f-SAEM: A fast Stochastic Approximation of the EM algorithm for nonlinear mixed effects models

Belhal Karimi[a,b,*], Marc Lavielle[a,b], Eric Moulines[a,b]

[a] *CMAP, Ecole Polytechnique, route de Saclay, 91120 Palaiseau, France*

[b] *INRIA Saclay, 1 Rue Honoré d'Estienne d'Orves, 91120 Palaiseau, France*

Abstract

The ability to generate samples of the random effects from their conditional distributions is fundamental for inference in mixed effects models. Random walk Metropolis is widely used to perform such sampling, but this method is known to converge slowly for medium dimensional problems, or when the joint structure of the distributions to sample is spatially heterogeneous. The main contribution consists of an independent Metropolis-Hastings (MH) algorithm based on a multidimensional Gaussian proposal that takes into account the joint conditional distribution of the random effects and does not require any tuning. Indeed, this distribution is automatically obtained thanks to a Laplace approximation of the incomplete data model. Such approximation is shown to be equivalent to linearizing the structural model in the case of continuous data. Numerical experiments based on simulated and real data illustrate the performance of the proposed methods. For fitting nonlinear mixed effects models, the suggested MH algorithm is efficiently combined with a stochastic approximation version of the EM algorithm for maximum likelihood estimation of the global parameters.

Keywords: MCMC, Stochastic approximation, EM, mixed effects, Laplace approximation

1. Introduction

Mixed effects models are often adopted to take into account the inter-individual variability within a population (see (Lavielle, 2014) and the references therein). Consider a study with N individuals from a same population. The vector of observations y_i associated to each individual i is assumed to be a realisation of a random variable which

[*] Corresponding author.
Email address: `belhal.karimi@inria.fr` (Belhal Karimi)

depends on a vector of random individual parameters ψ_i. Then, inference on the individual parameter ψ_i amounts to estimate its conditional distribution given the observed data y_i.

When the model is a linear (mixed effects) Gaussian model, then this conditional distribution is a normal distribution that can explicitly be computed (Verbeke, 1997). For more complex distributions and models, Monte Carlo methods must be used to approximate this conditional distribution. Most often, direct sampling from this conditional distribution is inefficient and it is necessary to resort to a Markov chain Monte Carlo (MCMC) method for obtaining random samples from this distribution. Yet, MCMC requires a tractable likelihood in order to compute the acceptance ratio. When this computation is impossible, a pseudo-marginal Metropolis Hastings (PMMH) has been developed in (Andrieu et al., 2009) and consists in replacing the posterior distribution evaluated in the MH acceptance rate by an unbiased approximation. An extension of the PMMH is the particle MCMC method, introduced in (Andrieu et al., 2010), where a Sequential Monte Carlo sampler (Doucet et al., 2000) is used to approximate the intractable likelihood at each iteration. For instance, this method is relevant when the model is SDE-based (see (Donnet and Samson, 2013)). In a fully Bayesian setting, approximation of the posterior of the global parameters can be used to approximate the posterior of the individual parameters using Integrated Nested Laplace Approximation (INLA) introduced in (Rue et al., 2009). When the goal is to do approximate inference, this method has shown great performances mainly because it approximates each marginal separately as univariate Gaussian distribution. In this paper, we focus on developing a method to perform exact inference and do not treat the case of approximate inference algorithms such as the Laplace EM or the First Order Conditional Estimation methods (Wang, 2007) that can introduce bias in the resulting parameters.

Note that generating random samples from $\mathtt{p}_i(\psi_i|y_i;\theta)$ is useful for several tasks to avoid approximation of the model, such as linearisation or Laplace method. Such tasks include the estimation of the population parameters θ of the model by a maximum likelihood approach, i.e. by maximizing the observed incomplete data likelihood $\mathtt{p}(y_1,\dots y_N;\theta)$ using the Stochastic Approximation of the EM algorithm (SAEM) algorithm combined with a MCMC procedure (Kuhn and Lavielle, 2004). Lastly, sampling from the conditional distributions $\mathtt{p}_i(\psi_i|y_i;\theta)$ is also known to be useful for model building. Indeed, in (Lavielle and Ribba, 2016), the authors argue that methods for model assessment and model validation, whether graphical or based on statistical tests, must use samples of the conditional distribution $\mathtt{p}_i(\psi_i|y_i;\theta)$ to avoid bias.

Designing a fast mixing sampler for these distributions is therefore of utmost importance to perform Maximum Likelihood Estimation (MLE) using the SAEM algorithm. The most common MCMC method for nonlinear mixed effects (NLME) models is the *random walk Metropolis* (RWM) algorithm (Robert and Casella, 2010; Roberts et al., 1997; Lavielle, 2014). This method is implemented in software tools such as Monolix, NONMEM, the saemix R package (Comets et al., 2017) and the nlmefitsa Matlab function. Despite its simplicity, it has been successfully used in many classical examples of pharmacometrics. Nevertheless, it can show its limitations when the dependency structure of the individual parameters is complex. Yet, maintaining an optimal acceptance rate (advocated in Roberts and Rosenthal (1997)) most often implies very small moves and therefore a very large number of iterations in medium and high dimensions since no information of the geometry of the target distribution is used.

The Metropolis-adjusted Langevin algorithm (MALA) uses evaluations of the gradient of the target density for proposing new states which are accepted or rejected using the Metropolis-Hastings algorithm (Roberts and Tweedie, 1996; Stramer and Tweedie, 1999). Hamiltonian Monte Carlo (HMC) is another MCMC algorithm that exploits information about the geometry of the target distribution in order to efficiently explore the space by selecting transitions that can follow contours of high probability mass (Betancourt, 2017). The No-U-Turn Sampler (NUTS) is an extension to HMC that allows an automatic and optimal selection of some of the settings required by the algorithm, (Brooks et al., 2011; Hoffman and Gelman, 2014). Nevertheless, these methods may be difficult to use in practice, and are computationally involved, in particular when the structural model is a complex ODE based model. The algorithm we propose is an independent Metropolis-Hastings (IMH) algorithm, but for which the proposal is a Gaussian approximation of the target distribution. For general data model (i.e. categorical, count or time-to-event data models or continuous data models), the Laplace approximation of the incomplete pdf $\mathtt{p}_i(y_i;\theta)$ leads to a Gaussian approximation of the conditional distribution $\mathtt{p}_i(\psi_i|y_i;\theta)$.

In the special case of continuous data, linearisation of the model leads, by definition, to a Gaussian linear model for which the conditional distribution of the individual parameter ψ_i given the data y_i is a multidimensional normal distribution that can be computed. Therefore, we design an independent sampler using this multivariate Gaussian distribution to sample from the target conditional distribution and embed this procedure in an exact inference algorithm, the SAEM, to speed the convergence of the vector of estimations of the global parameters $\hat{\theta}$.

The paper is organised as follows. Mixed effects models for continuous and noncontinuous data are presented in Section 2. The standard MH for NLME models is described in Section 3. The proposed method, called the nlme-IMH, is introduced in Section 4 as well as the f-SAEM, a combination of this new method with the SAEM algorithm for estimating the population parameters of the model. Numerical examples illustrate, in Section 5, the practical performances of the proposed method, both on a continuous pharmacokinetics (PK) model and a time-to-event example. A Monte Carlo study confirms that this new SAEM algorithm shows a faster convergence to the maximum likelihood estimate.

2. Mixed Effect Models

2.1. Population approach and hierarchical models

In the sequel, we adopt a population approach, where we consider N individuals and n_i observations per individual i. The set of observed data is $y = (y_i, 1 \leq i \leq N)$ where $y_i = (y_{ij}, 1 \leq j \leq n_i)$ are the observations for individual i. For the sake of clarity, we assume that each observation y_{ij} takes its values in some subset of $\mathbb{R}$. The distribution of the n_i-vector of observations y_i depends on a vector of individual parameters ψ_i that takes its values in a subset of $\mathbb{R}^p$.

We assume that the pairs (y_i, ψ_i) are mutually independent and consider a parametric framework: the joint distribution of (y_i, ψ_i) is denoted by $\mathtt{p}_i(y_i, \psi_i; \theta)$, where θ is the vector of parameters of the model. A natural decomposition of this joint distribution reads

$$\mathtt{p}_i(y_i, \psi_i; \theta) = \mathtt{p}_i(y_i|\psi_i; \theta)\mathtt{p}_i(\psi_i; \theta) \ , \tag{1}$$

where $\mathtt{p}_i(y_i|\psi_i; \theta)$ is the conditional distribution of the observations given the individual parameters, and where $\mathtt{p}_i(\psi_i; \theta)$ is the so-called population distribution used to describe the distribution of the individual parameters within the population.

A particular case of this general framework consists in describing each individual parameter ψ_i as the sum of a typical value ψ_{pop} and a vector of individual random effects η_i:

$$\psi_i = \psi_{\text{pop}} + \eta_i \ . \tag{2}$$

In the sequel, we assume that the random effects are distributed according to a multivariate Gaussian distribution: $\eta_i \sim_{\text{i.i.d.}} \mathcal{N}(0, \Omega)$. Extensions of this general model are detailed in Appendix A.

2.2. Continuous data models

A regression model is used to express the link between continuous observations and individual parameters:

$$y_{ij} = f_i(t_{ij}, \psi_i) + \varepsilon_{ij} \ , \tag{3}$$

where y_{ij} is the j-th observation for individual i measured at index t_{ij}, ε_{ij} is the residual error. It is assumed that for any index t, $\psi \rightarrow f_i(t, \psi)$ is twice differentiable in ψ.

We start by assuming that the residual errors are independent and normally distributed with zero-mean and a constant variance σ^2. Let $t_i = (t_{ij}, 1 \leq n_i)$ be the vector of observation indices for individual i. Then, the model for the observations reads:

$$y_i|\psi_i \sim \mathcal{N}(f_i(\psi_i), \sigma^2 \mathtt{Id}_{n_i \times n_i}) \quad \text{where} \quad f_i(\psi_i) = (f_i(t_{i,1}, \psi_i), \ldots, f_i(t_{i,n_i}, \psi_i)) \ . \tag{4}$$

If we assume that $\psi_i \sim_{\text{i.i.d.}} \mathcal{N}(\psi_{\text{pop}}, \Omega)$, then the parameters of the model are $\theta = (\psi_{\text{pop}}, \Omega, \sigma^2)$.

Remark 1. *An extension of this model consists in assuming that the variance of the residual errors is not constant over time, i.e., $\varepsilon_{ij} \sim \mathcal{N}(0, g(t_{ij}, \psi_i)^2)$. Such extension includes proportional error models ($g = bf$) and combined error models ($g = a + bf$) (Lavielle, 2014) but the proposed method remains the same whatever the residual error model is.*

2.3. Noncontinuous data models

Noncontinuous data models include categorical data models (Savic et al., 2011; Agresti, 1990), time-to-event data models (Mbogning et al., 2015; Andersen, 2006), or count data models (Savic et al., 2011). A categorical outcome y_{ij} takes its value in a set $\{1, \dots, L\}$ of L categories. Then, the model is defined by the conditional probabilities $(\mathbb{P}(y_{ij} = \ell|\psi_i), 1 \leq \ell \leq L)$, that depend on the vector of individual parameters ψ_i and may be a function of the time t_{ij}.

