Available online at www.sciencedirect.com
SCIENCE DIRECT®
Computational Statistics & Data Analysis 49 (2005) 1020–1038
COMPUTATIONAL STATISTICS & DATA ANALYSIS
www.elsevier.com/locate/csda

Maximum likelihood estimation in nonlinear mixed effects models

E. Kuhn, M. Lavielle*

Université Paris Sud, Bât. 425, 91405 Orsay, France

Received 10 September 2003; received in revised form 15 July 2004; accepted 16 July 2004
Available online 12 August 2004

Abstract

A stochastic approximation version of EM for maximum likelihood estimation of a wide class of nonlinear mixed effects models is proposed. The main advantage of this algorithm is its ability to provide an estimator close to the MLE in very few iterations. The likelihood of the observations as well as the Fisher Information matrix can also be estimated by stochastic approximations. Numerical experiments allow to highlight the very good performances of the proposed method.

Keywords: Mixed effects model; Nonlinear model; Maximum likelihood estimation; EM algorithm; SAEM algorithm

1. Introduction

The mixed effects models were introduced mainly for modeling responses of individuals that have the same global behavior with individual variations (see the book of Pinheiro and Bates (2000) and the many references therein, for example). In fact, we consider that all the responses follow a common known functional form that depends on unknown effects. Some of them are fixed (i.e. the same for all the individuals), the others are random, so they depend on the individuals (or on sub-groups of the population). Then, the model has two types of parameters: global parameters that correspond to the fixed effects and parameters which vary among the population that correspond to the random effects. This kind of observations are usually the result of repeated measurements: some individuals are

* Corresponding author. Tel.: +31-1-69-15-57-43; fax: +31-1-69-15-72-34.
E-mail addresses: estelle.kuhn@math.u-psud.fr (E. Kuhn), marc.lavielle@math.u-psud.fr (M. Lavielle).

0167-9473/$ - see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2004.07.002

repeatedly observed under different experimental conditions. This approach seems to be adapted to many situations, particularly in the fields of pharmacokinetic, biological growth, epidemiology or econometry.

Let us consider here the following general nonlinear mixed effects model:

$$y_{ij} = g(\boldsymbol{\phi_i}, \boldsymbol{\beta}, x_{ij}) + h(\boldsymbol{\phi_i}, \boldsymbol{\beta}, x_{ij})\varepsilon_{ij}, \quad 1 \leqslant i \leqslant n, \quad 1 \leqslant j \leqslant m_i, \tag{1}$$

where y_{ij} is the jth observation of the ith individual, at some known instant x_{ij}. Here, n is the number of individuals and m_i is the number of observations of individual i. The within-group errors (ε_{ij}) are supposed to be i.i.d. Gaussian random variables with mean zero and unknown variance σ^2. The model is nonlinear means that g or h are nonlinear functions of $\boldsymbol{\phi_i}$. The random vector $\boldsymbol{\phi_i}$ is modelized by

$$\boldsymbol{\phi_i} = \boldsymbol{A_i}\boldsymbol{\mu} + \boldsymbol{\eta_i} \quad \text{with} \quad \boldsymbol{\eta_i} \sim_{\text{i.i.d.}} N(0, \boldsymbol{\Gamma}),$$

where $\boldsymbol{\mu}$ is an unknown vector of population parameters. The individual matrix $\boldsymbol{A_i}$ is supposed to be known. The vector $\boldsymbol{\beta}$ denotes also unknown population parameters, which do not appear in the random effect $\boldsymbol{\phi_i}$. We suppose that the ε_{ij} and the $\boldsymbol{\eta_i}$ are mutually independent.

Our purpose is to propose a method for computing the maximum likelihood estimate of the unknown parameter vector $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\mu}, \boldsymbol{\Gamma}, \sigma^2)$ and to compare this method with other existing methods, particularly those based on the maximum likelihood approach.

In the case of a linear model, the estimation of the unknown parameters can be treated with the usual EM algorithm (Dempster et al., 1977) or with a Newton–Raphson algorithm (Pinheiro and Bates, 2000). A nonlinear function is often more suitable for modeling the physical problem, but requires a specific approach for estimating the parameters. Different methods, based generally on linearization of the log-likelihood, were suggested for dealing with nonlinear models. A Laplace approximation was proposed by Edward F. Vonesh in Vonesh (1996), a Bayesian approach was proposed by Racine-Poon (1985), Wakefield et al. (1994), Wakefield (1996). Walker (1996) uses a Monte-Carlo EM algorithm, whereas a simulated pseudo maximum likelihood estimator for these specific models is developed by Concordet and Nunez (2002).

In this paper, we show that the SAEM algorithm (stochastic approximation version of EM) is very efficient for computing the maximum likelihood estimate of $\boldsymbol{\theta}$. This iterative procedure consists at each iteration, in successively simulating the random effects with the conditional distribution, and updating the unknown parameters of the model. This algorithm was shown to converge under very general conditions by Delyon et al. (1999). When this algorithm is coupled with a MCMC procedure for the simulation step, Kuhn and Lavielle (2004) also established the convergence of the algorithm toward the MLE. Furthermore, the observed likelihood and the Fisher Information matrix can also be estimated by using also a stochastic approximation procedure. This method has the very nice advantage to converge very quickly to a neighborhood of the Maximum Likelihood Estimate. Then, only a few seconds are required for computing a MLE confidence interval, in any of the usual models used in the practice. The SAEM can be used for estimating homoscedastic models, but also heteroscedastic models. For the latter, the parameters related to the fixed effects are estimated in a Bayesian context in term of their expectations. By the way, the SAEM could also be used in an empirical Bayesian context for estimating the prior distribution of

the parameters of the model. Nevertheless, any comparison with any Bayesian estimation method is beyond the scope of the present paper.

Section 2 presents the EM and the SAEM algorithms. We briefly recall the main convergence results. We illustrate the proposed method in Section 3 with the very simple example of orange data, reported in Pinheiro and Bates (2000). The model used for this data is linear with respect to the random effects $\boldsymbol{\phi_i}$. Then the exact MLE can be computed with EM, and compared to the SAEM algorithm. Section 4 is dedicated to the comparison of SAEM with other popular methods of estimation through two numerical examples.

2. Algorithms proposed for maximum likelihood estimation

Any mixed effects model can be seen as an usual missing data problem. Indeed, the observed data are the y_{ij}, for $1 \leqslant i \leqslant n$ and $1 \leqslant j \leqslant m_i$, whereas the random effects $\boldsymbol{\phi_i}$, for $1 \leqslant i \leqslant n$, are the nonobserved data. Then, the complete data of the model are $(\boldsymbol{y}, \boldsymbol{\phi}) = (y_{ij}, \boldsymbol{\phi_i})_{1 \leqslant i \leqslant n, 1 \leqslant j \leqslant m_i}$. In the sequel, we will make the following assumption concerning the model:

- (H0) For any $1 \leqslant i \leqslant n$ and any $1 \leqslant j \leqslant m_i$,

$$y_{ij} = g(\boldsymbol{\phi_i}, \boldsymbol{\beta}, x_{ij}) + h(\boldsymbol{\phi_i}, \boldsymbol{\beta}, x_{ij})\varepsilon_{ij}, \tag{2}$$

  where g and h can be decomposed as

$$g(\boldsymbol{\phi_i}, \boldsymbol{\beta}, x_{ij}) = g_1(\boldsymbol{\phi_i}, x_{ij})g_2(\boldsymbol{\beta}, x_{ij}), \tag{3}$$

$$h(\boldsymbol{\phi_i}, \boldsymbol{\beta}, x_{ij}) = h_1(\boldsymbol{\phi_i}, x_{ij})h_2(\boldsymbol{\beta}, x_{ij}). \tag{4}$$

Assumption (H0) is equivalent to suppose that the complete data likelihood $f(\boldsymbol{y}, \boldsymbol{\phi}; \boldsymbol{\theta})$ belongs to the curved exponential family i.e. it can be written as

$$f(\boldsymbol{y}, \boldsymbol{\phi}; \boldsymbol{\theta}) = \exp\{-\Psi(\boldsymbol{\theta}) + \langle \tilde{S}(\boldsymbol{y}, \boldsymbol{\phi}), \Phi(\boldsymbol{\theta}) \rangle\}, \tag{5}$$

where Ψ and Φ denote two functions of the unknown parameter $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\mu}, \boldsymbol{\Gamma}, \sigma^2)$ defined on a subset Θ of $\mathbb{R}^p$ and $\langle \cdot, \cdot \rangle$ the scalar product and where $\tilde{S}(\boldsymbol{y}, \boldsymbol{\phi})$ is known as the minimal sufficient statistics of the complete model, taking its value in a subset $\mathcal{S}$ of $\mathbb{R}^m$.

