of this family of models through the EM algorithm is presented. The problem of selecting a particular mixture of linear mixed models is considered using penalized likelihood criteria. Illustrative Monte Carlo experiments are presented and an application to the clustering of gene expression profiles is detailed. All those experiments highlight the interest of linear mixed model mixtures to take into account data variability in a cluster analysis context.

Key words: cluster analysis; gene expression profile; linear model; mixture model; penalized likelihood criteria; random effect

Data and software link available from: http://stat.uibk.ac.at/SMIJ
Received April 2004; **revised** April 2005; **accepted** April 2005

1 Introduction

Microarrays are one of the domains in experimental molecular biology that offer the ability to measure the expression levels of a large amount of genes simultaneously, up to several thousands (Duggan *et al.*, 1999). Gene expression profiles are used to study gene function in cellular processes. Because of the large number of genes and the complexity of biological networks, clustering is often used to find co-regulated and functionally related groups of genes. Among the most used methods, we can cite hierarchical classification (Eisen *et al.*, 1998), self-organizing maps (Tamayo *et al.*, 1999) and the *K*-means algorithm (Tavazoie *et al.*, 1999). More recently, some authors have based cluster analysis of gene expression profiles on multivariate Gaussian mixtures (Ghosh and Chinnaiyan, 2002; Yeung *et al.*, 2001).

Address for correspondence: O Martin, INRA, Unité Protéomique, 2, Place Viala, 34060 Montpellier Cédex 1, France. E-mail: martinol@ensam.inra.fr

 10.1191/1471082X05st096oa

Data sets from multiple experiments are represented by an expression matrix in which each column represents a single microarray experiment. A row represents the expression vector for a particular gene and is called the gene expression profile. Columns can represent different time points in a particular experimental condition or different factor levels (occurrence of some disease for instance). In most cases, the logarithm of the expression ratio between the experimental condition and a reference condition for each gene is computed. As shown in Lee *et al.* (2000), any single microarray experiment is subject to substantial variability and replicate measures are needed to provide a reliable analysis of gene expression. However, as far as we know, most clustering studies of gene expression profiles did not take into account the variability of gene expression profiles and did not consider repeated data to derive genes clusters.

Standard clustering algorithms and finite mixture models are not tailored for taking into account repeated data. Two common attitudes when facing this variability problem in cluster analysis of gene expression profiles are the following:

- Neglecting the problem and clustering genes from a single measure of the variables for each gene.
- Restricting the variability to a mean effect and clustering genes from the mean values of independent repeated measures for each gene.

Both attitudes can be expected to be unsatisfactory. The first one clearly jeopardizes the analysis as soon as the variability is important. The second one is assuming that the variability does not depend on covariates or on the genes and can be unrealistic. A notable effort to deal with repeated data in a cluster analysis of genes is the software of Yeung *et al.* (2003) where several *ad hoc* procedures are proposed to downweight genes with noisy measures.

We propose a more formal approach, embedded in the model-based cluster analysis context. It is aiming to take into account *random effects* carried by repeated measures in a proper way. In statistical analysis, variability in data is classically related to random effects. Linear mixed models (LMM) are devoted to analyse those random effects from repeated data (Searle *et al.*, 1992). Examples of LMM applications in gene expression profile analysis are given in Wolfinger *et al.* (2001) and Efron *et al.* (2000). In this article, we propose to take into account the variability of measurements for mixture distributions with linear mixed models.

Assuming that the measures are repeated R times for a time series of T points and for each gene i ($i \in \{1, \ldots, I\}$), the repetition r of the log-ratio measure for the ith gene at time t is denoted y_{itr}. In order to represent the differential expression within a cluster k and to take into account repeated measures, a possible representation is

$$y_{itr}^k = \beta_{kt} + \xi_{it}^k + \varepsilon_{itr}^k.$$

In this model, the fixed effect β_{kt} gives the intensity of the differential expression at time t in cluster k. Owing to the important number of genes the gene effect is considered as a random effect. The term ξ_{it}^k represents a random effect, which is added to measures of gene i at time t in cluster k. This term can be regarded as the variation between gene expression profiles and the cluster centre. Finally, the error term ε_{itr}^k represents the experimental error.

This type of approach can lead to numerous models for clustering gene expression profiles from repeated data. It is considered in this article, which is organized as follows. Section 2 is devoted to the presentation of linear mixed models. It is focused on models of interest for analysing gene expression data. Mixture of mixed models and its estimation through the EM algorithm is presented in Section 3, and the EM equations are detailed for a particular model. Section 4 is devoted to the presentation of illustrative numerical Monte Carlo experiments and an application on the formation of wood tissues is detailed. A short concluding section summarizes the main points of this article and gives some perspectives for future work.

2 Linear mixed models for gene expression profile

To simplify the general presentation of linear mixed models, hereafter abbreviated LMM, we focus attention in Section 2.1 on a particular linear mixed model. In Section 2.2, alternative LMM are presented.

2.1 An example of LMM

Recall that the rth repetition of an expression log-ratio between two experimental conditions at time t for gene i is denoted y_{itr}. It is assumed that R repetitions of expression ratios are recorded at T different times for I independent genes. The linear mixed model, taking into account a covariance structure in the repetitions we consider here, is

$$y_{itr} = \beta_t + \xi_{it} + \varepsilon_{itr} \tag{2.1}$$

where β_t represents the fixed effect of time, $\xi_{it} \sim \mathcal{N}(0, \tau^2)$ is the random effect of gene i at instant t, and $\varepsilon_{itr} \sim \mathcal{N}(0, \sigma^2)$ is the error measure.

It is important to understand that the measurements of two different genes are supposed to be independent. However, the covariance structure between two log-ratios is as follows (noting $\delta_k^{k'} = 1$ if $k = k'$ and 0 otherwise):

$$\operatorname{cov}(y_{itr}, y_{i't'r'}) = \tau^2 \delta_i^{i'} \delta_t^{t'} + \sigma^2 \delta_i^{i'} \delta_t^{t'} \delta_r^{r'}. \tag{2.2}$$

In this model, the covariance between the repetitions of a gene at the same instant is not null and is equal to τ^2. The variance of an observation is $\tau^2 + \sigma^2$. For any given gene, the correlation between the measurements at two different instants is null.

As seen in model (2.1), linear mixed models are aiming to analyse the variability that is evident in data by including both fixed and random effects (Searle *et al.*, 1992). The general equation of a LMM is

$$\boldsymbol{y} = \underbrace{\boldsymbol{X\beta}}_{\text{fixed effects part}} + \underbrace{\boldsymbol{U\xi}}_{\text{random effects part}} + \boldsymbol{\epsilon} \tag{2.3}$$

where

- $\boldsymbol{y}$ is the random vector of N observations,
- $\boldsymbol{X}_{(N,p)}$ and $\boldsymbol{U}_{(N,q)}$ are known design matrices,
- $\boldsymbol{\beta}$ is the fixed effect vector, of size p, to be estimated,
- $\boldsymbol{\xi} = (\boldsymbol{\xi}_1', \ldots, \boldsymbol{\xi}_H')'$ is the vector of the H random effects, with $\boldsymbol{\xi}_h$ of size q_h for $h = 1, \ldots, H$, such that the $\boldsymbol{\xi}_h$'s are independent, $\boldsymbol{\xi}_h \sim \mathcal{N}(0, \tau_h^2 \mathbf{Id}_{q_h})$, for $h = 1, \ldots, H$, and the variances τ_h^2, $h = 1, \ldots, H$ are to be estimated,
- $\boldsymbol{\epsilon}$ is a random vector of residuals of size N such that $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \sigma^2 \mathbf{Id}_N)$ with σ^2 to be estimated and ε is independent of each $\boldsymbol{\xi}_h$.

The canonical equation (2.3) applies for model (2.1) with

- $\boldsymbol{y} = (\boldsymbol{y}_1', \ldots, \boldsymbol{y}_I')'$ where $\boldsymbol{y}_i$ is the random vector of measures for gene i. The size of $\boldsymbol{y}_i$ is RT,
- $\boldsymbol{\xi} = (\xi_{11}, \ldots, \xi_{1T}, \ldots, \xi_{I1}, \ldots, \xi_{IT})'$ vector of size IT of random effects ($H = 1$), and $\boldsymbol{\xi} \sim \mathcal{N}(0, \tau^2 \mathbf{Id}_{IT})$,
- the design matrix

$$\boldsymbol{U}_{(N,\,IT)} = \begin{bmatrix} \mathbf{1}_R & \mathbf{0}_R & \cdots & \mathbf{0}_R \\ \mathbf{0}_R & \ddots & \cdots & \vdots \\ \vdots & \cdots & \ddots & \mathbf{0}_R \\ \mathbf{0}^R & \cdots & \mathbf{0}_R & \mathbf{1}_R \end{bmatrix} \tag{2.4}$$

where $\mathbf{1}_R$ and $\mathbf{0}_R$ denote respectively the vectors $(1, \ldots, 1)'$ and $(0, \ldots, 0)'$ of size R,

- $\boldsymbol{\epsilon}$: random vector of errors of size $N = IRT$ and $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \sigma^2 \mathbf{Id}_N)$,
- $\boldsymbol{\beta}$: unknown fixed effects vector of size T,
- and the design matrix $\boldsymbol{X}_{(N,T)}$ of the form

$$\boldsymbol{X}_{(N,T)} = \begin{bmatrix} \dot{\boldsymbol{X}} \\ \vdots \\ \dot{\boldsymbol{X}} \end{bmatrix} \tag{2.5}$$

where

$$\dot{\boldsymbol{X}}_{(RT,T)} = \begin{bmatrix} \mathbf{1}_R & \mathbf{0}_R & \cdots & \mathbf{0}_R \\ \mathbf{0}^R & \ddots & \cdots & \vdots \\ \vdots & \cdots & \ddots & \mathbf{0}_R \\ \mathbf{0}_R & \cdots & \mathbf{0}_R & \mathbf{1}_R \end{bmatrix}. \tag{2.6}$$

Linear mixed models are incomplete data models where the missing data are the realizations ξ_{it}, $i = 1, \ldots, I$; $t = 1, \ldots, T$ of the random effect. The maximum likelihood

parameter estimates can be derived for instance with the EM algorithm (Dempster *et al.*, 1977; McLachlan and Krishnan, 1997), which consists of maximizing iteratively the conditional expectation of the complete likelihood,

$$\begin{aligned} l(\boldsymbol{\beta}, \tau^2, \sigma^2 | \boldsymbol{y}, \boldsymbol{\xi}) = & -\frac{1}{2}(N + IT)\ln(2\pi) - \frac{1}{2}N\ln(\sigma^2) - \frac{1}{2}IT\ln(\tau^2) \\ & -\frac{1}{2}\frac{(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{U}\boldsymbol{\xi})'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{U}\boldsymbol{\xi})}{\sigma^2} \\ & -\frac{1}{2}\frac{\boldsymbol{\xi}'\boldsymbol{\xi}}{\tau^2}, \end{aligned} \tag{2.7}$$

knowing the observations and a current value of the parameters.

