Fast Maximum Likelihood Estimation Algorithm Using Efficient MCMC proposal

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

March 23, 2018

Population models are widely used in domains like pharmacometrics 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. When the conditional expectation of the complete log likelihood is hard to compute, the Maximum Likelihood estimates are obtained using a stochastic version of the EM algorithm. Yet, this method implies being able to sample from the posterior distribution of the parameters given the observed data. A Markov Chain Monte Carlo procedure can be used to perform this simulation.
Our contribution consists in accelerating this posterior sampling in order to improve the overall parameter estimation algorithm convergence properties. For both continuous and non continuous data models, we build a proposal for the MCMC procedure that takes into account the multidimensional and covariance structure of the individual parameters. This proposal stems from an approximation of the true posterior distribution using a Taylor expansion of the structural model, for continuous data models, or the log likelihood otherwise, around the computable mode of the true posterior distribution.
We give experimental results on real and simulated pharmacokinetics datasets showing the effectiveness of our technique.

1 Introduction

We consider a complete model (y, ψ) where the realisations of y are observed and ψ is the missing data. When the complete model $\mathtt{p}(y, \psi, \theta)$ is parametric, the goal is to compute the maximum likelihood (ML) estimate of the parameter of the incomplete likelihood:

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

With the incomplete likelihood defined as:

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

When the direct derivation of this expression is hard, several methods use the complete model to iteratively find the quantity of interest. The EM algorithm has been the object of considerable interest since its presentation by Dempster, Laird and Rubin in 1977, see [DR77]. It has been relatively effective in context of maximum likelihood estimation of parameters of incomplete model (unobserved or more). This algorithm is monotonic in likelihood making it a stable tool to work with. This two steps algorithm consists in maximizing an auxiliary quantity that is the expectation of the complete log-likelihood with respect to the conditional distribution over the missing variable conditioned on the observed data and the current parameter estimate (also called the posterior distribution), see [WU83] for more details. Yet, when the quantity computed at the E-step involves unfeasible computations, new methods have been developed in order to by-pass the issue. Most of them alleviate the computation of the expectation using approximates. The Monte Carlo EM (MCEM) algorithm, first introduced in [WT90], 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 [Lav93]. 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 proven in [EKu15].
In this article, we aim at modifying the simulation step of the algorithm and propose an alternative MCMC procedure in order to accelerate the global convergence. Current solutions include using a Random Walk Metropolis algorithm that builds the Markov Chains without taking into consideration the covariance structure of the random effects. Indeed, this MCMC procedure consists in proposing, along each dimension, a new candidate sampled from a univariate Gaussian distribution centered in the current state with an adaptive variance, see [Y A05].
More recently, Stochastic Gradient MCMC methods, see [Y M15], 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, see [G O98] and Hamiltonian dynamics yields to the Hamiltonian Monte Carlo Algorithm, see [M G11]. Those methods scale well in high dimension but can be hard to tune and implement.
Our contribution consists in using an approximation of the conditional distribution of the individual parameters given the observations as a proposal distribution. In the context of NLME Models, when the observed outcomes are continuous, this approximation consists in linearising the non linear structural model around the Maximum a Posteriori. When the outcomes are non continuous, we use Laplace approximation of the incomplete log-likelihood. Those approximations result in new proposals that considerably accelerate the SAEM algorithm.

We present numerical examples of this new method to several models such as continuous models or repeated time-to-event data as developed in [CL15].

2 Notations and Models

2.1 Population and hierarchical model approach

In the sequel, we adopt a population approach where we consider several observations per individual. We denote by N the number of individuals in the population and n_i the number of observations per individual i. Let us define the observed data $y = (y_i, 1 \leq i \leq N)$ where $y_i = (y_{ij}, 1 \leq j \leq n_i)$ is the vector of observations y_{ij} that take their values in a subset of $\mathbb{R}^l$. The distribution of the vector of observations y_i depends on the vector of individual parameters ψ_i where $(\psi_i, 1 \leq i \leq N)$ take their values in a subset of $\mathbb{R}^p$.
We also assume that the couples (y_i, ψ_i) are mutually independent and consider a parametric framework where the distribution of the couple (y_i, ψ_i) is denoted by $\mathtt{p}(y_i, \psi_i; \theta)$. A natural decomposition of this joint distribution consists in writing:

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

Where $\mathtt{p}(\psi_i; \theta)$ is the so-called population distribution used to describe the distribution of the individual parameters within the population.
We can define the incomplete likelihood noted $L(\theta; y)$ by:

$$L(\theta; y) \triangleq \mathtt{p}(y; \theta) = \prod_{i=1}^{N} \mathtt{p}(y_i; \theta) \tag{4}$$

The ML estimate of θ is thus defined by:

$$\hat{\theta}_{ML} = \arg\max_{\theta \in \Theta} L(\theta, y) \tag{5}$$

2.2 Mixed Effect Models

A particular case of this general framework consists in describing each individual parameters ψ_i as a composition of fixed effects, common to the whole population, and random effects as follows:

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

Where u is a strictly monotonic transformation applied on the individual parameters ψ_i and can be the identity function (ψ_i is thus normally distributed), the logarithmic function (ψ_i has a log-normal distribution) or the logit or probit transformations. The vector ψ_{pop} is an unknown vector of fixed effects and η_i are the random effects. There are several possible options for defining the distribution of η_i. In the sequel, we will consider a multivariate Gaussian distribution $\eta_i \sim \mathcal{N}(0, \Omega)$.

An extension to this model consists in adding covariates to illustrate observed inter-individuals variability, as in [Lav15]:

$$u(\psi_i) = u(\psi_{pop}) + C_i\beta + \eta_i \tag{7}$$

With β a new vector of fixed effects and C_i a matrix of individual covariates.
In the following, we will use either the parameters ψ_i or the Gaussian transformed parameters $u(\psi_i)$ (normally distributed according to (6)).

2.2.1 Continuous data models

When the observations are continuous, the link between the observations and the individual parameters can be written as the following regression:

$$y_{ij} = f(t_{ij}, \psi_i) + \epsilon_{ij} \tag{8}$$

Where y_{ij} is the j-th observation for individual i measured at time t_{ij}, ϵ_{ij} is the residual error, f is the structural model and is a continuous and twice differentiable function of ψ_i.
We assume that the residual errors are independent and normally distributed with a constant variance σ^2, thus the hierarchical model rewrites:

$$\begin{aligned} y_i|\psi_i &\sim \mathcal{N}(f(t_i, \psi_i), \sigma^2 \mathtt{Id}_{n_i \times n_i}) \\ u(\psi_i) &\sim \mathcal{N}(u(\psi_{pop}), \Omega) \end{aligned} \tag{9}$$

Consequently, the overall model parameter to estimate is $\theta = (\psi_{pop}, \Omega, \sigma^2)$.
We remark that if the transformation u and the structural model f are both linear with respect to ψ_i, the model is a so-called linear mixed effects model.
An extension to this model consists in considering a regression model with heteroscedastic errors which means that the variance of the residual errors differs among the observations:

$$\epsilon_{ij} \sim \mathcal{N}(0, g(t_{ij}, \psi_i)^2 \mathtt{Id}_{n_i \times n_i}) \tag{10}$$

The error model can also be one of proportional, combined or exponential [EL17] but this choice does not impact the method developed in the sequel.

