# A Fast Markov chain Monte Carlo sampling for Maximum Likelihood Estimation

Belhal Karimi, Marc Lavielle, Eric Moulines
CMAP, Ecole Polytechnique, Universite Paris-Saclay, 91128 Palaiseau, France
Xpop Team, Inria, France
belhal.karimi@inria.fr

May 1, 2018

Population models are widely used in domains like pharmacometrics, economy or sociology where we need to model phenomena observed in each set of individuals. The population approach can be formulated in statistical terms using mixed effect models. In this context, sampling from the conditional distribution of the latent individual parameters given the observed data is crucial to perform many tasks. A direct draw is most of the time impossible. Sampling methods are thus used to obtain the desired samples. We consider a Markov Chain Monte Carlo procedure for sampling the random effects and/or estimating their conditional distribution. The choice of the proposal distribution is critical mainly for multidimensional space. New techniques such as SDE-based or Hamiltonian dynamics may be efficient but are difficult to tune and are costly. We propose the use of a multidimensional Gaussian proposal that takes into account the covariance structure of the random effects we want to infer and does not require any tuning. Numerical experiments based on simulated and real data highlight the very good performances of the proposed methods.

## 1 Introduction

We consider a complete data model $\mathtt{p}(y, \psi)$ where $y$ is the observed data and $\psi$ the latent data. We also consider parametric models $\mathtt{p}(y, \psi; \theta)$.

Many problems in computational statistics require sampling from the conditional distribution $\mathtt{p}(\psi|y; \theta)$. For instance, maximum likelihood (ML) estimation of $\theta$ consists in finding the model estimate $\hat{\theta}_{ML}$ defined as follows:

$$\hat{\theta}_{ML} = \arg\max_{\theta} \mathtt{p}(y; \theta) \tag{1}$$

where

$$\mathtt{p}(y : \theta) = \int \mathtt{p}(y, \psi; \theta) \mathtt{d}\psi \tag{2}$$

The Monte Carlo EM (MCEM) algorithm, first introduced in [39], approximates this quantity by a Monte Carlo integration. A Robbins Monroe type approximation can be used to evaluate that latter quantity after the simulation step, that is the SAEM algorithm described in [17]. When the posterior distribution of the individual parameters given the observed data is not tractable, sampling from this latter is impossible. The SAEM algorithm is thus coupled with an MCMC procedure to sample latent data from the posterior distribution. Convergence of such an algorithm has been proved in [11].

With the Bayesian modelling approach, another level of randomness is added to the model parameter and the general purpose of this approach is to find the parameter that is likely to have generated the observed data, i.e. estimating $\mathtt{p}(\theta|y)$ which boils down to sampling from $\mathtt{p}(\psi|y,\theta)$.

Obtaining realizations of the conditional distribution $\mathtt{p}(\psi|y;\theta)$ can turn out to be useful for several statistical tests and diagnostic plots used for models assessment. Furthermore, a classical way to compare models with each other consists in computing the observed likelihood by importance sampling which thus requires to sample from this intractable posterior distribution.

Markov Chain Monte Carlo (MCMC) is a general method for the simulation of distributions known up to a constant of proportionality [30, 31]. They consist in constructing a Markov chain that has the desired distribution as its stationary distribution. For instance, Random Walk Metropolis algorithms [9] that builds a chain, drawing samples from a Gaussian proposal distribution conditionally on the current state of the chain, are popular to perform this task. Many improvements have been made on this algorithm such as using an adaptive covariance proposal at each iteration, see [42].

More recently, Stochastic Gradient MCMC methods [43] leverage continuous dynamics to define a transition kernel that efficiently explores a target distribution. For instance, Langevin dynamics yields to the Metropolis Adjusted Langevin Algorithm [14] and Hamiltonian dynamics yields to the Hamiltonian Monte Carlo Algorithm [20]. Those methods scale well in high dimension but can be hard to tune and implement.

Alternatively, variational methods arose, as approximate inference procedures for high-dimensional distributions. For instance, variational bayes [38, 33] extended the EM algorithm when the expectation computed at the E-step was intractable and the dimension of the parameter space was too big. Variational algorithms can also be used to compute a better proposal for an MCMC algorithm [36, 24], in particular, it has been shown to speed up HMC algorithm [7].

In this article we present a faster method to sample from this intractable conditional distribution and use this faster procedure during the simulation step of the SAEM algorithm to perform faster ML estimation. We propose a new Gaussian proposal for a Metropolis-Hastings sampler [30]. In the context of Non Linear Mixed Effects (NLME) Models, when the observed outcomes are continuous, this approximation amounts to linearising the non linear structural model around the Maximum a Posteriori. When the outcomes are non continuous, we

use the Laplace approximation of the incomplete log-likelihood $\mathtt{p}(y,\theta)$ to obtain a Gaussian approximation of the conditional distribution $\mathtt{p}(\psi|y,\theta)$.

We present numerical examples of this new method used to perform ML estimation using the SAEM algorithm. We illustrate these performances on several models such as continuous pharmacokinetics (PK) models [37] or repeated time-to-event data [6]. A Monte Carlo study is presented to justify the effectiveness of our technique.

# 2 Mixed Effect Models

## 2.1 Population approach and hierarchical models

Here, we adopt a population approach where we consider several individuals and several observations per individual. We denote by $N$ the number of individuals in the population and $n_i$ the number of observations for 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 will assume 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 $\psi_i$ that takes its values in a subset of $\mathbb{R}^p$.

We naturally adopt a probabilistic framework by treating the $y_i$'s and the $\psi_i$'s as random variables. More precisely, we assume that the pairs $(y_i, \psi_i)$ are mutually independent and consider a parametric framework where the joint distribution of $(y_i, \psi_i)$ is denoted by $\mathtt{p}(y_i, \psi_i; \theta)$, where $\theta$ is the vector of parameters of the model. A natural decomposition of this joint distribution writes

$$\mathtt{p}(y_i, \psi_i; \theta) = \mathtt{p}(y_i|\psi_i; \theta)\mathtt{p}(\psi_i; \theta) \tag{3}$$

where $\mathtt{p}(y_i|\psi_i; \theta)$ is the conditional distribution of the observations, given the individual parameters, and where $\mathtt{p}(\psi_i; \theta)$ 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 parameters $\psi_i$ as a typical value $\psi_{\text{pop}}$, common to the whole population, and a vector individual random effects $\eta_i$:

$$\psi_i = \psi_{\text{pop}} + \eta_i \tag{4}$$

In the sequel, we will assume a multivariate Gaussian distribution for the random effects: $\eta_i \sim_{\text{i.i.d.}} \mathcal{N}(0, \Omega)$.

Several extension of model (4) are also possible. We can assume for instance that transformed individual parameters are normally distributed:

$$u(\psi_i) = u(\psi_{\text{pop}}) + \eta_i \tag{5}$$

where $u$ is a strictly monotonic transformation applied on the individual parameters $\psi_i$. Examples of such transformation are the logarithmic function

(assuming that $\psi_i$ is log-normally distributed), the logit and the probit transformations [18]. In the following, we will use either the original parameter $\psi_i$ or the Gaussian transformed parameter $u(\psi_i)$.

Another extension of model (4) consists in introducing individual covariates in order to explain part of the inter-individual variability:

$$u(\psi_i) = u(\psi_{\text{pop}}) + C_i\beta + \eta_i \tag{6}$$

where $C_i$ is a matrix of individual covariates. Here, the fixed effects are the vector of coefficients $\beta$ and the vector of typical parameters $\psi_{\text{pop}}$.

