INSTITUT NATIONAL DE LA STATISTIQUE ET DES ETUDES ECONOMIQUES
Série des Documents de Travail du CREST
(Centre de Recherche en Economie et Statistique)

n° 2002-13

Laplace Expansions in MCMC Algorithms for Latent Variable Models

C. GUIHENNEUC[1]
J. ROUSSEAU[2]

[1] INSERM U170, 16 Avenue P-V Couturier, 94807 Villejuif Cédex and Université Paris V.
Email : guihenneuc@biomedicale.unive-paris5.fr

[2] CREST, Timbre J340, 3 Avenue Pierre Larousse, 92245 Malakoff Cédex, and Université Paris V.
Email : judith.rousseau@ensae.fr

Laplace expansions in MCMC algorithms for latent variable models

By

Chantal Guihenneuc*[1] and Judith Rousseau[2]

[1] INSERM U170, 16 avenue P-V Couturier, 94 807 Villejuif Cedex

and Laboratoire de Statistique, Université Paris V, France

guihenneuc@biomedicale.univ-paris5.fr

[2] Laboratoire de Statistique, Université Paris V,

45 rue des St-Pères, 75 006 Paris and CREST, Malakoff, France.

rousseau@ensae.fr

Laplace expansions in MCMC algorithms for latent variable models

By

Chantal Guihenneuc and Judith Rousseau

Résumé

Il est nécessaire, dans les modèles de chaines de Markov cachées, d'avoir recours à des méthodes de simulations, de type Monte Carlo par Chaine de Markov (MCMC), pour obtenir des estimateurs bayésiens ou des régions de confiances bayésiennes. Le nombre de paramètres à simuler étant en general, de l'ordre du nombre d'observations - lorsque l'on considère la représentation par variables latentes - les algorithmes MCMC deviennent rapidement assez lents. Dans cet article, nous proposons une manière d'accélerer l'algorithme en diminuant le nombre de paramètres à simuler. Pour ce faire, nous utilisons une approximation de Laplace pour intégrer les paramètres de nuisances, à chaque itération de l'algorithme. La loi cible est donc modifiée. Nous démontrons que, en variation totale, la vraie loi cible et son approximation sont très proches, lorsque le nombre d'observations est grand (l'approximation est en $O(n^{-1})$). Pour illustrer ce résultat théorique nous effectuons des simulations. Les simulations montrent notamment que l'approximation que nous proposons se comporte extrêmement bien, tout au moins dans les exemples considérés.

Summary

To obtain Bayes estimates such as the posterior mean or bayesian confidence regions, in Hidden Markov models , it is necessary to simulate the posterior distribution using a MCMC algorithm. These algorithms get slower as the number of observations increases, specially in this case of latent variables. To improve the convergence of the algorithm, we propose to decrease the number of parameters to simulate at each iteration by using a Laplace approximation on the nuisance parameters. We therefore study, theoretically the impact that such an approximation has on the target posterior distribution. We prove that the distance between the true target distribution and the approximated one becomes essentially of order $O(n^{-1})$ as the number of observations increases. A simulation study illustrates the theoretical results. It turns out, that the approximated algorithm behaves extremely well, at least in the example considered in the paper, which is driven by a study on HIV patients.

1 Introduction

1.1 Motivations

As the complexity of the models covered by statistical inference increases, the need of new computional tools gets increasingly pressing. In this respect, Markov chain Monte Carlo (MCMC) methods have been widely developped in the last decade and have enhanced the use of complex models in different types of applications, typically using bayesian inference, see Robert and Casella (1999). In a bayesian approach, samples produced by MCMC algorithms are quite appropriate to approximate many aspects of the posterior distributions using ergodic averages. Hidden Markov models (HMM) constitute a widely studied class of complex models. They have been used in many areas as a convenient representation of weakly dependent heterogeneous phenomena. They are specific latent variable models where the completed model is directed by an unobserved Markov process S. When the state space of S is continuous, these models are usually called state space models such as in Econometrics, in stochastic volatility models (Shephard and Pitt (1997), Hamilton (1989), Chib (1996)) or in Signal processing (Hodgson (1999), Rabiner (1989)). HMM's also have a large ranging number of applications, when the state space of S is discrete : in Genetics as DNA sequence modelling (Rabiner (1989), Durbin et al (1998), Muri (1998)) and in medical areas (Guihenneuc et al (2000), Kirby and Spiegelhalter (1994)). This work has been motivated by biomedical applications but can be generalised to other domains of applications. In medical area, multistate models, i.e. finite state space HMM's, have been increasingly used to model and to characterize the progression of diseases. The definition of the states is generally based on the discretisation of continuous markers as the decline of CD4 cell counts for HIV patients. These markers are usually subject to great variability, so that the observed trajectories give a noisy representation of the true trajectories. The states are therefore considered as unobserved, leading to a hidden Markov modelisation.

