Non linear Mixed Effects Models: Bridging the gap between Independent Metropolis Hastings and Variational Inference

Belhal Karimi, Marc Lavielle and Eric Moulines [1]

Abstract

Variational inference and MCMC methods have been two popular methods in order to sample from a posterior distribution. Whereas the former extends the computation feasibility to higher dimension, the latter takes advantage of nice convergence properties to the exact posterior distribution. In this work we'll draw the parallel between a famous MCMC scheme called the Independent Metropolis Hastings and Variational inference. We'll explain our work on both Linear and Non-linear Gaussian cases. In the non linear case, a new proposal will be introduced motivated by a faster convergence of the Markov chain.

1. Introduction

We consider a complete model (y,z) where the realizations of y are observed and z is the missing data. When the complete model $p(y, z, \theta)$ is parametric, the goal is to compute the maximum likelihood (ML) estimate of the parameter of this joint distribution.

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

When the direct derivation of this expression is hard, several methods use the complete model to iteratively find the quantity of interest. The EM algorithm has been the object of considerable interest since its presentation by Dempster, Laird and Rubin in 1977. It has been relatively effective in context of maximum likelihood estimation of parameters of incomplete model (unobserved or more). This algorithm is monotonic in likelihood making it a stable tool to work with.

Yet, when the quantity computed at the E-step involves infeasible computations, new methods have been developed in order to by-pass the issue. The stochastic EM algorithm (Celeux & Diebolt, 1985) has been proposed in the context of mixture problem and involves splitting the E-step in a first simulation of the latent variables step and then a direct evaluation of the complete log model. A Robbins Monroe type approximation can be used to evaluate that latter quantity after the simulation step, that is the SAEM algorithm (Lavielle, 1993; E.Moulines, 2007).

2. Model and notations

We study a classical missing data problem where:

- The observed data is a continuous random variable $Y = (Y_i, 1 \leq i \leq N)$ that has observed values $(y_i, 1 \leq i \leq N)$ in $\mathcal{Y}$
- The latent data is a continuous random variable $Z = (Z_i, 1 \leq i \leq N)$ that takes on the values $(z_i, 1 \leq i \leq N)$ in $\mathcal{Z}$ and consists in N independent variables
- The components Y_i are generated independently of each other and from their corresponding Z_i
- $\log p(y, \theta)$ is the incomplete data log-likelihood
- $\log p(y, z, \theta)$ is the complete data log-likelihood and obtained by augmenting the observed data with the missing data
- We'll call $P_{Y_i, Z_i, \theta}$ and $P_{Z_i|Y_i, \theta}$ the probability distributions associated to the densities $p(y_i, z_i, \theta)$ and $p(z_i|y_i, \theta)$

3. Maximum likelihood estimation

Our problem joins a familiar class of problem in computational statistics that consists in maximizing the following quantity:

$$\log p(y, \theta) = \int \log p(y, z, \theta) \mu(dz) \tag{2}$$

When this quantity can not be computed in closed form, many algorithms use iterative procedure to find the

[1]Ecole Polytechnique, France. Correspondence to: Belhal Karimi <belhal.karimi@polytechnique.edu>.

ICML 2017 Workshop on Implicit Models. Copyright 2017 by the author(s).

maximum likelihood parameter estimate. Among those techniques, the EM algorithm (Dempster & Rubin, 1977). This two steps algorithm consists in maximizing an auxiliary quantity that is the expectation of the complete log-likelihood with respect to the conditional distribution over the missing variable conditioned on the current parameter estimate (also called the posterior distribution). Several alternatives have been developed throughout the past decades. Most of them alleviate the computation of the expectation using approximates. The MCEM algorithm (Celeux & Diebolt, 1985) approximate this quantity by a Monte Carlo integration, the SAEM algorithm (B. Delyon & Moulines, 1999) uses a stochastic approximation of this quantity.

In those both cases, we need to be able to simulate from the posterior distribution $P(Z_i|Y_i, \theta)$. In most of the cases, this probability density function is intractable. As a result, variants include MCMC or Variational Inference engines to sample from this distribution. That's where we are focusing. In the sequel, the parameter θ is thus fixed to a certain value θ_0 that will remains unchanged.

