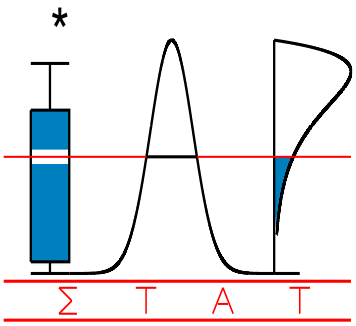
T E C H N I C A L

R E P O R T

0227

AN APPROXIMATE APPROACH TO FIT A LINEAR MIXED MODEL WITH A FINITE NORMAL MIXTURE AS RANDOM-EFFECTS DISTRIBUTION AND ITS SAS IMPLEMENTATION

A. KOMÁREK, G. VERBEKE and G. MOLENBERGHS



I A P S T A T I S T I C S

N E T W O R K

INTERUNIVERSITY ATTRACTION POLE

http://www.stat.ucl.ac.be/IAP

An Approximate Approach to Fit a Linear Mixed Model with a Finite Normal Mixture as Random-Effects Distribution and its SAS Implementation

Arnošt KOMÁREK, Geert VERBEKE and Geert MOLENBERGHS [1]

ABSTRACT. This paper describes an approximate method to compute maximum likelihood estimates of the parameters in the linear mixed model with a finite normal mixture as random-effects distribution. The proposed method uses an EM algorithm with an approximate M step which can be performed using procedures designed to fit a common linear mixed model. This approach enables, among others, to include easily various covariance structures of the residuals and random effects in the model. The suggested method has been implemented as a SAS macro which is briefly introduced and illustrated on data on heights of schoolgirls.

KEY WORDS. Maximum likelihood estimate; Repeated measurement.

1. INTRODUCTION

A linear mixed model is a frequently used tool for describing longitudinal continuous data. Its random effects are usually assumed to be normally distributed. Unfortunately, this basic assumption can very often be violated. This will occur, for example, if an important categorical covariate is omitted from the fixed part of the model. That is why Verbeke and Lesaffre (1996) and Verbeke and Molenberghs (2000, chap. 12) proposed to

[1]Arnošt Komárek is Research Assistant, and Geert Verbeke is Associate Professor, Biostatistical Centre, Catholic University of Leuven, U.Z. St. Rafaël, Kapucijnenvoer 35, B-3000 Leuven, Belgium (E-mail: arnost.komarek@med.kuleuven.ac.be). Geert Molenberghs is Professor, Center for Statistics, University of Limburg, Gebouw D, B-3590 Diepenbeek, Belgium.

assume that random effects are distributed according to a finite normal mixture. The advantage of this approach is, among others, the fact that many continuous distributions can be well approximated by a finite normal mixture illustrating that the proposed model is generally applicable. On the other hand, a big disadvantage is a lack of available computational tools to fit such models in practice.

In this article, we concentrate on describing an approximate method to compute maximum likelihood estimates of the linear mixed model with a finite normal mixture as random effects distribution which can work up to an arbitrary level of accuracy. Moreover, it can be quite easily implemented using common software for the linear mixed models. A SAS macro called `HetMixed` based on the procedure `PROC MIXED` has been developed and will be introduced and illustrated in the paper.

After defining the model in Section 2, we show in Section 3 how the estimates can be computed using the EM algorithm and how the most difficult part of it, the M step, can be performed approximately using tools designed for the classical linear mixed model. This SAS macro will be briefly introduced in Section 4 and illustrated in Section 5. Data introduced by Goldstein (1979) and analyzed in Verbeke and Lesaffre (1996) will be reanalyzed.

2. MODEL FORMULATION

This section introduces the concept of the linear mixed model with a finite normal mixture as random-effects distribution that was proposed by Verbeke and Lesaffre (1996) and also described by Verbeke and Molenberghs (2000, chap. 12). In accordance with the therminology used by these authors, the linear mixed model with a finite normal mixture as random effects distribution will be referred to the heterogeneity linear mixed model. It can be seen as an extension of the classical linear mixed model which will be called the homogeneity linear mixed model.

Let the random variable Y_{ik} denote the (possibly transformed) response of interest, for the ith individual measured at time t_{ik}, $i = 1, \ldots, N$, $k = 1, \ldots, n_i$, and let $\boldsymbol{Y}_i$ be the n_i-dimensional vector of all repeated measurements for the ith subject, that is, $\boldsymbol{Y}_i = (Y_{i1}, \ldots, Y_{in_i})^T$. The heterogeneity linear mixed model starts from similar relationship as the homogeneity model, that is from

$$\boldsymbol{Y}_i = \begin{pmatrix} \mathbb{X}_i & \mathbb{Z}_i \end{pmatrix} \begin{pmatrix} \boldsymbol{\beta}^F \\ \boldsymbol{\beta}^R \end{pmatrix} + \mathbb{Z}_i \boldsymbol{b}_i + \boldsymbol{\varepsilon}_i, \tag{1}$$

where $\mathbb{X}_i$ and $\mathbb{Z}_i$ are $(n_i \times p)$, respectively $(n_i \times q)$ matrices of known covariates, modeling how the response evolves over time for the ith subject. Further, $\boldsymbol{\beta}^F$ and $\boldsymbol{\beta}^R$ are p-dimensional, respectively q-dimensional vectors of unknown regression parameters. Variables $\boldsymbol{b}_i$ are subject-specific q-dimensional random effects, and $\boldsymbol{\varepsilon}_i$ is an n_i-dimensional vector of residual components ε_{ik}, $k = 1, \ldots, n_i$. All $\boldsymbol{\varepsilon}_i$ are assumed to be independent and normally distributed with mean vector zero and covariance matrix Σ_i.

