Extension of the SAEM algorithm to left-censored data in nonlinear mixed-effects model: Application to HIV dynamics model

Adeline Samson, Marc Lavielle, France Mentré

Computational Statistics & Data Analysis 2006;51(3):1562-1574

Extension of the SAEM algorithm to left-censored data in nonlinear mixed-effects model: application to HIV dynamics model

Adeline Samson[1], Marc Lavielle[2], France Mentr[1]

[1] INSERM U738, Paris, France; University Paris 7, Paris, France
[2] Paris-Sud University, Laboratoire de Mathématiques, Orsay, France

Abstract

The reduction of viral load is frequently used as a primary endpoint in HIV clinical trials. Non-linear mixed-effects models are thus proposed to model this decrease of the viral load after initiation of treatment and to evaluate the intra- and inter-patient variability. However, left censoring due to quantification limits in the viral load measurement is an additional challenge in the analysis of longitudinal HIV data. An extension of the Stochastic Approximation Expectation-Maximization (SAEM) algorithm is proposed to estimate parameters of these models. This algorithm includes the simulation of the left-censored data in a right-truncated Gaussian distribution. Simulation results show that the proposed estimates are less biased than the usual naive methods of handling such data: omission of all censored data points, or imputation of half the quantification limit to the first point below the limit and omission of the following points. The viral load measurements obtained in the TRIANON-ANRS81 clinical trial are analyzed with this method and a significant difference is found between the two treatment groups of this trial.

KEYWORDS: Accept-reject algorithm; HIV dynamics; Left-censored data; MCMC algorithm; Nonlinear mixed-effects models; SAEM algorithm

1 Introduction

HIV viral load is a widespread marker of the evolution of HIV infected patients (1); the reduction in HIV viral load is frequently used as the primary endpoint in clinical trials to evaluate the efficacy of anti-viral treatments (see for example 2; 3; 4; 5; 6; 7; 8). Non-linear mixed-effects models (NLMEM) can be used in these longitudinal studies to exploit the richness of the dynamics seized by repeated measurements and to account for inter- and intra-patient variability in viral load measurements. In addition, understanding the mechanism of the large inter-patient variability may help in making appropriate clinical decisions and providing individualized treatment. Unfortunately, all available assays of viral load measurements have a low limit of quantification (LOQ), generally between 20 and 400 copies/ml. Besides, the proportion of subjects with a viral load below LOQ has increased with the introduction of highly active antiretroviral treatments. Working with such left-censored data complicates the study of longitudinal viral load data. This issue is common in other longitudinal studies with LOQ, such as pharmacokinetics or pharmacodynamics, which also widely use NLMEM.

This paper aims to develop a reliable inference based on maximum likelihood (ML) theory for HIV dynamics models with left-censored viral load and NLMEM. It is indeed important to obtain reliable estimates of the viral dynamic parameters, that can be used to evaluate antiviral therapies through comparison of treatment groups.

To address the estimation problem in longitudinal data analysis containing censored values, naive procedures such as omitting the censored data or imputing a fixed value (e.g., the quantification limit or half the limit) are combined with usual estimation methods of mixed models (see Beal (9) for a comparison of classical procedures in NLMEM). However, the statistical properties of such procedures are unclear. More inventive approaches propose multiple imputations of the censored values, by substituting a reasonable guess for each missing value. For example, in linear mixed models, Hughes (10) proposes a Monte-Carlo version of the Expectation Maximization (EM) algorithm (11), taking into account the censored values as missing data. Hughes (10) shows that his approach significantly reduces the bias

associated with naive imputation procedures. Jacqmin-Gadda et al. (5) propose a direct maximization of the likelihood using an iterative process for linear mixed models as well, including an autoregressive error model. They combine two optimization algorithms, the Simplex and the Marquardt algorithms.

For nonlinear mixed models, the problem is more complex, the estimation of such model parameters being difficult even without censored observations. Indeed, because of the non linearity of the regression function in the random effects, the likelihood of NLMEM cannot be expressed in a closed form. Consequently, several authors propose some widely used likelihood approximation methods, such as linearization algorithms, which are implemented in the NONMEM software and in the nlme function of Splus and R software (12; 13); or Laplacian or Gaussian quadrature algorithms, which are implemented in the NLMIXED Macro of SAS (14). Wu and Wu propose a multiple imputation method for missing covariates in NLMEM based on a linearization algorithm (15). However, none of these algorithms based on likelihood approximation can be considered as fully established theoretically. A different point of view can be taken, the individual parameters and the censored values being considered as non-observed data. The EM algorithm is then the most adapted tool to estimate incomplete data models. Because of the nonlinearity of the model, stochastic versions of the EM algorithm are proposed. Wu (16; 17) introduces MCEM algorithms, with a Monte Carlo approximation of the Expectation step, adapted to both NLMEM and the censoring problem of observations and covariates. This Monte Carlo implementation is based on samples independently and identically distributed from the conditional density, requiring Markov Chain Monte Carlo (MCMC) procedures. The replication choice of the Monte Carlo sample is a central issue to guarantee convergence and this remains an open problem. Wu proposes an "exact" MCEM (17) but emphasizes that this MCEM algorithm is very slow to converge. Indeed simulations of these large samples at each iteration are time consuming. To address this computational problem, Wu also proposes an approximate MCEM (16; 17) using a linearization of the model leading to an approximate ML method.

As an alternative to address both the point-wise convergence and the computational problem, stochastic approximation versions of EM (SAEM) are proposed for NLMEM with no censored-values (18; 19). This algorithm requires a simulation of only one realization of the missing data at each iteration, avoiding the computational difficulty of independent sample sim-

ulation occurring in the MCEM and shortening the time to simulate. In addition, point-wise almost sure convergence of the estimate sequence to a local maximum of the likelihood is proved by Delyon et al. (18) under conditions satisfied by models from the exponential family. Girard and Mentré (20) propose a comparison of these estimation methods in NLMEM using a blind analysis, showing the accuracy of the SAEM algorithm in comparison with other methods. Especially, the computational convergence of the SAEM algorithm is clearly faster than those of the MCEM algorithm. However, this current SAEM algorithm is only appropriate for NLMEM without censored-values.

The first objective of the present paper is thus to extend the SAEM algorithm to handle left-censored data in NLMEM as an exact ML estimation method. We include in the extended SAEM algorithm the simulation of the left-censored data with a right-truncated Gaussian distribution. We prove the convergence of this extended SAEM algorithm under general conditions. The second objective of this paper is to illustrate this algorithm with a simulation study in the HIV dynamics context. Furthermore, we compare the extended SAEM algorithm with more classical approaches to handle left-censored data such as omission or imputation of the censored data, on the same simulation study.

