An improved SAEM algorithm for maximum likelihood estimation in mixtures of non linear mixed effects models

Marc Lavielle, Cyprien Mbogning

Statistics and Computing manuscript No.
(will be inserted by the editor)

An improved SAEM algorithm for maximum likelihood estimation in mixtures of non linear mixed effects models.

Marc Lavielle · Cyprien Mbogning

Received: date / Accepted: date

Abstract We propose a new methodology for maximum likelihood estimation in mixtures of non linear mixed effects models (NLMEM). Such mixtures of models include mixtures of distributions, mixtures of structural models and mixtures of residual error models. Since the individual parameters inside the NLMEM are not observed, we propose to combine the EM algorithm usually used for mixtures models when the mixture structure concerns an observed variable, with the Stochastic Approximation EM (SAEM) algorithm, which is known to be suitable for maximum likelihood estimation in NLMEM and also has nice theoretical properties. The main advantage of this hybrid procedure is to avoid a simulation step of unknown group labels required by a "full" version of SAEM. The resulting MSAEM (Mixture SAEM) algorithm is now implemented in the Monolix software. Several criteria for classification of subjects and estimation of individual parameters are also proposed. Numerical experiments on simulated data show that MSAEM performs well in a general framework of mixtures of NLMEM. Indeed, MSAEM provides an estimator close to the maximum likelihood estimator in very few iterations and is robust with regard to initialization. An application to pharmacokinetic (PK) data demonstrates the potential of the method for practical applications.

Keywords SAEM algorithm · Maximum likelihood estimation · Mixture models · Non linear mixed effects model · Monolix

M. Lavielle
Laboratoire de Mathématiques d'Orsay (LMO), Bat 425, 91405 Orsay cedex, France
& Inria Saclay, POPIX team
E-mail: marc.lavielle@inria.fr

C. Mbogning
Laboratoire de Mathématiques d'Orsay (LMO), Bat 425, 91405 Orsay cedex, France
& Inria Saclay, POPIX team
& LIMSS, Ecole Nationale Supérieure Polytechnique (ENSP), 8390 Yaoundé, Cameroun
E-mail: cyprien.mbogning@math.u-psud.fr

1 Introduction

Mixed effects models are frequently used for modeling longitudinal data when data is obtained from different individuals originating from a same population. Indeed, these models allow to describe both the within subject variability (the variability within each individual profile) and the between subject variability (the variability of the individual parameters). One complicating factor arises when the data is obtained from a population with some underlying heterogeneity. If we assume that the population consists of several homogeneous subpopulations, a straightforward extension of the mixed effects model is a finite mixture of mixed effects models.

Different types of mixtures of mixed effects models are considered in the literature, with different estimation algorithms appropriate to each situation. *Mixtures of distributions* assume that non-observed individual parameters come from different sub-populations. Such models are considered for instance in (Frühwirth-Schnatter 2006; De la Cruz, Quintana, and Marshall 2008). A linear mixed-effects model with heterogeneity in the random-effects is considered in (Verbeke and Lesaffre 1996) and a EM algorithm is proposed for maximizing the observed likelihood. The same model is used by (Proust and Jacqmin-Gadda 2005) but the MLE is obtained with the Marquardt algorithm. (Ketchum, Best, and Ramakrishnan 2012) propose to extend this model to a within-subject mixture model for analyzing

heart rate variability. The distribution of the residual errors is also a mixture in this model. (Ng, McLachlan, Wang, Ben-Tovim, and Ng 2006) consider normal mixtures in linear mixed effects models for clustering correlated gene-expression profiles. In these different mixtures of mixed effects models, extensions of the EM algorithm can be derived for computing the MLE since all these models are linear. Extensions to non linear models are much less frequent, mainly due to the fact that the maximization of the observed likelihood is complex. (Hou, Li, Zhang, Huang, and Wu 2008) propose a non linear mixed-effect mixture model for functional mapping of dynamic traits. They use a linearization approximation method by using the first-order Taylor expansion to approximate the non linear expectation function (Lindstrom and Bates 1990). (Wang, Schumitzky, and D'Argenio 2007; Wang, Schumitzky, and D'Argenio 2009) propose a Monte Carlo EM (MCEM) algorithm with importance sampling to deal with the intractable E step of the EM algorithm in non linear mixtures and avoid any model linearization.

In a non linear mixed effects model, the heterogeneity of the structural model cannot be adequately explained just by the inter-patient variability of certain parameters. It is therefore necessary to introduce a diversity of the structural models themselves (Lavielle, Mesa, Chatel, and Vermeulen 2010). *Between-subject model mixtures* assume that there exist sub-populations of individuals. Here, various structural models describe the response of the different sub-populations, and each subject belongs to one sub-population. One can imagine for example different structural models for responders, non responders and partial responders to a given treatment. *Within-subject model mixtures* assume that there exist sub-populations (of cells, viruses, etc.) within each patient. Again, differing structural models describe the response of the different sub-populations, but the proportions of each sub-population depend on the patient.

Our goal is to propose new methods for maximum likelihood estimation (MLE) of population parameters in a very general context of mixtures of NLMEM, including mixture of distributions and mixture of structural models. In the classical NLMEM framework, (Kuhn and Lavielle 2005) proposed the SAEM (Stochastic Approximation EM) algorithm which incorporates a simulation step of the unobserved individual parameters and a stochastic approximation of several statistics between the E and M steps. SAEM is recognized as a very powerful tool for NLMEM, known to accurately estimate population parameters and also to have good theoretical properties (Delyon, Lavielle, and Moulines 1999; Kuhn and Lavielle 2004; Allassonnière, Kuhn, and Trouvé 2010). On the other hand, the EM algorithm is widely used for "standard" mixtures models, *i.e.* when the mixture structure concerns some observed variable. We refer the reader to (Roeder and Wasserman 1997; McLachland and Peel 2000; Frühwirth-Schnatter 2006) and references therein for more details about mixture models. We then propose to combine the EM algorithm for mixture models, with the SAEM algorithm for NLMEM. The use of the resulting Mixed SAEM (MSAEM) instead of the SAEM itself avoids a simulation step of the unobserved group labels and significantly improves the results in term of stability and accuracy.

Section 2 of this paper describes the non linear mixed effects model for continuous data and different mixtures of models, including mixtures of distributions, mixtures of residual error models and mixtures of structural models. Section 3 is devoted to a description of the proposed MSAEM algorithm for maximum likelihood estimation in a mixture of NLMEM. Several numerical examples in Section 4 illustrate the performance of MSAEM.

2 Mixtures in non linear mixed-effects models

2.1 Non linear mixed-effects model

Mixed-effects models can address a wide class of data including continuous, count, categorical and time-to-event data. We will focus here on continuous data models. Modelling such data leads to using NLMEM as hierarchical models. At a first level, each individual has its own parametric regression model, known as the structural model, each identically defined up to a set of unknown individual parameters. At a second level, each set of individual parameters is assumed to be randomly drawn from some unknown population distribution. The model can then be defined as follow:

$$y_{ij} = f\left(x_{ij};\varphi_i\right) + g\left(x_{ij};\varphi_i,\xi\right)\varepsilon_{ij}, \tag{1}$$

where

- $y_{ij} \in \mathbb{R}$ denotes the j-th observation for the i-th individual, $1 \leq i \leq N$ and $1 \leq j \leq n_i$. $y_i = (y_{ij})$ is the vector of observations for the i-th individual.
- N is the number of individuals and n_i the number of observations for the i-th individual.
- x_{ij} denotes a vector of regression variables (for longitudinal data, x will generally be time).
- φ_i is the d-vector of individual parameters of individual i. We assume that all the φ_i are drawn from the same population distribution. We limit ourselves

to Gaussian models of the form:

$$\varphi_i = h(\mu, c_i) + \Sigma^{-1/2}\eta_i, \tag{2}$$

where h is a function which describes the covariate model: μ a vector of fixed-effects and c_i a vector of known covariates, $\eta_i \sim_{i.i.d} \mathcal{N}(0, I_d)$ a vector of standardized random effects and Σ the inter-individual variance-covariance matrix.

- $\varepsilon_{ij} \sim \mathcal{N}(0,1)$ denote the residual errors and are independent of individual parameters φ_i.
- f is a function defining the structural model and g a function defining the (possibly heteroscedastic) residual error model.
- $\theta = (\mu, \Sigma, \xi)$ is the complete set of population parameters.

The model is therefore completely defined by the joint probability distribution of the observations $y = (y_i)$ and the individual parameters $\varphi = (\varphi_i)$ which admits this hierarchical decomposition:

$$p(y_i, \varphi_i; \theta) = p(y_i|\varphi_i; \theta)p(\varphi_i; \theta) \tag{3}$$

Many problems are related with the use of these models: estimation of the population parameters θ with their standard errors, calculation of the likelihood of the observations for model selection or hypothesis testing purpose, estimation of the individual parameters (φ_i),...

When the model is linear, *i.e.* when the observations (y_{ij}) are Normally distributed, then the likelihood of the observations can be computed in a closed form and the EM algorithm can be used for maximizing this likelihood. On the other hand, when the structural model is not a linear function of the random effects and/or when the residual error model is not a linear Gaussian model, then the model of the observations is not linear anymore and the likelihood cannot be maximized using EM.

An alternative method consists of taking a first order Taylor expansion of the model function around the conditional modes of the random effects (Lindstrom and Bates 1990). Others have proposed the use of Gaussian quadrature rules (Davidian and Giltinan 1993).