Detailed formulas for the EM algorithm for linear mixed models can be found, for instance, in Searle *et al.* (1992, Section 8.3) or in Trottier (1998).

2.2 Alternative models

Model (2.1) introduces a random effect of the gene i at instant t. For a same gene, this random effect changes at each time. A natural alternative model would be to assume an additional random effect of a gene not depending of time. Thus a possible generalization of model (2.1) is a two random effects model

$$y_{itr} = \beta_t + (\xi_i + \xi_{it}) + \varepsilon_{itr} \tag{2.8}$$

where $\xi_i \sim \mathcal{N}(0, \omega^2)$ is the random effect of the gene i on the expression log-ratio.

This model is the most complex linear mixed model that will be considered in this article. The covariance between two different instants for the same gene is equal to ω^2. At a fixed time, the covariance between the repetitions of a gene is $\omega^2 + \tau^2$. The variance of an observation is $\omega^2 + \tau^2 + \sigma^2$. This covariance structure is summarized in

$$\mathrm{cov}(y_{itr}, y_{i't'r'}) = \omega^2\delta_i^{i'} + \tau^2\delta_i^{i'}\delta_t^{t'} + \sigma^2\delta_i^{i'}\delta_t^{t'}\delta_r^{r'}. \tag{2.9}$$

On the other side, a simpler model than (2.1) is

$$y_{itr} = \beta_t + \xi_i + \varepsilon_{itr}. \tag{2.10}$$

It is a one random effect model assuming that there is specific gene random effect. The covariance between two measures of any given gene is equal to ω^2, independently of time and repetition. The variance of an observation is $\omega^2 + \sigma^2$. This model assumes that the correlation between two repeated measures does not depend on time:

$$\mathrm{cov}(y_{itr}, y_{i't'r'}) = \omega^2\delta_i^{i'} + \sigma^2\delta_i^{i'}\delta_t^{t'}\delta_r^{r'}. \tag{2.11}$$

This assumption can be regarded as unrealistic in many situations.

Finally, several LMM are possible to describe gene expression log-ratios:

- Model E1 described by equation (2.10) with one random effect of gene i;
- Model E2 described by equation (2.1) with one random effect of gene i at time t;
- Model E3 described by equation (2.8) with two random effects,

We can also consider model E0 with no random effect. It is described by the equation

$$y_{itr} = \beta_t + \varepsilon_{itr}. \tag{2.12}$$

3 Mixture of linear mixed models

In this section, the finite mixture model is extended to the LMM context in order to propose a model-based cluster analysis tool for repeated data. Since Gaussian mixture is the most employed mixture model, especially in a cluster analysis context (see for instance Yeung *et al.* (2001) and Ghosh and Chinnaiyan (2002) for gene expression clustering), we restrict attention to this model. In a multivariate Gaussian mixture model, it is assumed that an observation $\boldsymbol{y}$ is arising from the mixture distribution

$$f(\boldsymbol{y}) = \sum_{k=1}^{K} p_k \varphi(\boldsymbol{y}|\boldsymbol{\mu}_k, \Gamma_k) \tag{3.1}$$

where $p_k \geq 0$, $k = 1, \ldots, K$ are the mixing proportions verifying the constraint $\sum_{k=1}^{K} p_k = 1$, $\varphi(\cdot\,|\boldsymbol{\mu}_k, \Gamma_k)$ being the density of a Gaussian distribution with mean vector $\boldsymbol{\mu}_k$ and variance matrix Γ_k. Consequently, knowing the mixture component C_k from which an observation arises, its conditional distribution is a Gaussian distribution with mean $\boldsymbol{\mu}_k$ and variance matrix Γ_k.

3.1 Mixture model for repeated gene expression data

In this section, we focus attention on Gaussian mixture models related to linear mixed model E2. To take into account repeated data in the mixture framework, we simply add the assumption that the repeated measures of a gene belong to the same mixture component. This natural assumption allows us to embed LMM in the mixture framework. The specific LMM assumptions to be added in the mixture model (3.1) concern the component mean $\boldsymbol{\mu}_k$ and variance matrix Γ_k. It is assumed that $\boldsymbol{y}^k$, the vector of observations arising from mixture component C_k obeys an LMM equation of the form

$$\boldsymbol{y}^k = \boldsymbol{X}^k \boldsymbol{\beta}_k + \boldsymbol{U}^k \boldsymbol{\xi}^k + \boldsymbol{\epsilon}^k \tag{3.2}$$

$\boldsymbol{\beta}_k$ being the fixed effect vector, $\boldsymbol{\xi}^k$ the random effect vector, $\boldsymbol{X}^k$ and $\boldsymbol{U}^k$ the design matrices.

The observations are arising from one of the K components and those that come from component C_k define a random vector $\boldsymbol{y}^k$ of size $N_k = I_k TR$ where I_k is the number of genes belonging to C_k. For E2 model at hand, vectors $\boldsymbol{y}^k$, $k = 1, \ldots, K$, verify the equation

$$y^k_{itr} = \beta_{kt} + \xi^k_{it} + \epsilon^k_{itr}. \tag{3.3}$$

Thus

$$\boldsymbol{y}^k \sim \mathcal{N}_{N_k}(\boldsymbol{X}^k\boldsymbol{\beta}_k, \tau^2_k \boldsymbol{U}^k(\boldsymbol{U}^k)' + \sigma^2_k \mathbf{Id}_{N_k}) \tag{3.4}$$

where

- $\boldsymbol{X}^k_{(N_k,T)}$ is a design matrix with the same structure as the design matrix $\boldsymbol{X}$ defined in (2.5) and (2.6);
- $\boldsymbol{\beta}_k$ is the fixed effect vector of size T for component C_k, $\boldsymbol{\beta}_k = (\beta_{kt}, t = 1, \ldots, T)$;
- $\boldsymbol{U}^k_{(N_k,I_kT)}$ is a design matrix with the same structure as the design matrix $\boldsymbol{U}$ defined in (2.4);
- τ^2_k is the random effect variance for component C_k;
- σ^2_k is the residual variance specific to each component.

Furthermore, the random effect vector of size T, $\boldsymbol{\xi}^k_i = (\xi^k_{it}, t = 1, \ldots, T)$, and $\boldsymbol{\xi}^k$ are assumed independent.

In order to estimate the parameters of an LMM mixture, we consider the maximum likelihood approach. We make use of the EM methodology that takes into account the incomplete structure of the data. Here missing data are of two types: 1) the indicator vectors $z = (z_i, i = 1, \ldots, I)$ of gene memberships to the mixture components: $z_i = (z^1_i, \ldots, z^K_i)$ with $z^k_i = 1$ if $i \in C_k$ and 0 otherwise, 2) the random effects $\boldsymbol{\xi}^k_i$, $i = 1, \ldots, I$, for each mixture component. The EM algorithm is detailed in the Appendix for LMM mixture model involving E2.

3.2 Remarks

We have considered above a mixture model where parameters $\boldsymbol{\beta}_k$, τ^2_k and σ^2_k are dependent on k. In some situations, it can be useful to constrain the mixture parameters to be fixed upon components and many alternative mixture models can be considered:

- M1: $(\boldsymbol{\beta}_k, k = 1, \ldots, K; \tau^2, \sigma^2)$;
- M2: $(\boldsymbol{\beta}_k, \tau^2_k, k = 1, \ldots, K; \sigma^2)$;
- M3: $(\boldsymbol{\beta}_k, \tau^2_k, \sigma^2_k, k = 1, \ldots, K)$, namely the above mentioned model.

Deriving the EM formulas for those alternative models does not involve any technical difficulty (see for instance Celeux *et al.* (2002) for a presentation of the EM algorithm with model M2).

Obviously, those three mixture model structures can be considered with different assumptions on the random effects. Instead of model E2, model E3, E1 or even model E0 can be considered.

For choosing both an LMM and a mixture model structure, we favoured the BIC (Bayesian Information Criterion) criterion (Schwarz, 1978). This criterion has been proved to be efficient in model selection (Kass and Raftery, 1995) and appears to be one of the most relevant criteria on a practical ground for choosing the number of components in a mixture model (Fraley and Raftery, 1998; Roeder and Wasserman, 1997). The BIC criterion has been defined in a noninformative Bayesian framework to approximate the integrated likelihood of a model. For a model M, BIC (a criterion to be minimized) is minus maximum log-likelihood for model M plus $(\nu_M/2)\ln(n)$,ν_M being the number of free parameters in model M, and n the sample size. Another interesting criterion in the cluster analysis context is the ICL (Integrated Completed Likelihood) criterion, which is an *á la* BIC approximation of the integrated complete likelihood (Biernacki *et al.*, 2000; McLachlan and Peel 2000, Sections 6.10 and 6.11). This criterion will be considered in the numerical experiments too, as well as the classical AIC criterion of Akaike (Akaike, 1974).

4 Numerical experiments

In this section, results of numerical experiments on both simulated and real data sets are reported. Simulation experiments are aiming to assess the ability of the EM algorithm to correctly estimate LMM mixture parameters. Experiments on a real data set, concerning wood formation, aim to highlight the interest of LMM mixture model for gene expression profiles clustering from repeated data.

4.1 Monte Carlo experiments

For each Monte Carlo experiment, we generated 100 samples from each type of simulated data. Two E2–M2 mixture models, denoted (A) and (B) in the following, have been simulated. In both cases, I is fixed to 200, the number T of instants was three and the number R of repetitions is four and a three component E2–M2 mixture model is considered. The mixing proportions were $p_1 = 0.3$, $p_2 = 0.5$ and $p_3 = 0.2$. Fixed effect parameters were $\boldsymbol{\beta}_1 = (0, 0, 2)'$, $\boldsymbol{\beta}_2 = (-1, 0, -1)'$ and $\boldsymbol{\beta}_3 = (1, 2, 0)'$. The random effect variances were $\tau_1^2 = 0.2$, $\tau_2^2 = 0.5$ and $\tau_3^2 = 1$. Models (A) and (B) only differ by the error measure variance: for model (A), it is $\sigma^2_{(A)} = 2$ and $\sigma^2_{(B)} = 3$ for model (B).

