Journal of Multivariate Analysis 100 (2009) 1938–1951

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

journal homepage: www.elsevier.com/locate/jmva

A mixture model-based approach to the clustering of exponential repeated data

M.J. Martinez *, C. Lavergne, C. Trottier

Institut de Mathématiques et de Modélisation de Montpellier, Université Montpellier II, Place Eugène Bataillon, 34095 Montpellier Cedex 5, France

ARTICLE INFO

Article history:
Received 5 April 2007
Available online 3 May 2009

AMS subject classifications:
62J12
62F10
62H12

Keywords:
Generalized linear model
Random effect
Mixture model
EM-algorithm
Metropolis–Hastings algorithm

ABSTRACT

The analysis of finite mixture models for exponential repeated data is considered. The mixture components correspond to different unknown groups of the statistical units. Dependency and variability of repeated data are taken into account through random effects. For each component, an exponential mixed model is thus defined. When considering parameter estimation in this mixture of exponential mixed models, the EM-algorithm cannot be directly used since the marginal distribution of each mixture component cannot be analytically derived. In this paper, we propose two parameter estimation methods. The first one uses a linearisation specific to the exponential distribution hypothesis within each component. The second approach uses a Metropolis–Hastings algorithm as a building block of a general MCEM-algorithm.

1. Introduction

In the past decades, finite mixture models have been extensively developed in the literature. Surveys on these issues can be found in [1–3]. In such finite mixture models, it is assumed that a sample of observations arises from a specified number of underlying groups or classes with unknown proportions and according to a specific form of distribution in each of them. A large number of distributions from the exponential family have been considered such as normal, Poisson, exponential [4]. Wedel and DeSarbo [5] have proposed a mixture of generalized linear models which contains the previously proposed mixtures as special cases, as well as a host of other parametric specifications theretofore not dealt with in the literature. The use of these mixture models can be, in particular, a way to consider unexpected variance in GLM's or also a way to take into account underlying unobserved latent variables forming groups or classes. More recently, Celeux, Martin and Lavergne [6] have proposed a mixture of linear mixed models (LMM) in a microarray data analysis context. The introduction of random effects allowed them to take into account the variability of gene expression profiles from repeated microarray experiments. In our work, we consider the analysis of finite mixtures for exponential repeated data. The mixture components correspond to different possible states of the statistical units. Dependency and variability of exponential repeated data are taken into account through exponential mixed models defined for each mixture component. In the field of the Health Sciences, applications may concern the modelling of lengths of repeated hospital stays for patients belonging to unknown clusters. Another example is the analysis of elimination times after repeated absorptions of a drug by patients not being controlled a priori.

Concerning parameter estimation in the proposed mixture of exponential mixed models, the use of the EM-algorithm which allows us to take into account the incomplete structure of the data is considered [7]. But the algorithm presented by

* Corresponding author.
E-mail address: martinez@math.univ-montp2.fr (M.J. Martinez).

0047-259X/$ – see front matter © 2009 Elsevier Inc. All rights reserved.
doi:10.1016/j.jmva.2009.04.006

Celeux et al. [6] for a mixture of linear mixed models cannot be transferred to the present setup because the marginal distribution of each mixture component cannot be analytically derived. Thus, we propose two parameter estimation methods. The first one uses a linearisation specific to the exponential distribution hypothesis associated with each mixture component. The second approach is adapted from the algorithm presented by McCulloch [8] for generalized linear mixed models (GLMM) [9] and uses a Metropolis–Hastings step [10] to allow construction of an MCEM-algorithm. This algorithm can be adapted to a mixture of any generalized linear mixed models.

The paper is organized as follows. After a description of the model hypotheses in Section 2, we outline the EM-algorithm presented by Celeux et al. for a mixture of linear mixed models in Section 3. This step makes easier the description of the developed algorithm in the exponential case. In Section 4, we describe the two proposed parameter estimation methods. In Section 5, we study the behaviour of these approaches on simulations. We also suggest in this last section a method to determine the appropriate number of mixture components. Finally, in Section 6, in order to illustrate the developed methods, we fit several mixture models to children with autism data set.

2. Mixture of exponential mixed models: Model definition

Consider $y = (y_1', \ldots, y_I')'$ a vector of observations where y_i is associated with the ith statistical unit. Each y_i contains the n_i repetitions y_{ij}. Consider also different components $\mathcal{C}_k$, $k = 1, \ldots, K$, corresponding to different groups of the statistical units. We assume that all repeated measures of a statistical unit belong to the same component and we define the indicator vectors $z_i = (z_{i1}, \ldots, z_{iK})$, $i = 1, \ldots, I$, with $z_{ik} = 1$ if unit $i \in \mathcal{C}_k$ and 0 otherwise.

To take into account dependency and variability of repeated data, we consider for each component an exponential mixed model with a random effect associated with each unit. This leads to a mixture of exponential mixed models and the density of Y_i may be written as follows:

$$f(y_i|\theta, p) = \sum_{k=1}^{K} p_k f_k(y_i|\theta_k)$$

where the p_k's are mixing weights with $0 < p_k < 1$ for $k = 1, \ldots, K$ and $\sum_{k=1}^{K} p_k = 1$, and $f_k(.|\theta_k)$ denotes the density function of the marginal distribution associated with the exponential mixed model with unknown parameters $\theta_k = (\beta_k, \sigma_k^2)$. Note that this marginal distribution cannot be analytically derived. Indeed, distributional assumptions are made conditionally on the non-observed random effects and an integral calculus which is not feasible except for particular distributions has to be made in order to derive the marginal distribution.

More precisely, given the mixture component $\mathcal{C}_k$ from which unit i arises and given the unobserved random effect ξ_i, the components Y_{ij} are assumed to be independent and exponentially distributed:

$$(Y_{ij}|\xi_i, Z_{ik} = 1) \sim \mathcal{E}xp(\mu_{ij}^k) \quad \text{with} \begin{cases} \mu_{ij}^k = \exp(x_{ij}'\beta_k + u_{ij}\xi_i) \\ (\xi_i|Z_{ik} = 1) \sim \mathcal{N}(0, \sigma_k^2) \end{cases}$$

where

- $\forall i, i' \in \{1, \ldots, I\}^2$ $i \neq i'$, ξ_i and $\xi_{i'}$ are assumed to be independent,
- β_k is the $q \times 1$ fixed effect parameter vector associated with component $\mathcal{C}_k$,
- σ_k^2 is the variance of the random effect associated with component $\mathcal{C}_k$.

For simplicity, with a slightly abusive vectorial notation, we write

$$(Y_i|\xi_i, Z_{ik} = 1) \sim \mathcal{E}xp(\mu_i^k) \quad \text{with} \begin{cases} \mu_i^k = \exp(X_i\beta_k + U_i\xi_i) \\ (\xi_i|Z_{ik} = 1) \sim \mathcal{N}(0, \sigma_k^2) \end{cases}$$

where

$$X_i = \begin{pmatrix} x_{i1}' \\ \vdots \\ x_{in_i}' \end{pmatrix} \quad \text{and} \quad U_i = (u_{i1}, \ldots, u_{in_i})'$$

are the $n_i \times q$ and $n_i \times 1$ known design matrices.

Thus, we focus here on a mixture model-based approach to the clustering of exponential repeated data.

Remark. We have previously noted that the density function of the marginal distribution cannot be derived in a closed form. Note that its general expression is given by

$$f_k(y_i|\theta_k) = \int \prod_{j=1}^{n_i} f(y_{ij}|\xi_i, z_{ik} = 1, \theta_k) f(\xi_i|z_{ik} = 1, \theta_k) \, \mathrm{d}\xi_i$$

with

$$f(y_{ij}|\xi_i, z_{ik} = 1, \theta_k) = \frac{1}{\exp(x_{ij}'\beta_k + u_{ij}\xi_i)} \exp\left\{-\frac{y_{ij}}{\exp(x_{ij}'\beta_k + u_{ij}\xi_i)}\right\}.$$

3. EM-algorithm for a mixture of linear mixed models

In this section, we outline the maximum likelihood estimation approach for a mixture of linear mixed models using the EM-algorithm presented by Celeux et al. [6]. The EM-methodology takes into account the incomplete structure of the data [11]. Missing data are here of two types: the indicator vectors z_i, $i = 1, \ldots, I$ of unit memberships to the mixture components and the random effects ξ_i, $i = 1, \ldots, I$.

