Statistics and Computing manuscript

Estimation in Nonlinear Mixed–Effects Models Using Heavy–Tailed Distributions

Cristian Meza · Felipe Osorio · Rolando De la Cruz

Received: December 7, 2009 / Accepted: October 18, 2010

Abstract Nonlinear mixed–effects models are very useful to analyze repeated measures data and are used in a variety of applications. Normal distributions for random effects and residual errors are usually assumed, but such assumptions make inferences vulnerable to the presence of outliers. In this work, we introduce an extension of a normal nonlinear mixed–effects model considering a subclass of elliptical contoured distributions for both random effects and residual errors. This elliptical subclass, the scale mixtures of normal (SMN) distributions, includes heavy–tailed multivariate distributions, such as Student–*t*, the contaminated normal and slash, among others, and represents an interesting alternative to outliers accommodation maintaining the elegance and simplicity of the maximum likelihood theory. We propose an exact estimation procedure to obtain the maximum likelihood estimates of the fixed–effects and variance components, using a stochastic approximation of the EM algorithm. We compare the performance of the normal and the SMN models with two real data sets.

Keywords Mixed–effects model · Outliers · Scale mixtures of normal distributions · SAEM algorithm · Random effects

C. Meza
Departamento de Estadística, CIMFAV, Universidad de Valparaíso, Gran Bretaña 1091. Valparaíso, CHILE.
Tel.: +56-32-2508170
Fax: +56-32-2508322
E-mail: cristian.meza@uv.cl

F. Osorio
Departamento de Estadística, CIMFAV, Universidad de Valparaíso, Gran Bretaña 1091. Valparaíso, CHILE.
E-mail: felipe.osorio@uv.cl

R. De la Cruz
Departamento de Salud Pública, Escuela de Medicina, and Departamento de Estadística, Facultad de Matemáticas, Pontificia Universidad Católica de Chile,
Marcoleta 434, Casilla 114D, Santiago, CHILE.
E-mail: rolando@med.puc.cl

1 Introduction

Analysis of longitudinal data is an essential issue in biological, agricultural, environmental, and medical applications, and many methodologies have already been proposed in the framework of linear and nonlinear mixed–effects models to analyze such data (see Davidian and Giltinan, 1995; Vonesh and Chinchilli, 1997; Demidenko, 2004). In this work we focus on nonlinear mixed–effects models (NLMEMs), which have recently become very popular. NLMEMs are mixed–effects models in which the intraindividual model relating the response variable to time is nonlinear in the parameters. With the development of novel estimation procedures (Davidian and Giltinan, 2003; Kuhn and Lavielle, 2005), they are widely used in longitudinal studies. Their main field of application is in pharmacokinetic research, to analyze within–subject pharmacokinetic processes of absorption, distribution and elimination governing drug concentrations. They have also been widely applied for the modelling of growth traits in various agricultural and laboratory species, such as mice, chickens, cattle, pigs and trees.

Several methods for estimating the parameters in NLMEMs have been proposed (see Vonesh and Chinchilli, 1997; Pinheiro and Bates, 2000; Davidian and Giltinan, 2003, for a review). The estimation of NLMEMs raises specific problems, even if the random effects and the errors are normal, because the likelihood of the model typically cannot be expressed in closed form. Several approximations to the log-likelihood have been proposed. One is the first–order conditional estimation method (Beal and Sheiner, 1992). This method linearizes the nonlinear response function of the model with a first order Taylor series expansion about current estimates of the fixed–effects and the zero means of random effects. The resulting linearized model is then fitted by the ML technique. Lindstrom and Bates (1990) proposed a more accurate approximation to the nonlinear response function of the

model by expanding the nonlinear response function of the model at the current estimates of fixed–effects and random effects. It has been shown that Lindstrom and Bates' (1990) procedure is equivalent to solving a set of estimating equations where the estimating functions are approximate first derivatives of a Laplace approximation of the log–likelihood of the original model (Wolfinger, 1993; Vonesh, 1996; Wolfinger and Lin, 1997).

These likelihood approximations often perform well if the number of the intraindividual measurements is not small and the variability of random effects is not large. However, when some of the individuals have sparse data or the variability of the random effects is large, there are considerable errors in approximating the likelihood function via these approximations (Davidian and Giltinan, 1995; Pinheiro and Bates, 1995; Lindstrom and Bates, 1990). This has motivated the use of exact maximum likelihood methods, such as the EM algorithm. In particular, the Monte Carlo EM (MCEM) algorithm (Wei and Tanner, 1990), in which the E step is approximated by using simulated samples from the exact conditional distribution of the random effects given the observed data, has been used for estimation in mixed–effects models. An MCEM algorithm for exact maximum likelihood estimation of a class of NLMEMs is given by Walker (1996). Walker's (1996) procedure is computationally very intensive because it uses Monte Carlo simulation to evaluate the integrals in the E-step. In order to reduce the computational intensity and complexity of the Walker's algorithm, Wang (2007) implemented the MCEM algorithm using samples obtained via importance sampling from a mixture distribution chosen for its simplicity of form, facility for sampling and efficiency. A stochastic version of the EM algorithm (SAEM) using stochastic approximations, proposed by Delyon, et al. (1999), proved to be more computationally efficient than a classical MCEM algorithm thanks to recycling of simulations from one iteration to the next in the smoothing phase of the algorithm. Moreover, as pointed out by Jank (2006) the SAEM algorithm, unlike the MCEM, converges with a fixed and typically small simulation size. Recently, Kuhn and Lavielle (2005) showed that the SAEM algorithm is very efficient for computing the ML estimate in NLMEMs.

The aim of this paper is to propose an exact estimation procedure in an extension of normal NLMEMs considering a subclass of elliptical contoured distributions for both random effects and residual errors. The class of scale mixtures of normal distributions (Andrews and Mallows, 1974) has received much attention in recent years, particularly because they include distributions with longer-than-normal tails, such as the Student-*t*, the contaminated normal and slash, among others, and they present good properties that allow for accommodating extreme and outlying observations better than the models under normality assumption. In the context of mixed–effects models, several authors have proposed using heavy-tailed distributions to accommodate outliers. For instance, Welsh and Richardson (1997) make a review of procedures for robust estimation using multivariate symmetrical distributions. Pinheiro, et al. (2001), Lin and Lee (2006) and Staudenmayer, et al. (2009) studied robust approaches to estimation in which both random effects and errors have multivariate Student-*t* distributions, while Choy and Smith (1997), Rosa, et al. (2003) and Rosa, et al. (2004) discussed Markov chain Monte Carlo (MCMC) implementations considering a Bayesian formulation. However, few alternatives have been studied for outlier accommodation in the context of nonlinear mixed–effects models. To date, Yeap and Davidian (2001) is the only reference. They proposed a two-stage approach for robust estimation in NLMEMs when outliers are present within and between individuals. As in Yeap and Davidian (2001), our proposal allows for accommodation and identification of both types of outliers. A feature of the proposed model is that the computational aspect is simplified by considering a hierarchical version of the model. This also allows for using a stochastic version of the EM algorithm and the SAEM algorithm, extending previous works under Gaussian assumptions proposed by Kuhn and Lavielle (2005) and Lavielle and Meza (2007).

The article is organized as follows. In Section 2, we describe the family of heavy-tailed distributions and EM and SAEM algorithms used in this work. Section 3 presents the nonlinear mixed–effects model with heavy-tailed distributions. The maximum likelihood estimation procedure using the SAEM algorithm is described in Section 4. Estimation of the likelihood and standard errors is discussed in Section 5. In Section 6, the application of the proposed methodology is illustrated with real data. Finally, some conclusions are given and possible future work is discussed in Section 7.

2 Preliminaries

In this Section we review the subclass of elliptical contoured distributions, more specifically, the scale mixture of multivariate normal distributions. As well, we describe two stochastic versions of the EM algorithm for ML estimation. The first algorithm is the stochastic approximation expectation maximization (SAEM) algorithm and the second is a parameter expansion version of the SAEM algorithm.

2.1 Scale Mixture of Multivariate Normal Distributions

A random vector $\boldsymbol{Y}$ in $\mathbb{R}^m$ is presumed to have a distribution that is a scale mixture of multivariate normal distributions (Andrews and Mallows, 1974) with parameters, $\boldsymbol{\mu} \in \mathbb{R}^m$, $\boldsymbol{\Lambda}$ a $(m \times m)$ positive definite symmetric matrix, and H a

(unidimensional) probability distribution function, such that $H(0) = 0$, if its density function is

$$p(\boldsymbol{y}) = \int_0^\infty N_m(\boldsymbol{y}; \boldsymbol{\mu}, \kappa^{-1}\boldsymbol{\Lambda})\, \mathrm{d}H(\kappa)$$
$$= |2\pi\boldsymbol{\Lambda}|^{-1/2} \int_0^\infty \kappa^{m/2} \exp\{-\tfrac{1}{2}\kappa D^2\}\, \mathrm{d}H(\kappa). \quad (1)$$

where $N_m(\cdot; \boldsymbol{\mu}, \boldsymbol{\Lambda})$ is the m–dimensional normal density with parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Lambda}$, and $D^2 = (\boldsymbol{y} - \boldsymbol{\mu})^T \boldsymbol{\Lambda}^{-1}(\boldsymbol{y} - \boldsymbol{\mu})$. We shall use the notation $SMN_m(\boldsymbol{\mu}, \boldsymbol{\Lambda}; H)$ to indicate that $\boldsymbol{Y}$ has density (1). When the mixture distribution function H is degenerate, $SMN_m(\boldsymbol{\mu}, \boldsymbol{\Lambda}; H)$ is a normal distribution.

The *SMN* distributions (1) has a convenient stochastic representation

$$\boldsymbol{Y} \stackrel{\mathrm{d}}{=} \boldsymbol{\mu} + \kappa^{-1/2}\boldsymbol{Z}, \quad (2)$$

where $\boldsymbol{Z} \sim N_m(\boldsymbol{0}, \boldsymbol{\Lambda})$ is independent of the mixture variable $\kappa \sim H(\boldsymbol{\nu})$, and $\boldsymbol{\nu}$ is a scalar or vector valued parameter. Note that the expression (2) provides a useful tool for random number generation and for theoretic purposes. Another form that represents the distribution (2) is the following two-stage hierarchical representation

$$\boldsymbol{Y}|\kappa \sim N_m(\boldsymbol{\mu}, \kappa^{-1}\boldsymbol{\Lambda}),$$
$$\kappa \sim H(\boldsymbol{\nu}). \quad (3)$$

From (3), using the iterative law of expectation yields

$$\mathrm{E}(\boldsymbol{Y}) = \mathrm{E}(\mathrm{E}(\boldsymbol{Y}|\kappa)) = \boldsymbol{\mu},$$
$$\mathrm{cov}(\boldsymbol{Y}) = \mathrm{E}(\mathrm{cov}(\boldsymbol{Y}|\kappa)) + \mathrm{cov}(\mathrm{E}(\boldsymbol{Y}|\kappa)) = \mathrm{E}(\kappa^{-1})\boldsymbol{\Lambda}.$$

The class of *SMN* distributions provides a group of thick-tailed distributions that are often useful for robust inference. Some of these are: the multivariate Student–*t* distribution, multivariate slash distribution, multivariate contaminated normal distribution and multivariate exponential power distribution. This class of distributions have been applied in the context of regression models (see for instance Lange and Sinsheimer, 1993; Liu, 1996) as well as in linear mixed–effects models (Choy and Smith, 1997; Rosa, et al., 2003, 2004), obtaining robust estimates against outlying observations.

2.2 The Algorithms

2.2.1 The EM and SAEM Algorithms

The EM (Dempster, et al., 1977) is a popular iterative algorithm for calculating parameter estimates via ML in models with missing data or in models that can be formulated as such. In circumstances as those prevailing here, the maximization of log-likelihood function based on the observed data $\boldsymbol{Y}_{\mathrm{obs}}$, denoted by $\ell_{\mathrm{o}}(\boldsymbol{\theta}; \boldsymbol{Y}_{\mathrm{obs}}) = \log p(\boldsymbol{Y}_{\mathrm{obs}}; \boldsymbol{\theta})$, is difficult to perform. The EM algorithm proceeds in two steps:

i) E-Step: replace the observed likelihood by the likelihood of a complete data set and compute its conditional expectation

$$Q(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}}^{(k)}) = \mathrm{E}\{\ell_{\mathrm{c}}(\boldsymbol{\theta}; \boldsymbol{Y}_{\mathrm{com}})|\boldsymbol{Y}_{\mathrm{obs}}, \widehat{\boldsymbol{\theta}}^{(k)}\},$$

where $\widehat{\boldsymbol{\theta}}^{(k)}$ is the estimate of $\boldsymbol{\theta}$ at the kth iteration;

ii) M-Step: maximize it with respect to $\boldsymbol{\theta}$ obtaining $\widehat{\boldsymbol{\theta}}^{(k+1)}$.

Each iteration of the EM algorithm increases the likelihood function $\ell_{\mathrm{o}}(\boldsymbol{\theta}; \boldsymbol{Y}_{\mathrm{obs}})$ and the EM sequence $\{\widehat{\boldsymbol{\theta}}^{(k)}\}$ converges to a stationary point of the observed likelihood under mild regularity conditions (Wu, 1983; Vaida, 2005).

However, in some applications of the EM algorithm the E–step cannot be obtained analytically and has to be calculated using a simulation. 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 order to reduce the amount of required simulations compared to the MCEM algorithm, the SAEM algorithm proposed by Delyon, et al. (1999) replaces the E–step of EM by a stochastic approximation procedure, while the M–step is unchanged. The SAEM algorithm consists at each iteration, in successively simulating the random effects with the conditional distribution, and updating the unknown parameters of the model. Thus, at iteration k, SAEM is as follows:

E–Step:

- *Simulation–step*: draw $\boldsymbol{q}^{(k,\ell)}$, $(\ell = 1, \ldots, m)$ from the conditional distribution $p(\cdot|\boldsymbol{Y}_{\mathrm{obs}}, \widehat{\boldsymbol{\theta}}^{(k-1)})$.
- *Stochastic approximation*: update $Q(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}}^{(k)})$ according to

$$Q(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}}^{(k)}) = Q(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}}^{(k-1)}) + \delta_k \Big[\frac{1}{m}\sum_{\ell=1}^{m} \log p(\boldsymbol{Y}_{\mathrm{obs}}, \boldsymbol{q}^{(k,\ell)}; \boldsymbol{\theta}) - Q(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}}^{(k-1)})\Big],$$

where δ_k is a smoothing parameter, i.e. a decreasing sequence of positive numbers as presented by Kuhn and Lavielle (2004).

M–Step:

- *Maximization–step*: update $\widehat{\boldsymbol{\theta}}^{(k)}$ according to

$$\widehat{\boldsymbol{\theta}}^{(k+1)} = \arg\max_{\theta} Q(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}}^{(k)}).$$

In other words, SAEM performes a Monte Carlo E–step, like MCEM, but with a small and fixed Monte Carlo sample sizes ($m \leq 10$), which then is combined in a "smooth" way with the previous step of the algorithm (in calculation of the Q function and/or of the parameter estimates).

