Efficient Metropolis Hastings sampling for nonlinear mixed effects models

Belhal Karimi and Marc Lavielle

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, in contrast with more sophisticated samplers such as the Metropolis Adjusted Langevin Algorithm or the No-U-Turn Sampler that involve costly tuning runs or intensive computation. 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. Numerical experiments based on real data highlight the very good performances of the proposed method for continuous data model.

1 Introduction

Mixed effects models are reference models when the inter-individual variability that can exist within the same population is considered (see [9] and the references therein). Given a population of individuals, the probability distribution of the series of observations for each individual depends on a vector of individual parameters. For complex priors on these individual parameters or

Belhal Karimi
Inria, Paris, France, e-mail: belhal.karimi@inria.fr

Marc Lavielle
Inria, Paris, France, e-mail: marc.lavielle@inria.fr

models, Monte Carlo methods must be used to approximate the conditional distribution of the individual parameters given the observations. 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.

Designing a fast mixing sampler is of utmost importance for several tasks in the complex process of model building. The most common MCMC method for nonlinear mixed effects models is the *random walk Metropolis* algorithm [14, 15, 9]. 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, maintaining an optimal acceptance rate (advocated in [15]) most often implies very small moves and therefore a very large number of iterations in high dimension. 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 [16, 18]. 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 [3]. 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, [8, 12]. Nevertheless, these methods may be difficult to use in practice, and are computationally involved, in particular when the structural model is a complex ODE based model.

The algorithm we propose is a Metropolis-Hastings algorithm, but for which the proposal is a very good approximation of the the target distribution. In the case of continuous data, linearisation of the model leads, by definition, to a Gaussian linear model for which the conditional distribution of the individual parameter given the data is a multidimensional normal distribution that can be calculated. For noncontinuous data model (i.e. categorical, count or time-to-event data models), the Laplace approximation of the incomplete pdf leads to a Gaussian approximation of the conditional distribution of the individual parameter given the observations.

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 that each observation y_{ij} takes its values in some subset of $\mathbb{R}$. The distribution of the n_i-vector of observations y_i depends on a vector of individual parameters ψ_i that takes its values in a subset of $\mathbb{R}^p$. We assume that the pairs (y_i, ψ_i) are mutually independent and consider a parametric framework: the joint distribution of (y_i, ψ_i) is denoted by $\mathtt{p}(y_i, \psi_i; \theta)$, where θ is the vector of parameters of the model. A natural decomposition of this joint distribution writes $\mathtt{p}(y_i, \psi_i; \theta) = \mathtt{p}(y_i|\psi_i; \theta)\mathtt{p}(\psi_i; \theta)$, 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 of individual random effects η_i: $\psi_i = \psi_{\text{pop}} + \eta_i$. 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 extensions of this model are straightforward, considering for instance transformation of the normal distribution, or adding individual covariates in the model.

2.2 Continuous data models

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

$$y_{ij} = f(t_{ij}, \psi_i) + \varepsilon_{ij} \ , \tag{1}$$

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 $\theta = (\psi_{\text{pop}}, \Omega, \sigma^2)$.

3 Sampling from conditional distributions

The 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, $\hat{\theta}_{\text{ML}} = \arg\max_{\theta \in \Theta} \mathcal{L}(\theta, y)$, where $\mathcal{L}(\theta, y) \triangleq \log \mathtt{p}(y; \theta)$. The

stochastic approximation version of EM [7] is an iterative procedure for ML estimation that requires to generate one or several realisations of this conditional distribution at each iteration of the algorithm.

Once the ML estimate $\hat{\theta}_{\text{ML}}$ has been computed, the observed Fisher information matrix $I(\hat{\theta}_{\text{ML}}, y) = -\nabla^2_\theta \mathcal{L}(\hat{\theta}_{\text{ML}}, y)$ can be derived thanks to the Louis formula [10]. 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 $\mathtt{p}(\psi_i|y_i;\hat{\theta}_{\text{ML}})$. Then, several statistical tests and diagnostic plots used for models assessment are based on realisations of the conditional distribution $\mathtt{p}(\psi_i|y_i;\hat{\theta}_{\text{ML}})$, rather than the mode of this distribution, in order to provide unbiased tests and plots.