After describing the model and the notations (Section 2), Section 3 describes the extended SAEM algorithm. Section 4 reports the simulation study and its results. We simulate datasets using the bi-exponential model for HIV dynamics proposed by Ding and Wu (6), and evaluate the statistical properties of the extended SAEM parameter estimates and the classical approaches. Particularly, we evaluate 2 comparison group tests, the Wald test and the likelihood ratio test, provided by the SAEM algorithm. We then apply the extended SAEM algorithm to the TRIANON-ANRS 81 clinical trial of HIV treatment in Section 5. The aim of this new analysis of the TRIANON data is to show the ability of NLMEM to describe the evolution of the viral load and to test a treatment's effect between the 2 treatment groups, in the presence of left-censored observations. Section 6 concludes the article with some discussion.

2 Models and notations

Let us define $y_i = (y_{i1}, \ldots, y_{in_i})^t$ where y_{ij} is the response value for individual i at time t_{ij}, $i = 1, \ldots, N$, $j = 1, \ldots, n_i$, and let $y = (y_1, \ldots, y_N)$. Let us define an NLMEM as follows

$$\begin{aligned} &y_{ij} = f(\phi_i, t_{ij}) + g(\phi_i, t_{ij})\, \varepsilon_{ij}, \\ &\varepsilon_i \sim \mathcal{N}(0, \sigma^2 I_{n_i}) \quad \text{and} \\ &\phi_i = X_i\mu + b_i, \text{with} \;\; b_i \sim \mathcal{N}(0, \Omega), \end{aligned}$$

where $f(\cdot)$ and/or $g(\cdot)$ are nonlinear functions of ϕ_i, $\varepsilon_i = (\varepsilon_{i1}, \ldots, \varepsilon_{in_i})^t$ represents the residual error, ϕ_i is a p-vector of individual parameters, μ is the $k \times p$-matrix of fixed effects, X_i is the k-vector of known covariates, b_i is a p-vector of random effects independent of ε_i, σ^2 is the residual variance, I_{n_i} the identity matrix of size n_i and Ω quantifies the covariance of the inter-individual random effects.

Because of assay limitation, when data y_{ij} are inferior to the limit of quantification (LOQ), we do not observe y_{ij} but only the censored value LOQ. These data are usually named left-censored data. Let denote $I_{obs} = \{(i,j)|y_{ij} \geq LOQ\}$ and $I_{cens} = \{(i,j)|y_{ij} \leq LOQ\}$ the index sets of the uncensored and censored observations respectively. For $(i,j) \in I_{cens}$, let $y_{ij}^{cens} = y_{ij}$ denote the unknown value of the censored observation j of subject i. Let denote y_i^{cens} the vector of censored observations of subject i. Finally, we observe

$$y_{ij}^{obs} = \begin{cases} y_{ij} & \text{if} \quad (i,j) \in I_{obs}, \\ LOQ & \text{if} \quad (i,j) \in I_{cens}. \end{cases}$$

We denote $y_i^{obs} = (y_{i1}^{obs}, \ldots, y_{in_i}^{obs})$ as the observations of subject i and $y^{obs} = (y_1^{obs}, \ldots, y_N^{obs})$ the total observations dataset.

The maximum likelihood estimation is based on the log-likelihood function $L(y^{obs}\,; \theta)$ of the response y^{obs} with $\theta = (\mu, \Omega, \sigma^2)$ the vector of all the parameters of the model

$$L(y^{obs}\,; \theta) = \log\left(\prod_{i=1}^{N} \int p(y_i^{obs}, y_i^{cens}, \phi_i; \theta)\, d\phi_i\, dy_i^{cens}\right), \tag{1}$$

where $p(y_i^{obs}, y_i^{cens}, \phi_i;\, \theta\,)$ is the likelihood of the complete data $(y_i^{obs}, y_i^{cens}, \phi_i)$ of the i-th subject. Because the random effects ϕ_i and the censored observations y_i^{cens} are unobservable and the regression functions are nonlinear, the

foregoing integral has no closed form. The complete likelihood of the i-th subject is equal to:

$$p(y_i^{obs}, y_i^{cens}, \phi_i; \theta) = \prod_{(i,j)\in I_{obs}} p(y_{ij}^{obs}|\phi_i;\theta)p(\phi_i;\theta) \prod_{(i,j)\in I_{cens}} p(y_{ij}^{cens}|\phi_i;\theta)p(\phi_i;\theta),$$

with

$$\begin{array}{rl} p(y_{ij}^{obs}|\phi_i;\theta) = \pi(y_{ij}^{obs}; f(\phi_i,t_{ij}), \sigma^2 g^2(\phi_i,t_{ij}))\ \mathbb{1}_{y_{ij}\geq LOQ}, & \text{if} \quad (i,j)\in I_{obs} \quad \text{and} \\ p(y_{ij}^{cens}|\phi_i;\theta) = \pi(y_{ij}^{cens}; f(\phi_i,t_{ij}), \sigma^2 g^2(\phi_i,t_{ij}))\ \mathbb{1}_{y_{ij}\leq LOQ}, & \text{if} \quad (i,j)\in I_{cens}, \end{array}$$

where $\pi(x; m, v)$ is the probability density function of the Gaussian distribution with mean m and variance v, evaluated at x.

3 Estimation algorithm

3.1 The SAEM algorithm

The EM algorithm introduced by Dempster et al. (11) is a classical approach to estimate parameters of models with non-observed or incomplete data. Let us briefly cover the EM principle. Let z be the vector of non-observed data. The complete data of the model is (y, z). The EM algorithm maximizes the $Q(\theta|\theta') = E(L_c(y, z; \theta)|y; \theta')$ function in 2 steps, where $L_c(y, z; \theta)$ is the log-likelihood of the complete data. At the m-th iteration, the E step is the evaluation of $Q_m(\theta) = Q(\theta \,|\, \widehat{\theta}_{m-1})$, whereas the M step updates $\widehat{\theta}_{m-1}$ by maximizing $Q_m(\theta)$. For cases in which the E step has no analytic form, Delyon et al. (18) introduce a stochastic version SAEM of the EM algorithm which evaluates the integral $Q_m(\theta)$ by a stochastic approximation procedure. The authors prove the convergence of this algorithm under general conditions if $L_c(y, z; \theta)$ belongs to the regular curved exponential family

$$L_c(y, z; \theta) = -\Lambda(\theta) + \langle S(y, z), \Phi(\theta)\rangle,$$

where $\langle ., . \rangle$ is the scalar product, Λ and Φ are 2 functions of θ and $S(y, z)$ is the minimal sufficient statistic of the complete model. The E step is then divided into a simulation step (S step) of the missing data $z^{(m)}$ under the conditional distribution $p(z|y; \widehat{\theta}_{m-1})$ and a stochastic approximation step (SA step) using $(\gamma_m)_{m\geq 0}$ a sequence of positive numbers decreasing to 0. This SA

step approximates $E\left[S(y,z)|\widehat{\theta}_{m-1}\right]$ at each iteration by the value s_m defined recursively as follows

$$s_m = s_{m-1} + \gamma_m(S(y, z^{(m)}) - s_{m-1}).$$

The M step is thus the update of the estimates $\widehat{\theta}_{m-1}$

$$\widehat{\theta}_m = \arg\max_{\theta\in\Theta} \left(-\Lambda(\theta) + \langle s_m, \Phi(\theta)\rangle\right).$$