We define here the function $L : \mathcal{S} \times \Theta \to \mathbb{R}$ as

$$L(\boldsymbol{s}, \boldsymbol{\theta}) = -\Psi(\boldsymbol{\theta}) + \langle \boldsymbol{s}, \Phi(\boldsymbol{\theta}) \rangle,$$

the likelihood of the observations $q(\boldsymbol{y}; \boldsymbol{\theta})$ as

$$q(\boldsymbol{y}; \boldsymbol{\theta}) = \int f(\boldsymbol{y}, \boldsymbol{\phi}; \boldsymbol{\theta}) \, \mathrm{d}\boldsymbol{\phi}$$

and the log-likelihood of the observations $l = \log q$.

Our purpose is then to compute a value of $\boldsymbol{\theta}$ which maximizes the function l for fixed observations $\boldsymbol{y}$.

2.1. The EM algorithm

Assuming the exponential form (5) of the complete likelihood, the kth iteration of EM reduces to the following two steps:

- *E-step*: evaluate the quantity $\boldsymbol{s_{k+1}} = E[\tilde{S}(\boldsymbol{y}, \boldsymbol{\phi})|\boldsymbol{y}; \boldsymbol{\theta_k}]$.
- *M-step*: compute $\boldsymbol{\theta_{k+1}} = \text{Argmax}_{\boldsymbol{\theta}}\, L(\boldsymbol{s_{k+1}}, \boldsymbol{\theta})$.

The complete likelihood f has the following analytical expression:

$$f(\boldsymbol{y}, \boldsymbol{\phi}; \boldsymbol{\theta}) \propto |\boldsymbol{\Gamma}|^{-n/2} \prod_{i,j} (\sigma^2 h^2(\boldsymbol{\phi_i}, \boldsymbol{\beta}, x_{ij}))^{-1/2} \times \exp\left\{ -\frac{1}{2} \sum_i (\boldsymbol{\phi_i} - \boldsymbol{A_i}\boldsymbol{\mu})^t \boldsymbol{\Gamma}^{-1} (\boldsymbol{\phi_i} - \boldsymbol{A_i}\boldsymbol{\mu}) - \frac{1}{2\sigma^2} \sum_{i,j} \left(\frac{y_{ij} - g(\boldsymbol{\phi_i}, \boldsymbol{\beta}, x_{ij})}{h(\boldsymbol{\phi_i}, \boldsymbol{\beta}, x_{ij})} \right)^2 \right\}. \tag{6}$$

If the function g depends linearly on $\boldsymbol{\phi}$ and if the function h does not depend on the random effects $\boldsymbol{\phi}$, this joint distribution is Gaussian, and the E-step can be performed in a close form. Following Dempster et al. (1977) and Wu (1983), convergence of EM is ensured if the following regularity assumptions are checked:

- (H1) The parameter vector $\boldsymbol{\theta}$ belongs to a parameter space Θ, which is an open subset of $\mathbb{R}^p$.
- (H2) The functions g and h are twice continuously differentiable with respect to $\boldsymbol{\beta}$.
- (H3) Let us define for all $\boldsymbol{\beta}$ and for all $\boldsymbol{\phi}$

$$\Lambda(\boldsymbol{\beta}, \boldsymbol{\phi}) = \frac{1}{2\sigma^2} \sum_{ij} \left(\frac{y_{ij} - g(\boldsymbol{\phi_i}, \boldsymbol{\beta}, x_{ij})}{h(\boldsymbol{\phi_i}, \boldsymbol{\beta}, x_{ij})} \right)^2 + \sum_{ij} \log |h(\boldsymbol{\phi_i}, \boldsymbol{\beta}, x_{ij})|.$$

There exists a continuously differentiable function $\hat{\beta}$ such that for all $\boldsymbol{\phi}$:

$$\Lambda(\hat{\beta}(\boldsymbol{\phi}), \boldsymbol{\phi}) \leqslant \Lambda(\boldsymbol{\beta}, \boldsymbol{\phi}).$$

Theorem 1. *Assume that assumptions* (H0)–(H3) *hold. Then, the sequence* $(\boldsymbol{\theta_k})$ *obtained from the EM algorithm converges to a stationary point of the log-likelihood l of the observations* $\boldsymbol{y}$ (*i.e. a point where the derivative of l is* 0).

In cases where the function h depends on the random effects or where the function g does not depend linearly on the random effects, we propose an alternative which is a stochastic version of the EM algorithm.

2.2. A stochastic version of the EM algorithm

2.2.1. Description and general convergence result of the algorithm

The stochastic approximation version of the standard EM algorithm, proposed by Delyon et al. (1999) consists in replacing the usual E-step of EM by a stochastic procedure composed of two steps: first a simulation step of the missing data under the conditional distribution, second a stochastic approximation step:

- *Simulation-step*: draw $\boldsymbol{\phi}^{(\boldsymbol{k+1})}$ from the conditional distribution $p(\cdot|\boldsymbol{y};\boldsymbol{\theta_k})$.
- *Stochastic approximation*: update $\boldsymbol{s_k}$ according to

$$\boldsymbol{s_{k+1}} = \boldsymbol{s_k} + \gamma_k(\tilde{S}(\boldsymbol{y},\boldsymbol{\phi}^{(\boldsymbol{k+1})}) - \boldsymbol{s_k}). \tag{7}$$

- *Maximization-step*: update $\boldsymbol{\theta_k}$ according to

$$\boldsymbol{\theta_{k+1}} = \text{Arg}\max_{\boldsymbol{\theta}} L(\boldsymbol{s_{k+1}},\boldsymbol{\theta}).$$

When the simulation step cannot be directly performed, Kuhn and Lavielle (2004) propose to combine this algorithm with a MCMC procedure: the sequence $(\boldsymbol{\phi}^{(\boldsymbol{k})})$ is a Markov Chain with transition kernels $(\boldsymbol{\Pi_{\theta_k}})$. Then, the simulation step becomes:

- *Simulation-step*: using $\boldsymbol{\phi}^{(\boldsymbol{k})}$, draw $\boldsymbol{\phi}^{(\boldsymbol{k+1})}$ from the transition probability $\boldsymbol{\Pi_{\theta_k}}(\boldsymbol{\phi}^{(\boldsymbol{k})},\cdot)$.

Delyon et al. (1999) and Kuhn and Lavielle (2004) have shown that SAEM converges with probability 1 to a maximum (local or global) of the log-likelihood of the observations l under very general conditions. We just present here the conditions important from a practical point of view (see the two papers mentioned above for more details concerning the technical conditions):

- (SAEM1) For all k in $\mathbb{N}$, $\gamma_k \in [0,1]$, $\sum_{k=1}^{\infty}\gamma_k = \infty$ and $\sum_{k=1}^{\infty}\gamma_k^2 < \infty$.
- (SAEM2) For any $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, the transition kernel $\boldsymbol{\Pi_\theta}$ generates a uniformly ergodic chain which invariant probability is $p(\cdot|\boldsymbol{y};\boldsymbol{\theta})$.

Theorem 2. *Assume that assumptions* (H0)–(H3), (SAEM1) *and* (SAEM2) *hold, together with some standard regularity conditions. Then, the sequence* $(\boldsymbol{\theta_k})$ *obtained from the SAEM algorithm converges with probability* 1 *to a maximum* (*local or global*) $\widehat{\boldsymbol{\theta}}^{\text{MLE}}$ *of the log-likelihood* l *of the observations.*

Remark 1. It is well known that, according to the initial guess, EM can converge to a local maximum, a saddle-point, or a minimum of l. The stochastic behavior of SAEM ensures the convergence to a maximum of l. Nevertheless, this maximum is not necessary the global maximum of l. We could think in a simulated annealing version of SAEM, to prevent convergence to a local maximum (see Lavielle and Moulines (1997) for an application to the deconvolution problem).

Remark 2. One of the technical conditions required to prove the convergence of SAEM is the compactness of the support of $p(\cdot|\boldsymbol{y};\boldsymbol{\theta})$. In the case of the nonlinear mixed effects model we have chosen to consider here, the nonobserved data $\boldsymbol{\phi}$ follow a Gaussian distribution and theoretically take their values over an infinite set which is not compact. From a practical point of view, this assumption is not a restriction, since in the practice, any Gaussian random variable takes its values in a (very large) compact set.

Remark 3. Consider the homoscedastic model where the function h defined in (2) is constant and equal to 1. In this particular case, the assumptions are quite simpler. Indeed, convergence of SAEM is ensured if the function $\Lambda_1(\boldsymbol{\phi},\boldsymbol{\beta}) = \sum_{ij}(y_{ij} - g(\boldsymbol{\phi_i},\boldsymbol{\beta},x_{ij}))^2$ possesses for all $\boldsymbol{\phi}$ a minimum denoted $\hat{\beta}(\boldsymbol{\phi})$ such that the function $\hat{\beta}$ is continuously differentiable.