We have just described the part of the heterogeneity model that is the same as for the homogeneity model. The former one differs from the latter in assumptions on subject-specific effects $\boldsymbol{b}_i$. They are assumed to be independent under both models. The homogeneity model considers them to be normally distributed with mean vector zero and covariance matrix $\mathbb{D}$. The heterogeneity model is obtained by replacing this distributional assumption by a mixture of a prespecified number g of q-dimensional normal distributions with mean vectors $\boldsymbol{\mu}_j$ and covariance matrices $\mathbb{D}$, i.e.,

$$\boldsymbol{b}_i \sim \sum_{j=1}^{g} \pi_j N(\boldsymbol{\mu}_j, \mathbb{D}), \tag{2}$$

with $\sum_{j=1}^{g} \pi_j = 1$. A more general case would assume different covariance matrices $\mathbb{D}_1, \ldots, \mathbb{D}_g$ for each component of the mixture. However, this leads to the infinite likelihood as pointed out by McLachlan and Basford (1988). In order to avoid numerical problems in the estimating procedure, we will assume $\mathbb{D}_1 = \cdots = \mathbb{D}_g = \mathbb{D}$.

Vectors $\boldsymbol{W}_i = (W_{i1}, \ldots, W_{ig})^T$ can now be defined as follows. The term $W_{ij} = 1$ if $\boldsymbol{b}_i$ is sampled from the jth component of the mixture and 0 otherwise, $j = 1, \ldots, g$. The distribution of $\boldsymbol{W}_i$ is then described by

$$P(W_{ij} = 1) = E(W_{ij}) = \pi_j,$$

which is called *the prior probability* to be sampled from component j. Expected values of $\boldsymbol{b}_i$ can then easily be obtained as

$$E(\boldsymbol{b}_i) = E\big(E[\boldsymbol{b}_i|\boldsymbol{W}_i]\big) = E\left(\sum_{j=1}^{g} \boldsymbol{\mu}_j W_{ij}\right) = \sum_{j=1}^{g} \pi_j \boldsymbol{\mu}_j.$$

The expectation of the response is then

$$E(\boldsymbol{Y}_i) = E(\mathbb{X}_i\boldsymbol{\beta}^F + \mathbb{Z}_i\boldsymbol{\beta}^R + \mathbb{Z}_i\boldsymbol{b}_i + \boldsymbol{\varepsilon}_i) = \mathbb{X}_i\boldsymbol{\beta}^F + \mathbb{Z}_i\boldsymbol{\beta}^R + \mathbb{Z}_i \sum_{j=1}^{g} \pi_j \boldsymbol{\mu}_j.$$

Note that the model is overparametrized and therefore the additional constraint

$$\sum_{j=1}^{g} \pi_j \boldsymbol{\mu}_j = \boldsymbol{0} \tag{3}$$

is needed. Then the marginal mean of the response equals $E(\boldsymbol{Y}_i) = \mathbb{X}_i\boldsymbol{\beta}^F + \mathbb{Z}_i\boldsymbol{\beta}^R$ which is the same as in the case of the homogeneity model.

Model (1) with assumptions (2) can also be rewritten as a hierarchical Bayes model

$$\begin{aligned} &\boldsymbol{Y}_i|\boldsymbol{b}_i \sim N(\mathbb{X}_i\boldsymbol{\beta}^F + \mathbb{Z}_i\boldsymbol{\beta}^R + \mathbb{Z}_i\boldsymbol{b}_i, \Sigma_i), \\ &\boldsymbol{b}_i|\boldsymbol{\mu} \sim N(\boldsymbol{\mu}, \mathbb{D}), \\ &\boldsymbol{\mu} \in \{\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_g\}, \qquad \text{with } P(\boldsymbol{\mu} = \boldsymbol{\mu}_j) = \pi_j. \end{aligned} \tag{4}$$

This expression might be useful when the heterogeneity model is used for classification of individual profiles into one of g populations. The underlying data generating mechanism can be viewed as a two step process. First, the population is chosen and second, response is generated according to the chosen population. In practice, one can wish to reveal the

first step of this mechanism and try to classify an individual with observed response vector $\boldsymbol{Y}$ into one of the populations.

3. ESTIMATION OF THE HETEROGENEITY MODEL

3.1 General Concept

Estimation of the unknown parameters in the heterogeneity model is based on the marginal distribution of the observations $\boldsymbol{Y}_i$. Under (1) and (2), this distribution is

$$\boldsymbol{Y}_i \sim \sum_{j=1}^{g} \pi_j N(\mathbb{X}_i\boldsymbol{\beta}^F + \mathbb{Z}_i\boldsymbol{\beta}^R + \mathbb{Z}_i\boldsymbol{\mu}_j, \mathbb{V}_i), \qquad \text{with } \mathbb{V}_i = \mathbb{Z}_i\mathbb{D}\mathbb{Z}_i^T + \Sigma_i.$$

Let $\boldsymbol{\pi}$ be the vector of component probabilities (i.e., $\boldsymbol{\pi}^T = (\pi_1, \ldots, \pi_g)$) and let $\boldsymbol{\gamma}$ be the vector of all other unknown parameters (i.e., $\boldsymbol{\beta}^F$, $\boldsymbol{\beta}^R$, the components of the matrices $\mathbb{D}$ and Σ_i). Further, let $\boldsymbol{\theta}^T = (\boldsymbol{\pi}^T, \boldsymbol{\gamma}^T)$ denote the vector of all unknown parameters that are to be estimated. The method of maximum likelihood can be used to find the requested estimates. The likelihood function corresponding to the marginal distribution of the observations $\boldsymbol{Y}_i$ is of the form

$$L^*(\boldsymbol{\theta}|\boldsymbol{y}) = \prod_{i=1}^{N} \left\{ \sum_{j=1}^{g} \pi_j f_{ij}(\boldsymbol{y}_i|\boldsymbol{\gamma}) \right\}, \tag{5}$$

where $\boldsymbol{y}^T = (\boldsymbol{y}_1^T, \ldots, \boldsymbol{y}_N^T)$ is the vector containing all observed response values and f_{ij} is the density of an n_i-dimensional normal distribution $N(\mathbb{X}_i\boldsymbol{\beta}^F + \mathbb{Z}_i\boldsymbol{\beta}^R + \mathbb{Z}_i\boldsymbol{\mu}_j, \mathbb{V}_i)$.

Note that the likelihood function (5) is invariant under the $g!$ possible permutations of the mean vectors and corresponding probabilities of the components of the mixture. However, this lack of identifiability can easily be overcome by imposing some constraint on the parameters. For example, the constraint

$$\pi_1 \geq \pi_2 \geq \cdots \geq \pi_g \tag{6}$$

suggested by Aitkin and Rubin (1985) can be used. The likelihood is then maximized without the restriction, and the component labels are permuted afterward to satisfy (6).