A complete methodology for NLMEM is implemented in the Monolix software, including the Stochastic Approximation of EM (SAEM) proposed in (Delyon, Lavielle, and Moulines 1999). This algorithm is becoming a reference method for maximum likelihood estimation in NLMEM. Indeed, it is now implemented in NONMEM (software widely used for PKPD applications), Matlab (nlmefitsa.m) and R (saemix package).

2.2 Mixtures of mixed effects models

The simplest way to model a finite mixture model is to introduce a label sequence $(z_i; 1 \leq z_i \leq N)$ that takes its values in $\{1, 2, \ldots, M\}$ and is such that $z_i = m$ if subject i belongs to sub-population m.

In some situations, the label sequence (z_i) is known and can then be used as a categorical covariate in the model. We will address in the following the more challenging situation where this sequence is unknown. We therefore consider that (z_i) is a sequence of independent random variables taking values in $\{1, 2, \ldots, M\}$. A simple model might assume that the (z_i) are identically distributed: for $m = 1, \ldots, M$,

$$\mathbb{P}(z_i = m) = \pi_m. \tag{4}$$

But more complex models deserve to be considered for practical applications, for instance, the introduction of covariates for defining each individual's probabilities.

In its most general form, a mixture of mixed effects models assumes that there exist M joint distributions $p_1, \ldots, p_M$ and M vector of parameters $\theta_1, \ldots \theta_M$ such that the joint distribution defined in (3) now decomposes into

$$p(y_i, \varphi_i; \theta) = \sum_{m=1}^{M} \mathbb{P}(z_i = m)\, p_m(y_i, \varphi_i, \theta_m) \tag{5}$$

The mixture can then concern the distribution of the individual parameters $p(\varphi_i; \theta)$ and/or the conditional distribution of the observations $p(y_i|\varphi_i; \theta)$. Let us see some examples of such mixtures models:

i) A latency structure can be introduced at the level of the individual parameters assuming a Gaussian mixture model. This mixture model assumes that there exist $\mu_1, \Sigma_1, \ldots, \mu_M, \Sigma_M$ such that

$$\varphi_i = \sum_{m=1}^{M} \mathbb{1}_{z_i=m} h(\mu_m, c_i) + \Sigma_m^{-1/2}\eta_i. \tag{6}$$

The shape of the mixture depends on the structure of the variance-covariance matrices Σ_m. In the standard case of Gaussian mixture models, $h(\mu_m, c_i) = \mu_m$. Then, $p(\varphi_i; \theta) = \sum_{m=1}^{M} \pi_m \Phi(\varphi_i; \mu_m, \Sigma_m)$, where Φ denotes the d-dimensional Gaussian probability distribution function (pdf). We refer to (Banfield and Raftery 1993) or (Celeux and Govaert 1995) for a detailed presentation of such models. Gaussian mixture models are widely used for supervised and unsupervised classification in many applications. But even if the model itself is standard, its use in the context of NLMEM requires particular attention since the individual parameters are not observed. In other words, we aim to create clusters of non observed parameters.

ii) A latency structure can also be introduced at the level of the conditional distribution of the observations (y_{ij}):

$$y_{ij} = f\left(x_{ij};\varphi_i,z_i\right) + g\left(x_{ij};\varphi_i,z_i,\xi\right)\varepsilon_{ij}. \tag{7}$$

A mixture of conditional distributions therefore reduces to a mixture of residual errors and/or a mixture of structural models.

A mixture of residual error models has the general form:

$$g\left(x_{ij};\varphi_i,z_i,\xi\right) = \sum_{m=1}^{M} \mathbb{1}_{z_i=m} g_m\left(x_{ij};\varphi_i,\xi_m\right). \tag{8}$$

As an example, a mixture of constant error models assumes that

$$y_{ij} = f\left(x_{ij};\varphi_i\right) + \sum_{m=1}^{M} \mathbb{1}_{z_i=m}\xi_m\varepsilon_{ij}. \tag{9}$$

Between subject model mixtures (BSMM) assume that the structural model is a mixture of M different structural models:

$$f\left(\cdot;\varphi_i,z_i\right) = \sum_{m=1}^{M} \mathbb{1}_{z_i=m} f_m\left(\cdot;\varphi_i\right). \tag{10}$$

This model is relevant for example to distinguish different types of response to the same treatment. See (Lavielle, Mesa, Chatel, and Vermeulen 2010) for an application to HIV where different viral kinetics models are used to classify treated patients into responders, non-responders and rebounders on the basis of their viral load profiles.

2.3 Log-likelihood of mixture models

The completed data is (y,φ,z), where y and (φ,z) are respectively the observed and unobserved data. The complete log-pdf for subject i is

$$\ell\left(y_i,\varphi_i,z_i;\theta\right) = \tag{11}$$
$$\sum_{m=1}^{M} \mathbb{1}_{z_i=m}\left(\ell_m\left(y_i,\varphi_i;\theta_m\right) + \log\mathbb{P}\left(z_i=m\right)\right),$$

$\ell_m\left(y_i,\varphi_i;\theta_m\right)$ being the log-pdf of the pair of variables (y_i,φ_i) in group G_m defined by

$$G_m = \left\{i, 1\le i\le N \text{ such that } z_i = m\right\}.$$

In the case of a mixture of Gaussian distributions as described in (6), $\theta_m = (\xi,\mu_m,\Sigma_m)$ and the complete log-pdf becomes

$$\ell_m\left(y_i,\varphi_i;\theta_m\right) = \ell\left(y_i|\varphi_i;\xi\right) + \ell_m\left(\varphi_i;\mu_m,\Sigma_m\right).$$

For the mixture of structural models (BSMM) defined in (10), $\theta_m = (\xi,\mu,\Sigma)$ and

$$\ell_m\left(y_i,\varphi_i;\theta_m\right) = \ell_m\left(y_i|\varphi_i;\xi\right) + \ell\left(\varphi_i;\mu,\Sigma\right),$$

while for the mixture or error models defined in (8), $\theta_m = (\xi_m,\mu,\Sigma)$ and

$$\ell_m\left(y_i,\varphi_i;\theta_m\right) = \ell\left(y_i|\varphi_i;\xi_m\right) + \ell\left(\varphi_i;\mu,\Sigma\right).$$

The pdf associated to any combination of these different mixture models is straightforward to derive.

In the following, for the sake of clarity, we will make the assumption that the pdf ℓ_m belongs to the exponential family: there exists a function ψ of θ_m and a minimal sufficient statistic $T(y_i,\varphi_i)$ such that

$$\ell_m\left(y_i,\varphi_i;\theta_m\right) = \left\langle T\left(y_i,\varphi_i\right),\theta_m\right\rangle - \psi\left(\theta_m\right). \tag{12}$$

According to (4), if we assume that $\pi_m = \mathbb{P}\left(z_i=m\right)$, then

$$\ell\left(y_i,\varphi_i,z_i;\theta\right) = \tag{13}$$
$$\left(\mathbb{1}_{z_i=m}\sum_{m=1}^{M}\left\langle T\left(y_i,\varphi_i\right),\theta_m\right\rangle + \log(\pi_m) - \psi\left(\theta_m\right)\right).$$

Then, the pdf of (y,φ,z) also belongs to the exponential family:

$$\ell\left(y,\varphi,z;\theta\right) = \left\langle S\left(y,\varphi,z\right),\theta\right\rangle - \psi\left(\theta\right), \tag{14}$$

where

$$S\left(y,\varphi,z\right) = \tag{15}$$
$$\left(\sum_{i=1}^{n}\mathbb{1}_{z_i=m}, \sum_{i=1}^{n}\mathbb{1}_{z_i=m}T\left(y_i,\varphi_i\right); 1\le m\le M\right).$$

We will take advantage of this representation for our description of the proposed stochastic EM-like algorithms. Indeed, computing any (conditional) expectation of $\ell\left(y,\varphi,z;\theta\right)$ reduces to computing the (conditional) expectation of $S\left(y,\varphi,z\right)$.

Some statistical properties of the MLE for NLMEM can be derived (Online Resource).

3 Algorithms proposed for maximum likelihood estimation

We aim to estimate θ by maximizing the likelihood of the observations (y_i). As mentioned above, we are in the general framework of incomplete data where EM-type algorithms are known to be efficient.

First of all, we assume that the complete likelihood $\mathcal{L}(\theta;y,\varphi,z)$ can be maximized when the complete data is observed. In other words, there exists a function $\hat{\theta}$ such that for any (y,φ,z),

$$\hat{\theta}(S(y,\varphi,z)) = \arg\max\left\{\left\langle S\left(y,\varphi,z\right),\theta\right\rangle - \psi\left(\theta\right)\right\}. \tag{16}$$

3.1 The EM algorithm

Since φ and z are not observed, the EM algorithm replaces $S(y,\varphi,z)$ by its conditional expectation (Dempster, Laird, and Rubin 1977). Then, given some initial value $\theta^{(0)}$, iteration k of the EM algorithm updates $\theta^{(k-1)}$ into $\theta^{(k)}$ with the two following steps:

- **E-step :** evaluate the quantity

$$s_k = \mathbb{E}\left(S\left(y,\varphi,z\right)|y;\theta^{(k-1)}\right).$$

- **M-step:** with respect to (16), compute

$$\theta^{(k)} = \hat{\theta}(s_k).$$

Unfortunately, in the framework of non linear mixed-effects models, there is no explicit expression for the E-step since the relationship between observations y and individual parameters φ is non linear. Several authors have proposed stochastic versions of the EM algorithm which attempt to solve the problem. (Wei and Tanner 1990) proposed the Monte Carlo EM (MCEM) algorithm in which the E-step is replaced by a Monte Carlo approximation based on a large number of independent simulations of the missing data. In recent work, (Wang, Schumitzky, and D'Argenio 2007; Wang, Schumitzky, and D'Argenio 2009) also proposed an MCEM algorithm with importance sampling.