When the simulation step cannot be directly performed, Kuhn and Lavielle (2004) propose combining this algorithm

with a MCMC procedure: the sequence $\{\boldsymbol{q}^{(k)}\}$ is a Markov chain with transition kernels $\{\boldsymbol{\Pi}_{\widehat{\theta}^{(k)}}\}$. Then, the simulation step becomes:

- *Simulation-step*: draw $\boldsymbol{q}^{(k,\ell)}$, $(\ell = 1,\ldots,m)$, from the transition probability $\boldsymbol{\Pi}_{\widehat{\theta}^{(k)}}(\boldsymbol{q}^{(k-1)},\cdot)$.

As argued by Jank (2006) and recently by Meza, et al. (2009), an important issue is the analysis of the SAEM algorithm convergence. To run SAEM, the user must fix several constants as the number of total iterations and the number of iterations before starting the smoothing step of the SAEM algorithm. In order to define these constants, we can use a graphic approach based on the likelihood difference from one iteration to the next and monitor SAEM by estimating its progress towards $\widehat{\boldsymbol{\theta}}$ by using the property of increasing likelihood of the EM algorithm (see Meza, et al., 2009, for more details). Then, the number of iteration can be fixed and the smoothing step can be defined but it is important to note that this procedure implies to run the SAEM algorithm twice.

2.2.2 The PX–EM and PX–SAEM Algorithms

A major drawback of the EM algorithm is its slow convergence in some situations. To circumvent this limitation Liu, et al. (1998) proposed the so-called parameter-expanded EM (PX–EM) algorithm. Technically, the PX-EM algorithm expands the complete data model parameterized by $\boldsymbol{\theta}$, to a larger model parameterized by $\boldsymbol{\Theta}$ with $\boldsymbol{\Theta} = (\boldsymbol{\theta}, \boldsymbol{\alpha})$ where $\boldsymbol{\alpha}$ is a working parameter. To use the PX-EM algorithm two conditions must be satisfied: i) a many-to-one reduction function $R : \boldsymbol{\Theta} \to R(\boldsymbol{\Theta})$ that preserves the original observed data model, and ii) a value $\boldsymbol{\alpha}_0$ of $\boldsymbol{\alpha}$ that preserves the original complete data model (see Liu, et al., 1998, for more details).

Operationally, the PX-EM algorithm, like EM, consists of two steps. In particular, the *PX-E* step computes the conditional expectation of the log-likelihood

$$Q(\boldsymbol{\Theta}|\widehat{\boldsymbol{\Theta}}^{(k)}) = \mathrm{E}\{\ell_{\mathrm{c}}(\boldsymbol{\Theta};\boldsymbol{Y}_{\mathrm{com}})|\boldsymbol{Y}_{\mathrm{obs}},\widehat{\boldsymbol{\Theta}}^{(k)}\}, \tag{4}$$

where $\ell_{\mathrm{c}}(\boldsymbol{\Theta};\boldsymbol{Y}_{\mathrm{com}}) = \log p_X(\boldsymbol{Y}_{\mathrm{obs}},\boldsymbol{q};\boldsymbol{\Theta})$, is the expanded complete data log-likelihood and $\widehat{\boldsymbol{\Theta}}^{(k)} = (\widehat{\boldsymbol{\theta}}^{(k)},\boldsymbol{\alpha}_0)$. The *PX–M* step then maximizes (4) with respect to the expanded parameters

$$\widehat{\boldsymbol{\Theta}}^{(k+1)} = \arg\max_{\Theta} Q(\boldsymbol{\Theta}|\widehat{\boldsymbol{\Theta}}^{(k)})$$

and $\boldsymbol{\theta}$ is updated via $\widehat{\boldsymbol{\theta}}^{(k+1)} = R(\widehat{\boldsymbol{\Theta}}^{(k+1)})$.

In order to improve the convergence of the stochastic approximation version of EM, Lavielle and Meza (2007) propose adapting the PX-EM algorithm to the SAEM algorithm. Like the PX-EM, the PX-SAEM algorithm is a parameter expansion version of SAEM. Each iteration of PX-SAEM is broken down into three steps: the Simulation step, the Stochastic Approximation step of SAEM using the expanded model and the PX-M step of PX-EM. Thus, at iteration k, the algorithm can be described as

PX-E Step:

- *PX Simulation step*: draw $\boldsymbol{q}^{(k,\ell)}$, $(\ell = 1,\ldots,m)$ from the conditional distribution $p_X(\cdot|\boldsymbol{Y}_{\mathrm{obs}},\widehat{\boldsymbol{\Theta}}^{(k-1)} = (\widehat{\boldsymbol{\theta}}^{(k-1)},\boldsymbol{\alpha}_0))$.
- *PX Stochastic approximation*: update $Q(\boldsymbol{\Theta}|\widehat{\boldsymbol{\Theta}}^{(k)})$ according to

$$Q(\boldsymbol{\Theta}|\widehat{\boldsymbol{\Theta}}^{(k)}) = Q(\boldsymbol{\Theta}|\widehat{\boldsymbol{\Theta}}^{(k-1)}) + \delta_k\Big[\frac{1}{m}\sum_{\ell=1}^{m}\log p_X(\boldsymbol{Y}_{\mathrm{obs}},\boldsymbol{q}^{(k,\ell)};\boldsymbol{\Theta}) - Q(\boldsymbol{\Theta}|\widehat{\boldsymbol{\Theta}}^{(k-1)})\Big],$$

  where $\{\delta_k\}$ is a decreasing sequence of positive numbers.

PX-M Step:

- *PX Maximization-step*: update $\widehat{\boldsymbol{\Theta}}^{(k)}$ according to

$$\widehat{\boldsymbol{\Theta}}^{(k+1)} = \arg\max_{\Theta} Q(\boldsymbol{\Theta}|\widehat{\boldsymbol{\Theta}}^{(k)})$$

  and apply the reduction function to obtain $\widehat{\boldsymbol{\theta}}^{(k+1)} = R(\widehat{\boldsymbol{\Theta}}^{(k+1)})$ and $\widehat{\boldsymbol{\Theta}}^{(k+1)} = (\widehat{\boldsymbol{\theta}}^{(k+1)},\alpha_0)$.

As with the SAEM algorithm, when the simulation step cannot be performed directly, the simulation step becomes:

- *PX Simulation step*: draw $\boldsymbol{q}^{(k,\ell)}$, $(\ell = 1,\ldots,m)$, from the transition probability $\boldsymbol{\Pi}_{\widehat{\Theta}^{(k)}}(\boldsymbol{q}^{(k-1)},\cdot)$.

Lavielle and Meza (2007) show with numerical examples that the PX-SAEM algorithm substantially improves the speed of convergence toward the maximum likelihood estimate for linear and nonlinear mixed–effects models, in terms of reducing the number of iterations and the computing time.

3 Nonlinear mixed–effects models with heavy–tailed distributions

Let $\boldsymbol{Y}_i = (Y_{i1},\ldots,Y_{in_i})^T$ denote the response vector for subject i and $\boldsymbol{f}_i(\boldsymbol{z}_i,\boldsymbol{\phi}_i) = (f(\boldsymbol{z}_{i1},\boldsymbol{\phi}_i),\ldots,f(\boldsymbol{z}_{in_i},\boldsymbol{\phi}_i))^T$ a nonlinear vector–valued differentiable function of a vector–valued mixed–effects parameter $\boldsymbol{\phi}_i$ and a vector of covariates $\boldsymbol{z}_i$. A nonlinear mixed–effects model can then be expressed as

$$\boldsymbol{Y}_i = \boldsymbol{f}_i(\boldsymbol{z}_i,\boldsymbol{\phi}_i) + \boldsymbol{\epsilon}_i, \qquad i = 1,\ldots,n, \tag{5}$$

with the mixed–effects parameter $\boldsymbol{\phi}_i$ modeled as

$$\boldsymbol{\phi}_i = \boldsymbol{A}_i\boldsymbol{\beta} + \boldsymbol{B}_i\boldsymbol{b}_i, \qquad \text{with} \qquad \boldsymbol{b}_i \overset{\text{ind}}{\sim} N_q(\boldsymbol{0}, \boldsymbol{\Gamma}),$$

where $\boldsymbol{A}_i$ and $\boldsymbol{B}_i$ are known design matrices of size $r \times p$ and $r \times q$ that possibly depend on the subject and some covariable values, $\boldsymbol{\beta}$ is a p–dimensional vector of fixed–effects, $\boldsymbol{b}_i$ is a q–dimensional vector of random effects, $\boldsymbol{\Gamma} = \boldsymbol{\Gamma}(\boldsymbol{\gamma})$ is a positive definite matrix, structured by the vector $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_K)^T$, and $\boldsymbol{\epsilon}_i$ is an n_i–dimensional vector of within–subject errors. The $\boldsymbol{\epsilon}_i$ are assumed to be independent, with distribution $N_{n_i}(\boldsymbol{0}, \sigma^2\boldsymbol{I}_{n_i})$ independent of the $\boldsymbol{b}_i$. Most studies using NLMEMs assume normal distribution for random effects and within–subject errors, but such assumptions make the model sensitive to outliers. As discussed in previous works (see, for example Pinheiro, et al., 2001; Yeap and Davidian, 2001), an interesting feature of mixed–effects models is that outliers may occur either at the level of the within–subject error $\boldsymbol{\epsilon}_i$, called $\boldsymbol{\epsilon}$–outliers, or at the level of random effects $\boldsymbol{b}_i$, called $\boldsymbol{b}$–outliers. In the first case, some unusual within–subject values are observed, whereas in the second case some unusual subjects are observed.

In this paper, instead of normal assumptions in the nonlinear mixed–effects model (5) we replace the multivariate normal distributions with the scale mixture of multivariate normal distributions, which allows for outlier accommodation. Thus the model is expressed as

$$\begin{aligned} \boldsymbol{Y}_i|\boldsymbol{\phi}_i &\overset{\text{ind}}{\sim} SMN_{n_i}(\boldsymbol{f}_i(\boldsymbol{z}_i, \boldsymbol{\phi}_i), \sigma^2\boldsymbol{I}_{n_i}; H_1), \\ \boldsymbol{\phi}_i &\overset{\text{ind}}{\sim} SMN_r(\boldsymbol{A}_i\boldsymbol{\beta}, \boldsymbol{B}_i\boldsymbol{\Gamma}\boldsymbol{B}_i^T; H_2), \qquad i = 1, \ldots, n. \end{aligned} \tag{6}$$

With model (6), we are considering a robust estimation framework for $\boldsymbol{\beta}$, $\boldsymbol{\Gamma}$ and σ^2. Using the stochastic representation (2) we can rewrite the model (6) in a hierarchical form

$$\begin{aligned} \boldsymbol{Y}_i|\boldsymbol{\phi}_i, \kappa_i &\overset{\text{ind}}{\sim} N_{n_i}(\boldsymbol{f}_i(\boldsymbol{z}_i, \boldsymbol{\phi}_i), \kappa_i^{-1}\sigma^2\boldsymbol{I}_{n_i}), \\ \boldsymbol{\phi}_i|\tau_i &\overset{\text{ind}}{\sim} N_r(\boldsymbol{A}_i\boldsymbol{\beta}, \tau_i^{-1}\boldsymbol{B}_i\boldsymbol{\Gamma}\boldsymbol{B}_i^T), \\ \kappa_i &\overset{\text{ind}}{\sim} H_1(\boldsymbol{\nu}), \quad \tau_i \overset{\text{ind}}{\sim} H_2(\boldsymbol{\eta}), \qquad i = 1, \ldots, n, \end{aligned} \tag{7}$$

where $\boldsymbol{\nu}$ and $\boldsymbol{\eta}$ are a scalar or vector valued parameter of the mixture distribution, and κ_i and τ_i are assumed to be mutually independent.

Pinheiro, et al. (2001) propose a robust hierarchical linear mixed–effects model in which the random effects and the within-subject errors have multivariate Student-t distributions. They present several comparable and efficient EM-type algorithms for computing the ML estimates and illustrate the robustness of the model via a real example and some simulations. Although these distributional assumptions allow for the accommodation of outliers, in its formulation it is assumed that the mixture distributions for the two sources of variability in the model have the same shape and share the same parameters. Some authors (see, for instance Lin and Lee, 2007; Jara, et al., 2008) emphasize that this assumption may be very difficult to justify. In this work, we have adopted the approach suggested by Rosa, et al. (2004) and Staudenmayer, et al. (2009) (see also Jara, et al., 2008), where we assume that the mixture variables associated with errors and random effects are different. This allows a more direct comparison to the estimation procedure in the two stages proposed by Yeap and Davidian (2001). Moreover, the estimation procedure based on the EM algorithm has the advantage of allowing for the identification of outliers by examining the conditional distribution of mixture variables given the observed data. In the next section, we describe the maximum likelihood procedure in the model (6) via SAEM.

4 ML estimation in NLMEMs using a stochastic version of the EM algorithm

In this section we consider the maximum likelihood (ML) estimation of the parameters in the nonlinear mixed–effects model with heavy-tailed distributions using SAEM and PX-SAEM algorithms. In particular, we show the implementation of the algorithms to the case where the error term and the random effects follow a multivariate Student-t and multivariate slash distribution, respectively.