Table 1 displays the mean and, in parentheses, the standard error of the estimate parameters obtained with the EM algorithm. The EM algorithm has been initiated from a hierarchical clustering computed with the Ward criterion (Ward, 1963). Table 2 provides the classification error rate using the maximum *a posteriori* (MAP) decision rule from the estimate parameter values $\hat{\boldsymbol{p}}, \hat{\boldsymbol{\theta}}$ obtained with EM. This decision rule consists of assigning all the measures of gene i to mixture component $k(i)$ such that

$$k(i) = \arg\max_k \widehat{t_i(k)} \tag{4.1}$$

where $\widehat{t_i(k)} = P(i \in C_k | \boldsymbol{y}_i, \hat{\boldsymbol{p}}, \hat{\boldsymbol{\theta}})$.

Table 1 Parameter estimation values with EM from 100 simulated models (A) and (B)

	Parameter	Model	Component 1	Component 2	Component 3
(1)	Fixed effects $t=1$		$\beta_{11}=0$	$\beta_{21}=-1$	$\beta_{31}=1$
(2)		(A): $\sigma^2=2$	0.029 (0.124)	−1.007 (0.125)	1.071 (0.274)
(2)		(B): $\sigma^2=3$	0.019 (0.141)	−1.044 (0.156)	1.035 (0.420)
(1)	Fixed effects $t=2$		$\beta_{12}=0$	$\beta_{22}=0$	$\beta_{32}=2$
(2)		(A): $\sigma^2=2$	−0.008 (0.131)	0.004 (0.131)	2.061 (0.373)
(2)		(B): $\sigma^2=3$	−0.009 (0.163)	−0.014 (0.163)	2.030 (0.401)
(1)	Fixed effects $t=3$		$\beta_{13}=2$	$\beta_{23}=-1$	$\beta_{33}=0$
(2)		(A): $\sigma^2=2$	1.994 (0.153)	−0.992 (0.123)	0.008 (0.256)
(2)		(B): $\sigma^2=3$	1.970 (0.208)	−0.999 (0.165)	−0.037 (0.330)
(1)	Proportions		$p_1=0.3$	$p_2=0.5$	$p_3=0.2$
(2)		(A): $\sigma^2=2$	0.301 (0.030)	0.501 (0.041)	0.197 (0.045)
(2)		(B): $\sigma^2=3$	0.308 (0.043)	0.487 (0.060)	0.204 (0.069)
(1)	Random effects		${\tau_1}^2=0.2$	${\tau_2}^2=0.5$	${\tau_3}^2=1$
(2)		(A): $\sigma^2=2$	0.211 (0.092)	0.484 (0.117)	0.901 (0.270)
(2)		(B): $\sigma^2=3$	0.216 (0.134)	0.449 (0.153)	0.866 (0.303)
	Error measure				
(2)		(A): $\sigma^2=2$		2.005 (0.069)	
(2)		(B): $\sigma^2=3$		2.995 (0.095)	

(1) Simulated parameter values.
(2) Mean and (standard error) for parameter estimations.

Table 1 shows that sensible estimates of the model parameters are obtained in both situations (A) and (B). As expected, the estimation accuracy depends on the random effect variances: the greater the variance, the greater the estimation standard error. In the same manner, Table 2 shows that the classification error rate increases with the random effect variances. And, as σ^2 increases (model (B)), the variance of the parameter estimates increases even if the mean estimates remain good. In the same way, comparing the results in Table 2 for models (A) and (B), we note that the classification error rate increases with σ^2.

To assess the role of repetitions for estimating LMM mixture models, we carry out additional Monte Carlo experiments. They consist of 100 replications of model (A) but the number of repetitions is $R=2$ instead of $R=4$. We denoted (A′) this occurrence of model (A). The results for model (A′) displayed in Table 2 clearly show that the classification error rate decreases with the number of repetitions. This confirms our opinion that it is important properly to take into account repetitions to get relevant clustering structures for highly variable data sets.

Table 2 Classification error rates from 100 simulations for models (A), (B) and (A′).

	Model (A) $R=4$, $\sigma^2=2$	Model (B) $R=4$, $\sigma^2=3$	Model (A′) $R=2$, $\sigma^2=2$
Cluster 1	8.30	21.57	14.62
Cluster 2	6.38	21.16	13.82
Cluster 3	23.65	28.88	34.55

4.2 Wood formation DNA microarray analysis

Data we considered in this subsection have been gathered to study the mechanisms involved in wood formation (Hertzberg *et al.*, 2001). We first give a brief presentation on biological and technical aspects of the considered data set. More information concerning biological results, materials and methods can be found in the article of Hertzberg *et al.* (2001).

4.2.1 The data

Hertzberg *et al.* (2001) have studied the developing secondary xylem of poplar by analysing the profiles of 2995 expressed sequence tags (EST). The high organization of secondary xylem allows different developmental zones to be distinguished easily and a unique tissue-specific transcript profile for a well-defined developmental gradient to be determined. This property of wood-forming tissues allowed five tissue samples (A, B, C, D, E) and also a phloem sample (Phl) to be collected. The biological description of these tissues is given in Hertzberg *et al.* (2001).

Expression profiles for $T = 6$ different tissues (Phl, A, B, C, D, E) with DNA chips are to be analysed. To determine the steady-state mRNA levels at specific stages during the ontogeny of wood formation, 30 μm thick sections have been sampled through the wood development region. Those samples have been analysed by using a spotted cDNA-microarray consisting of the 2995 ESTs. Expressions profiles have been obtained by incorporating the Cy5 fluorophor in the experiment samples (Phl, A, B, C, D, E) and the Cy3 fluorophor in the reference sample, namely a mixture of samples A–E. For each experiment, replications have been carried out in order to obtain four measures for each gene. Thus, for each gene, we got $R = 4$ repetitions for the couple of measures (Cy5, Cy3). In the next section, we detail the pretreatment we achieved.

4.2.2 Pretreatment

Before clustering gene expression profiles, two stages appeared to be necessary: data normalization and gene selection.

- *Data normalization*. To remove systematic biases from microarray data, due to technical and biological problems, data must be normalized. For this task, we chose the well-known approach proposed in Yang *et al.* (2001). This normalization allows for a correction depending on intensity level of the spot and is different for each print-tip used to set down the probes on the glass.
- *Gene selection*. Cluster analysis is to be performed on the logarithm of the fluorescent intensities ratios log(Cy5/Cy3) between the two samples. The clustering is usually performed on a subset of genes showing changes between the experiments and the control. For instance, in Hertzberg *et al.* (2001), a hierarchical clustering is presented on the 539 genes showing at a least 8-fold differential expression. A less drastic 4-fold level criterion has been used to select 870 genes.

For this data set of 870 genes, the performances of 154 different mixture models have been compared. They concerned 126 different LMM mixtures, the nine models E*i*–M*j*,

for $i = 1, 2, 3$ and $j = 1, 2, 3$ with $K = 2$ to $K = 15$ components, and the 28 models E0–M1 and E0–M3 with $K = 2$ to $K = 15$ components.

4.2.3 *EM initialization*

EM solution can highly depend on its starting position, especially in a multivariate context. This jeopardizes statistical analysis of mixture. Spurious or insensible estimates can be derived from some starting values, especially for mixture models involving many parameters (see for instance McLachlan and Peel, 2000). Thus, the EM algorithm should be applied from a wide choice of starting values to be ensured to get a sensible maximum. To attenuate the starting value problem of EM, we made use of a strategy consisting of ten short runs of EM from different k-means results, followed by a long run of EM from the 'short run' solution providing the highest log-likelihood. By 'short run' of EM, we mean that the EM algorithm is stopped after a few iterations (ten in the present numerical experiments). This kind of approach has been proved to be efficient in many cases (Berchtold, 2004; Biernacki *et al.*, 2003).

Figure 1 displays the log-likelihood for the different models in competition. As expected, the log-likelihood increases with the model complexity. At least, it shows that, for each considered model, EM is not trapped in a suboptimal or spurious solution.

4.2.4 *Model selection*

The 154 models have been compared with penalized log-likelihood criteria AIC, BIC and ICL. For simplicity, we do not report their values. All those criteria show a marked superiority of mixture model M3, which allows for a different error measure variance for each mixture component. This fact illustrates the variability problems occurring with microarray data. Figure 2 displays AIC, BIC and ICL values for the models E0–M3, E1–M3, E2–M3 and E3–M3. The three criteria strongly support model E2–M3 and E3–M3. Since they provide quite similar values for both models, we chose the most parsimonious model E2–M3. From Figure 2, it appears that those criteria, even ICL, indicate no evidence for a particular number of components. Here, the overlapping of the groups highlights the difficulty to choose an appropriate number of components. We chose to focus on the 13 component mixture for the following reason. With 13 components the smallest proportion is equal to 3.8% (33 genes) and with 14 components, it is 2.2% (19 genes). Since the vector parameter $(p_k, \beta_k, \tau_k, \sigma_k)$, associated to component k is a nine-dimension vector, it did not seem reasonable to select a 14 component mixture. The parameter estimations for $K = 13$ are given in Table 3.

Before presenting the clustering associated to the 13 component mixture model E2–M3, it is of interest to compare models E0–M3 (no random effect) and E2–M3. Table 4 gives log-likelihood values for the two models E0 and E2 in the no mixture case ($K = 1$) and in the $K = 13$ component mixture case. Comparing these values with the log-likelihood of the null model (without fixed effect), which is $-34\,342$, it appears clearly that taking into account random effects leads to a dramatic improvement of the model.