In a time-to-event data model, the observations are the times at which events occur. An event may be one-off (e.g., death, hardware failure) or repeated (e.g., epileptic seizures, mechanical incidents). To begin with, we consider a model for a one-off event. The survival function $S(t)$ gives the probability that the event happens after time t:

$$S(t) \triangleq \mathbb{P}(T > t) = \exp\left\{-\int_0^t h(u)\mathrm{d}u\right\} \ , \tag{5}$$

where h is called the hazard function. In a population approach, we consider a parametric and individual hazard function $h(\cdot, \psi_i)$. The random variable representing the time-to-event for individual i is typically written T_i and may possibly be right-censored. Then, the observation y_i for individual i is

$$y_i = \left\{ \begin{array}{ll} T_i & \text{if } T_i \leq \tau_c \\ "T_i > \tau_c" & \text{otherwise ,} \end{array} \right. \tag{6}$$

where τ_c is the censoring time and "$T_i > \tau_c$" is the information that the event occurred after the censoring time.

For repeated event models, times when events occur for individual i are random times $(T_{ij}, 1 \leq j \leq n_i)$ for which conditional survival functions can be defined:

$$\mathbb{P}(T_{ij} > t|T_{i(j-1)} = t_{i(j-1)}) = \exp\left\{-\int_{t_{i(j-1)}}^t h(u, \psi_i)\mathrm{d}u\right\} \ . \tag{7}$$

Here, t_{ij} is the observed value of the random time T_{ij}. If the last event is right censored, then the last observation y_{i,n_i} for individual i is the information that the censoring time has been reached "$T_{i,n_i} > \tau_c$". The conditional pdf of $y_i = (y_{ij}, 1 \leq n_i)$ reads (see (Lavielle, 2014) for more details)

$$\mathtt{p}_i(y_i|\psi_i) = \exp\left\{-\int_0^{\tau_c} h(u, \psi_i)\mathrm{d}u\right\} \prod_{j=1}^{n_i - 1} h(t_{ij}, \psi_i) \ . \tag{8}$$

3. Sampling from conditional distributions

3.1. The conditional distribution of the individual parameters

Once the conditional distribution of the observations $\mathtt{p}_i(y_i|\psi_i; \theta)$ and the marginal distribution of the individual parameters ψ_i are defined, the joint distribution $\mathtt{p}_i(y_i, \psi_i; \theta)$

and the conditional distribution $\mathtt{p}_i(\psi_i|y_i;\theta)$ are implicitly specified. This conditional distribution $\mathtt{p}_i(\psi_i|y_i;\theta)$ plays a crucial role for inference in NLME models.

One of the main task is to compute the maximum likelihood (ML) estimate of θ

$$\hat{\theta}_{\text{ML}} = \arg\max_{\theta \in \mathbb{R}^d} \mathcal{L}(\theta, y) \ , \tag{9}$$

where $\mathcal{L}(\theta, y) = \log \mathtt{p}(y;\theta)$. In NLME models, this optimization is solved by using a surrogate function defined as the conditional expectation of the complete data log-likelihood (McLachlan and Krishnan, 2007). The SAEM is an iterative procedure for ML estimation that requires to generate one or several samples from this conditional distribution at each iteration of the algorithm. Once the ML estimate $\hat{\theta}_{\text{ML}}$ has been computed, the observed Fisher information matrix noted $I(\hat{\theta}_{\text{ML}}, y) = -\nabla^2_\theta \mathcal{L}(\hat{\theta}_{\text{ML}}, y)$ can be derived thanks to the Louis formula (Louis, 1982) which expresses $I(\hat{\theta}_{\text{ML}}, y)$ in terms of the conditional expectation and covariance of the complete data log-likelihood. Such procedure also requires to sample from the conditional distributions $\mathtt{p}_i(\psi_i|y_i;\hat{\theta}_{\text{ML}})$ for all $i \in \{1, \dots, N\}$.

Samples from the conditional distributions might also be used to define several statistical tests and diagnostic plots for models assessment. It is advocated in (Lavielle and Ribba, 2016) that such samples should be preferred to the modes of these distributions (also called *Empirical Bayes Estimate*(EBE), or *Maximum a Posteriori Estimate*), in order to provide unbiased tests and plots. For instance, a strong bias can be observed when the EBEs are used for testing the distribution of the parameters or the correlation between random effects.

In short, being able to sample individual parameters from their conditional distribution is essential in nonlinear mixed models. It is therefore necessary to design an efficient method to sample from this distribution.

3.2. The Metropolis-Hastings Algorithm

Metropolis-Hasting (MH) algorithm is a powerful MCMC procedure widely used for sampling from a complex distribution (Brooks et al., 2011). To simplify the notations, we remove the dependency on θ. For a given individual $i \in \{1, \dots, N\}$, the MH algorithm, to sample from the conditional distribution $\mathtt{p}_i(\psi_i|y_i)$, is described as:

Algorithm 1 Metropolis-Hastings algorithm

```
Initialization: Initialize the chain sampling $\psi_i^{(0)}$ from some initial distribution $\xi_i$ .
Iteration k: given the current state of the chain $\psi_i^{(k-1)}$:
    1. Sample a candidate $\psi_i^c$ from a proposal distribution $q_i(\,\cdot\,|\psi_i^{(k-1)})$.
    2. Compute the MH ratio:
           $$\alpha(\psi_i^{(k-1)}, \psi_i^c) = \frac{\mathtt{p}_i(\psi_i^c|y_i)}{\mathtt{p}_i(\psi_i^{(k-1)}|y_i)} \frac{q_i(\psi_i^{(k-1)}|\psi_i^c)}{q_i(\psi_i^c|\psi_i^{(k-1)})} \,. \tag{10}$$
    3. Set $\psi_i^{(k)} = \psi_i^c$ with probability $\min(1, \alpha(\psi_i^c, \psi_i^{(k-1)})$ (otherwise, keep $\psi_i^{(k)} = \psi_i^{(k-1)}$).
```

Under weak conditions, $(\psi_i^{(k)}, k \geq 0)$ is an ergodic Markov chain whose distribution converges to the target $\mathtt{p}_i(\psi_i|y_i)$ (Brooks et al., 2011).

Current implementations of the SAEM algorithm in Monolix (Chan et al., 2011), saemix (R package) (Comets et al., 2017), nlmefitsa (Matlab) and NONMEM (Beal and Sheiner, 1980) mainly use the same combination of proposals. The first proposal is an independent MH algorithm which consists in sampling the candidate state directly from the prior distribution of the individual parameter ψ_i. The MH ratio then reduces to $\mathtt{p}_i(y_i|\psi_i^c)/\mathtt{p}_i(y_i|\psi_i^{(k)})$ for this proposal.

The other proposals are component-wise and block-wise random walk procedures (Metropolis et al., 1953) that updates different components of ψ_i using univariate and multivariate Gaussian proposal distributions. These proposals are centered at the current state with a diagonal variance-covariance matrix; the variance terms are adaptively adjusted at each iteration in order to reach some target acceptance rate (Atchadé and Rosenthal, 2005; Lavielle, 2014). Nevertheless, those proposals fail to take into account the nonlinear dependence structure of the individual parameters.

A way to alleviate these problems is to use a proposal distribution derived from a discretised Langevin diffusion whose drift term is the gradient of the logarithm of the target density leading to the Metropolis Adjusted Langevin Algorithm (MALA). The MALA proposal is a multivariate Gaussian with the following mean $\mu_{i,\mathrm{MALA}}^{(k)}$ and covariance matrix Γ_{MALA}:

$$\mu_{i,\mathrm{MALA}}^{(k)} = \psi_i^{(k)} + \gamma \nabla_{\psi_i} \log \mathtt{p}_i(\psi_i^{(k)}|y_i) \quad \text{and} \quad \Gamma_{\mathrm{MALA}} = 2\gamma \mathsf{I}_p \tag{11}$$

where γ is a positive stepsize and I_p is the identity matrix in $\mathbb{R}^{p\times p}$. These methods appear to behave well for complex models but still do not take into consideration the multidimensional structure of the individual parameters. Recent works include efforts in that direction, such as the Anisotropic MALA for which the covariance matrix of the proposal depends on the gradient of the target measure (Allassonniere and Kuhn, 2013), the Tamed Unadjusted Langevin Algorithm (Brosse et al., 2017) based on the coordinate-wise taming of superlinear drift coefficients and a multidimensional extension of the Adaptative Metropolis algorithm (Haario et al., 2001) simultaneously estimating the covariance of the target measure and coercing the acceptance rate, see (Vihola, 2012).

The MALA algorithm is a special instance of the Hybrid Monte Carlo (HMC), introduced in (Neal et al., 2011); see (Brooks et al., 2011) and the references therein, and consists in augmenting the state space with an auxiliary variable p, known as the velocity in Hamiltonian dynamics. This algorithm belongs to the class of data augmentation methods. Indeed, the potential energy is augmented with a kinetic energy, function of an added auxiliary variable. The MCMC procedure then consists in sampling from this augmented posterior distribution. All those methods aim at finding the proposal q that accelerates the convergence of the chain. Unfortunately they are computationally involved (even in small and medium dimension settings, the computation of the gradient or the Hessian can be overwhelming) and can be difficult to implement (stepsizes and numerical derivatives need to be tuned and implemented).

We see in the next section how to define a multivariate Gaussian proposal for both continuous and noncontinuous data models, that is easy to implement and that takes into account the multidimensional structure of the individual parameters in order to accelerate the MCMC procedure.

4. The nlme-IMH and the f-SAEM

In this section, we assume that the individual parameters $(\psi_1, \ldots, \psi_N)$ are independent and normally distributed with mean $(m_1, \ldots, m_n)$ and covariance Ω. The MAP estimate, for individual i, is the value of ψ_i that maximizes the conditional distribution $\mathtt{p}_i(\psi_i|y_i, \theta)$:

$$\hat{\psi}_i = \arg\max_{\psi_i \in \mathbb{R}^p} \mathtt{p}_i(\psi_i|y_i) = \arg\max_{\psi_i \in \mathbb{R}^p} \mathtt{p}_i(y_i|\psi_i)\mathtt{p}_i(\psi_i) \; . \tag{12}$$

4.1. Proposal based on Laplace approximation

For both continuous and noncontinuous data models, the goal is to find a simple proposal, a multivariate Gaussian distribution in our case, that approximates the target distribution $\mathtt{p}_i(\psi_i|y_i)$. For general MCMC samplers, it is shown in (Roberts and Rosenthal, 2011) that the mixing rate in total variation depends on the expectation of the acceptance ratio under the proposal distribution which is also directly related to the ratio of the proposal to the target in the special case of independent samplers (see (Mengersen and Tweedie, 1996; Roberts and Rosenthal, 2011)). This observation naturally suggests to find a proposal which approximates the target. de Freitas et al. (2001) advocates the use a multivariate Gaussian distribution whose parameters are obtained by minimizing the Kullback-Leibler divergence between a multivariate Gaussian variational candidate distribution and the target distribution. In (Andrieu and Thoms, 2008) and the references therein, an adaptative Metropolis algorithm is studied and reconciled to a KL divergence minimisation problem where the resulting multivariate Gaussian distribution can be used as a proposal in a IMH algorithm. Authors note that although this proposal might be a sensible choice when it approximates well the target, it can fail when the parametric form of the proposal is not sufficiently rich. Thus, other parametric forms can be considered and it is suggested in (Andrieu et al., 2006) to consider mixtures, finite or infinite, of distributions belonging to the exponential family.