The log-likelihood function corresponding to the likelihood (5) is given by

$$l^*(\boldsymbol{\theta}|\boldsymbol{y}) = \sum_{i=1}^{N} \log \left\{ \sum_{j=1}^{g} \pi_j f_{ij}(\boldsymbol{y}_i|\boldsymbol{\gamma}) \right\}. \tag{7}$$

It is quite difficult to maximize this function and the EM algorithm introduced by Dempster, Laird and Rubin (1977) can be used to compute the desired estimates. The response vectors $\boldsymbol{Y}_i$ along with the (unobserved) population indicators $\boldsymbol{W}_i$ can be seen as complete data whereas the vectors $\boldsymbol{Y}_i$ alone can be viewed as incomplete data since information containing population membership is missing. The likelihood function (5) corresponds then to the incomplete data. The likelihood function that would have been obtained if values $\boldsymbol{w}_i = (w_{i1}, \ldots, w_{ig})^T$ of population indicators $\boldsymbol{W}_i$ had been observed equals

$$L(\boldsymbol{\theta}|\boldsymbol{y}, \boldsymbol{w}) = \prod_{i=1}^{N} \prod_{j=1}^{g} \{\pi_j f_{ij}(\boldsymbol{y}_i|\boldsymbol{\gamma})\}^{w_{ij}}, \tag{8}$$

where $\boldsymbol{w}^T = (\boldsymbol{w}_1^T, \ldots, \boldsymbol{w}_N^T)$ is the vector containing all hypothetically observed population indicators. The log-likelihood function corresponding to (8) then has the more tractable form

$$l(\boldsymbol{\theta}|\boldsymbol{y}, \boldsymbol{w}) = \sum_{i=1}^{N} \sum_{j=1}^{g} w_{ij} \{\log \pi_j + \log f_{ij}(\boldsymbol{y}_i|\boldsymbol{\gamma})\}.$$

Maximizing $l(\boldsymbol{\theta}|\boldsymbol{y}, \boldsymbol{w})$ with respect to $\boldsymbol{\theta}$ yields estimates that depend on the unobserved ("missing") indicators $\boldsymbol{w}$. The EM algorithm offers a solution to this problem by maximizing the expected value of $l(\boldsymbol{\theta}|\boldsymbol{y}, \boldsymbol{w})$, rather than $l(\boldsymbol{\theta}|\boldsymbol{y}, \boldsymbol{w})$ with respect to $\boldsymbol{\theta}$, where the expectation is taken over all unobserved w_{ij}. The conditional expectation of $l(\boldsymbol{\theta}|\boldsymbol{y}, \boldsymbol{w})$, given the observed data vector $\boldsymbol{y}$, is calculated within the E step (expectation step) of each iteration of the EM algorithm. The obtained expected log-likelihood function is then maximized within the M step (maximization step) of the algorithm.

Let $\boldsymbol{\theta}^{(t)}$ be the current estimate for $\boldsymbol{\theta}$, and let $\boldsymbol{\theta}^{(t+1)}$ stand for the updated estimate, obtained from one further iteration of the EM algorithm. The following E and M steps have to be executed to compute the updated estimate.

The E step. The conditional expectation of $l(\boldsymbol{\theta}|\boldsymbol{y},\boldsymbol{w})$, given the observed data vector $\boldsymbol{y}$ is given by

$$
\begin{aligned}
Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) &= E\big[l(\boldsymbol{\theta}|\boldsymbol{y},\boldsymbol{w})\big|\boldsymbol{y},\boldsymbol{\theta}^{(t)}\big] \\
&= \sum_{i=1}^{N}\sum_{j=1}^{g} p_{ij}(\boldsymbol{\theta}^{(t)})\big\{\log\pi_j + \log f_{ij}(\boldsymbol{y}_i|\boldsymbol{\gamma})\big\}.
\end{aligned} \tag{9}
$$

The terms $p_{ij}(\boldsymbol{\theta}^{(t)})$ are called *the posterior probabilities* for the ith individual to belong to the jth component of the mixture and can easily be computed using Bayes' theorem as

$$
\begin{aligned}
p_{ij}(\boldsymbol{\theta}^{(t)}) &= E\big[W_{ij}\big|\boldsymbol{y}_i,\boldsymbol{\theta}^{(t)}\big] = P\big(W_{ij}=1\big|\boldsymbol{y}_i,\boldsymbol{\theta}^{(t)}\big) = \\
&= \frac{\pi_j^{(t)} f_{ij}(\boldsymbol{y}_i|\boldsymbol{\gamma}^{(t)})}{\sum_{k=1}^{g}\pi_k^{(t)} f_{ik}(\boldsymbol{y}_i|\boldsymbol{\gamma}^{(t)})}.
\end{aligned} \tag{10}
$$

The M step. The objective function $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$ has to be maximized with respect to $\boldsymbol{\theta}$ to get the updated estimate $\boldsymbol{\theta}^{(t+1)}$. Expression (9) is the sum of two terms:

$$
Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) = Q_1(\boldsymbol{\pi}|\boldsymbol{\theta}^{(t)}) + Q_2(\boldsymbol{\gamma}|\boldsymbol{\theta}^{(t)}),
$$

where

$$
Q_1(\boldsymbol{\pi}|\boldsymbol{\theta}^{(t)}) = \sum_{i=1}^{N}\sum_{j=1}^{g} p_{ij}(\boldsymbol{\theta}^{(t)})\log\pi_j, \tag{11}
$$

$$
Q_2(\boldsymbol{\gamma}|\boldsymbol{\theta}^{(t)}) = \sum_{i=1}^{N}\sum_{j=1}^{g} p_{ij}(\boldsymbol{\theta}^{(t)})\log f_{ij}(\boldsymbol{y}_i|\boldsymbol{\gamma}). \tag{12}
$$

The first term depends only on the parameter $\boldsymbol{\pi}$, the second one only on the parameter $\boldsymbol{\gamma}$. Hence, it is possible to maximize each of these terms separately to find a maximum for Q. Q_1 is maximized for

$$
\pi_j^{(t+1)} = \frac{1}{N}\sum_{i=1}^{N} p_{ij}(\boldsymbol{\theta}^{(t)}).
$$

These estimates are equal to an average of posterior probabilities for all subjects belonging to a given population.