2.2.2 Non continuous data models

Non continuous data models include categorical data [RL11; Agr07], time to event data [CL15; And06], count data [RL11] models.
When the outcome is categorical, we denote by y_{ij} the observation that takes its value in a set $\{1, \dots, K\}$ of K categories. Thus, the model is defined by the probabilities $\mathbb{P}(y_{ij} = k|\psi_i)$ that are function of the couple (t_{ij}, ψ_i).
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 one-off event. The survival function $S(t)$ gives the probability that the event happens after time t:

$$\begin{aligned} S(t) &\triangleq \mathbb{P}(T > t) \\ &= e^{-\int_0^t h(u)du} \end{aligned} \tag{11}$$

Where h is called the hazard function.
In the population and hierarchical model approach, we will consider a parametric

and individual hazard function $h(., \psi_i)$.
Depending on the application, the length of time to this event may be called the survival time or the failure time. In general, we simply say time-to-event. 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 a single event right censoring model where the observations are:

$$y_i = \begin{cases} T_i & \text{if } T_i < t_c \\ "T_i > t_c" & \text{otherwise} \end{cases} \tag{12}$$

Where t_c is the censoring time and "$T_i > t_c$" is the information that the event occurred after the censoring time. The probabilities of the event for each individual are then defined as:

$$\mathbb{P}(T_i > t) = e^{-\int_0^t h(u,\psi_i)du} \tag{13}$$

For repeated event models, the observations are a sequence of time T_{ij} at which events occurred for individual i and the information that the censoring time has been reached "$T_{i,n_i} > t_c$". The probabilities of events rewrite:

$$\mathbb{P}(T_{ij} > t | T_{i,j-1} = t_{i,j-1}) = e^{-\int_{t_{i,j-1}}^t h(u,\psi_i)du} \tag{14}$$

3 Posterior sampling

3.1 Metropolis Hastings Algorithm

Metropolis-Hastings (MH) are a powerful class of inference algorithms that belong to the family of MCMC methods. MH algorithms are used to sample from a posterior distribution $\pi(.|y_i)$ for which direct sampling is difficult. It consists in iteratively constructing a Markov Chain:

- Drawing candidate state ψ_i^c from a proposal distribution q.
- Computing the MH ratio:

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

  Where ψ_i is the current state of the chain.

- Accepting ψ_i^c with probability $\min(1, \alpha(\psi_i^c, \psi_i))$

Under some general conditions [MT96; GG97; RT96], the convergence of the MH sampler to the stationary distribution has been proven.
The choice of the proposal distribution q plays a crucial role in the convergence behaviour of the chain and one can heuristically acknowledge that the closer the proposal is to the target distribution, the faster the chain will converge.
Current implementation of the SAEM algorithm in Monolix ([PL11]), saemix (R package) ([EL17]), nlmefitsa (Matlab) and NONMEM ([SB09]) uses a combination of three proposal distributions. The first one samples the candidate

state from the prior distribution of the individual parameters, that is independent of the current state, we call it an independent MH. Then the second and third, called Random Walk Metropolis (RWM) ([N M53]), updates the chain states by component using univariate Gaussian proposal distributions for each dimension of the state. These proposals are centered in the current state and have a variance Γ that adapts ([Y A05]) to the acceptance rate at each iteration. High acceptance rates can be achieved with these methods by proposing smaller transitions, however larger amounts of time will then be required to make long traversals of the state space. Ideal algorithm would make large transitions that are accepted with high probability.
Nevertheless, the univariate structure of those proposals:

- does not scale well in high dimension
- 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) ([GO 96; O S99]) and the Hamiltonian Monte Carlo (HMC) which implementation can be found as the "No U-Turns Sampler" in STAN [HG14; AG15].
The MALA consists in proposing a new state ψ_i^c using the gradient of the target measure at the current state ψ_i^m:

$$\psi_i^c \sim \mathcal{N}(\psi_i^m - \gamma_m \nabla_{\psi_i} \log \pi(\psi_i^m), 2\gamma_m) \tag{16}$$

Where $(\gamma_m)_{m>0}$ is a sequence of positive integers. It is a particular case of the RWM with a drift term [Y M15] 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 [T M12; Y A05]. 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 [EKu13] 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 [AP16].
On the other hand, the HMC algorithm, introduced in [SD87], 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 [M G11] suggests taking 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 hard to implement (the stepsizes and numerical derivatives need to be tuned and implemented).
In this article, we describe a method to construct a proposal for both continuous and non continuous data models that takes into account the multidimensional structure of the individual parameters in order to accelerate the MCMC procedure.

3.2 Proposals construction for MH algorithm

3.2.1 Maximum A Posteriori

The Maximum A Posteriori (MAP) (also known as Empirical Bayes Estimate (EBE)) is the value that maximizes the posterior distribution $\mathtt{p}(\psi_i|y_i,\theta)$ for a fixed θ:

$$\begin{aligned}\hat{\psi}_i &= \arg\max_{\psi_i} \mathtt{p}(\psi_i|y_i,\theta) \\ &= \arg\max_{\psi_i} \mathtt{p}(y_i|\psi_i,\theta)\mathtt{p}(\psi_i,\theta)\end{aligned} \tag{17}$$

3.2.2 Continuous Data Models

In the case where the model is described by (8) and ψ_i is normally distributed, our new method is based on the linearisation of the structural model around the MAP.
The Taylor expansion of f around $\hat{\psi}_i$, for all individual $i \in [1, N]$ gives:

$$f(t_i,\psi_i) \approx f(t_i,\hat{\psi}_i) + \nabla_{\psi_i} f(t_i,\hat{\psi}_i).^\top(\psi_i - \hat{\psi}_i) \tag{18}$$

This development defines the following linear model between y_i and ψ_i:

$$y_i - f(t_i,\hat{\psi}_i) + \nabla_{\psi_i} f(t_i,\hat{\psi}_i).^\top\hat{\psi}_i = \nabla_{\psi_i} f(t_i,\hat{\psi}_i).^\top\psi_i + \epsilon_i \tag{19}$$

We show in appendix A that under this linear model, the posterior distribution $\psi_i|y_i$ is tractable and follows a Gaussian distribution $\mathcal{N}(\mu_i, \Gamma_i)$ with parameters:

$$\begin{aligned}\mu_i &= \hat{\psi}_i \\ \Gamma_i &= \left[\frac{\nabla_{\psi_i} f(\hat{\psi}_i).^\top \nabla_{\psi_i} f(\hat{\psi}_i)}{\sigma^2} + \Omega^{-1}\right]^{-1}\end{aligned} \tag{20}$$

We use this latter as a proposal distribution for the MCMC procedure when the model is described by (8).

Remark.1: We note that the mode of the posterior distribution under both non linear and linear models are equal.