In general, this optimization step is difficult and computationally expensive since it requires to approximate (using Monte Carlo integration for instance) the integral of the log-likelihood with respect to the variational candidate distribution.

Independent proposal 1. *We suggest a Laplace approximation of this conditional distribution as described in (Rue et al., 2009) which is the multivariate Gaussian distribution with mean* $\hat{\psi}_i$ *and variance-covariance*

$$\Gamma_i = \left(-\mathrm{H}_{\hat{\psi}_i} + \Omega^{-1} \right)^{-1} \ , \tag{13}$$

where $\mathrm{H}_{\hat{\psi}_i} \in \mathbb{R}^{p \times p}$ *is the Hessian of* $\log\left(\boldsymbol{p}_i(y_i|\psi_i)\right)$ *evaluated at* $\hat{\psi}_i$*.*

Mathematical details for computing this proposal are postponed to Appendix B. We use this multivariate Gaussian distribution as a proposal in our IMH algorithm introduced in the next section, for both continuous and noncontinuous data models.

Remark 2. *Note that the resulting proposal distribution is based on the assumption that, in model (2), the random effects* η_i *are normally distributed. When this assumption does not hold, our method exploits the same Gaussian proposal, where the variance* Ω *in (13) is calculated explicitely. Consider the following example: the random effects* η_i *in (2) are no longer distributed according to a multivariate Gaussian distribution but a multivariate Student distribution with* d *degrees of freedom, zero mean and a prior shape matrix* ξ *such that* $\eta_i \sim t_d(0, \xi)$*. Then the vector of parameters of the model is* $\theta = (\psi_{\text{pop}}, \Omega, \sigma^2)$ *where* $\Omega = \frac{d}{d-2}\xi$ *is the prior covariance matrix. In that case, our method uses the Independent proposal 1 and computes the MH acceptance ratio (10) with the corresponding multivariate Student density* $\boldsymbol{p}_i(\psi_i)$*.*

We shall now see another method to derive a Gaussian proposal distribution in the specific case of continuous data models (see (3)).

4.2. Nonlinear continuous data models

When the model is described by (3), the approximation of the target distribution can be done twofold: either by using the Laplace approximation, as explained above, or by linearizing the structural model f_i for any individual i of the population. using (3) and (12), the MAP estimate can thus be derived as:

$$\hat{\psi}_i = \arg \min_{\psi_i \in \mathbb{R}^p} \left(\frac{1}{\sigma^2} \| y_i - f_i(\psi_i) \|^2 + (\psi_i - m_i)' \Omega^{-1} (\psi_i - m_i) \right) \ . \tag{14}$$

where $f_i(\psi_i)$ is defined by (4) and A' is the transpose of the matrix A.

We linearize the structural model f_i around the MAP estimate $\hat{\psi}_i$:

$$f_i(\psi_i) \approx f_i(\hat{\psi}_i) + \mathrm{J}_{f_i(\hat{\psi}_i)}(\psi_i - \hat{\psi}_i) \ , \tag{15}$$

where $\mathrm{J}_{f_i(\hat{\psi}_i)} \in \mathbb{R}^{n_i \times p}$ is the Jacobian of f_i evaluated at $\hat{\psi}_i$. Defining $z_i := y_i - f_i(\hat{\psi}_i) + \mathrm{J}_{f_i(\hat{\psi}_i)}\, \hat{\psi}_i$, this expansion yields the following linear model:

$$z_i = \mathrm{J}_{f_i(\hat{\psi}_i)}\, \psi_i + \varepsilon_i \ . \tag{16}$$

We can directly use the definition of the conditional distribution under a linear model (see (45) in Appendix C) to get an expression of the conditional covariance Γ_i of ψ_i given z_i under (16):

$$\Gamma_i = \left(\frac{\mathrm{J}'_{f_i(\hat{\psi}_i)}\, \mathrm{J}_{f_i(\hat{\psi}_i)}}{\sigma^2} + \Omega^{-1} \right)^{-1} \ . \tag{17}$$

Using (14) and the definition of the conditional distribution under a linear model we obtain that $\mu_i = \hat{\psi}_i$ (See Appendix D for details). We note that the mode of the conditional distribution of ψ_i in the nonlinear model (3) is also the mode and the mean of the conditional distribution of ψ_i in the linear model (16).

Independent proposal 2. *In the case of continuous data models, we propose to use the multivariate Gaussian distribution, with mean $\hat{\psi}_i$ and variance-covariance matrix Γ_i defined by (17) as a proposal for an independent MH algorithm avoiding the computation of an Hessian matrix.*

We can note that linearizing the structural model is equivalent to using the Laplace approximation with the expected information matrix. Indeed:

$$\mathbb{E}_{y_i|\hat{\psi}_i}\left(-\mathrm{H}_{l(\hat{\psi}_i)} \right) = \frac{\mathrm{J}'_{f_i(\hat{\psi}_i)}\, \mathrm{J}_{f_i(\hat{\psi}_i)}}{\sigma^2} \ . \tag{18}$$

Remark 3. *When the model is linear, the probability of accepting a candidate generated with this proposal is equal to 1.*

Remark 4. *If we consider a more general error model, $\varepsilon_i \sim \mathcal{N}(0, \Sigma(t_i, \psi_i))$ that may depend on the individual parameters ψ_i and the observation times t_i, then the conditional variance-covariance matrix reads:*

$$\Gamma_i = \left(\mathrm{J}'_{f_i(\hat{\psi}_i)}\, \Sigma(t_i, \hat{\psi}_i)^{-1}\, \mathrm{J}_{f_i(\hat{\psi}_i)} + \Omega^{-1} \right)^{-1} \ . \tag{19}$$

Remark 5. *In the model (33), the transformed variable $\phi_i = u(\psi_i)$ follows a normal distribution. Then a candidate ϕ_i^c is drawn from the multivariate Gaussian proposal with parameters:*

$$\mu_i = \hat{\phi}_i \ , \tag{20}$$

$$\Gamma_i = \left(\frac{\mathrm{J}'_{f_i(u^{-1}(\hat{\phi}_i))}\, \mathrm{J}_{f_i(u^{-1}(\hat{\phi}_i))}}{\sigma^2} + \Omega^{-1} \right)^{-1} \ , \tag{21}$$

where $\hat{\phi}_i = \arg\max_{\phi_i \in \mathbb{R}^p} \mathtt{p}_i(\phi_i|y_i)$ *and finally the candidate vector of individual parameters is set to* $\psi_i^c = u^{-1}(\phi_i^c)$

These approximations of the conditional distribution $\mathtt{p}_i(\psi_i|y_i)$ lead to our nlme-IMH algorithm, an Independent Metropolis-Hastings (IMH) algorithm for NLME models. For all individuals $i \in \{1, \ldots, N\}$, the algorithm is defined as:

Algorithm 2 The nlme-IMH algorithm

```
Initialization: Initialize the chain sampling $\psi_i^{(0)}$ from some initial distribution $\xi_i$ .
Iteration t: Given the current state of the chain $\psi_i^{(t-1)}$:
    1. Compute the MAP estimate:
           $\hat{\psi}_i^{(t)} = \arg\max_{\psi_i \in \mathbb{R}^p} \mathtt{p}_i(\psi_i|y_i)$ .   (22)
    2. Compute the covariance matrix $\Gamma_i^{(t)}$ using either (13) or (17).
    3. Sample a candidate $\psi_i^c$ from a the independent proposal $\mathcal{N}(\hat{\psi}_i^{(t)}, \Gamma_i^{(t)})$ denoted $q_i(\cdot|\hat{\psi}_i^{(t)})$.
    4. Compute the MH ratio:
           $\alpha(\psi_i^{(t-1)}, \psi_i^c) = \frac{\mathtt{p}_i(\psi_i^c|y_i)}{\mathtt{p}_i(\psi_i^{(t-1)}|y_i)} \frac{q_i(\hat{\psi}_i^{(t)}|\psi_i^c)}{q_i(\psi_i^c|\hat{\psi}_i^{(t)})}$ .   (23)
    5. Set $\psi_i^{(t)} = \psi_i^c$ with probability $\min(1, \alpha(\psi_i^c, \psi_i^{(t-1)}))$ (otherwise, keep $\psi_i^{(t)} = \psi_i^{(t-1)}$).
```

This method shares some similiarities with (Titsias and Papaspiliopoulos, 2018) that suggests to perform a Taylor expansion of $\mathtt{p}_i(y_i|\psi_i)$ around the current state of the chain, leaving $\mathtt{p}_i(\psi_i)$ unchanged.

Remark 6. *Although a multivariate Gaussian proposal is used in our presentation of the nlme-IMH, other type of distributions could be adopted. For instance, when the target distribution presents heavy tails, a Student distribution with a well-chosen degree of freedom could improve the performance of the independent sampler. In such case, the parameters of the Gaussian proposal are used to shift and scale the Student proposal distribution and the acceptance rate* (23) *needs to be modified accordingly. The numerical applications in Section 5 are performed using a Gaussian proposal but comparisons with a Student proposal distribution are given in Appendix* E.1*.*

4.3. Maximum Likelihood Estimation

The ML estimator defined by (9) is computed using the Stochastic Approximation of the EM algorithm (SAEM) (Delyon et al., 1999). The SAEM algorithm is described as follows:

Algorithm 3 The SAEM algorithm

Initialization: θ_0, an initial parameter estimate and M, the number of MCMC iterations.

Iteration k: given the current model parameter estimate θ_{k-1}:

1. **Simulation step:** for $i \in \{1, \dots, N\}$, draw vectors of individual parameters $(\psi_1^{(k)}, \dots, \psi_N^{(k)})$ after M transitions of Markov kernels $(\Pi_1^{(k)}(\psi_1^{(k-1)}, \cdot), \dots, (\Pi_N^{(k)}(\psi_N^{(k-1)}, \cdot))$ which admit as unique limiting distributions the conditional distributions $(\mathtt{p}_1(\psi_1|y_1; \theta_{k-1}), \dots, \mathtt{p}_N(\psi_N|y_N; \theta_{k-1}))$,
2. **Stochastic approximation step:** update the approximation of the conditional expectation $\mathbb{E}\left[\log \mathtt{p}(y, \psi; \theta)|y, \theta_{k-1}\right]$:

$$Q_k(\theta) = Q_{k-1}(\theta) + \gamma_k \left(\sum_{i=1}^{N} \log \mathtt{p}_i(y_i, \psi_i^{(k)}; \theta) - Q_{k-1}(\theta) \right) , \tag{24}$$

   where $\{\gamma_k\}_{k>0}$ is a sequence of decreasing stepsizes with $\gamma_1 = 1$.
3. **Maximisation step:** Update the model parameter estimate:

$$\theta_k = \arg\max_{\theta \in \mathbb{R}^d} Q_k(\theta) \ . \tag{25}$$

The SAEM algorithm is implemented in most sofware tools for NLME models and its convergence is studied in (Delyon et al., 1999; Kuhn and Lavielle, 2004; Allassonniere and Kuhn, 2013). The practical performances of SAEM are closely linked to the settings of SAEM. In particular, the choice of the transition kernel Π plays a key role. The transition kernel Π is directly defined by the proposal(s) used for the MH algorithm.