Another EM type algorithm for mixtures of mixed effects models was proposed in (De la Cruz, Quintana, and Marshall 2008). They use an extensive Monte-Carlo integration procedure during the E step for computing the marginal distribution of the observations in each cluster. Unfortunately, the computational effort required by this method is prohibitive for most practical application since the structural model f needs to be evaluated T times (here T is the Monte-Carlo size), at each iteration of the algorithm and for each patient. Furthermore, the authors claim that their procedure converges if conditions that ensure the convergence of EM are fulfilled. This is true in "theory", with an infinite Monte-Carlo size, but nothing can be said in realistic conditions.

We will see in the next sections that the proposed modified SAEM algorithm offers appealing practical and theoretical properties. Indeed, convergence of the algorithm is demonstrated under general conditions. Moreover it is extremely fast and can be used for complex problems.

3.2 The SAEM algorithm

The stochastic approximation version of the EM algorithm, proposed by (Delyon, Lavielle, and Moulines 1999), consists of replacing the E-step by a stochastic approximation obtained using simulated data. Given some initial value $\theta^{(0)}$, iteration k of SAEM consists of the three following steps:

- **S-step:** draw $\left(z^{(k)},\varphi^{(k)}\right)$ with the conditional distribution $p\left(z,\varphi|y,\theta^{(k-1)}\right)$.
- **AE-step:** update s_k according to

$$s_k = s_{k-1} + \delta_k\left(S\left(y,\varphi^{(k)},z^{(k)}\right) - s_{k-1}\right). \tag{17}$$

- **M-step:** compute $\theta^{(k)} = \hat{\theta}(s_k)$.

Here, (δ_k) is a decreasing sequence. In the case of NLMEM, the simulation step cannot be directly performed, and a MCMC procedure can be used ((Kuhn and Lavielle 2004)). Convergence of the parameter sequence $(\theta^{(k)})$ toward a (local) maximum of the likelihood is ensured under general conditions (Delyon, Lavielle, and Moulines 1999; Kuhn and Lavielle 2004; Allassonnière, Kuhn, and Trouvé 2010).

This version of SAEM for mixtures of NLMEM was first implemented in the Monolix software. We have noticed that the algorithm tends to become unstable and produces poor estimations when the problem becomes difficult: small sample sizes, heteroscedastic models, overlap between mixture components, etc. This poor behavior is mainly due to the fact that the S-step of SAEM requires simulation of the categorical variable (z_i), which then impacts the M-step, leading to inference problems. Due to the well known label-switching phenomenon, as pointed out by (Celeux, Hurn, and Robert 2000), uniform ergodicity of the Markov chain $\left(\varphi^{(k)},z^{(k)}\right)$ is no longer guaranteed and convergence of SAEM can not be ensured. We also have noticed that some components of the mixture can disappear during iterations, mainly when these components are not well separated. In the next section, we propose a methodology that avoids simulation of these latent categorical covariates and exhibits improved practical behavior.

3.3 The MSAEM algorithm

We have seen that the E-step of the EM algorithm requires evaluating $\mathbb{E}\left(S\left(y,\varphi,z\right)|y;\theta^{(k-1)}\right)$. We have the following relation:

$$\mathbb{E}\left(S\left(y,\varphi,z\right)|y;\theta\right) = \mathbb{E}\left(\mathbb{E}\left(S\left(y,\varphi,z\right)|y,\varphi,\theta\right)|y;\theta\right).$$

Then, by setting

$$H\left(y,\varphi,\theta\right) = \mathbb{E}\left(S\left(y,\varphi,z\right)|y,\varphi,\theta\right), \tag{18}$$

the E-step of the EM algorithm at iteration k reduces to calculating

$$\mathbb{E}\left(H\left(y,\varphi,\theta^{(k-1)}\right)|y;\theta^{(k-1)}\right).$$

The underlying idea in this operation is to use a conditional distribution that only depends on φ and y but not the latent categorical covariates z. Then, φ becomes the only unobserved variable of the model. Nevertheless, introduction of the latent categorical variable remains very useful since it allows one to derive a manageable expression of the complete likelihood. Then, iteration k of MSAEM requires us to calculate $H\left(y,\varphi;\theta^{(k-1)}\right)$:

- **S-step:** draw $\varphi^{(k)}$ with the conditional distribution $p\left(.|y,\theta^{(k-1)}\right)$.
- **E-step:** compute $H\left(y,\varphi^{(k)};\theta^{(k-1)}\right)$ using (18).
- **AE-step:** update s_k according to

$$s_k = s_{k-1} + \delta_k\left(H\left(y,\varphi^{(k)},\theta^{(k-1)}\right) - s_{k-1}\right). \quad (19)$$

- **M-step :** compute $\theta^{(k)} = \hat{\theta}(s_k)$.

The simulation step of the MSAEM algorithm at iteration k consists of a few MCMC iterations with $p(\varphi|y;\theta^{(k)})$ as the stationary distribution. More precisely, we propose to use the Hasting-Metropolis algorithm, with various proposed kernels. Here, the N subjects are assumed to be independent and the same procedure is used for the N subjects, i.e., for $i = 1, 2, \ldots, N$.

A first kernel consists in using the marginal distribution $p(\varphi_i)$ for generating a candidate φ_i^c. Then, the probability of acceptance, i.e., the probability to move from φ_i to φ_i^c, reduces to

$$\alpha\left(\varphi_i,\varphi_i^c\right) = \min\left(1, \frac{p\left(y|\varphi_i^c;\theta^{(k)}\right)}{p\left(y|\varphi_i;\theta^{(k)}\right)}\right).$$

Another possible kernel is the random walk: $\varphi_i^c \mathcal{N}\left(\varphi^{(k-1)},\Omega\right)$, where Ω is a diagonal matrix which is adaptively adjusted in order to reach a given acceptance rate (typically 0.3). Different directions can be used by setting different elements of the diagonal of Ω to 0 during iteration. Here, the probability of acceptance is

$$\alpha\left(\varphi_i,\varphi_i^c\right) = \min\left(1, \frac{p\left(y,\varphi_i^c;\theta^{(k)}\right)}{p\left(y,\varphi_i;\theta^{(k)}\right)}\right).$$

Practical implementation of this algorithm requires to compute $p(\varphi_i)$ and $p(y|\varphi_i)$. Depending on the type of mixture model considered (mixture of distributions, mixture of structural models, mixture of residual error models,etc.), these two terms can easily be computed in a closed form.

Certain parameters need to be well chosen to improve the convergence of the algorithm, such as the total number of iterations K, the number of iterations of the MCMC procedure during the S-step, and the step-size sequence (δ_k). We remark that selection of the various settings of the algorithm is not a problem related to the particular extension of SAEM to mixture models considered here, but a general issue for practical implementation of SAEM. We will give some leads, but an in-depth discussion of the choice of these settings is beyond the scope of the paper.

Sequence (δ_k) has a strong impact on the speed of convergence of the algorithm. Fast convergence towards a neighborhood of the solution is obtained with a constant sequence $\delta_k = 1$ during the first K_1 iterations of SAEM. Then, the M-step of SAEM reduces to maximizing the complete log-likelihood: for $k = 1, \ldots, K_1$,

$$\theta^{(k)} = \arg\max \mathcal{L}\left(\theta;y,\varphi^{(k)}\right).$$

Thus, if we consider a mixture model, the M-step consists of estimating the components of the mixtures using the observations y and the simulated individual parameters $\varphi^{(k)}$. An EM can be used at iteration k for computing $\theta^{(k)}$. After converging to a neighborhood of the MLE, a decreasing step-size sequence (δ_k) will permit almost sure convergence of the algorithm to a maximum of the observed likelihood (Online Resource). For the numerical experiments presented below, (δ_k) decreases as $1/k$.

When the number of subjects N is small, convergence of the algorithm can be improved by combining the stochastic approximation with Monte-Carlo, i.e., by running R Markov chains in parallel instead of only one chain. The S-step now consists of generating R sequences $\varphi^{(k,1)}, \ldots, \varphi^{(k,R)}$, and (19) becomes

$$s_k = s_{k-1} + \delta_k\left(\frac{1}{R}\sum_{r=1}^{R} H\left(y,\varphi^{(k,r)},\theta^{(k-1)}\right) - s_{k-1}\right).$$

For the numerical experiments, we have set $R = 5$ with $N = 100$.