Current implementations of the MCMC algorithm, to which we will compare our new method, in Monolix [5], saemix (R package) [6], nlmefitsa (Matlab) and NONMEM [2] mainly use the same combination of proposals. The first proposal is an independent Metropolis-Hasting (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 [11] 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 [1, 9]. 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) [16, 18] and the Hamiltonian Monte Carlo (HMC) which implementation can be found for instance as the "No U-Turns Sampler" in STAN [8, 4].

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 Nonlinear continuous data models

We will consider the model (1) for continuous data and assume that the ψ_i's are normally distributed with mean ψ_{pop} and variance-covariance Ω. For a given parameter value θ, the MAP estimate, for individual i, is the value of ψ_i that maximises the conditional distribution $\mathtt{p}(\psi_i|y_i,\theta)$:

$$\hat{\psi}_i = \arg\max_{\psi_i} \mathtt{p}(\psi_i|y_i,\theta) = \arg\max_{\psi_i} \mathtt{p}(y_i|\psi_i,\theta)\mathtt{p}(\psi_i,\theta)$$

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

$$f_i(\psi_i) \approx f_i(\hat{\psi}_i) + \nabla f_i(\hat{\psi}_i)(\psi_i - \hat{\psi}_i) \;, \tag{2}$$

where $\nabla f_i(\hat{\psi}_i) = J_i$ is the Jacobian matrix of vector $f_i(\hat{\psi}_i)$. Defining $z_i \triangleq y_i - f_i(\hat{\psi}_i) + J_i\hat{\psi}_i$ yields a linear model $z_i = J_i\psi_i + \epsilon_i$ which tractable conditional distribution can be used for approximating $\mathtt{p}(\psi_i|y_i,\theta)$:

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

$$\mu_i = \hat{\psi}_i \quad \textit{and} \quad \Gamma_i = \left(\frac{\nabla f_i(\hat{\psi}_i)'\nabla f_i(\hat{\psi}_i)}{\sigma^2} + \Omega^{-1}\right)^{-1} . \tag{3}$$

We then use this normal distribution as a proposal for continuous models.

4.2 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 parameters ψ_i. In our context, we can write the marginal pdf $\mathtt{p}(y_i)$ that we aim to approximate as $\mathtt{p}(y_i) = \int e^{\log \mathtt{p}(y_i,\psi_i)} \mathrm{d}\psi_i$. Then, the Taylor expansion of $\log(\mathtt{p}(y_i,\psi_i)$ around the MAP $\hat{\psi}_i$ (that verifies by definition $\nabla \log \mathtt{p}(y_i,\hat{\psi}_i) = 0$) yields the Laplace approximation of $-2\log(\mathtt{p}(y_i))$ as follows:

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

We thus obtain the following approximation of $\log \mathtt{p}(\hat{\psi}_i|y_i)$:

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

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 $\hat{\psi}_i$.

Proposition 2. *Let (y_i, ψ_i) be a pair of random variables where ψ_i is normally distributed with variance-covariance matrix Ω. Let $\hat{\psi}_i$ be the MAP of the conditional distribution $\mathtt{p}(\psi_i|y_i)$. Then, this latter distribution can be approximated by a Gaussian distribution with mean $\hat{\psi}_i$ and variance-covariance*

$$\Gamma_i = -\nabla^2 \log \mathtt{p}(y_i, \hat{\psi}_i)^{-1} = \left(-\nabla^2 \log \mathtt{p}(y_i|\hat{\psi}_i) + \Omega^{-1}\right)^{-1} .$$

Remark: In the case of continuous outcomes, linearising the structural model is equivalent to using the Laplace approximation with the expected information matrix. Indeed $\mathbb{E}\left(-\nabla^2 \log \mathtt{p}(y_i|\hat{\psi}_i)|\psi_i = \hat{\psi}_i\right) = \frac{\nabla f_i(\hat{\psi}_i) \nabla f_i(\hat{\psi}_i)'}{\sigma^2}$.

