Fast MCMC sampling for nonlinear mixed effects models

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

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

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

Abstract

The ability to generate samples of the random effects from their conditional distributions is fundamental for inference in mixed effects models. Random walk Metropolis is widely used to conduct such sampling, but such a method can converge slowly for high dimension problems, or when the joint structure of the distributions to sample is complex. We propose a Metropolis–Hastings (MH) algorithm based on a multidimensional Gaussian proposal that takes into account the joint conditional distribution of the random effects and does not require any tuning. Indeed, this distribution is automatically obtained thanks to a Laplace approximation of the original model. We show that such approximation is equivalent to linearizing the model in the case of continuous data. Then, the proposal is as close to the target distribution as the model is close to a linear model. Numerical experiments based on simulated and real data highlight the very good performances of the proposed methods for continuous and time-to-event data models. In particular, we show that the proposed MH algorithm can be efficiently combined with a stochastic approximation version of EM for maximum likelihood estimation in nonlinear mixed effects models.

Keywords: mcmc, mle, mixed effects, laplace approximation

1. Introduction

Mixed effects models are reference models when considering the inter-individual variability that can exist within the same population (see (Lavielle, 2014) and the references therein). Consider a study with N individuals from a same population. For each individual i in the study, we have a series of observations y_i whose probability distribution depends on a vector of individual parameters ψ_i which is a realization of a random variable. Then, when Gaussian priors are used, inference on the individual parameter ψ_i means estimating its conditional distribution given the observed data y_i.

*Corresponding author

Email addresses: belhal.karimi@polytechnique.edu (Belhal Karimi), marc.lavielle@polytechnique.edu (Marc Lavielle)

When the model is a linear (mixed effects) Gaussian model (Verbeke, 1997), then this conditional distribution is a normal distribution that can be explicitly characterized. For more complex priors or models, Monte Carlo methods must be used to approximate this conditional distribution. Most often, direct sampling from this conditional distribution is impossible and it is necessary to have resort to a Markov chain Monte Carlo (MCMC) method for obtaining random samples from this distribution.

Note that generating random samples of the conditional distributions $\mathtt{p}(\psi_i|y_i)$ is extremely useful for several tasks in the complex process of model building, when we want to avoid approximation of the model, such as linearization or Laplace method. Such tasks include the estimation of the population parameters θ of the model, either by a maximum likelihood approach, i.e. by maximizing the observed pdf $\mathtt{p}(y_1, \dots y_N; \theta)$ using the SAEM algorithm combined with a MCMC procedure (Kuhn and Lavielle, 2004), or by a Bayesian method, i.e. by estimating $\mathtt{p}(\theta|y_1, \dots y_N)$ using a Gibbs sampler.

In (Lavielle and Ribba, 2016), the authors argue that methods for model assessment and model validation, whether graphical or based on statistical tests, must use samples of the conditional distributions $\mathtt{p}(\psi_i|y_i)$ to avoid estimation bias.

Designing a fast mixing sampler is of utmost importance. The most common MCMC method for nonlinear mixed effects models is the *random walk Metropolis* algorithm (Robert and Casella, 2010; Roberts et al., 1997; Lavielle, 2014). This method is implemented in software tools such as Monolix, NONMEM, the saemix R package and the nlmefitsa Matlab function.

Despite its simplicity, it has been successfully used in many classical examples of pharmacometry, when the number of random effects is not too large. Nevertheless, it can show its limitations when the parameter space to explore becomes large. In particular, maintaining an optimal acceptance rate (advocated in (Roberts et al., 1997)) most often implies very small moves and therefore a very large number of iterations. Therefore, if we want to adapt the MCMC to high-dimensional probability distributions of practical interest, we need to make better use of the geometry of the target distribution in order to explore the space faster.

The Metropolis-adjusted Langevin algorithm (MALA) uses evaluations of the gradient of the target density for proposing new states which are accepted or rejected using the Metropolis–Hastings algorithm (Roberts and Tweedie, 1996a; Stramer and Tweedie, 1999). Several variations have been proposed for improving the behavior of MALA by incorporating more information about the properties of the target distribution in the proposal, see for instance (Durmus et al., 2017) and (Girolami and Calderhead, 2011).

Hamiltonian Monte Carlo (HMC) is another MCMC algorithm that exploits information about the geometry of the target distribution for exploring efficiently the space by selecting transitions that can follow contours of high probability mass (Betancourt, 2017). The No-U-Turn Sampler (NUTS) is an extension to HMC that allows an automatic and optimal selection of some of the settings required by the algorithm, (Hoffman and Gelman, 2014; Neal et al., 2011).

Nevertheless, these methods may be difficult to use in practice, and are computation-

ally involved, in particular when the structural model is a complex ODE based model.

The algorithm we propose is a Metropolis-Hastings algorithm, but for which the choice of proposal is optimized by defining a proposal distribution, which approximates the target distribution. In the case of continuous data, linearization of the model leads, by definition, to a Gaussian linear model for which the conditional distribution of the individual parameter ψ_i given the data y_i is a multidimensional normal distribution that can be calculated. It is therefore this normal distribution that will be used for proposing a complete vector of individual parameters at each iteration. For noncontinuous data model (i.e. categorical, count or time-to-event data models), the Laplace approximation of the incomplete pdf $\mathtt{p}(y_i)$ leads to a Gaussian approximation of the conditional distribution $\mathtt{p}(\psi_i|y_i)$.