3.4 Some examples

3.4.1 Mixtures of Normal distributions

We consider here that the distributions of the individual parameters is a mixture of normal distributions

$$\varphi_i \sim_{\text{iid}} \sum_{m=1}^{M} \pi_m \mathcal{N}\left(\mu_m,\Sigma_m\right)$$

Let $p = (p_1, \ldots, p_M)$, $\mu = (\mu_1, \ldots, \mu_M)$ and $\Sigma = (\Sigma_1, \ldots, \Sigma_M)$. Then, the conditional log-pdf of the individual parameters φ is given by

$$\begin{aligned}\ell\left(\varphi|z;\mu,\Sigma\right) = &-\frac{1}{2}\sum_{i=1}^{N}\sum_{m=1}^{M} \mathbb{1}_{z_i=m}\left(d\log(2\pi) + \log|\Sigma_m|\right)\\ &-\frac{1}{2}\sum_{i=1}^{N}\sum_{m=1}^{M} \mathbb{1}_{z_i=m}(\varphi_i-\mu_m)'\Sigma_m^{-1}(\varphi_i-\mu_m),\end{aligned}$$

and the log-pdf of the labels z is:

$$\ell(z;\pi) = \sum_{i=1}^{N}\sum_{m=1}^{M} \mathbb{1}_{z_i=m}\log(\pi_m). \tag{20}$$

On the other hand, we consider a proportional residual model:

$$y_{ij} = f(x_{ij},\varphi_i) + \xi f(x_{ij},\varphi_i)\,\varepsilon_{ij}.$$

Then, the conditional log-pdf of the observations y is given by

$$\ell(y|\varphi,z;\xi) = -\sum_{i,j}\log\left(\xi f(x_{ij},\varphi_i)\right) - \frac{N_{\text{tot}}}{2}\log(2\pi) - \frac{1}{2\xi^2}\sum_{i,j}\left(\frac{y_{ij}-f(x_{ij},\varphi_i)}{f(x_{ij},\varphi_i)}\right)^2, \tag{21}$$

where $N_{\text{tot}} = \sum_{i=1}^{N} n_i$ is the total number of observations. Sufficient statistics of the complete model are:

$$S = (S_{1,m}, S_{2,m}, S_{3,m}, S_4; 1 \le m \le M),$$

where

$$S_{1,m} = \sum_{i=1}^{N} \mathbb{1}_{z_i=m} \tag{22}$$

$$S_{2,m} = \sum_{i=1}^{N} \mathbb{1}_{z_i=m}\varphi_i \tag{23}$$

$$S_{3,m} = \sum_{i=1}^{N} \mathbb{1}_{z_i=m}\varphi_i\varphi_i' \tag{24}$$

$$S_4 = \sum_{i=1}^{N}\sum_{j=1}^{n_i} \left(y_{ij}/f(x_{ij},\varphi_i) - 1\right)^2, \tag{25}$$

and the function $\hat{\theta}$ is given by:

$$\hat{\pi}_m(S) = S_{1,m}/N \tag{26}$$

$$\hat{\mu}_m(S) = S_{2,m}/S_{1,m} \tag{27}$$

$$\hat{\Sigma}_m(S) = \frac{S_{3,m}}{S_{1,m}} - \left(\frac{S_{2,m}}{S_{1,m}}\right)\left(\frac{S_{2,m}}{S_{1,m}}\right)' \tag{28}$$

$$\hat{b}(S) = \sqrt{S_4/N_{\text{tot}}}. \tag{29}$$

At iteration k, SAEM requires using simulated sequences $\varphi^{(k)}$ and $z^{(k)}$ for updating the set of statistics defined in (22-25) using the stochastic approximation scheme defined in (17).

Instead, at iteration k, MSAEM consists of using only the simulated sequence $\varphi^{(k)}$ for computing $H\left(y,\varphi;\theta^{(k-1)}\right)$ using (18), and updating the set of statistics using the stochastic approximation scheme defined in (19).

The minimal sufficient statistic here is $\mathbb{1}_{z_i=m}$, and the E-step of iteration k of MSAEM reduces to the evaluation of:

$$\begin{aligned}\gamma_{i,m}^{(k)} &= \mathbb{E}\left(\mathbb{1}_{z_i=m}|y_i,\varphi_i^{(k)};\theta^{(k-1)}\right)\\ &= \mathbb{P}\left(z_i=m|y_i,\varphi_i^{(k)};\theta^{(k-1)}\right)\\ &= \mathbb{P}\left(z_i=m|\varphi_i^{(k)};\mu^{(k-1)},\Sigma^{(k-1)},p^{(k-1)}\right)\\ &= \frac{\pi_m^{(k-1)}\nu(\varphi_i^{(k)};\mu_m^{(k-1)},\Sigma_m^{(k-1)})}{\sum_{r=1}^{M}\pi_r^{(k-1)}\nu(\varphi_i^{(k)};\mu_r^{(k-1)},\Sigma_r^{(k-1)})},\end{aligned}$$

where ν is the pdf of a Gaussian vector.

The zero-one variable $\mathbb{1}_{z_i=m}$ present in expressions when applying SAEM is replaced at iteration k in MSAEM by the probability $\mathbb{P}\left(z_i=m|\varphi_i^{(k)},\theta^{(k-1)}\right)$, and permits us to tackle the problems mentioned before. Then, the A-step of MSAEM reduces to:

$$s_{k,1,m} = s_{k-1,1,m} + \delta_k\left(\sum_{i=1}^{N}\gamma_{i,m}^{(k)} - s_{k-1,1,m}\right)$$

$$s_{k,2,m} = s_{k-1,2,m} + \delta_k\left(\sum_{i=1}^{N}\gamma_{i,m}^{(k)}\varphi_i^{(k)} - s_{k-1,2,m}\right)$$

$$s_{k,3,m} = s_{k-1,3,m} + \delta_k\left(\sum_{i=1}^{N}\gamma_{i,m}^{(k)}\varphi_i^{(k)}\varphi_i^{(k)\prime} - s_{k-1,3,m}\right)$$

$$s_{k,4} = s_{k-1,4} + \delta_k\left(\sum_{i,j}\left(\frac{y_{ij}-f(x_{ij},\varphi_i^{(k)})}{f(x_{ij},\varphi_i^{(k)})}\right)^2 - s_{k-1,4}\right).$$

Parameters are then updated using the function $\hat{\theta}$ defined above.

3.4.2 Mixtures of residual error models

Suppose now a proportional residual model in each group, given by $g_m(x_{ij},\varphi_i,\xi) = \xi_m f(x_{ij},\varphi_i)$, where $\xi = (\xi_1,\ldots,\xi_m)$. Then, the conditional log-pdf of the observations in group G_m is now:

$$\begin{aligned}\ell_m(y_i|\varphi_i;\xi) &= \ell(y_i|\varphi_i;\xi_m)\\ &= -\frac{1}{2\xi_m^2}\sum_{j=1}^{n_i}\left(\frac{y_{ij}-f(x_{ij},\varphi_i)}{f(x_{ij},\varphi_i)}\right)^2\\ &\quad -\sum_{j=1}^{n_i}\log\left(\xi_m f(x_{ij},\varphi_i)\right) - \frac{n_i}{2}\log(2\pi),\end{aligned}$$

On the other hand, if we assume the same normal distribution for all the individual parameters, then the log-pdf of the individual parameters is

$$\ell(\varphi_i;\mu,\Sigma) = -\frac{1}{2}(\varphi_i-\mu)'\Sigma^{-1}(\varphi_i-\mu) - \frac{d}{2}\log(2\pi) - \frac{1}{2}\log(|\Sigma|).$$

Here, the E-step requires computing:

$$\gamma_{i,m}^{(k)} = \mathbb{P}\left(z_i = m | y_i, \varphi_i^{(k)}, \theta^{(k-1)}\right) = \frac{\pi_m^{(k-1)} \ell\left(y_i | \varphi_i^{(k)}; \xi_m^{(k-1)}\right)}{\sum_{r=1}^{M} \pi_r^{(k-1)} \ell\left(y_i | \varphi_i^{(k)}; \xi_m^{(k-1)}\right)}.$$

In the A-step, we approximate several minimal sufficient statistics as follows:

$$\begin{aligned} s_{k,i,1,m} &= s_{k-1,i,1,m} + \delta_k \left(\gamma_{i,m}^{(k)} - s_{k-1,i,1,m}\right) \\ s_{k,2} &= s_{k-1,2} + \delta_k \left(\sum_{i=1}^{N} \varphi_i^{(k)} - s_{k-1,2}\right) \\ s_{k,3} &= s_{k-1,m} + \delta_k \left(\sum_{i=1}^{N} \varphi_i^{(k)} \varphi_i^{(k)'} - s_{k-1,m}\right) \\ s_{k,4,m} &= s_{k-1,4,m} \\ &+\delta_k \left(\sum_{i,j} \gamma_{i,m}^{(k)} \left(\frac{y_{ij} - f_m\left(x_{ij}, \varphi_i^{(k)}\right)}{f_m\left(x_{ij}, \varphi_i^{(k)}\right)}\right)^2 - s_{k-1,4,m}\right). \end{aligned}$$

In the M-step, we update parameters according to:

$$\begin{aligned} \pi_m^{(k)} &= \frac{1}{N} \sum_{i=1}^{N} s_{k,1,i,m} \\ \mu^{(k)} &= \frac{s_{k,2}}{N} \\ \Sigma^{(k)} &= \frac{s_{k,3}}{N} - \left(\frac{s_{k,2}}{N}\right)\left(\frac{s_{k,2}}{N}\right)' \\ \xi_m^{(k)} &= \sqrt{\frac{s_{k,4,m}}{\sum_{i=1}^{N} n_i s_{k,1,i,m}}}. \end{aligned}$$

3.5 Estimation of the individual parameters

For a given set of population parameters θ, we use each individual conditional distribution $p(z_i, \varphi_i | y_i, \theta)$ for estimating the latent variable z_i and the vector of individual parameters φ_i.