4. Background on Posterior sampling

4.1. Independent Metropolis Hastings

Metropolis-Hastings are a powerful class of inference algorithms that belong to the family of MCMC methods. This kind of algorithm constructs a Markov Chain by proposing candidate states sampled from a proposal distribution and then accepting or rejecting it according to the MH-step (see Algorithm 1 for detail). When this proposal is independent of the current state of the chain, we call the algorithm Independent Metropolis Hastings.

Algorithm 1 Independent Metropolis Hastings

Input: initial state Z_0, proposal distribution q, number of iterations M, target measure π
for $m = 1$ **to** M **do**
  $Z_m \sim q(Z)$
  $\alpha(Z_m, Z_{m-1}) = \frac{\pi(Z_m)q(Z_{m-1})}{\pi(Z_{m-1})q(Z_m)}$
  Accept Z_m with probability $\min(\alpha, 1)$
end for

We will explicitly write our target measure π in the different cases we'll deal with in the sequel.

4.2. Variational Inference

Variational methods approximates those intractable distributions by finding the best distribution minimizing a divergence criteria. Let $\mathcal{D}$ be a family of distributions over the latent variable Z_i. Variational Methods solve the following optimization problem:

$$q^* = \arg\min_{q \in \mathcal{D}} D_{KL}(q || P_{Z_i|Y_i,\theta_0}) \tag{3}$$

Which simplifies to:

$$q^* = \arg\max_{q \in \mathcal{D}} \mathcal{L}(q) \tag{4}$$

Where $\mathcal{L}(q) = \mathbb{E}_q(\log p(y_i, z_i, \theta_0)) - \mathbb{E}_q(\log q(z_i))$ is called the ELBO (Evidence Lower BOund).

The practical implementation of such an algorithm consisting in first, restricting the search space to a family of known distributions (in our case the Gaussian family). Then, performing a Monte Carlo integration of the gradient of the ELBO, see (R. Ranganath & Blei, 2013) and finally performing a gradient ascent as described by Algorithm 2. We restrict ourselves to the family of Gaussian distributions of mean μ and variance Γ. We run the variational inference on the mean and fix the value of Γ. As a result the problem now is written, if we denote q_μ the Gaussian density of mean μ:

$$\mu^* = \arg\max_{\mu \in \mathbb{R}} \mathcal{L}(\mu) \tag{5}$$

Where $\mathcal{L}(\mu) = \mathbb{E}_{q_\mu}(\log p(y_i, z_i, \theta_0)) - \mathbb{E}_{q_\mu}(\log q_\mu(z_i))$

Algorithm 2 Gradient Descent for VI

Input: number of iterations K, initial μ_0, stepsize ρ
Initialize $\mu^0 = \mu_0$.
for $k = 1$ **to** K **do**
  $\mu^k < -\mu^{k-1} + \rho \nabla_\mu \mathcal{L}$
end for
Return μ^K

5. Mixed effect models

In our domain of applications, mainly pharmacokinetic and pharmacodynamic (PK-PD) data analysis, we are mostly facing mixed effects models:

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

Where $f : \mathbb{R}^d \to \mathbb{R}$ is the structural model and is a function of ψ_i that can be linear or not, $\psi_i \in \mathbb{R}^d$ are the individual parameters, y_{ij} are the observations for individual i and $\epsilon_{ij} \sim \mathcal{N}(0, \Sigma)$. There can be more than one observation (subscript j) per individual. The parameters ψ_i are composed of a fixed part ψ_{pop} and a random one η_i:

$$\psi_i = \psi_{pop} + \eta_i \tag{7}$$

where $\eta_i \sim \mathcal{N}(0, \Omega)$.

In this context, our goal is to sample from the posterior $P(\psi_i|y_i, \theta_0)$ for all individuals. For simplicity we'll consider the centered random variable η_i in our study. Thus, the goal is to sample from $P(\eta_i|y_i, \theta_0) = P(y_i|\eta_i, \theta_0)P(\eta_i)$. For the purpose of the stochastic EM algorithm we can easily shift to $\psi_i = M(\eta_i, \psi_{pop})$.
We'll now separate the cases where the structural model is linear or not and explain how MCMC are Variational Inference can be applied to our problem.