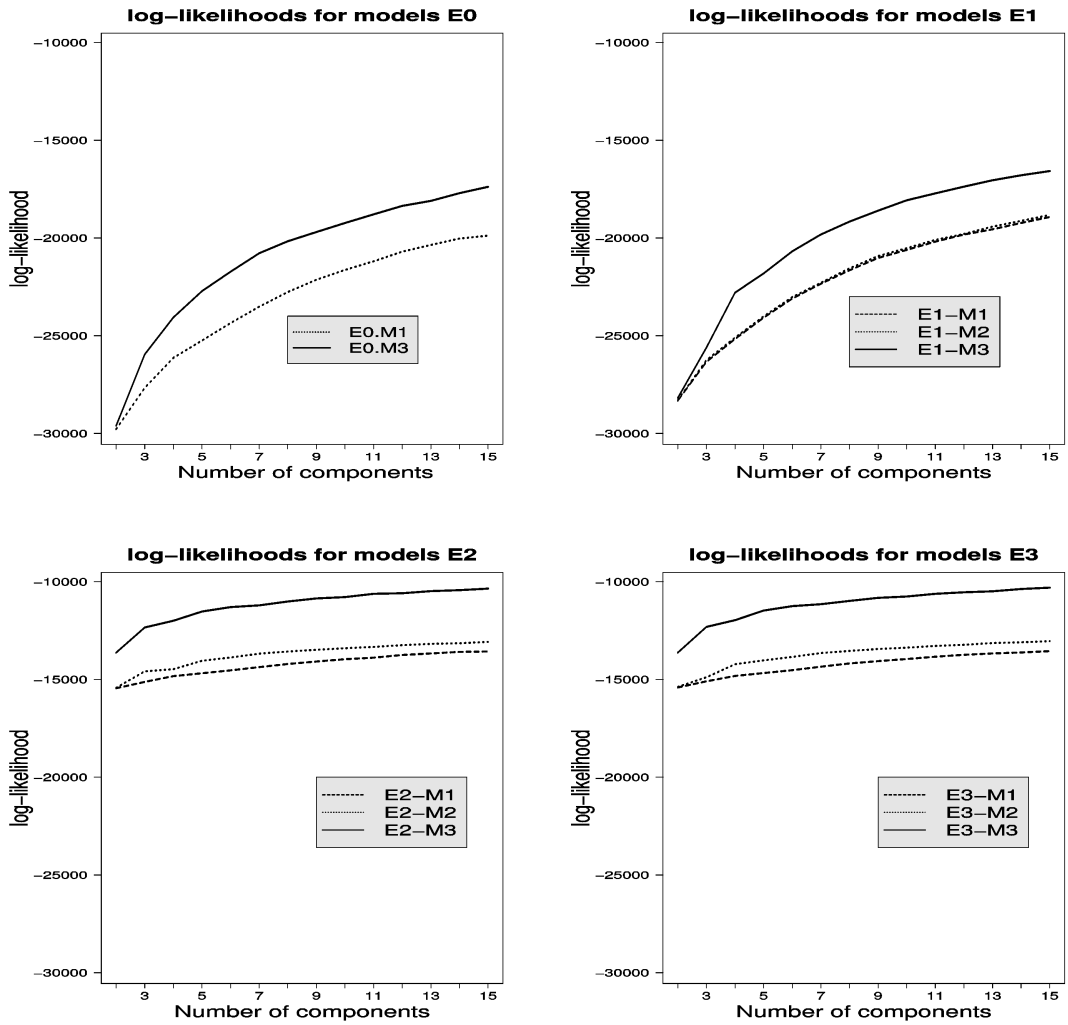

Figure 1 Maximum log-likelihood values for the different models with $K=2$ to 15. The four graphics represent the log-likelihood for the four models Ei ($i=1,\ldots,4$) combined with the different mixture models. In the top left panel, the log-likelihoods for the models E1–M2 and E1–M3 are so close that it is difficult to distinguish the two curves

4.2.5 Cluster analysis

The 13 clusters obtained from the MAP operator (4.1) for model E2-M3 are depicted in Figure 3. For each cluster and each gene, all the repetitions are represented.

Despite the fact that penalized log-likelihood criteria show that model E0–M3 is irrelevant for this data set and could lead to unreliable clustering of the genes, since it does not take into account differences in the gene variability, it is interesting to look at the results obtained wih this model for $K=13$ to analyse the consequences of neglecting

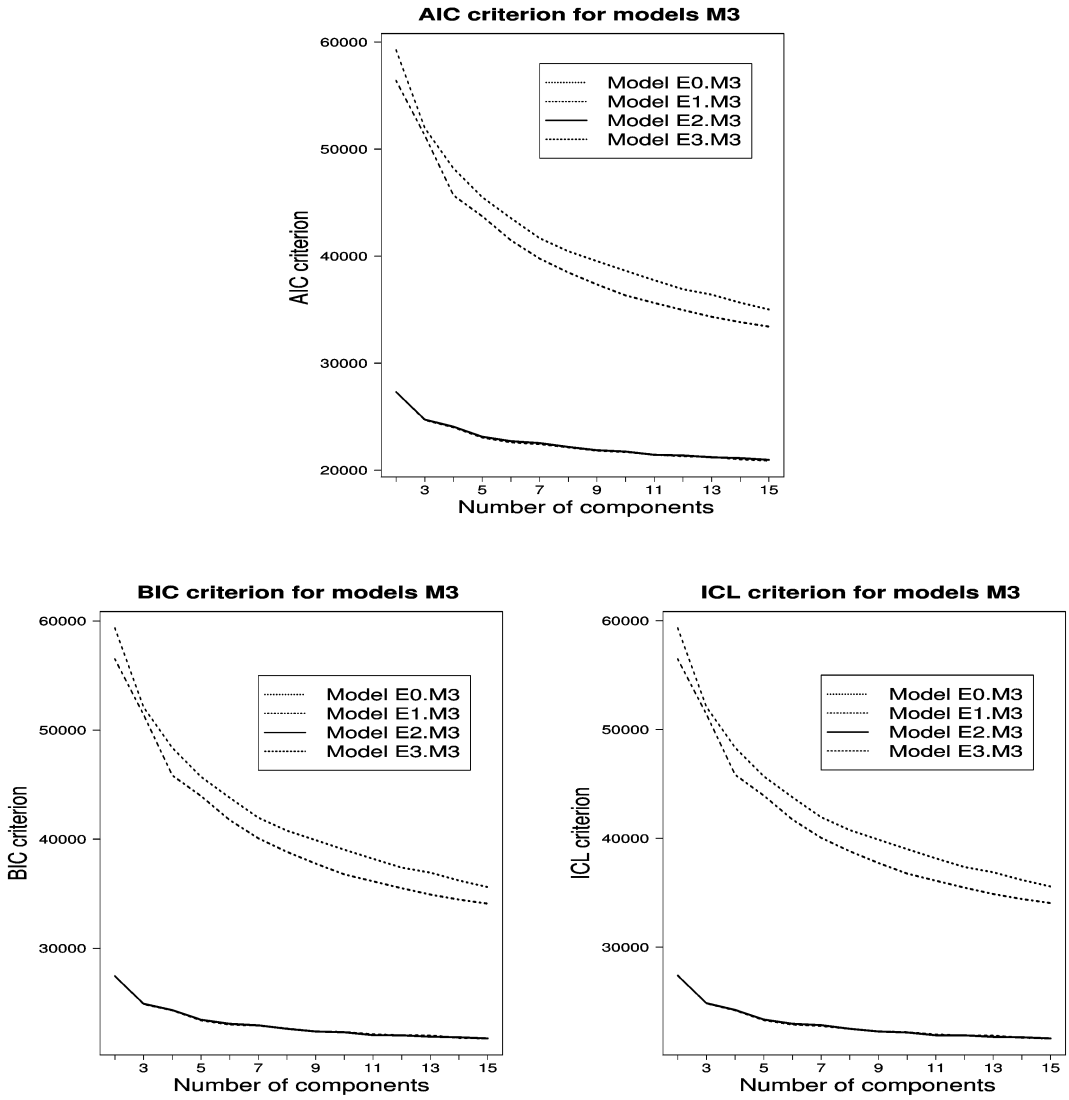

Figure 2 AIC, BIC and ICL criteria for the different models with $K=2$ to 15. The three graphics represent AIC (top), BIC (bottom left) and ICL (bottom right) criteria for the different mixture model M3 and for $K=2,\ldots,15$ components. The values for models E2–M3 and E2–M3 are so close that it is difficult to distinguish the curves for the three criteria

the possible random effects. The parameter estimations for model E0–M3 are given in Table 5 and Figure 4 represents the cluster profiles.

Table 6 compares the classifications derived from models E0–M3 and E2–M3. It gives the percentage of common genes between the clusters of the two classifications. This table deserves some remarks.

Table 3 Parameter estimations for the 13 component mixture with model E2–M3 for the wood formation dataset

Component	Parameter β	Proportion	Random effect std. dev.	Measure error
1	$\beta_1 = (-0.230, 0.347, 0.442, 0.437, 0.433, -1.474)'$	$p_1 = 0.077$	$\tau_1 = 0.462$	$\sigma_1 = 0.224$
2	$\beta_2 = (-1.749, -0.212, 0.023, 0.082, 0.137, -0.423)'$	$p_2 = 0.073$	$\tau_2 = 0.356$	$\sigma_2 = 0.219$
3	$\beta_3 = (0.086, -0.362, -0.246, 0.269, 1.621, 1.005)'$	$p_3 = 0.172$	$\tau_3 = 0.183$	$\sigma_3 = 0.143$
4	$\beta_4 = (-0.499, -0.808, -0.122, 1.611, 2.820, 2.253)'$	$p_4 = 0.038$	$\tau_4 = 0.767$	$\sigma_4 = 0.385$
5	$\beta_5 = (0.685, 0.541, 0.390, -0.268, -1.465, 0.020)'$	$p_5 = 0.078$	$\tau_5 = 0.401$	$\sigma_5 = 0.274$
6	$\beta_6 = (-2.214, -0.611, 0.242, 0.933, 0.934, 0.338)'$	$p_6 = 0.061$	$\tau_6 = 0.786$	$\sigma_6 = 0.348$
7	$\beta_7 = (2.942, 3.045, 2.907, 0.472, -1.708, 0.795)'$	$p_7 = 0.055$	$\tau_7 = 0.705$	$\sigma_7 = 0.711$
8	$\beta_8 = (1.911, 1.847, 1.245, 0.196, -0.940, -0.537)'$	$p_8 = 0.075$	$\tau_8 = 0.694$	$\sigma_8 = 0.212$
9	$\beta_9 = (-0.287, -0.385, -0.226, 0.343, 1.433, 1.297)'$	$p_9 = 0.105$	$\tau_9 = 0.531$	$\sigma_9 = 0.215$
10	$\beta_{10} = (1.027, -0.016, -0.524, -0.427, 1.320, 2.378)'$	$p_{10} = 0.067$	$\tau_{10} = 0.584$	$\sigma_{10} = 0.373$
11	$\beta_{11} = (1.328, 0.665, 0.157, -0.452, -0.157, 1.566)'$	$p_{11} = 0.092$	$\tau_{11} = 0.790$	$\sigma_{11} = 0.201$
12	$\beta_{12} = (1.721, 1.598, 1.117, -0.474, -1.567, 0.853)'$	$p_{12} = 0.063$	$\tau_{12} = 0.639$	$\sigma_{12} = 0.458$
13	$\beta_{13} = (1.586, 2.000, 1.730, 1.071, -0.050, -1.343)'$	$p_{13} = 0.044$	$\tau_{13} = 1.243$	$\sigma_{13} = 0.432$

Std. dev. = standard deviation.

