ESAIM: PS
January 2008, Vol. 12, p. 196–218
DOI: 10.1051/ps:2007045

ESAIM: Probability and Statistics
www.esaim-ps.org

PARAMETRIC INFERENCE FOR MIXED MODELS DEFINED BY STOCHASTIC DIFFERENTIAL EQUATIONS

Sophie Donnet[1] and Adeline Samson[2]

Abstract. Non-linear mixed models defined by stochastic differential equations (SDEs) are considered: the parameters of the diffusion process are random variables and vary among the individuals. A maximum likelihood estimation method based on the Stochastic Approximation EM algorithm, is proposed. This estimation method uses the Euler-Maruyama approximation of the diffusion, achieved using latent auxiliary data introduced to complete the diffusion process between each pair of measurement instants. A tuned hybrid Gibbs algorithm based on conditional Brownian bridges simulations of the unobserved process paths is included in this algorithm. The convergence is proved and the error induced on the likelihood by the Euler-Maruyama approximation is bounded as a function of the step size of the approximation. Results of a pharmacokinetic simulation study illustrate the accuracy of this estimation method. The analysis of the Theophyllin real dataset illustrates the relevance of the SDE approach relative to the deterministic approach.

Mathematics Subject Classification. 62M99, 62F10, 62F15, 62M09, 62L20, 65C30, 65C40, 62P10.

Received July 13, 2006. Revised April 14, 2007.

Introduction

In the context of biology, experimental studies often consist in repeated measurements of a biological criteria (drug concentration, viral concentration, hemodynamic response, etc.) obtained from a population of subjects. The statistical parametric approach commonly used to analyze this longitudinal data is through mixed models: the same regression function is used for all the subjects, but the regression parameters differ between the individuals. These models have the capacity to discriminate between the inter-subjects variability (the variance of the individual regression parameters) and the residual variability.

In mixed models, the regression function describes a time-dependent dynamic process deriving from physical, physiological or biological principles. These sophisticated modeling approaches involve the use of dynamic systems based on ordinary differential equations (ODEs). For instance, in pharmacokinetics which consist in the study of the drug evolution in an organism, the human body is assimilated to a set of compartments within which the drug flows. As a consequence, the drug concentration evolution is described through dynamic systems

Keywords and phrases. Brownian bridge, diffusion process, Euler-Maruyama approximation, Gibbs algorithm, incomplete data model, maximum likelihood estimation, non-linear mixed effects model, SAEM algorithm.

[1] Paris-Sud University, Laboratoire de Mathématiques, Orsay, France; UMR CNRS 8145, University Paris 5, Paris, France; adeline.samson@univ-paris5.fr

[2] INSERM U738, Paris, France; University Paris 7, Paris, France; UMR CNRS 8145, University Paris 5, Paris, France.

© EDP Sciences, SMAI 2008

Article published by EDP Sciences and available at http://www.esaim-ps.org or http://dx.doi.org/10.1051/ps:2007045

and the regression function is solution of an ODE. The estimation problem in the case where the ODE has no analytical solution has already been solved in [17].

However, most of the time, the studied biological process is not fully understood or too complex to be modeled deterministically. So, to account for time-dependent or serial correlated residual errors and to handle real life variations in model parameters occurring over time, mixed models described by stochastic differential equations (SDEs) have been introduced in the literature (see [32] or [43] for instance). These models are a natural extension of the models defined by ODEs, allowing to take into account errors associated with misspecifications and approximations in the dynamic system.

Hence, this paper deals with parameter estimation for a mixed model defined by a SDE:

$$y_{ij} = z(t_{ij}, \phi_i) + \varepsilon_{ij},$$

where y_{ij} is the observation of subject $i = 1, \ldots, I$ at time t_{ij}, $j = 0, \ldots, J_i$ and $(\varepsilon_{ij})_{i,j}$ is a sequence of i.i.d Gaussian random variables of variance σ^2, representing the measurement errors. The parameters ϕ_i are independently distributed with a density depending on a parameter β. The regression function z is a realization of a diffusion process defined as the solution of the SDE describing the observed dynamic process:

$$\mathrm{d}z(t, \phi) = F(z, t, \phi)\mathrm{d}t + \gamma \mathrm{d}B(t),$$

driven by a Brownian motion $\{B_t, t_0 \leq t \leq T\}$, a drift function F depending on the parameter ϕ and a volatility coefficient γ. If the volatility coefficient γ is zero, the SDE is an ODE, the SDE model parameter ϕ being obviously equivalent to the parameter of the corresponding ODE system, and therefore being interpreted in the same way. In such models, three fundamentally different types of noise have to be distinguished: the inter-subject variability, representing the variance of the individual parameters ϕ_i, the dynamic noise γ, reflecting the real random fluctuations around the corresponding theoretical dynamic model, and the measurement noise σ representing the uncorrelated part of the residual variability associated with assay, dosing and sampling errors, for instance, in a biological context.

The main objective of this paper is to develop a maximum likelihood method to estimate the parameter vector $\theta = (\beta, \gamma^2, \sigma^2)$ for these mixed models. This method will combine statistical tools developed for the estimation of diffusion processes, and for the estimation of mixed models. We first recall some standard estimation methods of diffusion processes, the estimation methods of mixed models being detailed thereafter.

The parametric estimation of a diffusion process is a key issue. Estimation of continuously observed diffusion processes is widely studied (see for instance [29, 36]). Statistical inference of discretely observed diffusion processes is a critical question. When the transition probability of the diffusion process is explicitly known, Dacunha-Castelle and Florens-Zmirou [12] propose a consistent maximum likelihood estimator. However, this transition density has generally no closed form and the estimation methods have to sidestep this difficulty. A short summary of such estimation methods is provided below (see [36, 40] for complete reviews). Explicit estimators based on the minimization of suitable contrasts are proposed by [11, 21] and [24] and results on the asymptotical distribution of the estimators are given. Analytical methods include those of Bibby and Sorensen [9], Sorensen [41] – using estimating functions, Poulsen [35] – using a numerical solution of the Kolmogorov equation – or Aït-Sahalia [1] – based on an analytical non-Gaussian approximation of the likelihood function. Other methods approximate the transition density *via* simulation. They consider the unobserved paths as missing data and introduce a set of auxiliary latent data points between every pair of observations. Along these auxiliary latent data points, the process can be finely sampled using the Gaussian Euler-Maruyama approximation to evaluate the likelihood function *via* numerical integration as proposed by Pedersen [33] and Elerian *et al.* [19], or to evaluate the posterior distribution in a Bayesian analysis again *via* numerical integration, as discussed by Eraker [20] and Roberts and Stramer [37]. In this context and for both maximum likelihood and Bayesian estimations, standard Markov Chain Monte-Carlo (MCMC) methods are used to sample the process with the conditional distributions. However, the convergence rate of these estimation methods decreases with the increase in number of latent data points. Different solutions are proposed to overcome this difficulty: Eraker

[20] suggests the sampling of only one element at a time, while Elerian *et al.* [19] propose to sample block-wise with an importance sampling algorithm. Roberts and Stramer [37] take a slightly different approach as they sample transformations of the diffusion process. To sidestep the Euler-Maruyama approximation, Beskos *et al.* [7, 8] develop an exact simulation method of the diffusion process, applicable even without any analytical form of the transition density. This algorithm can be included in a Monte-Carlo procedure to approximate the likelihood function for a classical estimation and in a Gibbs algorithm for a Bayesian inference. However, this exact simulation method is only adapted for time-homogeneous SDEs, which is frequently not the case when studying biological dynamical systems. Furthermore, even under the conditions defined by Beskos *et al.* [7, 8], this exact method requires the inclusion of accept-reject algorithms, which are difficult to implement in the general case of non-linear SDEs and often require a large computational time. Therefore an Euler-Maruyama approximation approach is considered in this paper.

The above-cited papers do not take into account the observation noise on the collected data. In the case of continuously observed stochastic processes with additive noise, Dembo and Zeitouni propose an EM algorithm [14]. The problem of the parameter estimation of discretely observed diffusion processes with additive measurement noise is evoked in few papers and is not completely solved. In the particular case of linear SDEs, the Kalman filter [38] or the EM algorithms [39] can be used. When the observed process is a Gaussian martingale (and can be seen as an Hidden Markov model), Douc and Matias [18] or Gloter and Jacod [22, 23] exhibit estimators and study their theoretical properties. Unfortunately, these explicit forms of maximum likelihood estimates are limited to the linear SDEs case. Furthermore, these methods are not adapted to the context of mixed models where the parameters of the SDEs are random variables.

The theory for mixed models is widely developed for deterministic models (ODEs). For linear mixed models, maximum likelihood estimation has been well studied [5, 34]. In this context, Ditlevsen *et al.* [16] propose an estimation method adapted to linear mixed model defined by linear SDES, but their example is restricted to the case where the transition probability has explicit expression. For non-linear mixed models, the maximum likelihood estimation is more complex: the likelihood cannot be expressed in a closed form because of the non-linearity of the regression function in the individual random parameters. Several authors propose some widely used likelihood approximation methods, such as linearization algorithms [5, 30], Laplacian or Gaussian quadrature algorithms [45]. However none of these algorithms based on likelihood approximation can be considered as fully established theoretically. In this context, Overgaard *et al.* [32] and Tornoe *et al.* [43] have introduced SDEs in non-linear mixed models, using an extended Kalman filter of the diffusion process, with linearization-based estimation algorithm. The convergence of their algorithm is not proved. An alternative to linearization or approximation of the likelihood is to consider the individual parameters as non-observed data. The EM algorithm is then the most adapted tool to estimate incomplete data models. Because of the non-linearity of the model, stochastic versions of the EM algorithm are proposed. Monte-Carlo EM (MCEM) algorithms have been proposed, with a Monte-Carlo approximation of the E-step [10, 44], but these algorithms may have computational problems, such as slow or even no convergence. As an alternative to address the computational problem, Delyon *et al.* propose a stochastic approximation version of EM (SAEM) [13, 27]. Pointwise almost sure convergence of the estimate sequence to a local maximum of the likelihood has been proved under general conditions [13]. Kuhn and Lavielle [26] propose to combine the SAEM algorithm with a MCMC procedure for the individual parameter simulation, which is not direct in the case of non-linear mixed models. To our knowledge, these estimation methods are not yet extended to mixed models defined by SDEs.

Our main purpose is thus to propose an efficient algorithmic estimation method of the vector of parameters θ together with theoretical convergence results. We consider an approximate statistical model, of which the regression term is the Euler-Maruyama discretized approximate diffusion process of the SDE. The parameter inference is then performed on this new model, using a stochastic version of the EM algorithm. Section 1 describes the setup of the problem which is considered in this paper, detailing the diffusion process and its Euler-Maruyama approximation. The estimation algorithm is presented in Section 2. This section details a tuned MCMC procedure supplying both theoretical and computational convergence properties. The error on the estimation induced by the Euler-Maruyama scheme is quantified in Section 3. In Section 4, the estimation

algorithm is applied to a non-linear mixed effects model issued from pharmacokinetics. Section 5 concludes with some discussion.

1. Data and model

1.1. Mixed model defined by SDEs

Let $y = (y_{ij})_{i=1..I, j=0..J_i}$ denote the vector of the observations for subject i measured at time t_{ij} with $t_{i0} \leq t_{i1} \leq \ldots \leq t_{iJ_i} \leq T$. Let the data y be described by the following statistical model $\mathcal{M}$:

$$\left.\begin{array}{rcll} y_{ij} & = & z(t_{ij}, \phi_i) + \varepsilon_{ij}, & 1 \leq i \leq I, 0 \leq j \leq J_i \\ \phi_i & \sim_{i.i.d.} & \pi(\cdot, \beta), & \\ \varepsilon_{ij} & \sim_{i.i.d} & \mathcal{N}(0, \sigma^2), & \\ \mathrm{d}z(t, \phi_i) & = & F(z, t, \phi_i)\,\mathrm{d}t + \gamma \mathrm{d}B(t), & (1) \\ z(t_0, \phi_i) & = & z_0(\phi_i), & \end{array}\right\} \quad (\mathcal{M})$$

where $\phi_i \in \mathbb{R}^d$ is the individual parameter for subject i, randomly distributed with the density π, depending on the parameter $\beta \in \mathbb{R}^p$ and $\varepsilon = (\varepsilon_{ij})_{i=1\ldots I, j=0\ldots J_i}$ represents the measurement error, with a measurement noise variance σ^2. The regression term $z(t, \phi_i)$ for subject i is a realization of the diffusion process $z : \mathbb{R} \longrightarrow \mathbb{R}$ defined by the equation (1), with B a one-dimensional Brownian motion, γ the volatility coefficient and the function $F : \mathbb{R} \times [t_0, T] \times \mathbb{R}^d \longrightarrow \mathbb{R}$ a known measurable drift function, non-linearly depending on the non-observed parameter ϕ_i. The initial condition z_0 of this process is a deterministic known function of the random parameter ϕ_i (this deterministic function can be a constant).