Unfortunately, the term (12) cannot be maximized analytically as can the first one. It will immediately be shown how an approximate optimization of Q_2 can be obtained using the common software for fitting the homogeneity linear mixed models, such as the SAS procedure `PROC MIXED` or the R/Splus function `lme`. Function Q_2 is to be maximized with respect to $\boldsymbol{\gamma}$. If posterior probabilities $p_{ij}(\boldsymbol{\theta}^{(t)})$ are integers, function (12) would be a log-likelihood for the homogeneity model based on observations from $\sum_{i=1}^{N}\sum_{j=1}^{g} p_{ij}(\boldsymbol{\theta}^{(t)})$ individuals. Note that maximization of (12) with respect to $\boldsymbol{\gamma}$ is equivalent to maximization of

$$A \cdot Q_2(\boldsymbol{\gamma}|\boldsymbol{\theta}^{(t)}) = \sum_{i=1}^{N}\sum_{j=1}^{g} A \cdot p_{ij}(\boldsymbol{\theta}^{(t)}) \log f_{ij}(\boldsymbol{y}_i|\boldsymbol{\gamma})$$

for an arbitrary positive constant A. Further, numbers $A \cdot p_{ij}(\boldsymbol{\theta}^{(t)})$ can be arbitrarily close to integers by choosing A sufficiently large. In practice, their rounded values can be used to approximate the function $A \cdot Q_2(\boldsymbol{\gamma}|\boldsymbol{\theta}^{(t)})$. Let $a_{ij}(\boldsymbol{\theta}^{(t)})$ denote integers such that

$$a_{ij}(\boldsymbol{\theta}^{(t)}) \doteq A \cdot p_{ij}(\boldsymbol{\theta}^{(t)}) \tag{13}$$

and let

$$Q_2^A(\boldsymbol{\gamma}|\boldsymbol{\theta}^{(t)}) = \sum_{i=1}^{N}\sum_{j=1}^{g} a_{ij}(\boldsymbol{\theta}^{(t)}) \log f_{ij}(\boldsymbol{y}_i|\boldsymbol{\gamma}). \tag{14}$$

The function $Q_2^A(\boldsymbol{\gamma}|\boldsymbol{\theta}^{(t)})$ can be interpreted as the log-likelihood function for an appropriate homogeneity linear mixed model which corresponds to the observations taken on $\sum_{i=1}^{N}\sum_{j=1}^{g} a_{ij}(\boldsymbol{\theta}^{(t)})$ mutually independent individuals. Note that the ith response vector $\boldsymbol{Y}_i$ from the original data set appears $\sum_{j=1}^{g} a_{ij}(\boldsymbol{\theta}^{(t)})$ times in a data set which corresponds to the desirable homogeneity model. At the same time, the marginal distribution of $a_{ij}(\boldsymbol{\theta}^{(t)})$ response vectors $\boldsymbol{Y}_i$ out of their $\sum_{j=1}^{g} a_{ij}(\boldsymbol{\theta}^{(t)})$ replications follows the n_i-dimensional normal distribution $N(\mathbb{X}_i\boldsymbol{\beta}^F + \mathbb{Z}_i\boldsymbol{\beta}^R + \mathbb{Z}_i\boldsymbol{\mu}_j, \mathbb{V}_i)$, with $\mathbb{V}_i = \mathbb{Z}_i\mathbb{D}\mathbb{Z}_i^T + \Sigma_i$. At this moment, common software for homogeneity linear mixed models is able to compute

updated approximate estimates of $\boldsymbol{\gamma}$. The higher the value of A is used, the better the approximation is obtained. One has to take into account only present computational possibilities.

When implementing this method, one also has to consider the constraint (3) of the form $\sum_{j=1}^{g} \pi_j \boldsymbol{\mu}_j = \mathbf{0}$ that was exposed to the population means at the beginning of this paper. Fortunately, it is not too difficult to ensure that this constraint is satisfied since the originally restricted q-dimensional parameters $\boldsymbol{\beta}^R, \boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_g$ can be replaced by unrestricted q-dimensional parameters $\boldsymbol{\delta}_1, \ldots, \boldsymbol{\delta}_g$ using the relationship

$$\boldsymbol{\delta}_j = \boldsymbol{\beta}^R + \boldsymbol{\mu}_j, \qquad j = 1, \ldots, g.$$

In fact, parameters $\boldsymbol{\delta}_j$ express real population means, whereas parameters $\boldsymbol{\mu}_j$ represent the contrasts between a population mean and the overall mean $\boldsymbol{\beta}^R$. Restriction (3) also gives the way to compute $\boldsymbol{\beta}^R$ from $\boldsymbol{\delta}_1, \ldots, \boldsymbol{\delta}_g$, that is

$$\boldsymbol{\beta}^R = \sum_{j=1}^{g} \pi_j \boldsymbol{\delta}_j.$$

3.2 Empirical Bayes Inference

The random effects $\boldsymbol{b}_i$ in model (1) are assumed to be random variables and that is why they cannot be estimated in a standard way. So called *Empirical Bayes* (EB) estimates $\hat{\boldsymbol{b}}_i$ can be used for random-effects inference. It will be shown immediately that they can quite easily be obtained using the common software for fitting the homogeneity linear mixed models. Let us denote the estimate of $\boldsymbol{\theta}$ parameters obtained using the EM algorithm described in the previous section as $\hat{\boldsymbol{\theta}}$. The EB estimate $\hat{\boldsymbol{b}}_i$ of the random effects is then given by

$$\hat{\boldsymbol{b}}_i = \hat{\boldsymbol{b}}_i(\hat{\boldsymbol{\theta}}) = E[\boldsymbol{b}_i | \boldsymbol{Y}_i = \boldsymbol{y}_i, \boldsymbol{\theta} = \hat{\boldsymbol{\theta}}].$$

The expected value is based on a posterior distribution derived from the model (4) using Bayesian techniques. See Gelman *et al.* (1995).