Mixed effects models for continuous and noncontinuous data are presented Section 2. The standard MH for nonlinear mixed effects models and the new proposed method are described Section 3 and Section 4, respectively. How to combine this new method with the SAEM algorithm for estimating the population parameters of the model is described Section 5. Numerical examples illustrate in Section 6 the very good practical performances of the proposed method, both on a continuous pharmacokinetics (PK) model and a time-to-event example. A Monte Carlo study confirms that this new SAEM algorithm converges extremely fast to the maximum likelihood estimate.

2. Mixed Effect Models

2.1. Population approach and hierarchical models

We will adopt a population approach in the sequel, where we consider N individuals and n_i 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 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 ψ_i that takes its values in a subset of $\mathbb{R}^p$.

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

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

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 ψ_i as a typical value ψ_{pop}, and a vector individual random effects η_i:

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

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 (2) are also possible. We can assume for instance that transformed individual parameters are normally distributed:

$$u(\psi_i) = u(\psi_{\text{pop}}) + \eta_i \ , \tag{3}$$

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

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

$$u(\psi_i) = u(\psi_{\text{pop}}) + C_i\beta + \eta_i \ , \tag{4}$$

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

2.2. Continuous data models

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

$$y_{ij} = f(t_{ij}, \psi_i) + \varepsilon_{ij} \ , \tag{5}$$

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 assumed to be a twice differentiable function of ψ_i.

We start by assuming that the residual errors are independent and normally distributed with zero-mean and a constant variance σ^2. Let $t_i = (t_{ij}, 1 \leq n_i)$ be the vector of observation 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 $f_i(\psi_i) = (f(t_{i,1}, \psi_i), \ldots, f(t_{i,n_i}, \psi_i))$.

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 ψ_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:

$$\varepsilon_{ij} \sim \mathcal{N}(0, g(t_{ij}, \psi_i)^2) \ . \tag{6}$$

Such extension includes proportional error models ($g = bf$) and combined error models ($g = a + bf$) (Lavielle, 2014) but the proposed method remains the same whatever the residual error model is.

2.3. Non continuous data models

Non continuous data models include categorical data models (Savic et al., 2011; Agresti, 1990), time to event data model (Mbogning et al., 2015; Andersen, 2006), or count data models (Savic et al., 2011) 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 depend on the vector of individual parameters ψ_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). 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:

$$\begin{aligned} S(t) &\triangleq \mathbb{P}(T > t) \\ &= e^{-\int_0^t h(u)\mathrm{d}u} , \end{aligned} \tag{7}$$

where h is called the hazard function.

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

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

$$y_i = \begin{cases} T_i & \text{if } T_i < \tau_c \\ \text{"}T_i > \tau_c\text{"} & \text{otherwise} , \end{cases} \tag{8}$$

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

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

$$\mathbb{P}(T_{ij} > t|T_{i,j-1} = t_{i,j-1}) = e^{-\int_{t_{i,j-1}}^t h(u,\psi_i)\mathrm{d}u} . \tag{9}$$

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 (Lavielle, 2014) for more details) that the conditional pdf of $y_i = (y_{ij}, 1 \leq n_i)$ writes

$$\mathtt{p}(y_i|\psi_i) = \exp\left\{-\int_0^{\tau_c} h(u, \psi_i)\mathrm{d}u\right\} \prod_{j=1}^{n_i-1} h(t_{ij}, \psi_i) . \tag{10}$$

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 ψ_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.

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 to compute the maximum likelihood (ML) estimate of θ

$$\hat{\theta}_{\mathrm{ML}} = \arg\max_{\theta\in\Theta} \mathcal{L}(\theta, y) \;, \tag{11}$$

where $\mathcal{L}(\theta, y) = \log \mathtt{p}(y;\theta)$. 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

$$I(\hat{\theta}_{\mathrm{ML}}, y) = -\nabla^2_\theta \mathcal{L}(\hat{\theta}_{\mathrm{ML}}, y) \tag{12}$$

can be derived thanks to the Louis formula (Louis, 1982). 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 design a sensible method to sample from this distribution.

3.2. The Metropolis Hastings Algorithm

Metropolis-Hasting (MH) algorithm is a powerful MCMC procedure widely used for sampling from a posterior distribution for which direct sampling is difficult. Since the following remains true for any model parameter θ, we choose not to mention it to avoid cumbersome notations. In our case, iteration k of the MH algorithm consists in

- Drawing a candidate ψ_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:

$$\alpha(\psi_i^c, \psi_i^{(k-1)}) = \frac{\mathtt{p}(\psi_i^c|y_i)}{\mathtt{p}(\psi_i^{(k-1)}|y_i)} \frac{q(\psi_i^{(k-1)}|\psi_i^c)}{q(\psi_i^c|\psi_i^{(k-1)})} \,. \tag{13}$$

- 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 $\mathtt{p}(\psi_i|y_i)$ (Mengersen and Tweedie, 1996; Roberts et al., 1997; Roberts and Tweedie, 1996b).

The choice of the proposal transition density 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, regarding independent sampler for instance, 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 (de Freitas et al., 2001; Wainwright and Jordan, 2008; Salimans et al., 2015; Zhang et al., 2018).

Current implementations of the SAEM algorithm in Monolix (Chan et al., 2011), saemix (R package) (Comets et al., 2017), nlmefitsa (Matlab) and NONMEM (Beal and Sheiner, 1980) mainly use the same combination of proposals. The first proposal is an independent MH algorithm which consists in sampling the candidate state directly from the marginal distribution of the individual parameter ψ_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 (Metropolis et al., 1953) that updates different components of ψ_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 (Atchadé and Rosenthal, 2005; Lavielle, 2014).