Traditionally, there are essentially two ways to calculate posterior quantities of interest: asymptotic expansions or simulations. In HMM's, the number of parameters is proportional to the number of observations and therefore, asymptotic expansions such as Laplace expansions are not valid. It is then necessary to compute the posterior distribution via a MCMC algorithm. However, the larger the number of observations, the larger the number of parameters and thus, the longer we have to run the algorithm to compute the posterior distribution, for instance.

In the same time, had we been able to use an asymptotic approximation, such as a Laplace approximation of the posterior distribution, the better it would have been. We therefore study in this paper a way to combine those two approaches in order to accelerate the MCMC algorithm.

1.2 The HMM model

Denote X the observations and S the hidden states and denote $\mathcal{L}$ any distribution, so that $\mathcal{L}(\lambda|X, S, \underline{\theta})$ represents the conditional distribution of λ given $X, S, \underline{\theta}$, for instance. We assume, as is traditionally the case in HMM's, that the observations X, conditionally on latent variables S, are independent and distributed according to some family of distributions indexed by a parameter $\underline{\theta}$. The distribution of the latent process S depends on a parameter λ. Assume that λ is the parameter of interest, and let π and h be the priors on $\underline{\theta}$ and λ respectively. The aim is thus to simulate the posterior distribution of λ given X or of (λ, S) given X. The hierarchical structure of HMM's induces a natural Gibbs algorithm, which would be constructed as follows:

1. $\lambda^t \sim \mathcal{L}(\lambda|X, S^{t-1}, \underline{\theta}^{t-1})$

2. $S^t \sim \mathcal{L}(S|X, \lambda^t, \underline{\theta}^{t-1})$

3. $\underline{\theta}^t \sim \mathcal{L}(\underline{\theta}|X, S^t, \lambda^t)$.

We denote M_0 this algorithm, which we call the Gibbs algorithm. It often happens that the correlations between S and $\underline{\theta}$ are very strong, in other words that the knowledge of S (and X) implies a good knowledge of $\underline{\theta}$. In this case, the Gibbs algorithm has poor mixing properties. This phenomenon occurs in particular in medical models, such as the HIV model proposed by Guihenneuc et al (2000) or in the Ion channel model considered by Hodgson (1999). It is then important to get rid of $\underline{\theta}$, which is a nuisance parameter.

In this paper, we consider the following type of HMM's : the data consist of observed values X_{ij} where i indexes the individual and j the follow-up point, $1 \leq i \leq n$, $1 \leq j \leq n_i$. The X_{ij}'s are independent conditionally on the unobserved random variables $S_{ij} = s$, with distribution P_{θ_s}. We assume that P_{θ_s} has a density with respect to Lebesgue measure denoted by $f_{\theta_s}(X)$, $s = 1, ..., k$, where $\theta_s \in \Theta_s$, and Θ_s is an open subset of $\mathbb{R}^{p_s}$. The densities f_{θ_s} may differ by more than the

parameter θ_s, they can belong to different parametric families. The latent process is defined as follows : the individuals are independent, and for each indiviual i, the S_{ij}'s, $j = 1, ..., n_i$, are continuous time markov processes depending on a parameter $\lambda \in L$. This structure is driven by medical applications but the markov property is actually not necessary and we could use any kind of dependence structure to model S, as can be seen in Section 2. In this regard, a motivating illustration of such HMM concerns the HIV model proposed by Guihenneuc *et al.* (2000). There, the latent process S represents the health progression throughout 6 transient unobserved states. The observed process is a biological marker (CD4 cell counts), which has a great within individual variability. In this model, a seventh state is considered, which corresponds to the the AIDS's status, based on clinical symptoms. This state is therefore perfectly observable. S is modeled by a Markov process on $\{1, ..., 7\}$, for which λ_{ij} represents the transition rate to state j starting from state i. The conditional distributions for S are therefore given by:

$$pr(S_{ij}|\lambda, S_{ij-1}) = (\exp \Lambda dt_{ij})_{S_{ij-1}S_{ij}} \quad \text{and} \quad pr(S_{i1} = s) = \delta(s) > 0, \tag{1}$$

where Λ is the infinitesimal transition matrix. The error process is supposed to be Gaussian.