5.1. Gaussian linear case

For simplicity, we omit the number of observations per individual. The model can be written as:

$$y_i = A_i\psi_i + \epsilon_i \tag{8}$$

Where A_i is a design matrix and $\psi_i = \psi_{pop} + \eta_i$. In this case, the true posterior is tractable and we can easily calculate that:

$$\eta_i|y_i \sim \mathcal{N}(m, G) \tag{9}$$

Where $m = \Gamma A_i'\Sigma^{-1}(y_i - A_i\psi_{pop})$ and $G = (A_i'\Sigma^{-1}A_i + \Omega^{-1})^{-1}$

In this case, we can sample directly from the true posterior distribution. Thus, using MCMC or Variational Inference has no sense. We can verify that those both methods, whether we propose in the context of an MCMC with the true posterior or we apply a gradient descent over the mean of a candidate distribution, will converge to the right distribution.

5.2. Gaussian non linear case

The model can be written as:

$$y_i = f(\psi_i) + \epsilon_i \tag{10}$$

Where $f : \mathbb{R}^d \to \mathbb{R}$ is a non linear function and $\psi_i = \psi_{pop} + \eta_i$. The posterior distribution $P(\eta_i|y_i, \theta_0)$ is intractable in this case.
We are constructing a new independent proposal distribution for our Metropolis Hastings algorithm based on the linearization of this model around the maximum a posteriori (MAP) $\hat{\psi}_i$.

$$\hat{\psi}_i = \arg\max_{\psi_i} P(\psi_i|y_i; \theta_0) \tag{11}$$

which can rewrite, considering the distribution of the components of the model:

$$\begin{aligned}\hat{\psi}_i = \arg\max_{\psi_i} \quad & (y_i - f(\psi_i))'\Sigma^{-1}(y_i - f(\psi_i)) \\ & + (\psi_i - \psi_{pop})'\Omega^{-1}(\psi_i - \psi_{pop})\end{aligned} \tag{12}$$

The Taylor expansion around this point gives the following expression:

$$y_i = f(\hat{\psi}_i) + \nabla_\psi f(\hat{\psi}_i)(\psi_i - \hat{\psi}_i) + \epsilon_i \tag{13}$$

Writing that $\psi_i = \psi_{pop} + \eta_i$ and developing the relation in order to have a linear expression in η_i, we obtain:

$$y_i - f(\hat{\psi}_i) - \nabla_\psi f(\hat{\psi}_i)(\psi_{pop} - \hat{\psi}_i) = \nabla_\psi f(\hat{\psi}_i)\eta_i + \epsilon_i \tag{14}$$

We can now write the posterior distribution $\eta_i|y_i$ and show that $\eta_i|y_i \sim \mathcal{N}(\mu_{lin}, \Gamma_{lin})$ where $\mu_{lin} = \mathbb{E}(\eta_i|y_i) = \hat{\eta}_i = \hat{\psi}_i - \psi_{pop}$ and $\Gamma_{lin}^{-1} = \nabla_\psi f(\hat{\psi}_i)'\Sigma^{-1}\nabla_\psi f(\hat{\psi}_i) + \Omega^{-1}$

This new distribution will serve as an independent (independent of the current state of the chain since always centered in $\hat{\eta}_i$) proposal for our MH algorithm.
The other aspect we want to develop is to know if a variational approach where our candidate distribution would be taken in the family of the Gaussian distributions with variance Γ_{lin} and with mean μ on which we'll do perform the gradient ascent, would converge to the same mean $\hat{\eta}_i$. In other words, we want to highlight the equivalence of those two methods based on the same approximation (that the proposal is a Gaussian distribution even though, due to the non-linearity of the model, the true posterior does not belong to a known family of distribution).