A first estimate is the Maximum a Posteriori (MAP) which is obtained by maximizing this joint conditional distribution with respect to (z_i, φ_i):

$$(\hat{z}_i, \hat{\varphi}_i) = \arg\max_{(z_i, \varphi_i)} p\left(z_i, \varphi_i | y, \theta\right) \tag{30}$$

Such a maximization is not straightforward and requires performing a two-step procedure:

1) For $m = 1, \ldots, M$ compute

$$\hat{\varphi}_{i,m} = \arg\max_{\varphi_i} p(y_i | \varphi_i, z_i = m; \theta) p(\varphi_i | z_i = m; \theta) \tag{31}$$

2) Compute

$$\hat{m}_i = \arg\max_{m} p(y_i, \hat{\varphi}_{i,m} | z_i = m; \theta) \mathbb{P}\left(z_i = m; \theta\right) \tag{32}$$

and set

$$(\hat{z}_i, \hat{\varphi}_i) = (\hat{m}_i, \hat{\varphi}_{i,\hat{m}_i}). \tag{33}$$

Another estimate of the latent covariate z_i maximizes the marginal conditional distribution:

$$\hat{z}_i = \arg\max_{m} \mathbb{P}\left(z_i = m | y_i; \theta\right), \tag{34}$$

where

$$\begin{aligned} \mathbb{P}\left(z_i = m | y_i, \theta\right) &= \mathbb{E}\left(\mathbb{P}\left(z_i = m | y_i, \varphi_i, \theta\right) | y_i, \theta\right) \\ &= \mathbb{E}\left(\gamma_{i,m} | y_i, \theta\right), \end{aligned} \tag{35}$$

which can be estimated using the stochastic approximation procedure described in Sections 3.4.1 and 3.4.2.

Instead of maximizing the conditional distribution for estimating φ_i, an alternative is to compute the conditional mean $\hat{\varphi}_i = \mathbb{E}\left(\varphi_i | y_i; \theta\right)$.

We remark that

$$\mathbb{E}\left(\varphi_i | y_i; \theta\right) = \sum_{m=1}^{M} \mathbb{E}\left(\varphi_i | y_i, z_i = m; \theta\right) \mathbb{P}\left(z_i = m | y_i; \theta\right).$$

Then, estimating the conditional expectation of the individual parameters requires estimating the conditional probabilities $\mathbb{P}\left(z_i = m | y_i; \theta\right)$ and the conditional means in each group, $\mathbb{E}\left(\varphi_i | y_i, z_i = m; \theta\right)$.

We have seen above how to estimate $\mathbb{P}\left(z_i = m | y_i; \theta\right)$ using stochastic approximation. On the other hand, $\mathbb{E}\left(\varphi_i | y_i, z_i = m; \theta\right)$ can easily be estimated by MCMC.

4 Numerical experiments

A simulation study was conducted to evaluate the performance of the proposed algorithm for estimating the parameters of the different non linear mixed-effects mixture models. We used a pharmacokinetics (PK) model for these numerical experiments. The vector of individual PK parameters of subject i is

$$\varphi_i = \left(\log(ka_i), \log(V_i), \log(Cl_i)\right), \tag{36}$$

where V_i is the volume of distribution, Cl_i the clearance and ka_i the absorption rate constant of the subject. We define φ_i as the set of log-parameters, since log-normal distributions will be used for describing the inter-subject variability of these PK parameters.

The structural model is an oral administration PK model with one compartment, first order absorption

and linear elimination. The plasmatic concentration of drug predicted by the model is:

$$f(\varphi_i, x_{ij}) = \frac{D_i k a_i}{V_i\left(ka_i - \frac{Cl_i}{V_i}\right)}\left(e^{-\frac{Cl_i}{V_i}x_{ij}} - e^{-ka_i x_{ij}}\right). \tag{37}$$

Here, (x_{ij}) are the measurement times of subject i, and D_i the dose administered at time 0. We use the same amount of drug $D_i = 1000$mg and the same measurement times for the N subjects (times are in hours):

$$x_i = (0.25,\ 1,\ 2.5,\ 6,\ 16,\ 26,\ 72)$$

For each of the four examples considered, $L = 100$ datasets were simulated and the parameters were estimated using MSAEM. Let $\theta^\star$ be the parameter used for the simulation and $\hat{\theta}_\ell$ be the estimated parameter obtained with the ℓ-th simulated dataset. The following quantities were computed:

- the relative estimation errors (in %): for $1 \leq \ell \leq L$,

$$\text{REE}_\ell = \frac{\hat{\theta}_\ell - \theta^\star}{\theta^\star} \times 100.$$

- The relative root mean square error (in %):

$$\text{RRMSE} = \frac{\sqrt{\frac{1}{L}\sum_{\ell=1}^{L}(\hat{\theta}_\ell - \theta^\star)^2}}{\theta^\star} \times 100.$$

4.1 Mixtures of distributions

We assume a proportional error model for the observed concentration:

$$y_{ij} = f(\varphi_i, x_{ij}) + \xi f(\varphi_i, x_{ij})\varepsilon_{ij}, \tag{38}$$

where $\varepsilon_{ij} \sim \mathcal{N}(0,1)$ and $\xi = 0.2$. We assume that ka_i and Cl_i are log-normally distributed:

$$\log(ka_i) \sim \mathcal{N}(\mu_1, \sigma_1^2),$$
$$\log(Cl_i) \sim \mathcal{N}(\mu_3, \sigma_3^2),$$

with $\mu_1 = 1$, $\mu_3 = 4$, $\sigma_1^2 = 0.04$ and $\sigma_3^2 = 0.04$. Mixtures of distributions will be used for V_i. Scenarios 1 and 2 assume a homoscedastic model:

$$\log(V_i) \sim p_1\mathcal{N}(\mu_{21}, \sigma_2^2) + p_2\mathcal{N}(\mu_{22}, \sigma_2^2),$$

with the following numerical values

S1: $p_2 = 0.7$, $\mu_{21} = 30$, $\mu_{22} = 70$, $\sigma_2^2 = 0.04$,
S2: $p_2 = 0.7$, $\mu_{21} = 30$, $\mu_{22} = 50$, $\sigma_2^2 = 0.04$.

The difference between the two means is significantly reduced in Scenario 2 compared to Scenario 1.

Scenario 3 assumes an heteroscedastic model:

$$\log(V_i) \sim p_1\mathcal{N}(\mu_{21}, \sigma_{21}^2) + p_2\mathcal{N}(\mu_{22}, \sigma_{22}^2),$$

with

S3: $p_2 = 0.7$, $\mu_{21} = 30$, $\mu_{22} = 50$, $\sigma_{21}^2 = 0.08$, $\sigma_{22}^2 = 0.04$.

Fig. 1 displays the probability distribution functions of $\log(V_i)$ under the three scenarios. The distributions are well separated in Scenario 1. The overlapping between the two distributions increases in Scenario 2 since the two distributions become closer. Increasing one of the variances in Scenario 3 further increases this overlapping.

Fig. 2 displays the distribution of the observed concentration in both groups under each scenario. Medians and 90% confidence intervals are used to summarize these distributions.

Results obtained with the MSAEM algorithm are displayed in Fig. 3. We show the distribution of the relative estimation errors (REE_ℓ) for each parameter under each scenario, with $N = 100$ and $N = 1000$ subjects. Relative root mean square errors are presented in Table 1. These are compared with those obtained in differing situations, i.e., when φ and/or z are known.

Results obtained with scenario S1 are very similar whether or not z is known. Indeed, the two components of the mixture are well separated here, and the conditional probabilities of belonging to each class are close to 0 or 1. The results deteriorate with scenarios S2 and S3 for the parameters of the mixture which are much less-well estimated when z is unknown. It is also interesting to notice that the other model parameters are little affected by knowledge of z.

Lastly, we remark that the differences when individual parameters (φ_i) are known or not have little impact on the estimation of the parameters of the mixture. The most difficult parameter to estimate when (φ_i) is unknown is the variance of $\log(ka_i)$. This is a purely statistical issue related to the quantity of information in the data, and independent of the mixture model: few observations are available during the absorption phase which makes it difficult to estimate the absorption rate constant ka. The boxplots confirm that parameters are better estimated with $N = 1000$, but even with $N = 100$, we do not see any bias in the estimation of the parameters, with the exception perhaps of the variances of the mixture in scenario S3, which are poorly estimated.

Fig. 4 provides a graphical illustration of the probability of correct classification in both groups for the different scenarios and $N = 1000$ subjects. For each of the $K = 100$ runs, the probabilities of correct classification were ranked in increasing order and the median of these $K = 100$ ranked sequences was computed. The same procedure was then repeated, but assuming that the individual parameters (φ_i) were known. Fig.4 compares these two medians. As expected, for each scenario, the probabilities of correct classification are greater when

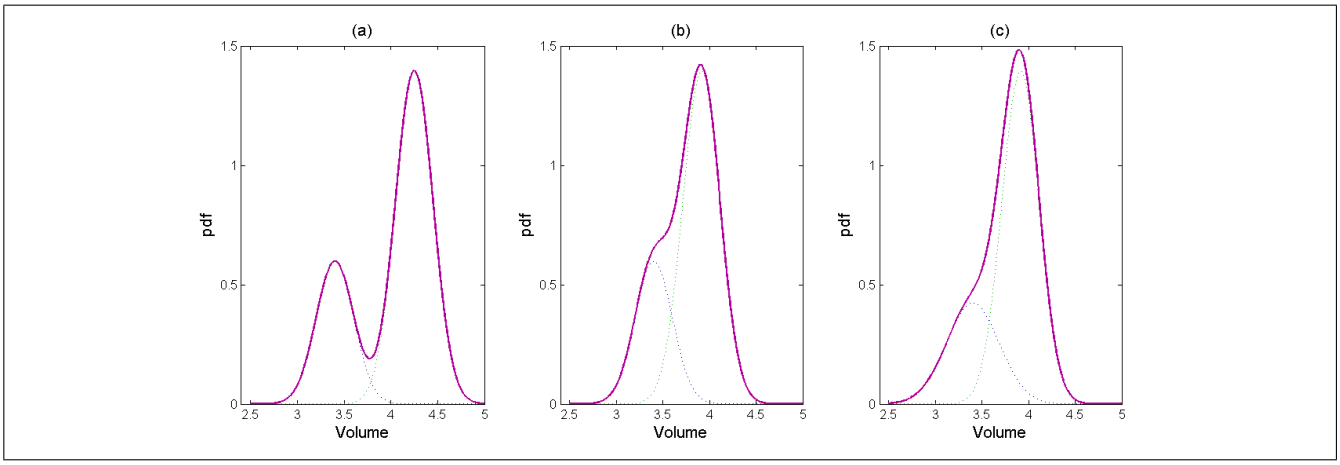

Fig. 1: Probability distribution function of the log-volume, (a) scenario S1: $\mu_1 = 30$, $\mu_2 = 70$, $\sigma^2_{21} = \sigma^2_{22} = 0.04$; (b) scenario S2: $\mu_1 = 30$, $\mu_2 = 50$, $\sigma^2_{21} = \sigma^2_{22} = 0.04$; (c) scenario S3: $\mu_1 = 30$, $\mu_2 = 50$, $\sigma^2_{21} = 0.08$, $\sigma^2_{22} = 0.04$.