Our objective is to propose a maximum likelihood estimation method of the parameters vector θ, where $\theta = (\beta, \gamma^2, \sigma^2)$ belongs to some open subset Θ of the Euclidean space $\mathbb{R}^{p+2}$. As the I random parameters ϕ_i and the I random trajectories $z(t, \phi_i)$ are not observed, this statistical problem can be viewed as an incomplete data model. The observable vector y is thus considered as part of a so-called complete vector (y, z, ϕ).

Remark 1.

- In Sections 1, 2 and 3, for notation conveniences, the observation times are assumed to be the same for all subjects: $t_{ij} = t_j$, for all $j = 0, \ldots, J$ and for all $i = 1, \ldots, I$. All the developments proposed in these sections can be obviously extended to the case of unbalanced design. This is especially the case in Section 4, where the estimation algorithm is applied on a real dataset with different observation times for all subjects.
- This work can be extended to a statistical model with a regression function being equal to $g(z(t))$, with g a linear or non-linear function, *i.e.*

$$y_{ij} = g(z(t_j, \phi_i)) + \varepsilon_{ij}, \qquad 1 \leq i \leq I, 0 \leq j \leq J.$$

  However, for the simplicity's sake, we only consider the case $g(z(t, \phi)) = z(t, \phi)$ in this paper.
- The identifiability of this model is a complex problem which is beyond the scope of this paper. However, for some simple examples, the parameters identifiability can be proved.

1.2. Diffusion model

The diffusion process, solution of SDE (1) is defined on a filtered probability space $(\mathcal{O}, \mathcal{F}, \mathcal{F}_t, \mathbb{P})$. Statistical inference makes sense only if the existence and uniqueness of a solution of the SDE (1) for all $z(t_0)$, ϕ and γ

is ensured. Sufficient conditions of existence and uniqueness are the following globally Lipschitz, linear growth and boundedness conditions:

Assumption **(A0)**:

(1) For all $\phi \in \mathbb{R}^d$, for all $0 < R < \infty$, there exists $0 < K_R < \infty$ such that for all $t_0 \leq t \leq T$, for all $x, x' \in \mathbb{R}$ with $|x| \leq R$, $|x'| \leq R$

$$|F(x,t,\phi) - F(x',t,\phi)| \leq K_R|x - x'|.$$

(2) For all $\phi \in \mathbb{R}^d$, for all $0 < T < \infty$, there exists a constant $0 < C_T < \infty$ such that for all $t_0 \leq t \leq T$, for all $x \in \mathbb{R}$

$$|F(x,t,\phi)| \leq C_T(1 + |x|).$$

(3) γ is a non null constant.

Under this assumption, for any $t_0 < t < T$, the distribution of $z(t)$ conditioned by the filtration $\mathcal{F}_{t-}$ is absolutely continuous with respect to the Lebesgue measure on $\mathbb{R}$ ($\mathcal{F}_{t-}$ being the filtration generated by $\{z(s), s < t\}$). This distribution is denoted $p_{z|\phi}(\cdot|\phi;\gamma^2)$ in the following. As a consequence, both y and (y,z,ϕ) have density functions, denoted respectively $p_y(y;\theta)$ and $p_{y,z,\phi}(y,z,\phi;\theta)$ depending on the parameter θ.

1.3. **Introduction of an approximate statistical model**

For common SDEs, the diffusion density $p_{z|\phi}$ has generally no closed form. Consequently neither the likelihood of the observed data $p_y(y;\theta)$ nor the likelihood of the complete data $p_{y,z,\phi}(y,z,\phi;\theta)$ have analytical forms, which further complicates the parameters estimation. To overcome this difficulty, an approximate statistical model, based on the Euler-Maruyama approximation of the diffusion process is introduced.

1.3.1. *Euler-Maruyama approximation of the diffusion process*

The Euler-Maruyama scheme is one of the simplest discrete-time approximation of a diffusion process leading to Gaussian approximations of the transition densities. If the time intervals between the observation instants are too great to obtain a good approximation of the transition density, a natural approach is to introduce a set of auxiliary latent data points between every pair of observations, as first proposed by Pedersen [33]. Let $t_0 = \tau_0 < \tau_1 < \ldots < \tau_n < \ldots < \tau_N = t_J$ denote the deduced discretization of the time interval $[t_0, t_J]$. Let us assume that, for all $j = 0, \ldots, J$, there exists an integer n_j verifying $t_j = \tau_{n_j}$, with $n_0 = 0$ by definition. Let $(h_n)_{1\leq n\leq N}$ be the sequence of the step sizes defined as $h_n = \tau_n - \tau_{n-1}$. Let $h = \max_{1\leq n\leq N} h_n$ be the maximal step size.

Then the diffusion process denoted w and supplied by the Euler-Maruyama approximation of the SDE is described by the following iterative scheme: for a fixed ϕ_i, $w_0(\phi_i) = z_0(\phi_i)$, and for $n = 1 \ldots N$,

$$\begin{aligned} h_n &= \tau_n - \tau_{n-1}\,, \\ w_{i,n} &= w_{i,n-1} + h_n\, F(w_{i,n-1}, \tau_{n-1}, \phi_i) + \gamma\sqrt{h_n}\,\xi_n\,, \\ \xi_n &\sim_{i.i.d} \mathcal{N}(0,1), \end{aligned} \tag{2}$$

where $w_{i,n}$ denotes the realization of the process at time τ_n for the parameter ϕ_i. Consequently, $(w_{i,n_0}, \ldots, w_{i,n_J})$ is an approximation of the original diffusion process at observations instants $(z(t_0,\phi_i), \ldots, z(t_J,\phi_i))$. In the following, let $w_i = (w_{i,n})_{n=0\cdots N}$ denote the realizations vector of the process at the discrete times $(\tau_n)_{n=0\cdots N}$.

1.3.2. *Approximate statistical model*

Using this approximation of the diffusion process provided by the Euler-Maruyama scheme of step size h, an approximate statistical model denoted model $\mathcal{M}_h$ is defined as:

$$\left.\begin{array}{rcll} y_{ij} & = & w_{i,n_j} + \varepsilon_{ij}, & 1 \leq i \leq I, 0 \leq j \leq J\,, \\ \phi_i & \sim_{i.i.d} & \pi(\cdot;\beta)\,, & \\ \varepsilon_{ij} & \sim_{i.i.d} & \mathcal{N}(0,\sigma^2)\,, & \\ w_0(\phi_i) & = & z_0(\phi_i), & \\ h_n & = & \tau_n - \tau_{n-1}\,, & \\ w_{i,n} & = & w_{i,n-1} + h_n\, F(w_{i,n-1},\tau_{n-1},\phi_i) + \gamma\sqrt{h_n}\,\xi_n\,, & 1 \leq n \leq N, \\ \xi_n & \sim_{i.i.d} & \mathcal{N}(0,1)\,, & \end{array}\right\} \quad (\mathcal{M}_h)$$

where $w_{i,n_j} = w(t_j,\phi_i)$ is a realization of the Euler-Maruyama approximated diffusion process defined in (2). On this model $\mathcal{M}_h$, y results from the partial observation of the complete data (y,w,ϕ) with $w = (w_i)_{i=1,\ldots,I}$.

Remark 2. In this data augmentation framework, the choice of the discretization grid $(\tau_n)_{0\leq n\leq N}$ is a central issue to guarantee the fast convergence of the estimation algorithms. Indeed, on the one hand, a small step size h ensures a fine Gaussian diffusion approximation. However, on the other hand, it increases the volume of missing data (w,ϕ), which can lead to arbitrarily poor convergence properties of the algorithms when the missing data volume widely exceeds the volume of actually observed data y. Furthermore, the time intervals between two observations can be strongly different. Therefore, for practical purposes and to prevent unbalanced volumes of missing data, we propose to adjust the step sizes for each single time interval.

In the following, the distributions referring to the model $\mathcal{M}_h$ are denoted q while those referring to the model $\mathcal{M}$ are denoted p. On $\mathcal{M}_h$, the observation vector y is distributed with density distribution $q_y(y;\theta)$, which has no closed form because of the SDE non-linearity with respect to ϕ. But by enriching the observed data with the missing data, and by the Markov property of the diffusion process, the complete data likelihood is analytically known:

$$\begin{aligned} q_{y,w,\phi}(y,w,\phi;\theta) &= \prod_{i=1}^{I}\left[q_{y|w}(y_i|w_i;\sigma^2)\ \prod_{n=1}^{N} q_{w|\phi}(w_{i,n}|w_{i,n-1},\phi_i;\gamma^2)\ \pi(\phi_i;\beta)\right] \\ &= \prod_{i=1}^{I}\left[q_{y|w}(y_i|w_i;\sigma^2)\ \prod_{n=1}^{N} d(w_{i,n};\ w_{i,n-1} + h_n\, F(w_{i,n-1},\tau_{n-1},\phi_i),\ \gamma^2 h_n)\ \pi(\phi_i;\beta)\right], \end{aligned}$$

where $d(.;m,v)$ denotes the Gaussian density with mean m and variance v. As a consequence, the estimation of θ can be performed on the model $\mathcal{M}_h$.

2. Maximum likelihood estimation on the model $\mathcal{M}_h$

In this section, we propose a maximum likelihood estimation method, the vector of parameters θ being thus estimated as the maximizing value of the likelihood $q_y(.\,;\theta)$.

2.1. Stochastic versions of the EM algorithm

The Expectation Maximization (EM) algorithm proposed by Demspter *et al.* [15] takes advantage of the incomplete data model structure. We consider that the observed data y are the partial observations of the complete data (y,x) with x the vector of the non-observed data. The EM algorithm is useful in situations where the direct maximization of $\theta \to q_y(.\,;\theta)$ is more complex than the maximization of $\theta \to Q(\theta|\theta')$, with:

$$Q(\theta|\theta') = E_{x|y}\left[\log q_{y,x}(y,x;\theta)|y;\theta'\right].$$

The EM algorithm is an iterative procedure: at the k-th iteration, the E-step is the evaluation of $Q_k(\theta) = Q(\theta \,|\, \widehat{\theta}_{k-1})$ while the M-step updates $\widehat{\theta}_{k-1}$ by maximizing $Q_k(\theta)$. For cases where the E-step has no closed form, Delyon *et al.* [13] propose the Stochastic Approximation EM algorithm (SAEM) replacing the E-step by a stochastic approximation of $Q_k(\theta)$. The E-step is thus divided into a simulation step (S-step) of the non-observed data $x^{(k)}$ with the conditional distribution $q_{x|y}(.\ |y;\widehat{\theta}_{k-1})$ and a stochastic approximation step (SA-step):

$$Q_k(\theta) = Q_{k-1}(\theta) + \alpha_k \left[\log\left(q_{y,x}(y, x^{(k)}; \widehat{\theta}_{k-1})\right) - Q_{k-1}(\theta)\right],$$

where $(\alpha_k)_{k\in\mathbb{N}}$ is a sequence of positive numbers decreasing to zero. One of the advantage of the SAEM algorithm is the low-level dependence on the initialization θ_0, due to the stochastic approximation of the E-step.

The distribution $q_{x|y}(.|y;\widehat{\theta}_{k-1})$ is likely to be a complex distribution, as for the model $\mathcal{M}_h$, resulting in the impossibility of a direct simulation of the non-observed data x. For such cases, Kuhn and Lavielle [26] suggest to realize the simulation step *via* a MCMC scheme by constructing a Markov chain with an unique stationary distribution $q_{x|y}(\cdot|y;\widehat{\theta}_{k-1})$ at the k-th iteration. They prove the convergence of the estimates sequence provided by this SAEM algorithm towards a maximum of the likelihood under general conditions and in the case where $q_{y,x}$ belongs to a regular curved exponential family.