We propose a fast version of the SAEM algorithm using our resulting independent proposal distribution called the f-SAEM. The simulation step of the f-SAEM is achieved using the nlme-IMH algorithm (see algorithm 2) for all individuals $i \in \{1, \dots, N\}$ and the next steps remain unchanged. In practice, the number of transitions M is small since the convergence of the SAEM does not require the convergence of the MCMC at each iteration (Kuhn and Lavielle, 2004). In the sequel, we carry out numerous numerical experiments to compare our nlme-IMH algorithm to state-of-the-art samplers and assess its relevance in a MLE algorithm such as the SAEM.

5. Numerical Examples

5.1. A pharmacokinetic example

5.1.1. Data and model

32 healthy volunteers received a 1.5 mg/kg single oral dose of warfarin, an anticoagulant normally used in the prevention of thrombosis (O'Reilly and Aggeler, 1968). Figure 1 shows the warfarin plasmatic concentration measured at different times for these patients (the single dose was given at time 0 for all the patients).

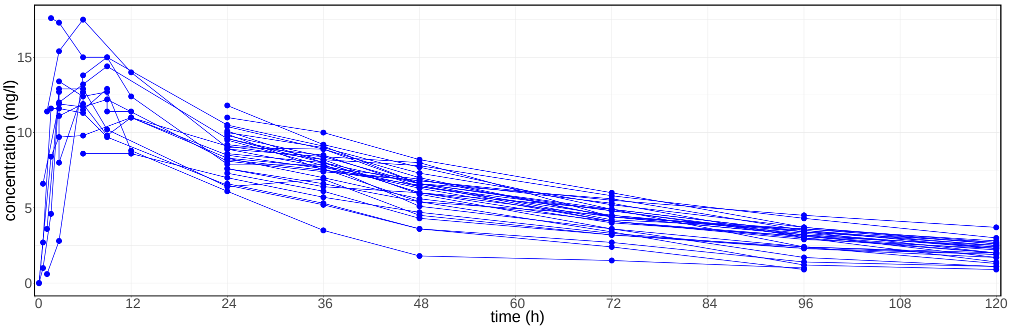

Figure 1: Warfarin concentration (mg/l) over time (h) for 32 subjects

We consider a one-compartment pharmacokinetics (PK) model for oral administration, assuming first-order absorption and linear elimination processes:

$$f(t, ka, V, k) = \frac{D\, ka}{V(ka-k)}(\mathrm{e}^{-ka\, t} - \mathrm{e}^{-k\, t}) \ , \tag{26}$$

where ka is the absorption rate constant, V the volume of distribution , k the elimination rate constant, and D the dose of drug administered. Here, ka, V and k are PK parameters that can change from one individual to another. Let $\psi_i = (ka_i, V_i, k_i)$ be the vector of individual PK parameters for individual i. The model for the j-th measured concentration, noted y_{ij}, for individual i writes:

$$y_{ij} = f(t_{ij}, \psi_i) + \varepsilon_{ij} \ . \tag{27}$$

We assume in this example that the residual errors are independent and normally distributed with mean 0 and variance σ^2. Lognormal distributions are used for the three PK parameters:

$$\log(ka_i) \sim \mathcal{N}(\log(ka_{\text{pop}}), \omega^2_{ka}) \ , \log(V_i) \sim \mathcal{N}(\log(V_{\text{pop}}), \omega^2_V) \ , \log(k_i) \sim \mathcal{N}(\log(k_{\text{pop}}), \omega^2_k) \ . \tag{28}$$

This is a specific instance of the nonlinear mixed effects model for continuous data described in Section 2.2. We thus use the multivariate Gaussian proposal whose mean and covariance are defined by (47) and (17). In such case the gradient can be explicitly computed. Nevertheless, for the method to be easily extended to any structural model, the gradient is calculated using Automatic Differentiation (Griewank and Walther, 2008) implemented in the R package "Madness" (Pav, 2016).

5.1.2. MCMC Convergence Diagnostic

We study in this section the behaviour of the MH algorithm used to sample individual parameters from the conditional distribution $\mathtt{p}_i(\psi_i|y_i;\theta)$. We consider only one of the 32 individuals for this study and fix θ close to the ML estimate obtained with the SAEM algorithm, implemented in the saemix R package (Comets et al., 2017): $ka_{\text{pop}} = 1$, $V_{\text{pop}} = 8$, $k_{\text{pop}} = 0.01$, $\omega_{ka} = 0.5$, $\omega_V = 0.2$, $\omega_k = 0.3$ and $\sigma^2 = 0.5$.

We run the classical version of MH implemented in the saemix package and for which different transition kernels are used successively at each iteration: independent proposals from the marginal distribution $\mathtt{p}_i(\psi_i)$, component-wise random walk and block-wise random walk. We compare it to our proposed algorithm 2.

We run 20 000 iterations of these two algorithms and evaluate their convergence by looking at the convergence of the median for the three components of ψ_i. We see Figure 2 that, for parameter k_i, the sequences of empirical median obtained with the two algorithms converge to the same value, which is supposed to be the theoretical median of the conditional distribution. It is interesting to note that the empirical median with the nlme-IMH converge very rapidly. This is interesting in the population approach framework because it is mainly the median values of each conditional distribution that are used to infer the population distribution. Autocorrelation plots, Figure 2, highlight slower mixing of the RWM whereas samples from the nlme-IMH can be considered independent few iterations after the chain has been initialized. Comparison for all three dimensions of the individual parameter ψ_i using a Student proposal distribution can be found in Appendix E.1.

The Mean Square Jump Distance (MSJD) as well as the Effective Sample Size (ESS) of the two methods are reported in Table 5. MSJD is a measure used to diagnose the mixing of the chain. It is calculated as the mean of the squared euclidean distance between every point and its previous point. Usually, this quantity indicates if the chain is moving enough or getting stuck at some region and the ESS is a quantity that estimates the number of independent samples obtained from a chain. Larger values of those two quantities for our method show greater performance of the sampler in comparison with the RWM.

Table 1: MSJD and ESS per dimension.

	MSJD			**ESS**		
	ka_i	V_i	k_i	ka_i	V_i	k_i
RWM	0.009	0.002	0.006	1728	3414	3784
nlme-IMH	0.061	0.004	0.018	13694	14907	19976

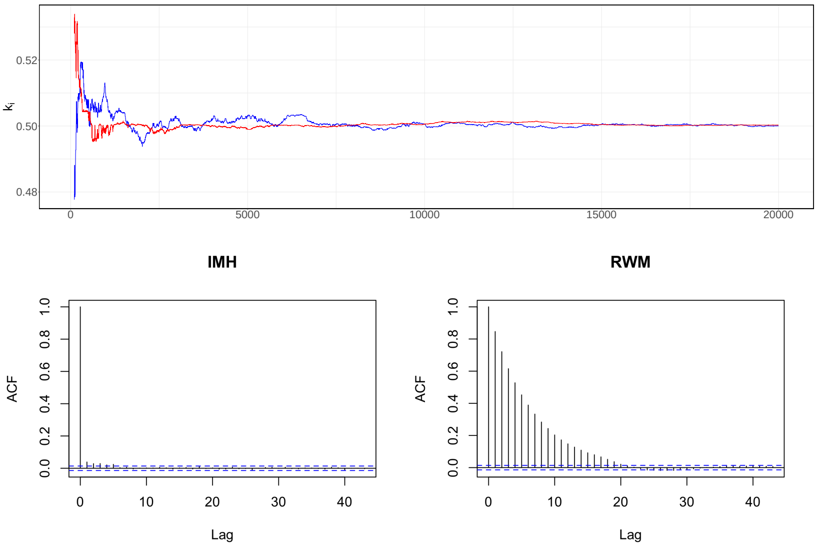

Figure 2: Modelling of the warfarin PK data. Top plot: convergence of the empirical medians of $\mathrm{p}_i(k_i|y_i;\theta)$ for a single individual. Comparison between the reference MH algorithm (blue) and the nlme-IMH (red). Bottom plot: Autocorrelation plots of the MCMC samplers for parameter k_i.

Comparison with state-of-the-art methods: We then compare our new approach to the three following samplers: an independent sampler that uses variational approximation as proposal distribution (de Freitas et al., 2001), the MALA (Roberts and Tweedie, 1996) and the No-U-Turn Sampler (Hoffman and Gelman, 2014).

The same design and settings (dataset, model parameter estimate, individual) as in section 5.1.2 are used throughout the following experiments.

5.1.2.a. Variational MCMC algorithm

The Variational MCMC algorithm (de Freitas et al., 2001) is a MCMC algorithm with independent proposal. The proposal distribution is a multivariate Gaussian distribution whose parameters are obtained by a variational approach that consists in minimising the Kullback Leibler divergence between a multivariate Gaussian distribution $q(\psi_i, \delta)$, and the target distribution for a given model parameter estimate θ noted $\mathrm{p}_i(\psi_i|y_i, \theta)$. This problem boils down to maximizing the so-called Evidence Lower Bound ELBO(θ) defined as:

$$\mathrm{ELBO}(\delta) \triangleq \int q(\psi_i, \delta) \left(\log \mathrm{p}_i(y_i, \psi_i, \theta) - \log q(\psi_i, \delta)\right) \mathrm{d}\psi_i \ . \tag{29}$$

We use the Automatic Differentiation Variational Inference (ADVI) (Kucukelbir et al., 2015) implemented in RStan (R Package (Stan Development Team, 2018)) to obtain the vector of parameters noted δ_{VI} defined as:

$$\delta_{VI} \triangleq \underset{\delta \in \mathbb{R}^p \times \mathbb{R}^{p \times p}}{\arg\max} \ \mathrm{ELBO}(\delta) \ .$$

The algorithm stops when the variation of the median of the objective function falls below the 1% threshold. The means and standard deviations of our nlme-IMH and the Variational MCMC proposals compare with the posterior mean (calculated using the NUTS (Hoffman and Gelman, 2014)) as follows:

Table 2: Means and standard deviations.

	Means			**Stds**		
	ka_i	V_i	k_i	ka_i	V_i	k_i
Variational proposal	0.90	7.93	0.48	0.14	0.03	0.07
Laplace proposal	0.88	7.93	0.52	0.18	0.04	0.09
NUTS (ground truth)	0.91	7.93	0.51	0.18	0.05	0.09

We observe that the mean of the variational approximation is slightly shifted from the estimated posterior mode (see table 2 for comparison) whereas a considerable difference lies in the numerical value of the covariance matrix obtained with ADVI. The empirical standard deviation of the Variational MCMC proposal is much smaller than our new proposal defined by (17) (see table 2), which slows down the MCMC convergence.