Let $\underline{\theta} = (\theta_1, ..., \theta_k) \in \Theta = \Theta_1 \times \cdots \Theta_k$, $\pi(\underline{\theta})$, $h(\lambda)$ the priors on Θ and L, respectively. We also denote $\pi_s(\theta_s)$ the marginal prior of θ_s, $s = 1, ..., k$. If we want to characterize the progression of the hidden process S, λ is then the parameter of interest, if we want to reconstruct the individual trajectories, then S is the parameter of interest, we can also consider (λ, S) as the parameter of interest, but $\underline{\theta}$ is generally a nuisance parameter. We are thus interested in the posterior distribution of the parameter of interest, for instance $\pi(\lambda|X)$, the posterior density of λ, to determine Bayes estimates such as the posterior mean or the posterior median of each λ_{ij}, or to construct confidence regions, such as HPD regions. This posterior distribution is obviously not available in close form and we must simulate it, using an MCMC algorithm. Indeed, the posterior distribution of interest has the following form when λ is the parameter of interest:

$$\begin{aligned} \pi(\lambda|X) &= \sum_{S \in \mathcal{S}} \pi(\lambda, S|X) \\ &\propto \sum_{S \in \mathcal{S}} \int_{\Theta} f(X|\underline{\theta}, S) d\pi(\underline{\theta}) pr(S|\lambda) h(\lambda), \end{aligned}$$

where $\mathcal{S}$ is the set of all possible cases for S. Although $\mathcal{S}$ is finite when the number of hidden states is finite, it becomes quickly very large, when the number of obser-

vations increases. Moreover, in most cases, the integral of the right hand side of the above equality, i.e. the integral over $\underline{\theta}$, cannot be obtained in close form. Therefore, it is usually necessary to use a Gibbs algorithm in the form of M_0. Within this structure, there are ways to improve the classical Gibbs algorithm, in particular when a Hasting-Metropolis step is needed as would often happen when simulating λ, conditionally on all the other variables, see for instance Tierney and Mira (1999). The improvement, here, would however only be on the Hasting-Metropolis step, which is not the problem we focus on.

In this paper, we propose an other way to accelerate the algorithm, while changing slightly the target distribution. An other interest that, we believe, such work might induce, is as a first step in studying how an asymptotic expansion can reasonably be done inside steps of an MCMC algorithm, and the effect it has globally on the simulations. Usually Laplace expansions are used to build up good proposal. Here the aim is therefore quite different.

The paper is constructed as follows. In Section 2 we present our approximated algorithm and how it is constructed. We then present theoretical results which validate such an approach, since we prove that our limiting target distribution is close to the true target distribution as the number of observations goes to infinity. In Section 3 we present simulations to illustrate the theoretical results. These simulations shed lights on some features of the Gibbs algorithm.

2 Laplace expansion

2.1 The Laplace algorithm

To begin with, we define some notations used throughout the paper: Let $l_n(\underline{\theta})$ be the log-likelihood, conditional on S, i.e. the log-likelihood of the completed model and let $\hat{\underline{\theta}} = (\hat{\theta}_1, ..., \hat{\theta}_k) = \underline{\hat{\theta}}(X, S)$ be the conditional maximum likelihood estimate. The differentiation operator will be denoted by D, i.e. for any function $g : \mathbb{R}^p \longrightarrow \mathbb{R}^q$ with $p, q \geq 1$, $D^\nu g(z)$ is the ν-th derivative of g with respect to z, where $\nu = (\nu_1, ..., \nu_p)$, $\nu_i \geq 0$. We also denote $|\nu| = \nu_1 + ... + \nu_p$. For simplicity's sake we also denote $Dg(z)$ the vector of first derivatives and $D_2 g(z)$ the matrix of second derivatives of g. Let J be the non normalized empirical Fisher information matrix of the completed model, i.e. $J = -D_2 l_n(\hat{\underline{\theta}})$ and let $|J|$ be its determinant. Finally, $||\mu_1 - \mu_2||_{TV}$ denotes the total variation norm of $\mu_1 - \mu_2$ and $\psi = \log \pi(\theta)$; recall that n_s denotes the number of observations in the state s, when the vector S

is known.