Given the mixture component $\mathcal{C}_k$ from which unit i arises, Y_i is here assumed to be normally distributed and modelled by:

$$(Y_i|Z_{ik} = 1) = X_i\beta_k + U_i\xi_i + \varepsilon_i$$

where

- $(\xi_i|Z_{ik} = 1) \sim \mathcal{N}(0, \sigma_k^2)$,
- ε_i is the $n_i \times 1$ error vector assumed to be normally distributed: $\varepsilon_i \sim \mathcal{N}(0, \tau^2 I_{n_i})$ with I_{n_i} the identity matrix of order n_i.
- $\forall i \in \{1, \ldots, I\}$, ε_i and ξ_i are assumed to be independent and $\forall i \neq i' \in \{1, \ldots, I\}^2$, ξ_i and $\xi_{i'}$, respectively ε_i and $\varepsilon_{i'}$, are assumed to be independent,
- β_k, σ_k^2, X_i and U_i are defined as previously.

Thus the distribution of Y_i is a mixture of linear mixed models defined by

$$f(y_i|\theta, p) = \sum_{k=1}^{K} p_k f_k(y_i|\theta_k)$$

where $p = (p_1, \ldots, p_K)$ are the mixing weights, $\theta = (\theta_1, \ldots, \theta_K)$ with $\theta_k = (\beta_k, \sigma_k^2, \tau^2)$ the linear mixed model parameters associated with component $\mathcal{C}_k$, and $f_k(y_i|\theta_k)$ denotes the density function of the Gaussian distribution of Y_i with mean $X_i\beta_k$ and variance matrix $\Gamma_{k,i} = \tau^2 I_{n_i} + \sigma_k^2 U_i U_i'$. In their paper, Celeux et al. [6] consider a mixture model where all parameters are dependent on component $\mathcal{C}_k$. We consider here a mixture model where the parameters β_k and σ_k^2 depend on component $\mathcal{C}_k$ whereas the residual variance τ^2 is the same for all mixture components since this particular situation will be needed in Section 4.1.

The log-likelihood associated with the complete data (y, z, ξ) is given by

$$L(\theta, p|y, z, \xi) = \sum_{i=1}^{I}\sum_{k=1}^{K} z_{ik}\Big\{\ln p_k + \ln f(y_i, \xi_i|z_{ik} = 1, \theta_k)\Big\}$$

where $\ln f(y_i, \xi_i|z_{ik} = 1, \theta_k)$ can be written as

$$\ln f(y_i|\xi_i, z_{ik} = 1, \theta_k) + \ln f(\xi_i|z_{ik} = 1, \theta_k)$$

with

- $$\begin{aligned}\ln f(y_i|\xi_i, z_{ik} = 1, \theta_k) &= -\frac{1}{2}\Big(n_i \ln 2\pi + n_i \ln \tau^2 + \frac{\varepsilon_i'\varepsilon_i}{\tau^2}\Big)\\ &= -\frac{1}{2}\Big\{n_i \ln 2\pi + n_i \ln \tau^2 + \frac{(y_i - X_i\beta_k - U_i\xi_i)'(y_i - X_i\beta_k - U_i\xi_i)}{\tau^2}\Big\}\end{aligned}$$
- $$\ln f(\xi_i|z_{ik} = 1, \theta_k) = -\frac{1}{2}\Big(\ln 2\pi + \ln \sigma_k^2 + \frac{\xi_i^2}{\sigma_k^2}\Big).$$

At iteration $[t + 1]$, the E-step consists of computing the expectation of the complete data log-likelihood given the observed data and a current value of the parameters $(\theta^{[t]}, p^{[t]})$:

$$\begin{aligned}Q(\theta, p|\theta^{[t]}, p^{[t]}) &= E\Big[L(\theta, p|y, z, \xi)|y, \theta^{[t]}, p^{[t]}\Big]\\ &= \sum_{i=1}^{I}\sum_{k=1}^{K} t_k^{[t]}(y_i) \ln p_k - \frac{1}{2}\sum_{i=1}^{I}\sum_{k=1}^{K} t_k^{[t]}(y_i)\left\{(n_i + 1)\ln 2\pi + n_i \ln \tau^2 + \ln \sigma_k^2 + \frac{E_{c\,k}^{[t]}(\varepsilon_i'\varepsilon_i)}{\tau^2} + \frac{E_{c\,k}^{[t]}(\xi_i^2)}{\sigma_k^2}\right\}\end{aligned}$$

where

$$E_{c\,k}^{[t]}(.) = E(.|y_i, z_{ik} = 1, \theta_k^{[t]})$$

and

$$t_k^{[t]}(y_i) = P(Z_{ik}=1|y_i,\theta^{[t]},p^{[t]}) = \frac{p_k^{[t]} f(y_i|z_{ik}=1,\theta_k^{[t]})}{f(y_i|\theta^{[t]},p^{[t]})} = \frac{p_k^{[t]} f_k(y_i|\theta_k^{[t]})}{\sum_{l=1}^{K} p_l^{[t]} f_l(y_i|\theta_l^{[t]})}$$

denotes the estimated posterior probability that unit i arises from component $\mathcal{C}_k$.

The M-step consists of maximising $Q(\theta,p|\theta^{[t]},p^{[t]})$. It leads to the following explicit expressions for $k=1,\ldots,K$:

$$p_k^{[t+1]} = \frac{\sum_{i=1}^{I} t_k^{[t]}(y_i)}{I}$$

$$\beta_k^{[t+1]} = \left(\sum_{i=1}^{I} t_k^{[t]}(y_i) X_i' X_i\right)^{-1} \sum_{i=1}^{I} t_k^{[t]}(y_i)\left\{\tau^{2[t]} X_i' \Gamma_{k,i}^{[t]-1}(y_i - X_i\beta_k^{[t]}) + X_i' X_i \beta_k^{[t]}\right\}$$

$$\sigma_k^{2[t+1]} = \frac{1}{\sum_{i=1}^{I} t_k^{[t]}(y_i)} \sum_{i=1}^{I} t_k^{[t]}(y_i)\left\{\sigma_k^{4[t]}(y_i - X_i\beta_k^{[t]})' \Gamma_{k,i}^{[t]-1} U_i U_i' \Gamma_{k,i}^{[t]-1}(y_i - X_i\beta_k^{[t]}) + \sigma_k^{2[t]} - \sigma_k^{4[t]}\mathrm{tr}(\Gamma_{k,i}^{[t]-1} U_i U_i')\right\}$$

$$\tau^{2[t+1]} = \frac{1}{n}\sum_{i=1}^{I}\sum_{k=1}^{K} t_k^{[t]}(y_i)\left\{\tau^{4[t]}(y_i - X_i\beta_k^{[t]})' \Gamma_{k,i}^{[t]-1} \Gamma_{k,i}^{[t]-1}(y_i - X_i\beta_k^{[t]}) + n_i\tau^{2[t]} - \tau^{4[t]}\mathrm{tr}(\Gamma_{k,i}^{[t]-1})\right\}.$$

Details can be found in [6].