Let us detail the sufficient statistics needed for evaluation at the SA step of the extended SAEM algorithm for the nonlinear mixed models previously presented. The sufficient statistics are $S^{(1)} = \sum_{i=1}^{N} \phi_i$, $S^{(2)} = \sum_{i=1}^{N} \phi_i^2$ and $S^{(3)} = \sum_{i,j}(y_{ij} - f(\phi_i, t_{ij}))^2$, where $y_{ij} = y_{ij}^{obs}$ if $(i,j) \in I_{obs}$ and $y_{ij} = y_{ij}^{cens}$ if $(i,j) \in I_{cens}$. Therefore, at the m-th iteration of SAEM the M-step reduces to

$$\begin{aligned}
\widehat{\mu}_m &= \frac{1}{N} s_m^{(1)}, \\
\widehat{\omega^2}_m &= \frac{1}{N} s_m^{(2)} - (s_m^{(1)})^2 \quad \text{and} \\
\widehat{\sigma^2}_m &= \frac{1}{NJ} s_m^{(3)}.
\end{aligned}$$

In cases in which the simulation of the non-observed vector z cannot be directly performed, Kuhn and Lavielle (19) propose to combine this algorithm with a Markov Chain Monte Carlo (MCMC) procedure. The convergence of this SAEM algorithm is ensured under general conditions; the 2 main conditions are presented below (see Kuhn and Lavielle 21, for technical conditions)

(SAEM 1) For any $\theta \in \Theta$, the Gibbs algorithm generates a uniformly ergodic chain which invariant probability is $p(z|y;\theta)$.

(SAEM 2) For all m in the integer set $\mathbb{N}^*$, $\gamma_m \in [0,1]$, $\sum_{m=1}^{\infty} \gamma_m = \infty$ and $\sum_{m=1}^{\infty} \gamma_m^2 < \infty$.

For NLMEM with left-censored data, the nonobserved vector is $z = (\phi, y^{cens})$, with $\phi = (\phi_1, \ldots, \phi_N)$ being the individual parameters vector and $y^{cens} = (y_1^{cens}, \ldots, y_N^{cens})$ the left-censored data vector. The S step of the SAEM algorithm is the simulation of the missing data (ϕ, y^{cens}) under the posterior distribution $p(\phi, y^{cens}|y^{obs};\theta)$. This step can be performed by use of a Gibbs sampling algorithm. At the m-th iteration of the SAEM algorithm, the Gibbs algorithm is thus divided into 2 steps

1. Simulation of $\phi^{(m)}$ by use of a Metropolis-Hastings (M-H) algorithm constructing a Markov Chain $\phi^{(m)}$ with $p(\cdot|y^{obs}, y^{cens(m-1)}; \widehat{\theta}_{m-1})$ as the unique stationary distribution,

2. Simulation of $y^{cens(m)}$ with the posterior right-truncated Gaussian distribution $p(\cdot|y^{obs}, \phi^{(m)}; \widehat{\theta}_{m-1})$.

Consequently, under assumptions (SAEM1) and (SAEM2) and general additional conditions, by applying the convergence theorem of Kuhn and Lavielle (19) the estimate sequence $(\widehat{\theta}_m)_{m\geq 0}$ produced by the extended SAEM algorithm converges towards a (local) maximum of the likelihood $L(y^{obs}; .)$.

Samson et al. (22) propose to estimate the likelihood function with use of an importance sampling procedure and detail its implementation. They estimate the Fisher information matrix combining a stochastic approximation approach and the Louis' missing information principle (23): the Hessian of the log-likelihood of the observed data can be obtained almost directly from the simulated missing data (see Kuhn and Lavielle 19, for more implementation details). We adapt their estimates of the likelihood and the Fisher information matrix to the extended SAEM algorithm, to implement the 2 comparison group tests, the Wald test and the likelihood ratio test.

3.2 Computational aspects

The convergence of the SAEM algorithm is ensured under the 2 assumptions (SAEM 1) and (SAEM 2), which require careful choices of the implementation of the Gibbs algorithm and the stochastic approximation step size respectively.

3.2.1 Gibbs algorithm

The convergence of the Gibbs algorithm depends on the M-H algorithm generating ϕ and the simulation method generating y^{cens}.

At the m-th iteration of the SAEM algorithm, the M-H algorithm proceeds as follows: a candidate ϕ^c is simulated with a proposal distribution $q_{\widehat{\theta}_{m-1}}$. The candidate is accepted (i.e. $\phi^{(m)} = \phi^c$), with the acceptation probability ρ

$$\rho = \min\left(\frac{p(\phi^c|y^{obs}, y^{cens(m-1)}; \widehat{\theta}_{m-1})}{p(\phi^{(m-1)}|y^{obs}, y^{cens(m-1)}; \widehat{\theta}_{m-1})} \frac{q_{\widehat{\theta}_{m-1}}(\phi^c|\phi^{(m-1)})}{q_{\widehat{\theta}_{m-1}}(\phi^{(m-1)}|\phi^c)}, 1\right),$$

and the candidate is rejected (i.e. $\phi^{(m)} = \phi^{(m-1)}$), with probability $1 - \rho$.

We propose the 3 following proposal distributions $q_{\widehat{\theta}_{m-1}}$ for the M-H procedure

1. $q^{(1)}_{\widehat{\theta}_{m-1}}$ is the prior distribution of ϕ, that is, the Gaussian distribution $\mathcal{N}(\widehat{\mu}_{m-1}, \widehat{\Omega}_{m-1})$,

2. $q^{(2)}_{\widehat{\theta}_{m-1}}$ is the multidimensional random walk $\mathcal{N}(\phi^{(m-1)}, \lambda\widehat{\Omega}_{m-1})$, where λ is a scaling parameter chosen to ensure a sufficient acceptation rate,

3. $q^{(3)}_{\widehat{\theta}_{m-1}}$ is a succession of p unidimensional Gaussian random walks: each component of ϕ is successively updated.