6. Experiments

6.1. One-compartment model for Theophylline

This section develops the application of those two methods on a Pharmacokinetics (PK) example. Beforehand, the standard approach is to approximate the body as a simple compartment models. In this example we will focus on a one-compartment model for theophylline following oral dose D at time $t = 0$ leading to description of concentration $y(t_i)$ at time $t_i \geq 0$ (i varies from 1 to N and denote the individual of the population):

$$y_i = y(t_i) = f(\psi_i) + \epsilon_i \tag{15}$$

With :

$$f(\psi_i) = \frac{D(k_a)_i}{V_i((k_a)_i - (C_l)_i/V_i)}(e^{(-(k_a)_i t_i)} - e^{(-\frac{(C_l)_i}{V_i}t_i)}) \tag{16}$$

Where $(k_a)_i$ is the fractional rate of absorption for individual i, $(C_l)_i$ is the clearance rate for individual i and V_i is the volume of distribution for individual i and D is the dose injected.
In our notation, the complete model is $p(y_i, \psi_i, \theta_0)$ where $\psi_i = ((k_a)_i, (C_l)_i, V_i)$ is the vector of individual parameters where each component is composed of a fixed effect term and a random effect (a centered Gaussian with same variance Ω) and $\theta_0 = (\Omega_0, \Sigma_0$ (with $\epsilon_i \sim \mathcal{N}(0, \Sigma_0)$ and

$(k_a)_i = (k_a)_{pop} + \eta_{(k_a)_i}$ and $\eta_{(k_a)_i} \sim \mathcal{N}(0, \Omega_0)$).
Our goal is to simulate for instance from the posterior distribution $P((k_a)_i|y_i, \theta_0)$. As we said above we can work on the distribution $P(\eta_{(k_a)_i}|y_i, \theta_0)$ and equivalently for the others parameters.
Following the method of linearization of this model as described above, we obtain faster convergence of the MH algorithm with this new independent proposal than our reference Random Walk Metropolis consisting in three successive kernels proposing with a Gaussian centered in the current state of the chain and whose variance adapts with respect to the optimum acceptance rate.

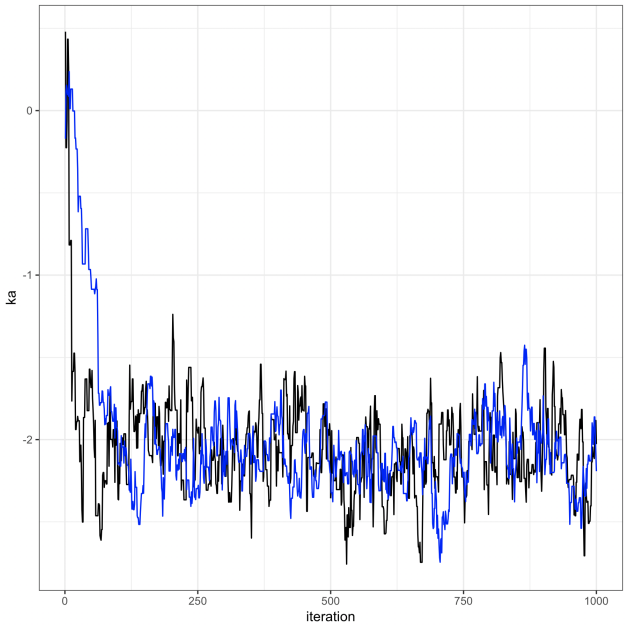

Figure 1. MCMC samples. RWM in blue and Independent MH with our new proposal in black. 1000 iterations of MCMC iterations. Plotting the posterior distribution $P(\eta_{(k_a)_i}|y_i, \theta_0)$ for a random individual i

7. Discussion

This work is still in progress. Our goals are multiple. First of all, following (E.Kuhn, 2015) setting convergence properties for the SAEM algorithm coupled with an MCMC procedure, we are aiming at setting up similar properties for the SAEM coupled with a variational inference procedure (See also (A. Gunawardana, 2005) for the Variational EM algorithm). Moreover, we would like to theoretically draw some parallels between those two methods. And finally, many PK-PD models consider not continuous but categorical or count data. In this case, an alternative to linearization has to be found in order to construct such a proposal.
Also, in our context of applying this method to the SAEM algorithm, while calculating the MAP once for one MCMC run is not costly, doing it K times (K being the number of SAEM iterations) can be. We are investigating a way to calculate the MAP once at the beginning and then apply a single step of gradient descent at each SAEM iteration in order to slowly move the MAP estimate after each update of the posterior distribution towards a better approximation.