4. Estimation in a mixture of exponential mixed models

We come back now to the parameter estimation for the mixture of exponential mixed models presented in Section 2. In this context, the use of the EM-algorithm is not directly possible. The complete data log-likelihood associated to this model is given by

$$L(\theta,p|y,\xi,z) = \sum_{i=1}^{I}\sum_{k=1}^{K} z_{ik}\ln p_k + \sum_{i=1}^{I}\sum_{k=1}^{K} z_{ik}\ln f(y_i|\xi_i,z_{ik}=1,\theta_k) + \sum_{i=1}^{I}\sum_{k=1}^{K} z_{ik}\ln f(\xi_i|z_{ik}=1,\theta_k)$$

with

- $$\ln f(y_i|\xi_i,z_{ik}=1,\theta_k) = \sum_{j=1}^{n_i}\ln f(y_{ij}|\xi_i,z_{ik}=1), = -\sum_{j=1}^{n_i}\left\{x_{ij}'\beta_k + u_{ij}\xi_i + \frac{y_{ij}}{\exp(x_{ij}'\beta_k + u_{ij}\xi_i)}\right\}$$

  because the y_{ij}'s are independent conditionally on ξ_i.

- $$\ln f(\xi_i|z_{ik}=1,\theta_k) = -\frac{1}{2}\left(\ln 2\pi + \ln\sigma_k^2 + \frac{\xi_i^2}{\sigma_k^2}\right).$$

In this case, at iteration $[t+1]$, the expectation of the complete data log-likelihood given the observed data and a current value of the parameters $(\theta^{[t]},p^{[t]})$ is given by

$$Q(\theta,p|\theta^{[t]},p^{[t]}) = \sum_{i=1}^{I}\sum_{k=1}^{K} t_k^{[t]}(y_i)\ln p_k - \sum_{i=1}^{I}\sum_{k=1}^{K} t_k^{[t]}(y_i)\sum_{j=1}^{n_i}\left[x_{ij}'\beta_k + u_{ij}E_{c\,k}^{[t]}(\xi_i) + \exp(-x_{ij}'\beta_k)E_{c\,k}^{[t]}\left[\exp(-u_{ij}\xi_i)\right]y_{ij}\right]$$
$$- \frac{1}{2}\sum_{i=1}^{I}\sum_{k=1}^{K} t_k^{[t]}(y_i)\left[\ln 2\pi + \ln\sigma_k^2 + \frac{E_{c\,k}^{[t]}(\xi_i^2)}{\sigma_k^2}\right],$$

with $E_{c\,k}^{[t]}(.)$ and $t_k^{[t]}(y_i)$ are defined as in Section 3. Thus, the EM-algorithm leads to formulae depending on conditional expectations $E_{c\,k}^{[t]}(\xi_i^2)$, $E_{c\,k}^{[t]}\left[\exp(-u_{ij}\xi_i)\right]$ and posterior probabilities $t_k^{[t]}(y_i)$, $i=1,\ldots,I$, $k=1,\ldots,K$. Because of the non-availability of the marginal distribution for each mixture component, probabilities $t_k^{[t]}(y_i)$ cannot be derived in closed form.

Furthermore, neither the conditional expectation $E_{c\,k}^{[t]}(\xi_i^2)$ nor $E_{c\,k}^{[t]}\left[\exp(-u_{ij}\xi_i)\right]$ can be computed too since these calculations involve the unknown conditional distribution of ξ_i given y_i. We propose two parameter estimation methods which allow us to get round these problems.

4.1. A method based on linearisation

This first approach is a conceptually simple method which involves two steps: a linearisation specific to the exponential mixed model [12] associated with each mixture component and the use of the EM-algorithm for parameter estimation in a mixture of linear mixed models.

Knowing the component $\mathcal{C}_k$, the conditional distribution associated with statistical unit i is given by

$$\forall j \in \{1, \ldots, n_i\} \quad (Y_{ij}|\xi_i, Z_{ik} = 1) \sim \mathcal{E}xp(\mu_{ij}^k),$$

or equivalently:

$$\forall j \in \{1, \ldots, n_i\} \quad \frac{Y_{ij}}{\mu_{ij}^k} \sim \mathcal{E}xp(1),$$

thus

$$\forall j \in \{1, \ldots, n_i\} \quad \ln(Y_{ij}) - \ln(\mu_{ij}^k) \sim \mathcal{G}umbel,$$

where the Gumbel density function is defined by $\forall t \in \mathbb{R}\ f(t) = \exp(t - \exp(t))$ with mean $\gamma = -0.57722$ and variance $\frac{\pi^2}{6}$.

This enables us to write:

$$\ln(Y_{ij}) - \ln(\mu_{ij}^k) = \gamma + \varepsilon_{ij} \quad \text{where } E(\varepsilon_{ij}) = 0 \text{ and } \operatorname{var}(\varepsilon_{ij}) = \frac{\pi^2}{6}.$$

Using vectorial notation, we obtain

$$\ln(Y_i) - \ln(\mu_i^k) = \gamma + \varepsilon_i \quad \text{where } E(\varepsilon_i) = 0_{n_i} \text{ and } \operatorname{var}(\varepsilon_i) = \frac{\pi^2}{6} I_{n_i}.$$

Defining the variable $D_i = \ln(Y_i) - \gamma$, we end up with the linearised model:

$$D_i = X_i\beta_k + U_i\xi_i + \varepsilon_i$$

with 0-mean error vector ε_i and known variance matrix $\frac{\pi^2}{6} I_{n_i}$, which is considered as a linear mixed model $\mathcal{M}_k$ for the data $d_i = \ln(y_i) - \gamma$ given the component $\mathcal{C}_k$.

Finally, we use the EM-algorithm to estimate the parameters of the mixture of linear mixed models defined by

$$h(d_i|\theta, p) = \sum_{k=1}^{K} p_k h_k(d_i|\theta_k)$$

where $h_k(d_i|\theta_k)$ is the Gaussian density function with mean vector $X_i\beta_k$ and variance matrix $\Gamma_{k,i} = \frac{\pi^2}{6} I_{n_i} + \sigma_k^2 U_i U_i'$. In this approach, note that vector d_i is derived from the data y_i whatever the component $\mathcal{C}_k$ from which unit i arises and without any use of the current value of the parameters.

The parameter estimation for this mixture of linear mixed models using the EM-algorithm described in Section 3 leads to the following expressions for $k = 1, \ldots, K$:

$$p_k^{[t+1]} = \frac{\sum_{i=1}^{I} t_k^{[t]}(d_i)}{I}$$

$$\sigma_k^{2[t+1]} = \frac{1}{\sum_{i=1}^{I} t_k^{[t]}(d_i)} \sum_{i=1}^{I} t_k^{[t]}(d_i)\left\{\sigma_k^{4[t]}\,(d_i - X_i\beta_k^{[t]})'\,\Gamma_{k,i}^{[t]-1} U_i\, U_i'\,\Gamma_{k,i}^{[t]-1}\,(d_i - X_i\beta_k^{[t]}) + \sigma_k^{2[t]} - \sigma_k^{4[t]}\operatorname{tr}(\Gamma_{k,i}^{[t]-1} U_i U_i')\right\}$$

$$\beta_k^{[t+1]} = \left(\sum_{i=1}^{I} t_k^{[t]}(d_i) X_i' X_i\right)^{-1} \sum_{i=1}^{I} t_k^{[t]}(d_i)\left\{\frac{\pi^2}{6} X_i'\,\Gamma_{k,i}^{[t]-1}(d_i - X_i\beta_k^{[t]}) + X_i' X_i \beta_k^{[t]}\right\}$$

where

$$t_k^{[t]}(d_i) = \frac{p_k^{[t]}\, h_k(d_i|\theta_k^{[t]})}{\sum_{l=1}^{K} p_l^{[t]}\, h_l(d_i|\theta_l^{[t]})}.$$

Note that this linearisation method uses the pseudo-likelihood techniques [13,14] which here consist of approximating the original GLMM by a linear mixed model for pseudo-data. Then we use the well-known theory for linear mixed models as outlined in Section 3. The advantage of this approach is that it is fast since it is not required to integrate out the unobserved random effects. A disadvantage is that no true log-likelihood is used and the corresponding estimators are pseudo-maximum likelihood estimators.