## 2.2 Continuous data models

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

$$y_{ij} = f(t_{ij}, \psi_i) + \varepsilon_{ij} \tag{7}$$

Where $y_{ij}$ is the j-th observation for individual $i$ measured at time $t_{ij}$, $\varepsilon_{ij}$ is the residual error, $f$ is the structural model and is a continuous and twice differentiable function of $\psi_i$.

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

$$y_i|\psi_i \sim \mathcal{N}(f_i(\psi_i), \sigma^2 \mathtt{Id}_{n_i \times n_i})$$

where $f_i(\psi_i) = (f(t_{i,1}, \psi_i), \ldots, f(t_{i,n_i}, \psi_i))$, or equivalently,

$$\mathtt{p}(y_i|\psi_i; \sigma^2) = (2\pi\sigma^2)^{\frac{n_i}{2}} e^{-\frac{1}{2\sigma^2}\|y_i - f_i(\psi_i)\|}$$

If we assume that $\psi_i \sim_{\text{i.i.d.}} \mathcal{N}(\psi_{\text{pop}}, \Omega)$, then the parameters of the model are $\theta = (\psi_{\text{pop}}, \Omega, \sigma^2)$. Furthermore, if the structural model $f$ is linear with respect to $\psi_i$, then the model is a so-called *linear mixed effects model*.

An extension of this model consists in assuming that the variance of the residual errors is not constant over time:

$$\varepsilon_{ij} \sim \mathcal{N}(0, g(t_{ij}, \psi_i)^2) \tag{8}$$

Such extension includes proportional error models ($g = bf$) and combined error models ($g = a + bf$) [18] but the proposed method remains the same whatever the residual error model.

## 2.3 Non continuous data models

Non continuous data models include categorical data models [28, 3], time to event data model [6, 4], or count data models [28] models.

A categorical outcome $y_{ij}$ takes its value in a set $\{1, \ldots, L\}$ 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 depends on the vector of individual parameters $\psi_i$ and may be function of the time $t_{ij}$.

In 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, strikes). To begin with, we will consider a model for a single event. The survival function $S(t)$ gives the probability that the event happens after time $t$:

$$\begin{aligned} S(t) &\triangleq \mathbb{P}(T > t) \\ &= e^{-\int_0^t h(u)du} \end{aligned} \tag{9}$$

where $h$ is called the hazard function.

In a population approach, we will consider a parametric and individual hazard function $h(., \psi_i)$.

The random variable representing the time-to-event for individual $i$ is typically written $T_i$.
A particular case of this model is to consider the time-to-event $T_i$ of a single event which is right censored:

$$y_i = \begin{cases} T_i & \text{if } T_i < \tau_c \\ \text{"}T_i > \tau_c\text{"} & \text{otherwise} \end{cases} \tag{10}$$

where $\tau_c$ 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:

$$\mathbb{P}(T_{ij} > t|T_{i,j-1} = t_{i,j-1}) = e^{-\int_{t_{i,j-1}}^t h(u,\psi_i)du} \tag{11}$$

Here, $t_{i,j}$ is the observed value of the random time $T_{i,j}$. 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 "$T_{i,n_i} > \tau_c$".

Then, we can show (see [18] for more details) that the conditional pdf of $y_i = (y_{ij}, 1 \leq n_i)$ writes

$$\mathrm{p}(y_i|\psi_i) = \exp\left\{-\int_0^{\tau_c} h(u, \psi_i)du\right\} \prod_{j=1}^{n_i-1} h(t_{ij}, \psi_i) \tag{12}$$

# 3 Sampling from conditional distributions

## 3.1 The conditional distribution of the individual parameters

Once the conditional distribution of the observations $\mathtt{p}(y_i|\psi_i;\theta)$ and the marginal distribution of the individual parameters $\psi_i$ are defined, the joint distribution $\mathtt{p}(y_i,\psi_i;\theta)$, but above all the conditional distribution $\mathtt{p}(\psi_i|y_i;\theta)$ are implicitly defined.

Indeed, this conditional distribution $\mathtt{p}(\psi_i|y_i;\theta)$ plays a crucial role in most methods used for inference in nonlinear mixed effects models.

One of the main task to perform is undoubtedly the estimation of the parameters of the model $\theta$, i.e., for instance, the calculation of the maximum likelihood (ML) estimate of $\theta$

$$\hat{\theta}_{\mathrm{ML}} = \arg\max_{\theta\in\Theta} \mathcal{L}(\theta, y). \tag{13}$$

where $\mathcal{L}(\theta, y) = \mathtt{p}(y;\theta)$. The stochastic approximation version of EM (SAEM) is an iterative procedure for ML estimation that requires to generate one or several realizations of this conditional distribution at each iteration of the algorithm.

Once the ML estimate $\hat{\theta}_{\mathrm{ML}}$ has been computed, the observed Fisher information matrix

$$I(\hat{\theta}_{\mathrm{ML}}, y) = -\nabla^2_\theta \mathcal{L}(\hat{\theta}_{\mathrm{ML}}, y) \tag{14}$$

can be derived thanks to the Louis formula [19]. This method is based on conditional expectations that cannot be explicitly calculated, but can be approximated by Monte-Carlo simulation. Such procedure requires to sample extensively from the conditional distribution $\mathtt{p}(\psi_i|y_i;\hat{\theta}_{\mathrm{ML}})$.

Then, several statistical tests and diagnostic plots used for models assessment are based on realizations of the conditional distribution $\mathtt{p}(\psi_i|y_i;\hat{\theta}_{\mathrm{ML}})$, rather than the mode of this distribution, in order to provide unbiased tests and plots.

Finally, the observed likelihood $\mathcal{L}(\hat{\theta}_{\mathrm{ML}}, y)$, used to compare models with each other, can be estimated by an importance sampling method. An estimate of the conditional distribution makes it possible to build a good proposal for this Monte Carlo method.

In short, being able to sample individual parameters from their conditional distribution is essential in nonlinear mixed models. It is therefore imperative to have at our disposal a good simulation tool.

## 3.2 The Metropolis Hastings Algorithm

Metropolis-Hasting (MH) algorithm is a powerful MCMC procedure widely used for sampling from a posterior distribution for which direct sampling is difficult. In our case, iteration $k$ of the MH algorithm consists in

- Drawing a candidate $\psi_i^c$ from a proposal distribution $q(\cdot|\psi_i^{(k-1)})$ where $\psi_i^{(k-1)}$ is the current state of the chain.
- Computing the MH ratio:

$$\alpha(\psi_i^c, \psi_i^{(k-1)}) = \frac{\pi(\psi_i^c|y_i)}{\pi(\psi_i^{(k-1)}|y_i)} \frac{q(\psi_i^{(k-1)}|\psi_i^c)}{q(\psi_i^c|\psi_i^{(k-1)})} \tag{15}$$

- Setting $\psi_i^{(k)} = \psi_i^c$ with probability $\min(1, \alpha(\psi_i^c, \psi_i^{(k-1)})$ (otherwise, keep $\psi_i^{(k)} = \psi_i^{(k-1)}$).

Under some general conditions, $(\psi_i^{(k)}, k \geq 0)$ is an ergodic Markov chain which distribution converges to the target distribution $\mathtt{p}(\psi_i|y_i)$ [23, 13, 32].

The choice of the proposal distribution $q$ plays a crucial role in the convergence behavior of the chain and one can heuristically acknowledge that the closer the proposal is to the target distribution, the faster the chain will converge. Ineed, this heuristic is mainly shared by variational inference practicioners that compute a Gaussian proposal that is optimally close, in the sense of Kullback-Leibler divergence, to the target distribution [36, 24, 7].