5 A pharmacokinetic example

5.1 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 [13], for who we measure warfarin plasmatic concentration at different times. We will consider a one-compartment pharmacokinetics (PK) model for oral administration, assuming first-order absorption and linear elimination processes:

$$f(t, ka, V, k) = \frac{D\, ka}{V(ka - k)}(e^{-ka\, t} - e^{-k\, t}) , \tag{4}$$

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 $y_{ij} = f(t_{ij}, \psi_i) + \varepsilon_{ij}$. 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. The model that is used here for these data is therefore the nonlinear 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 linearisation of the structural model f proposed in (4). 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.

5.2 MCMC Convergence Diagnostic

We will consider only one of the 32 individuals for this study and fix θ to some arbitrary value, close to the Maximum Likelihood (ML) estimate obtained with SAEM (saemix R package [6]): $ka_{\text{pop}} = 1$, $V_{\text{pop}} = 8$, $k_{\text{pop}} = 0.01$, $\omega_{ka} = 0.5$, $\omega_V = 0.2$, $\omega_k = 0.3$ and $\sigma^2 = 0.5$. First, we compare the classical version of MH implemented in the saemix package composed of: independent draw using the marginal distribution $\mathtt{p}(\psi_i)$, component-wise random walk and block-wise random walk used successively with our new proposal.

We ran 20 000 iterations of these two algorithms looked 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 1 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.

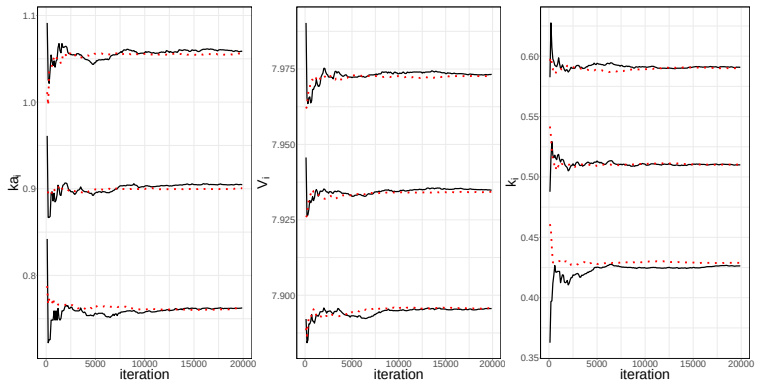

Fig. 1 Modelling of the warfarin PK data: convergence of the empirical quantiles of order 0.1, 0.5 and 0.9 of $\mathtt{p}(\psi_i|y_i;\theta)$ for a single individual. The reference MH algorithm (Random walk) is in black and solid and the new version is in red and dotted.

Then, we implement the MALA, which proposal, at iteration k, is defined by $\psi_i^c \sim \mathcal{N}(\psi_i^{(k)} - \gamma_k \nabla \log \pi(\psi_i^{(k)}), 2\gamma_k)$. The stepsize ($\gamma = 10^{-2}$) is constant and is tuned such that the optimal acceptance rate of 0.57 is reached [15]. The gradient of the log posterior distribution $\nabla \log \pi(\psi_i^{(k)})$ is calculated nu-

merically at each iteration of the MCMC algorithm by finite difference using the following relationship: $\nabla \log \pi(\psi_i^{(k)}) = \nabla \log \mathbf{p}(y_i|\psi_i^{(k)}) + \nabla \log \mathbf{p}(\psi_i^{(k)})$. Figure 2 highlights good convergence of a well-tuned MALA. Quantiles stabilisation is quite similar to our method for all orders and dimensions. Yet, tuning the MALA involved costly runs that does not scale well with the dimension. The NUTS [8] uses a recursive algorithm to build a set of candidate samples that spans a broad range of the target distribution without requiring the user to choose how many steps it wants to execute. The NUTS, implemented in rstan (R Package [17]), is fast and steady and presents similar, or even better convergence behaviors for some quantiles and dimension, than our method (see Figure 2).