2.2. Extension of the SAEM algorithm to the model $\mathcal{M}_h$

In the particular case of the model $\mathcal{M}_h$, the non-observed data vector is equal to $x = (w, \phi)$. The estimate sequence obtained by the SAEM algorithm on the model $\mathcal{M}_h$ is denoted by $(\widehat{\theta}_{h,k})_{k\geq 0}$. As the simulation under the conditional distribution $q_{w,\phi|y}$ can not be performed directly, the SAEM algorithm combined with a MCMC procedure is applied to the model $\mathcal{M}_h$ to estimate the model parameter θ. To ensure the convergence of the SAEM algorithm, the model $\mathcal{M}_h$ is assumed to fulfill some regular conditions:

Assumption (**A1**):

(1) $\pi(.\,;\beta)$ is such that $q_{y,w,\phi}$ belongs to the exponential family:

$$\log q_{y,w,\phi}(y, w, \phi; \theta) = -\psi(\theta) + \langle S_h(y, w, \phi), \nu(\theta)\rangle\,,$$

where ψ and ν are two functions of θ, $S_h(y, w, \phi)$ is known as the minimal sufficient statistics of the complete model, taking its value in a subset $\widetilde{\mathcal{S}}$ of $\mathbb{R}^m$ and $\langle\cdot,\cdot\rangle$ is the scalar product on $\mathbb{R}^m$.

(2) $\pi(.\,;\beta)$ fulfills standard regularity conditions such that assumptions **(M2–M5)** of [13] hold.

(3) $\beta \longmapsto \pi(\phi;\beta)$ is of class $\mathcal{C}^m$ for all $\phi \in \mathbb{R}^d$, where m is the dimension of $\widetilde{\mathcal{S}}$.

Remark 3. Assumption (**A1**) is checked by a wide family of probability distributions π, such as Gaussian distributions, etc.

Under the assumption (**A1**), the SA-step of the SAEM algorithm reduces to the approximation of $E\left[S_h(y, w, \phi)|y;\theta'\right]$. Iteration 0 of the SAEM algorithm consists in the initialization of $\theta_{h,0}$ and the approximation $s_{h,0}$ of $E\left[S_h(y, w, \phi)|y;\theta_{h,0}\right]$. The k-th iteration of the SAEM algorithm is thus

- S-Step: a realization of the non-observed data $(w^{(k)}, \phi^{(k)})$ is generated through M iterations of a MCMC procedure providing an uniformly ergodic Markov chain with $q_{w,\phi|y}(\cdot|y;\widehat{\theta}_{h,k-1})$ as unique stationary distribution;
- SA-Step: $s_{h,k-1}$ is updated using the following stochastic approximation scheme:

$$s_{h,k} = s_{h,k-1} + \alpha_k(S_h(y, w^{(k)}, \phi^{(k)}) - s_{h,k-1});$$

- M-Step: $\widehat{\theta}_{h,k-1}$ is updated to maximize the complete log-likelihood:

$$\widehat{\theta}_{h,k} = \arg\max_{\theta}\left(-\psi(\theta) + \langle s_{h,k}, \nu(\theta)\rangle\right).$$

For example, the sufficient statistics corresponding to σ^2 and γ^2 are:

$$S_h^{(1)}(y,w,\phi) = \frac{1}{I(J+1)}\sum_{i=1}^{I}\sum_{j=0}^{J}(y_{ij}-w_{i,n_j})^2,$$
$$S_h^{(2)}(y,w,\phi) = \frac{1}{IN}\sum_{i=1}^{I}\sum_{n=1}^{N}\frac{\left(w_{i,n}-h_nF(w_{i,n-1},\tau_{n-1},\phi_i)\right)^2}{h_n},$$

the SA-step is

$$s_{h,k}^{(1)} = s_{h,k-1}^{(1)} + \alpha_k(S_h^{(1)}(y,w^{(k)},\phi^{(k)}) - s_{h,k-1}^{(1)}),$$
$$s_{h,k}^{(2)} = s_{h,k-1}^{(2)} + \alpha_k(S_h^{(2)}(y,w^{(k)},\phi^{(k)}) - s_{h,k-1}^{(2)}),$$

and the M-step for σ^2 and γ^2 at iteration k reduces to $\widehat{\sigma^2}_{h,k} = s_{h,k}^{(1)}$ and $\widehat{\gamma^2}_{h,k} = s_{h,k}^{(2)}$. The sufficient statistics for β depend on the distribution $\pi(.\,;\beta)$.

2.3. Convergence of the SAEM algorithm on the model $\mathcal{M}_h$

Let denote Π_θ the transition probability of the Markov chain generated by the MCMC algorithm. Following [26], the convergence of the SAEM algorithm combined with MCMC is ensured under the following additional assumption:

Assumption (**A2**):

(1) The chain $(w^{(k)},\phi^{(k)})_{k\geq 0}$ takes its values in a compact set $\mathcal{E}$ of $\mathbb{R}^N\times\mathbb{R}^d$.

(2) For any compact subset V of Θ, there exists a real constant L such that for any (θ,θ') in V^2

$$\sup_{\{(w,\phi),(w',\phi')\}\in\mathcal{E}} |\Pi_\theta\,(w',\phi'|w,\phi) - \Pi_{\theta'}\,(w',\phi'|w,\phi)| \leq L\|\theta-\theta'\|_{\mathbb{R}^{p+2}}.$$

(3) The transition probability Π_θ supplies an uniformly ergodic chain whose invariant probability is the conditional distribution $q_{w,\phi|y}(\cdot\,;\theta)$, *i.e.*

$$\exists K_\theta\in\mathbb{R}^+,\quad \exists\rho_\theta\in]0,1[\quad|\quad \forall k\in\mathbb{N}\quad \|\Pi_\theta^k(\cdot|w,\phi)-q_{w,\phi|y}(\cdot;\theta)\|_{TV}\leq K_\theta\rho_\theta^k,$$

where $\|\cdot\|_{TV}$ is the total variation norm. Furthermore,

$$K=\sup_{\theta\in\Theta}K_\theta<\infty \quad\text{and}\quad \rho=\sup_{\theta\in\Theta}\rho_\theta<1.$$

(4) The function S_h is bounded on $\mathcal{E}$.

A MCMC procedure fulfilling the assumption (**A2-3**) of uniform ergodicity for the Markov Chain generated by Π_θ is proposed in Section 2.4.

Theorem 1. *Let assumptions (**A0-A1-A2**) hold. Let (α_k) be a sequence of positive numbers decreasing to* 0 *such that for all k in $\mathbb{N}$, $\alpha_k\in[0,1]$, $\sum_{k=1}^\infty\alpha_k=\infty$ and $\sum_{k=1}^\infty\alpha_k^2<\infty$.*

Assuming the sequence $(s_{h,k})_{k\geq 1}$ takes its values in a compact set, the sequence $(\widehat{\theta}_{h,k})_{k\geq 1}$ obtained by the SAEM algorithm on the model $\mathcal{M}_h$ converges almost surely towards a (local) maximum $\theta_{h,\infty}$ of the likelihood q_y.

Proof. The convergence of the estimates towards a local maximum of the likelihood function q_y is ensured by the result of [26]. Indeed, assumption (**A1-1**) ensures the exponentiality of the model (assumption (**M1**) of [26]). Assumption (**A1-2**) implies assumptions (**M2-M5**) of [26]. Assumption (**SAEM1'**) is verified by the sequence (α_k). Assumption (**SAEM2**) is due to assumption (**A1-3**). Finally, the required assumption (**SAEM3'**) is resumed in assumption (**A2**). □

Remark 4. If the compactness on $(s_{h,k})_{k\geq 0}$ is not checked or difficult to check, the algorithm can be stabilized using a projection of the stochastic approximation sequence on varying bounds proposed by Andrieu and Moulines [2].

2.4. Simulation of the non-observed data using a MCMC procedure

At the k-th iteration of the SAEM algorithm, given an estimate $\widehat{\theta}_{h,k-1}$, a realization of the non-observed data $(w^{(k)}, \phi^{(k)})$ is generated through the succession of M iterations of a MCMC procedure. MCMC procedures construct a Markov chain with $q_{w,\phi|y}(.|y\,;\widehat{\theta}_{h,k-1})$ as the invariant distribution, by proposing candidates (ϕ^c, w^c) with any proposal density Q. However, sampling all the missing data at the same time can lead to poor convergence properties. Therefore, a hybrid Gibbs algorithm is implemented and realized successively M times, the m-th iteration being written as:

(1) for $i = 1, \ldots, I$, generation of $\phi_i^{(m)}$, using a Metropolis-Hastings (M-H) procedure with Q_1 as proposal density and such that $q_{\phi|y,w}(.\,|y_i, w_i^{(m-1)}; \widehat{\theta}_{h,k-1})$ is the invariant distribution;

(2) for $i = 1, \ldots, I$, generation of $w_i^{(m)}$, using a M-H procedure with Q_2 as proposal distribution and such that $q_{w|y,\phi}(.\,|y_i, \phi_i^{(m)}; \widehat{\theta}_{h,k-1})$ is the invariant distribution.

A careful choice of the proposal densities Q_1 and Q_2 will help the algorithm to quickly explore the parameters space. In the following, some proposal densities of which efficiency is proved on numerical examples are detailed. To simplify the notation, the parameter $\widehat{\theta}_{h,k-1}$ is omitted since this simulation is performed for a fixed $\widehat{\theta}_{h,k-1}$.

2.4.1. *Proposal distributions*

(1) Simulation of the candidate ϕ_i^c can be carried out with the prior density π which allows an efficient exploration of the space of parameters. This leads to an independent M-H algorithm. An alternative consists in generating a candidate in a neighborhood of $\phi_i^{(m-1)}$, $\phi_i^c = \phi_i^{(m-1)} + \eta_i$ with $\eta_i \sim \mathcal{N}(0, \delta)$ and where δ is a scaling parameter on which the algorithm convergence depends. This results in the so-called random-walk M-H algorithm (see for example [6]).

(2) A trajectory candidate w_i^c can be generated using the Euler-Maruyama scheme which corresponds to the prior distribution. An alternative to simulate w_i^c consists in splitting the vector w_i into two parts $(w_{i,n_0}, \cdots, w_{i,n_J})$ and $w_{i,aux}$, the former being the process observed at times $(t_j)_{j=0\cdots J}$ and the latter being the process observed at the auxiliary latent times excluding the observation times. The simulation of $(w_{i,n_0}^c, \cdots, w_{i,n_J}^c)$ can be performed with random walk distributions: $w_{i,n_j}^c = w_{i,n_j}^{(m-1)} + \eta_i'$ where $\eta_i' \sim \mathcal{N}(0, \delta')$ and δ' is a scaling parameter chosen to ensure good convergence properties. As proposed by Pedersen [33], the trajectory at the auxiliary times $w_{i,aux}^c$ can be generated using an unconditioned distribution but it would have poor convergence properties. A more appropriate strategy consists in generating a candidate $w_{i,aux}^c$ using Brownian bridges, conditioning the proposed bridge on the events $(w_{i,n_j}^c)_{j=0\cdots J}$, as suggested by Eraker [20] or Roberts and Stramer [37]. More precisely, for $n_{j-1} < n < n_j$, w_{i,τ_n} is simulated with:

$$w_{i,\tau_n}^c = w_{i,n_{j-1}}^c + \frac{w_{i,n_j}^c - w_{i,n_{j-1}}^c}{t_j - t_{j-1}}(\tau_n - t_{j-1}) + \overline{B}_{\tau_n},$$

where $\overline{B}$ is a standard Brownian bridge on $[0, 1]$ equal to zero for $t = 0$ and $t = 1$, which can be easily simulated.

2.4.2. *Uniform ergodicity of the MCMC procedure*

In case of a cyclic combination, the uniform ergodicity of the Markov Chain is ensured if one of the proposal distributions satisfies a minoration condition (Props. 3 and 4 of [42]). However, by Corollary 4 of [42], an independent M-H algorithm verifies the minoration condition if the weight function $q_{w,\phi|y}(w, \phi|y)/R(w, \phi)$ is bounded, where R denotes the proposal distribution for the couple (w, ϕ). This is obviously the case when

$R(w,\phi) = q_{w,\phi}(w,\phi)$. As a consequence, the Metropolis-Hastings algorithm proposed in part 2.4.1 fulfills the conditions of uniform ergodicity required by the SAEM algorithm.

3. Survey of the error induced by the Euler-Maruyama approximation

The estimation method proposed in this paper generates two distinct types of errors on the parameters estimate that have to be controlled.