Current implementations of the SAEM algorithm in Monolix ([27]), saemix (R package) ([10]), nlmefitsa (Matlab) and NONMEM ([33]) 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 marginal distribution of the individual parameter $\psi_i$. The MH ratio then reduces to $\mathtt{p}(y_i|\psi_i^c)/\mathtt{p}(y_i|\psi_i^{(k)})$ for this proposal. The other proposals are component-wise and block-wise random walk procedures ([25]) that updates different components of $\psi_i$ using univariate and multivariate Gaussian proposal distributions.

These proposals are centered in the current state with a diagonal variance-covariance matrix, with variance terms which are adaptively adjusted at each iteration in order to reach some optimal acceptance rate ([42, 18]).

Nevertheless, the independent structure of those proposals has several drawbacks:

- such procedure is not suitable for sampling distributions in high dimension
- it does not take into account the covariance structure of the individual parameters

Major steps forward in this regard were made when a proposal process derived from a discretised Langevin diffusion with a drift term based on the gradient information of the target density was suggested in the Metropolis Adjusted Langevin Algorithm (MALA) ([15, 26]) and the Hamiltonian Monte Carlo (HMC) which implementation can be found for instance as the "No U-Turns Sampler" in STAN [16, 2].

The MALA consists in proposing a new state $\psi_i^c$ using the gradient of the target measure at the current state $\psi_i^{(k)}$:

$$\psi_i^c \sim \mathcal{N}(\psi_i^{(k)} - \gamma_k \nabla \log \pi(\psi_i^{(k)}), 2\gamma_k) \tag{16}$$

where $(\gamma_k)_{k>0}$ is a sequence of positive integers. It is a particular case of the RWM with a drift term [43] and a covariance matrix that is diagonal and isotropic (uniform in all directions). Likewise, the candidate state is accepted after computing the MH acceptance ratio. Several variants of this method have been developed in particular to optimize the covariance matrix of the proposal [35, 42].

These methods appear to scale well in high dimension but still does not take into consideration the multidimensional structure of the individual parameters. Recently a version where the covariance matrix of the proposal depends as well on the direction of the gradient of the target measure was derived in [12] and is called the Anisotropic MALA. Moreover, when the target measure is non-smooth and log-concave, algorithms using a Moreau-Yosida envelope of the non-smooth part of the target measure were developed in [1].

On the other hand, the HMC algorithm, introduced in [34], consists in augmenting the state space with an auxiliary variable $p$, known as the velocity in Hamiltonian dynamics. In Bayesian statistics, the goal being to sample from the conditional distribution of the individual parameters, the potential energy function is defined as the negated logarithm of this posterior distribution. Using HMC, we add to this potential energy a kinetic energy $V(p_i) = \frac{1}{2} p_i^T M^{-1} p_i$ function of the new auxiliary variable $p$ and $M$ called the mass matrix. This MCMC procedure will thus sample from this augmented posterior distribution calculated as the exponential of the sum of those two energy terms. We can choose the distribution of the auxiliary variable which is independent of the individual parameters, as we wish, specifying the distribution via the kinetic energy function $V(p_i)$. Current practice with HMC consists in using a quadratic kinetic energy, which leads the auxiliary variables to have a zero-mean multivariate Gaussian distribution with covariance $M$. Then the individual parameters of interest are obtained solving Hamiltonian dynamics and a MH ratio is calculated to accept or reject the augmented candidate state. Because of the independence of those two variables, the resulting individual parameters of the HMC algorithm are samples drawn from the desired posterior distribution.

The Riemann Manifold Hamiltonian Monte Carlo [20] suggests to take into consideration the curvature of the target distribution by assigning the covariance of the proposal distribution for the variable $p$ to be the Hessian of the target measure ($M_i(\psi_i) = \nabla^2 \pi(\psi_i | y_i)$).

All those methods aim at finding the proposal $q$ that will accelerate the convergence of the chain. Unfortunately they require a lot of computational resources (the calculus of the gradient or the Hessian can slow the algorithm) and can be difficult to implement (stepsizes and numerical derivatives need to be tuned and implemented).

We will see in the next section how to construct a proposal for both continuous and non continuous 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 A multivariate Gaussian proposal

## 4.1 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 first a linear relationship between the observations $y_i$ and the vector of individual parameters $\psi_i$:

$$y_i = A_i \psi_i + \varepsilon_i \tag{17}$$

where $A_i$ is the design matrix for individual $i$ and where $\psi_i$ is normally distributed around some value $m_i$

$$\psi_i \sim \mathcal{N}(m_i\,,\,\Omega)$$

.

Then, the conditional distribution of $\psi_i$ is a normal distribution:

$$\psi_i | y_i \sim \mathcal{N}(\mu_i\,,\,\Gamma_i)$$

where

$$\begin{aligned} \Gamma_i &= \left( \frac{A_i' A_i}{\sigma^2} + \Omega^{-1} \right)^{-1} \\ \mu_i &= \Gamma_i \left( \frac{A_i' y_i}{\sigma^2} + \Omega^{-1} m_i \right) \end{aligned} \tag{18}$$

Here, $\mu_i$ is the mode of the conditional distribution of $\psi_i$, known as the Maximum A Posteriori (MAP) estimate, or the Empirical Bayes Estimate (EBE) of $\psi_i$.

In the linear case, sampling $\psi_i$ from its conditional distribution, i.e. around the MAP estimate with an appropriate variance-covariance matrix, can be seen as a MH algorithm where a candidate is always accepted with probability 1. In other words, this conditional distribution is optimal as a proposal since only one iteration is required to reach the target distribution.

Let us see how to take advantage of this property in the non-linear case.

## 4.2 Nonlinear continuous data models

We will consider the model (7) for continuous data and assume that the $\psi_i$'s are normally distributed with mean $m_i$ and variance-covariance $\Omega$. For a given parameter value $\theta$, the MAP estimate, for individual $i$, is the value of $\psi_i$ that

maximizes the conditional distribution $\mathtt{p}(\psi_i|y_i,\theta)$:

$$\begin{aligned}\hat{\psi}_i &= \arg\max_{\psi_i} \mathtt{p}(\psi_i|y_i,\theta)\\ &= \arg\max_{\psi_i} \mathtt{p}(y_i|\psi_i,\theta)\mathtt{p}(\psi_i,\theta)\\ &= \arg\min_{\psi_i} \left(\frac{1}{\sigma^2}\|y_i - f(t_i,\psi_i)\|^2 + (\psi_i - m_i)'\Omega^{-1}(\psi_i - m_i)\right).\end{aligned} \tag{19}$$

In the linear model (17), we have seen that $\hat{\psi}_i$ is the conditional mean $\mu_i$ of a Gaussian distribution that can be computed in closed-form using (18). The use of an optimization procedure is necessary for computing $\hat{\psi}_i$ in the nonlinear case.

Once the MAP estimate $\hat{\psi}_i$ has been computed, the method is based on the linearisation of the structural model $f$ around $\hat{\psi}_i$:

$$f_i(\psi_i) \approx f_i(\hat{\psi}_i) + \nabla f_i(\hat{\psi}_i)(\psi_i - \hat{\psi}_i) \tag{20}$$

where $\nabla f_i(\hat{\psi}_i)$ is the Jacobian matrix of vector $f_i(\hat{\psi}_i)$.