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) (Roberts and Tweedie, 1996a; Stramer and Tweedie, 1999) and the Hamiltonian Monte Carlo (HMC) which implementation can be found for instance as the "No U-Turns Sampler" in STAN (Hoffman and Gelman, 2014; Carpenter et al., 2017).

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

$$\psi_i^c \sim \mathcal{N}(\psi_i^{(k)} - \gamma_k \nabla \log \pi(\psi_i^{(k)}), 2\gamma_k) \ , \tag{14}$$

where $(\gamma_k)_{k>0}$ is a sequence of positive integers. It is a particular case of the RWM with a drift term (Ma et al., 2015) 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 (Marshall and Roberts, 2012; Atchadé and Rosenthal, 2005).

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 (Allassonniere and Kuhn, 2013) 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 (Durmus and Moulines, 2017).

On the other hand, the HMC algorithm, introduced in (Duane et al., 1987), 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 (Girolami and Calderhead, 2011) 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 ($M_i(\psi_i) = \nabla^2 \pi(\psi_i | y_i)$).

Most recently, (Titsias and Papaspiliopoulos, 2018) proposed a large class of MCMC samplers based on the Taylor expansion around the current state of the chain of $\mathtt{p}(y|\psi)$, leaving $\mathtt{p}(\psi)$ unchanged. This leads to a multivariate Gaussian proposal distribution. MALA samplers are shown to be a special case of this class of procedures.

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 ψ_i:

$$y_i = A_i \psi_i + \varepsilon_i \ , \tag{15}$$

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

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

Then, the conditional distribution of ψ_i is a normal distribution:

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

where

$$\begin{aligned} \Gamma_i &= \left(\frac{A_i' A_i}{\sigma^2} + \Omega^{-1} \right)^{-1} \ , \\ \mu_i &= \Gamma_i \left(\frac{A_i' y_i}{\sigma^2} + \Omega^{-1} m_i \right) \ . \end{aligned} \tag{16}$$

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

In the linear case, sampling ψ_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 (5) for continuous data and assume that the ψ_i's are normally distributed with mean m_i and variance-covariance Ω. For a given parameter value θ, the MAP estimate, for individual i, is the value of ψ_i that maximizes the conditional distribution $\mathtt{p}(\psi_i|y_i,\theta)$:

$$\begin{aligned}\hat{\psi}_i &= \arg\max_{\psi_i} \mathtt{p}(\psi_i|y_i,\theta) \\ &= \arg\max_{\psi_i} \mathtt{p}(y_i|\psi_i,\theta)\mathtt{p}(\psi_i,\theta) \\ &= \arg\min_{\psi_i} \left(\frac{1}{\sigma^2}\|y_i - f(t_i,\psi_i)\|^2 + (\psi_i - m_i)'\Omega^{-1}(\psi_i - m_i)\right) .\end{aligned} \tag{17}$$

In the linear model (15), we have seen that $\hat{\psi}_i$ is the conditional mean μ_i of a Gaussian distribution that can be computed in closed-form using (16). 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 $\hat{\psi}_i$:

$$f_i(\psi_i) \approx f_i(\hat{\psi}_i) + \nabla f_i(\hat{\psi}_i)(\psi_i - \hat{\psi}_i) , \tag{18}$$

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

$$z_i = \nabla f_i(\hat{\psi}_i)\psi_i + \varepsilon_i , \tag{19}$$

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

Proposition 1. *Under linear model* (19)*, the conditional distribution of ψ_i is a Gaussian distribution with mean μ_i and variance-covariance Γ_i where*

$$\begin{aligned}\mu_i &= \hat{\psi}_i , \\ \Gamma_i &= \left(\frac{\nabla f_i(\hat{\psi}_i)'\nabla f_i(\hat{\psi}_i)}{\sigma^2} + \Omega^{-1}\right)^{-1} .\end{aligned} \tag{20}$$

We then propose to use this normal distribution as a proposal in our MCMC procedure when the model is described by (5). 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 ψ_i in the nonlinear model is also the mode and the mean of the conditional distribution of ψ_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 ψ_i and on the observation times (t_{ij}), then the variance-covariance matrix of the conditional distribution rewrites

$$\Gamma_i = \left(\nabla f_i(\hat{\psi}_i)'\Sigma(t_i, \hat{\psi}_i)^{-1}\nabla f_i(\hat{\psi}_i) + \Omega^{-1}\right)^{-1} . \tag{21}$$

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

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

where $\hat{\phi}_i = \arg\max_{\phi_i} \mathrm{p}(\phi_i|y_i, \theta)$ 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 ψ_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 (Wolfinger, 1993) or (Wang, 2007).

Proposition 2. *Let (y_i, ψ_i) be a pair of random variables where ψ_i is normally distributed with variance-covariance matrix Ω and where the conditional pdf of y_i is twice differentiable with respect to ψ_i. Let $\hat{\psi}_i$ be the value of ψ_i that maximizes the conditional distribution $\boldsymbol{p}(\psi_i|y_i)$. Then, the conditional distribution of ψ_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 $-\nabla^2 l(\hat{\psi}_i) = -\nabla^2 \log(\mathrm{p}(y_i|\hat{\psi}_i))$ 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

$$-\mathbb{E}\left(\nabla^2 \log l(\hat{\psi}_i)|\psi_i = \hat{\psi}_i\right) = \mathbb{E}\left(\nabla \log l(\hat{\psi}_i)\nabla \log l(\hat{\psi}_i)'|\psi_i = \hat{\psi}_i\right) . \tag{23}$$

Following (23), 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 θ 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 $\mathbb{E}\left[\log \mathtt{p}(y,\psi;\theta)|y,\theta^{k-1}\right]$:

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

where $\{\gamma_k\}_{k>0}$ is a sequence of decreasing stepsizes with $\gamma_1 = 1$.
Maximisation step updating the estimation of θ:

$$\theta^k = \arg\max_{\theta\in\Theta} Q_k(\theta) . \tag{25}$$

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 (Delyon et al., 1999; Kuhn and Lavielle, 2004).