4.1 ML estimation using the SAEM Algorithm

In order to implement the SAEM algorithm for maximum likelihood estimation in the nonlinear mixed–effects model (7), we consider the vector of complete data as $\boldsymbol{Y}_{\text{com}} = (\boldsymbol{Y}^T, \boldsymbol{q}^T)^T$, with $\boldsymbol{Y} = (\boldsymbol{Y}_1^T, \ldots, \boldsymbol{Y}_n^T)^T$ and $\boldsymbol{q} = (\boldsymbol{\phi}^T, \boldsymbol{\kappa}^T, \boldsymbol{\tau}^T)^T$ where $\boldsymbol{\phi} = (\boldsymbol{\phi}_1^T, \ldots, \boldsymbol{\phi}_n^T)^T$, $\boldsymbol{\kappa} = (\kappa_1, \ldots, \kappa_n)^T$ and $\boldsymbol{\tau} = (\tau_1, \ldots, \tau_n)^T$ represents the incomplete data. The complete data is then taken to be the vector $\boldsymbol{Y}$. Then it is easy to derive the complete data log-likelihood for model (7) as

$$\begin{aligned} \ell_{\text{c}}(\boldsymbol{\theta}; \boldsymbol{Y}_{\text{com}}) = &-\frac{N}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\kappa_i||\boldsymbol{Y}_i - \boldsymbol{f}_i(\boldsymbol{z}_i, \boldsymbol{\phi}_i)||^2 \\ &- \frac{n}{2}\log|\boldsymbol{B}_i\boldsymbol{\Gamma}\boldsymbol{B}_i^T| - \frac{1}{2}\sum_{i=1}^{n}\tau_i(\boldsymbol{\phi}_i - \boldsymbol{A}_i\boldsymbol{\beta})^T(\boldsymbol{B}_i\boldsymbol{\Gamma}\boldsymbol{B}_i^T)^{-1}(\boldsymbol{\phi}_i - \boldsymbol{A}_i\boldsymbol{\beta}) \\ &+ \sum_{i=1}^{n}\log H_1(\kappa_i; \boldsymbol{\nu}) + \sum_{i=1}^{n}\log H_2(\tau_i; \boldsymbol{\eta}) + C, \end{aligned} \tag{8}$$

where $\boldsymbol{\theta} = (\boldsymbol{\beta}^T, \boldsymbol{\gamma}^T, \sigma^2, \boldsymbol{\nu}, \boldsymbol{\eta})^T$, $N = \sum_{i=1}^{n} n_i$ and C is a constant that is independent of the parameter vector $\boldsymbol{\theta}$. Using simple algebra, we obtain that the expected complete data

log-likelihood function is

$$
\begin{aligned}
Q(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}}^{(k)}) &= \mathrm{E}\{\ell_{\mathrm{c}}(\boldsymbol{\theta};\boldsymbol{Y}_{\mathrm{com}})|\boldsymbol{Y},\widehat{\boldsymbol{\theta}}^{(k)}\} \\
&= -\frac{N}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\mathrm{E}\{\kappa_i||\boldsymbol{Y}_i - \boldsymbol{f}_i(\boldsymbol{z}_i,\boldsymbol{\phi}_i)||^2|\boldsymbol{Y},\widehat{\boldsymbol{\theta}}^{(k)}\} \\
&- \frac{n}{2}\log|\boldsymbol{B}_i\boldsymbol{\Gamma}\boldsymbol{B}_i^T| - \frac{1}{2}\sum_{i=1}^{n}\mathrm{E}\{\tau_i(\boldsymbol{\phi}_i - \boldsymbol{A}_i\boldsymbol{\beta})^T(\boldsymbol{B}_i\boldsymbol{\Gamma}\boldsymbol{B}_i^T)^{-1} \\
&\times(\boldsymbol{\phi}_i - \boldsymbol{A}_i\boldsymbol{\beta})|\boldsymbol{Y},\widehat{\boldsymbol{\theta}}^{(k)}\} + \sum_{i=1}^{n}\mathrm{E}\{\log H_1(\kappa_i;\boldsymbol{\nu})|\boldsymbol{Y},\widehat{\boldsymbol{\theta}}^{(k)}\} \\
&+ \sum_{i=1}^{n}\mathrm{E}\{\log H_2(\tau_i;\boldsymbol{\eta})|\boldsymbol{Y},\widehat{\boldsymbol{\theta}}^{(k)}\} + C, \qquad (9)
\end{aligned}
$$

Let $\boldsymbol{U}(\boldsymbol{\gamma}) = \partial Q(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}}^{(k)})/\partial\boldsymbol{\gamma}$ and $\boldsymbol{H}(\boldsymbol{\gamma}) = -\partial^2 Q(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}}^{(k)})/\partial\boldsymbol{\gamma}\partial\boldsymbol{\gamma}^T$. Assuming this, it is possible to switch the integration and differentiation operators, and thus we can update the parameter estimates as

$$
\begin{aligned}
\widehat{\boldsymbol{\beta}}^{(k+1)} &= \Big(\sum_{i=1}^{n} S_{1,i}^{(k)}\boldsymbol{A}_i^T(\boldsymbol{B}_i\widehat{\boldsymbol{\Gamma}}^{(k)}\boldsymbol{B}_i^T)^{-1}\boldsymbol{A}_i\Big)^{-1} \\
&\times\sum_{i=1}^{n}\boldsymbol{A}_i^T(\boldsymbol{B}_i\widehat{\boldsymbol{\Gamma}}^{(k)}\boldsymbol{B}_i^T)^{-1}\boldsymbol{S}_{2,i}^{(k)}, \qquad (10)
\end{aligned}
$$

$$
\widehat{\boldsymbol{\gamma}}^{(k+1)} = \widehat{\boldsymbol{\gamma}}^{(k)} + \{\boldsymbol{H}(\widehat{\boldsymbol{\gamma}}^{(k)})\}^{-1}\boldsymbol{U}(\widehat{\boldsymbol{\gamma}}^{(k)}), \qquad (11)
$$

$$
\widehat{\sigma}^{2(k+1)} = \frac{1}{N}\sum_{i=1}^{n} S_{7,i}^{(k)} \qquad (12)
$$

$$
\widehat{\boldsymbol{\nu}}^{(k+1)} = \arg\max_{\nu}\sum_{i=1}^{n}\mathrm{E}\{\log H_1(\kappa_i;\boldsymbol{\nu})|\boldsymbol{Y},\widehat{\boldsymbol{\theta}}^{(k)}\} \qquad (13)
$$

$$
\widehat{\boldsymbol{\eta}}^{(k+1)} = \arg\max_{\eta}\sum_{i=1}^{n}\mathrm{E}\{\log H_2(\tau_i;\boldsymbol{\eta})|\boldsymbol{Y},\widehat{\boldsymbol{\theta}}^{(k)}\}, \qquad (14)
$$

where $S_{1,i}^{(k)} = \mathrm{E}\{\tau_i|\boldsymbol{Y}_i,\widehat{\boldsymbol{\theta}}^{(k)}\}$, $\boldsymbol{S}_{2,i}^{(k)} = \mathrm{E}\{\tau_i\boldsymbol{\phi}_i|\boldsymbol{Y},\widehat{\boldsymbol{\theta}}^{(k)}\}$, the gradient vector, $\boldsymbol{U}(\boldsymbol{\gamma})$ and the Hessian matrix, $\boldsymbol{H}(\boldsymbol{\gamma})$ have elements given by

$$
\begin{aligned}
\frac{\partial Q(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}}^{(k)})}{\partial\gamma_s} &= -\frac{n}{2}\,\mathrm{tr}\Big\{(\boldsymbol{B}_i\boldsymbol{\Gamma}\boldsymbol{B}_i^T)^{-1}\boldsymbol{B}_i\frac{\partial\boldsymbol{\Gamma}}{\partial\gamma_s}\boldsymbol{B}_i^T\Big\} + \frac{1}{2}\sum_{i=1}^{n}S_{3,i}^{(k)} \\
-\frac{\partial^2 Q(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}}^{(k)})}{\partial\gamma_s\partial\gamma_t} &= \frac{n}{2}\,\mathrm{tr}\Big\{(\boldsymbol{B}_i\boldsymbol{\Gamma}\boldsymbol{B}_i^T)^{-1}\boldsymbol{B}_i\frac{\partial^2\boldsymbol{\Gamma}}{\partial\gamma_s\partial\gamma_t}\boldsymbol{B}_i^T\Big\} \\
&-\frac{n}{2}\,\mathrm{tr}\Big\{(\boldsymbol{B}_i\boldsymbol{\Gamma}\boldsymbol{B}_i^T)^{-1}\boldsymbol{B}_i\frac{\partial\boldsymbol{\Gamma}}{\partial\gamma_s}\boldsymbol{B}_i^T(\boldsymbol{B}_i\boldsymbol{\Gamma}\boldsymbol{B}_i^T)^{-1}\boldsymbol{B}_i\frac{\partial\boldsymbol{\Gamma}}{\partial\gamma_t}\boldsymbol{B}_i^T\Big\} \\
&+\frac{1}{2}\sum_{i=1}^{n}\{S_{4,i}^{(k)} + S_{5,i}^{(k)} - S_{6,i}^{(k)}\},
\end{aligned}
$$

respectively, for $s,t = 1,\ldots,K$, which must be evaluated at $\boldsymbol{\gamma} = \widehat{\boldsymbol{\gamma}}^{(k)}$ with

$$
\begin{aligned}
S_{3,i}^{(k)} &= \mathrm{E}\Big\{\tau_i(\boldsymbol{\phi}_i - \boldsymbol{A}_i\widehat{\boldsymbol{\beta}}^{(k)})^T(\boldsymbol{B}_i\boldsymbol{\Gamma}\boldsymbol{B}_i^T)^{-1}\boldsymbol{B}_i\frac{\partial\boldsymbol{\Gamma}}{\partial\gamma_s}\boldsymbol{B}_i^T(\boldsymbol{B}_i\boldsymbol{\Gamma}\boldsymbol{B}_i^T)^{-1} \\
&\quad\times(\boldsymbol{\phi}_i - \boldsymbol{A}_i\widehat{\boldsymbol{\beta}}^{(k)})|\boldsymbol{Y},\widehat{\boldsymbol{\theta}}^{(k)}\Big\} \\
S_{4,i}^{(k)} &= \mathrm{E}\Big\{\tau_i(\boldsymbol{\phi}_i - \boldsymbol{A}_i\widehat{\boldsymbol{\beta}}^{(k)})^T(\boldsymbol{B}_i\boldsymbol{\Gamma}\boldsymbol{B}_i^T)^{-1}\boldsymbol{B}_i\frac{\partial\boldsymbol{\Gamma}}{\partial\gamma_s}\boldsymbol{B}_i^T(\boldsymbol{B}_i\boldsymbol{\Gamma}\boldsymbol{B}_i^T)^{-1} \\
&\quad\times\boldsymbol{B}_i^T\frac{\partial\boldsymbol{\Gamma}}{\partial\gamma_t}\boldsymbol{B}_i^T(\boldsymbol{B}_i\boldsymbol{\Gamma}\boldsymbol{B}_i^T)^{-1}(\boldsymbol{\phi}_i - \boldsymbol{A}_i\widehat{\boldsymbol{\beta}}^{(k)})|\boldsymbol{Y},\widehat{\boldsymbol{\theta}}^{(k)}\Big\} \\
S_{5,i}^{(k)} &= \mathrm{E}\Big\{\tau_i(\boldsymbol{\phi}_i - \boldsymbol{A}_i\widehat{\boldsymbol{\beta}}^{(k)})^T(\boldsymbol{B}_i\boldsymbol{\Gamma}\boldsymbol{B}_i^T)^{-1}\boldsymbol{B}_i\frac{\partial\boldsymbol{\Gamma}}{\partial\gamma_t}\boldsymbol{B}_i^T(\boldsymbol{B}_i\boldsymbol{\Gamma}\boldsymbol{B}_i^T)^{-1} \\
&\quad\times\boldsymbol{B}_i^T\frac{\partial\boldsymbol{\Gamma}}{\partial\gamma_s}\boldsymbol{B}_i^T(\boldsymbol{B}_i\boldsymbol{\Gamma}\boldsymbol{B}_i^T)^{-1}(\boldsymbol{\phi}_i - \boldsymbol{A}_i\widehat{\boldsymbol{\beta}}^{(k)})|\boldsymbol{Y},\widehat{\boldsymbol{\theta}}^{(k)}\Big\} \\
S_{6,i}^{(k)} &= \mathrm{E}\Big\{\tau_i(\boldsymbol{\phi}_i - \boldsymbol{A}_i\widehat{\boldsymbol{\beta}}^{(k)})^T(\boldsymbol{B}_i\boldsymbol{\Gamma}\boldsymbol{B}_i^T)^{-1}\boldsymbol{B}_i\frac{\partial^2\boldsymbol{\Gamma}}{\partial\gamma_s\partial\gamma_t}\boldsymbol{B}_i^T \\
&\quad\times(\boldsymbol{B}_i\boldsymbol{\Gamma}\boldsymbol{B}_i^T)^{-1}(\boldsymbol{\phi}_i - \boldsymbol{A}_i\widehat{\boldsymbol{\beta}}^{(k)})|\boldsymbol{Y},\widehat{\boldsymbol{\theta}}^{(k)}\Big\} \\
S_{7,i}^{(k)} &= \mathrm{E}\{\kappa_i\|\boldsymbol{Y}_i - \boldsymbol{f}_i(\boldsymbol{z}_i,\boldsymbol{\phi}_i)\|^2|\boldsymbol{Y},\widehat{\boldsymbol{\theta}}^{(k)}\}.
\end{aligned}
$$

In the *simulation step* of the algorithm, at iteration k, we need to draw $\boldsymbol{q}_i^{(k)} = (\boldsymbol{\phi}_i^{(k)},\kappa_i^{(k)},\tau_i^{(k)})$ from the conditional distribution $p(\cdot|\boldsymbol{Y}_i,\widehat{\boldsymbol{\theta}}^{(k-1)})$. We propose to use the Gibbs sampler algorithm to simulate from this conditional distribution. At iteration k, the iteration s of the Gibbs sampler starts with $(\boldsymbol{\phi}_i^{(k,s)},\kappa_i^{(k,s)},\tau_i^{(k,s)})$ and makes the transition to $(\boldsymbol{\phi}_i^{(k,s+1)},\kappa_i^{(k,s+1)},\tau_i^{(k,s+1)})$ via the following scheme (for $i = 1,\ldots,n$)

- Sample $\boldsymbol{\phi}_i^{(k,s+1)}$ from $[\boldsymbol{\phi}_i|\boldsymbol{Y}_i,\kappa_i^{(k,s)},\tau_i^{(k,s)},\widehat{\boldsymbol{\theta}}^{(k)}]$
- Sample $\kappa_i^{(k,s+1)}$ from $[\kappa_i|\boldsymbol{Y}_i,\boldsymbol{\phi}_i^{(k,s+1)},\tau_i^{(k,s)},\widehat{\boldsymbol{\theta}}^{(k)}]$
- Sample $\tau_i^{(k,s+1)}$ from $[\tau_i|\boldsymbol{Y}_i,\boldsymbol{\phi}_i^{(k,s+1)},\kappa_i^{(k,s+1)},\widehat{\boldsymbol{\theta}}^{(k)}]$.

In the case of nonlinear mixed–effects model the full conditional for $\boldsymbol{\phi}_i$ is not available analytically. This suggests to carry out a Metropolis–Hastings (M–H) algorithm within each Gibbs step. In practice, at iteration k, to approximate the distribution of $\boldsymbol{\phi}|\boldsymbol{Y}_i,\kappa_i^{(k,s)},\tau_i^{(k,s)},\widehat{\boldsymbol{\theta}}^{(k)}$, three transition kernels were successively used: first, the conditional distribution of $\boldsymbol{\phi}_i$ given τ_i (for $i = 1,\cdots,n$) at iteration k, which is the Gaussian distribution $q_{\theta_{i,k}^{\star}}^{(1)} \sim N(\boldsymbol{A}_i\boldsymbol{\beta}^{(k)},\tau_i^{(k,s)^{-1}}\widehat{\boldsymbol{\Gamma}}^{(k)})$; second, the multidimensional random walk $q_{\theta_{i,k}^{\star}}^{(2)} \sim N(\boldsymbol{\phi}_{i,p}^{(k,s+1)}, \rho^2\tau_i^{(k,s)^{-1}}\widehat{\boldsymbol{\Gamma}}^{(k)})$, where ρ is a constant and $\boldsymbol{\phi}_{i,p}^{(k+1,s)}$ represents the simulation of the random effects at global iteration k and for the p-th iteration of M-H, with $p = 1,\cdots,m_2$; finally $q_{\theta_{i,k}^{\star}}^{(3)}$ is a succession of d unidimensional Gaussian random walks: each component of $\boldsymbol{\phi}$ are successively updated. In summary, the simulation–step at iteration k consists in running first m_1 iterations of the M–H algorithm with pro-

posal $q^{(1)}_{\theta^\star_{i,k}}$, then m_2 iterations with proposal $q^{(2)}_{\theta^\star_{i,k}}$ and finally m_3 iterations with proposal $q^{(3)}_{\theta^\star_{i,k}}$.