Remark.2: If we assume an heteroscedastic error model, i.e. $\epsilon_i \sim \mathcal{N}(0, \Sigma(\psi_i))$ with a covariance matrix that depends on the individual parameters and is not diagonal, then the parameters rewrite:

$$\begin{aligned}\mu_i &= \hat{\psi}_i \\ \Gamma_i &= \left[\nabla_{\psi_i} f(\hat{\psi}_i).^\top \Sigma(\hat{\psi}_i)^{-1} \nabla_{\psi_i} f(\hat{\psi}_i) + \Omega^{-1}\right]^{-1}\end{aligned} \tag{21}$$

Remark.3: If the transformed variable $\phi_i = u(\psi_i)$ is normally distributed, then the candidate state ϕ_i^c is drawn from the resulting proposal distribution with the following parameters:

$$\begin{aligned}\mu_i &= \hat{\phi}_i \\ \Gamma_i &= \left[\nabla_{\phi_i} f(\hat{\phi}_i).^\top \Sigma(\hat{\phi}_i)^{-1} \nabla_{\phi_i} f(\hat{\phi}_i) + \Omega^{-1}\right]^{-1}\end{aligned} \tag{22}$$

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

3.2.3 Non Continuous Data Models

As far as non continuous outcomes, there is no analytical relationship between the observations and the individual parameters and thus no linearisation can be applied. Here, the strategy to build an efficient proposal consists in using a Laplace approximation of the joint model as described in [Wol17] or [Y07].
Define $l(\psi_i) \triangleq \mathtt{p}(y_i|\psi_i)$. We show in appendix B that the proposal resulting from this approximation is a Gaussian distribution with the following parameters:

$$\begin{aligned}\mu_i &= \hat{\psi}_i \\ \Gamma_i &= \left[-\nabla^2 \log l(\hat{\psi}_i) + \Omega^{-1}\right]^{-1}\end{aligned} \tag{23}$$

In the case where the calculus of $\nabla^2 \log l(\hat{\psi}_i) = \nabla^2 \log \mathtt{p}(y_i|\hat{\psi}_i)$ is hard, we approximate the Fisher information $-\nabla^2 \log l(\hat{\psi}_i)$ by its conditional expectation

$-\mathbb{E}^{y_i|\hat{\psi}_i}\left[\nabla^2 \log l(\hat{\psi}_i)\right]$. And using the Fisher identity stated as:

$$\mathbb{E}^{y_i|\hat{\psi}_i}\left[\nabla^2 \log l(\hat{\psi}_i)\right] = -\mathbb{E}^{y_i|\hat{\psi}_i}\left[\nabla \log l(\hat{\psi}_i).\nabla \log l(\hat{\psi}_i)^\top\right] \tag{24}$$

we obtain:

$$\begin{aligned} -\nabla^2 \log l(\hat{\psi}_i) &\approx -\mathbb{E}^{y_i|\hat{\psi}_i}\left[\nabla^2 \log l(\hat{\psi}_i)\right] \\ &= \mathbb{E}^{y_i|\hat{\psi}_i}\left[\nabla \log l(\hat{\psi}_i).\nabla \log l(\hat{\psi}_i)^\top\right] \\ &\approx \frac{\nabla l(\hat{\psi}_i).\nabla l(\hat{\psi}_i)^\top}{l^2(\hat{\psi}_i)} \end{aligned} \tag{25}$$

The proposal resulting from this last approximation is a Gaussian distribution with the following parameters:

$$\begin{aligned} \mu_i &= \hat{\psi}_i \\ \Gamma_i &= \left[\frac{\nabla l(\hat{\psi}_i).\nabla l(\hat{\psi}_i)^\top}{l^2(\hat{\psi}_i)} + \Omega^{-1}\right]^{-1} \end{aligned} \tag{26}$$

Remark.4: In the case of continuous outcomes, using Laplace approximation of the incomplete log-likelihood results in the proposal obtained by linearising the structural model. Indeed:

$$\begin{aligned} \mathbb{E}^{y_i|\hat{\psi}_i}\left[-\nabla^2_{\psi_i} \log l(\hat{\psi}_i)\right] &= \mathbb{E}^{y_i|\hat{\psi}_i}\left[-\frac{1}{\sigma^2}\nabla^2_{\psi_i} f(\hat{\psi}_i).(y_i - f(\hat{\psi}_i))^\top + \frac{\nabla_{\psi_i} f(\hat{\psi}_i).\nabla_{\psi_i} f(\hat{\psi}_i)^\top}{\sigma^2}\right] \\ &= \frac{\nabla_{\psi_i} f(\hat{\psi}_i).\nabla_{\psi_i} f(\hat{\psi}_i)^\top}{\sigma^2} \end{aligned} \tag{27}$$

4 Maximum Likelihood Estimation

4.1 The SAEM Algorithm

In this incomplete data model context, the estimation algorithm consists in, at iteration k:

1. Sampling latent data $\psi_i^k \sim \mathtt{p}(\psi_i|y_i;\theta^{k-1})$ under the current model parameter estimate θ^{k-1} for $i \in [\![1, N]\!]$
2. Updating the stochastic approximation $Q_k(\theta)$ of the quantity $\mathbb{E}\left[\log \mathtt{p}(y,\psi;\theta)|y,\theta^{k-1}\right]$:

$$Q_k(\theta) = Q_{k-1}(\theta) + \gamma_k\left[\sum_{i=1}^{N} \log \mathtt{p}(y_i,\psi_i^k;\theta) - Q_{k-1}(\theta)\right] \tag{28}$$

   Where $\{\gamma_k\}_{k>0}$ is a sequence of positive stepsize with $\gamma_1 = 1$.

3. Maximisation step:

$$\theta^k = \arg\max_{\theta\in\Theta} Q_k(\theta) \tag{29}$$

The SAEM algorithm has been shown theoretically to converge to a maximum of the likelihood of the observations under very general conditions [BM99]. In the simulation step, since the relation between the observed data and the individual parameters can be non linear, sampling from the posterior distribution requires using an inference algorithm. Kuhn et al. in [EKu15] proved almost sure convergence of the sequence of parameters obtained by this algorithm coupled with an MCMC procedure during the simulation step.

In the stochastic approximation step, the sequence of decreasing positive integers γ_k controls the convergence of the algorithm. In practice, γ_k is set equal to 1 during the first K1 iterations to let the algorithm explore the parameter space without memory and to converge quickly to a neighbourhood of the ML estimate. Our contribution aims at accelerating the algorithm during this first part. The stochastic approximation is performed during the final K2 iterations where $\gamma_k = \frac{1}{k}$, ensuring the almost sure convergence of the estimate.

We focus on the simulation step of the algorithm, during the first K1 iterations, since the maximisation step is usually performed well and the challenge in those types of problem to sample efficiently from the intractable multidimensional conditional distribution $p(\psi_i|y_i, \theta)$. The coupling with an MCMC procedure suggests choices of kernel transition that could accelerate the convergence of the Markov Chain. We present in this article specific proposal distribution to speed the convergence of the maximum likelihood estimation. In the sequel, the SAEM algorithm refers to this MCMC coupled version using the proposals detailed in section 3.1. Theoretically, those Markov Chains are converging to the posterior distribution we want to sample from. Yet, in practice, one does not wait for convergence of the MCMC procedure at each iteration of the SAEM algorithm and only runs a couple of iterations.
Consequently, the goal is to accelerate the overall convergence behaviour with a fixed and low number of MCMC iterations.