The practical performances of SAEM are closely linked to the settings of SAEM. In particular, the choice of the transition kernel Π plays a key role.

The transition kernel Π is directly defined by the proposal(s) used for the MH algorithm. The new proposal based on a normal approximation of the conditional distribution of ψ_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 θ_{k-1} for all individuals i:
$$\hat{\psi}_i^{(k)} = \arg\max_{\psi_i} \mathtt{p}(\psi_i | y_i, \theta_{k-1}) \ . \tag{26}$$

2. Compute the covariance matrix $\Gamma_i^{(k)}$ such as:
$$\Gamma_i^{(k)} = \begin{cases} \left(\frac{\nabla f_i(\hat{\psi}_i^{(k)}) \nabla f_i(\hat{\psi}_i^{(k)})'}{\sigma^2} + \Omega^{-1} \right)^{-1} & \text{if the data model is continuous ,} \\ \left(\nabla l(\hat{\psi}_i^{(k)}) \nabla l(\hat{\psi}_i^{(k)})' + \Omega^{-1} \right)^{-1} & \text{otherwise .} \end{cases} \tag{27}$$

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 (γ_k). In practice, γ_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 (θ_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 (O'Reilly and Aggeler, 1968). Figure 1 shows the warfarin plasmatic concentration measured at different times for these patients (the single dose was given at time 0 for all the patients).

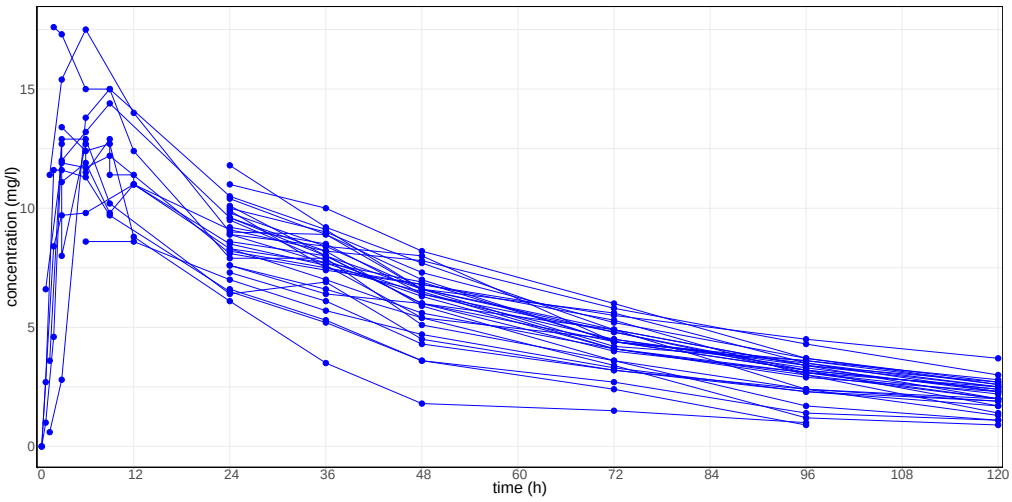

Figure 1: Warfarin concentration (mg/l) over time (h) for 32 subjects

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

$$f(t, ka, V, k) = \frac{D\, ka}{V(ka - k)}(e^{-ka\, t} - e^{-k\, t}) \;, \tag{28}$$

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

$$y_{ij} = f(t_{ij}, \psi_i) + \varepsilon_{ij} \;. \tag{29}$$

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

three PK parameters:

$$\begin{aligned} \log(ka_i) &\sim \mathcal{N}(\log(ka_{\text{pop}}), \omega_{ka}^2) \ , \\ \log(V_i) &\sim \mathcal{N}(\log(V_{\text{pop}}), \omega_V^2) \ , \\ \log(k_i) &\sim \mathcal{N}(\log(k_{\text{pop}}), \omega_k^2) \ . \end{aligned} \tag{30}$$

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 given by Proposition 1 and based on a linearization of the structural model f proposed in (28).

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 θ 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 ψ_i. Here, $\hat{q}_\alpha^{(k)}(\psi_{i,\ell})$ is the empirical quantile of order α of $(\psi_{i,\ell}^{(1)}, \psi_{i,\ell}^{(2)}, \ldots, \psi_{i,\ell}^{(k)})$ and ℓ denotes the component of the individual parameter.

We see Figure 2 that, for all α and all ℓ, the sequences of empirical quantiles $\hat{q}_\alpha^k(\psi_{i,\ell})$ 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.