In the case of an heteroscedastic model such that $g = h$, convergence of SAEM requires that the following function:

$$\Lambda_2(\boldsymbol{\phi},\boldsymbol{\beta}) = \frac{1}{2\sigma^2}\sum_{i,j}\left(\frac{y_{ij}}{g(\boldsymbol{\phi_i},\boldsymbol{\beta},x_{ij})} - 1\right)^2 + \sum_{ij}\log|g_1(\boldsymbol{\beta},x_{ij})|$$

possesses for all $\boldsymbol{\phi}$ a minimum denoted $\hat{\beta}(\boldsymbol{\phi})$ such that the function $\hat{\beta}$ is continuously differentiable.

Remark 4. The SAEM is useful for fitting models that belong to the exponential family. If this assumption is not satisfied, it is always possible to consider the vector of fixed parameters $\boldsymbol{\beta}$ as a random vector. Then, $\boldsymbol{\beta}$ is estimated in a Bayesian context in term of its expectation. An example of such extension is proposed Section 3.3.

2.2.2. Simulation of the missing data

When $\boldsymbol{\phi}^{(\boldsymbol{k+1})}$ cannot be exactly drawn from the conditional distribution $p(\cdot|\boldsymbol{y};\boldsymbol{\theta_k})$, we have to choose judiciously a transition $\Pi_{\boldsymbol{\theta}}$ that converges to the target distribution $p(\cdot|\boldsymbol{y};\boldsymbol{\theta})$. The Metropolis–Hastings algorithm provides a solution in general cases. Usually, $\Pi_{\boldsymbol{\theta}}$ will be defined as the succession of M iterations of a MCMC procedure and the simulation-step of iteration k consists in simulating $\boldsymbol{\phi}^{(\boldsymbol{k+1})}$ with the transition probability $\Pi_{\boldsymbol{\theta_k}}(\boldsymbol{\phi}^{(\boldsymbol{k})}, d\boldsymbol{\phi}^{(\boldsymbol{k+1})}) = P^M_{\boldsymbol{\theta_k}}(\boldsymbol{\phi}^{(\boldsymbol{k})}, d\boldsymbol{\phi}^{(\boldsymbol{k+1})})$, where

$$P_{\boldsymbol{\theta_k}}(\boldsymbol{\phi}, d\boldsymbol{\phi}') = r_{\boldsymbol{\theta_k}}(\boldsymbol{\phi},\boldsymbol{\phi}')\min\left\{\frac{p(\boldsymbol{\phi}'|\boldsymbol{y};\boldsymbol{\theta_k})r_{\boldsymbol{\theta_k}}(\boldsymbol{\phi}',\boldsymbol{\phi})}{p(\boldsymbol{\phi}|\boldsymbol{y};\boldsymbol{\theta_k})r_{\boldsymbol{\theta_k}}(\boldsymbol{\phi},\boldsymbol{\phi}')}, 1\right\}\mathrm{d}\phi'$$

for $\boldsymbol{\phi}' \neq \boldsymbol{\phi}$ and $P_{\boldsymbol{\theta_k}}(\boldsymbol{\phi},\{\boldsymbol{\phi}\}) = 1 - \int_{\boldsymbol{\phi}'\neq\boldsymbol{\phi}} P_{\boldsymbol{\theta_k}}(\boldsymbol{\phi}, d\boldsymbol{\phi}')$, where $r_{\boldsymbol{\theta}}(\boldsymbol{\phi},\boldsymbol{\phi}')$ is any aperiodic recurrent transition density.

We set $M = 1$ in the sequel, in order to avoid intricate notations. Extension to any value of M is straightforward. In the practice, we have remarked that the choice of M has very few influence on the speed of convergence of the algorithm. Very low correlations between the $\boldsymbol{\phi}^{(\boldsymbol{k})}$, and obtained with $M \gg 1$, are not useful to improve the convergence of the algorithm. We usually use $M = 5$ or 10 in the practice.

We can use the population distribution π as an instrumental distribution. Then, writing $f(\boldsymbol{y}, \boldsymbol{\phi}; \boldsymbol{\theta}) = d(\boldsymbol{y}|\boldsymbol{\phi}; \boldsymbol{\theta})\pi(\boldsymbol{\phi}; \boldsymbol{\theta})$, the acceptance probability only depends on the conditional distribution d of the observation $\boldsymbol{y}$:

$$P_{\boldsymbol{\theta_k}}(\boldsymbol{\phi}, d\boldsymbol{\phi}') = \pi(\boldsymbol{\phi}'; \boldsymbol{\theta_k}) \min\left\{\frac{d(\boldsymbol{y}|\boldsymbol{\phi}'; \boldsymbol{\theta_k})}{d(\boldsymbol{y}|\boldsymbol{\phi}; \boldsymbol{\theta_k})}, 1\right\} \mathrm{d}\boldsymbol{\phi}'.$$

In the case of the nonlinear mixed effects model, the kth step of this MCMC procedure consists in:

1. draw $\boldsymbol{\phi}' = (\boldsymbol{\phi}'_{\mathbf{1}}, \ldots, \boldsymbol{\phi}'_{\boldsymbol{n}})'$ i.i.d. with the prior distribution $\mathcal{N}(\boldsymbol{\mu_k}, \boldsymbol{\Gamma_k})$ and $\boldsymbol{u} = (u_1, \ldots, u_n)$ i.i.d. with the uniform distribution $\mathcal{U}([0, 1])$.
2. for $i = 1, \ldots, n$, compute

$$\Delta_i = \sum_{j=1}^{m_i} \left[\log\left(\frac{h_1(\boldsymbol{\phi}'_{\boldsymbol{i}}, x_{ij})}{h_1(\boldsymbol{\phi}_{\boldsymbol{i}}^{(\boldsymbol{k})}, x_{ij})}\right) + \frac{1}{2\sigma^2}\left(\frac{y_{ij} - g(\boldsymbol{\phi}'_{\boldsymbol{i}}, \boldsymbol{\beta_k}, x_{ij})}{h(\boldsymbol{\phi}'_{\boldsymbol{i}}, \boldsymbol{\beta_k}, x_{ij})}\right)^2 - \frac{1}{2\sigma^2}\left(\frac{y_{ij} - g(\boldsymbol{\phi}_{\boldsymbol{i}}^{(\boldsymbol{k})}, \boldsymbol{\beta_k}, x_{ij})}{h(\boldsymbol{\phi}_{\boldsymbol{i}}^{(\boldsymbol{k})}, \boldsymbol{\beta_k}, x_{ij})}\right)^2 \right].$$

3. for $i = 1, \ldots, n$, set

$$\boldsymbol{\phi}_{\boldsymbol{i}}^{(\boldsymbol{k+1})} = \boldsymbol{\phi}'_{\boldsymbol{i}} \quad \text{if } \Delta_i \leqslant \log(u_i),$$
$$\boldsymbol{\phi}_{\boldsymbol{i}}^{(\boldsymbol{k+1})} = \boldsymbol{\phi}_{\boldsymbol{i}}^{(\boldsymbol{k})} \quad \text{elsewhere.}$$

2.2.3. Some improvements of the algorithm

We have to choose the sequence of step sizes $(\gamma_k)_{k \geqslant 0}$ such that assumption (SAEM1) is checked: each γ_k must belong to $[0, 1]$, the series $\sum \gamma_k$ must diverge, and the series $\sum \gamma_k^2$ must converge. We must also take into account the influence of the step sizes on the convergence speed of the algorithm. In the practice, it is useful to choose the first step sizes equal to 1, in order to allow more flexibility during the first iterations: in fact, the initial guess $\boldsymbol{\theta_0}$ may be far from the maximum likelihood value we are looking for and the first iterations will require big variations of the sequence $(\boldsymbol{\theta_k})$. After converging to a neighborhood of the MLE, it is interesting to choose smaller step sizes in order to refine the estimation near the objective value and ensure the almost sure convergence of the algorithm. Finally, we recommend to set $\gamma_k = 1$ for $1 \leqslant k \leqslant K$ and $\gamma_k = (k - K)^{-1}$ for $k \geqslant K + 1$. In the practice K can be chosen between 50 and 100.

It is possible to reduce the variations of the sequence $(\boldsymbol{\theta_k})$ around the MLE, by averaging the sequence $(\boldsymbol{s_k})$ after iteration K as follows: (i) run SAEM with $\gamma_k = (k - K)^{-\alpha}$ for $k \geqslant K + 1$, where $0.5 < \alpha < 1$ (ii) compute $\overline{s}$ as an average of the sequence $(\boldsymbol{s_k})$, $k \geqslant K + 1$ and compute the maximum in $\boldsymbol{\theta}$ using $\overline{s}$. Some theoretical results concerning averaging of SAEM are given by Delyon et al. (1999). From a practical point of view, the improvement is slight. We think that it is not really useful for computing an estimate that possesses a variance much bigger than the improvement we can expect.