The true marginal distribution of $(\lambda, S)|X$ is given by :

$$\pi(\lambda, S|X) = \frac{\left\{\int_\Theta \prod_{s=1}^k \prod_{S_{ij}=s} f(X_{ij}|\theta_s)\pi(\underline{\theta})d\underline{\theta}\right\} pr(S|\lambda)h(\lambda)}{\int_L \sum_S \left\{\int_\Theta \prod_{s=1}^k \prod_{S_{ij}=s} f(X_{ij}|\theta_s)\pi(\underline{\theta})d\underline{\theta}\right\} pr(S|\lambda)h(\lambda)d\lambda}.$$

$\underline{\theta}$ is a nuisance parameter, that we want to avoid simulating. We therefore propose to replace the integral over $\underline{\theta}$, which is a finite dimensional parameter, by its Laplace approximation. The approximate marginal distribution of $(\lambda, S)|X$ would then be :

$$\hat{\pi}(\lambda, S|X) = \frac{\left\{\prod_{s=1}^k \prod_{S_{ij}=s} f(X_{ij}|\hat{\theta}_s)\pi(\hat{\underline{\theta}})J^{-1/2}\right\} pr(S|\lambda)\pi(\lambda)}{\int_L \sum_S \left\{\prod_{s=1}^k \prod_{S_{ij}=s} f(X_{ij}|\hat{\theta}_s)\pi(\hat{\underline{\theta}})J^{-1/2}\right\} pr(S|\lambda)\pi(\lambda)d\lambda}.$$

Because of the structure of the model, we have :

$$l_n(\underline{\theta}) = \sum_{s=1}^k \sum_{S_{ij}=s} \log f(X_{ij}|\theta_s) = \sum_{s=1}^k l_s(\theta_s).$$

The conditional model can therefore be separated into k submodels that work like standard independent and identically distributed models, for which the Laplace expansion is well known, see for instance Kass, Tierney and Kadane (1989).

Denote also, J_{n_s} the normalized conditional empirical Fisher information matrix associated with the group s, : $J_{n_s} = -n_s^{-1} D^2 l_s(\hat{\theta}_s)$, $s = 1, ..., k$ and $|J_{n_s}|$ its determinant. Let $g(X|S)$ be the marginal conditional density of X given S and $\hat{g}(X|S)$ its Laplace approximation, i.e.

$$g(X|S) = \int_\Theta e^{l_n(\underline{\theta})} \pi(\underline{\theta})d\underline{\theta}$$

and

$$\hat{g}(X|S) = (2\pi)^{p_1+...+p_k} \prod_{s=1}^k n_s^{-p_s/2} |J_{n_s}|^{-1/2} \prod_{S_{ij}=s} f(X_{ij}|\hat{\theta}_s)\pi(\hat{\underline{\theta}}).$$

We thus would have :$\hat{\pi}(\lambda, S|X) \propto \hat{g}(X|S)pr(S|\lambda)h(\lambda)$. However, since this approximation will be used at each iteration of the Gibbs algorithm, there will be cases, i.e. S, for which the approximation is quite poor. To avoid weird effects that could be caused with such S's we use instead as the approximated density of X given S : $\tilde{g}(X|S) = \mathbb{I}_B(S, X)\hat{g}(X|S)$, where $B = \{(X, S); g(X|S) = \hat{g}(X|S)(1 + O(n^{-a}))\}$, for some $a \in (1/2, 1)$, and where n is the number of individuals. a will be chosen as close to 1 as possible. We then have as the limiting target distribution : $\tilde{\pi}(\lambda, S|X) \propto \tilde{g}(X|S)pr(S|\lambda)h(\lambda)$. The new algorithm has thus the following structure : at the t-th iteration,

1. $\lambda^t \sim \pi(\lambda|X, S^{t-1})$ which is the true one,

2. $S^t \sim \tilde{\pi}(S|X, \lambda^t)$.