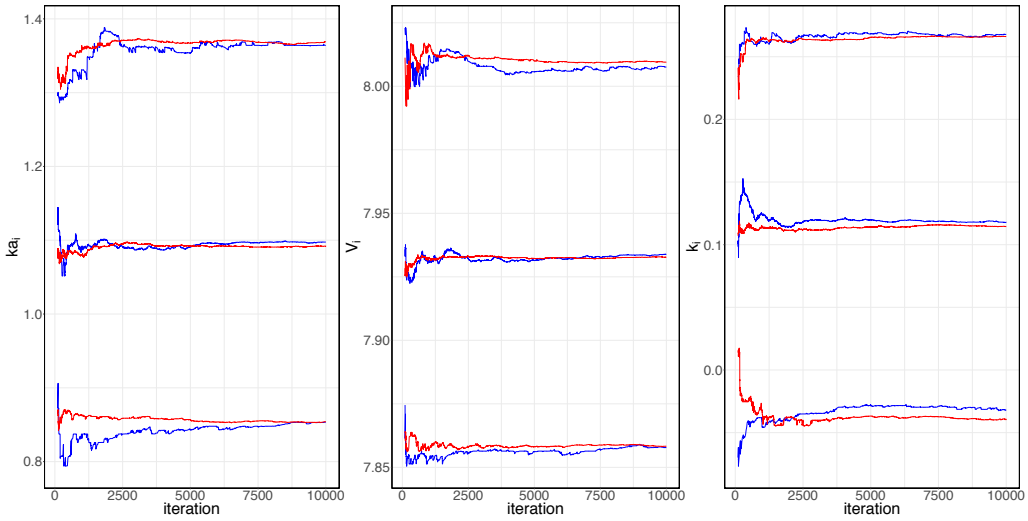

Figure 2: Modelling of the warfarin PK data: convergence of the empirical quantiles of order 0.1, 0.5 and 0.9 of $\mathtt{p}(\psi_i|y_i;\theta)$ for a single individual. The reference MH algorithm is in blue and the new version is in red.

6.1.3. Maximum likelihood estimation of the parameters

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

The stepsize γ_k 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.

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_{pop} and ω_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.

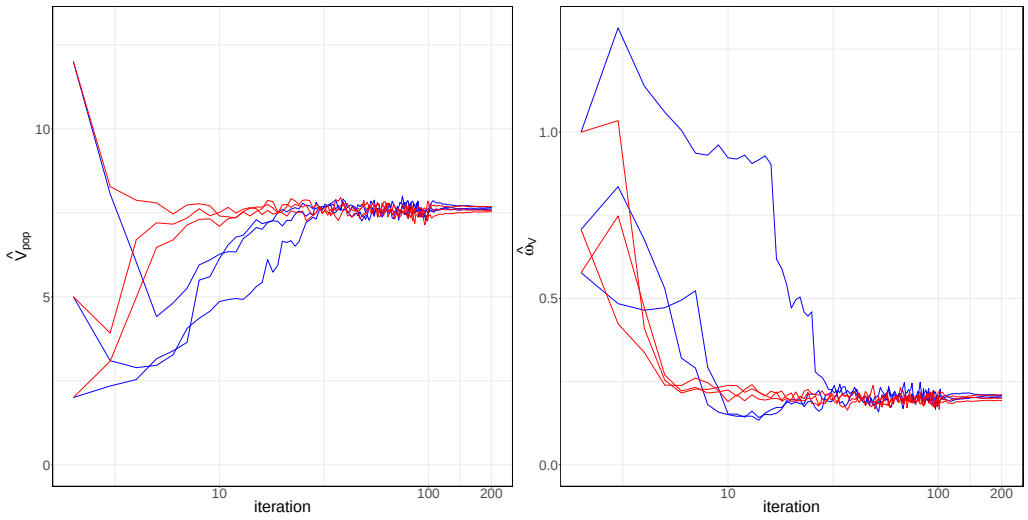

Figure 3: Estimation of the population PK parameters for the warfarin data: convergence of the sequences of estimates $(\hat{V}_{\text{pop},k}, 1 \leq k \leq 200)$ and $(\hat{\omega}_{V,k}, 1 \leq k \leq 200)$ obtained with SAEM and three different initial values using the reference MH algorithm (blue)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

6.1.4. Monte Carlo study

A Monte Carlo study confirmed 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 similar to 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, at each iteration k and for each

component ℓ of the parameter, the mean square distance over the replicates

$$E_k(\ell) = \frac{1}{M}\sum_{m=1}^{M}\left(\theta_k^{(m)}(\ell) - \theta_K^{(m)}(\ell)\right)^2 . \tag{31}$$

Figure 4 clearly shows that the use of the new proposal leads to a much faster convergence towards the maximum likelihood 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 distance decreases monotonically. The sequence of estimates approaches the target with each iteration, compared to the standard algorithm which makes twists and turns before converging.