Figure 3 shows the proposals marginals and the marginal posterior distribution for the individual parameters k_i and V_i. Biplot of the samples drawn from the two multivariate Gaussian proposals (our independent proposal and the variational MCMC proposal) as well as samples drawn from the posterior distribution (using the NUTS) are also presented in this figure. We conclude that both marginal and bivariate posterior distributions are better approximated by our independent proposal than the one resulting from a KL divergence optimization.

Besides similar marginal variances, both our independent proposal and the true posterior share a strong anisotropic nature, confirmed by the similar correlation values of table 6 (see Appendix E.1). Same characteristics are observed for the other parameters. Those highlighted properties leads to a better performance of the nlme-IMH versus the ADVI sampler as reflected in Figure 4. Larger jumps of the chain and bigger ESS show how effective the nlme-IMH is compared to the ADVI (see Table 3).

5.1.2.b. Metropolis Adjusted Langevin Algorithm (MALA) and No-U-Turn Sampler (NUTS)

We now compare our method to the MALA, which proposal is defined by (11). The gradient of the log posterior distribution $\nabla_{\psi_i} \log \mathrm{p}_i(\psi_i^{(k)}|y_i)$ is also calculated by Automatic Differentiation. In this numerical example, the MALA has been initialized at the MAP and the stepsize ($\gamma = 10^{-2}$) is tuned such that the acceptance rate of 0.57 is reached (Roberts and Rosenthal, 1997).

We also compare the implementation of NUTS (Hoffman and Gelman, 2014; Carpenter et al., 2017) in the RStan package to our method in Figure 4. Figure 4 highlights good convergence of a well-tuned MALA and the NUTS. nlme-IMH and NUTS mixing

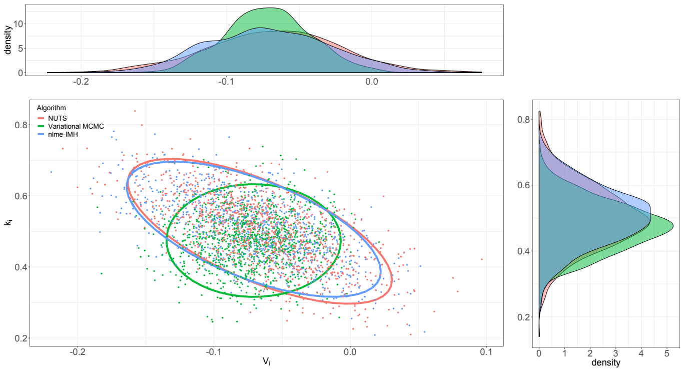

Figure 3: Modelling of the warfarin PK data: Comparison between the proposals of the nlme-IMH (blue), the Variational MCMC (green) and the empirical target distribution sampled using the NUTS (red). Marginals and biplots of the conditional distributions $k_i|y_i$ and $V_i|y_i$ for a single individual. Ellipses containing 90% of the data points are represented on the main plot.

properties, from autocorrelation plots in Figure 4 seem to be similar and much better than all of the other methods. Table 3 presents a benchmark of those methods regarding MSJD and ESS. Both nlme-IMH and NUTS have better performances here. For parameters ka and V, the ESS of the NUTS, presented as a gold standard sampler for this king of problem, are slightly higher than our proposed method.

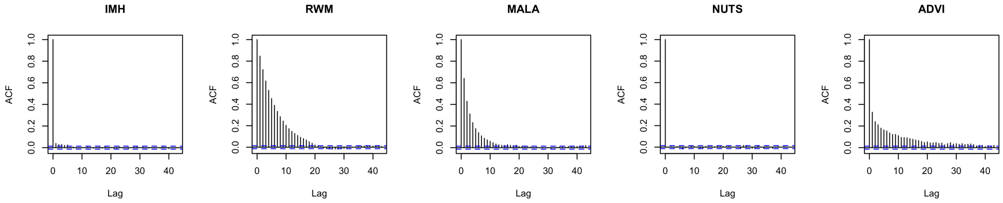

Figure 4: Modelling of the warfarin PK data: Autocorrelation plots of the MCMC samplers for parameter k_i.

Table 3: MSJD and ESS per dimension.

	MSJD			**ESS**		
	ka_i	V_i	k_i	ka_i	V_i	k_i
RWM	0.009	0.002	0.006	1728	3414	3784
nlme-IMH	0.061	0.004	0.018	13694	14907	19976
MALA	0.024	0.002	0.006	3458	3786	3688
NUTS	0.063	0.004	0.018	18684	19327	19083
ADVI	0.037	0.002	0.010	2499	1944	2649

In practice, those three methods imply tuning phases that are computationally involved , warming up the chain and a careful initialisation whereas our independent sampler is automatic and fast to implement. Investigating the asymptotic convergence behavior of those methods highlights the competitive properties of our IMH algorithm to sample from the target distribution.

Since our goal is to embed those samplers into a MLE algorithm such as the SAEM, we shall now study how they behave in the very first iterations of the MCMC procedure. Recall that the SAEM requires only few iterations of MCMC sampling under the current model parameter estimate. We present this non asymptotic study in the following section.

5.1.3. Comparison of the chains for the first 500 iterations

We produce 100 independent runs of the RWM, the nlme-IMH, the MALA and the NUTS for 500 iterations. The boxplots of the samples drawn at a given iteration threshold (three different thresholds are used) are presented Figure 5 against the ground truth for the parameter **ka**. The ground truth has been calculated by running the NUTS for 100 000 iterations.

For the three numbers of iteration (5,20,500) considered in Figure 5, the median of the nlme-IMH and NUTS samples are closer to the ground truth. Figure 5 also highlights that all those methods succeed in sampling from the whole distribution after 500 iterations.

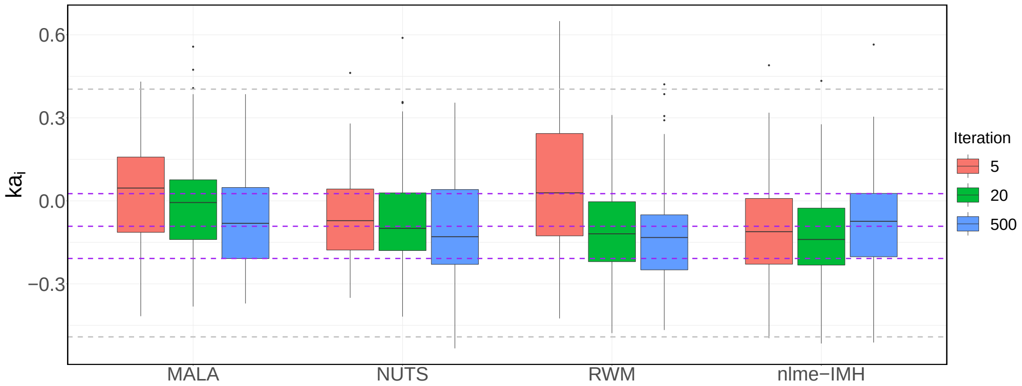

Figure 5: Modelling of the warfarin PK data: Boxplots for the RWM, the nlme-IMH, the MALA and the NUTS algorithm, averaged over 100 independent runs. The groundtruth median, 0.25 and 0.75 percentiles are plotted as a dashed purple line and its maximum and minimum as a dashed grey line.

We now use the RWM, the nlme-IMH and the MALA in the SAEM algorithm and observe the convergence of the resulting sequences of parameters.

5.1.4. Maximum likelihood estimation

We use the SAEM algorithm to estimate the population PK parameters ka_{pop}, V_{pop} and k_{pop}, the standard deviations of the random effects ω_{k_a}, ω_V and ω_k and the residual variance σ^2.

The stepsize γ_k is set to 1 during the first 100 iterations and then decreases as $1/k^a$ where $a = 0.7$ during the next 100 iterations.

Here we compare the standard SAEM algorithm, as implemented in the saemix R package, with the f-SAEM algorithm and the SAEM using the MALA sampler. In this example, the nlme-IMH and the MALA are only used during the first 20 iterations of the SAEM. The standard MH algorithm is then used.

Figure 6 shows the estimates of V_{pop} and ω_V computed at each iteration of these three variants of SAEM and starting from three different initial values. First of all, we notice that, whatever the initialisation and the sampling algorithm used, all the runs converge towards the maximum likelihood estimate. It is then very clear that the f-SAEM converges faster than the standard algorithm. The SAEM using the MALA algorithm for sampling from the individual conditional distribution presents a similar convergence behavior as the reference.

We can conclude, for this example, that sampling around the MAP of each individual conditional distribution is the key to a fast convergence of the SAEM during the first iterations.

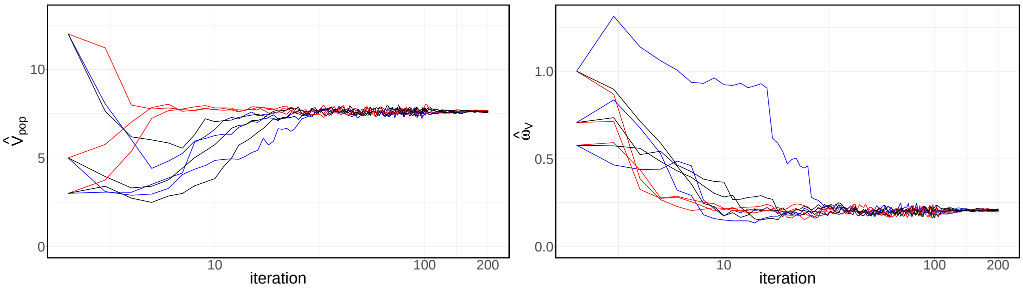

Figure 6: Estimation of the population PK parameters for the warfarin data: convergence of the sequences of estimates $(\hat{V}_{\text{pop},k}, 1 \leq k \leq 200)$ and $(\hat{\omega}_{V,k}, 1 \leq k \leq 200)$ obtained with SAEM and three different initial values using the reference MH algorithm (blue), the f-SAEM (red) and the SAEM using the MALA sampler (black).

5.1.5. Monte Carlo study

We conduct a Monte Carlo study to confirm the properties of the f-SAEM algorithm for computing the ML estimates.

$M = 50$ datasets have been simulated using the PK model previously used for fitting the warfarin PK data with the following parameter values: $ka_{\text{pop}} = 1$, $V_{\text{pop}} = 8$, $k_{\text{pop}} = 0.1$, $\omega_{ka} = 0.5$, $\omega_V = 0.2$, $\omega_k = 0.3$ and $\sigma^2 = 0.5$. The same original design with $N = 32$ patients and a total number of 251 PK measurements were used for all the simulated datasets. Since all the simulated data are different, the value of the ML estimator varies from one simulation to another. If we run K iterations of SAEM, the last element of the sequence $(\theta_k^{(m)}, 1 \leq k \leq K)$ is the estimate obtained from the m-th simulated dataset. To investigate how fast $(\theta_k^{(m)}, 1 \leq k \leq K)$ converges to $\theta_K^{(m)}$ we study how fast $(\theta_k^{(m)} - \theta_K^{(m)}, 1 \leq k \leq K)$ goes to 0. For a given sequence of estimates, we can then define, at each iteration k and for each component ℓ of the parameter, the mean square distance over the replicates

$$E_k(\ell) = \frac{1}{M} \sum_{m=1}^{M} \left(\theta_k^{(m)}(\ell) - \theta_K^{(m)}(\ell) \right)^2 . \tag{30}$$

Figure 7 shows using the new proposal leads to a much faster convergence towards the maximum likelihood estimate. Less than 10 iterations are required to converge with the f-SAEM on this example, instead of 50 with the original version. It should also be noted that the distance decreases monotonically. The sequence of estimates approaches the target at each iteration, compared to the standard algorithm which makes twists and turns before converging.