Then, an efficient simulation method has to be implemented to generate $y_{ij}^{cens(m)}$ for all $(i,j) \in I_{cens}$ with the right-truncated Gaussian distribution with mean $f(\phi_i^{(m)}, t_{ij})$, variance equal to $\widehat{\sigma^2}_{m-1}\, g^2(\phi_i^{(m)}, t_{ij})$ and truncated at the right by the value LOQ. We implement the accept-reject algorithm proposed by Robert (24) because of its simplicity and because it slightly improves upon previous algorithms developed by Gelfand and Smith (25). This algorithm is composed of the following steps

1. compute $\alpha = \frac{C+\sqrt{C^2+4}}{2}$,

2. simulate x with the translated exponential distribution $\mathcal{E}(\alpha, C)$ with density $p(x|\alpha, C) = \alpha \exp(-\alpha(x - C))\mathbb{1}_{x \geq C}$,

3. compute $\rho(x) = \exp(-(x - \alpha)^2/2)$,

4. simulate u with $\mathcal{U}_{[0,1]}$,

5. if $u \leq \rho(x)$, then keep x and compute $y_{ij}^{cens(m)} = f(\phi_i^{(m)}, t_{ij}) - x\widehat{\sigma}_{m-1} g(\phi_i^{(m)}, t_{ij})$,

   else return to step (2).

The simulation of x with the translated exponential distribution $\mathcal{E}(\alpha, C)$ in step (2) is performed by simulating u with a uniform distribution $\mathcal{U}[0,1]$ on the unit interval and then by computing $x = -\frac{1}{\alpha}\ln(1-u) + C$.

With the proposal distributions detailed below, this Gibbs algorithm converges and generates a uniformly ergodic chain with $p(\phi, y^{cens}|y^{obs}; \theta)$ as the stationary distribution, thus fulfilling the assumption (SAEM 1).

3.2.2 Stochastic approximation step size sequence

The sequence $(\gamma_m)_{m\geq 0}$ has to fulfill the assumption (SAEM 2). We recommend the use of $\gamma_m = 1$ during the first M_1 iterations $1 \leq m \leq M_1$, and $\gamma_m = (m - M_1)^{-1}$ during the last M_2 iterations. Indeed, the initial guess θ_0 may be far from the maximum likelihood value and the first iterations with $\gamma_m = 1$ allow for converging to a neighborhood of the maximum likelihood estimate. Furthermore, the inclusion of a hybrid Gibbs procedure (instead of a Metropolis-Hastings procedure in the classic SAEM algorithm) slows up the convergence of the extended SAEM algorithm. The convergence is monitored by graphical criterion. The choice of M_1 and M_2 values and $(\gamma_m)_{m\geq 0}$ are adapted according to the graphical convergence of all the parameter estimates.

4 Simulation study

4.1 Simulation settings

The first objective of this simulation study is to illustrate the main statistical properties of the extended SAEM algorithm in the context of HIV viral dynamics (bias, root mean square errors, group comparison tests). The second objective is to compare the extended SAEM algorithm to some of the classical approaches proposed to take into account a censoring process.

We use the bi-exponential model for initial HIV dynamics proposed by Ding and Wu (6) to simulate the datasets

$$f(\phi_i, t_{ij}) = \log_{10}(P_{1i}e^{-\lambda_{1i}t_{ij}} + P_{2i}e^{-\lambda_{2i}t_{ij}}).$$

This function is a simplified analytical solution of a differential system describing HIV viral load decrease during anti-retroviral treatment proposed by Perelson et al. (1). It has p=4 individual parameters: P_{1i}, P_{2i} are the baseline values and λ_{1i}, λ_{2i} represent 2-phase viral decay rates. These parameters are positive and distributed according to a log-normal distribution. Thus, ϕ_i and μ take the following values: $\phi_i = (\ln P_{1i}, \ln P_{2i}, \ln \lambda_{1i}, \ln \lambda_{2i})$ and $\mu = (\ln P_1, \ln P_2, \ln \lambda_1, \ln \lambda_2)$. We assume identical sampling times for all subjects: for all i in $1, \ldots, N$, $t_{ij} = t_j$ for $j = 1, \ldots, n$. Additive Gaussian random effects are assumed for each parameter with a diagonal covariance matrix Ω. Let $\omega^2 = (\omega_1^2, \omega_2^2, \omega_3^2, \omega_4^2)$ denote the vector of the variances of the

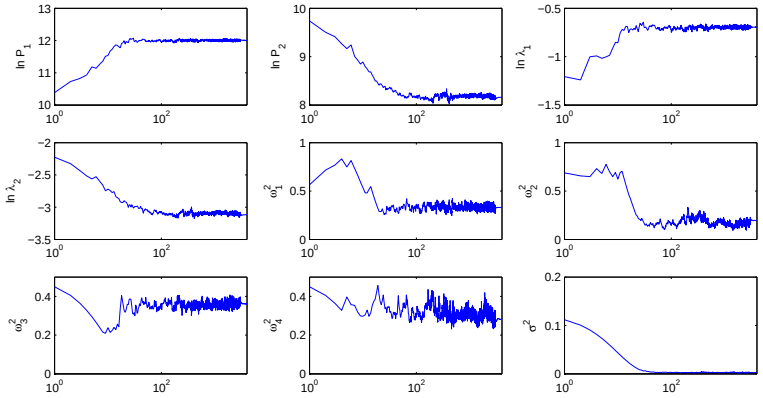

Figure 1: Convergence of the SAEM parameter estimates for one simulated dataset with $N = 40$ subjects (semi-log scale)

random effects. Additive Gaussian error is assumed with a constant variance σ^2 (i.e. $g(\phi_i, t_j) = 1$ for all i, j).

For the fixed effects, the values are those proposed by Ding and Wu (6): $\ln P_1 = 12$, $\ln P_2 = 8$, $\ln \lambda_1 = \ln(0.5)$, $\ln \lambda_2 = \ln(0.05)$. The inter-subject variability is identical for the 4 parameters: $\omega_1^2 = \omega_2^2 = \omega_3^2 = \omega_4^2 = 0.3$ corresponding to a variation coefficient of 55%, which is a realistic inter-subject variability in the context of HIV dynamics. We chose a variance $\sigma = 0.065$, which corresponds to a constant variation coefficient of 15% for the viral load. With the Matlab software, we generate N=40 total number of subjects with n=6 blood samples per patient, taken on days 1, 3, 7, 14, 28 and 56. We consider the same limit of quantification as Ding and Wu: $LOQ = \log_{10}(400) \approx 2.6$.

The convergence of the SAEM algorithm on a simulated dataset is illustrated in Figure 1. The initial estimates are arbitrarily chosen for all the parameters. During the first $M_1 = 3000$ iterations, the estimates converge to a neighborhood of the maximum likelihood. Then, smaller step sizes during $M_2 = 1000$ additional iterations ensure the almost sure convergence of the algorithm to the maximum likelihood estimate. We implement the extended SAEM algorithm in a Matlab function. It takes about 120 s for the extended SAEM algorithm to converge with 4000 iterations on a conventional Intel Pentium IV 2.8 GHz workstation.