- A wide similarity between clusters 1, 2 and 3 for the two models can be noticed. All these clusters present a small random effect ($\tau_1 = 0.462$, $\tau_2 = 0.356$, $\tau_3 = 0.183$) in model E2–M3. Thus, neglecting the random effect in these cases would not affect greatly the composition of the clusters.
- Clusters 4–12 present a weak agreement between the two classifications (percentages of common genes are between 35% and 71%). A more marked random effect can be observed for these clusters: neglecting the random effects for these clusters can lead to unreliable results.
- It appears that there is no relation between clusters 13 from models E0–M3 and E2–M3. Table 3 shows that the cluster 13 from model E2–M3 is the one with the greatest random effect ($\tau_1 3 = 1.243$). This cluster cannot be recovered when

Table 4 Log-likelihood values for models E0–M3 and E2–M3 with $K = 1$ and $K = 13$ components

	Model E0–M3	Model E2–M3
$K = 1$	−34257	−16267
$K = 13$	−18099	−10485

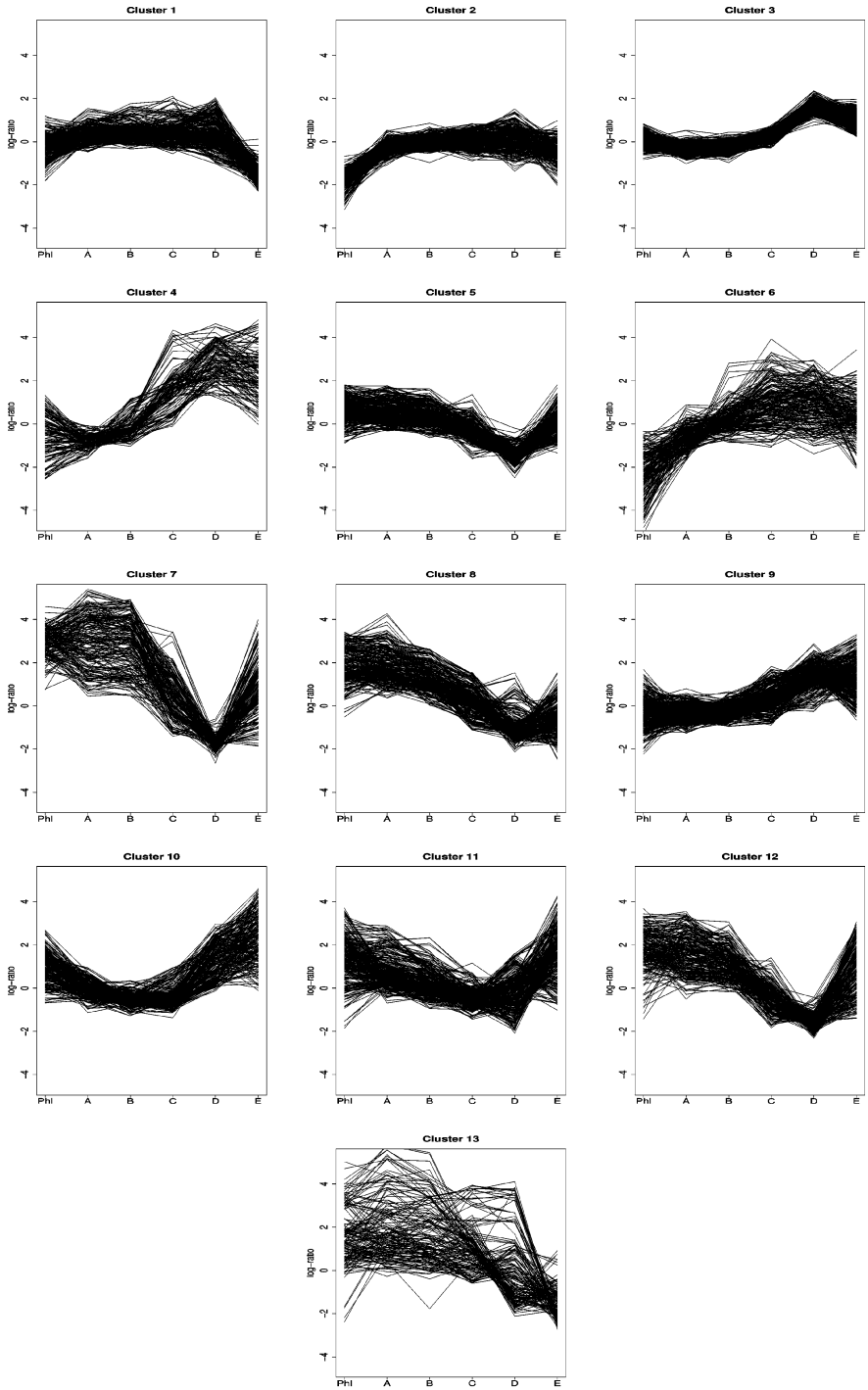

Figure 3 Gene expression profiles of the 13 clusters for model E2–M3

Table 5 Parameters estimation for the 13 components of the mixture model E0–M3

Component	Parameter β	Proportion	Measure error
1	$\beta_1 = (-0.210, 0.379, 0.461, 0.480, 0.435, -1.498)'$	$p_1 = 0.083$	$\sigma_1 = 0.529$
2	$\beta_2 = (-1.759, -0.225, 0.007, 0.056, 0.100, -0.388)'$	$p_2 = 0.081$	$\sigma_2 = 0.423$
3	$\beta_3 = (0.076, -0.370, -0.256, 0.266, 1.603, 0.971)'$	$p_3 = 0.174$	$\sigma_3 = 0.228$
4	$\beta_4 = (-0.598, -0.812, -0.079, 1.682, 3.105, 2.659)'$	$p_4 = 0.030$	$\sigma_4 = 0.813$
5	$\beta_5 = (0.763, 0.647, 0.469, -0.277, -1.468, 0.200)'$	$p_5 = 0.106$	$\sigma_5 = 0.533$
6	$\beta_6 = (-2.275, -0.563, 0.482, 1.358, 1.290, 0.167)'$	$p_6 = 0.053$	$\sigma_6 = 0.954$
7	$\beta_7 = (2.870, 3.221, 2.999, 0.824, -1.329, 0.170)'$	$p_7 = 0.070$	$\sigma_7 = 1.112$
8	$\beta_8 = (1.708, 1.542, 1.112, 0.282, -0.935, -0.917)'$	$p_8 = 0.087$	$\sigma_8 = 0.681$
9	$\beta_9 = (-0.008, -0.470, -0.267, 0.530, 1.675, 1.126)'$	$p_9 = 0.093$	$\sigma_9 = 0.517$
10	$\beta_{10} = (1.043, 0.089, -0.474, -0.431, 1.467, 2.974)'$	$p_{10} = 0.050$	$\sigma_{10} = 0.574$
11	$\beta_{11} = (1.779, 0.351, -0.205, -0.592, 0.370, 1.127)'$	$p_{11} = 0.054$	$\sigma_{11} = 0.614$
12	$\beta_{12} = (2.040, 1.971, 1.196, -0.581, -1.374, 1.714)'$	$p_{12} = 0.074$	$\sigma_{12} = 0.685$
13	$\beta_{13} = (-0.641, -0.085, -0.192, -0.288, 0.473, 1.699)'$	$p_{13} = 0.045$	$\sigma_{13} = 0.524$

random effects are neglected. And, this cluster could be interesting from a genetic point of view: the more deep the tissues are in the poplar, the more the genes of this cluster are underexpressed.

Prediction of random effects and cluster expression profiles. We have highlighted the important role of random effects in the clustering context. Models that do not take into account random effects can lead to unreliable interpretation. Now, with a mixture of LMM, it is important to analyse the random effects in each cluster. First, it is interesting to remove the random effects to see the variability due to error measure in each cluster. Secondly, it is interesting to predict the gene expression profiles in each cluster.

Removing the random effects consists of representing the profiles with $y_{itr} - \xi_{it}^k$, if gene i is assigned to cluster k, rather than y_{itr}. The unknown ξ_{it}^k's can be estimated with their Best Linear Unbiased Predictor (BLUP) $\hat{\xi}_{it}^k$ (Searle *et al.*, 1992). Figure 5 represents the resulting 13 profiles for model E2–M3 with all the repetitions for each gene. An interesting interpretation of the clustering built with a mixture of LMM can be deduced from examination of Figures 3 and 5. Figure 5 provides a clearer representation of the cluster expression profiles than Figures 3 and 4. And, comparison of Figure 3 and Figure 5 throws light on the respective roles of error measurements and random Effects in the clusters' variability. For instance, cluster 11 has a marked Profile with a wide random effect. On the other hand, it appears that cluster 7 has an important measure error variability, and so on. Moreover, if gene i is assigned to cluster k, a natural prediction of its differential expression Profile is $\beta_k + \hat{\xi}_{it}$.