The first type of error is induced by the estimation method itself. The estimation algorithm produces a sequence $(\widehat{\theta}_{h,k})_{k\geq 0}$ of estimates which converges towards $\theta_{h,\infty}$, a (local) maximum of the $\mathcal{M}_h$-likelihood $q_y(y;\cdot)$ function. The limiting distribution of the estimate is a rather delicate issue, which is beyond the scope of this paper. However the variance of this estimate $\widehat{\theta}_{h,k}$ is classically controlled by the standard error which is evaluated through the Fisher information matrix of the estimates. Kuhn and Lavielle [26] propose to estimate this Fisher information matrix by using the stochastic approximation procedure and the Louis' missing information principle [31]. This Fisher information matrix estimate can be adapted to the model $\mathcal{M}_h$ for a fixed value of the step size h. At the k-th iteration of the SAEM algorithm, the three following quantities are evaluated:

$$\Delta_k = \Delta_{k-1} + \alpha_k \left[\partial_\theta \log q_{y,w,\phi}(y, w^{(k)}, \phi^{(k)}; \widehat{\theta}_{h,k}) - \Delta_{k-1}\right] \tag{3}$$

$$\begin{aligned} G_k = G_{k-1} + \alpha_k \Big[&\partial_\theta^2 \log q_{y,w,\phi}(y, w^{(k)}, \phi^{(k)}; \widehat{\theta}_{h,k}) \\ &+\partial_\theta \log q_{y,w,\phi}(y, w^{(k)}, \phi^{(k)}; \widehat{\theta}_{h,k})\, \partial_\theta \log q_{y,w,\phi}(y, w^{(k)}, \phi^{(k)}; \widehat{\theta}_{h,k})^t - G_{k-1}\Big] \end{aligned} \tag{4}$$

$$H_k = \Delta_k \Delta_k^t - G_k. \tag{5}$$

As the sequence $(\widehat{\theta}_{h,k})_{k\geq 0}$ converges to the maximum of the likelihood, the sequence $(H_k)_{k\geq 0}$ converges to the observed Fisher information matrix. The diagonal elements of the inverse of this matrix provide estimates of the variance of $\widehat{\theta}_{h,k}$.

A second type of error is induced on the estimate by the Euler-Maruyama scheme. Indeed, for the reasons evoked in Section 1, the SAEM algorithm is applied to the model $\mathcal{M}_h$ instead of to the model $\mathcal{M}$: the algorithm maximizes the $\mathcal{M}_h$-likelihood function q_y instead of the $\mathcal{M}$-likelihood function p_y.

The aim of this section is to study this second type of error induced by the Euler-Maruyama scheme on the conditional distribution $q_{w,\phi|y}$ and on the likelihood function q_y. In Theorem 2, we propose bounds of this error as a function of the maximal step size of the Euler-Maruyama scheme h. In the following, an additional assumption holds:

Assumption (**A3**):

The function $F : \mathbb{R} \times [t_0, T] \times \mathbb{R}^d \longrightarrow \mathbb{R}$ is infinitely differentiable in the variable space and its partial derivatives of any order are uniformly bounded with respect to x and ϕ.

Theorem 2. *Let the assumptions (**A0**–**A3**) hold.*

(1) *Let z and w be the diffusion processes of the models $\mathcal{M}$ and $\mathcal{M}_h$ respectively, at the observation times: $z = (z(t_0), \cdots, z(t_J))$ and $w = (w(t_0), \cdots, w(t_J))$. Let $p_{z,\phi|y}$ and $q_{w,\phi|y}$ be the conditional distributions on the models $\mathcal{M}$ and $\mathcal{M}_h$ respectively. There exists a non-negative constant $C(y)$ dependent of y and a non-negative constant H_0, such that, for any $0 < h < H_0$,*

$$\left\|p_{z,\phi|y} - q_{w,\phi|y}\right\|_{TV} \leq C(y)h,$$

where $\|\cdot\|_{TV}$ denotes the total variation distance.

(2) *Let p_y and q_y be the likelihoods of the models $\mathcal{M}$ and $\mathcal{M}_h$ respectively. There exists a non-negative constant $C_2(y)$ dependent of y and independent of θ, such that for all $0 < h < H_0$,*

$$\sup_{\{\theta=(\beta,\gamma^2,\sigma^2),\gamma_0^2<\gamma^2<\Gamma_0^2\}} |p_y(y;\theta) - q_y(y;\theta)| \leq C_2(y)h.$$

Theorem 2 is proved in Appendix A. These results are based on the convergence rate of the transition densities proposed by Bally and Talay [4].

As a principal consequence of part (2) of this theorem, and assuming regularity hypotheses on the Hessians of the likelihoods of both models $\mathcal{M}$ and $\mathcal{M}_h$, the bias of the estimates induced by both the numerical approximation and the estimation algorithm, is controlled. More precisely, let introduce an additional assumption:
Assumption (**A4**):

(1) The likelihood functions $p_y(y;\theta)$ and $q_y(y;\theta)$ are twice differentiable.
(2) Let θ_∞ and $\theta_{h,\infty}$ be the maxima of functions $p_y(y;\theta)$ and $q_y(y;\theta)$ respectively. There exist two non-negative constants ε_1 and ε_2 such that for every $\theta \in \{\theta_{h,\infty} + t(\theta_{h,\infty} - \theta_\infty), t \in [0,1]\}$, and for every $x \in \mathbb{R}^{p+2}$

$$-x^t H_{p_y}(\theta)x \geq \varepsilon_1 \|x\|^2$$
$$-x^t H_{q_y}(\theta)x \geq \varepsilon_2 \|x\|^2,$$

where H_{p_y} and H_{q_y} are the Hermitian matrices of $p_y(y;\cdot)$ and $q_y(y;\cdot)$ respectively.

Corollaire 1. *Let assumptions (**A0**–**A4**) and the assumptions of Theorem 1 hold. Let θ_∞ and $\theta_{h,\infty}$ be the likelihood maxima of models $\mathcal{M}$ and $\mathcal{M}_h$ respectively. Let $(\widehat{\theta}_{h,k})_{k\geq 1}$ be the sequence of estimates obtained by the SAEM algorithm on the $\mathcal{M}_h$ model. Therefore, $(\widehat{\theta}_{h,k})_{k\geq 1}$ converges almost surely towards $\theta_{h,\infty}$ and there exists a non-negative constant M, independent of θ such that*

$$\|\theta_{h,\infty} - \theta_\infty\|^2 \leq Mh.$$

Corollary 1 is proved in Appendix A.

4. Theophyllin pharmacokinetic example

The estimation method developed in Section 2 is applied below to a pharmacokinetics example.

4.1. Pharmacokinetics

Pharmacokinetics (PK) studies the time course of drug substances in the organism. This can be described through dynamic systems, the human body being assimilated to a set of compartments within which the drug evolves with time. In general, these systems are considered in their deterministic version. However, in a recent book on PK modeling, Krishna [25] claims that the fluctuations around the theoretical pharmacokinetic dynamic model may be appropriately modeled by using SDEs rather than ODEs. Overgaard *et al.* [32] suggest the introduction of SDEs to consider serial correlated residual errors due for example to erroneous dosing, sampling history or structural model misspecification.

In the PK context, non-linear mixed-effects models are classically considered with a Gaussian distribution for the individual parameters: $\phi_i \sim \mathcal{N}(\mu, \Omega)$, for $i = 1, \ldots, I$. The assumption **(A1)** is fulfilled for this particular choice of distribution $\pi(. : \beta)$. In this case, the parameter β to estimate is $\beta = (\mu, \Omega)$. In the following, the hypothesis "$t_{ij} = t_j$ for all i", is not assumed and the observation times t_{ij} may differ between subjects.

In a deterministic approach, the regression function z is defined as the solution of a PK ordinary differential system: $\mathrm{d}z(t)/\mathrm{d}t = F(z(t), t, \phi)$ with $z(t_0) = Z_0$, each component of the vector ϕ having a PK meaning.

For example, a classic one compartment PK model with first order absorption and first order elimination is described by the following dynamic equation: $z_0 = 0$ and

$$\frac{\mathrm{d}z(t,\phi)}{\mathrm{d}t} = \frac{Dose \cdot K_a K_e}{Cl}\mathrm{e}^{-K_a t} - K_e z(t,\phi), \tag{6}$$

where z represents the drug concentration in blood, $Dose$ is the known drug oral dose received by the subject, K_e is the elimination rate constant, K_a is the absorption rate constant and Cl is the clearance of the drug. A stochastic differential system can be deduced from this ODE:

$$\mathrm{d}z(t,\phi) = \left(\frac{Dose \cdot K_a K_e}{Cl}\mathrm{e}^{-K_a t} - K_e z(t,\phi)\right)\mathrm{d}t + \gamma \mathrm{d}B_t, \tag{7}$$

where B_t is a Brownian motion and γ is the volatility coefficient of the SDE. This SDE fulfills assumptions **(A0)** and **(A3)**.

This SDE is linear and the law of the diffusion Z is analytically known. However, this diffusion is non-linear with respect to the individual parameter ϕ_i. Consequently, the likelihood of the corresponding non-linear mixed model has no analytical form and estimation methods such as the SAEM algorithm combined with MCMC schemes are needed. In practice, the assumption **(A2-1)** is always fulfilled as simulation of diffusions or individual parameters is always realized in a compact set.

4.2. Simulation study

The aim of this simulation study is to illustrate the accuracy (bias and root mean square errors) of the extended SAEM algorithm developed in Section 2 on a PK application.

We use the previous PK model to mimic the Theophyllin drug pharmacokinetic. The design of simulation is $I = 36$ subjects and nine blood samples per patient ($J = 8$), taken at 15 min, 30 min, 1, 2, 3.5, 5, 7, 9, 12 h after dosing. The drug oral dose ($Dose$) received by the subject is chosen arbitrarily between 3 and 6 mg. To prevent the parameters from taking unrealistic negative values, the vector $\phi \in \mathbb{R}^3$ is classically composed of the log parameters $\phi = (\log(K_e), \log(K_a), \log(Cl))$. The individual parameters $(\phi_i)_{i=1,\dots,I}$ are thus simulated with Gaussian distributions $\mathcal{N}(\mu, \Omega)$, with μ equal to $(-2.52, 0.40, -3.22)$ as proposed by [34]. A diagonal variance-covariance matrix Ω is assumed for the Gaussian distribution of ϕ. Let $\omega^2 = (\omega_1^2, \omega_2^2, \omega_3^2)$ denote the vector of these variances. The inter-subject variability is set equal for the three parameters: $\omega_1^2 = \omega_2^2 = \omega_3^2 = 0.01$. We set a volatility coefficient equal to $\gamma^2 = 0.2$ and an additive Gaussian measurement error $\sigma^2 = 0.1$. We generate 100 datasets with this protocol. To evaluate the accuracy of the estimates of $\theta = (\mu, \omega^2, \gamma^2, \sigma^2)$ produced by the SAEM algorithm, the estimation of the parameters is performed on the 100 datasets, simulated by the previous protocol, using the extension of the SAEM algorithm presented in Section 2.2.

The Euler-Maruyama scheme included in the SAEM algorithm is implemented on a grid with auxiliary latent data points introduced between each pair of observation instants as detailed in Section 1.3.1. The number of auxiliary points has to be chosen carefully because a volume of missing data too large can induce arbitrarily poor convergence properties of the Gibbs algorithm. In this example, we divide each time interval $[t_{i,j}, t_{i,j+1}]$ into 20 sub-intervals of equal length. This choice supplies a reasonable volume of missing data with respect to the volume of observed data, avoids unbalance between the observation-time intervals and proves its numerical efficiency in accurately approximating the solution of the SDE.