4.2 The Fast SAEM Algorithm

The fast version of the SAEM algorithm we propose in this article leaves the stochastic approximation update and the maximisation step of algorithm 4.1 unchanged but samples candidate states as follows:

1. Compute the MAP under the current model parameter estimate θ^{k-1} for all individuals i:
$$\hat{\psi}_i^k = \arg\max_{\psi_i} \mathrm{p}(\psi_i|y_i, \theta^{k-1}) \tag{30}$$

2. Compute the covariance matrix Γ_i^k such as:
$$\Gamma_i^k = \begin{cases} \left[\frac{\nabla f(\hat{\psi}_i^k).\nabla f(\hat{\psi}_i^k)^\top}{\sigma^2} + \Omega^{-1}\right]^{-1} & \text{if the data model is continuous} \\ \left[\frac{\nabla l(\hat{\psi}_i^k).\nabla l(\hat{\psi}_i^k)^\top}{l^2(\hat{\psi}_i^k)} + \Omega^{-1}\right]^{-1} & \text{otherwise} \end{cases} \tag{31}$$

3. Sample individual parameters ψ_i^k from the posterior distribution $p(\psi_i|y_i, \theta^{k-1})$ using an Independent MH algorithm with proposal $\mathcal{N}(\hat{\psi}_i^k, \Gamma_i^k)$

In practice this new proposal is very efficient during the first iterations of the algorithm and can be replaced by current RWM kernel transitions once the algorithm has converged to a neighbourhood of the ML estimate.

5 Numerical Examples

In this section we will compare the SAEM convergence properties of both methods on both continuous and non continuous data models using real and simulated datasets. Using simulated data has proven to be a powerful approach to derive properties of such algorithms [L A01]. In [Yao00], Yao defines and studies an online stochastic approximation scheme. In this situation, the number of observations goes to infinity and the sequence of estimates converges to the true value of the parameter.

5.1 A Pharmacokinetic example

5.1.1 The data

Warfarin is an anticoagulant normally used in the prevention of thrombosis and thromboembolism, the formation of blood clots in the blood vessels and their migration elsewhere in the body, respectively. In [RA 68], O'Reilly provide set of plasma warfarin concentrations and Prothrombin Complex Response in thirty normal subjects after a single loading dose. A single large loading dose of warfarin sodium, 1.5 mg/kg of body weight, was administered orally to all 32 subjects. Measurements were made each 12 or 24h. The dataset can be found in Monolix and simulx R package and is plotted in Figure 1.

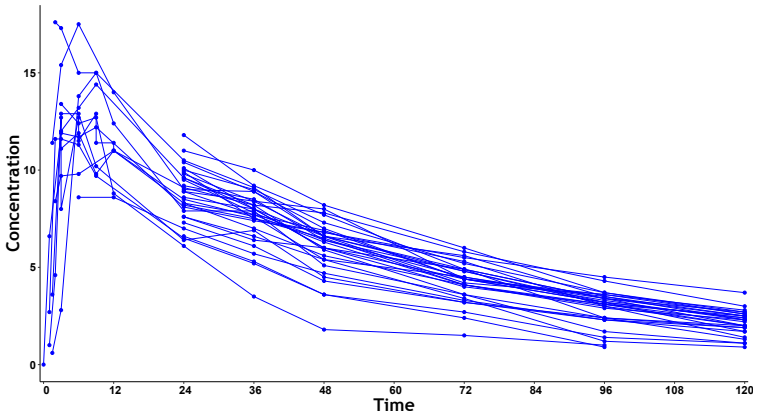

Figure 1: Warfarin concentration over time for 32 subjects

5.1.2 The model

This section develops the application of the new method on a Pharmacokinetics (PK) example. We present here a one compartment PK model for oral administration with first order absorption and linear elimination used for fitting the warfarin data. Beforehand, the standard approach is to approximate the

body as a simple compartment models. In this example we will focus on a one-compartment model for warfarin following oral dose D at time $t = 0$ leading to description of concentration y_{ij} at time $t_{ij} \geq 0$ (i varies from 1 to N and denote the individual of the population):

$$y_{ij} = f(t_{ij}, \psi_i) + \epsilon_{ij} \tag{32}$$

With :

$$f(t, \psi) = \frac{Dka}{V(ka - k)}(e^{-kat} - e^{-kt}) \tag{33}$$

Where ka is he absorption rate constant, k is the elimination rate constant, V is the volume of distribution and D is the dose administered.
In our notation, the complete model is $\mathtt{p}(y_i, \psi_i, \theta)$ where $\psi_i = (ka_i, V_i, k_i)$ is the vector of individual parameters. We apply a log transformation to each of the three variables. Then, $u(\psi_i) = (\log(ka_i), \log(V_i), \log(k_i))$ with:

$$\begin{aligned} \log(ka_i) &\sim \mathcal{N}(\log(ka_{pop}), \omega^2_{ka}) \\ \log(V_i) &\sim \mathcal{N}(\log(V_{pop}), \omega^2_V) \\ \log(k_i) &\sim \mathcal{N}(\log(k_{pop}), \omega^2_k) \end{aligned} \tag{34}$$

5.1.3 Population parameters estimation on the warfarin dataset

First of all, we run both algorithms on a the warfarin dataset described above. We run the algorithms three times starting from those following different initial values for fixed and random effects:

	ka_{pop}	V_{pop}	k_{pop}	ω_{k_a}	ω_V	ω_k
run 1	1	5	2	1	1	1
run 2	3	12	5	0.7	0.7	0.7
run 3	6	3	7	0.5	0.5	0.5

We run the algorithm during 200 iterations setting the stepsize of the stochastic approximation to 1 during $K1 = 100$ and a decreasing step size during $K2 = 100$. The proposal described by (19) is used for the first 10 iterations, switching to the combination of three random walk proposals for the rest of the run as described in section 3.1.
Figure 2 shows convergence of the fixed effect V_{pop} and the random effect ω_V for both the reference algorithm, using RWM in blue, and the accelerated algorithm in red. Each of the three runs starting from different initial values, we remark that using the proposal centered in the MAP and using the covariance of the structure presents the same convergence behaviour whereas the reference algorithm follows different paths slowing the convergence depending on the initial values. Of course, both algorithms converge to the same MLE.