The conditional expectation $E(y^{cens}|y^{obs})$ of the censored values can be

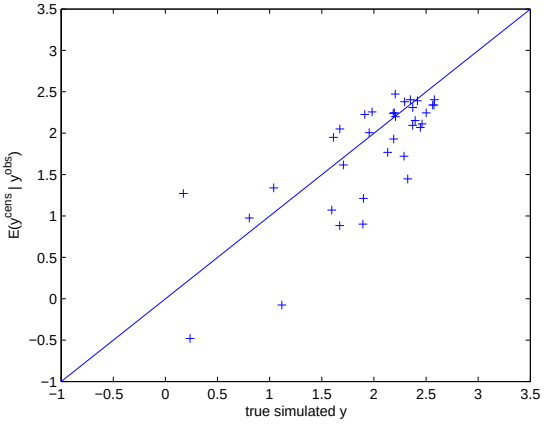

Figure 2: Expectation of the censored values $E(y^{cens}|y^{obs})$ evaluated by the extended SAEM algorithm as a function of the true simulated values y that are below the LOQ (2.6) on a simulated dataset.

evaluated from the posterior mean of the y^{cens} simulated during the last iterations of the extended SAEM algorithm. Figure 2 illustrates this evaluation on a simulated dataset: $E(y^{cens}|y^{obs})$ evaluated by SAEM is plotted as a function of the true simulated values y that are below the LOQ for this simulated dataset. The extended SAEM algorithm provides satisfactory expectation of these censored values.

4.2 Evaluation of estimates

Our aim is to evaluate and compare the estimates produced by the extended SAEM algorithm with those produced by 2 estimation approaches recommended in the presence of left-censored data. We fit the simulation model and compute the relative bias and relative root mean square error (RMSE) for each component of θ from 1000 replications of the trial described below.

We first assume that no censoring is present in the viral load. We estimate the datasets using the classical SAEM algorithm; this bias and RMSE are considered the benchmark for the comparison of the 3 methods on the censored datasets described below.

We then censor the simulated datasets by censoring observations that are below the LOQ. The censoring represents, on average, 0% of the observations

Table 1: Relative bias (%) and relative root mean square error (RMSE) (%) of the estimated parameters evaluated from 1000 simulated trials on the uncensored datasets (all data) with the SAEM algorithm and the left-censored datasets with 2 classic methods (M_1 and M_2) and with the extended SAEM algorithm (ML).

Parameters	Bias (%)				RMSE (%)			
	all data	left censored data			all data	left censored data		
		M_1	M_2	**ML**		M_1	M_2	**ML**
ln P_1	0.01	0.03	0.32	**0.03**	0.77	0.78	0.93	**0.77**
ln P_2	0.01	2.64	10.71	**0.23**	1.29	3.22	10.88	**1.63**
ln λ_1	0.98	2.67	12.94	**0.57**	12.47	12.55	19.76	**12.36**
ln λ_2	0.04	10.46	22.88	**0.62**	3.09	11.45	23.36	**3.98**
ω_1^2	0.28	3.69	37.51	**4.26**	24.17	26.55	49.60	**26.30**
ω_2^2	2.20	12.67	24.81	**6.21**	26.65	37.15	58.31	**37.70**
ω_3^2	1.97	6.85	12.53	**1.67**	22.48	23.03	31.01	**23.05**
ω_4^2	0.88	47.13	98.331	**6.59**	25.66	55.98	113.53	**36.85**
σ^2	0.51	10.31	440.77	**0.63**	16.34	26.24	453.24	**19.34**

at days 1 and 3, 0.07% at day 7, 2.81% at day 14, 26.96% at day 28 and 71.57% at day 56. First, we implement 2 classical approaches omitting or an imputing arbitrary value to the censored data. We name M_1 the naive approach, which omits all censored data. We then name M_2 the method recommended by several authors (9; 6; 26); for each patient, the first data below the LOQ is kept and imputed to LOQ/2, and then all the following censored data are omitted. We use the standard SAEM algorithm to fit the datasets for both the M_1 and M_2 methods. Second, we apply the extended SAEM algorithm presented in Section 3; this gives us the maximum likelihood (ML) estimates of the parameter θ from the original dataset y^{obs}.

The relative bias and RMSE obtained under the simulation model on the uncensored datasets with the classical SAEM algorithm are presented in Table 1 and referred as the "all data" estimates. These estimates have very small bias (<0.5% for the fixed effects, <5% for the variance parameters). The RMSE is really satisfactory for the fixed effects (<13%) and the variance parameters (<30%).

The relative bias and RMSE obtained on the censored datasets are presented in Table 1. Three of the fixed effects are estimated with bias by the

M_2 method, especially ln λ_2 (23%). The M_1 method reduces the bias for all the fixed effects but ln λ_2 still has a larger bias (10.5%) than before the censor. The bias of the variance parameters is increased with both the M_1 and M_2 methods, especially for ω_4^2 and σ^2 (47% and 98% for ω_4^2 and 10% and 440% of bias for σ^2 respectively). In contrast with these 2 methods, the extended SAEM algorithm provides estimates of all the parameters with small bias. The M_1 method gives a satisfactory RMSE except for λ_2 (11%) and ω_4^2 (56%). The M_2 method increases all the RMSE, especially for ω_4^2 (113%) and σ^2 (453.2%). The RMSE is satisfactory with the extended SAEM algorithm.

Every dataset is almost censored at days 28 or 56 during the second decay phase of the viral load decrease. This finding explains that the parameter estimates corresponding to this second decay rate (ln λ_2 and its variance ω_4^2) are the most affected by the censoring process. However, even with 71% of censoring at day 56, the bias and RMSE of the extended SAEM algorithm almost reach the uncensored dataset benchmark. This accuracy is also illustrated in Figure 3, which presents the distribution of the second decay rate parameters estimates (ln λ_2 and ln ω_4^2) for the 4 methods from the 1000 replications. This figure again points out the bad properties of the M_1 and M_2 methods, and reemphasizes that the extended SAEM algorithm reaches the exactness level of the estimation method applied to the uncensored datasets. The difference in the other parameters distributions between the 4 methods are similar.

The M_1 method provides estimates that are less biased than the M_2 method for all parameters. This finding can be explained by the design used for the simulation. The number of uncensored measurements is large enough to estimate quite accurately all the parameter by omitting all the censored data. Contrary to the M_1 method, which considers a partial dataset from the original dataset, the M_2 method is based on a modified partial dataset. Because the modification affects the data measured during the second decay, the parameter estimates of this second decay rate (ln λ_2 and its variance ω_4^2) are noticeably biased.