The MCEM (Monte-Carlo EM) was proposed by Wei and Tanner (1990), and used by Walker (1996) for nonlinear mixed effects models. This algorithm approximates the conditional expectation $Q_k(\boldsymbol{\theta}) = E(\log f(\boldsymbol{y}, \boldsymbol{\phi}; \boldsymbol{\theta})|\boldsymbol{y}; \boldsymbol{\theta}_k)$ in the *E-step* thanks to a Monte Carlo method. It requires to draw a sequence $(\boldsymbol{\phi}^{(k+1,\ell)})$, $1 \leqslant \ell \leqslant L$ at iteration k and to compute $\boldsymbol{s}_{k+1}$ as an average of the $\tilde{S}(\boldsymbol{y}, \boldsymbol{\phi}^{(k+1,j)})$. A good approximation requires a large number L of simulations at each iteration. We can combine MCEM with SAEM by adapting the approximation step (7) as follows:

$$\boldsymbol{s}_{k+1} = \boldsymbol{s}_k + \gamma_k \left(\frac{1}{L} \sum_{\ell=1}^{L} \tilde{S}(\boldsymbol{y}, \boldsymbol{\phi}^{(k+1,\ell)}) - \boldsymbol{s}_k \right). \tag{8}$$

With this new version of the algorithm, a small value of L (smaller than 10 in the practice) is enough to ensure very satisfactory results.

2.2.4. Estimation of the likelihood

The likelihood of the observations can be approximated via a Monte Carlo integration method (see Walker (1996) for example), using the fact that

$$q(\boldsymbol{y}; \boldsymbol{\theta}) = \int d(\boldsymbol{y}|\boldsymbol{\phi}; \boldsymbol{\theta}) \pi(\boldsymbol{\phi}; \boldsymbol{\theta}) \, \mathrm{d}\boldsymbol{\phi} \simeq \frac{1}{T} \sum_{t=1}^{T} d(\boldsymbol{y}|\boldsymbol{\phi}^{(t)}; \boldsymbol{\theta}), \tag{9}$$

where the $\boldsymbol{\phi}^{(t)}$ are drawn independently with the prior distribution $\pi(\cdot; \boldsymbol{\theta})$.

That can be useful if we want to study the behavior of the likelihood at each iteration of SAEM (or EM). At iteration k of the algorithm, a sequence $\boldsymbol{\phi}^{(k1)}, \boldsymbol{\phi}^{(k2)}, \ldots, \boldsymbol{\phi}^{(kT)}$ is drawn with the prior $\pi(\cdot; \boldsymbol{\theta}_k)$ and $q(\boldsymbol{y}; \boldsymbol{\theta}_k)$ is estimated using (9).

If the $\boldsymbol{\phi}^{(kt)}$ are drawn independently at each iteration, the estimated likelihood sequence $(q(\boldsymbol{y}; \boldsymbol{\theta}_k))$ will be very rife if T is not chosen large enough. A smooth curve can be obtained by using the same random numbers at each iteration: for example, if $\boldsymbol{\phi} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\tau})$, then we set $\boldsymbol{\phi}^{(kt)} = \boldsymbol{\mu}_k + \boldsymbol{\tau}_k Z^{(t)}$ for any k, and where the $Z^{(t)}$ are independent $\mathcal{N}(0, 1)$.

Another estimator of the likelihood is obtained by stochastic approximation, setting

$$q_k(\boldsymbol{y}; \boldsymbol{\theta}_k) = q_{k-1}(\boldsymbol{y}; \boldsymbol{\theta}_{k-1}) + \gamma_k \left(\frac{1}{T} \sum_{t=1}^{T} d(\boldsymbol{y}|\boldsymbol{\phi}^{(kt)}; \boldsymbol{\theta}_k)) - q_{k-1}(\boldsymbol{y}; \boldsymbol{\theta}_{k-1}) \right), \tag{10}$$

where the $\boldsymbol{\phi}^{(kt)}$ are independent. Indeed, when the sequence $(\boldsymbol{\theta}_k)$ converges almost surely to $\widehat{\boldsymbol{\theta}}^{\mathrm{MLE}}$, this sequence $(q_k(\boldsymbol{y}; \boldsymbol{\theta}_k))$ converges almost surely to $q(\boldsymbol{y}; \widehat{\boldsymbol{\theta}}^{\mathrm{MLE}})$.

2.2.5. Estimation of the variance of the estimates

Thanks to the maximum likelihood estimator obtained with the proposed algorithm, it is possible to obtain simultaneously an estimation of the Fisher Information matrix. Delyon et al. (1999) propose a method to estimate this matrix by using the fact that the gradient (the Fisher score function) and the Hessian (observed Fisher Information) of l can be obtained almost directly from the simulated missing data $\boldsymbol{\phi}$. Using the so-called Fisher identity,

the Jacobian of the log-likelihood of the observed data $l(\boldsymbol{\theta})$ is equal to the conditional expectation of the complete data likelihood:

$$\partial_\theta l(\boldsymbol{\theta}) \triangleq E[\partial_\theta \log f(\boldsymbol{y}, \boldsymbol{\phi}; \boldsymbol{\theta}) | \boldsymbol{y}; \boldsymbol{\theta}],$$

where ∂_θ denotes the differential with respect to $\boldsymbol{\theta}$. By analogy with the implementation of the SAEM algorithm, the following approximation scheme is proposed:

$$\boldsymbol{\Delta_k} = \boldsymbol{\Delta_{k-1}} + \gamma_k[\partial_\theta \log f(\boldsymbol{y}, \boldsymbol{\phi^{(k)}}; \boldsymbol{\theta_k}) - \boldsymbol{\Delta_{k-1}}].$$

Using the Louis's missing information principle (Louis, 1982), the Hessian of l at $\boldsymbol{\theta}$, is the observed Fisher Information matrix $\partial^2_\theta l(\boldsymbol{\theta})$ that may be expressed as

$$\partial^2_\theta l(\boldsymbol{\theta}) = E_\theta[\partial^2_\theta \log f(\boldsymbol{y}, \boldsymbol{\phi}; \boldsymbol{\theta})] + \mathrm{Cov}_\theta[\partial_\theta \log f(\boldsymbol{y}, \boldsymbol{\phi}; \boldsymbol{\theta})].$$

where $\mathrm{Cov}_\theta(\psi(\boldsymbol{\phi})) \triangleq E_\theta[(\psi(\boldsymbol{\phi}) - E_\theta(\psi(\boldsymbol{\phi})))(\psi(\boldsymbol{\phi}) - E_\theta(\psi(\boldsymbol{\phi})))^t]$. Using this expression, it is possible to derive the following stochastic approximation procedure to approximate $\partial^2_\theta l(\boldsymbol{\theta})$:

$$\begin{aligned}\boldsymbol{G_k} = \boldsymbol{G_{k-1}} &+ \gamma_k[\partial^2_\theta \log f(\boldsymbol{y}, \boldsymbol{\phi^{(k)}}; \boldsymbol{\theta_k}) + \partial_\theta \log f(\boldsymbol{y}, \boldsymbol{\phi^{(k)}}; \boldsymbol{\theta_k}) \\ &\times \partial_\theta \log f(\boldsymbol{y}, \boldsymbol{\phi^{(k)}}; \boldsymbol{\theta_k})^t - \boldsymbol{G}_{k-1}],\end{aligned}$$

$$\boldsymbol{H_k} = \boldsymbol{G_k} - \boldsymbol{\Delta_k}\boldsymbol{\Delta_k^t}.$$

Knowing that the algorithm proposed above converges to a limiting value $\boldsymbol{\theta^\star}$ and that l is regular enough, $(-\boldsymbol{H_k})$ converges to $-\partial^2_\theta l(\boldsymbol{\theta^\star})$. When l is an incomplete data log-likelihood function sufficiently smooth, the maximum likelihood estimator is asymptotically normal and the inverse of the observed Fisher Information matrix $-\partial^2_\theta l(\boldsymbol{\theta^\star})$ converges to the asymptotic covariance of the estimators.

3. Example of the orange trees

3.1. The model

We choose the example of the orange trees to illustrate our algorithm. This data was studied by Pinheiro and Bates (2000) and is available for example on S-plus. The data, shown in Fig. 1, consist in seven measurements of the trunk circumference of each of five orange trees. Inspection of Fig. 1 suggests that a "tree effect" is present.