5.1.4 MCMC Convergence Diagnostic

In this section, we will compare the convergence properties of the MCMC procedures used in the SAEM algorithm and in its fast version. We run both chains,

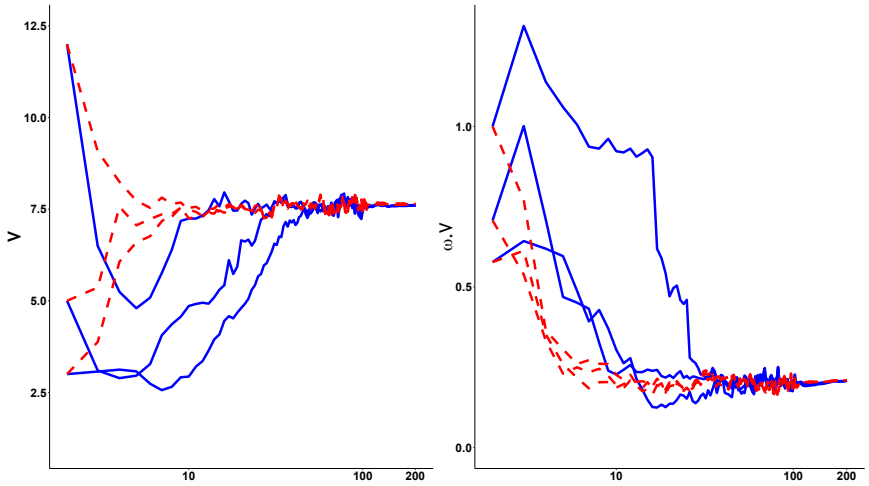

Figure 2: SAEM parameter estimates for the reference algorithm (blue) and the new method (red) initialized at three different values on Warfarin Dataset

under the obtained model parameter estimate, for 10000 iterations. Each iteration corresponds to 6 transition kernels. The original MH algorithm executes 2 transitions of each of the three kernels described in section 3.1 and the fast MH executes 6 transitions of the kernel given by (19), per iteration. For the sake of simplicity, we sample the Gaussian component η_i of the transformed individual parameter described in (6) for a given individual i. Given an MCMC batch $\{\eta_i^t\}_{t=0}^T$, at iteration T, we define the quantile q as:

$$\mathbb{P}(\eta_i^t < \alpha^T) = q \tag{35}$$

for any sample $\eta_i^t \in \{\eta_i^t\}_{t=0}^T$. We focus on three quantiles $q \in \{0.05, 0.5, 0.95\}$ and plot the empirical value α^T in figure 3.

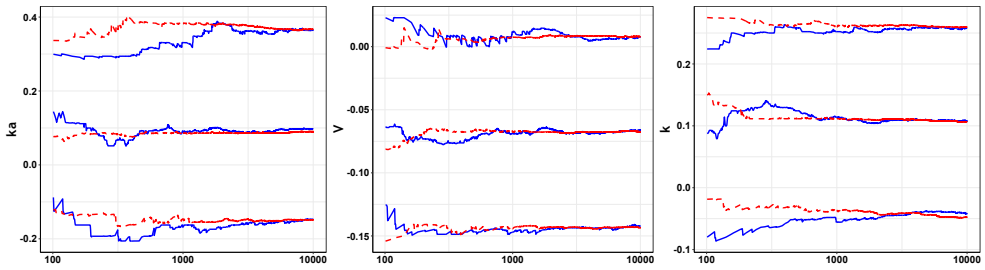

Figure 3: Quantiles convergence plots over the iterations for a single individual. RWM-MH in blue and fast MH in red.

5.1.5 Monte Carlo study

After having run the accelerated version of the algorithm on a real dataset and to justify the effectiveness of our technique, we present in this section a Monte

Carlo study that consists in generating several simulated datasets using the same generating values. Since each simulated dataset was generated with the same values, each run will converge to the same ML estimate. In practice we consider the ML estimate being the value of the parameter after K iterations. Those simulated datasets were generated using the following values: $(ka_{pop} = 1, V_{pop} = 8, k_{pop} = 0.1, \omega_{ka} = 0.5, \omega_V = 0.2, \omega_k = 0.3)$. We simulate 10 observations for each of the $N = 50$ individuals.

We perform a single run, using the prior values for K1 and K2, of the two methods on the $M = 50$ simulated datasets starting with initial values $(ka_{pop} = 1, V_{pop} = 10, k_{pop} = 1, \omega_{ka} = 1, \omega_V = 1, \omega_k = 1)$ and obtain a sequence $(\theta_K^{(m)}, 1 \leq m \leq M)$ of ML estimates. In order to compare the speed of convergence of the parameter estimate to the target value, we compute the error between the parameter $\theta_k^{(m)}$ at iteration k and its target value $\theta_K^{(m)}$ for each dataset m. The average error we will plot to compare both speed of convergence is defined as follows:

$$E_k = \frac{1}{M} \sum_{m=1}^{M} \left[\theta_k^{(m)} - \hat{\theta}^{(m)} \right]^2 \tag{36}$$

Figure 4 shows the evolution of this error at each iteration for both algorithms. We notice a considerable acceleration in the convergence to the MLE using our new method. The error reaches zero approximately twenty times faster than the original algorithm and decreases monotonically whereas the original algorithm parameter estimate appears to deviate from the MLE during the first iterations which yields to an increasing error during the first iterations as shown on figure 4.

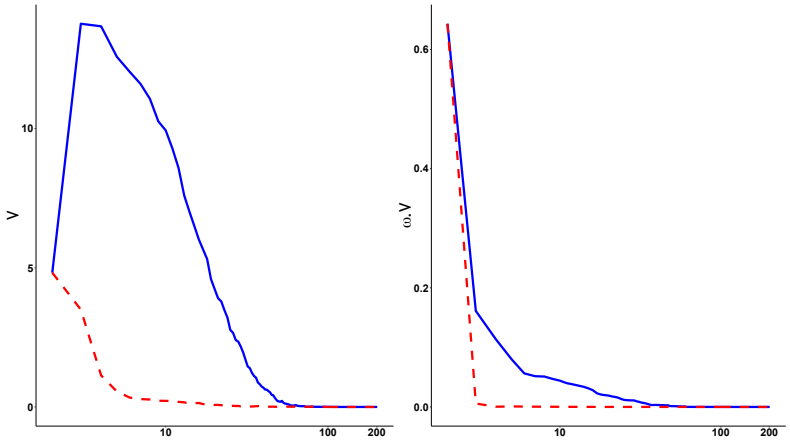

Figure 4: Averaged error E_k for V_{pop} and ω_V

5.2 Time to Event Data Model

5.2.1 The design and the model

Following 2.2.2, for any given hazard function h , the survival function S that represents the survival since the previous event at t_{j-1} , given here in terms of the cumulative hazard from t_{j-1} to t_j is defined by (11). We consider a repeated events design with a right-censoring time denoted $t_c = 20$.
The following Weibull model is used in this experiment:

$$h(t, \psi) = \frac{\beta}{\lambda} \left(\frac{t}{\lambda} \right)^{\beta - 1} \tag{37}$$

Where $\psi = (\lambda, \beta)$.
We apply a log transformation on the vector of individual parameters such as:

$$\begin{aligned} \log(\lambda_i) &\sim \mathcal{N}(\log(\lambda_{pop}), \omega_\lambda^2) \\ \log(\beta_i) &\sim \mathcal{N}(\log(\beta_{pop}), \omega_\beta^2) \end{aligned} \tag{38}$$

5.2.2 Population parameters estimation on a synthetic dataset

We generate a simulated dataset using the following generating values: ($\lambda_{pop} = 2, \omega_\lambda = 0.3, \beta_{pop} = 2, \omega_\beta = 0.3$) for $N = 50$ individuals. Three runs are plotted in Figure 5 starting from the following initial values for fixed and random effects:

	λ_{pop}	β_{pop}	ω_λ	ω_β
run 1	1	5	2	2
run 2	2	1	1	1
run 3	6	4	0.6	0.6

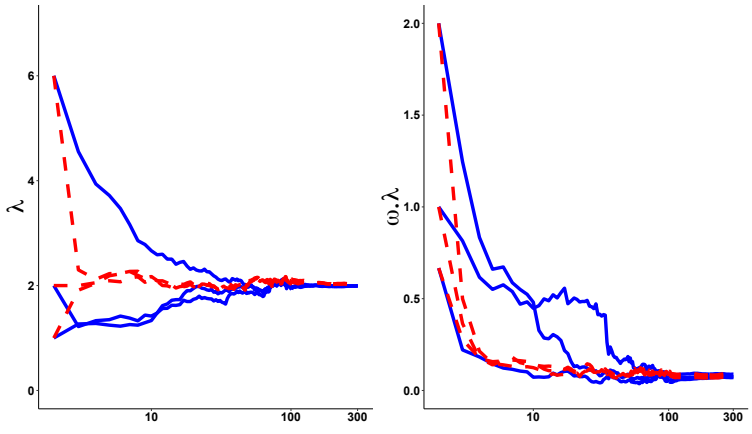

Figure 5: SAEM parameter estimates for the reference algorithm (blue) and the new method (red) initialized at three different values

We run the algorithm during 300 iterations setting the stepsize of the stochastic approximation to 1 during $K1 = 100$ and a decreasing step size during $K2 =$

200. The proposal built previously is used for the first 10 iterations, switching to the combination of three random walk proposals for the rest of the run as described in section 3.1.
Figure 5 shows same behaviour as in the continuous case. Indeed, regardless the initial values, the convergence behavior is similar and is in all three cases better than the algorithm using the RWM MCMC procedure.

5.2.3 MCMC Convergence Diagnostic

Similarly to the previous experiment, we compare the convergence properties of the MCMC procedures used in the SAEM algorithm and in its fast version for non continuous data model. We run both chains, under the model parameter estimate obtained in the previous section, for 5000 iterations. Each iteration corresponds to 6 transition kernels. The original MH algorithm executes 2 transitions of each of the three kernels described in section 3.1 and the fast MH executes 6 transitions of the kernel given by (26), per iteration. We sample the Gaussian component η_i of the transformed individual parameter described in (6) for a given individual i. Convergence of three quantiles $q \in \{0.05, 0.5, 0.95\}$ are plotted in figure 6.

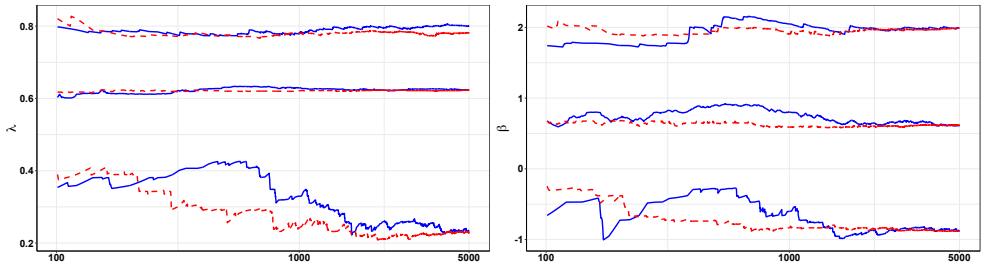

Figure 6: Quantiles convergence plots over the iterations for a single individual. RWM-MH in blue and fast MH in red.

5.2.4 Monte Carlo study

Likewise, we generate $M = 50$ synthetic datasets using the same generating values and run both algorithms on each dataset starting with the initial values $(\lambda_{pop} = 4, \omega_\lambda = 1.4, \beta_{pop} = 1, \omega_\beta = 1.4)$ for $N = 50$ individuals. The plot of the average average error defined in the previous section at each iteration can be found in Figure 7. Considerable improvement in terms of speed of convergence can be observed on this figure. The error reaches zero approximately ten times faster than the original algorithm and decreases monotonically whereas the original algorithm error appears to deviate from the MLE at some iteration and most importantly stays around a non null value for short period of time.

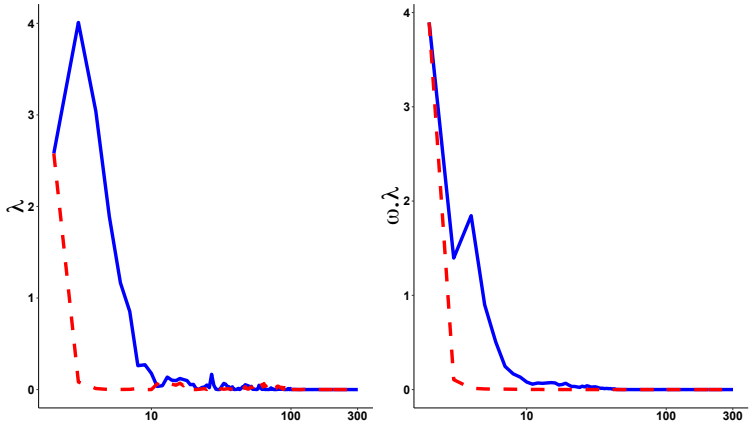

Figure 7: Averaged error E_k for λ_{pop} and ω_λ

6 Discussion

This article introduced the construction of new proposals for an MCMC procedure used to estimate Maximum Likelihood estimate in both continuous or non continuous mixed effects models. The method is based on the approximation of the posterior distribution of the individual parameters given the observed data. In both cases, the proposal is a Gaussian centered in the MAP and of covariance taking into consideration the covariance structure of the problem.
Since in the continuous case the approximation depends on a Taylor expansion of the structural model around a particular point, the efficiency of the resulting independent proposal could depend on how this Taylor expansion approximates well the function on the set of all parameters.
Moreover, the calculation of the MAP at each iteration can be costly in certain case, when the target distribution is strongly non-convex. In this case, an incremental version of the MCMC procedure, detailed in [Mai17], could be coupled to the new technique proposed in this article in order to handle massive data.

Mathematical Justifications

A Continuous Data Models

A.1 Reminder on the posterior distribution in Linear Gaussian Model

Consider any linear model:

$$y = X\beta + A\psi + e \tag{39}$$

with $\psi \sim \mathcal{N}(0, \Omega)$ and $e \sim \mathcal{N}(0, \Sigma)$
The conditional distribution $\psi|y$ is a Gaussian distribution of mean μ and covariance Γ defined by:

$$\begin{aligned} \Gamma &= (A^\top \Sigma^{-1} A + \Omega^{-1})^{-1} \\ \mu &= \Gamma A^\top \Sigma^{-1}(y - X\beta) \end{aligned} \tag{40}$$

A.2 Equality of the mode of the posterior distribution under the linear (19) and the non linear model (8)

Under the non linear model, we know that:

$$\begin{aligned} \hat{\psi}_i &= \arg\max_{\psi_i} \log \mathtt{p}(\psi_i | y_i, \theta) \\ &= \arg\max_{\psi_i} \log \mathtt{p}(y_i | \psi_i, \theta) + \log \mathtt{p}(\psi_i, \theta) \\ &= \arg\min_{\psi_i} \left(\frac{(y - f(\psi_i))^2}{\sigma^2} + (\psi_i - \psi_{pop}).^\top \Omega^{-1} (\psi_i - \psi_{pop}) \right) \end{aligned} \tag{41}$$