Biological comments. By analysing two metabolic pathways related to cell-wall formation and presented in Hertzberg *et al.*, 2001, page 14 736, we found some similar results to those described in this article. The first pathway concerns selected steps in the carbohydrate metabolism. For this pathway, we observed that five genes of the cluster 13 were implied in connected enzymatic reactions (three genes in the reaction

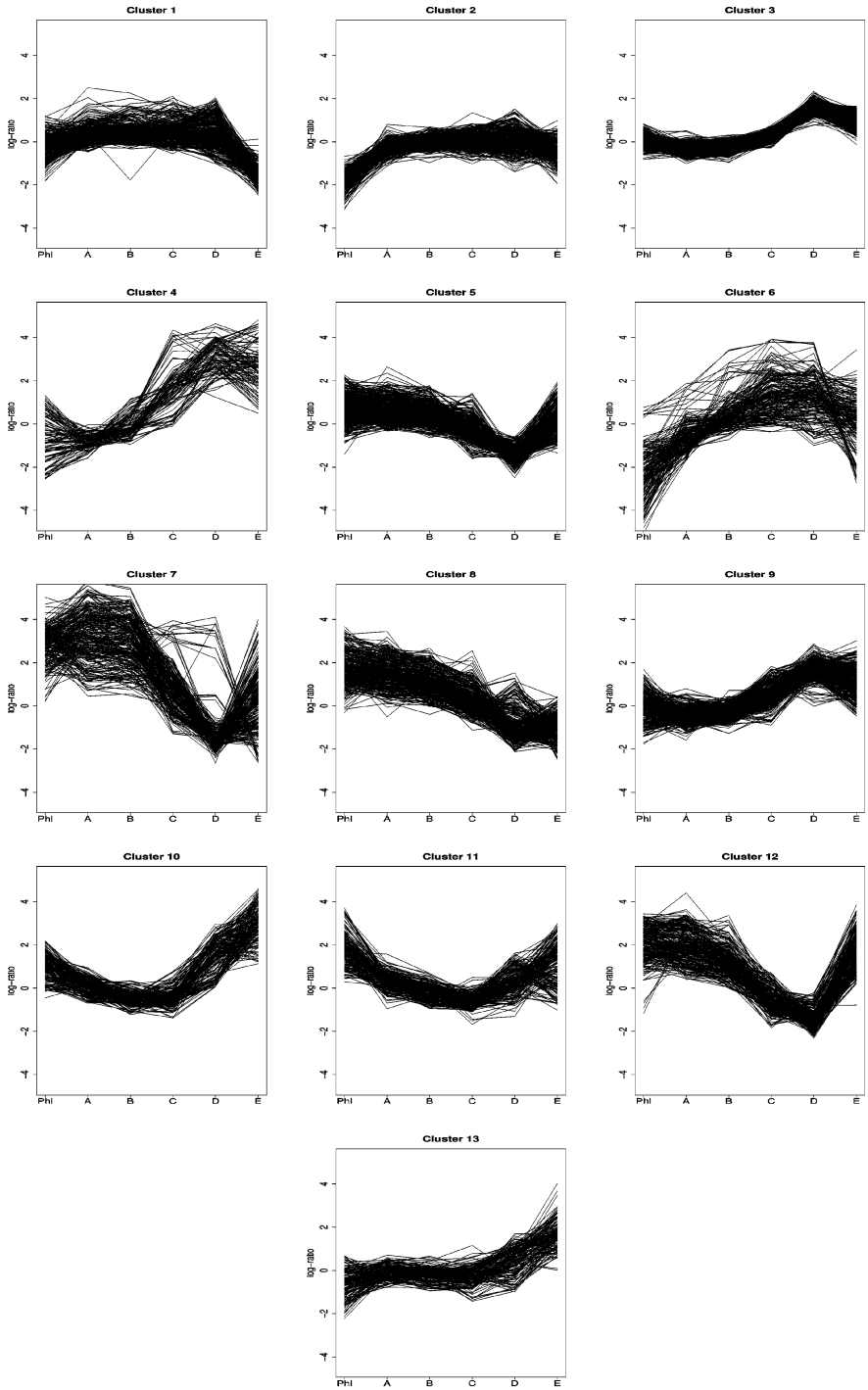

Figure 4 Gene expression profiles of the 13 clusters for model E0–M3

Table 6 Confusion table (in percentage) for the classifications derived from models E0-M3 and E2-M3

Model E2–M3	Model E0–M3 Cluster													Number of genes
	1	2	3	4	5	6	7	8	9	10	11	12	13	
1	**91.7**	0	0	0	0	0	0	0	0	0	0	0	0	66
2	0	**90.1**	0	0	0	0.9	0	0	0	0	0	0	0	65
3	0	0	**89.4**	0	0	0	0	0	3.2	0	0	0	0	150
4	0	0	0	**70.6**	0	1.3	0	0	6.7	0	0	0	0	32
5	0	0	0	0	**70.1**	0	0	1.4	0	0	0	0	0.9	71
6	0.8	5.2	0	0	0	**64.4**	0	0	2.4	3.7	0	0	3.5	51
7	0	0	0	0	0	0	**63.6**	0	0	0	0	4.8	0	47
8	0	0	0	0	4.6	0	5.9	**44.8**	0	0	0	5.8	0	64
9	0	0	4.2	1.7	0	2.2	0	0	**44.5**	3.8	0	0	17.9	93
10	0	0	0	0	0	0	0	0	7	**43.1**	13.8	0	6.5	59
11	0	0	0	0	4.9	0	0	0	0	7	**35.5**	16.5	8.3	78
12	0	0	0	0	7.9	0	0	11	0	0	1	**35.2**	0	56
13	4.8	0	0	0	0	3.7	13.8	17.7	0	0	1.2	0	**0**	38
Number of genes	72	70	153	26	94	46	61	75	79	44	48	63	39	870

EC4.2.2.2, one gene in the reaction EC3.1.1.11 and one gene in the reaction EC3.2.15). Concerning the metabolic pathway for lignin biosynthesis, the genes implied in the successive reactions EC1.14.13.11, EC6.2.1.12, EC1.2.1.44 and EC1.1.1.195 were in cluster 10 or in cluster 6. It could be thought of as surprising that connected reactions genes belong to clusters with quite different profiles. Our view is that the biological interpretation of this clustering remains difficult. But, it could be used to infer new biological hypotheses and new experiments to verify them.

4.2.6 The software

All the results have been obtained using R software (http://www.r-project.org) and a library, namely l3m, which has been developed for these mixture models, is available from the corresponding author upon request. Concerning the computation time, it depends on the number of genes and on the number of components. For the 870 genes of our data set, parameter estimations take about 15 minutes for model E2–M3 with $K = 13$ components. However, a great improvement could be expected by using C language rather than R language.

The parameters and information criteria are computed in the program l3m(). A simulated data set, namely dataL3M, of 100 statistical units is available in the library l3m in order to present the use of different functions. To obtain the parameter estimation for model E2–M3 with $K = 3$, two commands are necessary. First, the data set needs to be loaded using the command data(datal3m). Secondly, the model parameters' estimation is carried out with the command l3m(datal3m, model = 'E2.M3', K = 3).

The R command help() provides access to documentation on different functions of the library. For example, the command help(l3m) gives informations on different arguments and different results for the function l3m().

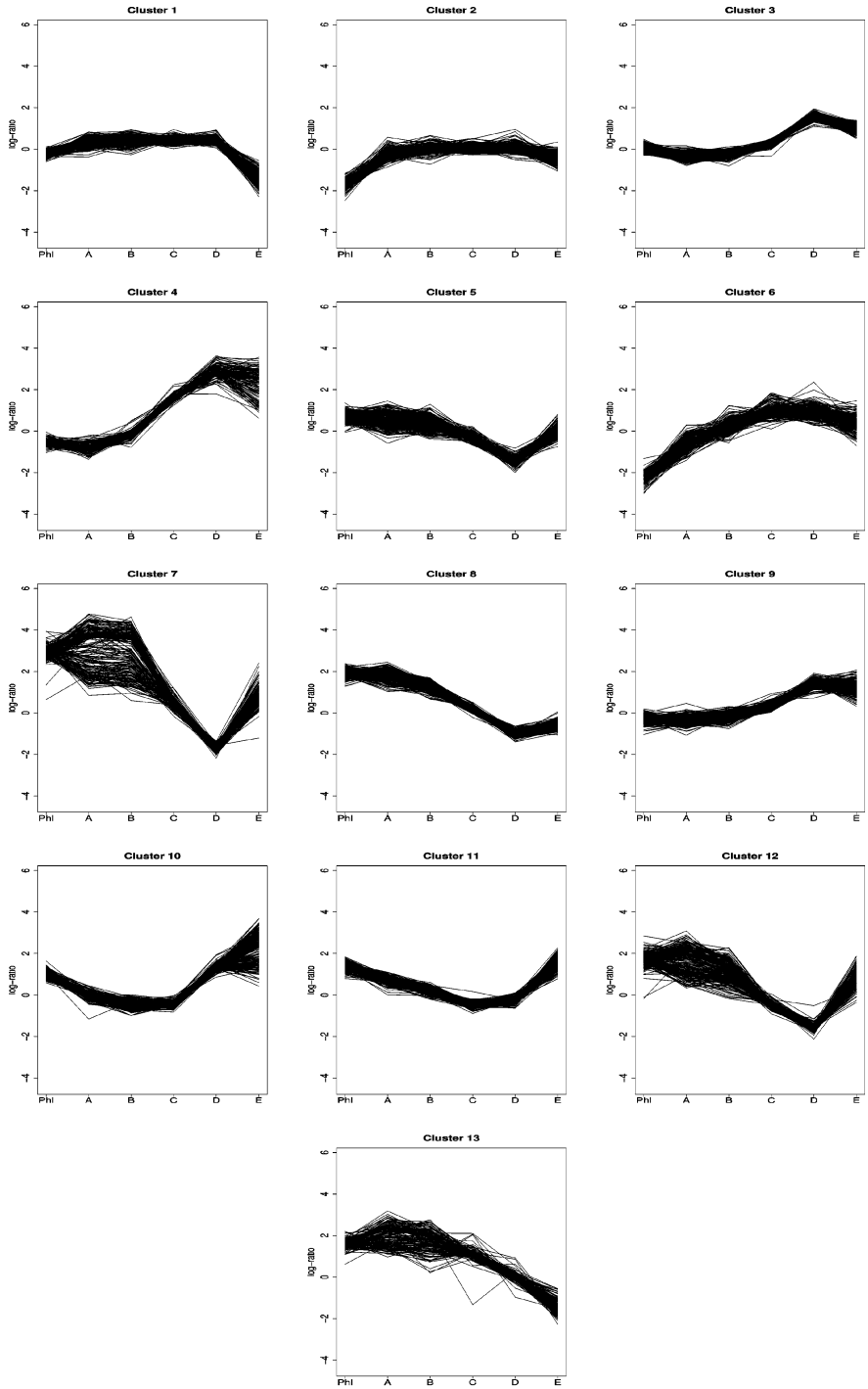

Figure 5 Cluster expression profiles

5 Discussion

A mixture of LMM models has been proposed and estimated with the EM algorithm. It can be useful in situations where repeated measures are available. In a cluster analysis context, it is expected to lead to more reliable clustering structures than standard model-based clustering since it allows profiting from the powerful LMM methodology in the mixture framework. And, in many situations, it can be crucial to distinguish the observations according to their variability.