4.2. An MCEM-algorithm

The proposed algorithm is adapted from the MCEM-algorithm presented by McCulloch [8] for generalized linear mixed models. Since expectations $E_{c\,k}^{[t]}(\xi_i^2)$ and $E_{c\,k}^{[t]}[\exp(-u_{ij}\xi_i)]$ and posterior probabilities $t_k^{[t]}(y_i)$ cannot be derived in closed form, our goal is to form Monte Carlo approximations of these quantities. To this aim, we incorporate a Metropolis–Hastings step into the EM-algorithm which does not require specification of the marginal distribution of Y_i. This leads us to draw values from the unknown conditional distribution of ξ_i given Y_i, $Z_{ik}=1$ and the current value $\theta_k^{[t]}$. One can then calculate Monte Carlo approximations of the two required expectations. In the same way, we draw values from the known distribution of ξ_i given $Z_{ik}=1$ and the current value $\theta_k^{[t]}$ in order to approximate marginal distribution $f_k(y_i|\theta_k^{[t]})$ by Monte Carlo methods and to calculate posterior probability $t_k^{[t]}(y_i)$. Before presenting the proposed algorithm in Section 4.2.2, we recall in Section 4.2.1 the Metropolis–Hastings step applied to our specific case.

4.2.1. The Metropolis–Hastings step

The Metropolis–Hastings algorithm [10] is certainly one of the most famous MCMC methods [15]. The aim of the MCMC methods is to generate samples from a target distribution π unavailable in closed form. To this end, a candidate distribution h (called the instrumental or proposal distribution) must be specified from which potential new values are drawn. Among samples generated from h, Metropolis–Hastings selects representative samples of the target distribution π using an acceptance/rejection method.

To define the proposed Metropolis–Hastings step, we need to specify the candidate distribution h. We propose to take h equal to the marginal distribution in class $\mathcal{C}_k$ of ξ_i given the current value $\theta_k^{[t]}$ [8]. Let $\xi_i^{[m]}$ be the previous draw from the conditional distribution of ξ_i given Y_i, $Z_{ik}=1$ and the current value $\theta_k^{[t]}$. The probability of accepting the new value ξ_i^* generated using the candidate distribution h is given by

$$\rho(\xi_i^{[m]},\xi_i^*)=\min\left\{1,\frac{f(\xi_i^*|y_i,z_{ik}=1,\theta_k^{[t]})h(\xi_i^{[m]})}{f(\xi_i^{[m]}|y_i,z_{ik}=1,\theta_k^{[t]})h(\xi_i^*)}\right\}$$

where the second term simplifies to:

$$\begin{aligned}\frac{f(\xi_i^*|y_i,z_{ik}=1,\theta_k^{[t]})h(\xi_i^{[m]})}{f(\xi_i^{[m]}|y_i,z_{ik}=1,\theta_k^{[t]})h(\xi_i^*)}&=\frac{f(\xi_i^*|y_i,z_{ik}=1,\theta_k^{[t]})f(\xi_i^{[m]}|z_{ik}=1,\theta_k^{[t]})}{f(\xi_i^{[m]}|y_i,z_{ik}=1,\theta_k^{[t]})f(\xi_i^*|z_{ik}=1,\theta_k^{[t]})}\\&=\frac{f(y_i|\xi_i^*,z_{ik}=1,\theta_k^{[t]})}{f(y_i|\xi_i^{[m]},z_{ik}=1,\theta_k^{[t]})}.\end{aligned}$$

By choosing h equal to the random effect distribution, probability ρ is simplified since the obtained formula only involves the specification of the conditional distribution of Y_i given ξ_i and the component $\mathcal{C}_k$ from which unit i arises.

4.2.2. The proposed MCEM-algorithm

Incorporating this Metropolis–Hastings step into the EM-algorithm gives the following Monte Carlo EM (MCEM) algorithm at iteration $[t+1]$:

(1) For $i=1,\ldots,I$ and $k=1,\ldots,K$, draw:
- M values $\xi_i^{[1]},\ldots,\xi_i^{[M]}$ from the distribution of $\xi_i|Y_i,Z_{ik}=1$ given the current value $\theta_k^{[t]}$ using the Metropolis–Hastings algorithm described above and use them to form Monte Carlo approximations of the two required expectations in the function $Q(\theta,p|\theta^{[t]},p^{[t]})$:

$$E_{c\,k}^{[t]}(\xi_i^2)\simeq\frac{1}{M}\sum_{m=1}^{M}\xi_i^{[m]2}$$

$$E_{c\,k}^{[t]}\big[\exp(-u_{ij}\xi_i)\big]\simeq\frac{1}{M}\sum_{m=1}^{M}\exp(-u_{ij}\xi_i^{[m]}).$$

Table 1
Parameter estimation results obtained with EM and MCEM in the Gaussian case on 100 simulated data sets.

		EM		MCEM	
Simulated values		Mean	s.d.	Mean	s.d.
	$p_1 = 0.6$	0.6006	0.0171	0.6006	0.0170
$\mathcal{C}_1$	$\beta_1 = -2$	−1.9872	0.1141	−1.9873	0.1140
	$\sigma_1^2 = 0.2$	0.1994	0.1201	0.1991	0.1166
	$p_2 = 0.4$	0.3994	0.0171	0.3994	0.0170
$\mathcal{C}_2$	$\beta_2 = 2$	2.0289	0.1737	2.0292	0.1733
	$\sigma_2^2 = 0.8$	0.7657	0.2897	0.7605	0.2855
	$\tau^2 = 2$	2.0210	0.1340	2.0216	0.1339

- N values $\xi_i^{[1]}, \ldots, \xi_i^{[N]}$ from the known distribution of ξ_i given $Z_{ik} = 1$ and the current value $\theta_k^{[t]}$ in order to approximate the marginal distribution:

$$\begin{aligned} f_k(y_i|\theta_k^{[t]}) &= f(y_i|z_{ik} = 1, \theta_k^{[t]}) \\ &= \int \prod_{j=1}^{n_i} f(y_{ij}|\xi_i, z_{ik} = 1, \theta_k^{[t]}) f(\xi_i|z_{ik} = 1, \theta_k^{[t]})\, \mathrm{d}\xi_i \\ &\approx \frac{1}{N} \sum_{n=1}^{N} \left\{ \prod_{j=1}^{n_i} f(y_{ij}|\xi_i^{[n]}, z_{ik} = 1, \theta_k^{[t]}) \right\} \end{aligned}$$

  and to obtain an approximation of the posterior probability $t_k^{[t]}(y_i)$.

(2) Then maximise the function $Q(\theta, p|\theta^{[t]}, p^{[t]})$ defined previously to obtain new parameter values $\theta^{[t+1]}$ and $p^{[t+1]}$.

5. Simulation results

In this section, we study the behaviour of the two estimation methods developed in Section 4 via simulations. These simulations are performed in a two-component mixture model context. In order to study the behaviour of the MCEM-algorithm, we first consider in Section 5.1 its use in the Gaussian case. In Section 5.2, we come back to the exponential case and we compare the relative performances of the two proposed methods. Finally, in Section 5.3, we discuss about the choice of the appropriate number of mixture components. We propose and compare two criteria associated with the two developed parameter estimation methods.