Following Pinheiro and Bates (2000), a logistic curve is used for modeling the trunk circumference y_{ij} of tree i at age x_j:

$$y_{ij} = g(x_j, \phi_i; \beta_1, \beta_2) + \varepsilon_{ij}, \quad 1 \leqslant i \leqslant n, \ \ 1 \leqslant j \leqslant m, \tag{11}$$

$$g(x_j, \phi_i; \beta_1, \beta_2) = \frac{\phi_i}{1 + \exp(-\frac{x_j - \beta_1}{\beta_2})}. \tag{12}$$

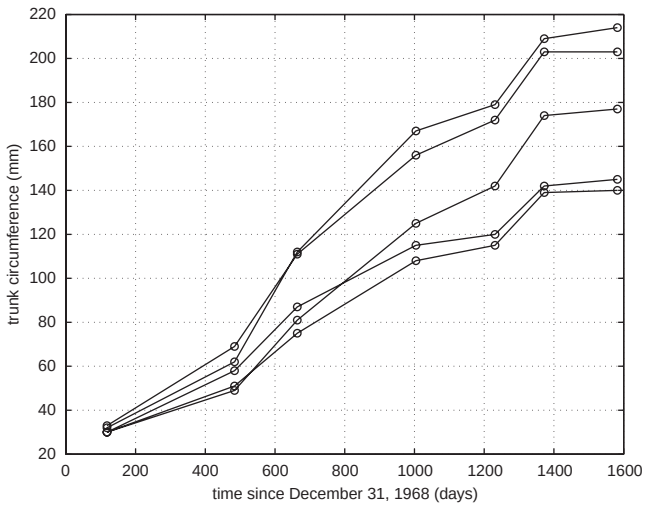

Fig. 1. Circumference of the five orange trees.

We suppose here that the ε_{ij} are independent $\mathcal{N}(0, \sigma^2)$ error terms. On the one hand, the asymptotic trunk circumference $\boldsymbol{\phi_i}$, is treated as a random effect, and is assumed to be Gaussian with mean μ and variance τ^2. On the other hand, the age at which the tree attains half of its asymptotic trunk circumference β_1, and the grow scale β_2 are treated as two fixed effects. Setting

$$g_1(\phi_i) = \phi_i \quad \text{and} \quad g_2(\beta_1, \beta_2, x_j) = \frac{1}{1 + \exp\left(-\frac{x_j - \beta_1}{\beta_2}\right)} \tag{13}$$

the likelihood of the complete model as the form

$$f(\boldsymbol{y}, \boldsymbol{\phi}; \boldsymbol{\theta}) = (2\pi\sigma^2)^{-nm/2}(2\pi\tau^2)^{-m/2} \times \exp\left[-\frac{1}{2\sigma^2}\sum_{i,j}(y_{ij} - g_1(\phi_i)g_2(\beta_1, \beta_2, x_j))^2 - \frac{1}{2\tau^2}\sum_i(\phi_i - \mu)^2\right], \tag{14}$$

where $\boldsymbol{\theta} = (\beta_1, \beta_2, \mu, \tau^2, \sigma^2)$.

3.2. The EM and SAEM algorithms

The nonobserved data is the sequence of random effects (ϕ_i). From (14), we deduce that the conditional distribution $p(\phi_i|\boldsymbol{y}; \boldsymbol{\theta})$ is Gaussian:

$$\phi_i|\boldsymbol{y}; \boldsymbol{\theta} \sim \mathcal{N}(u_i, V),$$

where

$$u_i = V\left(\frac{1}{\sigma^2}\sum_{j=1}^{n} y_{ij}\, g_2(\beta_1,\beta_2,x_j) + \frac{\mu}{\tau^2}\right) \quad \text{and} \quad V = \left(\frac{1}{\sigma^2}\sum_{j=1}^{n} g_2^2(\beta_1,\beta_2,x_j) + \frac{1}{\tau^2}\right)^{-1}. \tag{15}$$

Thus, the Expectation-step of EM and the Simulation-step of SAEM can be easily done. On the other hand, the minimal sufficient statistics function $\tilde{S}$ is

$$\tilde{S}(\boldsymbol{y},\boldsymbol{\phi}) = (\tilde{S}_{\boldsymbol{y}^2}, \tilde{S}_{\boldsymbol{\phi}}, \tilde{S}_{\boldsymbol{\phi}^2}, (\tilde{S}_{\boldsymbol{y}\boldsymbol{\phi}}(j))_{1\leqslant j\leqslant n})), \tag{16}$$

where

$$\tilde{S}_{\boldsymbol{y}^2} = \sum_{i=1}^{n}\sum_{j=1}^{m} y_{ij}^2, \quad \tilde{S}_{\boldsymbol{\phi}} = \sum_{i=1}^{n} \phi_i, \quad \tilde{S}_{\boldsymbol{\phi}^2} = \sum_{i=1}^{n} \phi_i^2, \quad \tilde{S}_{\boldsymbol{y}\boldsymbol{\phi}}(j) = \sum_{i=1}^{n} y_{ij}\phi_i. \tag{17}$$

The maximum likelihood estimate of $\boldsymbol{\theta}$ can be expressed as a function of $\tilde{S}(\boldsymbol{y},\boldsymbol{\phi})$:

$$(\hat{\beta}_1,\hat{\beta}_2) = \text{Arg}\min_{\beta_1,\beta_2}\left\{-2\sum_{j=1}^{m} g_2(\beta_1,\beta_2,x_j)\tilde{S}_{\boldsymbol{y}\boldsymbol{\phi}}(j) + \tilde{S}_{\boldsymbol{\phi}^2}\sum_{j=1}^{m} g_2^2(\beta_1,\beta_2,x_j)\right\}, \tag{18}$$

$$\hat{\mu} = \frac{\tilde{S}_{\boldsymbol{\phi}}}{n}, \tag{19}$$

$$\hat{\tau^2} = \frac{\tilde{S}_{\boldsymbol{\phi}^2}}{n} - \left(\frac{\tilde{S}_{\boldsymbol{\phi}}}{n}\right)^2, \tag{20}$$

$$\hat{\sigma^2} = \frac{1}{nm}\left(\tilde{S}_{\boldsymbol{y}^2} - 2\sum_{j=1}^{m} g_2(\hat{\beta}_1,\hat{\beta}_2,x_j)\tilde{S}_{\boldsymbol{y}\boldsymbol{\phi}}(j) + \tilde{S}_{\boldsymbol{\phi}^2}\sum_{j=1}^{m} g_2^2(\hat{\beta}_1,\hat{\beta}_2,x_j)\right). \tag{21}$$

The M-step of both algorithms requires the use of a Newton–Raphson algorithm for computing $(\hat{\beta}_1,\hat{\beta}_2)$.

The sequence of estimates $(\boldsymbol{\theta}_{\boldsymbol{k}}^{\text{EM}})$ and $(\boldsymbol{\theta}_{\boldsymbol{k}}^{\text{SAEM}})$ are displayed in Figs. 2 and 3, together with the log-likelihood sequences $\log q(\boldsymbol{y};\boldsymbol{\theta}_{\boldsymbol{k}}^{\text{EM}})$ and $\log q(\boldsymbol{y};\boldsymbol{\theta}_{\boldsymbol{k}}^{\text{SAEM}})$ that can easily be computed in a close form

$$q(\boldsymbol{y};\boldsymbol{\theta}) = (2\pi\sigma^2)^{-nm/2}\left(\frac{V}{2}\right)^{-m/2}\exp\left(-\frac{1}{2\sigma^2}\sum_{ij} y_{ij}^2 - \frac{m\mu^2}{2\tau^2} + \frac{A}{2}\sum_{i} u_i^2\right). \tag{22}$$

The estimation of the parameters after 100 and 1000 iterations with EM and SAEM are displayed in Table 1.

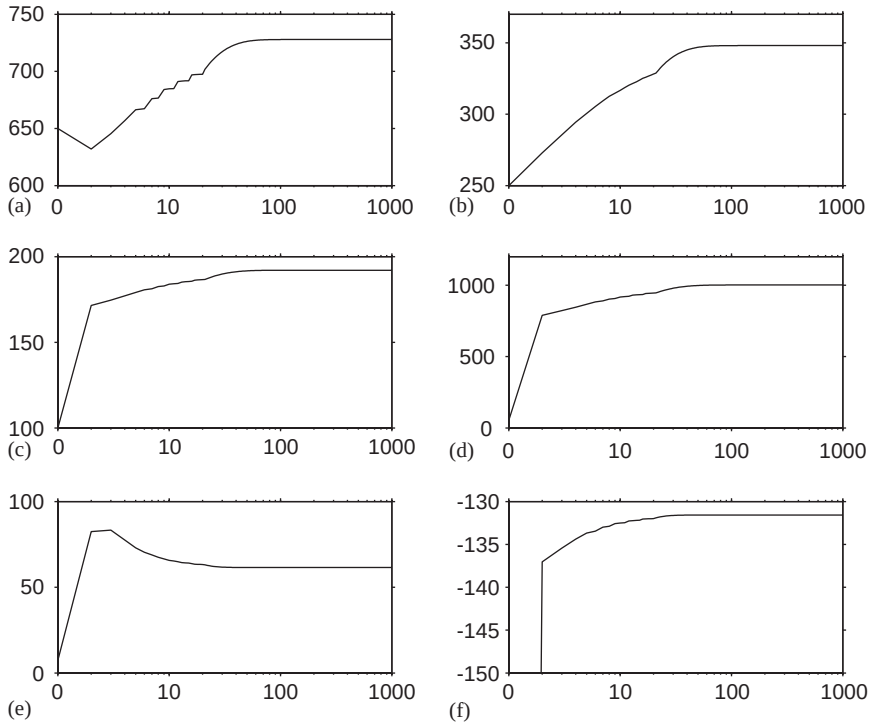

Fig. 2. Estimation of $\boldsymbol{\theta}$ using EM. A logarithmic scale is used for the x-axis. (a) (β_{1k}), (b) (β_{2k}), (c) (μ_k), (d) (τ_k^2), (e) (σ_k^2), (f) $\log q(\boldsymbol{y}; \boldsymbol{\theta}_{\boldsymbol{k}}^{\mathrm{EM}})$.