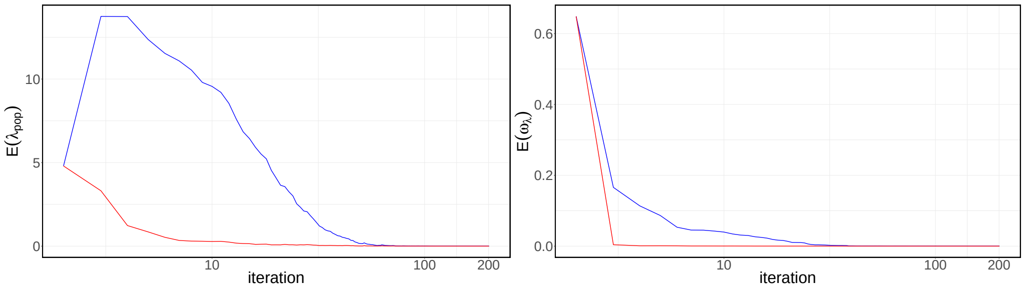

Figure 7: Convergence of the sequences of mean square distances $(E_k(V_{\text{pop}}), 1 \leq k \leq 200)$ and $(E_k(\omega_V), 1 \leq k \leq 200)$ for V_{pop} and ω_V obtained with SAEM on $M = 50$ synthetic datasets using the reference MH algorithm (blue) and the f-SAEM (red).

5.2. *Time-to-event Data Model*

5.2.1. *The model*

In this section, we consider a Weibull model for time-to-event data (Lavielle, 2014; Zhang, 2016). For individual i, the hazard function of this model is:

$$h(t, \psi_i) = \frac{\beta_i}{\lambda_i} \left(\frac{t}{\lambda_i} \right)^{\beta_i - 1} . \tag{31}$$

Here, the vector of individual parameters is $\psi_i = (\lambda_i, \beta_i)$. These two parameters are assumed to be independent and lognormally distributed:

$$\log(\lambda_i) \sim \mathcal{N}(\log(\lambda_{\text{pop}}), \omega_\lambda^2), \log(\beta_i) \sim \mathcal{N}(\log(\beta_{\text{pop}}), \omega_\beta^2) . \tag{32}$$

Then, the vector of population parameters is $\theta = (\lambda_{\text{pop}}, \beta_{\text{pop}}, \omega_\lambda, \omega_\beta)$.

Repeated events were generated, for $N = 100$ individuals, using the Weibull model (31) with $\lambda_{\text{pop}} = 10$, $\omega_\lambda = 0.3$, $\beta_{\text{pop}} = 3$ and $\omega_\beta = 0.3$ and assuming a right censoring time $\tau_c = 20$.

5.2.2. MCMC Convergence Diagnostic

Similarly to the previous section, we start by looking at the behaviour of the MCMC procedure used for sampling from the conditional distribution $\mathtt{p}_i(\psi_i|y_i;\theta)$ for a given individual i and assuming that θ is known. We use the generating model parameter in these experiments $(\theta = (\lambda_{\text{pop}} = 10, \beta_{\text{pop}} = 3, \omega_\lambda = 0.3, \omega_\beta = 0.3))$.

We run 12 000 iterations of the reference MH algorithm the nlme-IMH to estimate the median of the posterior distribution of λ_i. We see Figure 8 that the sequences of empirical medians obtained with the two procedures converge to the same value but the new algorithm converges faster than the standard MH algorithm. Autocorrelation plots, Figure 8, are also significantly showing the advantage of the new sampler as the chain obtained with the nlme-IMH is mixing almost ten times faster than the reference sampler.

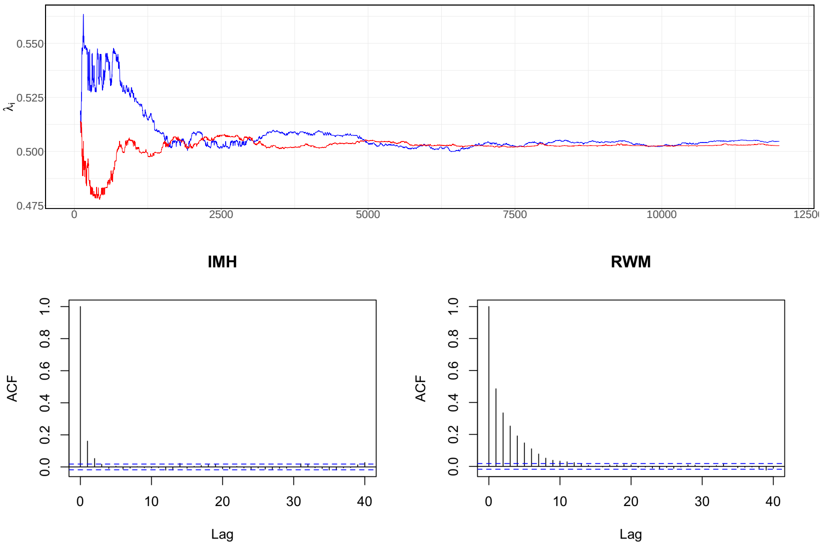

Figure 8: Time-to-event data modelling. Top plot: convergence of the empirical medians of $\mathtt{p}_i(\lambda_i|y_i;\theta)$ for a single individual. Comparison between the reference MH algorithm (blue) and the nlme-IMH (red). Bottom plot: Autocorrelation plots of the MCMC samplers for parameter λ_i.

Plots for the other parameter can be found in Appendix E.2. Comparisons with state-of-the-art methods were conducted as in the previous section. These comparisons led us to the same remarks as those made for the previous continuous data model both on the asymptotic and non asymptotic regimes.

Table 4: MSJD and ESS per dimension.

	MSJD		**ESS**	
	λ_i	β_i	λ_i	β_i
RWM	0.055	0.093	3061	1115
nlme-IMH	0.095	0.467	8759	8417

5.2.3. Maximum likelihood estimation of the population parameters

We run the standard SAEM algorithm implemented in the saemix package (extension of this package for noncontinuous data models is available on GitHub: https://github.com/belhal/saemix) and the f-SAEM on the generated dataset.

Figure 9 shows the estimates of λ_{pop} and ω_λ computed at each iteration of the two versions of the SAEM and starting from three different initial values. The same behaviour is observed as in the continuous case: regardless the initial values and the algorithm, all the runs converge to the same solution but convergence is much faster with the proposed method. The same comment applies for the two other parameters β_{pop} and ω_β.

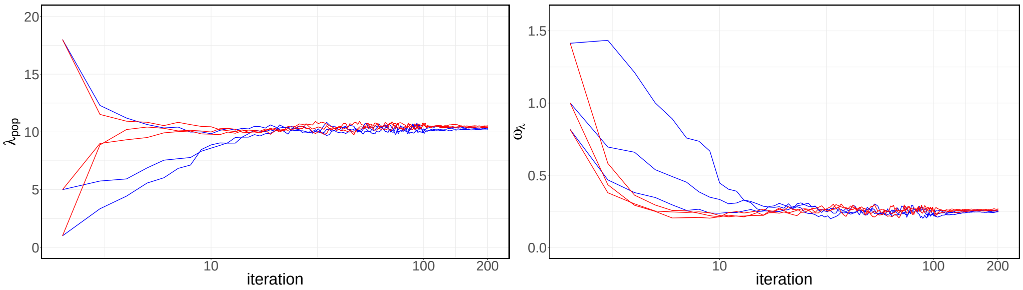

Figure 9: Population parameter estimation in time-to-event-data models: convergence of the sequences of estimates $(\hat{\lambda}_{\text{pop},k}, 1 \leq k \leq 200)$ and $(\hat{\omega}_{\lambda,k}, 1 \leq k \leq 200)$ obtained with SAEM and three different initial values using the reference MH algorithm (blue) and the f-SAEM (red).

5.2.4. Monte Carlo study

We now conduct a Monte Carlo study in order to confirm the good properties of the new version of the SAEM algorithm for estimating the population parameters of a time-to-event data model. $M = 50$ synthetic datasets are generated using the same design as above. Figure 10 shows the convergence of the mean square distances defined in (30) for λ_{pop} and ω_λ. All these distances converge monotonically to 0 which means that both algorithms properly converge to the maximum likelihood estimate, but very few iterations are required with the new version to converge while about thirty iterations are needed with the SAEM.

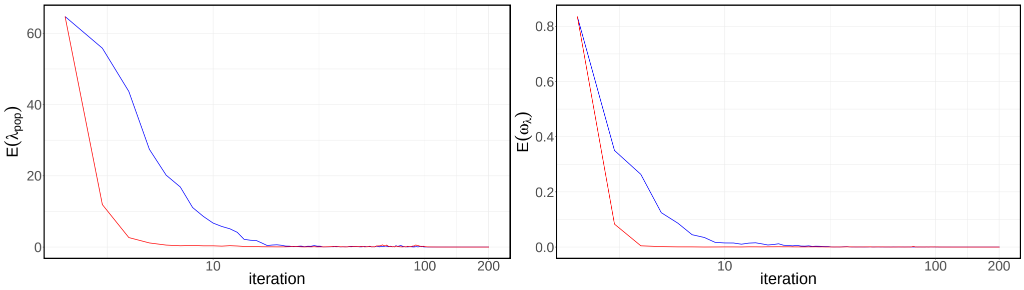

Figure 10: Convergence of the sequences of mean square distances $(E_k(\lambda_{\text{pop}}), 1 \leq k \leq 200)$ and $(E_k(\omega_\lambda), 1 \leq k \leq 200)$ for λ_{pop} and ω_λ obtained with SAEM from $M = 50$ synthetic datasets using the reference MH algorithm (blue) and the f-SAEM (red).

6. Conclusion and discussion

We present in this article an independent Metropolis-Hastings procedure for sampling random effects from their conditional distributions and a fast MLE algorithm, called the f-SAEM, in nonlinear mixed effects models.

The idea of the method is to approximate each individual conditional distribution by a multivariate normal distribution. A Laplace approximation makes it possible to consider any type of data, but we have shown that, in the case of continuous data, this approximation is equivalent to linearizing the structural model around the conditional mode of the random effects.

The numerical experiments demonstrate that the proposed nlme-IMH sampler converges faster to the target distribution than a standard random walk Metropolis. This practical behaviour is partly explained by the fact that the conditional mode of the random effects in the linearized model coincides with the conditional mode of the random effects in the original model. The proposal distribution is therefore a normal distribution centered around this MAP. On the other hand, the dependency structure in the conditional distribution of the random effects is well approximated by the covariance structure of the Gaussian proposal. So far, we have mainly applied our method to standard problems encountered in pharmacometrics and for which the number of random effects remains small. It can nevertheless be interesting to see how this method behaves in higher dimension and compare it with methods adapted to such situations such as MALA or HMC. Lastly, we have shown that this new IMH algorithm can easily be embedded in the SAEM algorithm for maximum likelihood estimation of the population parameters. Our numerical studies have shown empirically that the new transition kernel is effective in the very first iterations. It is of great interest to determine automatically and in an adaptive way an optimal scheme of kernel transitions combining this new proposal with the block-wise random walk Metropolis.