Defining $z_i = y_i - f_i(\hat{\psi}_i) + \nabla f_i(\hat{\psi}_i)\hat{\psi}_i$, this development yields the following linear model

$$z_i = \nabla f_i(\hat{\psi}_i)\psi_i + \varepsilon_i \tag{21}$$

that can be used for approximating the conditional distribution of $\psi_i$ using the following Proposition (the proof is in the Appendix A):

**Proposition 1.** *Under linear model* (21)*, the conditional distribution of* $\psi_i$ *is a Gaussian distribution with mean* $\mu_i$ *and variance-covariance* $\Gamma_i$ *where*

$$\begin{aligned}\mu_i &= \hat{\psi}_i\\ \Gamma_i &= \left(\frac{\nabla f_i(\hat{\psi}_i)'\nabla f_i(\hat{\psi}_i)}{\sigma^2} + \Omega^{-1}\right)^{-1}.\end{aligned} \tag{22}$$

We then propose to use this normal distribution as a proposal in our MCMC procedure when the model is described by (7). Then, the closer the model is to a linear model, the higher the probability of accepting a candidate generated with this proposal.

**Remarks**:

1. It is interesting to note that the mode of the conditional distribution of $\psi_i$ in the nonlinear model is also the mode and the mean of the conditional distribution of $\psi_i$ in the linear model.

2. If we consider a more general error model, $\varepsilon_i \sim \mathcal{N}(0, \Sigma(t_i,\psi_i))$ where the variance-covariance matrix is not necessarily diagonal and may depend on the individual parameters $\psi_i$ and on the observation times $(t_{ij})$, then the variance-covariance matrix of the conditional distribution rewrites

$$\Gamma_i = \left(\nabla f_i(\hat{\psi}_i)'\Sigma(t_i,\hat{\psi}_i)^{-1}\nabla f_i(\hat{\psi}_i) + \Omega^{-1}\right)^{-1} \tag{23}$$

3. If it is not $\psi_i$ itself, but the transformed variable $\phi_i = u(\psi_i)$ which is normally distributed, then a candidate $\phi_i^c$ is drawn from a Gaussian proposal with parameters:

$$\begin{aligned}\mu_i &= \hat{\phi}_i \\ \Gamma_i &= \left( \frac{\nabla_\phi f_i(u^{-1}(\hat{\phi}_i))' \nabla_\phi f_i(u^{-1}(\hat{\phi}_i))}{\sigma^2} + \Omega^{-1} \right)^{-1}\end{aligned} \tag{24}$$

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

## 4.3 Non continuous data models

As far as non continuous outcomes, the model for the observations of individual $i$ is the conditional distribution of $y_i$ given the set of individual paramers $\psi_i$. There is no analytical relationship between the observations and the individual parameters and thus no linearisation method can be directly applied. Here, the strategy to build an efficient proposal consists in using a Laplace approximation of the joint model as described in [40] or [41].

**Proposition 2.** *Let $(y_i, \psi_i)$ be a pair of random variables where $\psi_i$ is normally distributed with variance-covariance matrix $\Omega$ and where the conditional pdf of $y_i$ is twice differentiable with respect to $\psi_i$. Let $\hat{\psi}_i$ be the value of $\psi_i$ that maximizes the conditional distribution $\boldsymbol{p}(\psi_i|y_i)$. Then, the conditional distribution of $\psi_i$ can be approximated by a Gaussian distribution with mean $\hat{\psi}_i$ and variance-covariance*

$$\Gamma_i = \left( -\nabla^2 l(\hat{\psi}_i) + \Omega^{-1} \right)^{-1}$$

*where $l(\psi_i) \triangleq \log\left(\boldsymbol{p}(y_i|\psi_i)\right)$.*

The proof is postponed to Appendix B.

**Remarks**:

1. In our context, the objective is not to approximate a conditional distribution as well as possible, but, above all, to determine a proposal that is effective in practice. Then, several solutions are possible when the variance-covariance matrix $-\nabla^2 l(\hat{\psi}_i) = -\nabla^2 \log(\mathtt{p}(y_i|\hat{\psi}_i))$ is tricky to calculate.

   This matrix is the *observed* Fisher information matrix evaluated at $\psi_i = \hat{\psi}_i$. It can be replaced by its conditional expectation, i.e. by the *expected* Fisher information matrix

   $$-\mathbb{E}\left( \nabla^2 \log l(\hat{\psi}_i) | \psi_i = \hat{\psi}_i \right) = \mathbb{E}\left( \nabla \log l(\hat{\psi}_i) \nabla \log l(\hat{\psi}_i)' | \psi_i = \hat{\psi}_i \right) \tag{25}$$

   Following (25), this matrix can also be replaced by $\nabla l(\hat{\psi}_i) \nabla l(\hat{\psi}_i)'$ which only requires to compute -or numerically approximate- the gradient of $l$.

2. In the case of continuous outcomes, linearising the structural model is equivalent to using the Laplace approximation with the expected information matrix. Indeed:

$$\begin{aligned}
&\mathbb{E}\left(-\nabla^2 \log l(\hat{\psi}_i)|\psi_i = \hat{\psi}_i\right) \\
&= \mathbb{E}\left(-\frac{1}{\sigma^2}\nabla^2 f_i(\hat{\psi}_i)(y_i - f_i(\hat{\psi}_i))' + \frac{\nabla f_i(\hat{\psi}_i)\nabla f_i(\hat{\psi}_i)'}{\sigma^2}|\psi_i = \hat{\psi}_i\right) \\
&= \frac{\nabla f_i(\hat{\psi}_i)\nabla f_i(\hat{\psi}_i)'}{\sigma^2}
\end{aligned}$$

# 5 Maximum Likelihood Estimation

## 5.1 The SAEM Algorithm

Let us see how this new version of the MH algorithm can be combined with the SAEM algorithm for computing the ML estimate of $\theta$ defined as $\hat{\theta}_{\text{ML}} = \arg\max_{\theta\in\Theta} \mathtt{p}(y;\theta)$.

In this incomplete data model context, iteration $k$ of SAEM consists in the following steps:

**Simulation step** for $i = 1, 2, \ldots, N$, drawing the latent data $\psi_i^{(k)}$ from a transition probability $\Pi^{(k)}(\psi_i^{(k-1)}, \cdot)$ which admits as unique limiting distribution the conditional distribution $\mathtt{p}(\psi_i|y_i;\theta_{k-1})$,

**Stochastic approximation step** updating the approximation of the conditional expectation $\mathbb{E}\left(\log \mathtt{p}(y,\psi;\theta)|y,\theta^{k-1}\right)$:

$$Q_k(\theta) = Q_{k-1}(\theta) + \gamma_k\left(\sum_{i=1}^{N} \log \mathtt{p}(y_i,\psi_i^{(k)};\theta) - Q_{k-1}(\theta)\right) \tag{26}$$

where $\{\gamma_k\}_{k>0}$ is a sequence of decreasing stepsizes with $\gamma_1 = 1$.

**Maximisation step** updating the estimation of $\theta$:

$$\theta^k = \arg\max_{\theta\in\Theta} Q_k(\theta) \tag{27}$$

The SAEM algorithm is implemented in most sofware tools for nonlinear mixed effects models and is known to be fast and very efficient in practice. Furthermode, it has been shown to converge to some maximum of the likelihood of the observations under very general conditions [5, 11].

The practical performances of SAEM are closely linked to the settings of SAEM. These settings consist essentially in the choice of the sequence of stepsizes $(\gamma_k)$ and the choice of the transition kernel $\Pi$.