The use of different kernels permits to increase the convergence and to favour all kind of transition. The values of parameters ρ, m_1, m_2 and m_3 involved in this simulation procedure have to be chosen by the user. Few iterations of the M–H algorithm at each simulation–step are enough to converge and in practice, m_1, m_2 and m_3 are less than 10. The choice of ρ is more delicate as they play an important role in the random walk. The values of ρ are linked to the acceptance rate of the M–H algorithm, so they must be chosen to approximate the 'optimal acceptance rate'. Some theoretical and empirical results (see Roberts, et al., 1997; Roberts and Rosenthal, 2001) have shown that in high dimensions, under various regularity conditions, it is optimal to choose the scale parameter of the random walk such that the asymptotic acceptance rate of the M–H algorithm is approximately 0.234.

Once we draw a sequence $(\boldsymbol{\phi}_i^{(k,\ell)}, \kappa_i^{(k,\ell)}, \tau_i^{(k,\ell)})$, $\ell = 1,\ldots,m$, at iteration k, the conditional expectations $S^{(k)}_{h,i}$, $(h = 1,\ldots,7)$ in (10)-(12) are replaced with the following stochastic approximations:

$$S^{(k)}_{1,i} = S^{(k-1)}_{1,i} + \delta_k\Big(\frac{1}{m}\sum_{\ell=1}^{m}\tau_i^{(k,\ell)} - S^{(k-1)}_{1,i}\Big) \tag{15}$$

$$\boldsymbol{S}^{(k)}_{2,i} = \boldsymbol{S}^{(k-1)}_{2,i} + \delta_k\Big(\frac{1}{m}\sum_{\ell=1}^{m}\tau_i^{(k,\ell)}\boldsymbol{\phi}_i^{(k,\ell)} - \boldsymbol{S}^{(k-1)}_{2,i}\Big) \tag{16}$$

$$\begin{aligned} S^{(k)}_{3,i} = S^{(k-1)}_{3,i} &+ \delta_k\Big(\frac{1}{m}\sum_{\ell=1}^{m}\tau_i^{(k,\ell)}\boldsymbol{b}_i^{(k,\ell)^T}\boldsymbol{B}_i^T(\boldsymbol{B}_i\widehat{\boldsymbol{\Gamma}}^{(k)}\boldsymbol{B}_i^T)^{-1} \\ &\times \boldsymbol{B}_i\frac{\partial\boldsymbol{\Gamma}}{\partial\gamma_s}\boldsymbol{B}_i^T(\boldsymbol{B}_i\widehat{\boldsymbol{\Gamma}}^{(k)}\boldsymbol{B}_i^T)^{-1}\boldsymbol{B}_i\boldsymbol{b}_i^{(k,\ell)} - S^{(k-1)}_{3,i}\Big) \end{aligned} \tag{17}$$

$$\begin{aligned} S^{(k)}_{4,i} = S^{(k-1)}_{4,i} &+ \delta_k\Big(\frac{1}{m}\sum_{\ell=1}^{m}\tau_i^{(k,\ell)}\boldsymbol{b}_i^{(k,\ell)^T}\boldsymbol{B}_i^T(\boldsymbol{B}_i\widehat{\boldsymbol{\Gamma}}^{(k)}\boldsymbol{B}_i^T)^{-1} \\ &\times \boldsymbol{B}_i\frac{\partial\boldsymbol{\Gamma}}{\partial\gamma_s}\boldsymbol{B}_i^T(\boldsymbol{B}_i\widehat{\boldsymbol{\Gamma}}^{(k)}\boldsymbol{B}_i^T)^{-1}\boldsymbol{B}_i\frac{\partial\boldsymbol{\Gamma}}{\partial\gamma_t}\boldsymbol{B}_i^T(\boldsymbol{B}_i\widehat{\boldsymbol{\Gamma}}^{(k)}\boldsymbol{B}_i^T)^{-1} \\ &\times \boldsymbol{B}_i\boldsymbol{b}_i^{(k,\ell)} - S^{(k-1)}_{4,i}\Big) \end{aligned} \tag{18}$$

$$\begin{aligned} S^{(k)}_{5,i} = S^{(k-1)}_{5,i} &+ \delta_k\Big(\frac{1}{m}\sum_{\ell=1}^{m}\tau_i^{(k,\ell)}\boldsymbol{b}_i^{(k,\ell)^T}\boldsymbol{B}_i^T(\boldsymbol{B}_i\widehat{\boldsymbol{\Gamma}}^{(k)}\boldsymbol{B}_i^T)^{-1} \\ &\times \boldsymbol{B}_i\frac{\partial\boldsymbol{\Gamma}}{\partial\gamma_t}\boldsymbol{B}_i^T(\boldsymbol{B}_i\widehat{\boldsymbol{\Gamma}}^{(k)}\boldsymbol{B}_i^T)^{-1}\boldsymbol{B}_i\frac{\partial\boldsymbol{\Gamma}}{\partial\gamma_s}\boldsymbol{B}_i^T(\boldsymbol{B}_i\widehat{\boldsymbol{\Gamma}}^{(k)}\boldsymbol{B}_i^T)^{-1} \\ &\times \boldsymbol{B}_i\boldsymbol{b}_i^{(k,\ell)} - S^{(k-1)}_{5,i}\Big) \end{aligned} \tag{19}$$

$$\begin{aligned} S^{(k)}_{6,i} = S^{(k-1)}_{6,i} &+ \delta_k\Big(\frac{1}{m}\sum_{\ell=1}^{m}\tau_i^{(k,\ell)}\boldsymbol{b}_i^{(k,\ell)^T}\boldsymbol{B}_i^T(\boldsymbol{B}_i\widehat{\boldsymbol{\Gamma}}^{(k)}\boldsymbol{B}_i^T)^{-1} \\ &\times \boldsymbol{B}_i\frac{\partial^2\boldsymbol{\Gamma}}{\partial\gamma_s\partial\gamma_t}\boldsymbol{B}_i^T(\boldsymbol{B}_i\widehat{\boldsymbol{\Gamma}}^{(k)}\boldsymbol{B}_i^T)^{-1}\boldsymbol{B}_i\boldsymbol{b}_i^{(k,\ell)} - S^{(k-1)}_{6,i}\Big) \end{aligned} \tag{20}$$

$$\begin{aligned} S^{(k)}_{7,i} = S^{(k-1)}_{7,i} &+ \delta_k\Big(\frac{1}{m}\sum_{\ell=1}^{m}\kappa_i^{(k,\ell)}\|\boldsymbol{Y}_i - \boldsymbol{f}_i(\boldsymbol{z}_i,\boldsymbol{\phi}_i^{(k,\ell)})\|^2 \\ &- S^{(k-1)}_{7,i}\Big), \end{aligned} \tag{21}$$

where δ_k is the smoothing parameter, chosen in order to ensure the convergence of the algorithm and $\boldsymbol{B}_i\boldsymbol{b}_i^{(k,\ell)} = \boldsymbol{\phi}_i^{(k,\ell)} - \boldsymbol{A}_i\widehat{\boldsymbol{\beta}}^{(k)}$. Conditional expectations involved in equations (13) and (14) must be replaced with their respective stochastic approximations (see Section 4.2 for one particular case). Kuhn and Lavielle (2005) suggest that a small value of m (smaller than 10 in practice) is enough to ensure very satisfactory results.

It is important to note that, although in this kind of model, the $S^{(k)}_{h,i}$, with $h = 1,\ldots,7$, are no longer minimal sufficient statistics unlike in Kuhn and Lavielle (2005) since the complete likelihood does not belong to the exponential family, they have the same role. Indeed, they resume the relevant information of the complete likelihood for the *maximization step* and they permit to simplify the *stochastic approximation step* of SAEM. Furthermore, an important remark is that when the degrees of freedom ν and η are known, the complete likelihood $\ell_{\rm c}(\boldsymbol{\theta};\boldsymbol{Y}_{\rm com})$ belongs to the exponential family and we recover the MCMC–SAEM algorithm proposed by Kuhn and Lavielle (2005). On the other hand, we should note that Gu and Kong (1998) and recently Cai (2010) extended the Stochastic Approximation type algorithm and showed its convergence properties in more general context that the exponential family.

An important special case occurs when $\boldsymbol{\Gamma}$ is symmetric and positive definite matrix and the subject-specific parameters are modeled as $\boldsymbol{\phi}_i = \boldsymbol{A}_i\boldsymbol{\beta} + \boldsymbol{b}_i$. Then, we can replace (11) as follows

$$\widehat{\boldsymbol{\Gamma}}^{(k+1)} = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{S}_i^{(k)},$$

$$\boldsymbol{S}_i^{(k)} = \mathrm{E}\{\tau_i(\boldsymbol{\phi}_i - \boldsymbol{A}_i\widehat{\boldsymbol{\beta}}^{(k)})(\boldsymbol{\phi}_i - \boldsymbol{A}_i\widehat{\boldsymbol{\beta}}^{(k)})^T|\boldsymbol{Y},\widehat{\boldsymbol{\theta}}^{(k)}\}$$

where as before, the conditional expectation $\boldsymbol{S}_i^{(k)}$ is replaced by the stochastic approximation

$$\begin{aligned} \boldsymbol{S}_i^{(k)} = \boldsymbol{S}_i^{(k-1)} &+ \delta_k\Big(\frac{1}{m}\sum_{\ell=1}^{m}\tau_i^{(k,\ell)}(\boldsymbol{\phi}_i^{(k,\ell)} - \boldsymbol{A}_i\widehat{\boldsymbol{\beta}}^{(k)}) \\ &\times (\boldsymbol{\phi}_i^{(k,\ell)} - \boldsymbol{A}_i\widehat{\boldsymbol{\beta}}^{(k)})^T - \boldsymbol{S}_i^{(k-1)}\Big) \end{aligned}$$

which is equivalent to the approximations defined in (17)-(20).

In the next Section we describe the implementation of SAEM and PX-SAEM algorithms in one particular case, where the conditional response vector is assumed to follow a multivariate Student-*t* distribution and the vector of random effects follows a multivariate slash distribution.

4.2 The Student-*t*/Slash Nonlinear Mixed–Effects Model

Two distributions belonging to the *SMN* class are the slash (Rogers and Tuckey, 1972) and Student–*t* (Lange, et al.,

1989) distributions, which have been suggested as alternatives in robust modeling. The slash and Student–*t* distributions are obtained from (3) assuming that: $\kappa \sim Beta(\nu, 1)$ and $\kappa \sim Gamma(\nu/2, \nu/2)$, respectively, with densities

$$p(\kappa) = \nu\kappa^{\nu-1},\ 0 < \kappa < 1,\ \nu > 0, \quad \text{and}$$
$$p(\kappa) = \frac{(\nu/2)^{\nu/2}\kappa^{\nu/2-1}}{\Gamma(\nu/2)} \exp(-\tfrac{1}{2}\nu\kappa),\ \kappa, \nu > 0.$$

In each case, the parameter $\nu > 0$ corresponds to the degrees of freedom. We can also appreciate that it is possible to recover the Gaussian model when we consider $\nu \to \infty$.

The Student–*t*/slash nonlinear mixed–effects model is obtained from (7) assuming a Student–*t* for the conditional response vector and a slash distribution for the random effects. Thus it can be expressed as

$$\begin{aligned} \boldsymbol{Y}_i|\boldsymbol{\phi}_i, \kappa_i &\overset{\text{ind}}{\sim} N_{n_i}(\boldsymbol{f}_i(\boldsymbol{z}_i, \boldsymbol{\phi}_i), \kappa_i^{-1}\sigma^2\boldsymbol{I}_{n_i}), \\ \boldsymbol{\phi}_i|\tau_i &\overset{\text{ind}}{\sim} N_r(\boldsymbol{A}_i\boldsymbol{\beta}, \tau_i^{-1}\boldsymbol{\Gamma}), \\ \kappa_i &\overset{\text{ind}}{\sim} Gamma\left(\frac{\nu}{2}, \frac{\nu}{2}\right), \quad \tau_i \overset{\text{ind}}{\sim} Beta(\eta, 1), \end{aligned} \tag{22}$$

for $i = 1, \ldots, n$. The log–likelihood function for the model defined in (22) is given by

$$\begin{aligned} \ell_{\text{c}}(\boldsymbol{\theta}; \boldsymbol{Y}_{\text{com}}) &= -\frac{N}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\kappa_i\|\boldsymbol{Y}_i - \boldsymbol{f}_i(\boldsymbol{z}_i, \boldsymbol{\phi}_i)\|^2 \\ &- \frac{n}{2}\log|\boldsymbol{\Gamma}| - \frac{1}{2}\operatorname{tr}\boldsymbol{\Gamma}^{-1}\sum_{i=1}^{n}\tau_i(\boldsymbol{\phi}_i - \boldsymbol{A}_i\boldsymbol{\beta})(\boldsymbol{\phi}_i - \boldsymbol{A}_i\boldsymbol{\beta})^T \\ &+ n\Big\{\frac{\nu}{2}\log\Big(\frac{\nu}{2}\Big) - \log\Gamma\Big(\frac{\nu}{2}\Big)\Big\} + \frac{\nu}{2}\sum_{i=1}^{n}(\log\kappa_i - \kappa_i) - \sum_{i=1}^{n}\log\kappa_i \\ &+ n\log\eta + (\eta-1)\sum_{i=1}^{n}\log\tau_i + C. \end{aligned} \tag{23}$$

Dropping out all the terms that are not functions of $\boldsymbol{\theta}$, the relevant part of the expected complete data log-likelihood function for the Student-*t*/slash nonlinear mixed–effects model can be written as

$$\begin{aligned} Q(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}}^{(k)}) &= -\frac{N}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\mathrm{E}\{\kappa_i\|\boldsymbol{Y}_i - \boldsymbol{f}_i(\boldsymbol{z}_i, \boldsymbol{\phi}_i)\|^2|\boldsymbol{Y}, \widehat{\boldsymbol{\theta}}^{(k)}\} \\ &- \frac{n}{2}\log|\boldsymbol{\Gamma}| - \frac{1}{2}\operatorname{tr}\boldsymbol{\Gamma}^{-1}\sum_{i=1}^{n}\mathrm{E}\{\tau_i(\boldsymbol{\phi}_i - \boldsymbol{A}_i\boldsymbol{\beta})(\boldsymbol{\phi}_i - \boldsymbol{A}_i\boldsymbol{\beta})^T|\boldsymbol{Y}, \widehat{\boldsymbol{\theta}}^{(k)}\} \\ &+ n\Big\{\frac{\nu}{2}\log\Big(\frac{\nu}{2}\Big) - \log\Gamma\Big(\frac{\nu}{2}\Big)\Big\} \\ &+ \frac{\nu}{2}\sum_{i=1}^{n}(\mathrm{E}\{\log\kappa_i|\boldsymbol{Y}, \widehat{\boldsymbol{\theta}}^{(k)}\} - \mathrm{E}\{\kappa_i|\boldsymbol{Y}, \widehat{\boldsymbol{\theta}}^{(k)}\}) \\ &+ n\log\eta + \eta\sum_{i=1}^{n}\mathrm{E}\{\log\tau_i|\boldsymbol{Y}, \widehat{\boldsymbol{\theta}}^{(k)}\}. \end{aligned} \tag{24}$$

The solutions for $\boldsymbol{\beta}$, $\boldsymbol{\Gamma}$ and σ^2 in M step of the algorithm are given by equations (10)-(12), respectively, as defined in Section 4.1. The solutions for ν and η defined in equations (13) and (14) for this particular case, must satisfy:

$$\log\Big(\frac{\nu}{2}\Big) - \psi\Big(\frac{\nu}{2}\Big) + \frac{1}{n}\sum_{i=1}^{n}(S_{8,i}^{(k)} - S_{9,i}^{(k)}) = 0 \tag{25}$$

$$\widehat{\eta}^{(k+1)} = -n\Big/\sum_{i=1}^{n}S_{10,i}^{(k)}, \tag{26}$$

where $S_{8,i}^{(k)} = \mathrm{E}\{\log\kappa_i|\boldsymbol{Y}, \widehat{\boldsymbol{\theta}}^{(k)}\}$, $S_{9,i}^{(k)} = \mathrm{E}\{\kappa_i|\boldsymbol{Y}, \widehat{\boldsymbol{\theta}}^{(k)}\}$, $S_{10,i}^{(k)} = \mathrm{E}\{\log\tau_i|\boldsymbol{Y}, \widehat{\boldsymbol{\theta}}^{(k)}\}$, and $\psi(z) = \mathrm{d}\log\Gamma(z)/\mathrm{d}\,z$, denotes the digamma function. Note that $\widehat{\nu}^{(k+1)}$ can be obtained by solving (25) using a one-dimensional Newton-Raphson algorithm.