4.3 Application to group comparison

We consider that the subjects of each simulated trial belong to 2 different treatment groups of equal size (i.e. 20 subjects per group). We performed a Wald test and Likelihood ratio test (LRT) to test a difference between the treatment groups on the viral load decrease, especially on the first viral decay

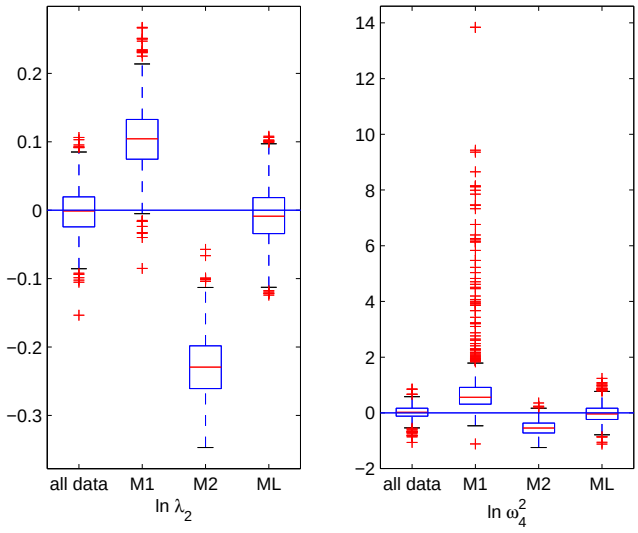

Figure 3: Boxplot of the second decay rate parameter estimates ($\ln \lambda_2$ and $\ln \omega_4^2$) for the 4 methods on the 1000 replications: the all-data method, the M_1 and M_2 methods, and ML, the extended SAEM algorithm.

rate, $\ln\lambda_1$, as proposed by Ding and Wu (6). We apply these tests using SAEM on the uncensored datasets and the extended SAEM algorithm on the censored datasets and evaluate their type I errors. We do not evaluate the type I errors obtained with the M_1 and M_2 methods on the censored datasets because the previous simulation study already illustrates their bad properties.

Let $G_i = 0$ denote a control treatment group subject and $G_i = 1$ an experiment treatment group subject. In this example, the vector of covariates X_i is $(1, G_i)$. Let β denote the treatment effect parameter on $\ln\lambda_1$.

$$\ln \lambda_{1i} = \ln \lambda_1 + \beta G_i.$$

In this case, the matrix of fixed effects is

$$\mu = \begin{pmatrix} \ln P_1 & \ln P_2 & \ln \lambda_1 & \ln \lambda_2 \\ 0 & 0 & \beta & 0 \end{pmatrix}.$$

We test by LRT or Wald test the hypothesis that the 2 treatments are equal, H_0: $\{\beta = 0\}$, versus the alternative hypothesis H_1: $\{\beta \neq 0\}$. Because the likelihood function is differentiable for every θ and the H_0 is locally

equivalent to a linear space, the LRT statistic is asymptotically chi-squared distributed. We thus compare the $2(L_1 - L_0)$ statistic with a 1 degree of freedom χ^2_1 distribution, where L_0 and L_1 are the log-likelihoods evaluated by importance sampling under H_0 and H_1, respectively. The importance sampling procedure is implemented by simulating a sample of size 5000 of the individual parameters ϕ_i with the Gaussian approximation of the posterior distribution, using estimates of the individual posterior mean $E(\phi_i|y_i^{obs})$ and the posterior variance $\mathrm{Var}(\phi_i|y_i^{obs})$ evaluated by the empirical mean and variance of the ϕ_i simulated by the SAEM algorithm during the last 500 iterations.

For the Wald test, the information Fisher matrix, whose inverse matrix's diagonal corresponds to the variance of the parameter estimates, is estimated by a stochastic approximation procedure during the iteration of the SAEM algorithm. We estimate the parameter $\hat{\beta}$ and its standard error $SE(\hat{\beta})$ under H_1. Under the hypothesis that likelihood is twice continuously differentiable for every θ, the Wald statistic is asymptotically chi-squared distributed. Therefore, we compare the statistic $\hat{\beta}^2/SE^2(\hat{\beta})$ with a χ^2_1 distribution. For both tests, the type I error is estimated by the proportion of trials for which H_0 is rejected as these datasets are simulated without any treatment effect.

The type I error of the Wald test is 4% for the classical SAEM algorithm on the uncensored datasets, and 5.9% for the LRT. We find similar results on the left-censored datasets using the extended SAEM algorithm. The type I error of the Wald test and the LRT are 4.1% and 5.4%, respectively using this algorithm. These again illustrate the good statistical properties of this extended SAEM algorithm.

5 Application to the Trianon (ANRS81) trial

We illustrate the extended SAEM algorithm on viral load data from the clinical trial TRIANON supported by the French Agence National de Recherche sur le Sida (ANRS). In this study, 144 patients infected with HIV-1, who were randomized into 2 treatment groups, undergo treatment for 72 weeks: 71 patients receive treatment A (lamivudine, d4T and indinavir) and 73 patients treatment B (nevirapine, d4T and indinavir). Viral load is measured at weeks 4 and 8 and every 8 weeks thereafter up through week 72. The HIV RNA assay used in this study has a limit of detection of 20 cp/ml. The

comparison of the log reduction of the viral load from baseline to week 72 between the 2 groups with use of a standard statistical approach involving intention to treat shows treatment A to be superior, although the authors expected a superiority of the 3-class regimen (treatment B). See Launay et al. (27) for a more complete description of the study design and results. The data are presented in Figure 4.

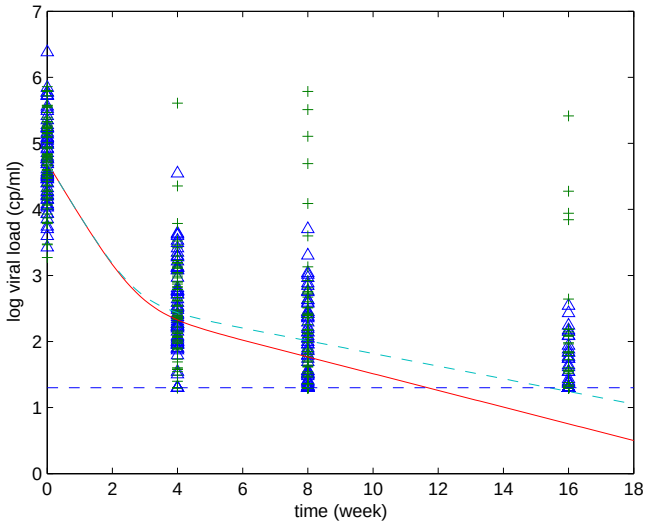

Figure 4: Observed individual viral load decreases in the 2 groups of patients of the TRIANON trial, with the predicted mean curves obtained with the extended SAEM algorithm in the 2 groups: (Δ), group A observations; (+), group B observations; plain line, group A prediction; dashed line, group B prediction; dotted line, LOQ level.