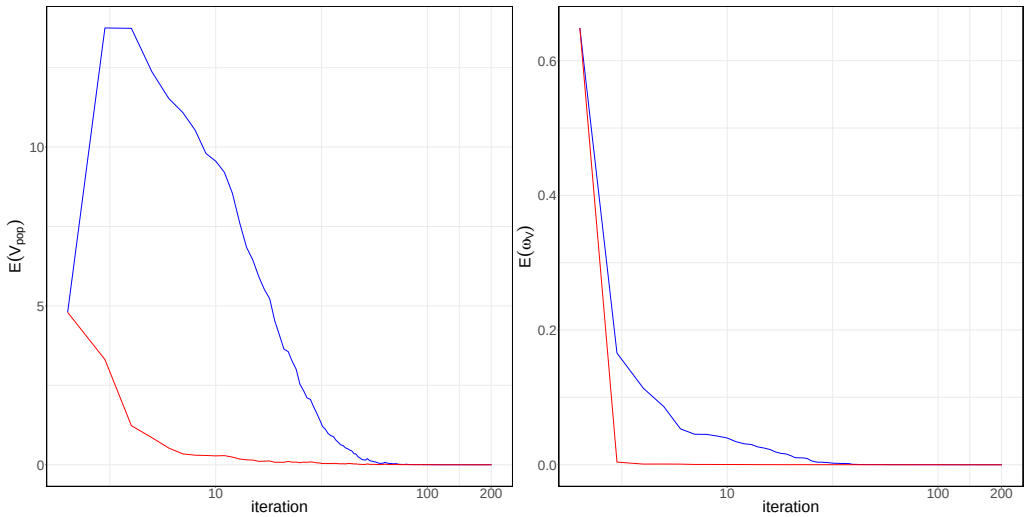

Figure 4: Convergence of the sequences of mean square distances $(E_k(V_{\text{pop}}), 1 \leq k \leq 200)$ and $(E_k(\omega_V), 1 \leq k \leq 200)$ for V_{pop} and ω_V obtained with SAEM on $M = 50$ synthetic datasets using the reference MH algorithm (blue) and the new proposal during the first 10 iterations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

6.2. Time to Event Data Model

6.2.1. The model

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

$$h(t, \psi_i) = \frac{\beta_i}{\lambda_i}\left(\frac{t}{\lambda_i}\right)^{\beta_i - 1} . \tag{32}$$

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

$$\begin{aligned} \log(\lambda_i) &\sim \mathcal{N}(\log(\lambda_{\text{pop}}), \omega_\lambda^2) , \\ \log(\beta_i) &\sim \mathcal{N}(\log(\beta_{\text{pop}}), \omega_\beta^2) . \end{aligned} \tag{33}$$

Then, the vector of population parameters is $\theta = (\lambda_{\text{pop}}, \beta_{\text{pop}}, \omega_\lambda^2, \omega_\beta^2)$.

Individual parameters for $N = 100$ individuals were generated using model (33) with $\lambda_{\text{pop}} = 10$, $\omega_\lambda = 0.3$, $\beta_{\text{pop}} = 3$ and $\omega_\beta = 0.3$. Then, repeated events were generated for each individual using the Weibull model (32) and assuming a right censoring time $\tau_c = 20$.

6.2.2. MCMC Convergence Diagnostic

Similarly to the previous section, we start by looking at the behavior of the MCMC procedure used for sampling from the conditional distribution $\mathtt{p}(\psi_i|y_i;\theta)$ for a given individual i and assuming that θ is known.

We ran 5 000 iterations of the reference MH algorithm and the new proposed algorithm for estimating quantiles of order 0.1, 0.5 and 0.9 of the conditional distributions of λ_i and β_i.

We see Figure 5 that the sequences of empirical quantiles obtained with the two procedures converge to the same value but the new algorithm converges much faster than the standard MH algorithm.

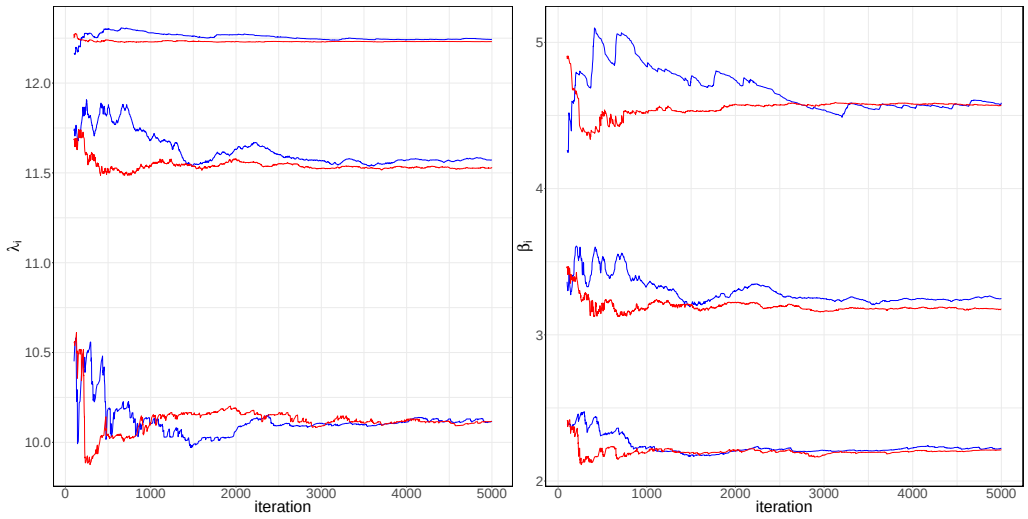

Figure 5: Convergence of the empirical quantiles of order 0.1, 0.5 and 0.9 of $\mathtt{p}(\psi_i|y_i;\theta)$ for a single individual. The reference MH algorithm is in blue and the new version is in red.

6.2.3. Maximum likelihood estimation of the parameters

We no longer consider θ known here. It is estimated using SAEM.

We used first the standard SAEM algorithm implemented in the saemix package (extension of this package for non continuous data models is available on GitHub: `https:`

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

We then introduced the new proposal only during the first five iterations of SAEM.

Figure 6 shows the estimates of λ_{pop} and ω_λ computed at each iteration of the two versions of the SAEM and starting from three different initial values. The same behaviour is observed again as in the continuous cas: regardless the initial values and the algorithm, all the runs converge to the same solution but convergence is much faster with the new method. The same comment applies for the two other parameters β_{pop} and ω_β.