In the *simulation step*, the full conditional distributions to implement the Gibbs sampler algorithm are given by (for $i = 1, \ldots, n$)

$$\boldsymbol{\phi}_i|\boldsymbol{Y}_i, \kappa_i, \tau_i; \boldsymbol{\theta} \propto \exp\Big\{-\frac{1}{2}(\kappa_i D_{\epsilon_i}^2 + \tau_i D_{\phi_i}^2)\Big\} \tag{27}$$

$$\kappa_i|\boldsymbol{Y}_i, \boldsymbol{\phi}_i, \tau_i; \boldsymbol{\theta} \sim Gamma\Big(\frac{n_i + \nu}{2}, \frac{D_{\epsilon_i}^2 + \nu}{2}\Big), \tag{28}$$

$$\tau_i|\boldsymbol{Y}_i, \boldsymbol{\phi}_i, \kappa_i; \boldsymbol{\theta} \sim Truncated\ Gamma\Big(\frac{r}{2} + \eta, \frac{D_{\phi_i}^2}{2}, t\Big) \tag{29}$$

where

$$\begin{aligned} D_{\epsilon_i}^2 &= \frac{1}{\sigma^2}||\boldsymbol{Y}_i - \boldsymbol{f}_i(\boldsymbol{z_i}, \boldsymbol{\phi}_i)||^2, \text{ and} \\ D_{\phi_i}^2 &= (\boldsymbol{\phi}_i - \boldsymbol{A}_i\boldsymbol{\beta})^T\boldsymbol{\Gamma}^{-1}(\boldsymbol{\phi}_i - \boldsymbol{A}_i\boldsymbol{\beta}). \end{aligned} \tag{30}$$

Here truncated Gamma variables (29) have a right truncation point at $t = 1$. Simulation of the independent right truncated gamma variables is performed using the accept-reject algorithm proposed by Philippe (1997). If the full conditional (27) is not available analytically, we then employ the Metropolis-Hastings algorithm to draw from it. As described in Section 4.1, the conditional expectations defined in (10)–(12) are replaced with the stochastic approximations (15)–(21) and additionally the conditional expectations in (25) and (26) must be replaced with the following stochastic approximations

$$S_{8,i}^{(k)} = S_{8,i}^{(k-1)} + \delta_k\Big(\frac{1}{m}\sum_{\ell=1}^{m}\log\kappa_i^{(k,\ell)} - S_{8,i}^{(k-1)}\Big), \tag{31}$$

$$S_{9,i}^{(k)} = S_{9,i}^{(k-1)} + \delta_k\Big(\frac{1}{m}\sum_{\ell=1}^{m}\kappa_i^{(k,\ell)} - S_{9,i}^{(k-1)}\Big), \tag{32}$$

$$S_{10,i}^{(k)} = S_{10,i}^{(k-1)} + \delta_k\Big(\frac{1}{m}\sum_{\ell=1}^{m}\log\tau_i^{(k,\ell)} - S_{10,i}^{(k-1)}\Big). \tag{33}$$

4.2.1 The PX-SAEM algorithm in the Student–t/Slash Nonlinear Mixed–Effects Model

EM-type procedures, including the SAEM algorithm, can present slow convergence in some situations. To deal with this problem, we decided to apply a parameter expansion version of SAEM (PX-SAEM) to the model defined in (7). In this specific model, we focus on the distribution of error and as discussed by Kent, et al. (1994), Meng and van Dyk (1997) and Liu, et al. (1998), we can modified the standard EM procedure for the multivariate Student-*t* distribution by

following an augmentation scheme including a working parameter α, thus obtaining the following model:

$$\begin{aligned}\boldsymbol{Y}_i|\boldsymbol{\phi}_i,\kappa_i &\overset{\text{ind}}{\sim} N_{n_i}(\boldsymbol{f}_i(\boldsymbol{z}_i,\boldsymbol{\phi}_i),\kappa_i^{-1}\sigma_\star^2\boldsymbol{I}_{n_i}),\\ \boldsymbol{\phi}_i|\tau_i &\overset{\text{ind}}{\sim} N_r(\boldsymbol{A}_i\boldsymbol{\beta}_\star,\tau_i^{-1}\boldsymbol{\Gamma}_\star),\\ \frac{\kappa_i}{\alpha} &\overset{\text{ind}}{\sim} Gamma(\nu/2,\nu/2), \quad \tau_i \overset{\text{ind}}{\sim} Beta(\eta,1),\end{aligned} \tag{34}$$

for $i = 1,\ldots,n$. As mentioned before, two conditions must be satisfied to use the PX-SAEM algorithm: i) a many-to-one reduction function R that preserves the original observed-data model and ii) a value α_0 of α that preserves the original complete-data model. In this context, $\boldsymbol{\Theta} = (\boldsymbol{\beta}_\star, \boldsymbol{\Gamma}_\star, \sigma_\star^2, \alpha)$, the reduction function is $R(\boldsymbol{\Theta}) = (\boldsymbol{\beta}_\star, \boldsymbol{\Gamma}_\star, \frac{\sigma_\star^2}{\alpha})$ and $\alpha_0 = 1$.

As usual, in this problem the observed-data model does not depend on working parameter α. Then we can set α at α_0 at the beginning of each simulation step. At iteration k of the PX–SAEM algorithm, the stochastic approximations are updated as in (15)–(18) and the unique difference between the standard SAEM and the PX–SAEM lies in the maximization step where parameters are updated by calculating $\widehat{\boldsymbol{\Theta}}^{(k+1)}$, which maximizes the conditional expectation of the complete likelihood, $Q(\boldsymbol{\Theta}|\widehat{\boldsymbol{\Theta}}^{(k)})$. The reduction function is then applied to obtain $\widehat{\boldsymbol{\theta}}^{(k+1)} = R(\widehat{\boldsymbol{\Theta}}^{(k+1)})$ and $\widehat{\boldsymbol{\Theta}}^{(k+1)} = (\widehat{\boldsymbol{\theta}}^{(k+1)}, \alpha_0)$.

For this specific NLMEM, the application of the reduction function in this PX version of SAEM leads to adjustment in the estimate of the variance error, producing minor changes in the maximization step of the previous SAEM algorithm. Indeed, we only need to replace the maximization step of σ^2 with:

$$\widehat{\sigma}^{2(k+1)} = \frac{\widehat{\sigma}_\star^{2(k+1)}}{\widehat{\alpha}^{(k+1)}},$$

where

$$\widehat{\sigma}_\star^{2(k+1)} = \frac{1}{N}\sum_{i=1}^{n} S_{7,i}^{(k)} \qquad \text{and} \qquad \widehat{\alpha}^{(k+1)} = \frac{1}{n}\sum_{i=1}^{n} S_{9,i}^{(k)}.$$

Here, $N = \sum_{i=1}^{n} n_i$ and, $S_{7,i}^{(k)}$ and $S_{9,i}^{(k)}$ are defined in (21) and (32), respectively.

Remark 1: Several authors have proposed to address the estimation of the shape parameters for the mixture variables in nonlinear regression models (see, for example Lange and Sinsheimer, 1993; Jamshidian, 1999). For linear mixed–effects models under Student-*t* errors, Welsh and Richardson (1997), Pinheiro, et al. (2001) and Lin (2008) have described procedures for the estimation of the degrees of freedom. Such works has focused on efficient algorithms based on the EM algorithm and its variants. Although the approach of these works has been quite successful in practice, in particular, for the Student–*t* distribution, Fernández and Steel (1999) have warned of potential problems that may arise in the estimation of degrees of freedom and they notice that in this case the function of log-likelihood is unbounded and indeed corresponds to a nonregular estimation problem. In addition, Lucas (1997) note that the parameter estimates are robust only against extreme observation in the case that the degrees of freedom are kept fixed. Thus, one alternative is to assume that the parameters associated with the mixture variables κ_i and τ_i are known.

Remark 2: Another alternative for selecting the parameters associated with the mixture variables is to follow the strategy proposed by Lange, et al. (1989) (see also, Staudenmayer, et al., 2009). In this work we follow the approach of Lange, et al. (1989) when it is appropriate. In particular, in the Student-*t*/Slash nonlinear mixed–effects model to estimate the degrees of freedom ν and η, the procedure is based on addressing the estimation of ν and η as a model selection problem. That is, for a grid of acceptable values of ν and η, we perform the estimation of $\boldsymbol{\theta} = (\boldsymbol{\beta}^T, \text{vech}^T\,\boldsymbol{\Gamma}, \sigma^2)^T$ using the SAEM algorithm or PX-SAEM described above. Next, we estimate the likelihood function of the model for the observed data as shown in equation (35). Finally, we choose ν and η such that the likelihood function in (35) is maximized. Remember that, as discussed previously, when the degree of freedom are known, the complete likelihood belongs to the exponential family and our procedure is equivalent to the MCMC–SAEM algorithm proposed by Kuhn and Lavielle (2005).

5 Estimation of the likelihood and standard errors

5.1 Estimation of the likelihood

Based on model (6), we obtain that the likelihood function of model for the observed data $\ell_{\text{o}}(\boldsymbol{\theta}; \boldsymbol{Y}_{\text{obs}})$ is defined as

$$\ell_{\text{o}}(\boldsymbol{\theta}; \boldsymbol{Y}_{\text{obs}}) = \int p(\boldsymbol{y}, \boldsymbol{\phi}; \boldsymbol{\theta})\,\mathrm{d}\boldsymbol{\phi} = \int p(\boldsymbol{y}|\boldsymbol{\phi}; \boldsymbol{\theta}) p(\boldsymbol{\phi}; \boldsymbol{\theta})\,\mathrm{d}\boldsymbol{\phi}, \tag{35}$$

where $p(\boldsymbol{y}|\boldsymbol{\phi}; \boldsymbol{\theta})$ and $p(\boldsymbol{\phi}; \boldsymbol{\theta})$ are the densities of the NLMEM given in (6). Following Meza, et al. (2009), we can compute this integral using an importance sampling scheme for any continuous distribution $\widetilde{p}$. Equation (35) can be represented as

$$\ell_{\text{o}}(\boldsymbol{\theta}; \boldsymbol{Y}_{\text{obs}}) = \int p(\boldsymbol{y}|\boldsymbol{\phi}; \boldsymbol{\theta}) \frac{p(\boldsymbol{\phi}; \boldsymbol{\theta})}{\widetilde{p}(\boldsymbol{\phi}; \boldsymbol{\theta})} \widetilde{p}(\boldsymbol{\phi}; \boldsymbol{\theta})\,\mathrm{d}\boldsymbol{\phi},$$

so that $\ell_{\text{o}}(\boldsymbol{\theta}; \boldsymbol{Y}_{\text{obs}})$ can be estimated by

$$\widehat{\ell}_{\text{o}}(\boldsymbol{\theta}; \boldsymbol{Y}_{\text{obs}}) = \frac{1}{M}\sum_{m=1}^{M} p(\boldsymbol{y}|\boldsymbol{\phi}_m; \boldsymbol{\theta}) \frac{p(\boldsymbol{\phi}_m; \boldsymbol{\theta})}{\widetilde{p}(\boldsymbol{\phi}_m; \boldsymbol{\theta})} \tag{36}$$

where $\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_M$ are drawn from $\widetilde{p}(\boldsymbol{\phi}; \boldsymbol{\theta})$. An efficient choice for $\widetilde{p}(\boldsymbol{\phi}; \boldsymbol{\theta})$ consists of the conditional distribution of $\boldsymbol{\phi}$ given the data $\boldsymbol{Y}$. In practice, during the last few iterations of SAEM, i.e. at convergence, we estimate empirically the conditional mean $\mathrm{E}(\boldsymbol{\phi}|\boldsymbol{Y}; \widehat{\theta})$ and the conditional variance $\mathrm{Var}(\boldsymbol{\phi}|\boldsymbol{Y}; \widehat{\theta})$ from the MCMC procedure. Then, we choose as sampling distribution $\widetilde{p}$ the distribution of random effects $\boldsymbol{\phi}$ defined in (6) with those moments.

5.2 Standard error approximation

Consider $\boldsymbol{U}_{\mathrm{o}}(\boldsymbol{\theta}) = \partial \ell_{\mathrm{o}}(\boldsymbol{\theta}; \boldsymbol{Y}_{\mathrm{obs}})/\partial \boldsymbol{\theta}$, and note that the score function for the observed data satisfies,

$$\boldsymbol{U}_{\mathrm{o}}(\boldsymbol{\theta}) = \mathrm{E}_{\theta}\{\boldsymbol{U}_{\mathrm{c}}(\boldsymbol{\theta}; \boldsymbol{Y}_{\mathrm{com}})|\boldsymbol{Y}_{\mathrm{obs}}\}, \tag{37}$$

with $\boldsymbol{U}_{\mathrm{c}}(\boldsymbol{\theta}; \boldsymbol{Y}_{\mathrm{com}}) = \partial \ell_{\mathrm{c}}(\boldsymbol{\theta}; \boldsymbol{Y}_{\mathrm{com}})/\partial \boldsymbol{\theta}$, the score function for the complete data. Louis (1982) showed that the observed information matrix $\boldsymbol{I}_{\mathrm{o}}(\boldsymbol{\theta}) = -\partial^2 \ell_{\mathrm{o}}(\boldsymbol{\theta}; \boldsymbol{Y}_{\mathrm{obs}})/\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T$, can be calculated from the information supplied by the EM algorithm, through the formula

$$\begin{aligned}\boldsymbol{I}_{\mathrm{o}}(\boldsymbol{\theta}) = &\mathrm{E}_{\theta}\{\boldsymbol{H}_{\mathrm{c}}(\boldsymbol{\theta}; \boldsymbol{Y}_{\mathrm{com}})|\boldsymbol{Y}_{\mathrm{obs}}\} + \boldsymbol{U}_{\mathrm{o}}(\boldsymbol{\theta})\boldsymbol{U}_{\mathrm{o}}^T(\boldsymbol{\theta}) \\ &- \mathrm{E}_{\theta}\{\boldsymbol{U}_{\mathrm{c}}(\boldsymbol{\theta}; \boldsymbol{Y}_{\mathrm{com}})\boldsymbol{U}_{\mathrm{c}}^T(\boldsymbol{\theta}; \boldsymbol{Y}_{\mathrm{com}})|\boldsymbol{Y}_{\mathrm{obs}}\}, \end{aligned} \tag{38}$$

where $\boldsymbol{H}_{\mathrm{c}}(\boldsymbol{\theta}; \boldsymbol{Y}_{\mathrm{com}}) = -\partial^2 \ell_{\mathrm{c}}(\boldsymbol{\theta}; \boldsymbol{Y}_{\mathrm{com}})/\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T$.