We denote M_L this algorithm which we call the Laplace algorithm. To validate this algorithm, we thus need to make sure that its target distribution is close to the true one, as n goes to infinity. The idea is the following, the Laplace expansion, ensures us in regular cases, that $g(X|S) = \tilde{g}(X|S)(1 + O(n^{-a}))$ for all S such that $(X, S) \in B$. The difference between the true and the approximated distribution of $(\lambda, S|X)$ (in total variation), will then be essentially of order n^{-a} except on B^c, the complementary set of B, which will be forgotten by our algorithm. To control this difference, we thus need to control $pr(B^c|X)$, which is done in the following section.

2.2 Validity of the approximation

In this Section, we present results ensuring that the approximated target distribution and the true target distributions are close to one an other. As was said in the previous Section, the conditional model (of X given S) can be separated into k independent and identically distributed models. The Laplace approximation, will then, mainly be a Laplace approximation in each submodel. To make sure that this approximation is good, we thus need to have enough observations in each model, i.e. in each state s, $s \in \{1, ..., k\}$. To do so we consider the following assumption on the underlying unobserve process S :

[H] : For all $s \leq k$, $i = 1 \cdots n$,

$$pr(\text{for some} \quad j \leq n_i; S_{ij} = s|\lambda) \geq c_0(\lambda) > 0, \quad \text{such that} \int_L c_0(\lambda)h(\lambda)d\lambda < \infty.$$

This hypothesis is not strong. In particular in the HIV example, we have, $pr(\text{for some} \quad j \leq n_i; S_{ij} = s|\lambda) > \delta(s)$, so that **[H]** is satisfied.

The results that are stated in this section are written for non compacts Θ. In the compact case things are simpler, we point out, throughout this section how things would simplify in the compact case. In particular, the assumptions **[A1]**-**[A6]** given below can be slightly simplified in the compact case.

In the following, we denote $E_\theta\{h(X)\}$ the expectation of $h(X)$ under P_θ

[A1] In each submodel, $s = 1, ..., k$, the log-likelihood, $\log f_{\theta_s}(x)$, is 4 times continuously differentiable in θ_s and satisfies: for $\nu = (\nu_1, ..., \nu_p) \in \mathbb{N}^p$, such that $|\nu| \leq 4$, there exists $\delta > 0$ and there exists $q > 2$ for which

$$\int E_{\theta_s} \left(\sup_{|\theta_s - \theta'_s| < \delta} |D^\nu \log f_{\theta'_s}(x)|^q \right) \pi_s(\theta_s)d\theta_s < \infty,$$

where π_s denotes the marginal prior density of θ_s.

[**A2**] In each submodel $s = 1, ..., k$, the information matrix $I_s(\theta_s)$ is definite positive, for all $\theta_s \in \Theta_s$, where

$$I_s(\theta_s) = -\int \frac{\partial^2 \log\left(f_{\theta_s}(x)\right)}{\partial\theta_s \partial\theta_s^t} f_{\theta_s}(x)dx,$$

is the Fisher information matrix per observation associated with the density $f_{\theta_s}(x)$.

[**A3**] For all $s = 1, ..., k$, there exists $0 < c < 1/2$ such that :

$$\int_{\Theta_s} P_{\theta_s}\left(|\hat{\theta}_s - \theta_s| > n_s^{-c}\right)\pi_s(\theta_s)d\theta_s \leq n_s^{-2}.$$

[**A4**] Let $\underline{\theta}_0 = (\theta_{0,1}, ..., \theta_{0,k})$ with $\theta_{0,s} \in \Theta_s$, $s = 1, ..., k$. For all $s = 1, ..., k$,

$$\int_{\Theta} pr\{\theta_s; |\theta_s - \theta_{0,s}| > n_s^{-c}, K_s(\theta_{0,s}, \theta_s) < 2\log n_s/n_s\}\pi_s(\theta_{0,s})d\theta_{0,s} \leq Cn_s^{-2-a},$$

and

$$\int_{\Theta} pr\left\{\theta_s; |\theta_s - \theta_{0,s}| > n_s^{-c}, K_s^2(\theta_{0,s}, \theta_s) < \frac{(2+a)\log n_s}{n_s} M_{2,s}(\theta_{0,s}, \theta_s)\right\}\pi_s(\theta_{0,s})d\theta_{0,s} \leq Cn_s^{-2-a},$$

where $K_s(\theta_{0,s}, \theta_s) = E_{\theta_{0,s}}\left(\log f_{\theta_{0,s}}(X) - \log f_{\theta_s}(X)\right)$ and

$$M_{2,s}(\theta_{0,s}, \theta_s) = [E_{\theta_{0,s}}\{(\log f_{\theta_s} - \log f_{\theta_{0,s}})^2\}]^{1/2}[E_{\theta_s}\{(\log f_{\theta_s} - \log f_{\theta_{0,s}})^2\}]^{1/2}.$$