Let us denote $\hat{\boldsymbol{\delta}}_j = \hat{\boldsymbol{\beta}}^R + \hat{\boldsymbol{\mu}}_j$, $j = 1, \ldots, g$ and $\hat{\boldsymbol{b}}_i^j = \hat{\mathbb{D}}\mathbb{Z}_i^T\hat{\mathbb{V}}_i^{-1}(\boldsymbol{y}_i - \mathbb{X}_i\hat{\boldsymbol{\beta}}^F - \mathbb{Z}_i\hat{\boldsymbol{\delta}}_j)$, $i = 1, \ldots, N$, while obtaining all 'hat' expressions by replacing the corresponding quantities by their estimate. It can be shown that the EB estimates of the random effects for the heterogeneity linear mixed model equal

$$\hat{\boldsymbol{b}}_i = \sum_{j=1}^{g} p_{ij}(\hat{\boldsymbol{\theta}})\hat{\boldsymbol{b}}_i^j + \sum_{j=1}^{g} p_{ij}(\hat{\boldsymbol{\theta}})\hat{\boldsymbol{\mu}}_j. \tag{15}$$

It easily follows that the quantities $\hat{\boldsymbol{b}}_i^j$ are standard EB estimates of random effects for $a_{ij}(\hat{\boldsymbol{\theta}})$ individuals with common response vector $\boldsymbol{Y}_i$ from the homogeneity linear mixed model that was used in the last iteration of the EM algorithm when maximizing Q_2^A function (14). This property can be advantageously used when computing EB estimates for the heterogeneity linear mixed model.

The EB estimates $\hat{\boldsymbol{b}}_i$ of the random effects are often used for diagnostic purposes, such as the detection of outliers, etc. More information concerning the use of the EB estimates can be found in Verbeke and Molenberghs (2000, chap. 7).

3.3 Classification

The heterogeneity model can perfectly serve for classification purposes of longitudinal profiles or clustering. However, such classification no longer must be based on the random-effects estimates. One should rather base such procedures on the posterior probabilities $p_{ij}(\hat{\boldsymbol{\theta}})$ evaluated in the estimate $\hat{\boldsymbol{\theta}}$ of the vector $\boldsymbol{\theta}$. Very common practice in mixture models is to classify the ith individual into the kth component for which $\max_{j=1,\ldots,g} p_{ij}(\hat{\boldsymbol{\theta}}) = p_{ik}(\hat{\boldsymbol{\theta}})$.

As already mentioned, the just described methodology for the computation of the maximum likelihood estimates of the heterogeneity linear mixed model can be quite easily implemented using existing procedures and functions for fitting homogeneity linear mixed models such as the SAS procedure `MIXED` or the R/Splus function `lme`. The main advantage of this approach is the fact that all covariance structures for matrices Σ_i and $\mathbb{D}$ that are offered by these functions and procedures can be used.

A SAS macro called `HetMixed`, for fitting heterogeneity models can be downloaded along with its manual from the URL of the Biostatistical Centre, K.U. Leuven:

`http://www.kuleuven.ac.be/biostat/research/software.htm`

The syntax of the macro is:

```
%HetMixed(DATA = , SUBJECT = , REPEATED = ,
   RESPONSE = , FIXED = , RANDOM = ,
   TYPEREP = simple, TYPERAND = un,
   G = 1, AMIN = 10, AMAX = &AMIN, ABY = 10,
   DECISWIT = 1, DECISBET = &DECISWIT, STOPWIT = 0.00001, STOPBET = 0.0001,
   MAXITER = 100, MIXEDMI = 50,
   INITPOST = , ENDPOST = , EB = , PIEPS = 0.1);
```

The macro was developed using the SAS version 8. A lot of effort was spent in making its syntax as similar as possible to the SAS procedure `MIXED`. For example, the `RANDOM, FIXED, SUBJECT, REPEATED` statements retain their meaning. All covariance structures for the matrices Σ_i and $\mathbb{D}$ that are available within the `MIXED` procedure can be specified within the `TYPEREP` and `TYPERAND` statements of the macro. The meaning of some other statements will be explained below, for the meaning of the remaining statements, we refer to the macro manual.

Let us now discuss some key features of the macro. To start EM, suitable initial estimates for $\boldsymbol{\theta}$ have to be found. Note that the knowledge of posterior probabilities $p_{ij}(\boldsymbol{\theta}^{(0)})$ defined by (10) is sufficient to compute the value of the objective function $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(0)})$ given by (9) that is to be maximized in the M step of the EM algorithm. This nice property is utilized by the macro and initial posterior probabilities $p_{ij}(\boldsymbol{\theta}^{(0)})$ for the ith individual to belong to the jth component of the mixture are used to start the EM algorithm. Such initial posterior probabilities can be given either by the user if the data set which contains them is specified by the `INITPOST` statement or they can be randomly generated.

As discussed before, the E step of the EM algorithm is quite straightforward, as well as updating the estimates of component probabilities $\boldsymbol{\pi}$. On the other hand, the computation of updated estimates of $\boldsymbol{\gamma}$ parameters is much more complicated and multiplication technique described earlier is used. At each iteration of the EM algorithm, an extended data set corresponding to the log-likelihood (14) of the homogeneity linear mixed model is created and subsequently updated estimates of the $\boldsymbol{\gamma}$ parameters are computed using `PROC MIXED`. Multiplication factors $a_{ij}(\boldsymbol{\theta}^{(t)})$, defined by (13), are computed using the user specified A. Note that the response vector $\boldsymbol{Y}_i$, $i = 1, \ldots, N$ has to be repeated $\sum_{j=1}^{g} a_{ij}(\boldsymbol{\theta}^{(t)})$ times to create the data set that is used to compute updated estimates of $\boldsymbol{\theta}$. This data set must be physically created since there is no direct introduction of weights in `PROC MIXED`. Thus a careful trade-off is needed between adequacy of the approximation (large A) and computation time (small A). Indeed, since the replicates specified by the weights have to be physically created, and computation time is related to the length of the dataset, one should be careful with values that are too large. It is quite desirable to start computation with lower A value and after the approximation provided by this value is too rough, to increase it. This idea is, among others, also implemented in the macro.