In microarray data analysis it could have many applications and could become a reference method for clustering gene expression profiles when the variability is important. Moreover, using the BLUP (Best Linear Unbiased Predictor) (Searle *et al.*, 1992) can be useful to build realistic profiles, providing a precise representation of each gene in its cluster.

Moreover, the analysis of LMM mixture models that can lead to interesting interpretations, useful for the practitioners since those models are powerful and parsimonious models.

As we have seen, the LMM mixture model can lead to numerous models that can be reliable in specific situations (for instance mixture components can have the same fixed effects but different random effects). This is the reason why it is important to propose an efficient way to select a reliable model. In this paper, we have proposed to choose an LMM mixture with the BIC criterion, or the ICL criterion when cluster analysis is the main concern. In the application we considered, those penalized log-likelihood criteria appeared to be useful in selecting a model, but not in selecting a sensible number of mixture components. The question of defining an appropriate penalized log-likelihood criterion to select a model would deserve future research. As shown in Vaida and Blanchard (2003), it is possible that the way of counting the effective number of parameters has to be modified when the focus is on conditional inference. Also, the approach proposed in Birgé and Massart (2001) could be of interest, especially in a moderate sample size setting.

We have presented the EM algorithm for an LMM mixture model using the ML method for estimating the vector parameter $\boldsymbol{\theta}$. An alternative method for estimating the parameters of an LMM model is the restricted maximum likelihood (REML) method, which can be regarded as a method of estimation of the variance components by maximizing the marginal likelihood obtained by integrating the likelihood over the fixed effect parameter $\boldsymbol{\beta}$. In the mixture context, we do not consider REML estimation. Actually, it seems that there is no sensitive difference between EM and REML estimates in LMM (Searle *et al.*, 1992, Section 6.7), and considering REML estimation in this context would involve technical complications without providing the ML estimation of the fixed effect parameters $\boldsymbol{\beta}_k$, $k = 1, \ldots, K$. This is a real drawback because the maximum likelihood value enters the composition of penalized likelihood criteria to select a parsimonious model. However, considering the REML approach could be of interest in the mixture context because it can provide more reliable estimates than ML for small proportion mixture components. Hence, considering the implementation of the REML approach using stochastic versions of EM (McLachlan and Krishnan 1997, Section 6.3) could be profitable.

Acknowledgements

We thank Dr. Xavier Gidrol and the members of the Service de Génomique Fonctionnelle (CEA, SGF, Evry, France) for several discussions on the DNA chips technology. We also thank the associate editor and the two referees for their helpful comments and suggestions to improve the presentation of this article.

References

Akaike H (1974) A new look at the statistical model identification. *IEEE Transactions on Automatic Control* **19**, 716–23.

Berchtold A (2004) Optimisation of mixture models: Comparison of different strategies. *Computational Statistics* **19**, 385–406.

Biernacki C, Celeux G and Govaert G (2000) Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE, Trans. on PAMI* **22**, 719–25.

Biernacki C, Celeux G and Govaert G (2003) Choosing starting values for the EM algorithm for getting the highest likelihood in multivariate gaussian mixture models. *Computational Statistics & Data Analysis* **41**, 561–75.

Birgé L and Massart P (2001) Gaussian model selection. *Journal of European Mathematical Society* **3**, 203–68.

Celeux G, Lavergne C and Martin O (2002) Mixture of linear mixed models – application to repeated data clustering. Inria Research Report 4566.

Dempster AP, Laird NM and Rubin DB (1977) Maximum likelihood for incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society B*, **39**, 1–38.

Duggan D, Bitnner M, Chen Y, Meltzer P and Trent J (1999) Expression profiling using cDNA microarrays. *Nature Genetics Supplement* **21**, 10–14.

Efron B, Tibshirani R, Goss V and Chu G (2000) Microarrays and their use in a comparative experiment. Dept Statistics, Stanford Univ, Nov 2000.

Eisen M, Spellman P, Brown P and Botstein D (1998) Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences* **95**, 14863–68.

Fraley C and Raftery AE (1998) How many clusters? Which clustering method? Answers via model-based cluster analysis. *Computer Journal* **41**, 578–88.

Ghosh D and Chinnaiyan A (2002) Mixture modelling of gene expression data from microarray experiments. *Bioinformatics* **18**, 275–86.

Hertzberg M, Aspeborg H, Schrader J, Andersson A *et al.* (2001) A transcriptional roadmap to wood formation. *PNAS* **98**(25), 14732–37.

Kass RE and Raftery AE (1995) Bayes factors. *Journal of the American Statistical Association* **90**, 733–95.

Lee M, Kuo F, Whitmore G and Sklar J (2000) Importance of replication in microarray gene expression studies: Statistical methods and evidence from repetitive cDNA hybridizations. *Proceedings of the National Academy of Sciences* **97**, 9834–39.

McLachlan G and Krishnan T (1997) *The EM algorithm and extensions*. Wiley. Q2

McLachlan G and Peel D (2000) *Finite mixture models*. Wiley. Q2

Roeder K and Wasserman L (1997) Practical Bayesian density estimation using mixtures of normals. *Journal of the American Statistical Association* **92**, 894–902.

Schwarz G (1978) Estimating the dimension of a model. *Annals of Statistics* **6**, 61–464. Q3

Searle S, Casella G and McCulloch C (1992) *Variance components*. John Wiley & Sons. Q2

Tamayo P, Slonim D, Mesirov J *et al.* (1999) Interpreting patterns of gene expresion with self-organizing maps: methods and application to hematopoietic differentiation. *Proceedings of the National Academy of Sciences* **96**, 2907–12.

Tavazoie S, Hughes J, Campbell M, Cho R and Church G (1999) Systematic determination of genetic network architecture. *Nature Genetics* **22**, 281–85.

Trottier C (1998) *Estimation dans les modèles linéaires généralisés à effets aléatoires*, PhD thesis, Institut National de Polytechnique de Grenoble.

Vaida F and Blanchard S (2003) Conditional Akaike information for mixed Effects models. Dept of Biostatistics, Harvard School of Public Health, 2003.

Ward J (1963) Hierarchical groupings to optimize an objective function. *Journal of the American Statistical Association* **58**, 234–44.

Wolfinger R, Gibson G, Wolfinger E *et al.* (2001) Assessing gene signifiance from cDNA microarray expression data via mixed models. *Journal of Computational Biology* **8**, 625–37.

Yang Y, Dudoit S, Luu P and Speed T (2001) Normalization for cDNA microarray data. *In* Bittner M, Chen Y, Dorsel A and Dougherty E eds. *Microarrays: Optical technologies and informatics, Proceedings of SPIE*, vol. 4266. Q2

Yeung K, Fraley C, Murua A, Raftery A and Ruzzo W (2001) Model-based clustering and data transformations for gene expression data. *Bioinformatics* **17**, 977–87.

Yeung K, Medvedovic M and Bumgarner R (2003) Clustering gene-expression data with repeated measurements. *Genome Biology* **4**. Q4

Appendix: EM algorithm for LMM mixture

We detail the EM algorithm for the LMM mixture model involving E2. We denote $\boldsymbol{p} = (p_1, \ldots, p_K)'$ the mixture proportion vector, $\boldsymbol{\theta}_k = (\boldsymbol{\beta}'_k, \tau_k^2, \sigma_k^2)'$ the parameter vector of the LMM associated to component C_k and $\boldsymbol{\theta} = (\boldsymbol{\beta}'_1, \ldots, \boldsymbol{\beta}'_K, \tau_1^2, \ldots, \tau_K^2, \sigma_1^2, \ldots, \sigma_K^2)'$. As noted above, there are two types of missing data: the indicator vectors $\boldsymbol{z} = (z_i, i = 1, \ldots, I)$ and the random effects $(\xi_{it}^k \; mmat = 1, \ldots, T)$, for each gene i in $\mathcal{C}_k$.

The log-likelihood associated to the complete data $(\boldsymbol{y}, \boldsymbol{z}, \boldsymbol{\xi})$ for this model is

$$l(\boldsymbol{\theta}, \boldsymbol{p}|\boldsymbol{y}, \boldsymbol{z}, \boldsymbol{\xi}) = \sum_{i=1}^{I} \sum_{k=1}^{K} z_i^k \ln (p_k \phi(\boldsymbol{y}_i, \boldsymbol{\xi}_i^k | \boldsymbol{\theta}_k)) \tag{A.1}$$

where the vector $\boldsymbol{y}_i$ of size TR contains all the measured ratios for gene i and where $\boldsymbol{\xi}_i^k = (\xi_{i1}^k, \ldots, \xi_{iT}^k)$ denotes the random effect vector of gene i in $\mathcal{C}_k$. Knowing that gene i is in C_k, $\boldsymbol{y}_i$ is a realization from a $\mathcal{N}(\dot{\boldsymbol{X}}\boldsymbol{\beta}_k, \Gamma_k)$ with $\Gamma_k = \tau_k^2 \dot{\boldsymbol{U}} \dot{\boldsymbol{U}}' + \sigma_k^2 \mathbf{Id}_{TR}$ where

$$\dot{\boldsymbol{U}}_{(TR,T)} = \begin{bmatrix} \mathbf{1}_R & \mathbf{0}_R & \cdots & \mathbf{0}_R \\ \mathbf{0}_R & \ddots & \cdots & \vdots \\ \vdots & \cdots & \ddots & \mathbf{0}_R \\ \mathbf{0}_R & \cdots & \mathbf{0}_R & \mathbf{1}_R \end{bmatrix} \text{ is the design matrix associated to } \boldsymbol{y}_i.$$

Therefore, the probability distribution function (pdf) ϕ of component C_k is a Gaussian distribution with mean

$$\begin{bmatrix} \dot{\boldsymbol{X}}\boldsymbol{\beta}_k \\ \mathbf{0}_T \end{bmatrix}$$

and variance matrix

$$\begin{bmatrix} \Gamma_k & \tau_k^2 \dot{U} \\ \tau_k^2 \dot{U}' & \tau_k^2 \mathrm{Id}_T \end{bmatrix}.$$