The transition kernel $\Pi$ is directly defined by the proposal(s) used for the MH algorithm. The new proposal based on a normal approximation of the conditional distribution of $\psi_i$ is extremely efficient but requires much more computational effort than standard procedures. It is therefore not recommended to use it systematically in all iterations of SAEM. It is indeed more appropriate to combine this proposal with other proposals such as the independent, the block-wise and the component-wise MH algorithms mentioned above and already used in the context of nonlinear mixed effects models.

Assume that our new proposal is used at iteration $k$, then the simulation step of SAEM decomposes as follows:

1. compute the MAP under the current model parameter estimate $\theta_{k-1}$ for all individuals $i$:
$$\hat{\psi}_i^{(k)} = \arg\max_{\psi_i} \mathtt{p}(\psi_i | y_i, \theta_{k-1}) \tag{28}$$

2. Compute the covariance matrix $\Gamma_i^{(k)}$ such as:
$$\Gamma_i^{(k)} = \begin{cases} \left( \frac{\nabla f_i(\hat{\psi}_i^{(k)}) \nabla f_i(\hat{\psi}_i^{(k)})'}{\sigma^2} + \Omega^{-1} \right)^{-1} & \text{if the data model is continuous} \\ \left( \nabla l(\hat{\psi}_i^{(k)}) \nabla l(\hat{\psi}_i^{(k)})' + \Omega^{-1} \right)^{-1} & \text{otherwise} \end{cases} \tag{29}$$

3. Run a small number of iterations of the MH algorithm with the proposal $\mathcal{N}(\hat{\psi}_i^{(k)}, \Gamma_i^{(k)})$.

It is important to stress that convergence of SAEM does not require to achieve the convergence of the MCMC algorithm to the stationary distribution at each iteration of SAEM during the simulation step. What is important is to ensure that the sequence of transition probabilities $(\Pi^{(k)})$ converges to a transition probability $\Pi_{\hat{\theta}_{\text{ML}}}$ which admits as unique limiting distribution the conditional distribution $\mathtt{p}(\psi_i | y_i; \hat{\theta}_{\text{ML}})$. For this reason, only a small number of iterations of the MH algorithm are performed at each iteration of SAEM, whatever the proposal being used.

The behavior of SAEM also depends on the choice of the sequence $(\gamma_k)$. In practice, $\gamma_k$ is usually set equal to 1 during the first $K_1$ iterations so that the algorithm can explore the parameter space without memory and converge quickly to a neighborhood of the ML estimate. The stochastic approximation is performed during the next $K_2$ iterations with $\gamma_k = 1/k^a$, ensuring the almost sure convergence of the sequence $(\theta_k)$ as soon as $0.5 < a \leq 1$. Monolix uses $\alpha = 0.7$ while saemix, nlmefitsa and NONMEM use $a = 1$. The new proposal mainly aims at accelerating the algorithm during this first stage of SAEM. Indeed, We have conducted numerous numerical experiments that tend to confirm that the use of the new multidimensional proposal is most effective during the very first iterations of SAEM. Once close to the solution, component-wise and block-wise random walks can be used in an efficient way.

# 6 Numerical Examples

## 6.1 A pharmacokinetic example

### 6.1.1 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 [29]. 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).

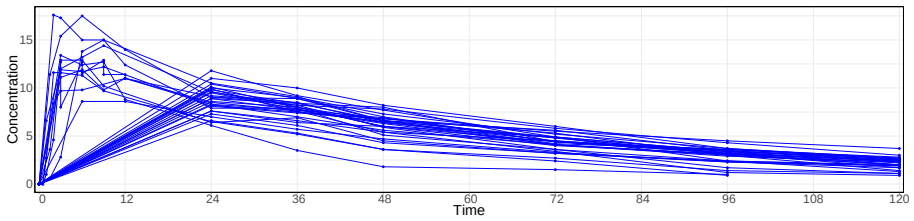


Figure 1: Warfarin concentration over time for 32 subjects

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

$$f(t, ka, V, k) = \frac{D\, ka}{V(ka - k)}(e^{-ka\, t} - e^{-k\, t}) \tag{30}$$

where $ka$ is the absorption rate constant, $V$ the volume of distribution , $k$ the elimination rate constant, and $D$ the dose administered.

Here, $ka$, $V$ and $k$ are PK parameters that can change from one individual to another. Then, 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 for individual $i$ writes

$$y_{ij} = f(t_{ij}, \psi_i) + \varepsilon_{ij} \tag{31}$$

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

$$\begin{aligned} \log(ka_i) &\sim \mathcal{N}(\log(ka_{\text{pop}}), \omega^2_{ka}) \\ \log(V_i) &\sim \mathcal{N}(\log(V_{\text{pop}}), \omega^2_{V}) \\ \log(k_i) &\sim \mathcal{N}(\log(k_{\text{pop}}), \omega^2_{k}) \end{aligned} \tag{32}$$

The model that is used here for these data is therefore the non-linear mixed effects model for continuous data described in Section 2.2. So we can use the proposal described given by Proposition 1 and based on a linearization of the structural model $f$ proposed in (30).

The structural model $f$ is quite simple in this example and the gradient could be calculated analytically. Nevertheless, for the method to be easily extended to any structural model, the gradient is calculated numerically by finite difference.

### 6.1.2 MCMC Convergence Diagnostic

We will start by examining the behavior of the MH algorithm used for sampling individual parameters from the conditional distribution $\mathtt{p}(\psi_i|y_i;\theta)$. We will consider only one of the 32 individuals for this study and fix $\theta$ to some arbitrary value, close to the ML estimate obtained with SAEM: $ka_{\text{pop}} = 1$, $V_{\text{pop}} = 8$, $k_{\text{pop}} = 0.01$, $\omega_{ka} = 0.5$, $\omega_V = 0.2$, $\omega_k = 0.3$ and $\sigma^2 = 0.5$.

First, we used the classical version of MH implemented in the saemix package and for which different transition kernels are used successively for each iteration: independent draw using the marginal distribution $\mathtt{p}(\psi_i)$, component-wise random walk and block-wise random walk.

We then replace the first independent proposal with our new proposal and keep the next two transition kernels. In practice, only a few iterations with this new combination of transition kernels is needed to converge close to the target value. The number of SAEM iterations using the new proposal will depends on the model and will be specified in the sequel.

We ran 10 000 iterations of these two algorithms and evaluated their convergence by looking at the convergence of the empirical quantiles of order 0.1, 0.5 and 0.9 for the three components of $\psi_i$. Here, $\hat{q}^k_\alpha(\psi_{i,\ell})$ is the empirical quantile of order $\alpha$ of $(\psi^{(1)}_{i,\ell}, \psi^{(2)}_{i,\ell}, \ldots, \psi^{(k)}_{i,\ell})$ and $\ell$ denotes the dimension of the individual parameter.

We see Figure 2 that, for all $\alpha$ and all $\ell$, the sequences of empirical quantiles $\hat{q}^k_\alpha(\psi_{i,\ell})$ obtained with the two algorithms converge to the same value, which is supposed to be the theoretical quantile of the conditional distribution.

The interest of the new proposal is shown here since we see that all the empirical quantiles obtained with this new algorithm converge faster than with the reference algorithm.

Finally, it is interesting to note that the empirical medians converge very rapidly. This is interesting in the population approach framework because it is mainly the central values of each conditional distribution that are used to infer the population distribution.

### 6.1.3 Maximum likelihood estimation of the parameters

Model parameters to estimate using the SAEM algorithm are the population PK parameters $ka_{\text{pop}}$, $V_{\text{pop}}$ and $k_{\text{pop}}$, the standard deviations of the random effects $\omega_{k_a}$, $\omega_V$ and $\omega_k$ and the residual variance $\sigma^2$.