[**A5**] There exists $0 < t < c$, such that $qt \geq 2$, with q defined in assumption [A1] and c in assumption [A3], satisfying: $pr\left(|I_s(\theta_s)|^{-1} > n_s^t/2\right) < n_s^{-2}$.

[**A6**] $\pi(\underline{\theta}) > 0$, for all $\underline{\theta} \in \Theta$, and π is twice continuously differentiable and satisfies the following conditions : for all $s \leq k$,

$$pr\left(\sup_{|\theta_s - \theta_{0,s}| < n_s^{-c}} |D\log\pi(\theta_s)| > n_s^t\right) \leq n_s^{-2},$$

and

$$pr\left(\sup_{|\theta_s - \theta_{0,s}| < n_s^{-c}} |D^2\log\pi(\theta_s)| > n_s^{2t}\right) \leq n_s^{-2},$$

with t defined in assumption [A5] and c in assumption [A3].

The first four conditions are usual in Laplace expansions. The fourth condition is expressed quite generally, as it is done in Bickel and Ghosh (1990). In regular models, it often requires fairly weak conditions on the prior, such as moment conditions. We have chosen this general expression because, depending on the model, the appropriate types of assumptions could be fairly different, even for very smooth models like the Gaussian (μ, σ) distribution, for instance, with $(\mu, \sigma) \in \mathbb{R} \times \mathbb{R}^+$, conditions such as those proposed by Ibragimov and Hasminskii (1981) are not really appropriate. Conditions [**A5**] and [**A6**] are conditions on the prior, and are needed to control the behaviour of the Laplace expansion when the parameter goes to the boundary of the set. When Θ is compact, it is enough to assume that terms are bounded, but when Θ is not compact, it is necessary to control the integrals. In the Gaussian case however, as in the HIV example, these conditions reduce to very simple conditions on the prior density, in the form : $pr(\sigma > cn/\log n^2) \leq n^{-2}$. Condition [**A6**] is equivalent, in the non compact case, to the type D_4 of priors defined in Ghosh *et al.* (1982).

In the previous Section we had defined B in a vague way : $B = \{(X,S); g(X|S) = \hat{g}(X|S)(1 + O(n^{-a}))\}$, for some $a > 0$. To be able to implement the algorithm, and to obtain a rigorous proof on its validity we now give an explicit expression of B: let $t = 2/q \in (0,1)$ with q defined in assumption [**A1**], $\beta \in (1/2, 1)$, let A_s be defined by : $A_s = \{\theta_s; l_s(\theta_s) - l_s(\hat{\theta}_s) > -\log n_s\} \cap \{|\theta_s - \hat{\theta}_s| > n^{-c}\}$, then

$$B(\beta,t,c) = \Big\{(X,S); n_s > n^\beta, pr(A_s) \leq n_s^{-1}, |D^\nu l_n(\hat{\theta}_s)| \leq n_s^{1+t}, \quad |\nu| \leq 3, |J_s| \geq n_s^{-t},$$
$$\sup_{|\theta_s - \hat{\theta}_s| < n_s^{-c}} |D^\nu l_n(\theta_s)| \leq n_s^{1+t}, |\nu| = 4, \sup_{|\theta_s - \hat{\theta}_s| < n_s^{-c}, \forall s} |D_2\psi(\underline{\theta})| \leq n_s^{2t}, D\psi(\hat{\underline{\theta}}) \leq n_s^t, \forall s \Big\},$$

where c is defined in assumption [**A3**]. Condition [**H**] implies that β can be as close to 1 as we want.

When B is defined as such, $a = \beta(1-3t)$. This definition of B could be made simpler in the compact case, in particular, we could drop the last two constrains on $\psi = \log \pi$.