Thus, we can to approximate the expectations in (37) and (38) using stochastic approximations

$$\boldsymbol{g}_k = \boldsymbol{g}_{k-1} + \delta_k\Big(\frac{1}{m}\sum_{\ell=1}^{m} \boldsymbol{U}_{\mathrm{c}}(\widehat{\boldsymbol{\theta}}^{(k)}; \boldsymbol{q}^{(k,\ell)}, \boldsymbol{Y}_{\mathrm{obs}}) - \boldsymbol{g}_{k-1}\Big),$$

$$\begin{aligned}\boldsymbol{J}_k = \boldsymbol{J}_{k-1} + \delta_k\Big(&\frac{1}{m}\sum_{\ell=1}^{m}\Big\{\boldsymbol{H}_{\mathrm{c}}(\widehat{\boldsymbol{\theta}}^{(k)}; \boldsymbol{q}^{(k,\ell)}, \boldsymbol{Y}_{\mathrm{obs}}) \\ &- \boldsymbol{U}_{\mathrm{c}}(\widehat{\boldsymbol{\theta}}^{(k)}; \boldsymbol{q}^{(k,\ell)}, \boldsymbol{Y}_{\mathrm{obs}})\boldsymbol{U}_{\mathrm{c}}^T(\widehat{\boldsymbol{\theta}}^{(k)}; \boldsymbol{q}^{(k,\ell)}, \boldsymbol{Y}_{\mathrm{obs}})\Big\} - \boldsymbol{J}_{k-1}\Big)\end{aligned}$$

where $\boldsymbol{q}^{(k,\ell)}, (\ell = 1, \ldots, m)$ are drawn from the conditional distribution $p(\cdot|\boldsymbol{Y}_{\mathrm{obs}}, \widehat{\boldsymbol{\theta}}^{(k-1)})$. Finally, the observed information matrix can be approximated as

$$\boldsymbol{H}_k = \boldsymbol{J}_k - \boldsymbol{g}_k \boldsymbol{g}_k^T,$$

at convergence $\boldsymbol{H}_k \to \boldsymbol{I}_{\mathrm{o}}(\widehat{\boldsymbol{\theta}})$, so that $\boldsymbol{H}_k^{-1}$ is an estimate of the covariance matrix of the parameter estimates (see Zhu and Lee, 2002; Cai, 2010).

6 Applications

6.1 Example 1: Theophylline kinetic data

Figure 1 shows data of a pharmacokinetic study, analyzed by Davidian and Giltinan (1995) and Pinheiro and Bates (1995), among others authors. In this experiment, we are interested in examining blood concentrations in twelve subjects after an oral dose of the anti-asthmatic agent theophylline was administered. Each patient received a dose D_i of the drug at time 0 and the jth serum concentration Y_{ij} of the ith patient is measured at times x_{ij}, with $i = 1, \ldots, 12$ and $j = 1, \ldots, 10$. The underlying pharmacokinetic processes are modeled by the following nonlinear mixed–effects model

$$Y_{ij} = \frac{D_i k_{ai}}{V_i(k_{ai} - Cl_i/V_i))}\left\{\exp\left(-\frac{Cl_i}{V_i}x_{ij}\right) - \exp(-k_{ai}x_{ij})\right\} + \epsilon_{ij}, \tag{39}$$

which is a first-order one-compartment model. In (39) k_{ai} is the absorption rate constants of subject i, V_i is the volume required to account for all drugs in the body of subject i and Cl_i is the clearance of subject i representing the volume of blood from which the drug is eliminated per unit time.

Because each of these parameters in (39) is necessarily positive to be meaningful, we will assume that the pharmacokinetic parameters for each subject $\boldsymbol{\phi}_i = (\log k_{ai}, \log V_i, \log Cl_i)^T$ are given by

$$k_{ai} = \exp(\beta_1 + b_{i1}),\ V_i = \exp(\beta_2 + b_{i2}),\ Cl_i = \exp(\beta_3 + b_{i3}).$$

First of all, we will consider an ML estimation of parameters of model (39) assuming that both the error terms and random effects follow a multivariate normal distribution, specifying the variance–covariance matrix of the random effects as: (i) unstructured $\boldsymbol{\Gamma}$, and (ii) $\boldsymbol{\Gamma} = \mathrm{diag}(\gamma_1, \gamma_2, \gamma_3)$.

This data set, widely studied, is implemented in the free software `MONOLIX` (Lavielle, 2005) including all the information and the parameters required for running SAEM for the model (39) under Gaussian assumptions. Using `MONOLIX` with these preset settings, we fitted model (39) using SAEM (including the simulated annealing step described in Lavielle, 2005) with $m = 10$ and with the following smoothing parameter

$$\delta_k = \begin{cases} 1, & \text{for } 1 \le k \le 300, \\ \frac{1}{k-300}, & \text{for } 301 \le k \le 500. \end{cases} \tag{40}$$

In order to select the variance–covariance matrix structure for the random effects which best describes the data, we use the Bayesian Information Criterion (BIC). The BIC for the fitted models (i) and (ii) are 362.78 and 346.82, respectively. Therefore, for this data set the best fit correspond to the diagonal variance–covariance matrix for the random effects. Henceforth all analysis will based considering a diagonal matrix for the random effects.

Figure 1 shows individual fits that reveal that there are some poorly fitted individuals. The estimated Mahalanobis distances $D^2_{\epsilon_i}$ and $D^2_{\phi_i}$, defined in (30), provide useful diagnostic statistics for identifying subjects with outlying observations (see, for example, Copt and Victoria-Feser, 2006). Note that under the Gaussian model (5) it is possible to show

that $D^2_{\epsilon_i} \sim \chi^2_{n_i}$ and $D^2_{\phi_i} \sim \chi^2_r$. Since $\mathrm{E}(D^2_{\epsilon_i}) = n_i$ and $\mathrm{E}(D^2_{\phi_i}) = r$, Pinheiro, et al. (2001) proposed the quantities $\widehat{D}^2_{\epsilon_i}/n_i$ and $\widehat{D}^2_{\phi_i}/r$ to identify outlying observations. These statistics have expected values equal to one. Figure 2 presents these diagnostic statistics, which suggests that individuals 5 and 9 are possibly ϵ–outlier and ϕ–outlier, respectively. Moreover, the Q–Q plot in Figure 3 confirms our suspicion of outliers in the random effects.

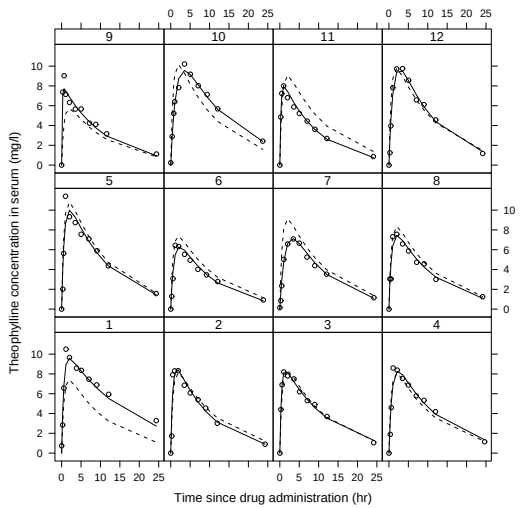

Fig. 1 Theophylline concentrations (in mg/L) for 12 patients and individual fits obtained with the SAEM algorithm under Gaussian assumptions on both random effects and the error term. Circles are observations. The solid lines are the individual fits using the individual parameters with the individual covariates and the dotted lines are the individual fits using the population parameters with the individual covariates.

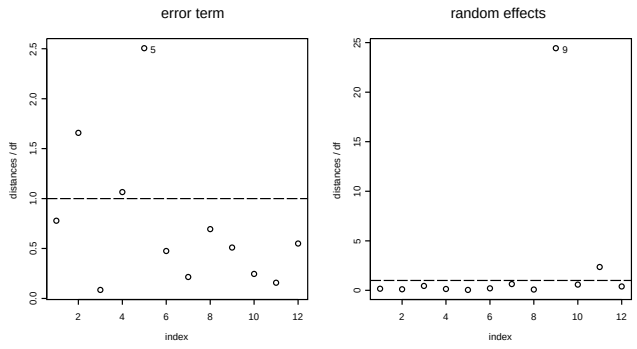

Fig. 2 Theophylline concentrations: Mahalanobis distances for residual vector and random effects for the Gaussian model.

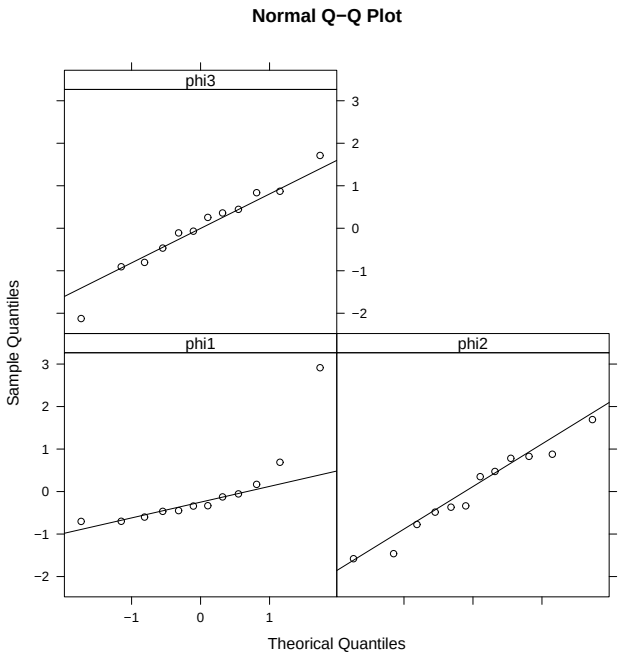

Fig. 3 Theophylline concentrations: Q-Q plots of random effects for the Gaussian model fitted with the SAEM algorithm.

It is well known that outlying observations may affect the estimation of the parameters under assumptions of normality. With the goal of accommodating outlying observations, we analyze this data set using the nonlinear mixed–effects model under heavy-tailed distributions. In the implementation of the SAEM algorithm, we use the same smoothing parameter described in (40) for the Gaussian model, starting the smoothing phase at iteration 300 and stopped the algorithm at iteration 500 with $m = 10$. Results of the estimation by the maximum likelihood method obtained using the SAEM algorithm for $\boldsymbol{\theta} = (\boldsymbol{\beta}^T, \mathrm{diag}^T\, \boldsymbol{\Gamma}, \sigma^2)^T$ considering several alternatives to the mixture distributions H_1 and H_2 are presented in Table 1 (standard errors in parenthesis). We focused on all possible combinations considering the Student–*t* and slash distributions. The estimation of the parameters for the mixture distributions H_1 and H_2 were chosen following the strategy proposed by Lange, et al. (1989). We note that for all models considered, the procedure selects small values for the parameters associated with mixture variables. These parameters act as tuning constants in robust estimation methods. In our case, we see that these choices provide adequate protection against outliers.

Note that the Gaussian model (5) is a particular case of model (6) where $\kappa_i = 1$, $\tau_i = 1$ and the mixing distributions H_1 and H_2 are degenerate. The likelihood ratio statistics for the Student–*t*, slash/Student–*t*, slash and Student–*t*/slash models against the normal model corresponding to $LR = 15.58$, $LR = 13.48$, $LR = 24.16$ and $LR = 24.50$, respectively. These results show that the NLMEM with heavy-

Table 1 Parameter estimates for Theophylline data under several fitted models.

Parameter $(\nu;\eta)$	Gaussian	Student–*t* (3;4)	slash/Student–*t* (1.5;3)
β_1	1.59 (0.32)	1.43 (0.44)	1.42 (0.41)
β_2	0.46 (0.02)	0.47 (0.03)	0.47 (0.04)
β_3	0.04 (0.00)	0.04 (0.00)	0.04 (0.00)
γ_1	0.44 (0.25)	0.36 (0.11)	0.36 (0.16)
γ_2	0.02 (0.01)	0.01 (0.02)	0.01 (0.02)
γ_3	0.03 (0.03)	0.03 (0.05)	0.03 (0.05)
σ^2	0.21 (0.06)	0.32 (0.10)	0.32 (0.10)
$\widehat{\ell}_o(\widehat{\boldsymbol{\theta}})$	-172.33	-164.54	-165.59
Parameter $(\nu;\eta)$	slash (1.25;1.5)	Student–*t*/slash (3.5;1.5)	Student–*t*/slash[a] (3.5;1.5)
β_1	1.44 (0.23)	1.46 (0.40)	1.48 (0.35)
β_2	0.47 (0.02)	0.47 (0.03)	0.47 (0.02)
β_3	0.04 (0.00)	0.04 (0.00)	0.04 (0.00)
γ_1	0.22 (0.09)	0.25 (0.13)	0.27 (0.18)
γ_2	0.01 (0.01)	0.01 (0.01)	0.01 (0.01)
γ_3	0.03 (0.03)	0.03 (0.05)	0.04 (0.03)
σ^2	0.21 (0.06)	0.32 (0.10)	0.30 (0.09)
$\widehat{\ell}_o(\widehat{\boldsymbol{\theta}})$	-160.25	-160.08	

[a] estimates using MCMC methods in `WinBUGS`

tailed distributions fits the data better than the normal NLMEM. Among the NLMEM with heavy-tailed distributions considered, we chose the Student–*t*/Sash NLMEM with parameters $\nu = 3.5$ and $\eta = 1.5$ for additionally analysis.

Table 2 Theophylline data: Estimated weights for the Student-*t*/slash model.

Subject	1	2	3	4	5	6
Residual errors (κ_i)	0.69	**0.37**	1.98	0.52	**0.26**	0.98
Random effects (τ_i)	**0.29**	0.69	0.73	0.70	0.71	0.68
Subject	7	8	9	10	11	12
Residual errors (κ_i)	1.54	0.75	0.93	1.37	1.87	0.85
Random effects (τ_i)	0.62	0.71	**0.29**	0.66	**0.42**	0.67

In order to detect outliers, the estimates of κ_i and τ_i are shown in Table 2. As we expected, patients 5 and 9 present small values of κ_i and τ_i, respectively, suggesting outlying observations at within-patient and random effect levels, that is, patient 5 is a $\boldsymbol{\epsilon}$–outlier and patient 9 is a $\boldsymbol{\phi}$–outlier. This is consistent with the diagnostic plot included in Figure 2. As well, Figure 1 suggests that patient 9 has an unusual growth pattern, and reveals that this patient has an unusually high Theophylline concentration serum at the time of the fourth measurement. As well, Table 2 reveals that patient 2 is a $\boldsymbol{\epsilon}$–outlier and patients 1 and 11 are $\boldsymbol{\phi}$–outlier, which cannot be concluded from Figure 2.