This new analysis of TRIANON data aims to evaluate the treatment effects on the evolution of the initial viral load decrease. We use the bi-exponential model presented in section 4 to fit the $\log_{10}$ viral load measurements until week 16. There are 64 (out of 275) and 65 (out of 281) observations, respectively, below the LOQ in group A and group B. We compare the extended SAEM algorithm with the usual M_2 method, the one recommended by Ding and Wu (6) to handle left-censored data. For both methods, we analyze the model under the null hypothesis (i.e. without treatment effect). We analyze then the 3 alternative hypotheses proposed by Ding and Wu: AH_1:

$\{\beta_{\lambda_1} \neq 0\}$; AH_2: $\{\beta_{\lambda_2} \neq 0\}$ and AH_3: $\{\beta_{\lambda_1} \neq 0\ \beta_{\lambda_2} \neq 0\}$. In group B, β_{λ_1} and β_{λ_2} are treatment effects added to $\ln\lambda_1$ and $\ln\lambda_2$ in group A

$$\begin{aligned} \ln \lambda_{1i} &= \ln \lambda_1 + \beta_{\lambda_1} G_i \quad \text{and} \\ \ln \lambda_{2i} &= \ln \lambda_2 + \beta_{\lambda_2} G_i, \end{aligned}$$

where $G_i = 0$ denotes a group A subject and $G_i = 1$ a group B subject. We use the one-dimensional Wald test to assess the AH_1 and AH_2 alternative hypotheses. We use a bi-dimensional Wald test for the two-dimensional vector $\beta = (\beta_{\lambda_1}, \beta_{\lambda_2})$ to assess the AH_3 hypothesis. We use the LRT to test all the nested models.

Using the M_2 method, the log-likelihoods are estimated at -617.24, -617.18, -617.0 and -616.92 under H_0, AH_1, AH_2 and AH_3, respectively. None of the 4 LRT is significant at 5%, and we find the same conclusions using the Wald test. With the extended SAEM algorithm, the log-likelihoods are estimated at -472.11, -467.59, -467.28 and -466.09 under H_0, AH_1, AH_2 and AH_3, respectively. The LRT are significant at 5%, except for the test of AH_2 vs AH_3. We find similar results using the Wald tests. Unsurprisingly, the likelihoods are not of the same order with both methods because they come from datasets with different numbers of observations: with the M_2 method, the left-censored data are omitted except for the first ones; with the extended SAEM algorithm, all the left-censored data are kept. The population parameter estimates (and their standard errors) of the final model under AH_2 for the extended SAEM algorithm are $\ln P_1$=10.8 (0.05), $\ln P_2$=6.39 (0.17), $\ln\lambda_1$=-1.30 (0.02), $\ln\lambda_2$=-3.18 (0.05), β_{λ_2}=-0.277 (0.08), ω_1^2=0.106 (0.03), ω_2^2=2.76 (0.46), ω_3^2=0.012 (0.01), ω_4^2=0.059 (0.02), and σ^2=0.38 (0.03). Figure 4 presents the curves predicted by this model, overlaid on the data. The censored data are plotted at the value LOQ. The predicted curves are below the LOQ at week 16 as the extended SAEM algorithm handles the censored data.

In conclusion, we find a significant difference between treatments using the extended SAEM algorithm but not with the recommended M_2 method. The superiority of the treatment A ($\beta_{\lambda_2} < 0$) is in concordance with the previous analysis of the TRIANON dataset (27). In addition, we are able to describe the evolution of the viral load and the treatments' effects. In our example, we find a trend for a faster viral load decrease under treatment A in the second phase.

6 Discussion

To analyse longitudinal data with left-censored responses, we propose a maximum likelihood estimation method that may be preferred over methods classically used with NLMEM. We extend the SAEM algorithm developed by Kuhn and Lavielle (19) and the monolix 1.1 Matlab function (http://mahery.math.u-psud.fr/∼lavielle/monolix) by including in the simulation step of the SAEM algorithm a simulation of the left-censored data with the right-truncated Gaussian distribution using an accept-reject algorithm proposed by Robert (24). This extended SAEM algorithm is available on the same web address. At the same time, the convergence of the algorithm is monitored by graphical criterion. An automatic implementation of a stopping criterion to optimize both the iterations number and the stochastic approximation step will be considered in the next extension.

We apply this extended SAEM algorithm to model the HIV viral load decrease. The simulation study illustrates the accuracy of our approach. We show that the bias and RMSE obtained by the extended SAEM algorithm are highly satisfactory for all parameters. They almost reach the benchmark obtained before censoring the datasets, although for the last observation time, 72% of the observations are below the LOQ. We consider 2 classical methods obtained either by omitting the data points below the limit or by imputing half the LOQ to left-censored data. We show that the bias and RMSE obtained by the extended SAEM algorithm are much reduced compared to these 2 approaches.

The analysis of the TRIANON dataset also demonstrates the ability of the extended SAEM algorithm to detect differences between 2 treatment groups. This example illustrates the necessity to handle carefully the left-censored data, as the usual approach fails to detect statistical difference between treatment groups. The bi-exponential model that we use is deduced from a differential equation model proposed by Perelson et al. (1) describing the global HIV dynamics with both the viral load decrease and the $CD4^+$ increase under treatment. Ding and Wu (6) show that this differential system has an analytic solution under the assumption that the non-infected $CD4^+$ cells concentration is constant. Because this assumption is not valid after several week's treatment, the authors recommend using this model only during the first weeks after beginning a treatment, before any rebound of the viral load due to multiple virus mutations. Thus, we consider only the first weeks of the HIV dynamics of TRIANON data. After several weeks, the dif-

ferential system has no more analytical solution. The exact SAEM algorithm could also be extended to this case but is out of the scope of this paper.