References

Agresti, A., 1990. Categorical data analysis. A Wiley-Interscience publication, Wiley, New York.

Allassonniere, S., Kuhn, E., 2013. Convergent Stochastic Expectation Maximization algorithm with efficient sampling in high dimension. Application to deformable template model estimation. arXiv preprint arXiv:1207.5938 .

Andersen, P.K., 2006. Survival Analysis. Wiley Reference Series in Biostatistics .

Andrieu, C., Doucet, A., Holenstein, R., 2010. Particle markov chain monte carlo methods. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 72, 269–342.

Andrieu, C., Moulines, É., et al., 2006. On the ergodicity properties of some adaptive mcmc algorithms. The Annals of Applied Probability 16, 1462–1505.

Andrieu, C., Roberts, G.O., et al., 2009. The pseudo-marginal approach for efficient monte carlo computations. The Annals of Statistics 37, 697–725.

Andrieu, C., Thoms, J., 2008. A tutorial on adaptive mcmc. Statistics and computing 18, 343–373.

Atchadé, Y.F., Rosenthal, J.S., 2005. On adaptive Markov chain Monte Carlo algorithms. Bernoulli 11, 815–828. doi:10.3150/bj/1130077595.

Beal, S., Sheiner, L., 1980. The NONMEM system. The American Statistician 34, 118–119.

Betancourt, M., 2017. A conceptual introduction to Hamiltonian Monte Carlo. arXiv preprint arXiv:1701.02434 .

Brooks, S., Gelman, A., Jones, G., Meng, X.L., 2011. Handbook of markov chain monte carlo. CRC press.

Brosse, N., Durmus, A., Moulines, É., Sabanis, S., 2017. The tamed unadjusted langevin algorithm. arXiv preprint arXiv:1710.05559 .

Carpenter, B., Gelman, A., Hoffman, M., Lee, D., Ben, G., Betancourt, M., Brubaker, M., Guo, J., Li, P., Riddell, A., 2017. Stan: A probabilistic programming language. Journal of Statistical Software 76.

Chan, P.L.S., Jacqmin, P., Lavielle, M., McFadyen, L., Weatherley, B., 2011. The use of the SAEM algorithm in MONOLIX software for estimation of population pharmacokinetic-pharmacodynamic-viral dynamics parameters of maraviroc in asymptomatic HIV subjects. Journal of Pharmacokinetics and Pharmacodynamics 38, 41–61.

Comets, E., Lavenu, A., Lavielle, M., 2017. Parameter estimation in nonlinear mixed effect models using saemix, an r implementation of the saem algorithm. Journal of Statistical Software 80, 1–42.

Delyon, B., Lavielle, M., Moulines, E., 1999. Convergence of a stochastic approximation version of the EM algorithm. Ann. Statist. 27, 94–128. doi:10.1214/aos/1018031103.

Donnet, S., Samson, A., 2013. Using pmcmc in em algorithm for stochastic mixed models: theoretical and practical issues. Journal de la Société Française de Statistique 155, 49–72.

Doucet, A., Godsill, S., Andrieu, C., 2000. On sequential monte carlo sampling methods for bayesian filtering. Statistics and computing 10, 197–208.

de Freitas, N., Højen-Sørensen, P., Jordan, M.I., Russell, S., 2001. Variational mcmc. Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence , 120–127.

Griewank, A., Walther, A., 2008. Evaluating derivatives: principles and techniques of algorithmic differentiation. volume 105. Siam.

Haario, H., Saksman, E., Tamminen, J., et al., 2001. An adaptive metropolis algorithm. Bernoulli 7, 223–242.

Hoffman, M.D., Gelman, A., 2014. The No-U-turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo. Journal of Machine Learning Research 15, 1593–1623.

Kucukelbir, A., Ranganath, R., Gelman, A., Blei, D., 2015. Automatic variational inference in stan, in: Cortes, C., Lawrence, N.D., Lee, D.D., Sugiyama, M., Garnett, R. (Eds.), Advances in Neural Information Processing Systems 28. Curran Associates, Inc., pp. 568–576.

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

Lavielle, M., 2014. Mixed effects models for the population approach: models, tasks, methods and tools. CRC press.

Lavielle, M., Ribba, B., 2016. Enhanced method for diagnosing pharmacometric models: random sampling from conditional distributions. Pharmaceutical research 33, 2979–2988.

Louis, T.A., 1982. Finding the observed information matrix when using the em algorithm. Journal of the Royal Statistical Society, Series B: Methodological 44, 226–233.

Mbogning, C., Bleakley, K., Lavielle, M., 2015. Joint modeling of longitudinal and repeated time-to-event data using nonlinear mixed-effects models and the SAEM algorithm. Journal of Statistical Computation and Simulation 85, 1512–1528. doi:10.1080/00949655.2013.878938.

McLachlan, G., Krishnan, T., 2007. The EM algorithm and extensions. volume 382. John Wiley & Sons.

Mengersen, K.L., Tweedie, R.L., 1996. Rates of convergence of the hastings and metropolis algorithms. Ann. Statist. 24, 101–121. doi:10.1214/aos/1033066201.

Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H., Teller, E., 1953. Equation of state calculations by fast computing machines. The Journal of Chemical Physics 21, 1087–1092. doi:10.1063/1.1699114.

Migon, H., Gamerman, D., Louzada, F., 2014. Statistical Inference: An Integrated Approach, Second Edition. Chapman & Hall/CRC Texts in Statistical Science, Taylor & Francis.

Neal, R.M., et al., 2011. Mcmc using hamiltonian dynamics. Handbook of Markov Chain Monte Carlo 2, 2.

O'Reilly, R.A., Aggeler, P.M., 1968. Studies on coumarin anticoagulant drugs initiation of warfarin therapy without a loading dose. Circulation 38, 169–177.

Pav, S.E., 2016. Madness: a package for multivariate automatic differentiation .

Robert, C.P., Casella, G., 2010. Metropolis–Hastings Algorithms. Springer New York, New York, NY. pp. 167–197. doi:10.1007/978-1-4419-1576-4_6.

Roberts, G.O., Gelman, A., Gilks, W.R., 1997. Weak convergence and optimal scaling of random walk metropolis algorithms. Ann. Appl. Probab. 7, 110–120. doi:10.1214/aoap/1034625254.

Roberts, G.O., Rosenthal, J.S., 1997. Optimal scaling of discrete approximations to langevin diffusions. J. R. Statist. Soc. B 60, 255–268.

Roberts, G.O., Rosenthal, J.S., 2011. Quantitative non-geometric convergence bounds for independence samplers. Methodology and Computing in Applied Probability 13, 391–403.

Roberts, G.O., Tweedie, R.L., 1996. Exponential convergence of langevin distributions and their discrete approximations. Bernoulli 2, 341–363.

Rue, H., Martino, S., Chopin, N., 2009. Approximate bayesian inference for latent gaussian models by using integrated nested laplace approximations. Journal of the royal statistical society: Series b (statistical methodology) 71, 319–392.

Savic, R.M., Mentré, F., Lavielle, M., 2011. Implementation and evaluation of the SAEM algorithm for longitudinal ordered categorical data with an illustration in pharmacokinetics-pharmacodynamics. The AAPS Journal 13, 44–53.

Stan Development Team, 2018. RStan: the R interface to Stan. URL: http://mc-stan.org/. r package version 2.17.3.

Stramer, O., Tweedie, R.L., 1999. Langevin-type models i: Diffusions with given stationary distributions and their discretizations*. Methodology And Computing In Applied Probability 1, 283–306. doi:10.1023/A:1010086427957.

Titsias, M.K., Papaspiliopoulos, O., 2018. Auxiliary gradient-based sampling algorithms. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 0. doi:10.1111/rssb.12269.

Verbeke, G., 1997. Linear mixed models for longitudinal data. Springer.

Vihola, M., 2012. Robust adaptive metropolis algorithm with coerced acceptance rate. Statistics and Computing 22, 997–1008.

Wang, Y., 2007. Derivation of various nonmem estimation methods. Journal of Pharmacokinetics and pharmacodynamics 34, 575–593.

Zhang, Z., 2016. Parametric regression model for survival data: Weibull regression model as an example. Ann Transl Med. 24.

Appendices

A. Extensions of model (2)

Several extensions of model (2) are also possible. We can assume for instance that the transformed individual parameters are normally distributed:

$$u(\psi_i) = u(\psi_{\text{pop}}) + \eta_i \ , \tag{33}$$

where u is a strictly monotonic transformation applied on the individual parameters ψ_i. Examples of such transformation are the logarithmic function (in which case the components of ψ_i are log-normally distributed), the logit and the probit transformations (Lavielle, 2014). In the following, we either use the original parameter ψ_i or the Gaussian transformed parameter $u(\psi_i)$.

Another extension of model (2) consists in introducing individual covariates in order to explain part of the inter-individual variability:

$$u(\psi_i) = u(\psi_{\text{pop}}) + C_i\beta + \eta_i \ , \tag{34}$$

where C_i is a matrix of individual covariates. Here, the fixed effects are the vector of coefficients β and the vector of typical parameters ψ_{pop}.

B. Calculus of the proposal in the noncontinuous case

Laplace approximation (see (Migon et al., 2014)) consists in approximating an integral of the form

$$I := \int \mathrm{e}^{v(x)} \, \mathrm{d}x \ , \tag{35}$$

where v is at least twice differentiable.