For the Student–*t*/slash NLMEM, we compared the SAEM algorithm with the Monte Carlo EM (MCEM) algorithm. To perform MCEM, following McCulloch (1997), we used a predeterminated sequence of Monte Carlo sample size M values: $M = 50, 200, 5000$ for iterations 1–19, 20–39 and 40 and over. The algorithm was stopped when iterates appeared to fluctuate randomly. Both algorithms were implemented in `MATLAB` 7 and run on an Intel Core 2 Quad PC computer at 2.40 Ghz and 8 GB of RAM. The results are summarized in Table 3, which shows the number of chains, sample sizes and user time to obtain the Student–*t*/slash maximum likelihood estimates. As expected, MCEM estimates are in good agreement with the SAEM algorithm, but although there are several strategies to improve the performance of the MCEM algorithm (see, for example, Wang, 2007), it is evident from the results that the MCEM algorithm requires considerable computational effort. As expected, the SAEM algorithm is much more efficient. Also, for comparison we fit the Student–*t*/slash NLMEM using the Bayesian software package `WinBUGS` (Spiegelhalter, et al., 1999). The values given for `WinBUGS` are means and standard deviations of the marginal posterior distributions obtained using the following non–informative priors $\boldsymbol{\beta} \sim N_3(\mathbf{0}, 10^6\boldsymbol{I}_3)$, $\sigma, \gamma_1^\star, \gamma_2^\star, \gamma_3^\star \sim U(0, 10^3)$, where $\gamma_l^\star = \sqrt{\gamma_l}$, $l = 1, 2, 3$. As we can see in Table 1 the estimates of all the parameters agree with each other with those obtained using the SAEM algorithm.

Table 3 Number of chains, sample sizes and user time to obtain the Student–*t*/slash maximum likelihood estimates for theophylline data.

Algorithm	Chains (M)	Total iteration	Time (sec.)
MCEM	50, 200, 5000[a]	300	3588
SAEM	10	500	15

[a]:For iterations 1–19, 20–39 and 40 and over

In order to increase the convergence speed of standard SAEM, we also applied the PX–SAEM algorithm to this data set using this specific *SMN* nonlinear mixed–effects model, by performing PX for the first 50 iterations, and standard SAEM afterwards. Results are presented in Figure (4) showing that the PX-SAEM algorithm increases the convergence speed of standard SAEM for this problem. Indeed, for the parameter σ^2, the PX-SAEM algorithm reached convergence four times as quickly as the standard algorithm.

6.2 Example 2: Guinea pigs data

The guinea pig data was discussed in Johansen (1984) and studied using nonlinear mixed effects models by several authors (see for example Lindstrom and Bates, 1990 and Lee and Xu, 2004). The experiment is as follows: 50 tissue sample were taken from the intestine of each of eight guinea pigs. Then for each guinea pigs, five tissue samples were assigned randomly to each of then different concentrations of *B*–methyl–glucoside. The uptake volume was measured in micromoles per milligram of fresh tissue per 2 minutes and

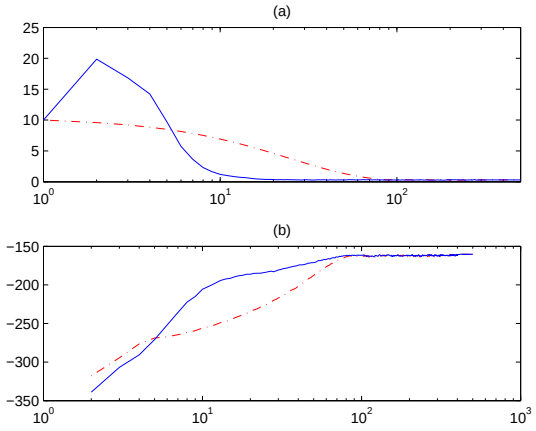

Fig. 4 (a): Sequences $\{\sigma^{2(k)}\}$ using SAEM (dotted line) and PX-SAEM (solid line) for the Student–*t*/slash model. A logarithmic scale is used for the number of iterations on the x-axis. (b) The observed log-likelihood sequences $\{\widehat{\ell}_o(\widehat{\boldsymbol{\theta}}^{(k)})\}$ obtained with SAEM (dotted line) and PX-SAEM (solide line) for the Student–*t*/slash model. A logarithmic scale is used for the number of iterations on the x-axis.

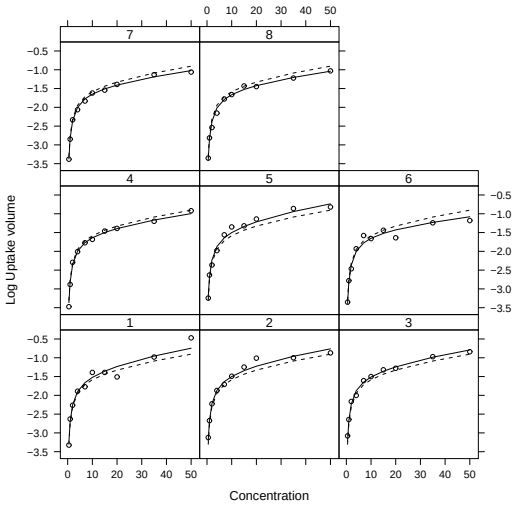

Fig. 5 Uptake volume of *B*–methyl–glucoside (in micromoles/mg) in tissue samples from 8 guinea pigs and individual fits obtained with the SAEM algorithm under Gaussian assumptions on both random effects and the error term. Circles are observations. The solid lines are the individual fits using the individual parameters and the dotted lines are the individual fits using the population parameters.

only the means of the five tissue samples at each concentration for each animal are reported. The data are plotted in Figure 5. Lindstrom and Bates (1990) proposed to model this dataset with the following nonlinear mixed–effects model:

$$\log(y_{ij}) = \log\left(\frac{\exp(\beta_0 + b_{i0})x_{ij}}{\exp(\beta_1 + b_{i1}) + x_{ij}} + \exp(\beta_2 + b_{i2})x_{ij}\right) + \epsilon_{ij} \tag{41}$$

where y_{ij} is the jth uptake volume for individual i, x_{ij} is the jth concentration level for individual i, $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)^T$ is a vector of fixed population effects, ϵ_{ij} is the error term and $\boldsymbol{b}_i = (b_{i0}, b_{i1}, b_{i2})^T$ is a vector of individual random effects, with $i = 1 \cdots, 8$ and $j = 1, \cdots, 10$.

Like the first example, in a first time we will consider an ML estimation of parameters of model (41) assuming that both the error terms and random effects follow a multivariate normal distribution, considering that the variance–covariance matrix of the random effects is $\boldsymbol{\Gamma} = \begin{pmatrix} \gamma_{11} & \gamma_{12} & \gamma_{13} \\ \gamma_{21} & \gamma_{22} & \gamma_{23} \\ \gamma_{31} & \gamma_{32} & \gamma_{33} \end{pmatrix}$.

Using MONOLIX, we fitted model (41) under Gaussian assumptions, with $m = 15$ and the following smoothing parameter

$$\delta_k = \begin{cases} 1, & \text{for } 1 \leq k \leq 500, \\ \frac{1}{k-500}, & \text{for } 501 \leq k \leq 1000. \end{cases} \tag{42}$$

Like the first example, Figure 5 reveals that the first animal is poorly fitted with this model which is confirmed by the estimated Mahalanobis distances (Figure 6). These distances suggest that the individual 1 is a possibly a ϵ-outlier.

In order to accommodate outlying observations, we analyzed this dataset using the NLMEM under heavy–tailed

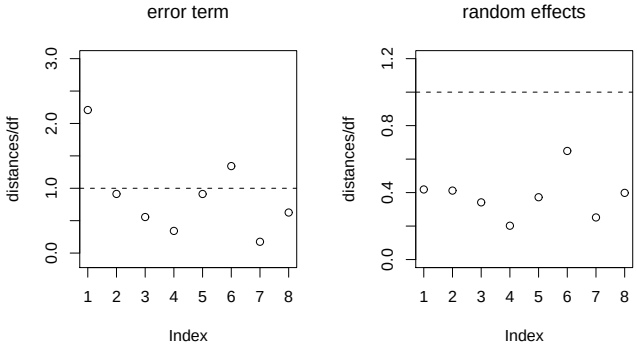

Fig. 6 Guinea pigs data: Mahalanobis distances for residual vector and random effects for the Gaussian model.

distributions. We applied SAEM with the same smoothing parameter described in (42) and we considered several combinations of Gaussian, Student-*t* and slash distributions for random effects and error terms, following the same strategy described in the first example, to apply and to choose the model that best fits the data. Among the NLMEM with heavy–tailed distributions considered, we chose the NLMEM

Table 4 Parameters estimates for Guinea pigs data

Parameters	Gaussian	Gaussian/Student–*t* $\eta = 2.1$	Gaussian/Student–*t*[a] $\eta = 2.1$
β_0	-1.60 (0.05)	-1.62 (0.03)	-1.60 (0.06)
β_1	0.88 (0.08)	0.86 (0.08)	0.88 (0.10)
β_2	-5.46 (0.13)	-5.44 (0.15)	-5.45 (0.15)
γ_{11}	0.002	0.002	0.002
γ_{22}	0.007	0.016	0.013
γ_{33}	0.082	0.064	0.066
γ_{12}	0.001	0.002	0.000
γ_{13}	-0.023	-0.025	0.001
γ_{23}	-0.002	-0.001	-0.018
σ^2	0.009	0.006	0.009
$\widehat{\ell}_o(\widehat{\boldsymbol{\theta}})$	71.34	64.93	

[a] estimates using MCMC methods in `WinBUGS`

where random effects $\boldsymbol{b}$ follow a multivariate normal distribution and the error terms follow a Student-*t* distribution with $\eta = 2.1$. Results of the estimation by maximum likelihood method for full Gaussian and Gaussian/Student–*t* models, obtained using the SAEM algorithm for $\boldsymbol{\theta} = (\boldsymbol{\beta}^T, \boldsymbol{\Gamma}, \sigma^2)^T$, are presented in Table 4. Also, we fit the Gaussian/Student–*t* NLMEM using the `WinBUGS` using the following non–informative priors $\boldsymbol{\beta} \sim N_3(\mathbf{0}, 10^3\boldsymbol{I}_3)$, $\sigma \sim U(0, 10^3)$, $\boldsymbol{\Gamma} \sim Wishart(10^3\boldsymbol{I}_3, 4)$. From Table 4 we can see that the results using the SAEM algorithm are in agreement with the estimates obtained using `WinBUGS` except that of the covariances of the random effects and the variance of the error term.

In order to detect outliers, the estimates of κ_i and τ_i are shown in Table 5. As we expected, animals 1 and 6 present small values of τ_i suggesting that animal 1 is a possible ϵ–outlier. This result shows the flexibility of the proposed NLMEM since this kind of model has the ability to adapt to outlying observations in this example.

Table 5 Guinea pigs data: Estimated weights for the Gaussian/Student-*t* model.

Subject	1	2	3	4
Residual errors (κ_i)	**0.31**	0.71	1.07	1.62
Random effects (τ_i)	1	1	1	1
Subject	5	6	7	8
Residual errors (κ_i)	0.62	**0.47**	1.82	0.86
Random effects (τ_i)	1	1	1	1

For this example, we compare SAEM with MCEM for this specific Gaussian/Student-*t* model, following the same strategy and the same predetermined sequence of Monte Carlo sample size used for the theophylline data. We observed that MCEM estimates are in good agreement with the SAEM algorithm but with a considerable computational effort since the SAEM algorithm takes almost 27 seconds for 1000 iterations while MCEM takes 7624 seconds to perform the same number of iterations.

7 Discussion

This paper considers an extension of NLMEMs where random effects and error term follow a large class of parametric distributions. The class of distributions we consider is scale mixtures of multivariate normal distributions that are often useful for robust inference. Therefore, this work represents a natural generalization of previous works of Pinheiro, et al. (2001); Lin and Lee (2006) and Lin (2008), for the nonlinear mixed–effects context. Thus, our propose is an alternative to the works of Yeap and Davidian (2001) and Yeap, et al. (2003).

We have implemented the stochastic approximation of the EM algorithm to obtain the maximum likelihood estimates of model parameters. The results obtained with the theophylline data show that the proposed algorithms are easy to implement and computationally efficient. A specific parameter expansion version of the SAEM algorithm was also proposed and was found to speed up the standard algorithm when used during the first iterations. The PX–SAEM algorithm was applied to σ^2, the variance error, resulting a very basic and simple modification of the standard SAEM algorithm. There exist several strategy to apply PX in this kind of model, for instance introducing working parameter in the random effects as proposed in Lavielle and Meza (2007) for Gaussian NLMEM, which potentially allow to increase more the speed of convergence. However, more complex strategies, i.e. more complex expanded models, involve more complexity in the maximization step, which leads to use maximization procedures. A balance between the gain in speed of convergence and complexity of the algorithm must be find to model longitudinal data sets using the kind of models introduced in this work.

The algorithms proposed can be easily adapted for use with existing software, such as `MONOLIX` (Lavielle, 2005). Extensions of this work to the case of REML estimation can be easily adapted following the approach developed in Meza, et al. (2007).