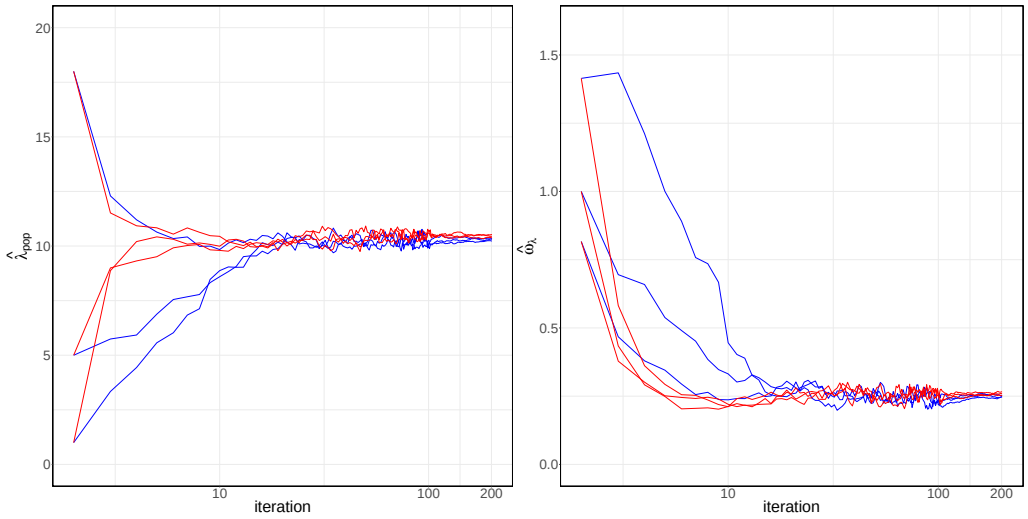

Figure 6: Population parameter estimation in time to event data models: convergence of the sequences of estimates $(\hat{\lambda}_{\text{pop},k}, 1 \leq k \leq 200)$ and $(\hat{\omega}_{\lambda,k}, 1 \leq k \leq 200)$ obtained with SAEM and three different initial values using the reference MH algorithm (blue) and the new proposal during the first 5 iterations (red).

6.2.4. Monte Carlo study

Once again, we have conducted a Monte Carlo study in order to confirm the good properties of the new version of the SAEM algorithm for estimating the population parameters of a time to event data model.

$M = 50$ synthetic datasets have been generated using the same design as previously used. Figure 7 shows the convergence of the mean square distances defined in (31) for λ_{pop} and ω_λ. All these distances converge monotonically to 0 which means that both algorithms properly converge to the maximum likelihood estimate, but very few iterations are required with the new version to converge while about thirty iterations are needed with the standard SAEM.

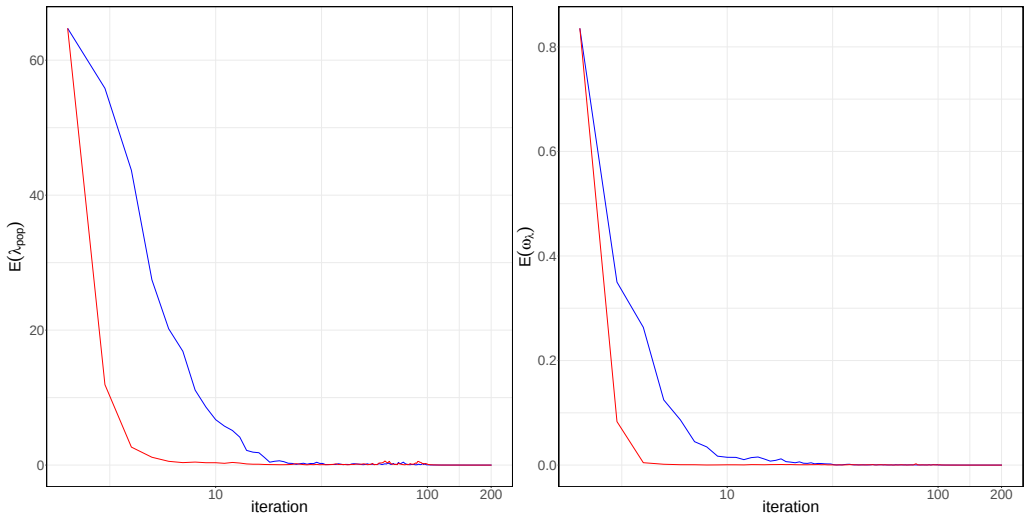

Figure 7: Convergence of the sequences of mean square distances $(E_k(\lambda_{\text{pop}}), 1 \leq k \leq 200)$ and $(E_k(\omega_\lambda), 1 \leq k \leq 200)$ for λ_{pop} and ω_λ obtained with SAEM from $M = 50$ synthetic datasets using the reference MH algorithm (blue) and the new proposal during the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

7. Conclusion and discussion

We presented in this article a Metropolis Hastings procedure for sampling random effects from their conditional distributions in nonlinear mixed effects models.

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

The numerical experiments that we have conducted seem to show that the proposed sampler converges to the target distribution extremely fast. This very good practical behavior is partly explained by the fact that the conditional mode of the random effects in the linearized model coincides with the conditional mode of the random effects in the original model. The proposal distribution is therefore a normal distribution centered around a "good" value. On the other hand, the dependency structure in the conditional distribution of the random effects is well approximated by the covariance structure of the proposal.

Nevertheless, it will be interesting in a future work to compare our approach with the variational inference method that outputs a proposal distribution which is optimal in the sense of Kullback Leibler divergence (Wainwright and Jordan, 2008).