The following second order Taylor expansion of the function v around a point x_0

$$v(x) \approx v(x_0) + \nabla v(x_0)(x - x_0) + \frac{1}{2}(x - x0)\nabla^2 v(x_0)(x - x0) \ , \tag{36}$$

provides an approximation of the integral I (consider a multivariate Gaussian probability distribution function which integral sums to 1):

$$I \approx \mathrm{e}^{v(x_0)} \sqrt{\frac{(2\pi)^p}{|-\nabla^2 v(x_0)|}} \exp\left\{ -\frac{1}{2}\nabla v(x_0)'\nabla^2 v(x_0)^{-1}\nabla v(x_0) \right\} \ . \tag{37}$$

In our context, we can write the marginal pdf $\mathtt{p}_i(y_i)$ that we aim to approximate as

$$\mathtt{p}_i(y_i) = \int \mathtt{p}_i(y_i, \psi_i) \mathrm{d}\psi_i \tag{38}$$

$$= \int \mathrm{e}^{\log \mathtt{p}_i(y_i,\psi_i)} \, \mathrm{d}\psi_i \ . \tag{39}$$

Then, let

$$\begin{aligned} v(\psi_i) &:= \log \mathtt{p}_i(y_i, \psi_i) && (40) \\ &= l(\psi_i) + \log \mathtt{p}_i(\psi_i) \ , && (41) \end{aligned}$$

and compute its Taylor expansion around the MAP $\hat{\psi}_i$. We have by definition that

$$\nabla \log \mathtt{p}_i(y_i, \hat{\psi}_i) = 0 \ ,$$

which leads to the following Laplace approximation of $\log \mathtt{p}_i(y_i)$:

$$-2 \log \mathtt{p}_i(y_i) \approx -p \log 2\pi - 2 \log \mathtt{p}_i(y_i, \hat{\psi}_i) + \log \left(\left| -\nabla^2 \log \mathtt{p}_i(y_i, \hat{\psi}_i) \right| \right) \ .$$

We thus obtain the following approximation of the logarithm of the conditional pdf of ψ_i given y_i evaluated at $\hat{\psi}_i$:

$$\log \mathtt{p}_i(\hat{\psi}_i | y_i) \approx -\frac{p}{2} \log 2\pi - \frac{1}{2} \log \left(\left| -\nabla^2 \log \mathtt{p}_i(y_i, \hat{\psi}_i) \right| \right) \ ,$$

which is precisely the log-pdf of a multivariate Gaussian distribution with mean $\hat{\psi}_i$ and variance-covariance $-\nabla^2 \log \mathtt{p}_i(y_i, \hat{\psi}_i)^{-1}$ with:

$$\begin{aligned} \nabla^2 \log \mathtt{p}_i(y_i, \hat{\psi}_i) &= \nabla^2 \log \mathtt{p}_i(y_i | \hat{\psi}_i) + \nabla^2 \log \mathtt{p}_i(\hat{\psi}_i) && (42) \\ &= \nabla^2 l(\hat{\psi}_i) + \Omega^{-1} \ . && (43) \end{aligned}$$

C. Linear continuous data models

Let $y_i = (y_{i,1}, \ldots, y_{i,n_i})'$ and $\varepsilon_i = (\varepsilon_{i,1}, \ldots, \varepsilon_{i,n_i})'$. Assume a linear relationship between the observations y_i and the vector of individual parameters ψ_i:

$$y_i = A_i \psi_i + \varepsilon_i \ , \qquad (44)$$

where $A_i \in \mathbb{R}^{n_i \times p}$ is the design matrix for individual i, ψ_i is normally distributed with mean $m_i \in \mathbb{R}^p$ and covariance $\Omega \in \mathbb{R}^{p \times p}$. Then, the conditional distribution of ψ_i given y_i is a normal distribution with mean μ_i and variance-covariance matrix Γ_i defined as:

$$\mu_i = \Gamma_i \left(\frac{A_i' y_i}{\sigma^2} + \Omega^{-1} m_i \right) \quad \text{where} \quad \Gamma_i = \left(\frac{A_i' A_i}{\sigma^2} + \Omega^{-1} \right)^{-1} \qquad (45)$$

Here, μ_i is the mode of the conditional distribution of ψ_i, known as the Maximum A Posteriori (MAP) estimate, or the Empirical Bayes Estimate (EBE) of ψ_i.

D. Conditional mode under the linearised model

Using (14), $\hat{\psi}_i$ satisfies:

$$-\frac{\mathrm{J}'_{f_i(\hat{\psi}_i)}}{\sigma^2}\left(y_i - f_i(\hat{\psi}_i)\right) + \Omega^{-1}(\hat{\psi}_i - m_i) = 0 \ , \tag{46}$$

which leads to the definition of the conditional mean μ_i of ψ_i given z_i, under the linearized model, by:

$$\begin{aligned}
\mu_i &= \Gamma_i \frac{\mathrm{J}'_{f_i(\hat{\psi}_i)}}{\sigma^2}\left(y_i - f_i(\hat{\psi}_i) + \mathrm{J}_{f_i(\hat{\psi}_i)}\,\hat{\psi}_i + \Omega^{-1} m_i\right) && (47)\\
&= \Gamma_i \left(\Omega^{-1}(\hat{\psi}_i - m_i) + \frac{\mathrm{J}'_{f_i(\hat{\psi}_i)}\,\mathrm{J}_{f_i(\hat{\psi}_i)}}{\sigma^2}\hat{\psi}_i + \Omega^{-1} m_i\right) && (48)\\
&= \Gamma_i \Gamma_i^{-1} \hat{\psi}_i = \hat{\psi}_i \ . && (49)
\end{aligned}$$

E. Numerical applications

E.1. A pharmacokinetic example

Table 5: MSJD and ESS per dimension.

	MSJD			**ESS**		
	ka_i	V_i	k_i	ka_i	V_i	k_i
RWM	0.009	0.002	0.006	1728	3414	3784
nlme-IMH (Gaussian)	0.061	0.004	0.018	13694	14907	19976
nlme-IMH (Student)	0.063	0.004	0.018	14907	19946	19856

Figures 11 and 12 highlight the performances of the RWM, the nlme-IMH using a Gaussian proposal distribution and a Student proposal. At iteration (t) of the MH algorithm, samples from the Student proposal distribution are obtained using the same parameters obtained in Proposal 2 as follows:

- Student samples $S_i^{(t)}$ are drawn from a student distribution with degree of freedom $k = 3$: $S_i^{(t)} \sim t(k)$
- Individual parameters $\psi_i^{(t)}$ are obtained using the mean and the covariance defined in Proposal 2 to shift and scale the obtained samples: $\psi_i^{(t)} = \hat{\psi}_i^{(t)} + S_i^{(t)}.\Gamma_i^{(t)}$

E.2. Time-to-event Data Model

Median convergence and autocorrelation plots of the RWM and our nlme-IMH methods for parameter β_i are presented in Figure 13. Same observations as for parameter λ_i can be made.

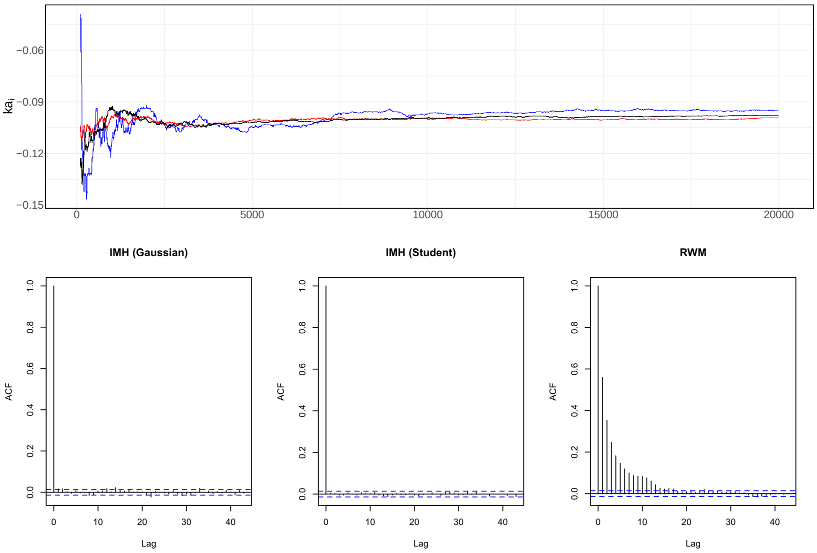

Figure 11: Modelling of the warfarin PK data. Top plot: convergence of the empirical medians of $\mathtt{p}_i(ka_i|y_i;\theta)$ for a single individual. Comparison between the reference MH algorithm (blue) and the nlme-IMH (red). Bottom plot: Autocorrelation plots of the MCMC samplers for parameter ka_i.

Table 6: Pairwise correlations of the proposals.

	ka_i, V_i	ka_i, k_i	V_i, k_i
Variational proposal	0.48	-0.28	-0.61
Laplace proposal	0.56	-0.39	-0.68
NUTS (ground truth)	0.55	-0.39	-0.68

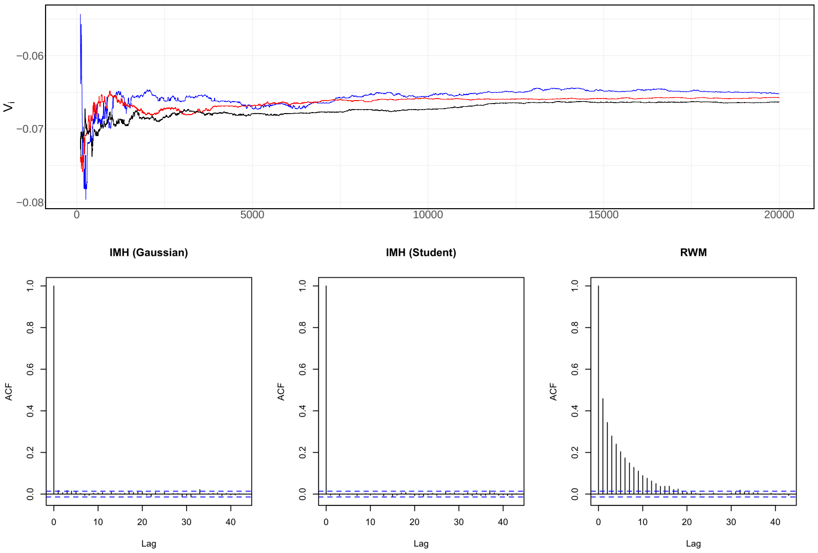

Figure 12: Modelling of the warfarin PK data. Top plot: convergence of the empirical medians of $\mathtt{p}_i(V_i|y_i;\theta)$ for a single individual. Comparison between the reference MH algorithm (blue) and the nlme-IMH (red). Bottom plot: Autocorrelation plots of the MCMC samplers for parameter V_i.

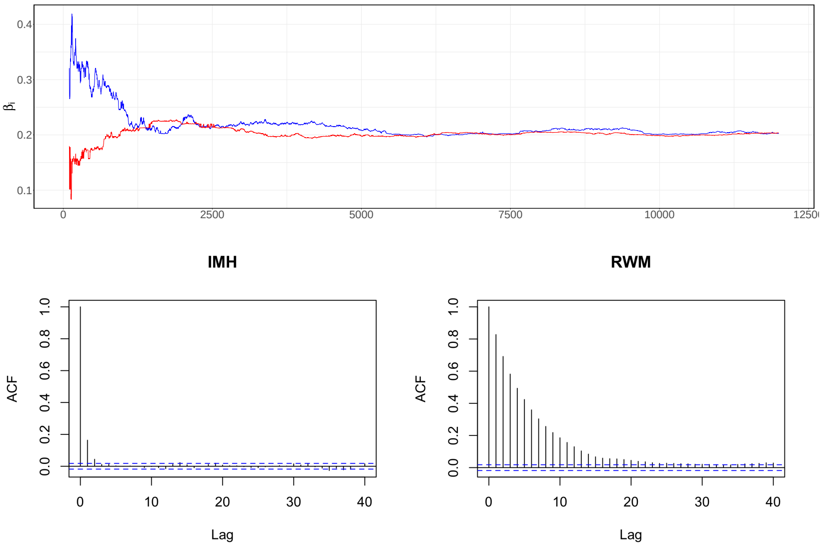

Figure 13: Time-to-event data modelling. Top plot: convergence of the empirical medians of $\mathtt{p}_i(\beta_i|y_i;\theta)$ for a single individual. Comparison between the reference MH algorithm (blue) and the nlme-IMH (red). Bottom plot: Autocorrelation plots of the MCMC samplers for parameter β_i.