An interesting feature of the proposed formulation is that the weight κ_i and τ_i can accommodate outliers for each of the sources of variability in the model. In fact, these weights can be used as tools for identifying outlying observations. Although the application addressed in our work reveals a very attractive performance for the identification of potential outliers, the influence of outliers on the maximum likelihood estimation still need to be investigated. To carry out this kind of analysis, it has been proposed to use cases deletion procedures (see, for example, Cook and Weisberg, 1982). A general approach for detecting outliers in regression models is the mean-shift outlier model (Cook and Weisberg, 1982). It has been shown that this approach is equivalent to the diagnostic analysis by elimination of observations in linear and nonlinear regression models under normal errors (Wei and

Shih, 1994). However, in our knowledge, few studies have been developed about the outlier detection using the mean-shift outlier model for nonnormal data or models with longitudinal structure (see, for example, Wei and Fung, 1999; Shi and Chen, 2008). To assess the influence of the ith subject on the estimates of $\boldsymbol{\theta}$, we can consider the mean-shift outlier model

$$\begin{aligned} \boldsymbol{Y}_i &= \boldsymbol{G}_d\boldsymbol{\psi} + \boldsymbol{f}_i(\boldsymbol{z}_i, \boldsymbol{\phi}_i) + \boldsymbol{\epsilon}_i, \\ \boldsymbol{Y}_j &= \boldsymbol{f}_j(\boldsymbol{z}_j, \boldsymbol{\phi}_j) + \boldsymbol{\epsilon}_j, \quad j \neq i, \end{aligned} \tag{43}$$

where $d = \{j_1, \ldots, j_d\}$ and $\boldsymbol{G}_d = (\boldsymbol{d}_{j_1}, \ldots, \boldsymbol{d}_{j_d})$ is a $n_i \times d$ matrix with $\boldsymbol{d}_{j_k}$ a n_i-dimensional vector with 1 at the j_kth position and zero elsewhere. Note that this formulation allows to assess whether the jth observation or groups of observations on the ith subject have an atypical behaviour by testing the hypothesis $H_0 : \boldsymbol{\psi} = \boldsymbol{0}$. In particular, we declare the ith subject as an $\boldsymbol{\epsilon}$-outlier considering $\boldsymbol{G}_d = \boldsymbol{I}_{n_i}$. We can carry out the above hypothesis test by fitting the model (43) and use the likelihood ratio test to compare models (43) versus (5). It is interesting to note that, as $\boldsymbol{\psi}$ is linearly incorporated in the mean-shift model defined in (43), the estimation of parameters requires a small modification of the procedure described in (10)-(14) and (15)-(18). Analogously we can make the detection of $\boldsymbol{b}$-outliers using the mean-shift outlier model for the subject-specific parameters,

$$\begin{aligned} \boldsymbol{\phi}_k &= \boldsymbol{G}_k\boldsymbol{\omega} + \boldsymbol{A}_k\boldsymbol{\beta} + \boldsymbol{B}_k\boldsymbol{b}_k, \\ \boldsymbol{\phi}_l &= \boldsymbol{A}_l\boldsymbol{\beta} + \boldsymbol{B}_l\boldsymbol{b}_l, \quad l \neq k, \end{aligned} \tag{44}$$

where the matrix $\boldsymbol{G}_k \in \mathbb{R}^{r\times s}$ has a definition similar to that given in (43). In this model, however, the parameter $\boldsymbol{\omega}$ is nonrandom and in order to test the hypothesis $H_0 : \boldsymbol{\omega} = \boldsymbol{0}$ using the likelihood ratio test, we must to consider a modification of the estimation process that requires to choose the matrix $\boldsymbol{B}_i$ described in (6) and (7) properly. It is possible to make an exhaustive search for outliers in a specific data set considering

$$H_0 : \boldsymbol{\psi} = \boldsymbol{0}, \text{ and/or, } \boldsymbol{\omega} = \boldsymbol{0}, \quad \text{for all } i, k = 1, \ldots, n, \tag{45}$$

but it is well known that carrying out this strategy to search for outliers may require a large computational burden. We recommend to adopt this approach to identify outliers for those observations that have received small weights by the estimation procedure or those identified as potential outliers in the plot of Mahalanobis distances. We expect that if the models with heavier tails than normal one proposed in (6) have the capacity to accommodate outliers, leading to the acceptance of the hypothesis in (45) whereas this situation may not be true under Gaussian errors.

Although for a specific observation the hypothesis (45) may be accepted, that is, is not an outlier. It may happen to have influence on other aspects of the model, in which case we requires more insightful methodologies to assess the influence of atypical observations. In particular, for nonlinear models with mixed–effects Lee and Xu (2004) and Russo, et al. (2009) conducted influence analyses considering the local influence method (Cook, 1986). This technique consists into study the effect of introducing small perturbations in the model (or data) using an appropriate measure of influence. The methodology has received increasing attention over the past 20 years mainly due to its flexibility to assess the model assumptions (see the discussion in Cook, 1997). Currently the authors work in the development of diagnostic techniques for the model proposed in this paper using both the local influence procedure and the mean-shift outlier model, thus extending the work of Osorio, et al. (2007) and Russo, et al. (2009).

In this paper we considered NLMEMs in which both random effects and error terms follow a SMN distribution. The class of SMN distributions provides a group of thick-tailed distributions that are often useful for robust inference, but in many applications the presence of skewness is detected in data sets. Therefore, it is necessary to consider flexible distributions that account for this issue. The scale mixtures of skew-normal distributions (Branco and Dey, 2001) is a class of skew-thick-tailed distributions, which extends the class of SMN distributions. Thus a generalization of this work is to consider the NLMEMs using scale mixtures of skew-normal distributions, thus extending the work of De la Cruz (2008) and De la Cruz and Branco (2009).

Acknowledgements The first author was partially supported by grants PBCT PSD-20, DIPUV 5/2007 and FONDECYT 11090024 . The second and third authors were partially supported by Fondo Nacional de Desarrollo Científico y Tecnológico - FONDECYT grants 11075071 and 11080017, respectively. We would also like to thank the reviewers for their constructive comments, which helped to substantially improve this manuscript.

References

Andrews, D.F., and Mallows, C.L.: Scale mixtures of normal distributions. J. R. Stat. Soc. Ser. B. 36, 99-102 (1974).

Beal, S.L., and Sheiner, L.B.: NONMEN User's Guide. Nonlinear Mixed–Effects Models for Repeated Measures Data. San Francisco: University of California (1992).

Branco, M.D., and Dey, D.K.: A general class of multivariate skew-elliptical distributions. J. Multivar. Anal. 79, 99-113 (2001).

Cai, L.: High-dimensional exploratory item factor analysis by a Metropolis-Hastings Robbins-Monro algorithm. Psychometrika. 75, 33-57 (2010).

Choy, S.T.B., and Smith, A.F.M.: Hierarchical models with scale mixtures of normal distributions. Test 6, 205-221 (1997).

Cook, R.D.: Assessment of local influence (with discussion). J. R. Stat. Soc. Ser. B. 48, 133-169 (1986).

Cook, R.D.: Local Influence. In Kotz, S., Read, C.B., and Banks, D.L. (eds.), Encyclopedia of Statistical Sciences, Update, vol. 1, pp. 380-385. Wiley (1997).

Cook, R.D., and Weisberg, S.: Residuals and Influence in Regression. Chapman & Hall, London (1982).

Copt, S., and Victoria-Feser, M.: High breakdown inference in the mixed linear model. J. Am. Stat. Assoc. 101, 292-300 (2006).

Davidian, M., and Giltinan, D.M.: Nonlinear Models for Repeated Measurements Data. Chapman & Hall, New York (1995).

Davidian, M., and Giltinan, D.M.: Nonlinear models for repeated measurements: An overview and update. J. Agric. Biol. Environ. Stat. 8, 387-419 (2003).

De la Cruz, R.: Bayesian non-linear regression models with skew-elliptical errors: Applications to the classification of longitudinal profiles. Computational Statistics and Data Analysis 53, 436-229 (2008).

De la Cruz, R., and Branco, M.D.: Bayesian analysis for nonlinear regression model under skewed errors, with application in growth curves. Biometrical Journal 51 (4), 588609 (2009).

Demidenko, E.: Mixed Models: Theory and Applications, Wiley, New York (2004).

Delyon, B., Lavielle, M., and Moulines, E.: Convergence of a stochastic approximation version of the EM algorithm. Ann. Stat. 27, 94-128 (1999).

Dempster, A.P., Laird, N.M. and Rubin, D.B.: Maximum likelihood from incomplete data via the EM algorithm (with discussion). J. R. Stat. Soc. Ser. B. 39, 1-38 (1977).

Fang, K.T., Kotz, S., and Ng, K.W.: Symmetric Multivariate and Related Distributions. Chapman & Hall, London (1990).

Fernández, C., and Steel, M.F.J.: Multivariate Student–*t* regression models: pitfalls and inference. Biometrika 86, 153-167 (1999).

Gu, M.G., and Kong, F.H.: A stochastic approximation algorithm with Markov chain Monte-Carlo method for incomplete data estimation problems. Proceeding of National Academy Sciences, USA 95, 7270-7274.

Jank, W.: Implementing and diagnosing the stochastic approximation EM algorithm. J. Comput. Graph. Stat. 15, 803-829 (2006).

Jamshidian, M.: Adaptive robust regression by using a nonlinear regression program. J Stat Softw. `http://www.jstatsoft.org/v04/i06` (1999).

Jara, A., Quintana, F., and San Martin, E.: Linear mixed models with skew–elliptical distributions: A Bayesian approach. Comput. Stat. Data Anal. 52, 5033-5045 (2008).

Johansen, S.: Functional Relations, Random Coefficients, and Nonlinear Regression with Application to Kinectic Data. Springer–Verlag, New York (1984).

Kent, J.T., Tyler, D.E., and Vardi, Y.: A curious likelihood identity for the multivariate *t* distribution. Commun. Stat. Simulat. C. 23, 441-453 (1994).

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

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

Lange, K.L., Little, R.J.A., and Taylor, J.M.G.: Robust statistical modeling using the *t* distribution. J. Am. Stat. Assoc. 84, 881-896 (1989).

Lange, K., and Sinsheimer, J.: Normal/Independent distributions and their applications in robust regression. J. Comput. Graph. Stat. 2, 175-198 (1993).

Lavielle, M.: Monolix User Guide Manual. `http://www.monolix.org` (2005).

Lavielle, M., and Meza, C.: A parameter expansion version of the SAEM algorithm. Stat. Comput. 17, 121-130 (2007).

Lee, S., and Xu, L.: Influence analyses of nonlinear mixed-effects models. Comput. Stat. Data Anal. 45, 321-341 (2004).

Lin, T.I.: Longitudinal data analysis using *t* linear mixed models with autoregressive dependence structures. J. Data Sci. 6, 333-355 (2008).

Lin, T.I., and Lee, J.C.: A robust approach to *t* linear mixed models applied to multiple sclerosis data. Stat. Med. 25, 1397-1412 (2006).

Lin, T.I., and Lee, J.C.: Estimation and prediction in linear mixed models with skew–normal random effects for longitudinal data. Stat. Med. 27, 1490-1507 (2007).

Lindstrom, M.J., and Bates, D.M.: Nonlinear mixed–effects models for repeated measures data. Biometrics 46, 673-787 (1990).

Little, R.J.A., and Rubin, D.B.: Statistical Analysis with Missing Data. Wiley, New York (2002).

Liu, C.: Bayesian robust multivariate linear regression with incomplete data. J. Am. Stat. Assoc. 91, 1219-1227 (1996).

Liu, C., Rubin, D. and Wu, Y.: Parameter expansion to accelerate EM: The PX-EM algorithm. Biometrika 85, 755-770 (1998).

Louis, T.A.: Finding the observed information matrix when using the EM algorithm. J. R. Stat. Soc. Ser. B. 44, 226-233 (1982).

Lucas, A.: Robustness of the Student *t* based -M-estimator. Commun. Stat. Theor. M. 26, 1165-1182 (1997).

McCulloch, C. E.: Maximum likelihood algorithms for generalized linear mixed models, J. Am. Stat. Assoc. 92, 162-170 (1997).

Meng, X.L. and van Dyk, D.A.: The EM algorithm - an old folk song sung to a fast new tune (with discussion). J. R. Stat. Soc. Ser. B. 59, 511-567 (1997).

Meza, C., Jaffrézic, F. and Foulley, J.L.: REML estimation of variance parameters in nonlinear mixed effects models using SAEM algorithm. Biometrical J. 49, 876-888 (2007).

Meza, C., Jaffrézic, F. and Foulley, J.L.: Estimation in the probit normal model for binary outcomes using the SAEM algorithm. Comput. Stat. Data Anal. 53, 1350-1360 (2009).

Osorio, F., Paula, G.A. and Galea, M.: Assessment of local influence in elliptical linear models with longitudinal structure. Comput. Stat. Data Anal. 51, 4354-4368 (2007).

Philippe, A.: Simulation of right and left truncated gamma distributions by mixtures. Stat. Comp. 7, 173-181 (1997).

Pinheiro, J., and Bates, D.M.: Approximations to the log–likelihood function in the nonlinear mixed–effects model. J. Comput. Graph. Stat. 4, 12-35 (1995).

Pinheiro, J., and Bates, D.M.: Mixed–Effects Models in S and S-PLUS. Springer, New York (2000).

Pinheiro, J., Liu, C., and Wu, Y.: Efficient algorithms for robust estimation in linear mixed–effects models using the multivariate *t* distribution. J. Comput. Graph. Stat. 10, 249-276 (2001).

Roberts, G.O., Gelman, A., and Gilks, W.: Weak convergence and optimal scaling of random walk metropolis algorithm. Ann. Applied Prob. 7, 110-120 (1997).

Roberts, G.O., and Rosenthal, J.S.: Optimal scaling of various metropolis–hastings algorithms. Stat. Science 16, 351-367 (2001).

Rogers, W.H., and Tuckey, J.W.: Understanding some long–tailed distributions. Stat. Neerl. 26, 211-226 (1972).

Rosa, G.J.M., Padovani, C.R., and Gianola, D.: Robust linear mixed models with Normal/Independent distributions and Bayesian MCMC implementation. Biometrical J. 45, 573-590 (2003).

Rosa, G.J.M., Gianola, D., and Padovani, C.R.: Bayesian longitudinal data analysis with mixed models and thick–tailed distributions using MCMC. J. Appl. Statist. 31, 855-873 (2004).

Russo, C.M., Paula, G.A., and Aoki, R.: Influence diagnostics in nonlinear mixed-effects elliptical models. Comput. Stat. Data Anal. 53, 4143-4156 (2009).

Shi, L., and Chen, G.: Detection of outliers in multilevel models. J. Stat. Plan. Infer. 138, 3189-3199 (2008).

Staudenmayer, J., Lake, E.E., and Wand, M.P.: Robustness for general design mixed models using the *t*–distribution. Stat. Model. 9, 235-255 (2009).

Spiegelhalter, D.J., Thomas, A., and Best, N.G.: Winbugs version 1.2 user manual. MRC Biostatistics Unit.

Vaida, F.: Parameter convergence for EM and MM algorithms. Stat. Sinica 15, 831-840 (2005).

Vonesh, E.F.: A note on the use of Laplace's approximation for nonlinear mixed-effects models. Biometrika 83, 447-452 (1996).

Vonesh, E.F., and Chinchilli, V.M.: Linear and Nonlinear Models for the Analysis of Repeated Measurements. Marcel Dekker, New York (1997).

Walker, S.: An EM algorithm for nonlinear random effects models. Biometrics 52, 934-944 (1996).

Wang, J.: EM algorithms for nonlinear mixed effects models. Comput. Stat. Data Anal. 51, 3244-3256 (2007).

Wei, W.H., and Fung, W.K.: The mean-shift outlier model in general weighted regression and its applications. Comput. Stat. Data Anal. 30, 429-441 (1999).

Wei, B., and Shih, J.: On statistical models for regression diagnostics. Ann. Inst. Statist. Math. 46, 267-278 (1994).

Wei, G., and Tanner, M.: A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms. J. Am. Stat. Assoc. 85, 699-704 (1990).

Welsh, A.H., and Richardson, A.M.: Approaches to the robust estimation of mixed models. In Maddala, G.S. and Rao, C.R. (eds.), Handbook of Statistics, vol. 15, pp. 343-384. Elsevier Science (1997).

Wolfinger, R.: Laplace's approximation for nonlinear mixed models. Biometrika 80, 791-795 (1993).

Wolfinger, R.D., and Lin, X.: Two Taylor–series approximation methods for nonlinear mixed models. Comput. Stat. Data Anal. 25, 465490 (1997).

Wu, C.-F.J.: On the convergence properties of the EM algorithm. Ann. Stat. 11, 95-103 (1983).

Yeap, B.Y., and Davidian, M.: Robust two–stage estimation in hierarchical nonlinear models. Biometrics 57, 266-272 (2001).

Yeap, B.Y., Catalano, P.J., Ryan, L.M., and Davidian, M.: Robust two stage approach to repeated measurements analysis of chronic ozone exposure in rats. J. Agric. Biol. Environ. Stat. 8, 438-454 (2003).

Zhu, H., and Lee, S.: Analysis of generalized linear mixed models via a stochastic approximation algorithm with Markov chain Monte-Carlo method. Stat. Comput. 12, 175-183 (2002).