Fast Maximum Likelihood Estimation method using efficient MCMC proposal

B. Karimi[1,2], **M. Lavielle**[1,2], **E. Moulines**[2]

INRIA[1], CMAP École Polytechnique[2]

belhal.karimi@polytechnique.edu

1 Problem statement

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.

Notations and Models

- **Population approach**. 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 $\mathbb{R}^l$.
- A natural decomposition of the joint distribution consists in writing:

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

- $\mathrm{p}(\psi_i; \theta)$ is the so-called **population distribution** used to describe the distribution of the individual parameters within the population.
- Incomplete log likelihood $L(\theta; y)$

$$L(\theta; y) \triangleq \mathrm{p}(y; \theta) = \prod_{i=1}^{N} \mathrm{p}(y_i; \theta) \tag{2}$$

- The ML estimate of θ is thus defined by:

$$\hat{\theta}_{ML} = \arg\max_{\theta \in \Theta} L(\theta, y) \tag{3}$$

- **Mixed Effect models**. Describing each individual parameters ψ_i as a composition of fixed effects, common to the whole population, and random effects as follows:

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

With β a new vector of fixed effects and C_i a matrix of individual covariates.

2 Maximum Likelihood Estimation

2.1 SAEM Algorithm coupled with MCMC procedure

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

Algorithm 1 SAEM algorithm

Initialisation: sample latent data $\psi_i^0 \sim \mathrm{p}(\psi_i|y_i; \theta^0)$ under a given model estimate θ^0,
Iteration k: given the current estimate θ^{k-1}:

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

$$Q_k(\theta) = Q_{k-1}(\theta) + \gamma_k \left[\sum_{i=1}^{N} \log \mathrm{p}(y_i, \psi_i^k; \theta) - Q_{k-1}(\theta)\right] \tag{5}$$

Where $\{\gamma_k\}_{k>0}$ is a sequence of positive stepsize with $\gamma_1 = 1$.
3. Set $\theta^k = \arg\max_{\theta \in \Theta} Q_k(\theta)$

Theorem 2.1: Convergence of the SAEM coupled with MCMC

With certain assumptions of ergodicity and smoothness of the transition kernel used in the MCMC:

1. if the complete model belongs to the exponential family and its sufficient statistics stay in a compact, then the results of convergence of (B. Delyon and Moulines(1999)) holds w.p.1.

2.2 Posterior sampling - Metropolis Hastings Algorithm

2.2.1 Continuous models

In the case where the outcomes are continuous and the individual parameters ψ_i are normally distributed the model is defined by:

$$y_{ij} = f(t_{ij}, \psi_i) + \epsilon_{ij} \tag{6}$$

and our new method is based on the linearisation of the structural model around the MAP defined as $\hat{\psi}_i = \arg\max_{\psi_i} \mathrm{p}(\psi_i|y_i, \theta)$.

Gaussian proposal for continuous models

- Taylor expansion of the structural model around the MAP:

$$f(t_i, \psi_i) \approx f(t_i, \hat{\psi}_i) + \nabla_{\psi_i} f(t_i, \hat{\psi}_i).^{\top}(\psi_i - \hat{\psi}_i) \tag{7}$$

- Resulting linear model between y_i and ψ_i:

$$\begin{aligned} &y_i - f(t_i, \hat{\psi}_i) + \nabla_{\psi_i} f(t_i, \hat{\psi}_i).^{\top}\hat{\psi}_i \\ &= \nabla_{\psi_i} f(t_i, \hat{\psi}_i).^{\top}\psi_i + \epsilon_i \end{aligned} \tag{8}$$

- Tractable conditional distribution $\psi_i|y_i$: Gaussian distribution $\mathcal{N}(\mu_i, \Gamma_i)$ with parameters:

$$\begin{aligned} \mu_i &= \hat{\psi}_i \\ \Gamma_i &= \left[\frac{\nabla_{\psi_i} f(\hat{\psi}_i).^{\top}\nabla_{\psi_i} f(\hat{\psi}_i)}{\sigma^2} + \Omega^{-1}\right]^{-1} \end{aligned} \tag{9}$$

2.2.2 Non Continuous 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 (Wolfinger(2017)) or (Y.(2007)).
Define $l(\psi_i) \triangleq \mathrm{p}(y_i|\psi_i)$. We can derive the following Gaussian proposal:

Gaussian proposal for non continuous models

- Laplace approximation, around the MAP, of:

$$\mathrm{p}(y_i, \theta) = \int e^{\log \mathrm{p}(y_i, \psi_i, \theta)} \mathrm{d}\psi_i \tag{10}$$