More precisely, the EM algorithm starts its first iteration with $A = A_1$ =`AMIN` and computes, using this value, numbers $a_{ij}(\boldsymbol{\theta}^{(0)})$. Similarly, the same $A = A_1$ is used in the

following iterations of the EM. Let m_1 be the number of iterations of the EM algorithm needed to achieve the convergence which is driven by the `DECISWIT` and `STOPWIT` options whose names come from the term *within* convergence. In order to find out whether used A was sufficiently large for function Q_2^A to approximate Q_2, the model will be refitted using increased value of A. Specifically, in the second run, the EM algorithm based on $A = A_2 = A_1 + \texttt{ABY}$, computes numbers $a_{ij}(\boldsymbol{\theta}^{(m_1)})$ using this A value and subsequently new parameter estimate $\boldsymbol{\theta}^{(m_1+1)}$. If this estimate is 'considerably different' from its previous value $\boldsymbol{\theta}^{(m_1)}$ the EM algorithm does not stop and continues using $A = A_2$ when computing the weights. The meaning of the term 'considerably different' is driven by the `DECISBET` and `STOPBET` options whose names are derived from the *between* convergence. The same principle is used after the convergence of the EM when using $A = A_2$ is achieved. So that, the whole iteration process consists of $\sum_{l=1}^{K} m_l$ iterations of the EM algorithm while each segment of m_l iterations uses the same A value equal to A_l and $A_l = A_{l-1} + \texttt{ABY}$. The *overall* convergence is reported when the *between* convergence criterion is satisfied for the first time. If the increased A value crosses `AMAX` when evaluating the *between* convergence the computation stops and no convergence is reported.

Both types of the convergence yield different stopping rules. The user can choose one of the three offered stopping rules specified by the `DECISWIT` and `DECISBET` options, respectively. The EM algorithm (either one of its sets with a specific value of A or the all iteration process) stops if

- $\left|Q\left(\boldsymbol{\theta}^{(t+1)}\middle|\boldsymbol{\theta}^{(t)}\right) - Q\left(\boldsymbol{\theta}^{(t)}\middle|\boldsymbol{\theta}^{(t-1)}\right)\right| < \varepsilon$ for two consecutive iterations;
- the *average* absolute difference between estimates of all parameters that are to be estimated in the two consecutive iterations is smaller than prespecified ε;
- the *maximal* absolute difference between estimates of all parameters that are to be estimated in the two consecutive iterations is smaller than prespecified ε.

Physically, two different values of ε are given by the `STOPWIT` and `STOPBET` options respectively. It should be emphasized that the approximation given by the A value used is taken into account only when maximizing the Q function. Its value used for the evaluation of the convergence is always computed exactly.

Recall that each increase of the A value provides better approximation to the objective function Q. Hence, after performing the M step of the EM algorithm when evaluating the *overall* convergence, the new estimates are closer to the real maximizer of this function than these obtained using the smaller value of A. Thus, one can expect that after using the higher value of A and computing one iteration of the EM algorithm, the objective function Q increases more than during several last iterations of the EM algorithm with the smaller value of A. That is why, one has to find a compromise which is satisfactory enough when computing an additional iteration of the EM algorithm with increased A constant and evaluating the *overall* convergence. Be aware of the fact that the estimates provided by the macro are *always* based on an approximation.

The macro provides, besides the estimates of model parameters, several additional quantities. The posterior probabilities (10) for the ith subject to belong to the jth component of the mixture can be saved in a data set prespecified by the `ENDPOST` statement. Similarly, empirical Bayes estimates (15) of random effects can be stored in a data set given by the `EB` option. The likelhood (5) and the log-likelihood (7) evaluated in the vector of estimates obtained are reported as well to enable the user for example the comparison of nested models.

5. EXAMPLE: HEIGHTS OF SCHOOLGIRLS

Growth curves of 20 girls with height measured on a yearly basis from age 6 to 10 were analyzed by Goldstein (1979, table 4.3, p. 101). The girls were classified according to the height of their mother (group A: $<$ 155 cm, group B: 155–164 cm, group C: $>$

164 cm). A significant (at 5%) group, as well as a significant group by age effect, were found. Because the group structure was obtained by discretizing the height of the mother at arbitrary points, which is very artificial, it may be useful to search for growth curve clusters, neglecting this a priori group structure.

A plot of the data is given in Figure 1. The heterogeneity linear mixed models for these growth curves have already been reported by Verbeke and Lesaffre (1996). They also have used an EM algorithm to compute the desired estimates thereby maximizing an exact Q_2 function (12) rather than the approximate Q_2^A function (14) within each M step of the algorithm. They have fitted a two and a three-component heterogeneity models describing linearly the evolution of the height and including random both intercept and slope. They have written a procedure using the GAUSS software to fit their models. This procedure was available for us. We will follow the same models to be able to compare exact ML estimates obtained by the GAUSS procedure and our approximate ML estimates. Precisely, the following model for the growth curves is assumed.

$$Y_{ij} = \beta_0^R + \beta_1^R t_{ij} + b_{i0} + b_{i1}\, t_{ij} + \varepsilon_{ij},$$

with $\Sigma_i = \sigma^2 I_5$ and unstructured matrix $\mathbb{D}$, $i = 1, \ldots, 20,\ j = 1, \ldots, 5$. At the same time, Y_{ij} denotes the height of the ith girl at the age t_{ij} and $t_{ij} = 6, \ldots, 10$.

The homogeneity model (i.e. the model with $g = 1$) can be fitted in SAS using the following syntax of `PROC MIXED`.

```
PROC MIXED DATA = girls METHOD = ml;
    CLASS id ageclss;
    MODEL height = age;
    RANDOM intercept age / TYPE = un subject = id;
    REPEATED ageclss / TYPE = simple subject = id;
RUN;
```

The estimates of a two-component heterogeneity model can be obtained using the macro `HetMixed` in the following way. Note that the variable concerning merely ones and corresponding to the intercept has to be created separately using a `DATA` step.

```
DATA girls; set girls;
    int = 1; RUN;

%HetMixed(DATA = girls,
    SUBJECT = id, REPEATED = ageclss, RESPONSE = height,
    FIXED = , RANDOM = int age,
    TYPEREP = simple, TYPERAND = un,
    G = 2, AMIN = 40, AMAX = 200, ABY = 40,
    DECISWIT = 1, DECISBET = 1, STOPWIT = 0.000001, STOPBET = 0.0001,
    MAXITER = 1000, ENDPOST = twopost);
```

An absent `INITPOST` statement indicates that the initial posterior probabilities are to be generated randomly.