And therefore $\hat{\psi}_i$ satisfies:

$$-\frac{\nabla_{\psi_i} f(\hat{\psi}_i)}{\sigma^2} \left[y_i - f(\hat{\psi}_i) \right] + \Omega^{-1} \hat{\psi}_i = 0 \tag{42}$$

Using the formula of the posterior mean given by the reminder (40), we compute

the posterior mean under the linear model and get:

$$\begin{aligned}
\mathbb{E}[\psi_i|y_i] &= \left(\frac{\nabla_{\psi_i} f(\hat{\psi}_i).^\top \nabla_{\psi_i} f(\hat{\psi}_i)}{\sigma^2} + \Omega^{-1}\right)^{-1} . \frac{\nabla_{\psi_i} f(\hat{\psi}_i)}{\sigma^2} . \left(y_i - f(\hat{\psi}_i) + \nabla_{\psi_i} f(\hat{\psi}_i).^\top \hat{\psi}_i\right) \\
&= \left(\frac{\nabla_{\psi_i} f(\hat{\psi}_i).^\top \nabla_{\psi_i} f(\hat{\psi}_i)}{\sigma^2} + \Omega^{-1}\right)^{-1} . \left(\underbrace{\frac{\nabla_{\psi_i} f(\hat{\psi}_i)}{\sigma^2}\left[y_i - f(\hat{\psi}_i)\right]}_{=\Omega^{-1}\hat{\psi}_i \text{ see (42)}} + \frac{\nabla_{\psi_i} f(\hat{\psi}_i)}{\sigma^2} \nabla_{\psi_i} f(\hat{\psi}_i).^\top \hat{\psi}_i\right) \\
&= \Gamma_i . \Gamma_i^{-1} . \hat{\psi}_i \\
&= \hat{\psi}_i
\end{aligned} \tag{43}$$

Finally, using (40) we have an expression for the conditional covariance Γ_i:

$$\Gamma_i = \left[\frac{\nabla_{\psi_i} f(\hat{\psi}_i).^\top \nabla_{\psi_i} f(\hat{\psi}_i)}{\sigma^2} + \Omega^{-1}\right]^{-1} \tag{44}$$

Which yields to the Gaussian proposal distribution with parameters:

$$\begin{aligned}
\mu_i &= \hat{\psi}_i \\
\Gamma_i &= \left[\frac{\nabla_{\psi_i} f(\hat{\psi}_i).^\top \nabla_{\psi_i} f(\hat{\psi}_i)}{\sigma^2} + \Omega^{-1}\right]^{-1}
\end{aligned} \tag{45}$$

B Non Continuous Data Models

B.1 Reminder on Laplace Approximation

Laplace approximation, see [DL14], consists in approximating an integral of the form:

$$I := \int e^{v(x)} dx \tag{46}$$

Where v is at least three times differentiable.

Based on a second order Taylor expansion of the function v around a point x_0 we get:

$$v(x) \approx v(x_0) + \nabla v(x_0)(x - x_0) + \frac{1}{2}(x - x0)\nabla^2 v(x_0)(x - x0) \tag{47}$$

Which results in an approximation of the integral I (consider a multivariate Gaussian which integral sums to 1):

$$I \approx e^{v(x_0)} \sqrt{\frac{(2\pi)^p}{|-\nabla^2 v(x_0)|}} e^{-1/2 \nabla v(x_0) \nabla^2 v(x_0)^{-1} \nabla v(x_0)} \tag{48}$$

B.2 Application to the population approach

In our context we can easily write the incomplete likelihood for individual i, that we are trying to approximate, for a given parameter estimate $\theta \in \Theta$, as:

$$\begin{aligned} \mathtt{p}(y_i, \theta) &= \int \mathtt{p}(y_i, \psi_i, \theta) \mathtt{d}\psi_i \\ &= \int e^{\log \mathtt{p}(y_i, \psi_i, \theta)} \mathtt{d}\psi_i \end{aligned} \tag{49}$$

In this case we identify $v(\psi_i) = \log \mathtt{p}(y_i, \psi_i, \theta)$. In the sequel we can lose dependency on the parameter θ since the model parameter stays constant throughout the MCMC. We'll consider it back when dealing with the SAEM algorithm. Also

$$\begin{aligned} \log \mathtt{p}(y_i, \psi_i) &= \log \mathtt{p}(y_i|\psi_i) + \log \mathtt{p}(\psi_i) \\ &= \log l(\psi_i) + \log \mathtt{p}(\psi_i) \end{aligned} \tag{50}$$

The following approximations are based on a Taylor expansion around the Maximum A Posteriori (MAP) (also known as Empirical Bayes Estimate (EBE)). As a result $\nabla \log \mathtt{p}(y_i, \hat{\psi}_i) = 0$ with $\hat{\psi}_i = \arg\max_{\psi_i} \mathtt{p}(y_i, \psi_i)$.

Using last derivation of the approximation of the integral I we can explicit an approximation of our incomplete log likelihood:

$$\begin{aligned} -2 \log \mathtt{p}(y_i) &\approx -p \log 2\pi - 2 \log \mathtt{p}(y_i, \hat{\psi}_i) + \log |-\nabla^2 \log \mathtt{p}(y_i, \hat{\psi}_i)| \\ &\approx -2 \log \mathtt{p}(y_i|\hat{\psi}_i) - 2 \log \mathtt{p}(\hat{\psi}_i) - p \log 2\pi + \log |-\nabla^2 \log \mathtt{p}(y_i, \hat{\psi}_i)| \end{aligned} \tag{51}$$

With a combination of this last approximation of the incomplete log-likelihood and the following Bayes rule:

$$\log \mathtt{p}(y_i) = \log \mathtt{p}(y_i|\hat{\psi}_i) + \log \mathtt{p}(\hat{\psi}_i) - \log \mathtt{p}(\hat{\psi}_i|y_i) \tag{52}$$

We can approximate $-\log \mathtt{p}(\hat{\psi}_i|y_i)$ by $-p \log 2\pi + \log |-\nabla^2 \log \mathtt{p}(y_i, \hat{\psi}_i)|$.
Notice that $-p \log 2\pi + \log |-\nabla^2 \log \mathtt{p}(y_i, \hat{\psi}_i)|$ is the value of a multivariate Gaussian distribution centered in $\hat{\psi}_i$ and of covariance $-\nabla^2 \log \mathtt{p}(y_i, \hat{\psi}_i)^{-1}$ evaluated in $\psi_i = \hat{\psi}_i$.
The Laplacian method consists in, through an approximation of the incomplete log likelihood, approaching the posterior distribution $\mathtt{p}(\psi_i|y_i)$ as a multivariate Gaussian distribution centered in $\hat{\psi}_i$ and of covariance $-\nabla^2 \log \mathtt{p}(y_i, \hat{\psi}_i)^{-1}$.
Where:

$$\nabla^2 \log \mathtt{p}(y_i, \hat{\psi}_i) = \nabla^2 \log l(\hat{\psi}_i) + \nabla^2 \log \mathtt{p}(\hat{\psi}_i) \tag{53}$$

With

$$\begin{aligned} \nabla^2 \log l(\hat{\psi}_i) &= \nabla^2 \log \mathtt{p}(y_i|\hat{\psi}_i) \\ \nabla^2 \log \mathtt{p}(\hat{\psi}_i) &= -\Omega^{-1} \end{aligned} \tag{54}$$

since the prior on the latent variable ψ_i is a Gaussian distribution of covariance Ω.
Consequently, the individual parameters, given the observed data, are approximated as samples from the following distribution:

$$\psi_i|y_i \sim \mathcal{N}\left(\hat{\psi}_i, \left[-\nabla^2 \log l(\hat{\psi}_i) + \Omega^{-1}\right]^{-1}\right) \tag{55}$$

References

[AG15] D. Lee A. Gelman and Y. Guo. "Stan: A probabilistic programming language for Bayesian inference and Optimization." In: *Journal Of Statistical Software* (2015).