To take into account the censored-data problem with NLMEM, Wu (7; 17) proposes MCEM algorithms. In his first paper, he proposes a MCEM with an M-step based on the linearization of the model, leading to an approximate maximum likelihood estimation method. In the second paper, he proposes an exact MCEM. However, he emphasizes computational problems, such as slow or even no convergence, especially when the dimension of the random effects is not small. Because the main problem of the MCEM is the simulation of large independent samples of the random effects at each iteration, Wu proposes complex sampling methods for the E-step. As an alternative, he also proposes an approximate MCEM, based on the linearization of the model for both the E- and M- steps, leading again to an approximate maximum likelihood estimation method. To avoid both the linearization step and the computational problem, the SAEM algorithm is a more adapted tool to estimate models with missing or non-observed data such as random effects or censored observations. Indeed, only one realization of the random effects has to be simulated at each iteration, sidestepping the computational problem of the E-step of the MCEM. The extended SAEM requires more iterations to reach the convergence than the standard SAEM, because of the inclusion of a more complex Gibbs algorithm. However, the extended SAEM is still less time consuming than the MCEM. As an example, Wu uses the same bi-exponential HIV dynamic model in his simulation study (i.e. a model with a random effect vector of size $p = 4$). Wu explains that it takes about 1 hour for the exact MCEM algorithm to converge, whereas the extended SAEM algorithm takes about 120 s to converge. The extended SAEM, which is a true maximum likelihood estimation method, is about 10 times faster than the approximate MCEM algorithm proposed by Wu (17). Wu proposes a PX-EM (28) version of its MCEM, which converges faster. The extended SAEM could also be combined with the PX-EM, gaining a similar rate of convergence. The method proposed by Jacqmin-Gadda et al. (5) for censored data analysed with linear mixed models could also be extended to the non-linear case. It would be interesting to compare these 2 exacts ML estimation methods.

We only focus on the left-censored data problem in the context of log viral load observations, but SAEM can be extended to other missing processes such as missing covariates, which Wu (17) includes in his MCEM. This requires making distribution assumptions for the incompletely observed covariates,

conditional on the completely observed covariates. This problem is beyond the scope of this article, but it may be solved by a highly similar approach.

In conclusion, the extended SAEM algorithm combines the statistical properties of an exact method together with computational efficiency. We thus recommend the use of this method in NLMEM with left-censored data.

Acknowledgments

The authors thank the scientific committee of the TRIANON-ANRS81 trial for giving us access to their patients' viral load measurements and especially Dr O. Launay, main investigator of TRIANON (Hospital Cochin University, Paris, France) and Dr J.P. Aboulker, methodological coordinator (INSERM SC10, Villejuif, France).

References

[1] A. Perelson, P. Essunger, Y. Cao, M. Vesanen, A. Hurley, K. Saksela, M. Markowitz, D. Ho, Decay characteristics of HIV-1 infected compartments during combination therapy, Nature 387 (1997) 188–191.

[2] H. Wu, A. Ding, V. De Gruttola, Estimation of HIV dynamic parameters, Stat. Med. 17 (1998) 2463–85.

[3] A. Ding, H. Wu, Relationships between antiviral treatment effects and biphasic viral decay rates in modeling HIV dynamics., Math. Biosci. 160 (1999) 63–82.

[4] A. Ding, H. Wu, A comparison study of models and fitting procedures for biphasic viral dynamics in HIV-1 infected patients treated with antiviral therapies., Biometrics 56 (2000) 293–300.

[5] H. Jacqmin-Gadda, R. Thiebaut, G. Chene, D. Commenges, Analysis of left-censored longitudinal data with application to viral load in HIV infection, Biostatistics 1 (2000) 355–368.

[6] A. Ding, H. Wu, Assessing antiviral potency of anti-HIV therapies in vivo by comparing viral decay rates in viral dynamic models, Biostatistics 2 (2001) 13–29.

[7] H. Wu, L. Wu, Identification of significant host factors for HIV dynamics modeled by non-linear mixed-effects models, Stat. Med. 21 (2002) 753–771.

[8] H. Wu, J.-T. Zhang, The study of long-term HIV dynamics using semi-parametric non-linear mixed-effects models, Stat. Med. 21 (2002) 3655–3675.

[9] S. Beal, Ways to fit a PK model with some data below the quantification limit, J. Pharmacokinet. Pharmacodyn. 28 (2001) 481–504.

[10] J. Hughes, Mixed effects models with censored data with applications to HIV RNA levels, Biometrics 55 (1999) 625–629.

[11] A. P. Dempster, N. M. Laird, D. B. Rubin, Maximum likelihood from incomplete data via the EM algorithm, J. R. Stat. Soc. B 39 (1977) 1–38.

[12] S. Beal, L. Sheiner, Estimating population kinetics, Crit. Rev. Biomed. Eng. 8 (1982) 195–222.

[13] M. Lindstrom, D. Bates, Nonlinear mixed-effects models for repeated measures data, Biometrics 46 (1990) 673–687.

[14] R. Wolfinger, Laplace's approximations for non-linear mixed-effect models, Biometrika 80 (1993) 791–795.

[15] H. Wu, L. Wu, A multiple imputation method for missing covariates in non-linear mixed-effects models with application to HIV dynamics, Stat. Med. 20 (2001) 1755–1769.

[16] L. Wu, A joint model for nonlinear mixed-effects models with censoring and covariates measured with error, with application to AIDS studies, J. Am. Stat. Assoc. 97 (2002) 955–964.

[17] L. Wu, Exact and approximate inferences for nonlinear mixed-effects models with missing covariates, J. Am. Stat. Assoc. 99 (2004) 700–709.

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

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

[20] P. Girard, F. Mentré, A comparison of estimation methods in nonlinear mixed effects models using a blind analysis, PAGE 14Abstr 834 [www.page-meeting.org/?abstract=834].

[21] E. Kuhn, M. Lavielle, Coupling a stochastic approximation version of EM with a MCMC procedure, ESAIM: P & S 8 (2005) 115–131.

[22] A. Samson, M. Lavielle, F. Mentré, The SAEM algorithm for group comparison tests in longitudinal data analysis based on non-linear mixed-effects model, submitted.

[23] T. A. Louis, Finding the observed information matrix when using the EM algorithm, J. R. Stat. Soc. B 44 (1982) 226–233.

[24] C. Robert, Simulation of truncated normal variables, Stat. Comput. 5 (1995) 121–125.

[25] A. E. Gelfand, A. F. M. Smith, Sampling-based approaches to calculating marginal densities, J. Am. Stat. Assoc. 85 (1990) 398–409.

[26] V. Duval, M. Karlsson, Impact of omission or replacement of data below the limit of quantification on parameter estimates in a two-compartment model, Pharm. Res. 19 (2002) 1835–1840.

[27] O. Launay, L. Grard, L. Morand-Joubert, P. Flandre, S. Guiramand-Hugon, V. Joly, G. Peytavin, A. Certain, C. Lvy, S. Rivet, C. Jacomet, J.-P. Aboulker, P. Yni, A. N. de Recherches sur le SIDA (ANRS) 081 Study Group Yni, Nevirapine or lamivudine plus stavudine and indinavir: examples of 2-class versus 3-class regimens for the treatment of human immunodeficiency virus type 1., Clin. Infect. Dis. 35 (2002) 1096–1105.

[28] C. Liu, D. B. Rubin, Y. N. Wu, Parameter expansion to accelerate EM: the PX-EM algorithm, Biometrika 85 (1998) 755–770.