The stepsize $\gamma_k$ was set to 1 during the first 100 iterations and then decreases

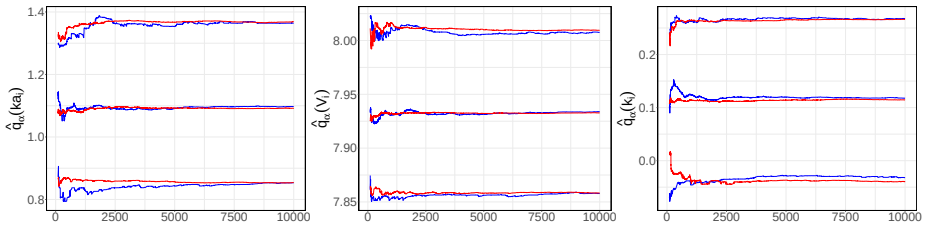


Figure 2: Modelling of the warfarin PK data: convergence of the empirical quantiles of order 0.1, 0.5 and 0.9 of $\mathtt{p}(\psi_i|y_i;\theta)$ for a single individual. The reference MH algorithm is in blue and the new version is in red.

as $1/k^a$ where $a = 0.7$ during the next 100 iterations.

Here we compare the standard SAEM algorithm, as implemented in saemix, with the version that uses the new proposal for the sampling of individual parameters in the simulation step. In this example, this new proposal is only used in the first 10 iterations of SAEM. The standard MH algorithm is then used.

Figure 3 shows the estimates of $V_{\text{pop}}$ and $\omega_V$ computed at each iteration of these two versions of SAEM and starting from three different initial values. First of all, we notice that, whatever the initialization and the sampling algorithm used, all the runs converge towards the same solution, i.e. towards the maximum likelihood estimate. It is then very clear that the new algorithm converges much faster and much more directly towards the solution than the standard algorithm.

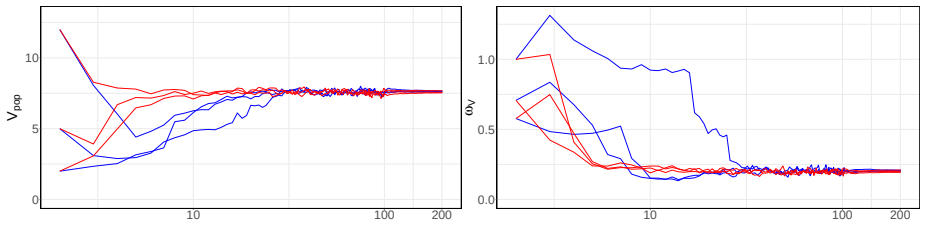


Figure 3: Modelling of the warfarin PK data: convergence of the sequences of estimates $(V_{\text{pop},k}, 1 \leq k \leq 200)$ and $(\omega_{V,k}, 1 \leq k \leq 200)$ obtained with SAEM and three different initial values using the reference MH algorithm (blue) and the new proposal during the first 10 iterations (red).

### 6.1.4 Monte Carlo study

A Monte Carlo study should confirm the very good properties of the new version of SAEM for estimating the parameters of the model.

$M = 50$ 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 was used for all the simulated datasets.

We are not interested here in the estimation errors, i.e. the differences between the values of the estimated parameters and those of the original parameters used for the simulation. Rather, we are interested in the convergence of the algorithm to the ML estimate.

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 $\theta_K^{(m)}$ is like studying 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 for each dimension $\ell$ of the parameter the average error over the replicates

$$E_k(\ell) = \frac{1}{M} \sum_{m=1}^{M} \left( \theta_k^{(m)}(\ell) - \theta_K^{(m)}(\ell) \right)^2 \tag{33}$$

Figure 4 shows very well that the use of the new proposal leads to a much faster convergence towards the ML estimate. Less than 10 iterations are required to converge with the new version of SAEM on this example, instead of 50 with the original version. It should also be noted that the error decreases monotonically. The sequence of estimates approaches the target with each iteration, compared to the standard algorithm which makes twists and turns before converging.

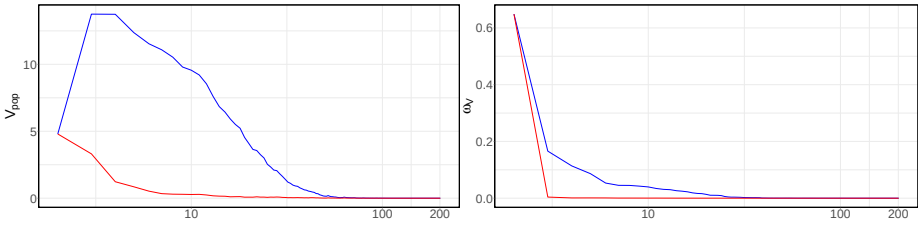


Figure 4: Convergence of the sequences of average errors $(E_k, 1 \leq k \leq 200)$ for $V_{\text{pop}}$ and $\omega_V$ obtained with SAEM on $M = 50$ synthetic datasets using the reference MH algorithm (blue) and the new proposal during the first 10 iterations (red).

## 6.2 Time to Event Data Model

### 6.2.1 The model

We will consider a repeated event model with right censoring ($\tau_c = 20$). We use the Weibull hazard function [18, 44] of parameter $\psi = (\lambda, \beta)$ defined by:

$$h(t, \psi) = \frac{\beta}{\lambda} \left( \frac{t}{\lambda} \right)^{\beta - 1} \tag{34}$$

A log transformation is applied on the individual parameters such as:

$$\begin{aligned} \log(\lambda_i) &\sim \mathcal{N}(\log(\lambda_{\text{pop}}), \omega_\lambda^2) \\ \log(\beta_i) &\sim \mathcal{N}(\log(\beta_{\text{pop}}), \omega_\beta^2) \end{aligned} \tag{35}$$

Formally, the model is defined by the conditional pdf of the $y_i$'s which can be derived using (12) and the Weibull model (34).

### 6.2.2 MCMC Convergence Diagnostic

Similarly to the previous section, we examine the behavior of the MCMC procedure for sampling the from the individual conditional distribution $\mathtt{p}(\psi_i|y_i;\theta)$ for a given individual $i$. A synthetic dataset is generated using the Weibull model with the following parameters value: $\lambda_{\text{pop}} = 10, \omega_\lambda = 0.3, \beta_{\text{pop}} = 3, \omega_\beta = 0.3$ for $N = 100$ individuals. In this section, the model parameter $\theta$ of the conditional distribution $\mathtt{p}(\psi_i|y_i;\theta)$ is equal to the parameter used for generating the dataset.

We ran 5 000 iterations of the two algorithms (using the extension of the saemix package to non continuous data models available at `https://github.com/belhal/saemix`), and evaluated the convergence of the empirical quantiles as explained previously.

We see Figue 5 that for all quantile order, the sequences of empirical quantiles converge to the same value. The new proposal presents a real advantage regarding the speed of convergence for all quantile orders including the empirical medians.

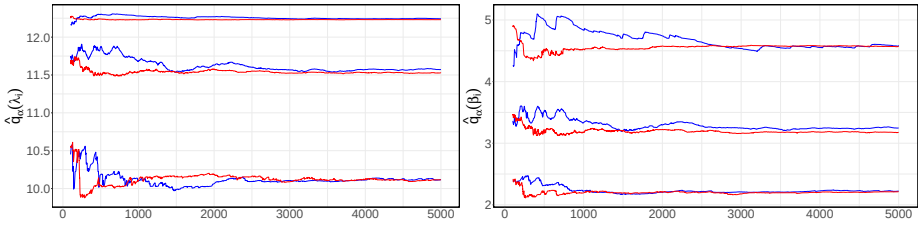


Figure 5: Convergence of the empirical quantiles of order 0.1, 0.5 and 0.9 of $\mathtt{p}(\psi_i|y_i;\theta)$ for a single individual. The reference MH algorithm is in blue and the new version is in red.