- We obtain:

$$\begin{aligned} -2\log \mathrm{p}(y_i) \approx &-2\log l(\hat{\psi}_i) - 2\log \mathrm{p}(\hat{\psi}_i) \\ &- p\log 2\pi + \log|-\nabla^2 \log \mathrm{p}(y_i, \hat{\psi}_i)| \end{aligned} \tag{11}$$

- Gaussian proposal:

$$\begin{aligned} \mu_i &= \hat{\psi}_i \\ \Gamma_i &= \left[-\nabla^2 \log l(\hat{\psi}_i) + \Omega^{-1}\right]^{-1} \end{aligned} \tag{12}$$

- Fisher Approximation:

$$-\nabla^2 \log l(\hat{\psi}_i) \approx \mathbb{E}^{y_i|\hat{\psi}_i}\left[\nabla^2 \log l(\hat{\psi}_i)\right] \tag{13}$$

- Combined to the Fisher identity, we obtain:

$$-\nabla^2 \log l(\hat{\psi}_i) \approx \frac{\nabla l(\hat{\psi}_i).\nabla l(\hat{\psi}_i)^{\top}}{l^2(\hat{\psi}_i)} \tag{14}$$

3 Numerical Application: Warfarin dataset

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. O'reilly(1968)), 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.

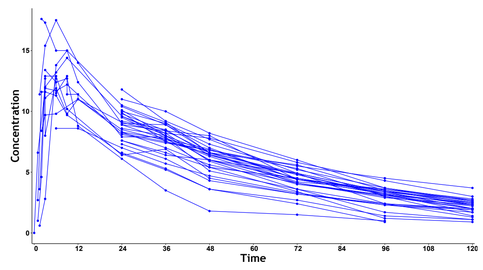

Figure 1: Warfarin concentration over time for 32 subjects.

PK model

$$y_{ij} = \frac{Dka}{V(ka - k)}(e^{-kat} - e^{-kt}) + \epsilon_{ij} \tag{15}$$

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 $\mathrm{p}(y_i, \psi_i, \theta)$ 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, $u(\psi_i) = (\log(ka_i), \log(V_i), \log(k_i))$.

- **Fast MCMC Convergence**

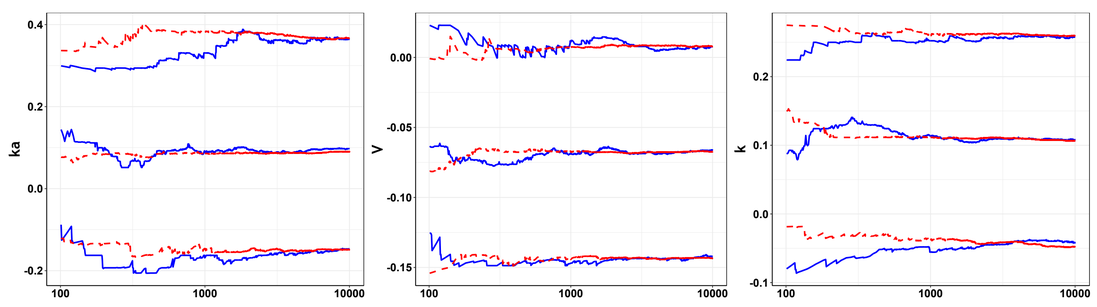

Figure 2: Convergence of quantiles (0.05, 0.5, 0.95).

- **Fast SAEM Convergence**

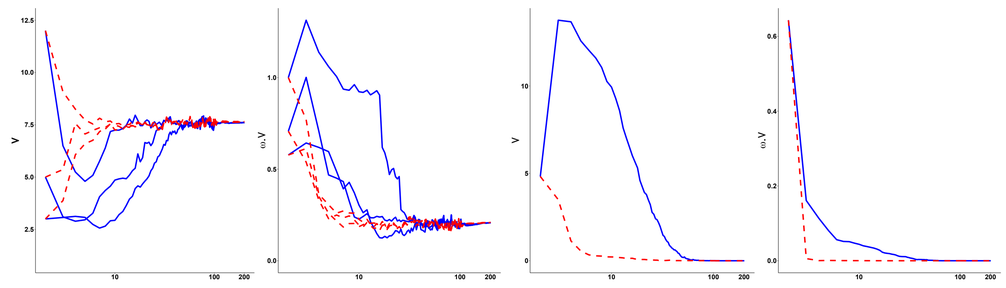

Figure 3: Runs on Warfarin dataset (Left) and average error on 100 synthetic datasets (Right) Fast method in red and reference in blue.

References

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

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

(RA. O'reilly(1968)) PM Aggeler RA. O'reilly. Studies on coumarin anticoagulant drugs. Initiation of warfarin therapy without a loading dose. 1968.

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

(Y.(2007)) Wang Y. Derivation of various NONMEM estimation methods.. *Journal of Pharmacokinetics and Pharmacodynamics*, 2007.