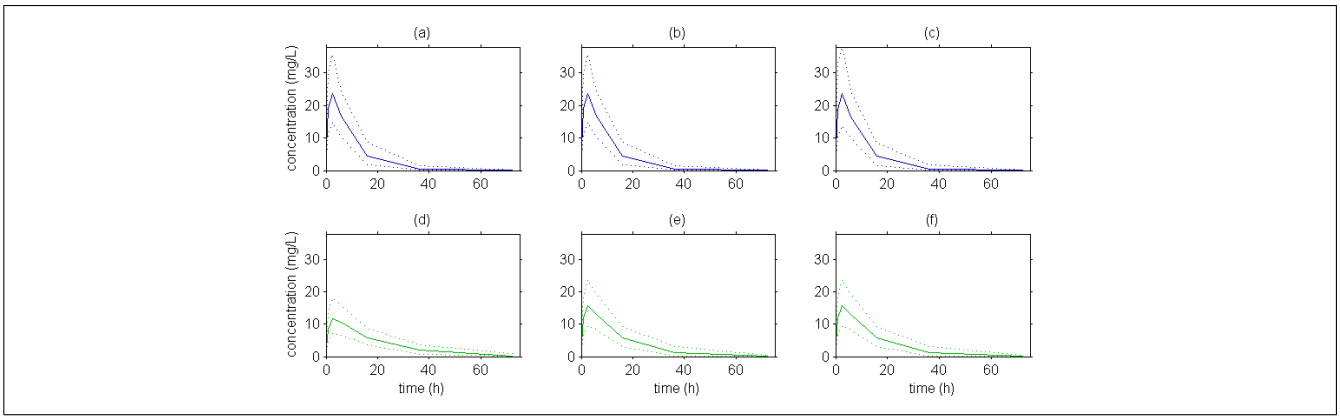

Fig. 2: Median (solid line) and 90% prediction interval (dotted line) of the observed concentration in different groups: (a-c) group 1, (d-f) group 2, and with different scenarios: (a)&(d) S1, (b)&(e) S2, (c)&(f).

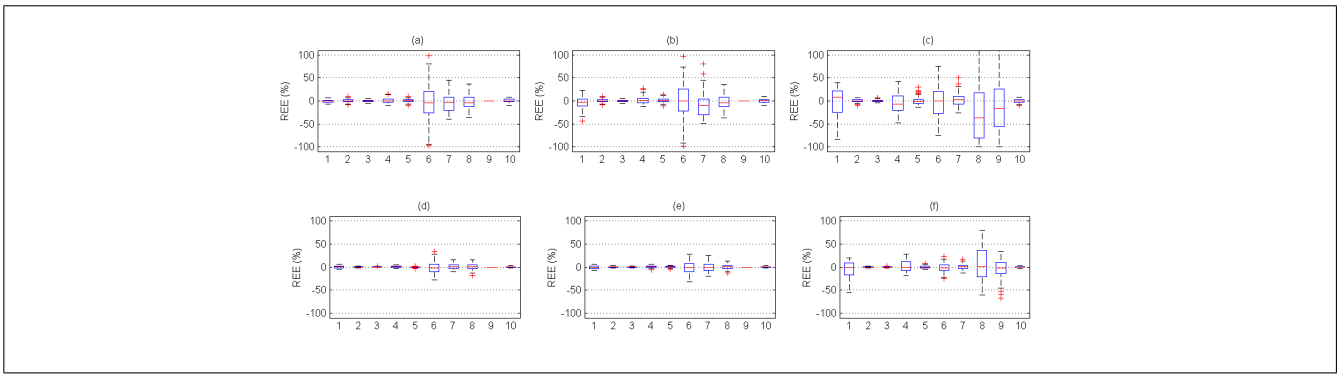

Fig. 3: Empirical distribution of the relative estimation errors (REE_ℓ) with different sample sizes: (a-c) $N = 100$, (d-f) $N = 1000$, and different scenarios: (a)&(d) S1, (b)&(e) S2, (c)&(f) S3. The estimated parameters are 1: p_2; 2: μ_1; 3: μ_{21}; 4: μ_{22}; 5: μ_3; 6: σ^2_1; 7: σ^2_{21}; 8: σ^2_{22} (only in S3, *i.e.* (c)&(f)); 9: σ^2_3; 10: ξ.

		N=100				N=1000			
θ	θ^*	z known φ known	z unknown φ known	z known φ unknown	z unknown φ unknown	z known φ known	z unknown φ known	z known φ unknown	z unknown φ unknown
p_2	0.7	6.74	6.95	6.07	6.87	2.04	2.17	2.19	2.21
μ_1	1	1.98	1.98	2.67	2.96	0.68	0.68	1.07	0.93
μ_{21}	30	3.30	3.95	4.25	5.35	1.24	1.40	1.50	1.73
μ_{22}	70	2.26	2.42	2.57	3.19	0.75	0.78	0.91	0.95
μ_3	4	1.75	1.75	2.13	2.24	0.62	0.62	0.69	0.65
σ_1^2	0.04	15.80	15.80	38.93	40.46	4.18	4.18	12.80	11.44
σ_2^2	0.04	13.48	13.99	14.67	18.91	5.51	5.95	5.09	5.38
σ_3^2	0.04	13.88	13.88	19.93	16.07	4.45	4.45	6.93	6.38
b	0.20	2.58	2.58	3.87	4.00	0.91	0.91	1.23	1.22

Table 1 Relative Root Mean Square Errors (RRMSE) in % of parameter estimates in Scenario 1, with $N = 100$ and $N = 1000$, assuming that φ and/or z are known or unknown.

		N=100				N=1000			
θ	θ^*	z known φ known	z unknown φ known	z known φ unknown	z unknown φ unknown	z known φ known	z unknown φ known	z known φ unknown	z unknown φ unknown
p_2	0.7	7.06	10.84	6.07	12.34	1.88	3.06	2.19	3.73
μ_1	1	1.92	1.92	2.64	2.98	0.56	0.56	1.11	1.08
μ_{21}	30	3.83	6.32	4.15	7.91	1.13	1.83	1.53	2.33
μ_{22}	50	2.63	3.83	2.45	4.91	0.71	1.04	0.93	1.65
μ_3	4	1.91	1.91	2.15	2.29	0.59	0.59	0.69	0.66
σ_1^2	0.04	14.23	14.23	37.85	38.09	4.48	4.48	12.48	11.34
σ_2^2	0.04	15.56	21.64	19.57	26.54	5.02	6.62	5.17	9.37
σ_3^2	0.04	13.91	13.91	14.92	16.11	4.64	4.64	6.57	5.18
b	0.20	2.59	2.59	3.88	4.08	0.84	0.84	1.23	1.28

Table 2 Relative Root Mean Square Errors (RRMSE) in % of parameter estimates in Scenario 2, with $N = 100$ and $N = 1000$, assuming that φ and/or z are known or unknown.

		N=100				N=1000			
θ	θ^*	z known φ known	z unknown φ known	z known φ unknown	z unknown φ unknown	z known φ known	z unknown φ known	z known φ unknown	z unknown φ unknown
p_2	0.7	0.00	18.81	6.06	32.17	0.00	10.76	2.18	17.30
μ_1	1	1.95	1.95	2.55	3.27	0.60	0.60	1.05	1.02
μ_{21}	30	6.01	15.77	5.75	22.80	1.62	8.34	1.98	12.12
μ_{22}	50	2.56	4.80	2.44	7.85	0.74	1.96	0.92	2.61
μ_3	4	1.85	1.85	2.16	2.24	0.63	0.63	0.70	0.74
σ_1^2	0.04	14.15	14.15	35.05	36.90	4.58	4.58	12.52	9.94
σ_{21}^2	0.08	24.62	53.45	30.95	64.80	8.18	24.84	11.14	34.71
σ_{22}^2	0.04	23.93	34.14	21.22	60.60	7.34	13.01	7.49	20.75
σ_3^2	0.04	15.77	15.77	14.64	15.20	4.02	4.02	5.32	5.96
b	0.20	3.07	3.07	3.82	3.30	0.78	0.78	1.27	1.17

Table 3 Relative Root Mean Square Errors (RRMSE) in % of parameter estimates in Scenario 3, with $N = 100$ and $N = 1000$, assuming that φ and/or z are known or unknown.