The implementation of the Gibbs procedure included in the SAEM algorithm requires subtle tuning in practice. In particular, the simulation of the diffusion process w on the auxiliary grid is highly critical. An unconditioned trajectory simulation with $q(w_{i,n_j}|w_{i,n_{j-1}};\theta)$ as proposed by Pedersen [33] provides poor numerical results in the case of this example. Indeed, a great number of these simulated trajectories produce large jumps $(w_{i,\tau_{n_j}} - w_{i,\tau_{n_j}-1})$. The probability of such trajectories being close to zero, it induces too low an acceptance rate. As suggested by Eraker [20] or Roberts and Stramer [37] and detailed in Section 2.4, a conditioned trajectory simulation through Brownian bridge distributions is preferred. Moreover, we update the missing trajectories at

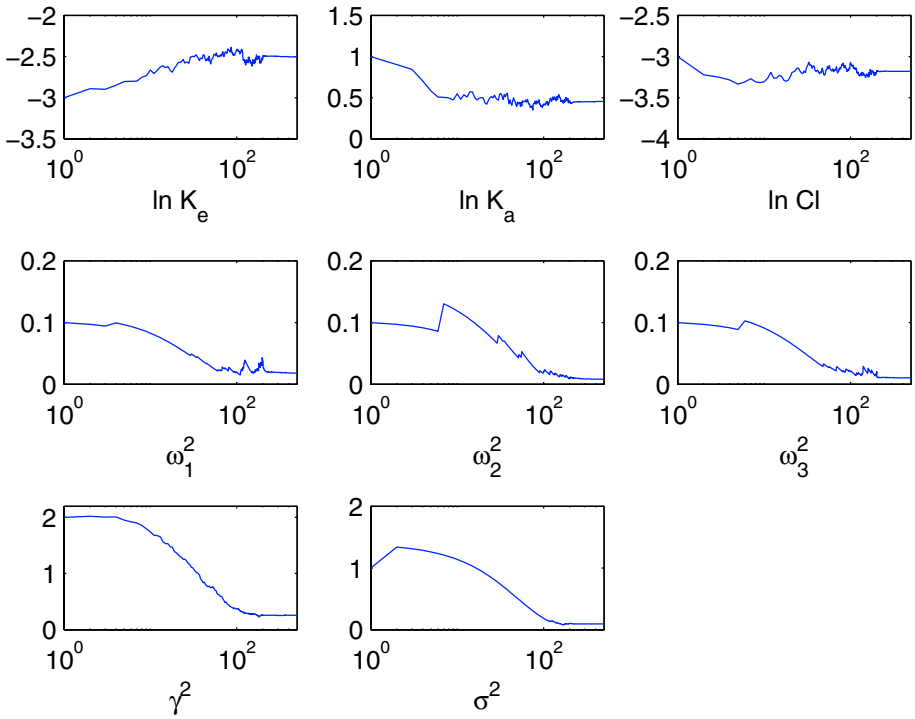

FIGURE 1. Evolution of the SAEM parameter estimates function of the iteration number in a logarithmic scale.

once for each subject, as recommended by Elerian *et al.* [19] to avoid a high level of rejection. In this example, we obtain acceptance rates in the neighborhood of 25%.

The implementation of the SAEM algorithm requires initial value θ_0 and the choice of the stochastic approximation sequence $(\alpha_k)_{k\geq 0}$. The initial values of the parameters are chosen arbitrarily and set to $\theta_0 = (-3, 1, -3, 0.1, 0.1, 0.1, 2, 1)$, the convergence of the SAEM algorithm few depending on the initialization. The step of the stochastic approximation scheme is chosen as recommended by Kuhn and Lavielle [27]: $\alpha_k = 1$ during the first iterations $1 \leq k \leq K_1$, and $\alpha_k = (k - K_1)^{-1}$ during the subsequent iterations. Indeed, the initial guess θ_0 might be far from the maximum likelihood value and the first iterations with $\alpha_k = 1$ allow the sequence of estimates to converge to a neighborhood of the maximum likelihood estimate. Subsequently, smaller step sizes during $K - K_1$ additional iterations ensure the almost sure convergence of the algorithm to the maximum likelihood estimate. We implement the extended SAEM algorithm with $K_1 = 200$ and $K = 500$ iterations. Figure 1 illustrates the convergence of the parameter estimates provided by the extended SAEM algorithm as a function of the iteration number in a logarithmic scale. During the first iterations of SAEM, the parameter estimates fluctuate, reflecting the Markov chain construction. After 200 iterations, the curves smooth out but still continue to converge towards a neighborhood of the likelihood maximum. Convergence is obtained after 500 iterations.

Let denote $\widehat{\theta}_r$ the estimate of θ obtained on the rth simulated dataset, for $r = 1, \ldots, 100$. The relative bias $\frac{1}{100}\sum_r \frac{(\widehat{\theta}_r - \theta)}{\theta}$ and relative root mean square error (RMSE) $\sqrt{\frac{1}{100}\sum_r \frac{(\widehat{\theta}_r - \theta)^2}{\theta^2}}$ for each component of θ are computed and presented in Table 1.

The estimates of the mean parameter μ have very low bias (<5%). The variance parameters have small bias (<9%) except γ^2, this variance parameter being slightly over-estimated (13%). The RMSE are very satisfactory for the mean parameter (<9%). The RMSE for the variance parameters are greater but still satisfactory (≤40%)

TABLE 1. Relative bias (%) and relative root mean square error (RMSE) (%) of the estimated parameters evaluated by the SAEM algorithm from 100 simulated trials with $I = 36$ subjects.

Parameters	Bias (%)	RMSE (%)
log K_e	0.42	−3.19
log K_a	4.14	8.95
log Cl	−0.23	−2.27
ω_1^2	3.83	40.03
ω_2^2	8.49	36.76
ω_3^2	−8.81	37.52
γ^2	13.02	21.31
σ^2	−4.44	18.79

in comparison to the small number of subjects ($I = 36$). The RMSE of σ^2 is particularly satisfactory (<20%) considering the complexity of the variability model.

In conclusion, even if this simulation study is performed on a complex model, the convergence of the extended SAEM algorithm towards the maximum likelihood neighborhood is computationally efficient. In addition despite the fact that the number of subjects is small, the extended SAEM algorithm all in all supplies accurate estimations of the parameters. Furthermore, the accuracy is comparable to that obtained with the classic SAEM algorithm for an ODE version of a mixed model *i.e.* for a model with one less variability level.

4.3. **A real data example**

The extended SAEM algorithm is used to estimate the PK parameters of the Theophyllin drug PK real dataset. This new analysis of the Theophyllin dataset aims at illustrating the advantage of the SDE approach over the ODE approach.

In this clinical trial, twelve subjects received a single oral dose of 3 to 6 mg of Theophyllin. Ten blood samples were taken around 15 min, 30 min, 1, 2, 3.5, 5, 7, 9, 12 and 24 h after dosing. The individual data are displayed in Figure 2. The Theophyllin PK is classically described by the one compartment model with first order absorption and first order elimination presented previously. We fit the Theophyllin data with the regression term successively defined as the solution (6) and then as that of the SDE (7).

In the ODE approach, the differential equation (6) has an explicit solution. Thus, the parameters estimates are obtained using the SAEM algorithm combined with a MCMC procedure proposed by Kuhn and Lavielle [26]. The individual concentration profiles are predicted by $\widehat{z}_{ij} = z(t_{ij}, \hat{\phi}_i)$ for all i and j where z is the solution of (6) and $\hat{\phi}_i$ is an estimation of the posterior mean $E(\phi_i|y_i;\hat{\theta})$ evaluated during the last iterations of the SAEM algorithm.

In the SDE approach, the same implementation of the extended SAEM algorithm as the one detailed for the simulation study (*i.e.* with an Euler-Maruyama approximation of the SDE) is used. The individual concentration predictions $E(w(t_{ij}, \phi_i)|y_i;\hat{\theta})$ for all i and j are evaluated by $\widehat{w}_{ij} = 1/100 \sum_{k=K-99}^{K} w^{(k)}(t_{ij}, \phi_i^{(k)})$ where $(w^{(k)}(t_{ij}, \phi_i^{(k)}))_{k=K-99,\ldots,K}$ are simulated under the conditional distribution $q_{w,\phi|y}(\ .\ |y_i;\widehat{\theta})$ during the 100 last iterations of the extended SAEM algorithm.

However, as the differential equation (7) is linear, an exact simulation of the diffusion process z can also be performed and be combined with the SAEM algorithm. The individual concentration predictions $E(z(t_{ij}, \phi_i)|y_i;\hat{\theta})$ for all i and j are evaluated by $\widehat{z}_{ij} = 1/100 \sum_{k=K-99}^{K} z^{(k)}(t_{ij}, \phi_i^{(k)})$ where $(z^{(k)}(t_{ij}, \phi_i^{(k)}))_{k=K-99,\ldots,K}$ are simulated under the conditional distribution $p_{z,\phi|y}(\ .\ |y_i;\widehat{\theta})$ during the 100 last iterations of the extended SAEM algorithm. As a consequence, on this particular example, the influence of the Euler approximation on the predictions can be estimated.

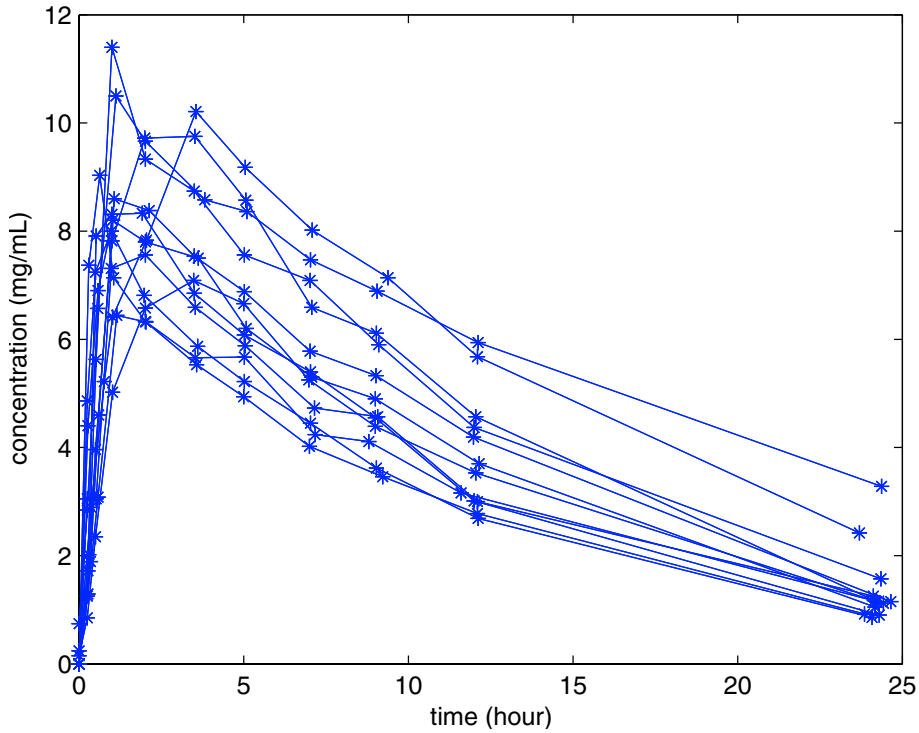

FIGURE 2. Individual concentrations for the pharmacokinetics of Theophyllin for 12 subjects.

The SAEM algorithm is implemented with 500 iterations for the 3 models. The step size for the Euler-Maruyama approximation is such that 100 latent points are introduced between each pair of observations instants. The step size h is decreased from the simulation study in order to assure the approximate model $\mathcal{M}_h$ to be close to the model $\mathcal{M}$.

The ODE and the two SDE predictions are overlaid on the data in Figure 3 for four typical subjects. Both ODE and SDE predicted curves for the other eight subjects are satisfactory and thus not presented here.

First, the results obtained by the Euler-Maruyama approximation and the exact simulation based approaches illustrate the control of the error induced by the Euler-Maruyama approximation (result of Th. 2), as the two predicted curves are almost always very closed. Moreover this graph illustrates that this error is insignificant in comparison of the contribution of the SDE approach. Secondly, for the subject 3, the ODE predicted curve is satisfactory as well as the two SDE predicted curves. For the subjects 1, 2 and 12, the ODE predicted curves miss some of the observed data, particularly the elimination phase of the drug. The SDE predicted curves improve all of these individual profiles, especially the elimination phase.

With the ODE approach, the variabilities are estimated as $\omega_1 = 0.003$, $\omega_2 = 0.653$, $\omega_3 = 0.167$, $\sigma = 0.709$. With the SDE approach, the variabilities are estimated as $\omega_1 = 0.001$, $\omega_2 = 0.639$, $\omega_3 = 0.001$, $\sigma = 0.466$ and $\gamma = 0.780$. Thus, the first two variabilities remain almost unchanged by the SDE approach whereas, as expected, σ is lower in the SDE approach than in the ODE approach. As a consequence, we observe a new decomposition of the various sources of variabilities of the data, which distinguishes the variability due to real fluctuations around the theoretical model (γ) from the residual variability (σ).