We can remark that EM has almost converged after 100 iterations. Thus, the value obtained with this algorithm can be considered as the maximum likelihood estimate of $\boldsymbol{\theta}$. In this example, the step size sequence (γ_k) used for SAEM was: $\gamma_k = 1$ for $1 \leqslant k \leqslant 100$ and $\gamma_k = (k - 99)^{-1}$ for $k \geqslant 100$. After some iterations, the SAEM algorithm has converged to a neighborhood of the MLE of $\boldsymbol{\theta}$. Since $\gamma_k = 1$ during the first iterations, no stochastic approximation are still performed, and $\boldsymbol{s_k} = \tilde{S}(\boldsymbol{y}, \boldsymbol{\phi}^{(\boldsymbol{k})})$. Thus, the behavior of $\boldsymbol{\theta}_{\boldsymbol{k}}^{\mathrm{SAEM}}$ remains quite perturbed until iteration 100. After that, the introduction of a decreasing step size allows the almost sure convergence of the sequence $\boldsymbol{\theta}_{\boldsymbol{k}}^{\mathrm{SAEM}}$ to $\widehat{\boldsymbol{\theta}}^{\mathrm{MLE}}$.

The estimation of the observed log-likelihood is displayed in Fig. 4. In this very simple example, the estimator, proposed in Section 2.2.4 can be compared to the true log-likelihood, since this one can be computed by using (22). When the same random Gaussian sequence is used at each iteration, for drawing $\boldsymbol{\phi}^{(\boldsymbol{k1})}, \boldsymbol{\phi}^{(\boldsymbol{k2})}, \ldots, \boldsymbol{\phi}^{(\boldsymbol{kT})}$, with $T = 100$, we can see that the estimated log-likelihood tends to increase at each iteration. We also remark that this sequence converges to an erroneous value. On the other hand, the stochastic approximation proposed in (10) converges to the true log-likelihood. In this example, the step sizes sequence (γ_k) is the same sequence used for estimating the parameters.

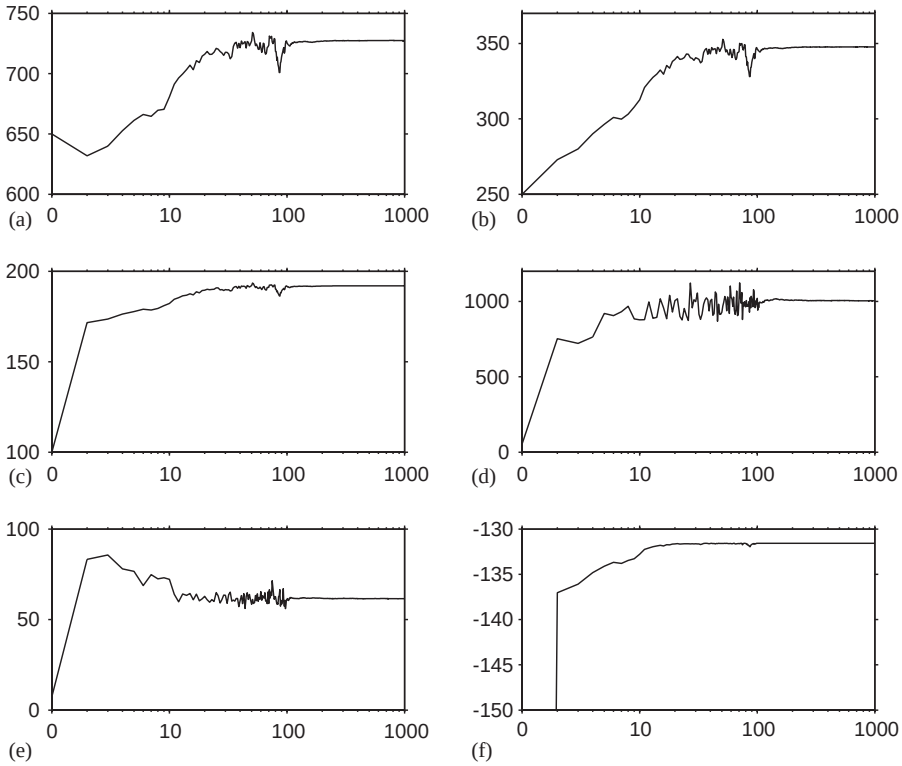

Fig. 3. Estimation of $\boldsymbol{\theta}$ using SAEM. A logarithmic scale is used for the x-axis. (a) (β_{1k}), (b) (β_{2k}), (c) (μ_k), (d) (τ_k^2), (e) (σ_k^2), (f) $\log q(\boldsymbol{y}; \boldsymbol{\theta}_{\boldsymbol{k}}^{\mathrm{SAEM}})$.

Table 1
Comparison of EM and SAEM estimates after 100 and 1000 iterations

Parameters	β_1	β_2	μ	τ^2	σ^2
$\boldsymbol{\theta_0}$	650	250	100	50	10
$\boldsymbol{\theta}_{100}^{\mathrm{EM}}$	727.89	348.06	192.05	1001.45	61.51
$\boldsymbol{\theta}_{1000}^{\mathrm{EM}}$	727.91	348.07	192.05	1001.49	61.51
$\boldsymbol{\theta}_{100}^{\mathrm{SAEM}}$	725.34	346.11	191.46	1020.26	60.56
$\boldsymbol{\theta}_{1000}^{\mathrm{SAEM}}$	727.36	347.67	191.93	1003.76	61.52

The Fisher Information of the MLE can also be estimated by using the stochastic approximation scheme presented in Section 2.2.5. We present in Table 2 the estimated standard deviation of each component of $\boldsymbol{\theta}^{\mathrm{EM}}$ and $\boldsymbol{\theta}^{\mathrm{SAEM}}$, obtained after 100 and 1000 iterations. We remark again that SAEM gives a good estimation in few iterations. Furthermore, 100

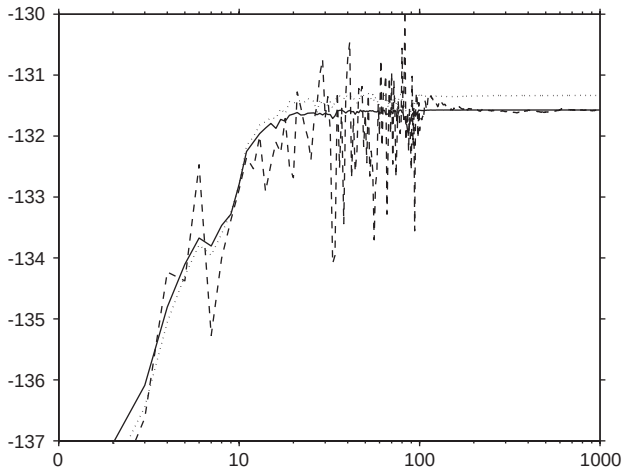

Fig. 4. Estimation of the log-likelihood of the observations. The values of $\log q(\boldsymbol{y}, \boldsymbol{\theta}_{\boldsymbol{k}}^{\mathrm{SAEM}})$ are represented in solid line. The estimated log-likelihood, computed from (9) using the same random sequence at each iteration, is displayed in dotted line. The stochastic approximations $\log q_k(\boldsymbol{y}; \boldsymbol{\theta}_{\boldsymbol{k}}^{\mathrm{SAEM}})$ of the log-likelihood are in dashed line.