(φ_i) is known, but it is interesting to notice that the difference is relatively small. As already mentioned, the difficulty of the estimation problem increases from Scenario 1 to Scenario 3. We can see that the difficulty of the classification problem also increases: it is obviously much more difficult to correctly classify the subjects under Scenario 3, where there is a lot of overlap, than under Scenario 1, where the two distributions are well separated.

4.2 Mixtures of error models

We still use the same PK model, but assuming now a mixture of residual error models:

$$\text{if } z_i = 1\ , \quad y_{ij} = f(\varphi_i, x_{ij}) + \xi_1 f(\varphi_i, x_{ij})\,\varepsilon_{ij},$$
$$\text{if } z_i = 2\ , \quad y_{ij} = f(\varphi_i, x_{ij}) + \xi_2 f(\varphi_i, x_{ij})\,\varepsilon_{ij},$$

with $\mathbb{P}(z_i = 1) = 0.3$ and $\xi_1 = 0.1$, $\xi_2 = 0.2$.

We assume that ka_i, V_i and Cl_i are log-normally distributed:

$$\log(ka_i) \sim \mathcal{N}\left(\mu_1, \sigma_1^2\right)$$
$$\log(V_i) \sim \mathcal{N}\left(\mu_2, \sigma_2^2\right)$$
$$\log(Cl_i) \sim \mathcal{N}\left(\mu_3, \sigma_3^2\right).$$

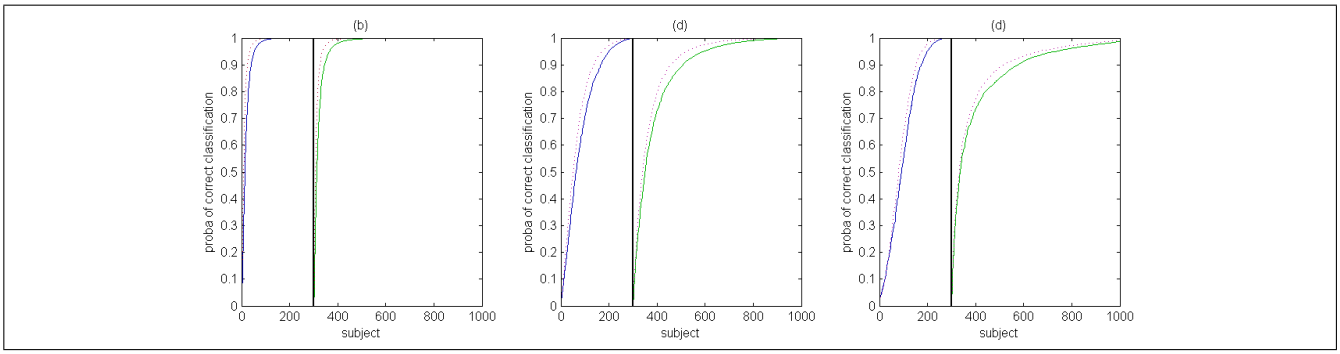

Fig. 4: Medians of the probabilities of correct classification ranked in increasing order in both groups for the different scenarios and $N = 1000$ subjects: (a) scenario S1 ; (b) scenario S2 ; (c) scenario S3. Group 1 is in blue (left) and group 2 in green (right). Solid line: the individual parameters (φ_i) are unknown ; dotted line: the individual parameters (φ_i) are known.

with $\mu_1 = 1$, $\mu_2 = 30$, $\mu_3 = 4$, $\sigma_1^2 = 0.04$, $\sigma_2^2 = 0.04$ and $\sigma_3^2 = 0.04$.

Numerical results are summarized in Table 4. The comments we can make about these results are similar to those for the previous examples: the fact that z is known or unknown affects mainly the estimation of parameters of the mixture model: proportions p_1 and p_2 and standard deviations ξ_1 and ξ_2.

5 An application to PK data

A drug X was orally administered to 199 patients. Each patients received one dose per day, during a period that varies between 1 and 14 days. The pharmacokinetics model that was shown to better describe the process is a 2 compartments model with linear absorption and linear elimination:

$$\dot{A}_d(t) = -k_a A_d(t)$$
$$\dot{A}_c(t) = k_a A_d(t) - k_e A_c(t) - k_{12} A_c(t) + k_{21} A_p(t)$$
$$\dot{A}_p(t) = k_{12} A_c(t) - k_{21} A_p(t)$$

where A_d is the amount of drug in the depot compartment, A_c the amount in the central compartment and A_p the amount in the peripheral compartment. There is no drug in any compartment before the administration of the drug: for any $t < 0$, $A_d(t) = A_c(t) = A_p(t) = 0$. If an amount D of drug is administrated at time τ, then $A_d(\tau^+) = A_d(\tau^-) + D$.

The concentration of drug $Cc = A_c/V$ is measured in the central compartment, where V is the volume of the central compartment.

Here, the individual PK parameters $(k_a, V, k_e, k_{12}, k_{21})$ are log-normally distributed. Then, the Gaussian vector φ_i is the vector of log-parameters. The variance-covariance matrix of φ_i is assumed to be diagonal.

The measured log-concentration is assumed to be normally distributed with a constant error model:

$$\log(y_{ij}) = \log(Cc(t_{ij}; \varphi_i)) + a\varepsilon_{ij}.$$

We first used MONOLIX to fit this NLMEM to the PK data. Observed concentration data from 4 patients with their concentrations predicted by the model are displayed Figure 5.

We then used a mixture of two log-normal distributions for modelling the distribution of the PK parameters. We used a forward strategy for selecting the best mixture model: we first assumed that only one of the five PK parameter distributions was a mixture of two distributions and we compared the five possible models with only one component modelled as a mixture. The model with the highest likelihood value was selected (k_e was selected). We then looked for a second parameter among the four remaining ones (k_{12} was selected), then a third parameter among the three remaining ones (V was selected), and lastly the best combination of four mixtures (k_a was selected). We then used the BIC criteria for comparing the six selected models, including the model without any mixture component and the one with all the parameters modelled with mixtures. All the results are summarized Table 5. The final model was a model assuming that the distributions of V, k_e and k_{12} are mixtures of log-normal distributions. The five estimated distributions of the five PK parameters are displayed Figure 6.

6 Discussion

There exist very few methods available for maximum likelihood estimation in mixtures of non linear mixed effects models. Methods implemented in the nlme R package and in NONMEM are based on a linearization

		N=100				N=1000			
θ	θ^*	z known φ known	z unknown φ known	z known φ unknown	z unknown φ unknown	z known φ known	z unknown φ known	z known φ unknown	z unknown φ unknown
p_2	0.3	6.73	11.76	6.06	20.97	2.01	2.93	2.18	4.80
μ_1	1	1.94	1.94	2.70	2.92	0.60	0.60	0.96	0.87
μ_2	30	1.98	1.98	2.28	2.29	0.64	0.64	0.72	0.73
μ_3	4	1.84	1.84	2.18	2.23	0.56	0.56	0.67	0.65
σ_1^2	0.04	15.94	15.94	26.00	36.91	4.67	4.67	10.66	9.33
σ_2^2	0.04	12.36	12.36	14.21	13.62	4.80	4.80	5.89	5.26
σ_3^2	0.04	15.38	15.38	15.23	16.07	4.74	4.74	4.95	4.57
b_1	0.10	5.36	9.41	6.38	19.60	1.84	2.55	2.49	5.08
b_2	0.20	3.17	4.05	4.28	14.35	0.96	1.14	1.30	1.97

Table 4 Relative Root Mean Square Errors (RRMSE) of parameter estimates in Scenario 3, with $N = 100$ and $N = 1000$, assuming that φ and/or z are known or unknown.

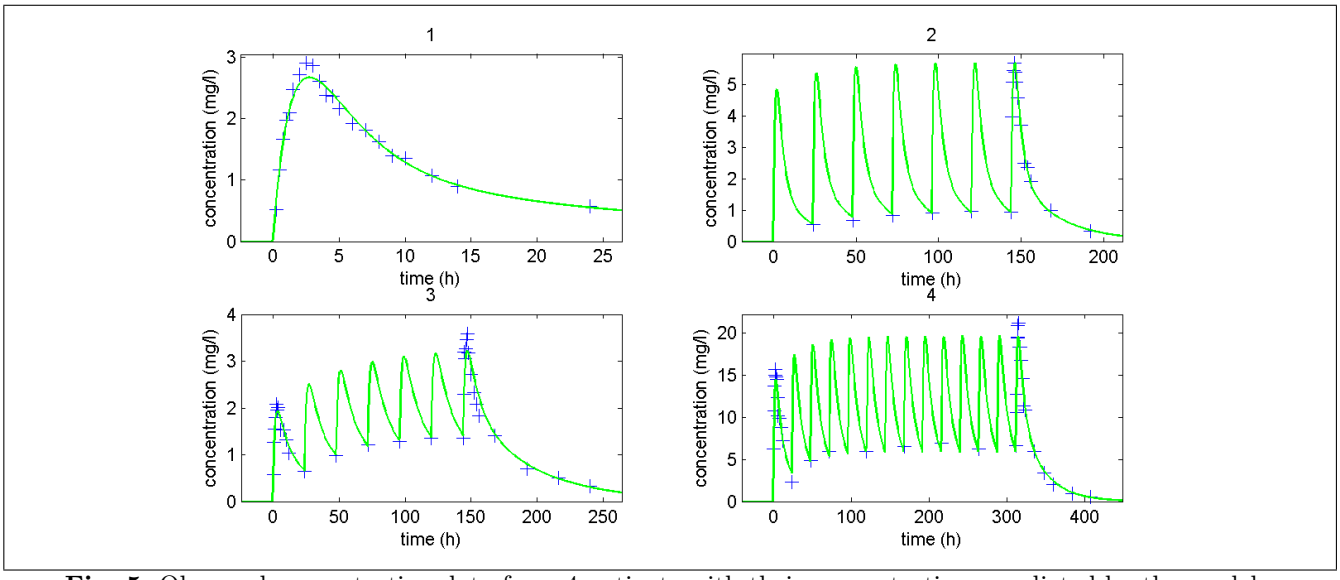

Fig. 5: Observed concentration data from 4 patients with their concentrations predicted by the model.