5. DISCUSSION

This paper proposes a maximum likelihood estimation method for mixed effects models defined by a discretely observed diffusion process including additive measurement noise. To that end, an approximate model $\mathcal{M}_h$ is

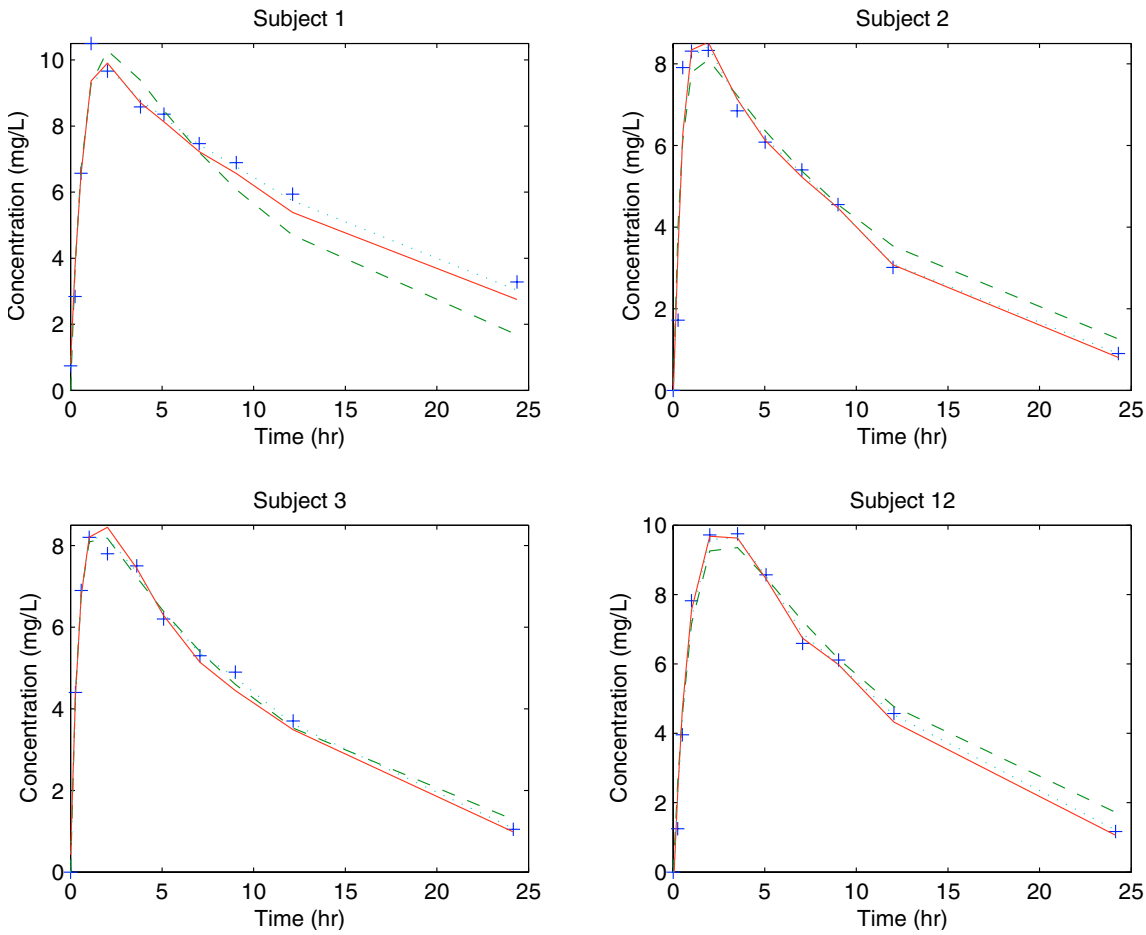

FIGURE 3. Individual concentration curves for subject 1, 2, 3 and 12, predicted by SAEM with the ODE approach (dotted line), the SDE approach based on the Euler-Maruyama approximation (plain line) and the SDE approach based on an exact simulation of the diffusion (dashed line) are overlaid on the data points for the pharmacokinetic of Theophyllin.

introduced, of which the regression term is evaluated using a Gaussian Euler-Maruyama approximation of maximal step size h. The SAEM algorithm, extended to this model, requires the simulation of the missing data (w, ϕ) with the conditional distribution $q_{w,\phi|y}$. The choice of the proposal distributions governs the convergence properties of the algorithm and thus is a key issue. A tuned MCMC procedure to perform this simulation is thus proposed, combining a hybrid Gibbs algorithm with independent or random walk Metropolis-Hastings schemes.

Moreover, we prove that the error induced by the Euler-Maruyama Gaussian approximation on the conditional distributions and the likelihoods decreases linearly when the step size h goes to zero. These results are proved under strong assumptions that could be probably be weakened.

When the step size h decreases and the number of auxiliary latent times increases, the simulation of the diffusion process becomes more difficult. Therefore, a trade-off between a small discretisation error and a small simulation variance has to be found. An extension of this work would be a theoretical result to find this trade-off. As this result is beyond the scope of this paper, we used in the examples empirical trade-offs.

The discretisation error is distinct from the error on the estimates induced by the estimation algorithms, which is classically evaluated through the Fisher information matrix. This Fisher information matrix is estimated by stochastic approximation using the Louis' missing information principle [31]. This matrix provides standard

errors of the estimates. The limiting distribution of the estimates has been described by Delyon *et al.* [13] for an averaged SAEM procedure. However this result can not be applied when the SAEM algorithm is combined with a MCMC algorithm because, the random noise in the stochastic approximation scheme is not anymore a martingale increment. In the case of SAEM-MCMC algorithm, the limiting distribution of the estimates is difficult to derive and this problem is beyond the scope of this paper.

The stochastic version of the EM algorithm SAEM proposed by Kuhn and Lavielle [26] is preferred to the Monte-Carlo EM (MCEM) developed by Wei and Tanner [44] because of its computational properties. Indeed, SAEM requires the generation of only one realization of the non-observed data at each iteration. In a context where the missing data have to be simulated by a MCMC method, decreasing the size of these missing data is a key issue to ensure acceptable computational times.

The accuracy of the extended SAEM algorithm is illustrated on a pharmacokinetic simulation study. The parameters are estimated with small bias and the mean square error are satisfactory given the complexity of the model. The relevance of the SDEs approach with respect to the deterministic one is exemplified on a real dataset based on a linear SDE. The comparison between the exact simulation of the SDE and the Euler-Maruyama approximation illustrates the accuracy of the discretisation approach. Therefore, this example justifies the use of this discretisation approach for the parameter estimation of more general models without explicit solution.

The same work can be carried out in a full Bayesian framework. In that context, a prior distribution is specified on the parameters θ and a tuned hybrid Gibbs sampling algorithm has to be developed to estimate the posterior distribution $p_{\theta|y}(\cdot|y)$. As before, the algorithms are performed on an approximate model defined by the Euler-Maruyama scheme, and the instrumental distributions proposed in this paper can be included in the Gibbs sampling algorithm. The ergodicity of the Markov chain generated by the MCMC algorithm on the approximate model can be proved under general assumptions. Furthermore, the errors induced by the introduction of the Euler-Maruryama scheme on the posterior distribution and on the posterior mean decrease linearly when the step size h of the numerical scheme goes to zero. The proofs of these results are almost the same as the ones given in this paper.

Appendix A. Proof of Theorem 2

A.1. **Proof of part 1**

We aim at bounding the total variation distance between $p_{z,\phi|y}$ and $q_{w,\phi|y}$ as a function of the step size h. $p_{z,\phi|y}$ and $q_{w,\phi|y}$ denote the joint posterior distributions of the diffusion and the individual parameters under models $\mathcal{M}$ and $\mathcal{M}_h$ respectively, (the quantity h is implicitly included in the notation $q_{w,\phi|y}$). The following notations are used: y_i is the vector of data of subject i at times $t_0, \dots, t_J$, x_i is a process trajectory of subject i observed at times $t_0, \dots, t_J$, *i.e.* $x_i = (x_{i,0}, \dots x_{i,J})$, $\forall i = 1 \dots I$ where $x_{i,j}$ is the observation for the subject i at time t_j.

Let us first decompose the quantity $\left\|p_{z,\phi|y} - q_{w,\phi|y}\right\|_{TV}$. Using the fact that the subjects $i = 1 \dots I$ are independent, we can write:

$$\begin{aligned}
\left\|p_{z,\phi|y} - q_{w,\phi|y}\right\|_{TV} &= \int \left| \prod_{i=1}^{I} p_{z,\phi|y}(x_i, \phi_i|y_i;\theta) - \prod_{i=1}^{I} q_{w,\phi|y}(x_i, \phi_i|y_i;\theta) \right| \mathrm{d}x_1 \dots \mathrm{d}x_I \mathrm{d}\phi_1 \dots \mathrm{d}\phi_I \\
&\leq \int \sum_{i=1}^{I} \left| p_{z,\phi|y}(x_i, \phi_i|y_i;\theta) - q_{w,\phi|y}(x_i, \phi_i|y_i;\theta) \right| \prod_{l=1}^{i-1} p_{z,\phi|y}(x_l, \phi_l|y_l;\theta) \\
&\quad \times \prod_{l=i+1}^{I} q_{w,\phi|y}(x_l, \phi_l|y_l;\theta) \mathrm{d}x_1 \dots \mathrm{d}x_I \mathrm{d}\phi_1 \dots \mathrm{d}\phi_I. \qquad (8)
\end{aligned}$$

We aim to bound every term of the sum of (8). The joint posterior distributions can be expressed as a function of the transition probabilities. As a consequence, the sketch of the proof is the following one:

(1) first, we propose two lemma (based on results of [28] and [3]) which supply bounds for the quantities $p_{z|\phi}\left(x_{i,j}|x_{i,j-1},\phi_i;\gamma^2\right)$ and $|q_{w|\phi}\left(x_{i,j}|x_{i,j-1},\phi_i;\gamma^2\right)-p_{z|\phi}\left(x_{i,j}|x_{i,j-1},\phi_i;\gamma^2\right)|$;
(2) then, we bound the quantities $\left|p_{z,\phi|y}(x_i,\phi_i|y_i;\theta)-q_{w,\phi|y}(x_i,\phi_i|y_i;\theta)\right|$ and $q_{w,\phi|y}(x_l,\phi_l|y_l;\theta)$ using the two previous lemma;
(3) in a third part, $q_{w,\phi|y}(x_l,\phi_l|y_l;\theta)$ is upper-bounded;
(4) at last, the total variation distance between $p_{z,\phi|y}$ and $q_{w,\phi|y}$ is proved to be equivalent to $O(h)$.

(1) We first recall the result deriving from [28] and [3] which supplies bounds for the quantity $p_{z|\phi}\left(x_{i,j}|x_{i,j-1},\phi_i;\gamma^2\right)$ independent of h.

Lemma 1. *There exists a non-negative constant $C_{1,J}$ independent of ϕ_i and γ^2, such that $\forall i=1\dots I$, $\forall j=1\dots J$,*

$$p_{z|\phi}\left(x_{i,j}|x_{i,j-1},\phi_i;\gamma^2\right)\leq C_{1,J}. \tag{9}$$

Proof. Following [28], there exists a non-negative constant $C_1(\phi_i,\gamma^2,t_j-t_{j-1})$ such that, $\forall i=1\dots I$, $\forall j=1\dots J$, $p_{z|\phi}\left(x_{i,j}|x_{i,j-1},\phi_i;\gamma^2\right)\leq C_1(\phi_i,\gamma^2,t_j-t_{j-1})$. This constant $C_1(\phi_i,\gamma^2,t_j-t_{j-1})$ depends on the volatility function and on the bounds of the derivatives of the drift and volatility functions, which are independent of ϕ_i under assumption (**A3**). As a consequence, assuming that γ^2 is contained in a compact $[\gamma_0,\Gamma_0]$, there exists a constant $C_1(t_j-t_{j-1})$ independent of γ^2 and ϕ_i such that, $\forall i=1\dots I$, $\forall j=1\dots J$, $p_{z|\phi}\left(x_{i,j}|x_{i,j-1},\phi_i;\gamma^2\right)\leq C_1(t_j-t_{j-1})$. Finally, $\forall i=1\dots I$, $\forall j=1\dots J$,

$$p_{z|\phi}\left(x_{i,j}|x_{i,j-1},\phi_i;\gamma^2\right)\leq \max_{j=1\dots J} C_1(t_j-t_{j-1}):=C_{1,J}.$$

□

Remark 5. If the time interval lengths $(t_j-t_{j-1})_{j=1\dots J}$ are independent of J, which is the case in practice, $C_{1,J}$ is independent of J.

We can obtain the same type of result for the quantity $|q_{w|\phi}\left(x_{i,j}|x_{i,j-1},\phi_i;\gamma^2\right)-p_{z|\phi}(x_{i,j}|x_{i,j-1},\phi_i;\gamma^2)|$.