Table 2
Estimation of the standard deviation of $\boldsymbol{\theta}^{\mathrm{EM}}$ and $\boldsymbol{\theta}^{\mathrm{SAEM}}$ obtained after 100 and 1000 iterations

Parameters	β_1	β_2	μ	τ^2	σ^2
$\hat{\sigma}(\boldsymbol{\theta}_{100}^{\mathrm{EM}})$	13.51	13.04	14.15	633.39	14.70
$\hat{\sigma}(\boldsymbol{\theta}_{1000}^{\mathrm{EM}})$	13.51	13.04	14.15	633.40	14.70
$\hat{\sigma}(\boldsymbol{\theta}_{100}^{\mathrm{SAEM}})$	12.89	12.51	13.83	604.53	13.41
$\hat{\sigma}(\boldsymbol{\theta}_{1000}^{\mathrm{SAEM}})$	13.51	13.04	14.15	633.40	14.70

iterations of SAEM, including the estimation of the likelihood and the Fisher Information matrix, only require about 1.5 s with a Pentium IV.

Of course, we cannot conclude about the performance of SAEM with only one realization of the algorithms. Indeed, the trajectories of EM and SAEM strongly depend on the value of $\boldsymbol{\theta_0}$. We ran 50 times the SAEM algorithm, with different random initial guess. Each value of $\boldsymbol{\theta_0}$ was independently drawn with a Gaussian distribution of mean $\widehat{\boldsymbol{\theta}}^{\mathrm{MLE}}$ and standard deviation $0.6\widehat{\boldsymbol{\theta}}^{\mathrm{MLE}}$. Then, we ran SAEM, and estimated the expected relative deviation between $\boldsymbol{\theta}_{\boldsymbol{k}}^{\mathrm{SAEM}}$ and $\widehat{\boldsymbol{\theta}}^{\mathrm{MLE}}$. For example, the expected relative deviation for the μ parameter, at iteration k, is approximated by

$$\mathrm{RDEV}_k(\mu) = \frac{1}{50} \sum_{\ell=1}^{50} \left| \frac{\mu_k^{(\ell)} - \hat{\mu}^{\mathrm{MLE}}}{\hat{\mu}^{\mathrm{MLE}}} \right| .$$

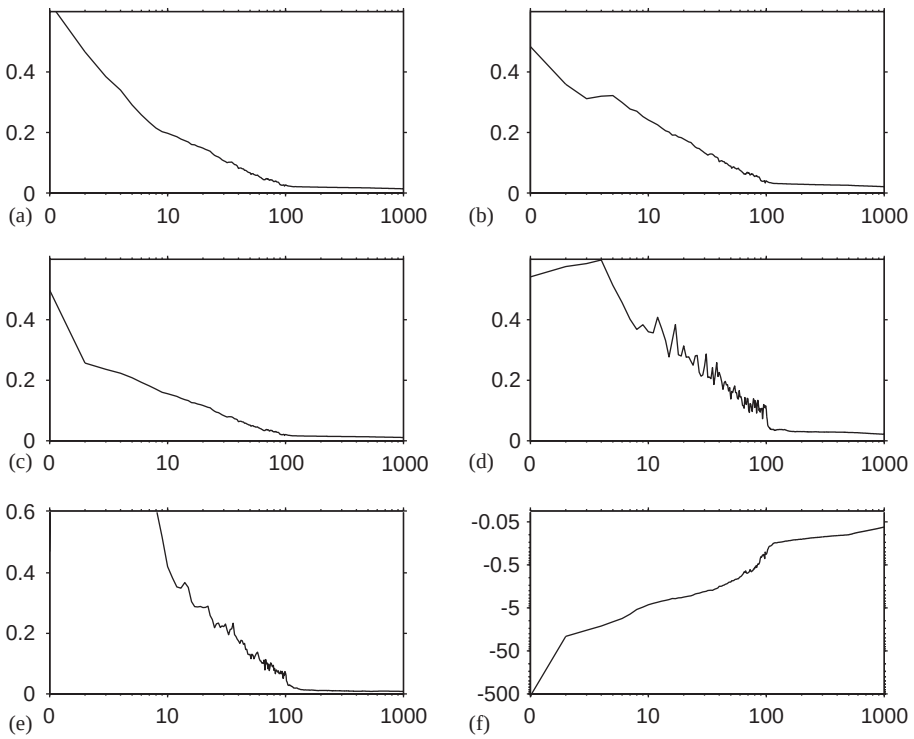

Fig. 5. (a–e) The five components of the mean deviation between $\boldsymbol{\theta}_{\boldsymbol{k}}^{\mathrm{SAEM}}$ and $\widehat{\boldsymbol{\theta}}^{\mathrm{MLE}}$. (f) The mean deviation between the log-likelihood $\log q(\boldsymbol{y}; \boldsymbol{\theta}_{\boldsymbol{k}}^{\mathrm{SAEM}})$ and the maximum log-likelihood $\log q(\boldsymbol{y}; \widehat{\boldsymbol{\theta}}^{\mathrm{MLE}})$.

We also estimated the expected absolute deviation between the log-likelihood $\log q(\boldsymbol{y}; \boldsymbol{\theta}_{\boldsymbol{k}}^{\mathrm{SAEM}})$ and the maximum log-likelihood $\log q(\boldsymbol{y}; \widehat{\boldsymbol{\theta}}^{\mathrm{MLE}})$ by

$$\mathrm{ADEV}_k(l) = \frac{1}{50} \sum_{\ell=1}^{50} (\log q(\boldsymbol{y}; \boldsymbol{\theta}_{\boldsymbol{k}}^{(\ell)}) - \log q(\boldsymbol{y}; \widehat{\boldsymbol{\theta}}^{\mathrm{MLE}})).$$

The results are displayed in Fig. 5.

We clearly see that a good estimation is expected after only 100 iterations. Indeed, only 100 iterations are enough if a relative precision of about 5% is required in the estimation of $\boldsymbol{\theta}$.

3.3. Extension to an heteroscedastic model

Let us consider now the following heteroscedastic model for this same example:

$$y_{ij} = \frac{\phi_i}{1 + \exp\left(-\frac{x_j - \beta_1}{\beta_2}\right)} (1 + \varepsilon_{ij}). \tag{23}$$

Table 3
Heteroscedastic model: mean value and standard deviation of the 50 estimations of $\boldsymbol{\theta}$, after 1000 iterations of SAEM

Parameters	β_1	β_2	μ	τ^2	σ^2
$\tau_1^2=\tau_2^2=10$					
Mean value	757.29	378.78	197.50	722.48	8.5×10^{-3}
Standard deviation	11.80	4.96	2.18	17.61	3×10^{-5}
$\tau_1^2=\tau_2^2=1000$					
Mean value	756.43	379.45	195.65	724.28	8.9×10^{-3}
Standard deviation	13.59	5.18	2.59	19.13	4×10^{-5}

We are outside the scope of the exponential model and SAEM cannot be used as before. The solution consists in regarding the fixed parameters (β_1, β_2) as the realization of a Gaussian random vector of mean (μ_1, μ_2) and diagonal covariance matrix with diagonal terms (τ_1^2, τ_2^2). As before, $\boldsymbol{\phi} = (\phi_i)$ is a sequence of i.i.d. Gaussian random variables of mean μ and variance τ^2.

It is important to notice that we do not change the model by doing that. The fixed effects remain fixed effects, since we still consider only one vector (β_1, β_2) for the all population.

The complete likelihood of this model is

$$\begin{aligned} f(\boldsymbol{y}, \boldsymbol{\phi}; \boldsymbol{\theta}) = {} & 2\pi\tau_1\tau_2(2\pi\sigma^2)^{-nm/2}(2\pi\tau^2)^{-n/2} \\ & \times \exp\left[-\frac{1}{2\sigma^2}\sum_{i,j}\left(\frac{y_{ij}}{g(\phi_i, \beta_1, \beta_2, x_j)} - 1\right)^2 - \sum_{i,j}\log(g(\phi_i, \beta_1, \beta_2, x_j)) \right. \\ & \left. - \sum_{i=1}^{n}\frac{(\phi_i-\mu)^2}{2\tau^2} - \frac{(\beta_1-\mu_1)^2}{2\tau_1^2} - \frac{(\beta_2-\mu_2)^2}{2\tau_2^2} \right], \end{aligned} \tag{24}$$

where $\boldsymbol{\theta} = (\mu_1, \mu_2, \tau_1^2, \tau_2^2, \mu, \tau^2, \sigma^2)$ is the new vector of hyperparameters. This model is now clearly exponential and we are able to apply the SAEM algorithm for estimating $(\mu_1, \mu_2, \mu, \tau^2, \sigma^2)$. Then we can use the estimation of (μ_1, μ_2) as estimation of (β_1, β_2).

We ran 50 times the SAEM algorithm (1000 iterations), with different random initial guess. The mean value and the standard deviation of the 50 estimations of $(\mu_1, \mu_2, \mu, \tau^2, \sigma^2)$ are displayed in Table 3 . It is interesting to remark that the estimation of the parameters is not influenced at all by the choice of (τ_1^2, τ_2^2). Indeed, the same results are obtained with $\tau_1^2=\tau_2^2=10$, or $\tau_1^2=\tau_2^2=1000$. The results of Table 3 clearly also show that convergence of SAEM does not depend on the initial value. Furthermore, the estimated values of (μ, β_1, β_2) are not very different from the values obtained with an additive model.