### 6.2.3 Maximum likelihood estimation of the parameters

Model parameters to estimate using the SAEM algorithm are the population parameters $\lambda_{\text{pop}}$, $\beta_{\text{pop}}$ and the standard deviations of the random effects $\omega_\lambda, \omega_\beta$. We generate a simulated dataset using the following generating values: $\lambda_{\text{pop}} = 10, \omega_\lambda = 0.3, \beta_{\text{pop}} = 3, \omega_\beta = 0.3$ for $N = 100$ individuals and right censoring time $\tau_c = 20$. The stepsize $\gamma_k$ was set to 1 during the first 100 iterations and then decreases as $1/k^a$ where $a = 0.7$ during the next 100 iterations.

We compare the standard SAEM algorithm using the extended version of the saemix package with the version that uses the new proposal for the sampling of the individual parameters in the simulation step. Here, the new proposal is only used the first 5 iterations of SAEM. The standard MH algorithm is then used.

Figure 6 shows the estimates of $\lambda_{\text{pop}}$ and $\omega_\lambda$ computed at each iteration of the two versions of the SAEM and starting from three different initial values. Same behaviour, as in the continuous case, may be identified here. Indeed, regardless the initial values, all the runs converge to the maximum likelihood estimate. Faster convergence is clearly observed in Figure 6.

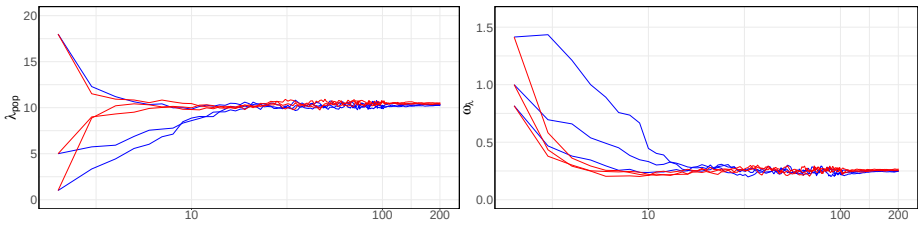


Figure 6: Modelling of simulated right censored time to event data: convergence of the sequences of estimates $(\lambda_{\text{pop},k}, 1 \leq k \leq 200)$ and $(\omega_{\lambda,k}, 1 \leq k \leq 200)$ obtained with SAEM and three different initial values using the reference MH algorithm (blue) and the new proposal during the first 5 iterations (red).

### 6.2.4 Monte Carlo study

Again, we conduct a Monte Carlo study to confirm the good properties of the new version of the SAEM algorithm for estimating the model parameters.

$M = 50$ synthetic datasets have been generated using the same design used in the previous section. Figure 7 shows the convergence of the average error defined by (33). Both algorithms converge monotonically to the ML estimate, yet very few iterations are required to converge with the newer version instead of 30 needed with the standard SAEM

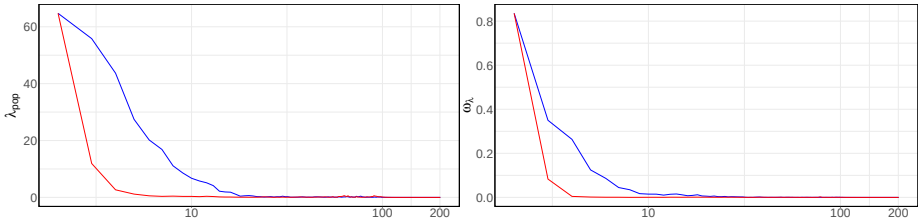


Figure 7: Convergence of the sequences of average errors $(E_k, 1 \leq k \leq 200)$ for $\lambda_{\text{pop}}$ and $\omega_\lambda$ obtained with SAEM on $M = 50$ synthetic datasets using the reference MH algorithm (blue) and the new proposal during the first 5 iterations (red).

# 7 Conclusion

We presented in this article a fast MCMC procedure embedded in the SAEM algorithm for performing ML estimation. Our new version is based a faster MCMC procedure for sampling from the individual conditional distribution $\mathsf{p}(\psi_i|y_i,\theta)$. In contrast with recent computationally expensive methods, ours take into consideration the covariance structure of the random effects through an approximation of this conditional distribution around its mode. Both real data analysis and Monte Carlo study showed the effectiveness of our approach.

These new proposals seem to approximate well the true conditional distributions, in the presented applications, yet, one can wonder to what extent these proposals are good approximations and can generalize to other models. Our future work consists in bridging the gap between our contribution and variational inference methods [38] that output an optimal proposal distribution, in the sense of Kullback Leibler divergence and see how those two proposals compare as far as MCMC convergence.

Incremental strategies, consisting in considering a mini-batch of the data at each iteration of the MCMC procedure have been developed in the recent years [22, 21] and seem to accelerate the convergence of the algorithm. When the training set size is too large, computing the MAP for each individual can be costly. Thus, incremental methods could allow our method to scale well where only a mini-batch of MAPs are computed.

# A Proof of Proposition 1

We can directly use (18) to get an expression of the conditional variance of $\psi_i$ under the linearized model:

$$\mathrm{Var}_{\mathrm{lin}}(\psi_i|y_i) = \left( \frac{\nabla f_i(\hat{\psi}_i) \nabla f_i(\hat{\psi}_i)'}{\sigma^2} + \Omega^{-1} \right)^{-1} \tag{36}$$

On the other hand, it was shown in Section 4.1 that the MAP is defined as

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

where $f_i(\psi_i)$ is the vector $(f(t_{i,1}, \psi_i), \ldots, f(t_{i,n_i}, \psi_i))$. Thus, $\hat{\psi}_i$ satisfies:

$$-\frac{\nabla f_i(\hat{\psi}_i)'}{\sigma^2} \left( y_i - f_i(\hat{\psi}_i) \right) + \Omega^{-1}(\hat{\psi}_i - m_i) = 0$$

Let $\Gamma_i = \mathrm{Var}_{\mathrm{lin}}(\psi_i|y_i)$. Using (18), we can now compute the conditional mean of $\psi_i$ under the linearized model:

$$\begin{aligned}
\mathbb{E}_{\mathrm{lin}}(\psi_i|y_i) &= \Gamma_i \frac{\nabla f_i(\hat{\psi}_i)'}{\sigma^2} \left( y_i - f_i(\hat{\psi}_i) + \nabla f_i(\hat{\psi}_i) \hat{\psi}_i + \Omega^{-1} m_i \right) \\
&= \Gamma_i \left( \Omega^{-1}(\hat{\psi}_i - m_i) + \frac{\nabla f_i(\hat{\psi}_i)' \nabla f_i(\hat{\psi}_i)}{\sigma^2} \hat{\psi}_i + \Omega^{-1} m_i \right) \\
&= \Gamma_i \Gamma_i^{-1} \hat{\psi}_i \\
&= \hat{\psi}_i
\end{aligned} \tag{37}$$

# B Proof of Proposition 2

Laplace approximation (see [8]) consists in approximating an integral of the form

$$I := \int e^{v(x)} dx \tag{38}$$

where $v$ is at least three times differentiable.