References

A. Gunawardana, W. Byrne. Convergence theorems for Generalized Alternating Minimization procedures. (English). 2005.

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

Celeux and Diebolt. The SEM Algorithm: a Probabilistic Teacher Algorithm Derived from the EM Algorithm for the Mixture Problem. (English). 1985.

Dempster, Laird and Rubin. Maximum likelihood from incomplete likelihood data via EM algorithm (with discussion). (English). *J. Roy. Statist. Soc. Ser.*, 1977.

E.Kuhn, M. Lavielle. Coupling a Stochastic Approximation version of EM with an MCMC Procedure. (English). 2015.

E.Moulines, O. Cappe. Online EM Algorithm for Latent Data Models. (English). 2007.

Lavielle, M. A stochastic algorithm for parametric and non-parametric estimation in the case of incomplete data . (English). 1993.

R. Ranganath, S. Gerrish and Blei, D.M. Black Box Variational Inference. (English). 2013.

Y., Wang. Derivation of various NONMEM estimation methods.. (English). 2007.

Appendices

A. Accelerate a stochastic version of the EM algorithm using this new proposal

Using (Y., 2007) first order conditional estimation approximation of the posterior distribution as an independent proposal in our MCMC procedure, can actually accelerate our algorithm, the Stochastic Approximation of the EM (SAEM) in the context of PK-PD models. In the sequel, we will study the behavior of this algorithm on a PK-PD model.

A.1. Yield Model

The data comes from 37 winter wheat experiments carried out between 1990 and 1996 on commercial farms near Paris, France. Each experiment was from a different site. Two soil types were represented, a loam soil and a chalky soil. Common winter wheat varieties were used. Each experiment consisted of five to eight different nitrogen fertiliser rates, for a total of 224 nitrogen treatments. Nitrogen fertilizer was applied in two applications during the growing season. For each nitrogen treatment, grain yield (adjusted to 150 g.kg-1 grain moisture content) was measured. In addition, end-of-winter mineral soil nitrogen (NO3- plus NH4+) in the 0 to 90 cm layer was measured on each site-year during February when the crops were tillering. Yield and end-of-winter mineral soil nitrogen measurements were in the ranges 3.44-11.54 t.ha-1 , and 40-180 kg.ha-1 respectively.
In this problem the sites are denoted by the index "i" and are the individuals in the dataset, the predictor is the dosage, the response is the grain yield and the covariate is the soil nitrogen.

We use a Linear Plateau model here and the the structural model is:

$$f(\psi_i) = \begin{cases} (Y_{max})_i + B_i * (t_i - (X_{max})_i) + \epsilon_i, & \text{if } t \geq (X_{ma} \\ (Y_{max})_i + \epsilon_i, & \text{otherwise} \end{cases} \tag{17}$$

Where $\psi_i = ((X_{max})_i, (Y_{max})_i, B_i)$ and $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$.

The SAEM algorithm consists in finding the maximum likelihood parameter estimate by first simulating the latent variable, here the variable ψ_i according to the posterior distribution $P_{\psi_i|y_i,\theta_k}$ at an iteration k and for all individuals i. Here, when the posterior is intractable we use an MCMC procedure. As said above, we are doing the MCMC over the random variable η_i such that:

$$\psi_i = \psi_{pop} + \eta_i \tag{18}$$

With ψ_{pop} being the fixed parameters and $\eta_i \sim \mathcal{N}(0, \Omega)$ the random effects.
The MLE problem consists then in finding the vector parameter $((X_{max})_{pop}, (Y_{max})_{pop}, B_{pop}, \omega_{X_{max}}, \omega_{Y_{max}}, \omega_B, \sigma)$
The reference algorithm, called the Random Walk Metropolis SAEM (RWM) consists in doing 6 RWM transition kernel proposing a candidate with the following kernel:

$$(\eta_i)_{candidate} \sim \mathcal{N}((\eta_i)_{current}, \Omega) \tag{19}$$

The first algorithm we implemented consists in doing 6 transition kernel proposing a candidate with the new kernel as follow:

$$(\eta_i)_{candidate} \sim \mathcal{N}(\hat{\eta}_i, (\nabla_\psi f(\hat{\psi}_i)' \Sigma^{-1} \nabla_\psi f(\hat{\psi}_i) + \Omega^{-1})^{-1}) \tag{20}$$

Where $\hat{\psi}_i = \psi_{pop} + \hat{\eta}_i$ is the MAP estimate that needs to be calculated at each SAEM parameter update.
Figure A.1 shows how fast this new algorithm is compared to the reference.