5.1. Preliminary results

We focus here on the use of the MCEM-algorithm in the case of mixtures of LMM for which the performances of the MCEM-algorithm can easily be compared to those of the EM-algorithm. We give in Table 1 the mean and standard deviation of the 100 estimated values obtained with both EM and MCEM on 100 simulated data sets. We set the number of statistical units I equal to 100 and we consider the same number of repetitions for each unit: $\forall i = 1, \ldots, I \;\; n_i = J = 6$. The mixing parameters are $p_1 = 0.6$ and $p_2 = 0.4$. The random effect variances are $\sigma_1^2 = 0.2$ and $\sigma_2^2 = 0.8$ and the residual variance is $\tau^2 = 2$. We consider here a unique fixed effect parameter by component: $\beta_1 = -2$ and $\beta_2 = 2$. For simplicity, we consider in these simulations $\forall i = 1, \ldots, I \;\; X_i = \mathbb{1}_J$ where $\mathbb{1}_J$ is the all-1 vector of length J.

Table 1 clearly shows that the MCEM-algorithm performs very close to the EM-algorithm. However, it is important to note that the MCEM-algorithm is numerically intensive. For the 100 simulation runs, the EM-algorithm implemented using R requires here only a few minutes whereas the MCEM-algorithm implemented in C takes a few hours. Note that this computation time will naturally become even more important in the exponential case.

The EM-algorithm is started from some initial values of the parameters. Recently, Seidel, Mosler and Alker [16] have demonstrated how different starting strategies can lead to different estimates in the context of fitting mixtures via the EM-algorithm. For high-dimensional data, initial values might be obtained through the use of some clustering algorithm. In this paper, the EM-algorithm has been initiated using the k-means method [3]. One way to reduce the starting value problem consists of first performing some number of short runs of EM from different k-means results. Then, the initial values are obtained from the short run solution providing the highest log-likelihood. This approach has been proved to be efficient in many cases [17].

Concerning each Metropolis–Hastings step in the MCEM-algorithm, one has to determine a simulation step for which the simulated chain effectively is in equilibrium, the so-called "burn-in" period. Then, the chain simulated before this period has to be discarded. In our simulations, various "burn-in" periods have been used. Finally, the results presented in this paper have been obtained using a "burn-in" period of 500 iterations. After discarding draws from the "burn-in" period, the chain was run for $M = 4000$ iterations at each Metropolis–Hastings step. Note that another method consists of increasing M as the EM-algorithm progresses. This approach has been used in particular by McCulloch [8]. Finally, concerning the Monte

Table 2
Parameter estimation results from 100 data sets simulated from model A.

		Model A			
		$J = 4$		$J = 8$	
Simulated values		Linear.	MCEM	Linear.	MCEM
	$p_1 = 0.6$	0.6017	0.5999	0.6002	0.6001
		(0.0039)	(0.0023)	(0.0012)	(0.0010)
$\mathcal{C}_1$	$\beta_1 = -3$	−2.9990	−2.9945	−3.0154	−3.0057
		(0.0878)	(0.0815)	(0.0921)	(0.0817)
	$\sigma_1^2 = 0.2$	0.2166	0.1960	0.1986	0.1977
		(0.1227)	(0.0834)	(0.0749)	(0.0645)
	$p_2 = 0.4$	0.3983	0.4001	0.3998	0.3999
		(0.0039)	(0.0023)	(0.0012)	(0.0010)
$\mathcal{C}_2$	$\beta_2 = 3$	3.0181	3.0042	3.0105	3.0121
		(0.1765)	(0.1641)	(0.1555)	(0.1588)
	$\sigma_2^2 = 0.8$	0.7419	0.7953	0.7798	0.7893
		(0.2490)	(0.2588)	(0.2455)	(0.2280)

Table 3
Parameter estimation results from 100 data sets simulated from model B.

		Model B			
		$J = 4$		$J = 8$	
Simulated values		Linear.	MCEM	Linear.	MCEM
	$p_1 = 0.6$	0.6634	0.6536	0.6481	0.6330
		(0.1340)	(0.0851)	(0.0856)	(0.0797)
$\mathcal{C}_1$	$\beta_1 = -1$	−0.8908	−0.9334	−0.9485	−0.9601
		(0.1845)	(0.1494)	(0.1599)	(0.1371)
	$\sigma_1^2 = 0.2$	0.2760	0.2376	0.2523	0.2330
		(0.2251)	(0.1466)	(0.1324)	(0.1119)
	$p_2 = 0.4$	0.3366	0.3464	0.3519	0.3670
		(0.1340)	(0.0851)	(0.0856)	(0.0797)
$\mathcal{C}_2$	$\beta_2 = 1$	1.2909	1.2202	1.2036	1.1504
		(0.5317)	(0.3047)	(0.3426)	(0.2895)
	$\sigma_2^2 = 0.8$	0.5437	0.6173	0.6195	0.6795
		(0.4568)	(0.3591)	(0.3257)	(0.3274)

Carlo sample size used in order to approximate the marginal distribution in the first step of the MCEM-algorithm, we have considered in all simulations $N = 3000$.

5.2. Comparison of the two proposed methods

In this section, we come back to the exponential case. As previously, simulations are performed to assess the ability of the proposed methods to estimate mixture parameters. We set the number of statistical units I equal to 100. The mixing parameters are $p_1 = 0.6$ and $p_2 = 0.4$ and the random effect variances are $\sigma_1^2 = 0.2$ and $\sigma_2^2 = 0.8$. First, we consider a fixed intercept by component but we define two models with more or less separated components:

- model A with $\beta_1 = -3$ and $\beta_2 = 3$,
- model B with $\beta_1 = -1$ and $\beta_2 = 1$.

In order to study the impact of the number of repetitions on the estimation quality, we also consider two values of J: $J = 4$ and $J = 8$.

Table 2 provides the mean and standard deviation of the estimations obtained from 100 samples generated from model A. Table 3 gives the results obtained for model B.

To complete our simulation study, additional simulations including nontrivial covariables into the linear predictor are performed too. We consider a fixed intercept and one covariable generated from the uniform distribution on [0, 1]. Thus, we define a third model:

- model C with $\beta_1 = \begin{pmatrix} -1 \\ 1 \end{pmatrix}$ and $\beta_2 = \begin{pmatrix} 1 \\ 3 \end{pmatrix}$.

As previously, to study the impact of the number of repetitions, we consider two values of J: $J = 4$ and $J = 8$.

Table 4 displays the estimation results obtained from 100 samples generated from model C.

Table 2 shows that both methods provide accurate parameter estimations. As expected, the precision of the estimation depends on the random effect variances: the greater the variance the greater the estimation's standard deviation. The number of repetitions also naturally influences the quality of the estimations. Note that the results obtained with the MCEM-algorithm are slightly better than those obtained with the method based on linearisation. Nevertheless, as noted previously,

Table 4
Parameter estimation results from 100 data sets simulated from model C.

		Model C			
		$J = 4$		$J = 8$	
Simulated values		Linear.	MCEM	Linear.	MCEM
	$p_1 = 0.6$	0.6269	0.5990	0.6021	0.5999
		(0.0444)	(0.0359)	(0.0316)	(0.0200)
	$\beta_{11} = -1$	−0.9433	−1.0018	−0.9889	−0.9946
$\mathcal{C}_1$		(0.1870)	(0.1562)	(0.1468)	(0.1112)
	$\beta_{12} = 1$	1.0042	0.9884	1.0077	1.0071
		(0.3485)	(0.2717)	(0.2276)	(0.1611)
	$\sigma_1^2 = 0.2$	0.2572	0.1893	0.2072	0.1934
		(0.1895)	(0.1251)	(0.0826)	(0.0654)
	$p_2 = 0.4$	0.3731	0.4010	0.3979	0.4000
		(0.0444)	(0.0359)	(0.0316)	(0.0200)
	$\beta_{21} = 1$	1.0650	0.9698	1.0331	1.0088
$\mathcal{C}_2$		(0.3191)	(0.2602)	(0.2316)	(0.1957)
	$\beta_{22} = 3$	3.0732	3.0120	2.9799	2.9933
		(0.4360)	(0.3535)	(0.2991)	(0.2364)
	$\sigma_2^2 = 0.8$	0.6036	0.7900	0.7702	0.7621
		(0.4027)	(0.4036)	(0.3585)	(0.2581)

Table 5
Average correct classification rates (with s.d.) from 100 simulations.