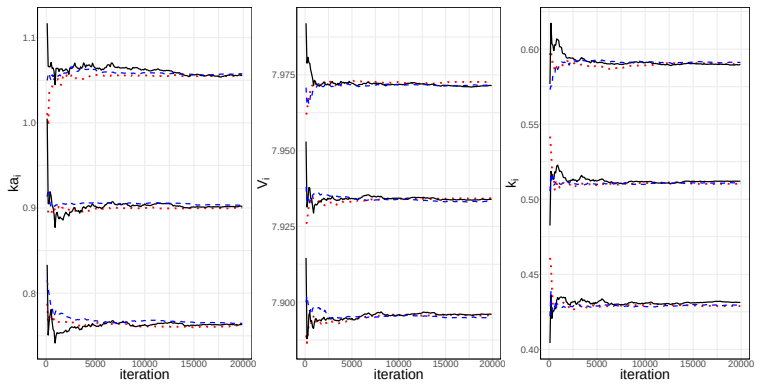

Fig. 2 Modelling of the warfarin PK data: convergence of the empirical quantiles of order 0.1, 0.5 and 0.9 of $\mathbf{p}(\psi_i|y_i;\theta)$ for a single individual. Our new MH algorithm is in red and dotted, the MALA is in black and solid and the NUTS is in blue and dashed.

6 Conclusion and discussion

We presented in this article an independent Metropolis-Hastings procedure for sampling random effects from their conditional distributions in nonlinear mixed effects models. The numerical experiments that we have conducted seem to show that the proposed sampler converges to the target distribution as fast as state-of-the-art samplers. This very good practical behaviour is partly explained by the fact that the conditional mode of the random effects in the linearised model coincides with the conditional mode of the random effects in the original model. Initial experiments embedding this fast and easy-to-implement MH algorithm within the SAEM algorithm [7], for Maximum Likelihood Estimation, indicate a drastically faster convergence behavior.

References

[1] Atchadé, Y. F. and Rosenthal, J. S. (2005). On adaptive Markov chain Monte Carlo algorithms. *Bernoulli*, 11(5):815–828.

[2] Beal, S. and Sheiner, L. (1980). The NONMEM system. *The American Statistician*, 34(2):118–119.

[3] Betancourt, M. (2017). A conceptual introduction to Hamiltonian Monte Carlo. *arXiv preprint arXiv:1701.02434*.

[4] Carpenter, B., Gelman, A., Hoffman, M., Lee, D., Ben, G., Betancourt, M., Brubaker, M., Guo, J., Li, P., and Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software*, 76(1).

[5] Chan, P. L. S., Jacqmin, P., Lavielle, M., McFadyen, L., and 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(1):41–61.

[6] Comets, E., Lavenu, A., and Lavielle, M. (2017). Parameter estimation in nonlinear mixed effect models using saemix, an r implementation of the saem algorithm. *Journal of Statistical Software*, 80(3):1–42.

[7] Delyon, B., Lavielle, M., and Moulines, E. (1999). Convergence of a stochastic approximation version of the EM algorithm. *Ann. Statist.*, 27(1):94–128.

[8] Hoffman, M. D. and Gelman, A. (2014). The No-U-turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, 15(1):1593–1623.

[9] Lavielle, M. (2014). *Mixed effects models for the population approach: models, tasks, methods and tools*. CRC press.

[10] 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.

[11] Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953). Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, 21(6):1087–1092.

[12] Neal, R. M. et al. (2011). MCMC using Hamiltonian dynamics. *Handbook of Markov Chain Monte Carlo*, 2(11).

[13] O'Reilly, R. A. and Aggeler, P. M. (1968). Studies on coumarin anticoagulant drugs initiation of warfarin therapy without a loading dose. *Circulation*, 38(1):169–177.

[14] Robert, C. P. and Casella, G. (2010). *Metropolis–Hastings Algorithms*, pages 167–197. Springer New York, New York, NY.

[15] Roberts, G. O., Gelman, A., and Gilks, W. R. (1997). Weak convergence and optimal scaling of random walk metropolis algorithms. *Ann. Appl. Probab.*, 7(1):110–120.

[16] Roberts, G. O. and Tweedie, R. L. (1996). Exponential convergence of langevin distributions and their discrete approximations. *Bernoulli*, 2(4):341–363.

[17] Stan Development Team (2018). RStan: the R interface to Stan. R package version 2.17.3.

[18] Stramer, O. and Tweedie, R. L. (1999). Langevin-type models i: Diffusions with given stationary distributions and their discretizations. *Methodology And Computing In Applied Probability*, 1(3):283–306.