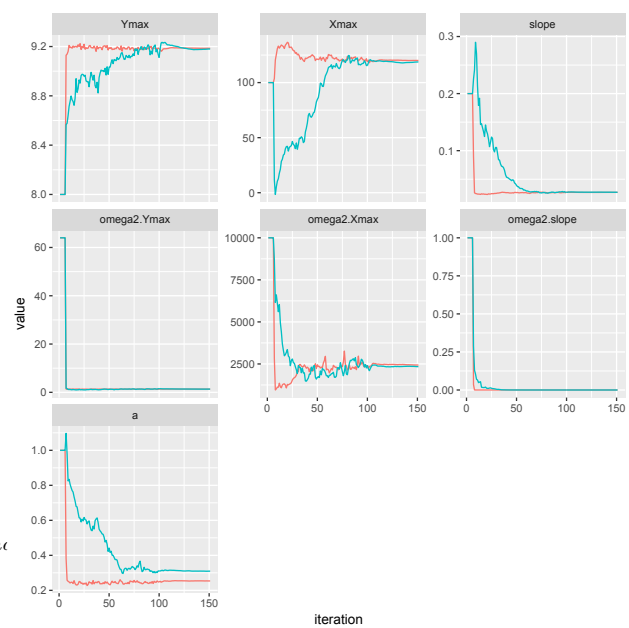

Figure 2. SAEM parameter estimates. In green the parameter estimate of the reference SAEM and in red the accelerated one using the new proposal

The calculation of the MAP being very costly we suggest an algorithm that consists in doing 6 transition kernel proposing a candidate with the new kernel as above during the first three iterations and then switching back to the regular RWM SAEM. See Figure A.1 to compare the reference SAEM and the new two algorithms.

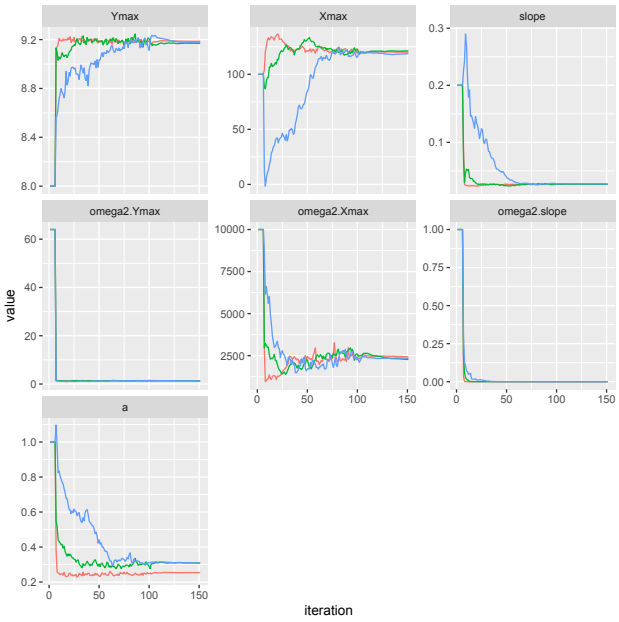

Figure 3. SAEM parameter estimates. In blue the parameter estimate of the reference SAEM, in red the algorithm computing the MAP at each iteration and in green the accelerated one using the new proposal for the first 3 iterations and switching to the regular one beyond

A.2. Discussion

Another work in progress towards this acceleration of the SAEM algorithm consists in calculating the MAP once during the first iteration of the SAEM and then moving the maximum a posteriori estimate towards the next one using a gradient descent step. If $\hat{\psi}_1$ is the MAP estimate at the first iteration of the SAEM, when $\theta = \theta_1$, then for iteration k

$$\hat{\psi}_k = \hat{\psi}_{k-1} + \rho \nabla p(\Psi | y, \theta_{k-1}) \tag{21}$$

This has been implemented but the results are not satisfying. The tuning of the stepsize ρ and the number of gradient descent steps could improve the convergence properties.