		Model A		Model B		Model C	
		$J = 4$	$J = 8$	$J = 4$	$J = 8$	$J = 4$	$J = 8$
Linear.	$\mathcal{C}_1$	99.98	100.00	93.37	96.38	98.25	98.32
		(0.17)	(0.00)	(12.68)	(4.50)	(2.94)	(3.54)
	$\mathcal{C}_2$	99.55	99.95	67.87	76.95	88.73	95.33
		(1.03)	(0.35)	(17.86)	(13.82)	(7.05)	(4.24)
MCEM	$\mathcal{C}_1$	99.96	100.00	96.15	96.52	97.70	98.85
		(0.23)	(0.00)	(6.13)	(4.93)	(2.87)	(1.72)
	$\mathcal{C}_2$	99.92	99.97	73.17	79.95	93.38	96.68
		(0.43)	(0.25)	(13.32)	(12.68)	(5.07)	(3.55)

the MCEM-algorithm is numerically intensive. A strategy to reduce computation times could be to use the estimates obtained with the method based on linearisation as initial values.

The results shown in Table 3 are obtained when the mixture components are less separated. They are not as adequate as the previous case but still reasonable. Remarks similar to those made for the first case can be made. We just note that the impact of the number of repetitions is more important in this case.

Finally, the results displayed in Table 4 are obtained using the mixture model with nontrivial covariates. These results keep being adequate. In spite of the fact that the MCEM-algorithm is numerically very intensive in this case, once again the results obtained with this algorithm are better than those obtained with the method based on linearisation with better precision.

Table 5 provides the average correct classification rate for each model using the maximum a posteriori probability (MAP) rule [18] from the estimate parameter values $\hat{p}, \hat{\theta}$. The MAP rule consists of assigning all the measures of unit i to the mixture component $\mathcal{C}_k$ such as

$$k = \underset{1 \le l \le K}{\operatorname{argmax}} \widehat{t_l(y_i)}$$

with $\widehat{t_l(y_i)} = P(Z_{il} = 1 | y_i, \hat{p}, \hat{\theta})$. The obtained results are globally satisfactory. Table 5 shows that the average correct classification rate decreases when the random effect variance increases. Moreover, we note that the variability of correct classification rates increases with the random effect variance. It also clearly shows that the average correct classification rate increases with the number of repetitions. Finally, we also note that the rates obtained with the MCEM-algorithm are slightly better on average than those obtained with the method based on linearisation with better precision.

5.3. Choice of the number of components

In the previous sections, the number of components is assumed to be known. However, in practical situations, this is mostly not the case and it thus becomes part of the estimation process. We take here a model choice point of view on this question. To determine the appropriate number of components, model selection criteria proposed by Martinez [19] and

Table 6
Choice of K for the simulated two-component mixture model A with criteria AIC^G, $\widehat{AIC}$, BIC^G and $\widehat{BIC}$.

	Model A					
	$J = 4$			$J = 8$		
K	1	2	3	1	2	3
AIC^G	0	94	6	0	95	5
$\widehat{AIC}$	1	86	13	0	93	7
BIC^G	0	100	0	0	100	0
$\widehat{BIC}$	1	99	0	0	100	0

Lavergne et al. [20] for generalized linear mixed models are here adapted to mixtures of generalized linear mixed models and compared from numerical experiments.

The first considered criterion is associated with the parameter estimation method based on the Gumbel linearisation developed in Section 4.1. This criterion is obtained by computing the general information criterion for mixture of linear mixed models on the transformed data d_i:

$$\begin{aligned} IC^G &= -2\sum_{i=1}^{I} \log h(d_i|\hat{\theta}, \hat{p}) + pen\ \nu \\ &= -2\sum_{i=1}^{I} \log \left\{ \sum_{k=1}^{K} \hat{p}_k h_k(d_i|\hat{\theta}_k) \right\} + pen\ \nu \end{aligned}$$

where pen denotes the penalty term and ν the number of free parameters of the model. The estimates $\hat{p}_k$ and $\hat{\theta}_k$ are obtained from the procedure defined in Section 4.1. Note that in all this section the letter G refers to Gumbel in the notation.

The second criterion is derived from a direct approximation of the log-likelihood by Monte Carlo methods via the MCEM-algorithm developed in Section 4.2:

$$\widehat{IC} = -2\sum_{i=1}^{I} \log \left(\sum_{k=1}^{K} \hat{p}_k \left\{ \frac{1}{N} \sum_{n=1}^{N} \left[\prod_{j=1}^{n_i} f(y_{ij}|\xi_i^{[n]}, z_{ik} = 1, \hat{\theta}_k) \right] \right\} \right) + pen\ \nu$$

where, as previously, pen denotes the penalty term and ν the number of free parameters of the model. Here $\hat{p}_k$ and $\hat{\theta}_k$ denote the estimates of p_k and θ_k obtained from the MCEM-algorithm and the N values $\xi_i^{[1]}, \ldots, \xi_i^{[N]}$ are generated from the distribution of ξ_i given $Z_{ik} = 1$ and $\hat{\theta}_k$.

For choosing the number of mixture components, we use the BIC (Bayesian Information Criterion) criterion [21]. This criterion is perhaps the most popular criteria for choosing the number of components in a mixture model [22]. Thus, we study the relative performances of criteria BIC^G and $\widehat{BIC}$ derived from general criteria IC^G and $\widehat{IC}$ by defining the penalty term pen equal to $\log(n)$, n being the sample size. We also consider criteria AIC^G and $\widehat{AIC}$ derived from the classical AIC (Akaike Information Criterion) criterion [23] with $pen = 2$.

In this experiment, 100 data sets are generated from the two-component mixture model A defined in Section 5.2. We estimate models with $K = 1, 2$ or 3 mixture components and we calculate the associated criteria values. We report in Table 6 the number of times each mixture model is selected. From these results, it appears that all criteria select the generating model most of the time when the mixture components are well separated. Criteria AIC^G and $\widehat{AIC}$ have a slight tendency to select a more complex model while BIC^G and $\widehat{BIC}$ nearly always select the generating model. Note that this tendency of AIC to select an overfitted model is well known in the literature. From Table 6, it also appears that criteria derived from Gumbel linearisation perform very close to criteria derived from a direct approximation when the mixture components are well separated.

A second experiment was performed using a two-component mixture model with less separated components. In this case, 100 data sets are generated from model B defined in Section 5.2. The obtained results are displayed in Table 7. It appears that criteria AIC^G and $\widehat{AIC}$ clearly prefer the two-component mixture model while BIC^G and $\widehat{BIC}$ have a tendency to underestimate the number of components, even more when the number of repetitions J is low. This better behaviour of AIC^G and $\widehat{AIC}$ is only due to their tendency to underpenalise the model complexity. It is important to note that the results obtained with BIC^G criterion are much better than those obtained with $\widehat{BIC}$ when the components are less separated.

Finally, Table 8 provides the choice of K for the simulated two-component mixture model C including nontrivial covariates with criteria AIC^G, $\widehat{AIC}$, BIC^G and $\widehat{BIC}$. Once again, it appears that all criteria select the generating model most of the time. In this particular third experiment, criteria AIC^G and $\widehat{AIC}$ have a slight tendency to select the more complex model while BIC^G and $\widehat{BIC}$ nearly always select the generating model.

In this section, note that we focused on BIC-type criterion which has been shown to have a satisfactory behaviour at a practical level (see Fraley and Raftery [24]). But the other criteria have been proposed in various situations. For instance, the ICL criterion [25] proved its relevancy to determine the number of mixture components. Unlike AIC and BIC criteria defined

Table 7
Choice of K for the simulated two-component mixture model B with criteria AIC^G, $\widehat{AIC}$, BIC^G and $\widehat{BIC}$.