To show the improvement of the results when the A value is being increased, Tables 1 and 2 report the estimates obtained after the approximate EM algorithm which was allowed to use different maximal A values, namely 10, 80, 160 and 200. Such results can be obtained by the macro by use of different values of the `AMAX` statement and sufficiently strict *between* convergence criterion. The convergence criterion in the last set of the iterations where the maximal A is used was chosen to be $\left|Q\left(\boldsymbol{\theta}^{(t+1)}\middle|\boldsymbol{\theta}^{(t)}\right) - Q\left(\boldsymbol{\theta}^{(t)}\middle|\boldsymbol{\theta}^{(t-1)}\right)\right| < 10^{-6}$ in two consecutive iterations. Both tables report also the estimates obtained by use of the GAUSS procedure of Verbeke and Lesaffre (1996).

As one can see, the $A = 10$ gives quite poor estimates. This poverty is particularly highlighted when comparing final log-likelihoods. However, the $A = 80$ results in estimates that are already quite close to what should be received. The estimated mean component

profiles based on the estimates obtained after the computation with maximal A equal to 200 are shown on Figure 2.

The fact that the user is able to indicate a good convergence even without the knowledge of the estimates computed using the exact EM algorithm which are generally unknown is illustrated in Table 3. The increase of the objective function Q defined by (9) when comparing the last iteration of the approximate EM algorithm with smaller A factor and the first iteration of the algorithm with this factor increased by 40 is reported in this table. In fact, this corresponds to the evaluation of the *between* or *overall* convergence of the algorithm as described in the previous section.

Reported log-likelihoods of a two and a three-component models might entice to perform a likelihood ratio test concerning the number g of mixture components. But this is not as a strightforward as it seems due to boundary problems as discussed by Ghosh and Sen (1985). Note that a classical likelihood ratio statistic does not follow a χ^2 distribution in this case.

6. DISCUSSION

Modelling repeated measures by the homogeneity linear mixed model is not always satisfactory since the assumed normal distribution of random effects might be violated. The homogeneity linear mixed model is also not useful for classification purposes. The so-called heterogeneity linear mixed model that allows us both to classify individual profiles and to create models with many other underlying distributions for random effects than just only the Gaussian one was therefore introduced. The distribution of random effects is assumed to be a mixture of normals which can well approximate many other continuous distributions. Note that the normality assumption for the random effects is violated, whenever an important categorical covariate has been omitted as a fixed effect in a linear

mixed model. Random effects then follow a mixture of g, possibly normal distributions, where g is the number of categories of the missing covariate.

Unfortunately, wider use of the heterogeneity linear mixed models was inhibited by insufficient software support. That is why we have proposed an approximate method as to compute maximum likelihood estimates of unknown parameters using available procedures and functions for the homogeneity linear mixed model. The proposed method has been implemented as a macro in the standard, commercially available software package SAS. The core procedure `MIXED` that comes with the software is used to perform the most involved part of the estimation procedure. We think that a flexible and stable procedure has resulted. Its main enticement comes from the fact that all covariance structures for both residual covariance matrices Σ_i and random effects covariance matrix $\mathbb{D}$ offered by the SAS `PROC MIXED` are equally available. The example has shown that when using sufficiently high value of A, a constant that drives a level of the approximation, estimates very close to the true ML estimates can be obtained.

REFERENCES

Aitkin, M., and Rubin, D. B. (1985), "Estimation and hypothesis testing in finite mixture models," *Journal of the Royal Statistical Society, Series B,* 47, 67-75.

Dempster, A. P., Laird, N. M., and Rubin, D.B. (1977), "Maximum likelihood from incomplete data via the EM algorithm (with discussion)," *Journal of the Royal Statistical Society, Series B,* 39, 1-38.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (1995), *Bayesian Data Analysis.* London: Chapman & Hall.

Ghosh, J. K., and Sen, P. K. (1985), "On the asymptotic performance of the log likelihood ratio statistic for the mixture model and related results." In: *Proceedings of the Berekely*

Conference in Honor of Jerzy Neyman and Jack Kiefer, Vol. 2, L.M. Le Cam and R.A. Olshen (Eds.). Monterey: Wadsworth, Inc., pp. 789-806.

Goldstein, H. (1979), *The Design and Analysis of Longitudinal Studies.* London: Academic Press.

McLachlan, G. J., and Basford, K. E. (1988), *Mixture Models: Inference and Applications to Clustering.* New York: Marcel Dekker.

Verbeke, G., and Lesaffre, E. (1996), "A linear mixed-effects model with heterogeneity in the random-effects population," *Journal of the American Statistical Association,* 91, 217-221.

Verbeke, G., and Molenberghs, G. (2000), *Linear Mixed Models for Longitudinal Data.* New York: Springer-Verlag.

TABLE 1. Heights of Schoolgirls, a two-component heterogeneity model. The parameter estimates computed using SAS macro `HetMixed` with different maximal A values and using GAUSS.