So far, we have mainly applied our method to standard problems encountered in pharmacometry, for example, and for which the number of random effects remains relatively limited. It will nevertheless be interesting to see how this method behaves in higher dimensions and compare it with methods adapted to such situations such as MALA or HMC.

Lastly, we have shown that this new MH algorithm can easily be embedded in the SAEM algorithm for maximum likelihood estimation of the population parameters. Our numerical studies have shown empirically that the new transition kernel is really effective in the very first iterations. It will then be interesting to determine automatically and in an adaptive way an optimal scheme of kernel transitions combining this new proposal with the block-wise random walk Metropolis.

References

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

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

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

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

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

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

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

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

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

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

Duane, S., Kennedy, A.D., Pendleton, B.J., Roweth, D., 1987. Hybrid Monte Carlo. Physics Letters B 195, 216 – 222. doi:10.1016/0370-2693(87)91197-X.

Durmus, A., Moulines, , 2017. Nonasymptotic convergence analysis for the unadjusted Langevin algorithm. Ann. Appl. Probab. 27, 1551–1587. doi:10.1214/16-AAP1238.

Durmus, A., Roberts, G.O., Vilmart, G., Zygalakis, K.C., 2017. Fast langevin based algorithm for mcmc in high dimensions. Ann. Appl. Probab. 27, 2195–2237. doi:10.1214/16-AAP1257.

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

Girolami, M., Calderhead, B., 2011. Riemann manifold langevin and hamiltonian monte carlo methods. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 73, 123–214.

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

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

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

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

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

Ma, Y.A., Chen, T., Fox, E., 2015. A complete recipe for stochastic gradient mcmc, in: Cortes, C., Lawrence, N.D., Lee, D.D., Sugiyama, M., Garnett, R. (Eds.), Advances in Neural Information Processing Systems 28. Curran Associates, Inc., pp. 2917–2925.

Marshall, T., Roberts, G., 2012. An adaptive approach to langevin mcmc. Statistics and Computing 22, 1041–1057. doi:10.1007/s11222-011-9276-6.

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

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

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

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

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

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

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

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

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

Roberts, G.O., Tweedie, R.L., 1996b. Geometric convergence and central limit theorems for multidimensional hastings and metropolis algorithms. Biometrika 83, 95–110. doi:10.1093/biomet/83.1.95.

Salimans, T., Kingma, D.P., Welling, M., 2015. Markov chain monte carlo and variational inference: Bridging the gap. JMLR 37, 1218–1226.

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

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

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

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

Wainwright, M.J., Jordan, M.I., 2008. Graphical models, exponential families, and variational inference. Found. Trends Mach. Learn. 1, 1–305. doi:10.1561/2200000001.

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

Wolfinger, R., 1993. Laplace's approximation for nonlinear mixed models. Biometrika 80, 791–795.

Zhang, C., Shahbaba, B., Zhao, H., 2018. Variational hamiltonian monte carlo via score matching. Bayesian Anal. 13, 485–506. doi:10.1214/17-BA1060.

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

Appendices

A. Proof of Proposition 1

We can directly use (16) to get an expression of the conditional variance of ψ_i under the linearized model:

$$\mathrm{Var}_{\mathrm{lin}}(\psi_i|y_i) = \left(\frac{\nabla f_i(\hat{\psi}_i)\nabla f_i(\hat{\psi}_i)'}{\sigma^2} + \Omega^{-1}\right)^{-1} . \tag{34}$$

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 $(f(t_{i,1},\psi_i),\ldots,f(t_{i,n_i},\psi_i))$ and m_i is defined in Section 4.1 as the mean of prior distribution of ψ_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 (16), we can now compute the conditional mean of ψ_i under the linearized model:

$$\begin{aligned}
\mathbb{E}_{\mathrm{lin}}(\psi_i|y_i) &= \Gamma_i\frac{\nabla f_i(\hat{\psi}_i)'}{\sigma^2}\left(y_i - f_i(\hat{\psi}_i) + \nabla f_i(\hat{\psi}_i)\hat{\psi}_i + \Omega^{-1}m_i\right) \\
&= \Gamma_i\left(\Omega^{-1}(\hat{\psi}_i - m_i) + \frac{\nabla f_i(\hat{\psi}_i)'\nabla f_i(\hat{\psi}_i)}{\sigma^2}\hat{\psi}_i + \Omega^{-1}m_i\right) \\
&= \Gamma_i\Gamma_i^{-1}\hat{\psi}_i \\
&= \hat{\psi}_i .
\end{aligned} \tag{35}$$

B. Proof of Proposition 2

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

$$I := \int e^{v(x)}\mathrm{d}x , \tag{36}$$

where v is at least three times differentiable.

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

$$v(x) \approx v(x_0) + \nabla v(x_0)(x - x_0) + \frac{1}{2}(x - x0)\nabla^2 v(x_0)(x - x0) , \tag{37}$$

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

$$I \approx e^{v(x_0)} \sqrt{\frac{(2\pi)^p}{|-\nabla^2 v(x_0)|}} \exp\left\{-\frac{1}{2}\nabla v(x_0)' \nabla^2 v(x_0)^{-1} \nabla v(x_0)\right\} . \tag{38}$$

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

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

Then, let

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

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 ψ_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

$$\begin{aligned} \nabla^2 \log(\mathtt{p}(y_i, \hat{\psi}_i)) &= \nabla^2 \log(\mathtt{p}(y_i|\hat{\psi}_i)) + \log(\mathtt{p}(\hat{\psi}_i)) \\ &= \nabla^2 l(\hat{\psi}_i) + \Omega^{-1} . \end{aligned} \tag{41}$$