	Model B					
	$J = 4$			$J = 8$		
K	1	2	3	1	2	3
AIC^G	9	90	1	2	94	4
$\widehat{AIC}$	7	87	6	2	90	8
BIC^G	58	42	0	18	82	0
$\widehat{BIC}$	86	14	0	66	34	0

Table 8
Choice of K for the simulated two-component mixture model C with criteria AIC^G, $\widehat{AIC}$, BIC^G and $\widehat{BIC}$.

	Model C					
	$J = 4$			$J = 8$		
K	1	2	3	1	2	3
AIC^G	0	87	13	0	81	19
$\widehat{AIC}$	0	89	11	0	79	21
BIC^G	0	99	1	0	100	0
$\widehat{BIC}$	0	100	0	0	100	0

from the integrated observed likelihood, this criterion is based on the integrated completed likelihood. The derivation of this criterion from the Gumbel linearisation seems to be possible following the example of Celeux et al. [6]. However, its derivation from the MCEM-algorithm is more complicated and needs more investigation in our particular mixture context where missing data are of two types.

To complete this section, note that the primary interest of this study is more to compare the behaviour of the two parameter estimation methods than to consider the model selection problem. Indeed, the criteria developed are associated with the two parameter estimation methods and thus we evaluate the impact of these methods on the choice quality of the number of components. In order to study in more detail the model selection problem, more extensive simulations have to be carried out with different true numbers of clusters and with for instance up to 5 estimated clusters for a two-component mixture. A detailed study of the model selection problem in mixtures of GLMM will be developed in a further work.

6. Application to children with autism data set

In order to illustrate the developed methods, we fit mixture models to autism data. More precisely, these data consist of imitative skills measurements collected on 77 children with autism aged from 5 to 7 years. These imitative abilities are measured using an imitation scale which assesses the ability of the children to imitate and recognize being imitated. This procedure generates the variable Imitation corresponding to scores from 0 to 28. For each child, 3 measurements were collected with a time interval of 6 months. Children considered in this study have a development age greater than or equal to 18 months for both fine and global motricity. For each child at every observation, psychological development age in months for the domain of communication and autonomy are measured using an adaptive behaviour scale. These measures generate two variables AgeEquCom and AgeEquAuton with values between 3 and 65. More information concerning materials and methods can be found in Pry et al. [26].

This is a two-level data set with 3 observations per child. The asymmetric distribution of the data with a strong proportion of low scores and a few large values suggests that these data may be modelled using an asymmetric distribution as the exponential distribution. Moreover, because of the repeated measurements collected for each child, a subject-specific random effect is introduced. Thus six mixtures of exponential mixed models ($K = 1, \ldots, 6$) are fitted to these data and compared. The different models include the two development age variables in the linear predictor giving

$$\eta_{ij}^k = \beta_{0k} + \beta_{1k}\,\mathrm{AgeEquCom}_{ij} + \beta_{2k}\,\mathrm{AgeEquAuton}_{ij} + \xi_i$$

for the mixture component $\mathcal{C}_k$.

The six fitted models have been compared with penalized log-likelihood criteria AIC and BIC. Table 9 displays log-likelihood, AIC and BIC values obtained for the 6 different models with the method based on linearisation. Clearly, the log-likelihood decreases from model M1 to M4 and then is stable for models M4, M5 and M6. The two criteria naturally support model M4. Since the log-likelihood is quite similar for models M4, M5 and M6, the criteria choose the most parsimonious model.

Table 10 displays the parameter estimations obtained for the different mixture models with the method based on linearisation. The results obtained with the MCEM-algorithm are closely similar but we do not report them for simplicity. These results illustrate the mechanism of the clustering procedure. At the beginning of the procedure, the introduction of new mixture components increases the log-likelihood and leads to new groupings of children with different parameter

Table 9
Log-likelihood, AIC and BIC values for the 6 models with $K = 1, \ldots, 6$.

	M1	M2	M3	M4	M5	M6
Log-lik.	−1069.924	−1010.386	−844.7942	−778.0898	−780.1397	−780.1397
AIC	2149.847	2038.772	1717.588	1594.180	1608.279	1618.279
BIC	2167.059	2069.754	1765.782	1659.586	1690.898	1718.110

Table 10
Parameter estimations obtained with the method based on linearisation for the 6 compared models for the children with autism data set.

		$\mathcal{C}_1$	$\mathcal{C}_2$	$\mathcal{C}_3$	$\mathcal{C}_4$	$\mathcal{C}_5$	$\mathcal{C}_6$
M2	p_k	0.2891	0.7109				
	β_{0k}	−1.8571	−0.7827				
	β_{1k}	0.4984	0.0021				
	β_{2k}	−0.5404	0.0882				
	σ_k^2	22.1732	3.2801				
M3	p_k	0.1060	0.2487	0.6453			
	β_{0k}	−53.7330	−0.3135	1.7363			
	β_{1k}	0.6096	0.3295	0.0090			
	β_{2k}	1.2010	−0.5181	0.0196			
	σ_k^2	64.8616	17.8186	0.0028			
M4	p_k	0.1053	0.1160	0.6618	0.1169		
	β_{0k}	−53.5381	−10.2783	2.1939	25.0901		
	β_{1k}	0.6117	−0.0206	0.0099	0.7648		
	β_{2k}	1.1982	0.0064	0.0013	−1.8783		
	σ_k^2	64.8672	1.1344	0.1321	63.3029		
M5	p_k	0.0924	0.1032	0.2215	0.4268	0.1562	
	β_{0k}	−51.7215	−10.9353	1.7348	1.7348	13.9261	
	β_{1k}	0.4387	0.0000	0.0090	0.0090	1.1834	
	β_{2k}	1.2799	0.0001	0.0196	0.0196	−1.5386	
	σ_k^2	59.3706	0.0000	0.0000	0.0000	80.6958	
M6	p_k	0.0924	0.1032	0.0144	0.4201	0.2137	0.1562
	β_{0k}	−51.7215	−10.9353	1.7348	1.7348	1.7348	13.9261
	β_{1k}	0.4387	0.0000	0.0090	0.0090	0.0090	1.1834
	β_{2k}	1.2799	0.0001	0.0196	0.0196	0.0196	−1.5386
	σ_k^2	59.3706	0.0000	0.0000	0.0000	0.0000	80.6958

estimations. Thus, for $K = 1, \ldots, 4$, introducing a new component appears to be reasonable and reveals some existing underlying structures that characterize the data. When the clustering procedure gets some stability, for each new component introduced in the model, the procedure only tends to subdivide one of the previous components and reestimate the parameters. Thus, from $K = 4$, it is clear that increasing the number of components does not make sense since the clustering procedure subdivides one component into two components with the same parameter estimations.

As in the simulations, note that the algorithm is initiated for each fitted model using the *k*-means method. Thus, we can note that the obtained results present a numerical robust stability from model M4. For instance, the parameter estimations obtained for models M5 and M6 are exactly the same although we did not use the estimations based on model M5 to initialise the algorithm for model M6. Clearly, the clustering procedure has subdivided one component of model M5 to get two new components with the same parameter estimations in model M6. The robust numerical behaviour of the clustering procedure is clearly displayed by the results and is certainly related to these particular data which provide a stable situation.

To conclude this section, the obtained clustering composed of 4 clusters can easily be explained by psychologists. As is to be expected, the biggest cluster $\mathcal{C}_3$ is composed of a large number of children with classical autism. The cluster $\mathcal{C}_2$ contains the children without any progress. According to the experts, this cluster matches to severe autism with genetic origins. A third cluster $\mathcal{C}_1$ is composed of children who are globally in progression with very small starting scores. Finally, this clustering also generates a last cluster $\mathcal{C}_4$ which can be interpreted as a group with strong heterogeneity, a kind of regression or at least a strong instability.