	Maximal A value				
Quantity	A=10	A=80	A=160	A=200	GAUSS
$\hat{\boldsymbol{\delta}}_1$	$\begin{pmatrix} 82.318 \\ 5.667 \end{pmatrix}$	$\begin{pmatrix} 82.804 \\ 5.389 \end{pmatrix}$	$\begin{pmatrix} 82.803 \\ 5.386 \end{pmatrix}$	$\begin{pmatrix} 82.804 \\ 5.386 \end{pmatrix}$	$\begin{pmatrix} 82.775 \\ 5.386 \end{pmatrix}$
$\hat{\boldsymbol{\delta}}_2$	$\begin{pmatrix} 82.802 \\ 5.784 \end{pmatrix}$	$\begin{pmatrix} 81.946 \\ 6.407 \end{pmatrix}$	$\begin{pmatrix} 81.939 \\ 6.421 \end{pmatrix}$	$\begin{pmatrix} 81.931 \\ 6.424 \end{pmatrix}$	$\begin{pmatrix} 82.065 \\ 6.419 \end{pmatrix}$
$\hat{\pi}_1$	.583	.682	.683	.683	.685
$\hat{\pi}_2$	.417	.318	.317	.317	.318
$\hat{\sigma}^2$	.476	.476	.476	.476	.469
$\widehat{\mathbb{D}}$	$\begin{pmatrix} 6.580 & -.082 \\ -.082 & .269 \end{pmatrix}$	$\begin{pmatrix} 6.455 & .123 \\ .123 & .048 \end{pmatrix}$	$\begin{pmatrix} 6.463 & .127 \\ .127 & .040 \end{pmatrix}$	$\begin{pmatrix} 6.463 & .129 \\ .129 & .040 \end{pmatrix}$	$\begin{pmatrix} 6.732 & .104 \\ .104 & .034 \end{pmatrix}$
Log-likelihood	-169.48	-166.80	-166.71	-166.70	-166.27

TABLE 2. Heights of Schoolgirls, a three-component heterogeneity model. The parameter estimates computed using SAS macro `HetMixed` with different maximal A values and using GAUSS.

	Maximal A value				
Quantity	A=10	A=80	A=160	A=200	GAUSS
$\hat{\boldsymbol{\delta}}_1$	$\begin{pmatrix} 81.028 \\ 5.755 \end{pmatrix}$	$\begin{pmatrix} 79.342 \\ 5.615 \end{pmatrix}$	$\begin{pmatrix} 79.380 \\ 5.604 \end{pmatrix}$	$\begin{pmatrix} 79.368 \\ 5.604 \end{pmatrix}$	$\begin{pmatrix} 79.457 \\ 5.599 \end{pmatrix}$
$\hat{\boldsymbol{\delta}}_2$	$\begin{pmatrix} 83.190 \\ 5.458 \end{pmatrix}$	$\begin{pmatrix} 84.275 \\ 5.330 \end{pmatrix}$	$\begin{pmatrix} 84.372 \\ 5.319 \end{pmatrix}$	$\begin{pmatrix} 84.393 \\ 5.316 \end{pmatrix}$	$\begin{pmatrix} 84.214 \\ 5.322 \end{pmatrix}$
$\hat{\boldsymbol{\delta}}_3$	$\begin{pmatrix} 82.239 \\ 6.183 \end{pmatrix}$	$\begin{pmatrix} 81.726 \\ 6.438 \end{pmatrix}$	$\begin{pmatrix} 81.672 \\ 6.456 \end{pmatrix}$	$\begin{pmatrix} 81.659 \\ 6.459 \end{pmatrix}$	$\begin{pmatrix} 81.655 \\ 6.470 \end{pmatrix}$
$\hat{\pi}_1$	.162	.199	.208	.210	.204
$\hat{\pi}_2$	.560	.505	.495	.494	.497
$\hat{\pi}_3$	.278	.295	.297	.297	.299
$\hat{\sigma}^2$	.485	.478	.477	.477	.455
$\widehat{\mathbb{D}}$	$\begin{pmatrix} 6.000 & .072 \\ .072 & .181 \end{pmatrix}$	$\begin{pmatrix} 2.870 & .379 \\ .379 & .042 \end{pmatrix}$	$\begin{pmatrix} 2.670 & .408 \\ .408 & .030 \end{pmatrix}$	$\begin{pmatrix} 2.603 & .416 \\ .416 & .028 \end{pmatrix}$	$\begin{pmatrix} 3.640 & .324 \\ .324 & .029 \end{pmatrix}$
Log-likelihood	-168.99	-166.06	-165.82	-165.78	-165.82

TABLE 3. Heights of Schoolgirls. The increase of the objective function Q when comparing the last iteration of the approximate EM algorithm with smaller A value and the first iteration of the algorithm with higher A value.

	A value								
Model	80	$\longrightarrow$	120	$\longrightarrow$	160	$\longrightarrow$	200	$\longrightarrow$	240
A two-component		.203		.105		.007		.005	
A three-component		.570		.303		.026		.014	

FIGURE 1. Heights of Schoolgirls. Growth curves of 20 schoolgirls from age 6 to 10.

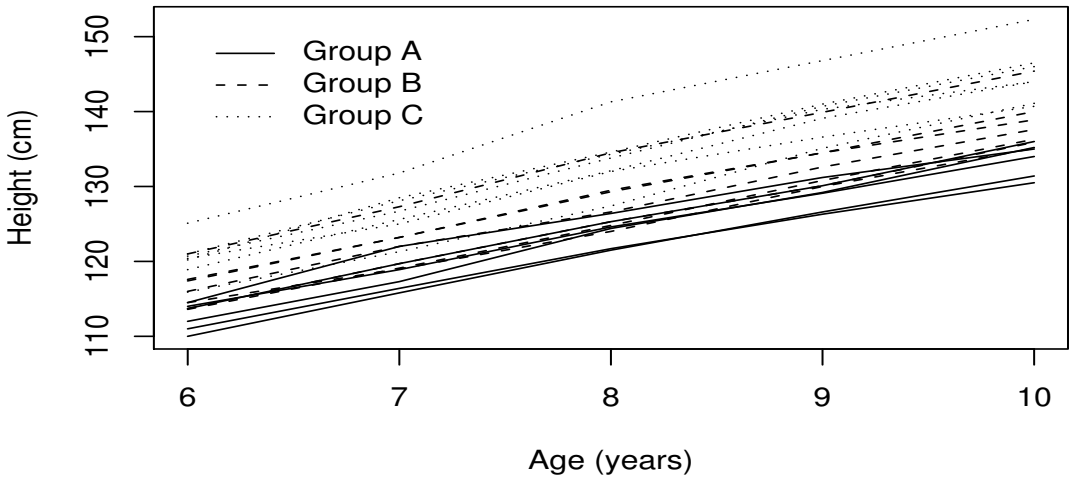


FIGURE 2. Heights of Schoolgirls. Estimated component means and probabilities, based on the heterogeneity models.