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

$$v(x) \approx v(x_0) + \nabla v(x_0)(x - x_0) + \frac{1}{2}(x - x0) \nabla^2 v(x_0)(x - x0) \tag{39}$$

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

$$I \approx e^{v(x_0)} \sqrt{\frac{(2\pi)^p}{|-\nabla^2 v(x_0)|}} \exp\left\{ -\frac{1}{2} \nabla v(x_0)' \nabla^2 v(x_0)^{-1} \nabla v(x_0) \right\} \tag{40}$$

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

$$\begin{aligned}\mathtt{p}(y_i) &= \int \mathtt{p}(y_i, \psi_i)\mathtt{d}\psi_i \\ &= \int e^{\log(\mathtt{p}(y_i,\psi_i))}\mathtt{d}\psi_i\end{aligned} \tag{41}$$

Then, let

$$\begin{aligned}v(\psi_i) &= \log(\mathtt{p}(y_i, \psi_i)) \\ &= l(\psi_i) + \log(\mathtt{p}(\psi_i))\end{aligned} \tag{42}$$

and we do the Taylor expansion around the MAP $\hat{\psi}_i$ that verifies by definition $\nabla \log \mathtt{p}(y_i, \hat{\psi}_i) = 0$:

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

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

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

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

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

# References


[1] A. Durmus, E. M., and Pereyra, M. Sampling from convex non continuously differentiable functions, when Moreau meets Langevin.

[2] A. Gelman, D. L., and Guo, Y. Stan: A probabilistic programming language for Bayesian inference and Optimization. *Journal Of Statistical Software* (2015).

[3] Agresti, A. An Introduction to Categorical Data Analysis. *Wiley Series in Probability and Statistics 423* (2007).

[4] Andersen, P. K. Survival Analysis. *Wiley Reference Series in Biostatistics* (2006).

[5] B. Delyon, M. L., and Moulines, E. Convergence of a stochastic approximation version of the EM algorithm. *The Annals of Statistics* (1999).

[6] C. Mbogning, K. B., and Lavielle, M. 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* (2015).

[7] C. Zhang, B. S., and Zhao, H. Variational hamiltonian monte carlo via score matching.

[8] D. Gamerman, H. M., and Louzada, F. Statistical Inference: An Integrated Approach, Second Edition.

[9] Durmus, A., Le Corff, S., Moulines, E., and Roberts, G. O. O. Optimal scaling of the Random Walk Metropolis algorithm under Lp mean differentiability. *Journal of Applied Probability 54*, 4 (2017), 1233 – 1260.

[10] E. Comets, A. L., and Lavielle, M. Parameter estimation in nonlinear mixed effect models using saemix. *Journal of Statistical Software* (2017).

[11] E.Kuhn, M. L. Coupling a Stochastic Approximation version of EM with an MCMC Procedure. *ESAIM: Probability and Statistics*, 8 (2015).

[12] E.Kuhn, S. A. Convergent Stochastic Expectation Maximization algorithm with efficient sampling in high dimension. Application to deformable template model estimation.

[13] G. O. Roberts, A. G., and Gilks, W. R. Weak Convergence and Optimal Scaling of Random Walk Metropolis Algorithms.. *Ann. Appl. Probab.* (1997).

[14] G. O. Roberts, J. S. R. Optimal scaling of discrete approximations to Langevin diffusions. *Journal of the Royal Statistical Society* (1998).

[15] G.O. Roberts, R. T. Exponential con- vergence of Langevin distributions and their discrete approximations. *Bernoulli* (1996).

[16] Hoffman, M., and Gelman, A. The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo. *Journal Of Machine Learning Research* (2014).

[17] Lavielle, M. A stochastic algorithm for parametric and non-parametric estimation in the case of incomplete data . *Elsevier Science* (1993).

[18] Lavielle, M. Mixed Effects Models for the Population Approach.

[19] Louis, T. Finding the Observed Information Matrix when using the EM algorithm. *JRSS 44* (1982).

[20] M. Girolami, B. C. Riemann manifold Langevin and Hamiltonian Monte Carlo methods. *Journal of the Royal Statistical Society* (2011).

[21] Maclaurin, D., and Adams, R. P. Firefly monte carlo: Exact MCMC with subsets of data. In *IJCAI* (2015), AAAI Press, pp. 4289–4295.

[22] Maire, F. Adaptive Incremental Mixture Markov chain Monte Carlo.

[23] Mengersen, K., and Tweedie, R. Rates of convergence of the Hastings and Metropolis algorithms. *The Annals of Statistics* (1996).

[24] N. de Freitas, P. Hojen-Sorensent, M. J., and Russell, J. Variational mcmc.

[25] N. Metropolis, A.W. Rosenbluth, M. R. Equations of state calculations by fast computing machine.

[26] O. Stramer, R. T. Langevin-typemodels i: Diffusions with given stationary distributions, and their discretizations. *Methodol. Comput. Appl. Prob.* (1999).

[27] P.L.S Chan, P. J., and Lavielle, M. 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* (2011).

[28] R. M. Savic, F. M., and Lavielle, M. Implementation and Evaluation of the SAEM Algorithm for Longitudinal Ordered Categorical Data with an Illustration in Pharmacokinetics–Pharmacodynamics. *The AAPS Journal* (2011).

[29] RA. O'reilly, P. A. Studies on coumarin anticoagulant drugs. Initiation of warfarin therapy without a loading dose.

[30] Robert, C. The metropolis–hastings algorithm.

[31] Robert, C. P., and Casella, G. *Monte Carlo Statistical Methods (Springer Texts in Statistics)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2005.

[32] Roberts, G., and Tweedie, R. Geometric convergence and central limit theorems for multidimensional Hastings and Metropolis algorithms. *Biometrika* (1996).

[33] S. Beal, L.B. Sheiner, A. B., and Bauer, R. NONMEM's User's Guide. *ICON Development Solutions* (2009).

[34] S. Duane, A.D. Kennedy Brian, J. P., and Duncan, R. Hybrid Monte Carlo. *Physics Letters B* (1987).

[35] T. Marshall, G. R. An adaptive approach to langevin MCMC. *Statistics and Computing* (2012).

[36] T. Salimans, D. P. K., and Welling, M. Markov chain monte carlo and variational inference: Bridging the gap.

[37] THT Nguyen, M-S Mouksassi, N. H. N. A. I. F. A. H. J. J. M. K. D. M. J. P. R. E. P. R. S. J. v. H. B. W. C. Z. E. C. F. M. Model Evaluation of Continuous Data Pharmacometric Models: Metrics and Graphics. *CPT Pharmacometrics Syst Pharmacol. 2* (2017), 87–109.

[38] Wainwright, M. J., and Jordan, M. I. Graphical models, exponential families, and variational inference. *Found. Trends Mach. Learn. 1* (2008), pp. 1–305.

[39] Wei, G., and Tanner, M. A Monte-Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms. *J. Amer. Statist. Assoc.* (1990).

[40] Wolfinger, R. Laplace's Approximation for Nonlinear Mixed Models. *Biometrika 80* (2017).

[41] Y., W. Derivation of various NONMEM estimation methods.. *Journal of Pharmacokinetics and Pharmacodynamics* (2007).

[42] Y. Atchade, J. R. On adaptive Markov chain Monte Carlo algorithms . *Bernoulli* (2005).

[43] Y. Ma, T. Chen, E. F. A Complete Recipe for Stochastic Gradient MCMC. *NIPS'15 Proceedings of the 28th International Conference on Neural Information Processing Systems* (2015).

[44] Zhang, Z. Parametric regression model for survival data: Weibull regression model as an example. *Ann Transl Med. 24* (2016).