7. Discussion and conclusions

In this paper, we define a new class of models for repeated data: mixtures of generalized linear mixed models. These models allow us to introduce a notion of heterogeneity in the GLMM. They take dependency and variability of repeated data into account through random effects defined within each mixture component. We proposed two parameter estimation methods: the MCEM-algorithm which can be used for mixtures of any generalized linear mixed models and the method

based on a linearisation specific to a mixture of exponential mixed models. These two methods are adaptations of the EM-algorithm getting round problems related to the direct use of it.

Simulations performed in the exponential case show that the two proposed parameter estimation methods globally perform well. They also show that the MCEM-algorithm gives slightly better results than the method based on linearisation. This behaviour difference is even greater in difficult situations with less separated mixture components or low number of repetitions. However, it is important to recall that, like all MCMC approaches, the MCEM-algorithm is numerically intensive since a large number of simulations is required at each iteration. In practice, a compiled programming language had to be used to reduce computation times. Moreover, even if this algorithm seems to perform well in practice, we still have not established theoretical results for convergence. On the contrary, the implementation of the method based on linearisation is fast and can easily be done with R for instance. Nevertheless, the use of this method is restricted to the exponential case. In this case, the estimates obtained with the method based on linearisation can be used as initial values for the MCEM-algorithm in order to reduce computation times.

In this paper, the focus is on the ML fitting of mixtures of GLMM via the EM-algorithm and the MCEM-algorithm. Although no theoretical properties have been established in this paper for this particular context, properties of MLEs for mixture models have been studied in various contexts. In particular, available results about consistency in a general context are to be reviewed in McLachlan and Peel [3]. Another issue concerns the identifiability of the parameters which is a necessary condition for the existence of consistent estimators. Although we obtained good results through our numerous simulations, theoretical results should be established in a further work in order to reinforce these simulation results. This question turns out to be difficult as outlined by Hennig [27]. In his paper, Hennig deals with the identifiability of the parameters of models for data generated by different linear regression distributions with Gaussian errors and considers in particular finite mixture models with random and fixed covariates. The counterexamples and the sufficient conditions for identifiability given in this paper could be an interesting starting point to investigate the identifiability question in our particular context.

Concerning the model selection problem, we developed two general criteria derived from the two proposed parameter estimation methods: IC^G and $\widehat{IC}$. The first simulations performed in the exponential case seem to show that BIC^G is generally preferable to $\widehat{BIC}$ in particular in difficult situations with less separated components. In practice, a possible strategy would be first to select the appropriate number of components using BIC^G which is better and faster than $\widehat{BIC}$. The parameters of the selected model could then be estimated using the MCEM-algorithm which seems to give slightly better results. However, a more detailed study of this model selection problem has to be performed in order to valid these first results.

Coming back to the numerically intensive problem of the MCEM-algorithm, it would be interesting to propose an intermediate version using simulation via a stochastic approximation in order to avoid calculations. A future work could adapt the method developed by Kuhn and Lavielle [28]. In their paper, Kuhn and Lavielle proposed an algorithm combining the stochastic approximation version of EM (SAEM) [29,30] with a Markov chain Monte Carlo procedure.

References

[1] D.M. Titterington, A.F.M. Smith, U.E. Makov, Statistical Analysis of Finite Mixture Distributions, John Wiley & Sons Ltd, Chichester, 1985.
[2] G.J. McLachlan, K.E. Basford, Mixture Models: Inference and Application to Clustering, Marcel Dekker, New York, 1988.
[3] G.J. McLachlan, D. Peel, Finite Mixture Models, Wiley-Interscience, New York, 2000.
[4] V. Hasselblad, Estimation of finite mixtures of distributions from the exponential family, Journal of the American Statistical Association 64 (1969) 1459–1471.
[5] M. Wedel, W.S. DeSarbo, A mixture likelihood approach for generalized linear models, Journal of Classification 12 (1995) 21–55.
[6] G. Celeux, O. Martin, C. Lavergne, Mixture of linear mixed models for clustering gene expression profiles from repeated microarray experiments, Statistical Modelling 5 (2005) 243–267.
[7] G.J. McLachlan, T. Krishnan, The EM Algorithm and Extensions, Wiley, New York, 1997.
[8] C.E. McCulloch, Maximum likelihood algorithms for generalized linear mixed models, Journal of the American Statistical Association 92 (1997) 162–170.
[9] C.E. McCulloch, S.R. Searle, Generalized, Linear and Mixed Models, in: Wiley Series in Probability and Statistics, 2001.
[10] W.K. Hastings, Monte Carlo sampling methods using Markov chains and their applications, Biometrika 57 (1970) 97–109.
[11] A.P. Dempster, N.M. Laird, D.B. Rubin, Maximum likelihood for incomplete data via the EM algorithm, Journal of the Royal Statistical Society B 39 (1977) 1–38.
[12] O.C. Gaudoin, C. Lavergne, J.L. Soler, A generalized geometric de-eutrophication software reliability model, IEEE Transactions on Reliability 43 (1994) 536–541.
[13] C. Gouriéroux, A. Monfort, Statistique et modèles économétriques, Economica (1989).
[14] C. Gouriéroux, A. Monfort, Pseudo-Likelihood Methods, in: Handbook of Statistics, vol. 11, Elsevier Science, North Holland, 1993.
[15] C.P. Robert, G. Casella, Monte Carlo Statistical Methods, 2nd ed., Springer-Verlag, New York, 2004.
[16] W. Seidel, K. Mosler, M. Alker, A cautionary note on likelihood ratio tests in mixture models, Annals of the Institute of Statistical Mathematics 52 (2000) 481–487.
[17] C. Biernacki, G. Celeux, G. Govaert, Choosing starting values for the EM algorithm for getting the highest likelihood in multivariate gaussian mixture models, Computational Statistics and Data Analysis 41 (2003) 561–575.
[18] M.H. DeGroot, Optimal Statistical Decisions, McGraw-Hill, New York, 1970.
[19] M.J. Martinez, Modèles linéaires généralisés à effets aléatoires: contributions au choix de modèle et au modèle de mélange, Ph.D. Thesis, Université Montpellier II, 2006.
[20] C. Lavergne, M.J. Martinez, C. Trottier, Empirical model selection in generalized linear mixed effects models, Computational Statistics 23 (2008) 99–109.
[21] G. Schwarz, Estimating the dimension of a model, Annals of Statistics 6 (1978) 461–464.
[22] R.E. Kass, A.E. Raftery, Bayes factors, Journal of the American Statistical Association 90 (1995) 773–795.
[23] H. Akaike, A new look at the statistical identification model, IEEE Transactions on Automatic Control 19 (1974) 716–723.

[24] C. Fraley, A.E. Raftery, Model-based clustering, discriminant analysis, and density estimation, Journal of the American Statistical Association 97 (2002) 611–631.
[25] C. Biernacki, G. Celeux, G. Govaert, Assessing a mixture model for clustering with the integrated completed likelihood, IEEE Transactions on Pattern Analysis and Machine Intelligence 22 (2001) 719–725.
[26] R. Pry, A.F. Petersen, A. Baghdadli, The relationship between expressive language level and psychological development in children with autism 5 years of age, Autism 9 (2005) 179–189.
[27] C. Hennig, Identifiability of models for clusterwise linear regression, Journal of Classification 17 (2000) 273–296.
[28] E. Kuhn, M. Lavielle, Coupling a stochastic approximation version of EM with an MCMC procedure, ESAIM: Probability and Statistics 8 (2004) 115–131.
[29] G. Celeux, J. Diebolt, A stochastic approximation type EM algorithm for the mixture problem, Stochastics and Stochastics Reports 41 (1992) 119–134.
[30] B. Delyon, M. Lavielle, E. Moulines, Convergence of a stochastic approximation version of the EM algorithm, Annals of Statistics 27 (1999) 94–128.