Parameters with a mixture distribution	BIC
–	1586.4
k_e	1559.4
(k_e, k_{12})	1534.2
(k_e, k_{12}, V)	1372.2
(k_e, k_{12}, V, k_a)	1376.4
(k_e, k_{12}, V, k_a, k_{21})	1391.4

Table 5 The six best models obtained for different number of mixture distributions and the corresponding Bayesian Information Criteria.

of the likelihood. These methods are known to pose real practical problems in several situations (bias, strong influence of the initial guess, poor convergence,...). Moreover the theoretical properties of the estimates obtained with these methods are unknown in most situations.

Some EM-types methods were developed for mixed models, including mixtures of NLMEM. These methods usually intend to approximate the E step of EM by a Monte-Carlo integration. The MCEM algorithm uses this Monte-Carlo integration for computing the conditional distribution $p(\varphi|y;\theta)$ while the procedure proposed in (De la Cruz, Quintana, and Marshall 2008) aims to integrate the joint distribution $p(\varphi, y;\theta)$ for computing the marginal distribution of y in each cluster. These methods can be drastically time-consuming when the structural model is complex, which is the case for most PKPD applications for instance.

We have proposed an extension of the SAEM algorithm for mixtures of mixed effects models. The model is very general, including mixtures of distributions, mixtures of residual error models and mixtures of structural models. Convergence of MSAEM toward a (local) maximum of the observed likelihood is obtained under very general conditions. The algorithm is fast mainly

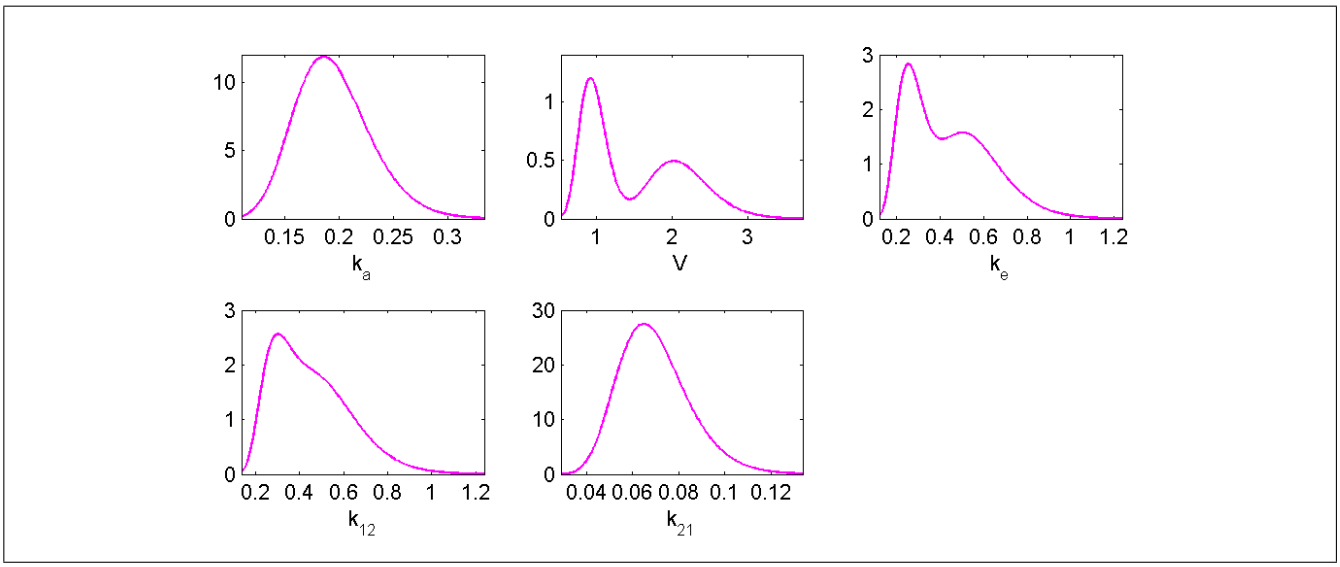

Fig. 6: Probability distribution functions of the five PK parameters. Distributions of V, k_e and k_{12} are mixtures of log-normal distributions ; distributions of k_a and k_{21} are log-normal distributions.

because the Monte-Carlo integration is replaced by a Stochastic Approximation. Indeed, only one Markov Chain needs to be drawn since the integration is performed over the iterations of the algorithm and not over the chains. Moreover, it exhibits very little sensitivity to the initial value, which is a very valuable property for practical applications. This algorithm for mixtures of NLMEM is now implemented in the MONOLIX software. MONOLIX is free for academic research and for students. Several demo examples including mixtures of NLMEM are available with the software.

Several extensions of the proposed method would be of particular interest. First, an optimal strategy for model building would be very useful, for selecting both the mixture structure and the number of clusters. Some specific tools for model assessment are also required for demonstrating that the selected model is capable to generate data similar to the observed ones. Lastly, we have only considered here the Maximum Likelihood approach for these models. Estimation in a Bayesian framework can also be done using posterior simulation via Markov chain Monte Carlo (MCMC) methods, see for example (Frühwirth-Schnatter 2006; De la Cruz, Quintana, and Marshall 2008).

Acknowledgment

The research leading to these results has received support from the Innovative Medicines Initiative Joint Undertaking under grant agreement n. 115156, resources of which are composed of financial contributions from the European Union's Seventh Framework Programme (FP7/2007-2013) and EFPIA companies in kind contribution. The DDMoRe project is also financially supported by contributions from Academic and SME partners.

References

Allassonnière, S., E. Kuhn, and A. Trouvé (2010). Construction of Bayesian deformable models via a stochastic approximation algorithm: A convergence study. *Bernouilli 16*, 641–678.

Banfield, J. D. and A. E. Raftery (1993). Model-based Gaussian and non Gaussian clustering. *Biometrics 49*, 803–821.

Celeux, G. and G. Govaert (1995). Gaussian parsimonious clustering models. *Pattern Recoguition 28*, 781–793.

Celeux, G., M. Hurn, and C. Robert (2000). Computational and inferential difficulties with mixtures posterior distribution. *J. American Statist. Assoc. 95*(3), 957–979.

Davidian, M. and D. M. Giltinan (1993). Some simple methods for estimating intraindividual variability in nonlinear random effects models. *Biometrics 49*, 59–73.

De la Cruz, R., F. Quintana, and G. Marshall (2008). Model based clustering for longitudinal data. *Computational Statistics and Data Analysis 52*(3), 1441–1457.

Delyon, B., M. Lavielle, and E. Moulines (1999). Convergence of a stochastic approximation version of the EM algorithm. *The Annals Of Statistics 27*, 94–128.

Dempster, A. P., Laird N. M. , and Rubin D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *JRSS 39*(B), 1–38.

Frühwirth-Schnatter, S. (2006). *Finite Mixture and Markov Switching Models*. New York: Springer.

Hou, W., H. Li, B. Zhang, M. Huang, and R. Wu (2008). A nonlinear mixed-effect mixture model for functional mapping of dynamic traits. *Heredity 101*, 321–328.

Ketchum, J., A. Best, and V. Ramakrishnan (2012). A within-subject normal-mixture model with mixed-effects for analyzing heart rate variability. *J. Biomet Biostat S7:013*.

Kuhn, E. and M. Lavielle (2004). Coupling a stochastic approximation version of EM with a MCMC procedure. *ESAIM P&S 8*, 115–131.

Kuhn, E. and M. Lavielle (2005). Maximum likelihood estimation in nonlinear mixed effects models. *Comput. Statist. Data Anal. 49*, 1020–1038.

Lavielle, M., H. Mesa, K. Chatel, and A. Vermeulen (2010). Mixture models and model mixtures with Monolix. In *Abstracts of the Annual Meeting of the Population Approach Group in Europe, Berlin.*

Lindstrom, M. and D. Bates (1990). Nonlinear mixed effects models for repeated measures data. *Biometrics 46*, 673–687.

McLachland, G. J. and D. Peel (2000). *Finite Mixture models.* New York: Wiley-Interscience.

Ng, S., G. McLachlan, K. Wang, L. Ben-Tovim, and S. Ng (2006). A mixture model with mixed effects components for clustering correlated gene-expression profiles. *Bioinformatics 22*, 1745–1752.

Proust, C. and H. Jacqmin-Gadda (2005). Estimation of linear mixed models with a mixture of distribution for the random effects. *Comput Methods Programs Biomed. 78*(2), 165–173.

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

Verbeke, G. and E. Lesaffre (1996). A linear mixed-effects model with heterogeneity in the random-effects population. *JASA 91*(433), 217–221.

Wang, X., A. Schumitzky, and D. Z. D'Argenio (2007). Non linear random effects mixture models : Maximum likelihood estimation via the EM algorithm. *Comput. Stat. Data Anal. 51*, 6614–6623.

Wang, X., A. Schumitzky, and D. Z. D'Argenio (2009). Population pharmacokinetic/pharmacodynamic mixture models via maximum a posteriori estimation. *Comput. Stat. Data Anal. 53*, 3907–3915.

Wei, G. and M. Tanner (1990). A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms. *Journal of the American Statistician Association 85*, 699–704.