Thus, we have

$$\begin{aligned} l(\boldsymbol{\theta}, \boldsymbol{p}|\boldsymbol{y}, z, \boldsymbol{\xi}) &= \sum_{i=1}^{I}\sum_{k=1}^{K} z_i^k \ln(p_k) + \sum_{i=1}^{I}\sum_{k=1}^{K} z_i^k \ln(\phi(\boldsymbol{y}_i, \boldsymbol{\xi}_i^k|\boldsymbol{\theta}_k)) \\ &= \sum_{i=1}^{I}\sum_{k=1}^{K} z_i^k \ln(p_k) + \sum_{i=1}^{I}\sum_{k=1}^{K} z_i^k h(\boldsymbol{\theta}_k|\boldsymbol{y}_i, \boldsymbol{\xi}_i^k)) \end{aligned} \tag{A.2}$$

where

$$\begin{aligned} h(\boldsymbol{\theta}_k|\boldsymbol{y}_i, \boldsymbol{\xi}_i^k) = &-\frac{1}{2}(TR + T)\ln(2\pi) - \frac{1}{2}TR\ln(\sigma_k^2) - \frac{1}{2}T\ln(\tau_k^2) \\ &-\frac{1}{2}\frac{(\boldsymbol{y}_i - \dot{\boldsymbol{X}}\boldsymbol{\beta}_k - \dot{\boldsymbol{U}}\boldsymbol{\xi}_i^k)'(\boldsymbol{y}_i - \dot{\boldsymbol{X}}\boldsymbol{\beta}_k - \dot{\boldsymbol{U}}\boldsymbol{\xi}_i^k)}{\sigma_k^2} \\ &-\frac{1}{2}\frac{(\boldsymbol{\xi}_i^k)'(\boldsymbol{\xi}_i^k)}{\tau_k^2}. \end{aligned} \tag{A.3}$$

E step

At iteration $q > 0$, this step consists of computing the expectation of the complete log-likelihood knowing the observed data and a current value of the parameters $\boldsymbol{\theta}^{[q]}$, $\boldsymbol{p}^{[q]}$, $[q]$ denoting the iteration index. In the LMM mixture context it is

$$\begin{aligned} \mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{p}|\boldsymbol{\theta}^{[q]}, \boldsymbol{p}^{[q]}) &= \mathbb{E}\big(l(\boldsymbol{\theta}, \boldsymbol{p}|\boldsymbol{y}, z, \boldsymbol{\xi})|\boldsymbol{y}, \boldsymbol{\theta}^{[q]}, \boldsymbol{p}^{[q]}\big) \\ &= \sum_{i=1}^{I}\sum_{k=1}^{K} t_i^{[q]}(k)\ln(p_k) \\ &\quad + \sum_{i=1}^{I}\sum_{k=1}^{K} t_i^{[q]}(k)\mathbb{E}[h(\boldsymbol{\theta}_k|\boldsymbol{y}_i, \boldsymbol{\xi}_i^k)|\boldsymbol{y}, \boldsymbol{\theta}^{[q]}] \end{aligned} \tag{A.4}$$

where

$$t_i^{[q]}(k) = P(i \in \mathcal{C}_k|\boldsymbol{y}_i, \boldsymbol{\theta}^{[q]}, \boldsymbol{p}^{[q]}) = \frac{p_k^{[q]}\varphi(\boldsymbol{y}_i|\boldsymbol{\theta}_k^{[q]})}{\sum_{l=1}^{K} p_l^{[q]}\varphi(\boldsymbol{y}_i|\boldsymbol{\theta}_l^{[q]})} \tag{A.5}$$

denotes the conditional probability that $\boldsymbol{y}_i$ arises from component C_k.

Since $(\boldsymbol{\xi}_i^k)'(\boldsymbol{\xi}_i^k)$ and $(\boldsymbol{y}_i - \dot{\boldsymbol{U}}\boldsymbol{\xi}_i^k)$, $1 \leq k \leq K$ are sufficient statistics for the complete model (Searle *et al.*, 1992), there is no need to compute the expectation $\mathbb{E}(l(\boldsymbol{\theta}, \boldsymbol{p}|\boldsymbol{y}, z, \boldsymbol{\xi})|\boldsymbol{y}, \boldsymbol{\theta}^{[q]}, \boldsymbol{p}^{[q]})$. To proceed to the maximization of $\mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{p}|\boldsymbol{\theta}^{[q]}, \boldsymbol{p}^{[q]})$, in the M step, we only need to compute the conditional expectation of those sufficient statistics $(\boldsymbol{\xi}_i^k)'(\boldsymbol{\xi}_i^k)$ and $(\boldsymbol{y}_i - \dot{\boldsymbol{U}}\boldsymbol{\xi}_i^k)$, knowing observed data $\boldsymbol{y}_i$ and a current value of the parameters $\boldsymbol{\theta}^{[q]}$ and $p^{[q]}$ (Dempster *et al.*, 1977).

Following (Trottier, 1998, page 49) and knowing that $\boldsymbol{y}_i \in C_k$, we obtain easily

$$\begin{aligned}\mathbb{E}(\boldsymbol{\xi}_i^{k'}\boldsymbol{\xi}_i^k|\boldsymbol{y}_i, \boldsymbol{\theta}^{[q]}) = \tau_k^4(\boldsymbol{y}_i - \dot{\boldsymbol{X}}\boldsymbol{\beta}_k)'\Gamma_k^{-1}\dot{\boldsymbol{U}}\dot{\boldsymbol{U}}'\Gamma_k^{-1}(\boldsymbol{y}_i - \dot{\boldsymbol{X}}\boldsymbol{\beta}_k) \\ + R\tau_k^2 - \tau_k^4\mathrm{tr}(\Gamma_k^{-1}\dot{\boldsymbol{U}}\dot{\boldsymbol{U}})\end{aligned} \tag{A.6}$$

and

$$\mathbb{E}(\boldsymbol{y}_i - \dot{\boldsymbol{U}}\boldsymbol{\xi}_i^k|\boldsymbol{y}_i, \boldsymbol{\theta}^{[q]}) = \dot{\boldsymbol{X}}\boldsymbol{\beta}_k + \sigma^2\Gamma_k^{-1}(\boldsymbol{y}_i - \dot{\boldsymbol{X}}\boldsymbol{\beta}_k). \tag{A.7}$$

M step

This step consists of finding the values maximizing $\mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{p}|\boldsymbol{\theta}^{[q]}, \boldsymbol{p}^{[q]})$. From (A.4), it leads to

$$p_k^{[q+1]} = \sum_{i=1}^{I} \frac{t_i^{[q]}(k)}{I}, \qquad \text{for } k = 1, \ldots, K, \tag{A.8}$$

and to solve the following log-likelihood equations for parameters $\boldsymbol{\beta}_k$, τ_k^2, σ_k^2, for $k = 1, \ldots, K$,

$$\sum_{i=1}^{I} t_i^{[q]}(k) \frac{\partial \mathbb{E}[h(\boldsymbol{\theta}_k|\boldsymbol{y}_i, \boldsymbol{\xi}_i^k)|\boldsymbol{y}, \boldsymbol{\theta}^{[q]}]}{\partial \boldsymbol{\beta}_k} = 0, \tag{A.9}$$

$$\sum_{i=1}^{I} t_i^{[q]}(k) \frac{\partial \mathbb{E}[h(\boldsymbol{\theta}_k|\boldsymbol{y}_i, \boldsymbol{\xi}_i^k)|\boldsymbol{y}, \boldsymbol{\theta}^{[q]}]}{\partial \tau_k^2} = 0, \tag{A.10}$$

and

$$\sum_{i=1}^{I} t_i^{[q]}(k) \frac{\partial \mathbb{E}[h(\boldsymbol{\theta}_k|\boldsymbol{y}_i, \boldsymbol{\xi}_i^k)|\boldsymbol{y}, \boldsymbol{\theta}^{[q]}]}{\partial \sigma_k^2} = 0. \tag{A.11}$$

Using the conditional expectations of the sufficient statistics (A.6) and (A.7), it leads to the following explicit formulas, for $k = 1, \ldots, K$,

$$\sigma_k^{2[q+1]} = \frac{1}{TR\sum_{i=1}^{I} t_i^{[q]}(k)} \sum_{i=1}^{I} t_i^{[q]}(k) \Big[\sigma^{4[q]} (\boldsymbol{y}_i - \dot{\boldsymbol{X}}\boldsymbol{\beta}_k^{[q]})' \Gamma_k^{-1[q]} \Gamma_k^{-1[q]} (\boldsymbol{y}_i - \dot{\boldsymbol{X}}\boldsymbol{\beta}_k^{[q]}) + RT\sigma_k^{2[q]} - \sigma_k^{4[q]} \mathrm{tr}\ (\Gamma_k^{-1[q]}) \Big], \tag{A.12}$$

$$\tau_k^{2[q+1]} = \frac{1}{T\sum_{i=1}^{I} t_i^{[q]}(k)} \sum_{i=1}^{I} t_i^{[q]}(k) \Big[\tau_k^{4[q]} (\boldsymbol{y}_i - \dot{\boldsymbol{X}}\boldsymbol{\beta}_k^{[q]})' \Gamma_k^{-1[q]} \dot{\boldsymbol{U}}\dot{\boldsymbol{U}}' \Gamma_k^{-1[q]} (\boldsymbol{y}_i - \dot{\boldsymbol{X}}\boldsymbol{\beta}_k^{[q]}) + T\tau_k^{2[q]} - \tau_k^{4[q]} \mathrm{tr}\ (\Gamma_k^{-1[q]} \dot{\boldsymbol{U}}\dot{\boldsymbol{U}}') \Big], \tag{A.13}$$

$$\boldsymbol{\beta}_k^{[q+1]} = \frac{1}{\sum_{i=1}^{I} t_i^{[q]}(k)} \sum_{i=1}^{I} t_i^{[q]}(k) \Big[\sigma_k^{2[q]} (\dot{\boldsymbol{X}}'\dot{\boldsymbol{X}})^{-1} \dot{\boldsymbol{X}}' \Gamma_k^{-1[q]} (\boldsymbol{y}_i - \dot{\boldsymbol{X}}\boldsymbol{\beta}_k^{[q]}) + \boldsymbol{\beta}_k^{[q]} \Big]. \tag{A.14}$$

HODDER
JOURNAL.....Statistical Modelling ARTID.....ST096OA

TO: CORRESPONDING AUTHOR

AUTHOR QUERIES - TO BE ANSWERED BY THE AUTHOR

The following queries have arisen during the typesetting of your manuscript. Please answer these queries by marking the required corrections at the appropriate point in the text.

Q1	Please provide running head	
Q2	Please provide location of publisher	
Q3	Page range correct?	
Q4	Please provide page range	
Q5	Please check quality of figs, 3, 4 and 5	