Lemma 2. *There exists a non-negative constant $C_{2,J}$ independent of ϕ_i and γ^2, such that $\forall i=1\dots I$, $\forall j=1\dots J$,*

$$|q_{w|\phi}\left(x_{i,j}|x_{i,j-1},\phi_i;\gamma^2\right)-p_{z|\phi}\left(x_{i,j}|x_{i,j-1},\phi_i;\gamma^2\right)|\leq C_{2,J}h. \tag{10}$$

Proof. Bally & Talay propose a bound for these quantities in [3]. More precisely, using the assumption (**A3**) and the fact that the volatility function is constant (thus the Hörmander's condition detailed in [3] is verified), for each subject i, there exists a non-negative constant $C_2(\phi_i,\gamma^2,t_j-t_{j-1})$ independent of h, $x_{i,j}$ and $x_{i,j-1}$ such that $|q_{w|\phi}\left(x_{i,j}|x_{i,j-1},\phi_i;\gamma^2\right)-p_{z|\phi}\left(x_{i,j}|x_{i,j-1},\phi_i;\gamma^2\right)|\leq C_2(\phi_i,\gamma^2,t_j-t_{j-1})h$. With the same arguments as those used in lemma 1, there exists a constant $C_{2,J}$ independent of ϕ_i and γ^2 bounding $C_2(\phi_i,\gamma^2,t_j-t_{j-1})$ for all ϕ_i and $\gamma^2\in[\gamma_0,\Gamma_0]$. □

(2) To bound the quantities $\left|p_{z,\phi|y}(x_i,\phi_i|y_i;\theta)-q_{w,\phi|y}(x_i,\phi_i|y_i;\theta)\right|$ of (8), we first apply the Bayes formula to obtain

$$\left|p_{z,\phi|y}(x_i,\phi_i|y_i;\theta)-q_{w,\phi|y}(x_i,\phi_i|y_i;\theta)\right|\leq$$
$$\frac{p_{y|z}(x_i;\sigma^2)\pi(\phi_i;\beta)}{p_y(y_i;\theta)}\left[\underbrace{\left|p_{z|\phi}(x_i|\phi_i;\gamma^2)-q_{w|\phi}(x_i|\phi_i;\gamma^2)\right|}_{b_i}+\overbrace{\frac{q_{w|\phi}(x_i|\phi_i;\gamma^2)}{\underbrace{q_y(y_i;\theta)}_{d_i}}}^{a_i}\underbrace{|p_y(y_i;\theta)-q_y(y_i;\theta)|}_{c_i}\right], \tag{11}$$

using the fact that the conditional distributions $p_{y|z}(y_i|x_i;\sigma^2)$ and $q_{y|w}(y_i|x_i;\sigma^2)$ are equal. Now, we bound the quantities a_i, b_i, c_i and finally d_i.

- To bound a_i, we write:

$$\begin{aligned} a_i &= q_{w|\phi}(x_i|\phi_i;\gamma^2) = \prod_{j=1}^{J} q_{w|\phi}\left(x_{i,j}|x_{i,j-1},\phi_i;\gamma^2\right) \qquad \text{(by the Markov property)} \\ &\leq \prod_{j=1}^{J}\left[\left|q_{w|\phi}\left(x_{i,j}|x_{i,j-1},\phi_i;\gamma^2\right) - p_{z|\phi}\left(x_{i,j}|x_{i,j-1},\phi_i;\gamma^2\right)\right| + p_{z|\phi}\left(x_{i,j}|x_{i,j-1},\phi_i;\gamma^2\right)\right] \\ &\leq (C_{1,J} + C_{2,J}h)^J \qquad \text{(by Lems. 1 and 2)}\ . \end{aligned} \tag{12}$$

- To bound b_i, we first use the Markov property to write:

$$\begin{aligned} b_i &= \left|p_{z|\phi}(x_i|\phi_i;\gamma^2) - q_{w|\phi}(x_i|\phi_i;\gamma^2)\right| \\ &= \left|\prod_{j=1}^{J} p_{z|\phi}\left(x_{i,j}|x_{i,j-1},\phi_i;\gamma^2\right) - \prod_{j=1}^{J} q_{w|\phi}\left(x_{i,j}|x_{i,j-1},\phi_i;\gamma^2\right)\right| \\ &\leq \sum_{j=1}^{J}\left|p_{z|\phi}\left(x_{i,j}|x_{i,j-1},\phi_i;\gamma^2\right) - q_{w|\phi}\left(x_{i,j}|x_{i,j-1},\phi_i;\gamma^2\right)\right| \\ &\quad \times \prod_{k=1}^{j-1} p_{z|\phi}\left(x_{i,k}|x_{i,k-1},\phi_i;\gamma^2\right) \prod_{k=j+1}^{J} q_{w|\phi}\left(x_{i,k}|x_{i,k-1},\phi_i;\gamma^2\right). \end{aligned} \tag{13}$$

Then, using the Lems. 1 and 2, inequality (13) becomes:

$$b_i \leq \sum_{j=1}^{J} C_{2,J}h \prod_{k=1}^{j-1} C_{1,J} \prod_{k=j+1}^{J} (C_{1,J} + C_{2,J}h) =: C_{3,J}h. \tag{14}$$

Like the constants $C_{1,J}$ and $C_{2,J}$, the constant $C_{3,J}$ is independent of ϕ_i and γ^2.

- To bound c_i, using the inequality (14), we have:

$$\begin{aligned} c_i &= |p_y(y_i;\theta) - q_y(y_i;\theta)| \\ &\leq \int\int p_{y|z}(x_i;\sigma^2)\underbrace{\left|p_{z|\phi}(x_i|\phi_i;\gamma^2) - q_{w|\phi}(x_i|\phi_i;\gamma^2)\right|}_{b_i}\pi(\phi_i;\beta)\mathrm{d}x_i\mathrm{d}\phi_i \\ &\leq C_{3,J}h\int\int p_{y|z}(x_i;\sigma^2)\pi(\phi_i;\beta)\mathrm{d}x_i\mathrm{d}\phi_i = C_{3,J}h. \end{aligned} \tag{15}$$

- The quantity $d_i = q_y(y_i;\theta)$ can be bounded from below. Indeed,

$$\begin{aligned} q_y(y_i;\theta) &\geq p_y(y_i;\theta) - \underbrace{|p_y(y_i;\theta) - q_y(y_i;\theta)|}_{c_i} \\ &\geq p_y(y_i;\theta) - C_{3,J}h \qquad \text{by the inequality (15)} \\ &\geq p_y(y_i;\theta) - C_{3,J}H_0 \qquad \text{with } h < H_0. \end{aligned}$$

At last, let $C_{5,y}$ denote $\min_{i=1\ldots I}(p_y(y_i;\theta)) - C_{3,J}H_0$. For H_0 small enough, $C_{5,y}$ is non-negative and we have

$$d_i = q_y(y_i;\theta) \geq C_{5,y} > 0. \tag{16}$$

- Finally, by inequations (12), (14), (15) and (16) we are able to upper-bound the inequation (11):

$$\begin{aligned}\left|p_{z,\phi|y}(x_i,\phi_i|y_i;\theta)-q_{w,\phi|y}(x_i,\phi_i|y_i;\theta)\right| &\leq \frac{p_{y|z}(x_i;\sigma^2)\pi(\phi_i;\beta)}{p_y(y_i;\theta)}\left[b_i+\frac{a_i}{d_i}c_i\right]\\ &\leq \frac{p_{y|z}(x_i;\sigma^2)\pi(\phi_i;\beta)}{p_y(y_i;\theta)}\left[C_{3,J}h+\frac{(C_{1,J}+C_{2,J}h)^J}{C_{5,y}}C_{3,J}h\right]. \qquad (17)\end{aligned}$$

(3) The bound on $q_{w,\phi|y}(x_i,\phi_i|y_i;\theta)$ can be deduced from inequalities (14) and (15):

$$\begin{aligned}q_{w,\phi|y}(x_i,\phi_i|y_i;\theta) &= \frac{p_{y|z}(x_i;\sigma^2)\pi(\phi_i;\beta)}{q_y(y_i;\theta)}q_{w|\phi}\left(x_i|\phi_i;\gamma^2\right)\\ &\leq \frac{p_{y|z}(x_i;\sigma^2)\pi(\phi_i;\beta)}{C_{5,y}}(C_{1,J}+C_{2,J}h)^J. \qquad (18)\end{aligned}$$

(4) The total variation distance can now be bounded. By inequation (8), we have:

$$\begin{aligned}\left\|p_{z,\phi|y}-q_{w,\phi|y}\right\|_{TV} \leq \int\cdots\int\Bigg(&\sum_{i=1}^{I}\left|p_{z,\phi|y}(x_i,\phi_i|y_i;\theta)-q_{w,\phi|y}(x_i,\phi_i|y_i;\theta)\right|\\ &\times\prod_{l=1}^{i-1}p_{z,\phi|y}(x_l,\phi_l|y_l;\theta)\prod_{l=i+1}^{I}q_{w,\phi|y}(x_l,\phi_l|y_l;\theta)\Bigg)\mathrm{d}x_1\ldots\mathrm{d}x_I\mathrm{d}\phi_1\ldots\mathrm{d}\phi_I.\end{aligned}$$

And so, using inequalities (17) and (18) we can write:

$$\begin{aligned}&\|p_{z,\phi|y}-q_{w,\phi|y}\|_{TV}\\ &\leq\int\cdots\int\Bigg(\sum_{i=1}^{I}\frac{p_{y|z}(x_i;\sigma^2)\pi(\phi_i;\beta)}{p_y(y_i;\theta)}\left[C_{3,J}h+\frac{(C_{1,J}+C_{2,J}h)^J}{C_{5,y}}C_{3,J}h\right]\\ &\quad\times\prod_{l=1}^{i-1}p_{z,\phi|y}(x_l,\phi_l|y_l;\theta)\prod_{l=i+1}^{I}\left[\frac{p_{y|z}(x_l;\sigma^2)\pi(\phi_l;\beta)}{C_{5,y}}(C_{1,J}+C_{2,J}h)^J\right]\Bigg)\mathrm{d}x_1\ldots\mathrm{d}x_I\mathrm{d}\phi_1\ldots\mathrm{d}\phi_I\\ &\leq\left[C_{3,J}h+\frac{(C_{1,J}+C_{2,J}h)^J}{C_{5,y}}C_{3,J}h\right]\sum_{i=1}^{I}\left(\frac{(C_{1,J}+C_{2,J}h)^J}{C_{5,y}}\right)^{I-i}\frac{1}{p_y(y_i;\theta)}\times\underbrace{\int p_{y|z}(x_i;\sigma^2)\mathrm{d}x_i}_{=1}\\ &\quad\times\underbrace{\int\pi(\phi_i;\beta)\mathrm{d}\phi_i}_{=1}\times\prod_{l=1}^{i-1}\underbrace{\int\int p_{z,\phi|y}(x_l,\phi_l|y_l;\theta)\mathrm{d}x_l\mathrm{d}\phi_l}_{=1}\times\prod_{l=i+1}^{I}\underbrace{\int\int\left[p_{y|z}(x_l;\sigma^2)\pi(\phi_l;\beta)\right]\mathrm{d}x_l\mathrm{d}\phi_l}_{=1}\\ &\leq\left[C_{3,J}h+\frac{(C_{1,J}+C_{2,J}h)^J}{C_{5,y}}C_{3,J}h\right]\sum_{i=1}^{I}\left(\frac{(C_{1,J}+C_{2,J}h)^J}{C_{5,y}}\right)^{I-i}\frac{1}{p_y(y_i;\theta)}=:C(y)h,\end{aligned}$$

where $C(y)$ is independent of h if $h<H_0$. □

A.2. **Proof of part 2**

In this part, we aim at bounding the distance between the likelihoods p_y and q_y of the models $\mathcal{M}$ and $\mathcal{M}_h$. We have proved in (15), that there exists $C_{3,J}$ such that:

$$|p_y(y_i;\theta) - q_y(y_i;\theta)| \leq C_{3,J}h,$$

where $C_{3,J}$ is a function of $C_{1,J}$ and $C_{2,J}$ (see equation (14)). $C_{1,J}$ and $C_{2,J}$ are independent of the parameters θ and so does $C_{3,J}$. As a consequence, we can bound the error induced by the numerical scheme on the likelihood functions:

$$\begin{aligned}|p_y(y;\theta) - q_y(y;\theta)| &= \left|\prod_{i=1}^{I} p_y(y_i;\theta) - \prod_{i=1}^{I} q_y(y_i;\theta)\right| \\ &\leq \sum_{i=1}^{I} |p_y(y_i;\theta) - q_y(y_i;\theta)| \prod_{l=1}^{i-1} p_y(y_l;\theta) \prod_{l=i+1}^{I} q_y(y_l;\theta).\end{aligned}$$