[Agr07] A. Agresti. "An Introduction to Categorical Data Analysis." In: *Wiley Series in Probability and Statistics* 423 (2007).

[And06] P. K. Andersen. "Survival Analysis." In: *Wiley Reference Series in Biostatistics* (2006).

[AP16] E. Moulines A. Durmus and M. Pereyra. "Sampling from convex non continuously differentiable functions, when Moreau meets Langevin." In: (2016).

[BM99] M. Lavielle B. Delyon and E. Moulines. "Convergence of a stochastic approximation version of the EM algorithm." In: *The Annals of Statistics* (1999).

[CL15] K. Bleakley C. Mbogning and M. Lavielle. "Joint modeling of longitudinal and repeated time-to-event data using nonlinear mixed-effects models and the SAEM algorithm." In: *Journal of Statistical Computation and Simulation* (2015).

[DL14] H.S. Migon D. Gamerman and F. Louzada. "Statistical Inference: An Integrated Approach, Second Edition." In: (2014).

[DR77] Laird Dempster and Rubin. "Maximum likelihood from incomplete likelihood data via EM algorithm (with discussion)." In: *J. Roy. Statist. Soc. Ser.* (1977).

[EKu13] S. Allassoniere E.Kuhn. "Convergent Stochastic Expectation Maximization algorithm with efficient sampling in high dimension. Application to deformable template model estimation." In: (2013).

[EKu15] M. Lavielle E.Kuhn. "Coupling a Stochastic Approximation version of EM with an MCMC Procedure." In: *ESAIM: Probability and Statistics* 8 (2015).

[EL17] A. Lavenu E. Comets and M. Lavielle. "Parameter estimation in nonlinear mixed effect models using saemix." In: *Journal of Statistical Software* (2017).

[G O98] J. S. Rosenthal G. O. Roberts. "Optimal scaling of discrete approximations to Langevin diffusions." In: *Journal of the Royal Statistical Society* (1998).

[GG97] A. Gelman G. O. Roberts and W. R. Gilks. "Weak Convergence and Optimal Scaling of Random Walk Metropolis Algorithms.." In: *Ann. Appl. Probab.* (1997).

[GO 96] R.L. Tweedie G.O. Roberts. "Exponential con- vergence of Langevin distributions and their discrete approximations." In: *Bernoulli* (1996).

[HG14] M. Hoffman and A. Gelman. "The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo." In: *Journal Of Machine Learning Research* (2014).

[L A01] F. Mentre L. Aarons M. O. Karlsson. “Role of modelling and simulation in Phase I drug development.” In: *European Journal of Pharmaceutical Sciences* (2001).

[Lav15] M. Lavielle. “Mixed Effects Models for the Population Approach.” In: (2015).

[Lav93] M. Lavielle. “A stochastic algorithm for parametric and non-parametric estimation in the case of incomplete data .” In: *Elsevier Science* (1993).

[M G11] B. Calderhead M. Girolami. “Riemann manifold Langevin and Hamiltonian Monte Carlo methods.” In: *Journal of the Royal Statistical Society* (2011).

[Mai17] F. Maire. “Adaptive Incremental Mixture Markov chain Monte Carlo.” In: (2017).

[MT96] K.L. Mengersen and R.L. Tweedie. “Rates of convergence of the Hastings and Metropolis algorithms.” In: *The Annals of Statistics* (1996).

[N M53] M.N Rosenbluth N. Metropolis A.W. Rosenbluth. “Equations of state calculations by fast computing machine.” In: (1953).

[O S99] R.L. Tweedie O. Stramer. “Langevin-typemodels i: Diffusions with given stationary distributions, and their discretizations.” In: *Methodol. Comput. Appl. Prob.* (1999).

[PL11] P. Jacqmin P.L.S Chan and M. Lavielle. “The use of the SAEM algorithm in MONOLIX software for estimation of population pharmacokinetic-pharmacodynamic-viral dynamics parameters of maraviroc in asymptomatic HIV subjects.” In: *Journal of Pharmacokinetics and Pharmacodynamics* (2011).

[RA 68] PM Aggeler RA. O’reilly. “Studies on coumarin anticoagulant drugs. Initiation of warfarin therapy without a loading dose.” In: (1968).

[RL11] F. Mentré R. M. Savic and M. Lavielle. “Implementation and Evaluation of the SAEM Algorithm for Longitudinal Ordered Categorical Data with an Illustration in Pharmacokinetics–Pharmacodynamics.” In: *The AAPS Journal* (2011).

[RT96] G.O. Roberts and R.L. Tweedie. “Geometric convergence and central limit theorems for multidimensional Hastings and Metropolis algorithms.” In: *Biometrika* (1996).

[SB09] A. Boekmann S. Beal L.B. Sheiner and R.J Bauer. “NONMEM’s User’s Guide.” In: *ICON Development Solutions* (2009).

[SD87] J. Pendleton S. Duane A.D. Kennedy Brian and R. Duncan. “Hybrid Monte Carlo.” In: *Physics Letters B* (1987).

[T M12] G.O. Roberts T. Marshall. “An adaptive approach to langevin MCMC.” In: *Statistics and Computing* (2012).

[Wol17] R. Wolfinger. “Laplace’s Approximation for Nonlinear Mixed Models.” In: *Biometrika* 80 (2017).

[WT90] G. Wei and M. Tanner. “A Monte-Carlo implementation of the EM algorithm and the poor man’s data augmentation algorithms.” In: *J. Amer. Statist. Assoc.* (1990).

[WU83] C. WU. " On the convergence properties of the EM algorithm." In: *The Annals of Statistics* 11 (1983).

[Y A05] J. Rosenthal Y. Atchade. "On adaptive Markov chain Monte Carlo algorithms ." In: *Bernoulli* (2005).

[Y M15] E.B. Fox Y. Ma T. Chen. "A Complete Recipe for Stochastic Gradient MCMC." In: *NIPS'15 Proceedings of the 28th International Conference on Neural Information Processing Systems* (2015).

[Y07] Wang Y. "Derivation of various NONMEM estimation methods.." In: *Journal of Pharmacokinetics and Pharmacodynamics* (2007).

[Yao00] J.-F. Yao. "On recursive estimation in incomplete data models." In: *Statistics* 34 (2000).