Table 4
Pharmacodynamic model: comparison of parameters estimates

Parameters	Exact	FOCE		LAP		EM		SAEM		$\hat{\sigma}(\widehat{\boldsymbol{\theta}}^{\mathrm{MLE}})$
μ_1	105	105.5	(1.8)	105.3	(1.6)	105.4	(1.7)	104.7	(1.5)	1.4
μ_2	12	12.2	(1.2)	12.4	(1.3)	12.3	(1.3)	11.8	(1.3)	1.0
μ_3	10	9.0	(2.8)	9.7	(2.4)	9.7	(1.6)	10.1	(0.9)	0.6
τ_1^2	64	59.7	(20.9)	58.4	(20.2)	60.0	(20.1)	62.1	(16.6)	14.5
τ_2^2	36	31.5	(11.0)	30.7	(11.9)	30.9	(10.7)	34.4	(10.8)	7.6
τ_3^2	12.25	6.6	(7.6)	13.3	(6.1)	10.1	(2.9)	11.2	(3.0)	2.8

The means and the estimated square roots of the MSE's between parenthesis, based on 50 simulations.

4. Comparisons with other methods

4.1. A pharmacodynamic model

We consider in this section the nonlinear population pharmacodynamic model used by Walker (1996), for comparing the MLEs obtained with EM algorithm to approximate MLEs obtained from the package NONMEM.

Simulated data are given by

$$y_{ij} = \phi_{1i} - \frac{\phi_{2i} x_j}{\phi_{3i} + x_j} + \varepsilon_{ij}, \quad 1 \leqslant i \leqslant n, \ 1 \leqslant j \leqslant m \tag{25}$$

with $n = 30$, $m = 6$, $x_1 = 0$, $x_2 = 5$, $x_3 = 10$, $x_4 = 20$, $x_5 = 40$ and $x_6 = 80$. The random effects and the additive noise are simulated with Gaussian distributions:

$$\phi_{1i} \sim_{\text{i.i.d.}} \mathcal{N}(105, 64), \quad \phi_{2i} \sim_{\text{i.i.d.}} \mathcal{N}(12, 36),$$
$$\phi_{3i} \sim_{\text{i.i.d.}} \mathcal{N}(10, 12.25), \quad \varepsilon_{ij} \sim_{\text{i.i.d.}} \mathcal{N}(0, 4).$$

According to Sheiner et al. (1991) and Walker (1996), this model can be used for the analysis of blood pressure $\boldsymbol{y}$ as a function of the dose d of an anti-hypertensive drug from a longitudinal study.

Walker (1996) compares different popular methods of estimation, such as first-order conditional estimation (FOCE) and LAPLACIAN methods of NONMEM. He computes the means and the standard errors (as the square roots of the MSE's) for these different estimators. Table 4 reproduces these values, with also the standard errors obtained with SAEM. We see that the EM algorithm of Walker and SAEM give similar results, but it is important here to remark that only 300 iterations of SAEM are performed for each single simulated data set. Then, computing time for SAEM is very much reduced, in comparison of the Monte-Carlo EM that requires to sample 10,000 random variates at each iteration and converges very slowly.

Table 4 also gives the estimation of the standard deviation of the MLE, using the approach proposed in Section 2.2.5. This method seems to be very accurate, since these values are close to the empirical standard deviation computed from the 50 simulations.

Table 5
Pharmacokinetic model: comparison of parameters estimates

Parameters	Exact	FOCE		Gauss		SPML		SAEM	
μ_1	20	20.08	(1.72)	19.88	(0.58)	20.38	(0.82)	20.01	(0.52)
μ_2	0.5	0.50	(0.09)	0.50	(0.04)	0.51	(0.06)	0.50	(0.01)
Γ_{11}	4	5.11	(7.03)	6.76	(4.78)	4.21	(2.96)	3.33	(1.64)
$\Gamma_{12} \times 100$	5.74	−4.38	(42.01)	13.20	(12.19)	−3.65	(15.40)	5.67	(3.67)
$\Gamma_{22} \times 1000$	3.28	10.00	(24.70)	3.78	(6.20)	9.61	(11.84)	3.40	(1.11)
σ^2	16	15.24	(1.70)	15.93	(2.05)	15.70	(1.73)	16.46	(1.78)

The means and the estimated square roots of the MSE's between parenthesis, based on 20 simulations.

4.2. A pharmacokinetic model

This example was proposed by Concordet and Nunez (2002). It is a kinetic population homoscedastic model, used for example for analyzing the concentration obtained after a constant drug diffusion. The data were simulated according to

$$y_{ij} = \phi_{i1}(1 - \exp[-\phi_{i2} t_j]) + \varepsilon_{ij},$$

where y_{ij} denotes the concentration on time $t_j = j$ for $1 \leqslant j \leqslant n = 7$ for individual i for $1 \leqslant i \leqslant m = 30$. The random effects $\boldsymbol{\phi_i}$ have a Gaussian distribution with mean $\boldsymbol{\mu} = (20, 0.5)$ and covariance matrix $\boldsymbol{\Gamma}$ with

$$\boldsymbol{\Gamma} = \begin{pmatrix} 4 & 0.0574 \\ 0.0574 & 0.00328 \end{pmatrix}.$$

The errors (ε_{ij}) are independent and follow a Gaussian distribution with mean zero and variance $\sigma^2 = 16$.

Following Concordet and Nunez (2002), the starting values of the parameters are chosen equal to their true value. Indeed, they do not study the optimization method, but the performances of the different estimates.

In their paper, Concordet and Nunez (2002) estimate the parameters of the model by maximizing a simulated pseudo-likelihood. We reproduce in Table 5 their simulation results, comparing FOCE, Gaussian quadrature and their simulated pseudo maximum likelihood (SPML) method. These results are compared with the SAEM algorithm. Here, the number of iterations in SAEM was taken to be 300.

For each algorithm, the means and the square roots of the MSE's were estimated, based on 20 simulations. We clearly see that the maximum likelihood estimate, obtained with SAEM has the smallest MSE's, for the mean parameter but also for the variance-covariance matrix. Indeed, Concordet and Nunez (2002) have shown that the SPML estimator is consistent. Their method provides a good estimation for large values of n, but not for a small number of individuals.

References

Concordet, D., Nunez, O.G., 2002. A simulated pseudo-maximum likelihood estimator for nonlinear mixed models. Comput. Statist. Data Anal. 39 (2), 187–201.

Delyon, B., Lavielle, M., Moulines, E., 1999. Convergence of a stochastic approximation version of the EM algorithm. Ann. Statist. 27 (1), 94–128.

Dempster, A.P., Laird, N.M., Rubin, D.B., 1977. Maximum likelihood from incomplete data via the EM algorithm. J. Roy. Statist. Soc. Ser. B 39 (1), 1–38 (with discussion).

Kuhn, E., Lavielle, M., 2004. Coupling a stochastic approximation version of EM with a MCMC procedure. ESAIM Probability and Statistics 8, 115–131.

Lavielle, M., Moulines, E., 1997. A simulated annealing version of the EM algorithm for non-Gaussian deconvolution. Statist. Comput. 7 (4), 229–236.

Louis, T.A., 1982. Finding the observed information matrix when using the EM algorithm. J. Roy. Statist. Soc. Ser. B 44 (2), 226–233.

Pinheiro, J.C., Bates, D.M., 2000. Mixed-Effects Models in S and S-PLUS. Springer, Berlin.

Racine-Poon, A., 1985. A Bayesian approach to nonlinear random effects models. Biometrics 41 (4), 1015–1023.

Sheiner, L., Hashimoto, Y., Beal, S., 1991. A simulation study comparing designs for dose ranging. Statist. Med. 10, 303–321.

Vonesh, E.F., 1996. A note on the use of Laplace's approximation for nonlinear mixed-effects models. Biometrika 83 (2), 447–452.

Wakefield, J., 1996. The Bayesian analysis of population pharmacokinetic models. J. Amer. Statist. Assoc. 91 (433), 62–75.

Wakefield, J., Smith, A., Racine-Poon, A., Gelfand, A., 1994. Bayesian analysis of linear and nonlinear population models by using the Gibbs sampler. J. Roy. Stat. Soc. Ser. C 43 (1), 201–221.

Walker, S., 1996. An EM algorithm for nonlinear random effects models. Biometrics 52 (3), 934–944.

Wei, G.C.G., Tanner, M.A., 1990. A Monte Carlo implementation of the EM algorithm and the Poor's Man's data augmentation algorithms. J. Amer. Statist. Assoc. 85 (411), 699–704.

Wu, C.-F.J., 1983. On the convergence properties of the EM algorithm. Ann. Statist. 11 (1), 95–103.