The quantity $p_y(y_i;\theta)$ can be bounded as follows:

$$\begin{aligned}p_y(y_i;\theta) &= \int p_{y|z}(y_i|x_i;\theta) p_{z|\phi}(x_i|\phi_i;\theta)\pi(\phi_i;\beta)\mathrm{d}x_i\mathrm{d}\phi_i \\ &\leq \int p_{y|z}(y_i|x_i;\theta) C_{1,J}^J \pi(\phi_i;\beta)\mathrm{d}x_i\mathrm{d}\phi_i \qquad \text{by Lemma 1} \\ &\leq C_{1,J}^J.\end{aligned}$$

As a consequence:

$$\begin{aligned}q_y(y_i;\theta) &\leq |p_y(y_i;\theta) - q_y(y_i;\theta)| + p_y(y_i;\theta) \\ &\leq C_{3,J}h + C_{1,J}^J \qquad \text{by inequality (15).}\end{aligned}$$

At last, there exists a non-negative constant $M_{I,J}$ such that, for every $\theta \in \left\{\theta = (\beta, \sigma^2, \gamma^2), \gamma_0^2 < \gamma^2 < \Gamma_0^2\right\}$,

$$\begin{aligned}|p_y(y;\theta) - q_y(y;\theta)| &\leq \sum_{i=1}^{I} C_{3,J}h(C_{1,J}^J)^{i-1}(C_{3,J}h + C_{1,J}^J)^{I-i-1} \\ &\leq M_{I,J}h. \end{aligned} \tag{19}$$

A.3. **Proof of corollary 1**

Let θ_∞ and $\theta_{h,\infty}$ be the maxima of $p_y(y;\theta)$ and $q_y(y;\theta)$ respectively. We aim at bounding the distance between θ_∞ and $\theta_{h,\infty}$ induced by the Euler-Maruyama approximation.

By the Taylor's theorem applied to $p_y(y;\theta)$ and $q_y(y;\theta)$, we can write:

$$\begin{aligned}p_y(y;\theta_{h,\infty}) &= p_y(y;\theta_\infty) + (\theta_{h,\infty} - \theta_\infty)^t \left[\int_0^1 (1-t) H_{p_y}\left\{\theta_{h,\infty} + t(\theta_{h,\infty} - \theta_\infty)\right\}\mathrm{d}t\right](\theta_{h,\infty} - \theta_\infty), \\ q_y(y;\theta_\infty) &= q_y(y;\theta_{h,\infty}) + (\theta_{h,\infty} - \theta_\infty)^t \left[\int_0^1 (1-t) H_{q_y}\left\{\theta_\infty + t(\theta_\infty - \theta_{h,\infty})\right\}\mathrm{d}t\right](\theta_{h,\infty} - \theta_\infty).\end{aligned}$$

As a consequence, we can write:

$$\begin{aligned}
p_y(y;\theta_\infty) - p_y(y;\theta_{h,\infty}) &+ q_y(y;\theta_{h,\infty}) - q_y(y;\theta_\infty) \\
&= -(\theta_{h,\infty}-\theta_\infty)^t \int_0^1 (1-t) H_{p_y}\left\{\theta_{h,\infty} + t(\theta_{h,\infty}-\theta_\infty)\right\} \mathrm{d}t (\theta_{h,\infty}-\theta_\infty) \\
&\quad -(\theta_{h,\infty}-\theta_\infty)^t \int_0^1 (1-t) H_{q_y}\left\{\theta_\infty + t(\theta_\infty-\theta_{h,\infty})\right\} \mathrm{d}t (\theta_{h,\infty}-\theta_\infty) \\
&\geq \varepsilon_1 \left\|\theta_{h,\infty}-\theta_\infty\right\|^2 \int_0^1 (1-t)\mathrm{d}t + \varepsilon_2 \left\|\theta_{h,\infty}-\theta_\infty\right\|^2 \int_0^1 (1-t)\mathrm{d}t \\
&\geq \frac{(\varepsilon_1+\varepsilon_2)}{2} \left\|\theta_{h,\infty}-\theta_\infty\right\|^2,
\end{aligned}$$

by assumption (**A4**). Therefore, we have :

$$\begin{aligned}
\left\|\theta_{h,\infty}-\theta_\infty\right\|^2 &\leq \frac{2}{(\varepsilon_1+\varepsilon_2)} \left(p_y(y;\theta_\infty) - p_y(y;\theta_{h,\infty})\right) + \left(q_y(y;\theta_{h,\infty}) - q_y(y;\theta_\infty)\right) \\
&\leq \frac{4}{(\varepsilon_1+\varepsilon_2)} \|p_y(y;.) - q_y(y;.)\|_\infty \\
&\leq \frac{4}{\varepsilon_1+\varepsilon_2} M_{I,J} h \qquad \text{by inequation (19).}
\end{aligned}$$

Finally,

$$\left\|\theta_{h,\infty}-\theta_\infty\right\|^2 \leq \frac{4}{\varepsilon_1+\varepsilon_2} M_{I,J} h.$$

References

[1] Y. Aït-Sahalia, Maximum likelihood estimation of discretely sampled diffusions: a closed-form approximation approach. *Econometrica* **70** (2002) 223–262.

[2] C. Andrieu and E. Moulines, On the ergodicity properties of some adaptive MCMC algorithms. *Ann. Appl. Prob.* **16** (2006) 1462–1505.

[3] V. Bally and D. Talay, *The law of the Euler Scheme for Stochastic Differential Equations: I. Convergence Rate of the Density.* Technical Report 2675, INRIA (1995).

[4] V. Bally and D. Talay, The law of the Euler scheme for stochastic differential equations (II): convergence rate of the density. *Monte Carlo Methods Appl.* **2** (1996) 93–128.

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

[6] J.E. Bennet, A. Racine-Poon and J.C. Wakefield, *MCMC for nonlinear hierarchical models.* Chapman & Hall, London (1996) 339–358.

[7] A. Beskos, O. Papaspiliopoulos, G.O. Roberts and P. Fearnhead, Exact and computationally efficient likelihood-based estimation for discretely observed diffusion processes. *J. R. Stat. Soc. B* **68** (2006) 333–382.

[8] A. Beskos and G.O. Roberts, Exact simulation of diffusions. *Ann. Appl. Prob.* **15** (2005) 2422–2444.

[9] B.M. Bibby and M. Sørensen, Martingale estimation functions for discretely observed diffusion processes. *Bernoulli* **1** (1995) 17–39.

[10] G. Celeux and J. Diebolt, The SEM algorithm: a probabilistic teacher algorithm derived from the EM algorithm for the mixture problem. *Computational. Statistics Quaterly* **2** (1985) 73–82.

[11] D. Dacunha-Castelle and M. Duflo, *Probabilités et statistiques. Tome 2.* Collection Mathématiques Appliquées pour la Maîtrise. [Collection of Applied Mathematics for the Master's Degree]. Masson, Paris, 1983. Problèmes à temps mobile. [Movable-time problems].

[12] D. Dacunha-Castelle and D. Florens-Zmirou, Estimation of the coefficients of a diffusion from discrete observations. *Stochastics* **19** (1986) 263–284.

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

[14] A. Dembo and O. Zeitouni, Parameter estimation of partially observed continuous time stochastic processes via the EM algorithm. *Stoch. Process. Appl.* **23** (1986) 91–113.
[15] A P. Dempster, N.M. Laird and D.B. Rubin, Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Statist. Soc. Ser. B* **39** (1977) 1–38. With discussion.
[16] S. Ditlevsen and A. De Gaetano, Mixed effects in stochastic differential equation models. *REVSTAT- Statistical Journal* **3** (2005) 137–153.
[17] S. Donnet and A. Samson, Estimation of parameters in incomplete data models defined by dynamical systems. *J. Stat. Plan. Inf.* (2007).
[18] R. Douc and C. Matias, Asymptotics of the maximum likelihood estimator for general hidden Markov models. *Bernoulli* **7** (2001) 381–420.
[19] O. Elerian, S. Chib and N. Shephard, Likelihood inference for discretely observed nonlinear diffusions. *Econometrica* **69** (2001) 959–993.
[20] B. Eraker, MCMC analysis of diffusion models with application to finance. *J. Bus. Econ. Statist.* **19** (2001) 177–191.
[21] V. Genon-Catalot and J. Jacod, On the estimation of the diffusion coefficient for multi-dimensional diffusion processes. *Ann. Inst. H. Poincaré Probab. Statist.* **29** (1993) 119–151.
[22] A. Gloter and J. Jacod, Diffusions with measurement errors. I. Local asymptotic normality. *ESAIM: PS* **5** (2001) 225–242.
[23] A. Gloter and J. Jacod, Diffusions with measurement errors. II. Optimal estimators. *ESAIM: PS* **5** (2001) 243–260 (electronic).
[24] M. Kessler, Estimation of an ergodic diffusion from discrete observations. *Scand. J. Statist.* **24** (1997) 211–229.
[25] R. Krishna, *Applications of Pharmacokinetic principles in drug development*. Kluwer Academic/Plenum Publishers, New York (2004).
[26] E. Kuhn and M. Lavielle, Coupling a stochastic approximation version of EM with a MCMC procedure. *ESAIM: PS* **8** (2004) 115–131.
[27] E. Kuhn and M. Lavielle, Maximum likelihood estimation in nonlinear mixed effects models. *Comput. Statist. Data Anal.* **49** (2005) 1020–1038.
[28] S. Kusuoka and D. Stroock, Applications of the malliavin calculus, part II. *J. Fac. Sci. Univ. Tokyo. Sect. IA, Math.* **32** (1985) 1–76.
[29] T. Kutoyants, *Parameter estimation for stochastic processes*. Helderman Verlag Berlin (1984).
[30] M.L. Lindstrom and D.M. Bates, Nonlinear mixed effects models for repeated measures data. *Biometrics* **46** (1990) 673–687.
[31] T.A. Louis, Finding the observed information matrix when using the EM algorithm. *J. Roy. Statist. Soc. Ser. B* **44** (1982) 226–233.
[32] R.V. Overgaard, N. Jonsson, C.W. Tornøe and H. Madsen, Non-linear mixed-effects models with stochastic differential equations: Implementation of an estimation algorithm. *J Pharmacokinet. Pharmacodyn.* **32** (2005) 85–107.
[33] A.R. Pedersen, A new approach to maximum likelihood estimation for stochastic differential equations based on discrete observations. *Scand. J. Statist.* **22** (1995) 55–71.
[34] J.C. Pinheiro and D.M. Bates, Approximations to the log-likelihood function in the non-linear mixed-effect models. *J. Comput. Graph. Statist.* **4** (1995) 12–35.
[35] R. Poulsen, Approximate maximum likelihood estimation of discretely observed diffusion process. *Center for Analytical Finance*, Working paper 29 (1999).
[36] B.L.S. Prakasa Rao, *Statistical Inference for Diffusion Type Processes*. Arnold Publisher (1999).
[37] G.O. Roberts and O. Stramer, On inference for partially observed nonlinear diffusion models using the Metropolis-Hastings algorithm. *Biometrika* **88** (2001) 603–621.
[38] F. Schweppe, Evaluation of likelihood function for gaussian signals. *IEEE Trans. Inf. Theory* **11** (1965) 61–70.
[39] H. Singer, Continuous-time dynamical systems with sampled data, error of measurement and unobserved components. *J. Time Series Anal.* **14** (1993) 527–545.
[40] H. Sørensen, Parametric inference for diffusion processes observed at discrete points in time: a survey. *Int. Stat. Rev* **72** (2004) 337–354.
[41] M. Sørensen, Prediction-based estimating functions. *Econom. J.* **3** (2000) 123–147.
[42] L. Tierney, Markov chains for exploring posterior distributions. *Ann. Statist.* **22** (1994) 1701–1762.
[43] C.W. Tornøe, R.V. Overgaard, H. Agersø, H.A. Nielsen, H. Madsen and E.N. Jonsson, Stochastic differential equations in NONMEM: implementation, application, and comparison with ordinary differential equations. *Pharm. Res.* **22** (2005) 1247–1258.
[44] G.C.G. Wei and M.A. Tanner, Calculating the content and boundary of the highest posterior density region via data augmentation. *Biometrika* **77** (1990) 649–652.
[45] R. Wolfinger, Laplace's approximation for nonlinear mixed models. *Biometrika* **80** (1993) 791–795.