We now state the main result of this section :

Theorem 1 *If* [**H**] *is satisfied and if the hypotheses* [**A1**] $-$ [**A7**] *are satisfied, the approximate target distribution is close to the true one in the following sense:*

$$||\hat{\pi}(\lambda, S|X) - \pi(\lambda, S|X)||_{TV} \leq Cn^{-a}, \tag{2}$$

except on a small set i.e.

$$P_{m(X)}(||\hat{\pi}(\lambda, S|X) - \pi(\lambda, S|X)||_{TV} > Cn^{-a}) \leq n^{-1},$$

where $P_{m(X)}$ denotes the probability under the marginal distribution of X.

By imposing stronger conditions, in particular by imposing bounds in the form n^{-h} for h greater than what is already imposed in assumptions **[A3]-[A6]**, we can obtain a better bound for $P_{m(X)}(||\hat{\pi}(\lambda, S|X) - \pi(\lambda, S|X)||_{TV} > Cn^{-a})$, as will appear clearly in the proof.

proof of Theorem 1: We have, for any borel set A on $L \times \mathcal{S}$:

$$
\begin{aligned}
&|\hat{\pi}(A|X) - \pi(A|X)| \\
&= \left| \frac{\sum_S \int_\lambda \mathbb{I}_A(\lambda, S)\mathbb{I}_B(X, s)\hat{g}(X|S)pr(S|\lambda)h(\lambda)d\lambda}{\sum_S \mathbb{I}_B(X, S)\hat{g}(X|S)pr(S)} - \frac{\sum_S \int_\lambda \mathbb{I}_A(\lambda, S)g(X|S)pr(S|\lambda)h(\lambda)d\lambda}{\sum_S g(X|S)pr(S)} \right| \\
&\leq \frac{\sum_S \int_\lambda \mathbb{I}_B(X, s)|\hat{g}(X|S) - g(X|S)|pr(S|\lambda)h(\lambda)d\lambda}{\sum_S \mathbb{I}_B(X, S)\hat{g}(X|S)pr(S)} \\
&\quad + \left| \frac{\sum_S \int_\lambda \mathbb{I}_A(\lambda, S)\mathbb{I}_B(X, s)g(X|S)pr(S|\lambda)h(\lambda)d\lambda}{\sum_S \mathbb{I}_B(X, S)\hat{g}(X|S)pr(S)} - \frac{\sum_S \int_\lambda \mathbb{I}_A(\lambda, S)g(X|S)pr(S|\lambda)h(\lambda)d\lambda}{\sum_S g(X|S)pr(S)} \right| \\
&\leq n^{-a} + pr(B^c|X) + \frac{\sum_S \int_\lambda \mathbb{I}_A(\lambda, S)\mathbb{I}_B(X, s)g(X|S)pr(S|\lambda)h(\lambda)d\lambda}{\sum_S g(X|S)pr(S)} \times \\
&\quad \left| \frac{\sum_S g(X|S)pr(S) - \sum_S \mathbb{I}_B(X, S)\hat{g}(X|S)pr(S)}{\sum_S \mathbb{I}_B(X, S)\hat{g}(X|S)pr(S)} \right| \\
&\leq 2n^{-a} + pr(B^c|X) + \frac{\sum_S \mathbb{I}_{B^c}(X, S)g(X|S)pr(S)}{\sum_S \mathbb{I}_B(X, S)\hat{g}(X|S)pr(S)}.
\end{aligned}
$$

When $(X, S) \in B$, $(1 - n^{-a})^{-1}g(X|S) \geq \hat{g}(X|S) \geq (1 + n^{-a})^{-1}g(X|S)$, so,

$$
\begin{aligned}
|\hat{\pi}(A|X) - \pi(A|X)| &\leq 2n^{-a} + \frac{\sum_S \mathbb{I}_{B^c}(X, S)g(X|S)pr(S)}{\sum_S \mathbb{I}_B(X, S)g(X|S)pr(S)}(1 + n^{-a}) \\
&\leq 2n^{-a} + (1 + n^{-a})\frac{pr(B^c|X)}{1 - pr(B^c|X)}.
\end{aligned}
$$

Therefore to prove Theorem 1, we just need to prove that

$$
P_{m(X)}\{pr(B^c|X) > n^{-a}\} \leq Cn^{-1}. \tag{3}
$$

We use Markov's inequality :

$$
\begin{aligned}
&P_{m(X)}\{pr(B^c|X) > n^{-a}\} \leq n^a pr(B^c) \\
&\quad = n^a pr(B^c \cap \{\exists s; n_s \leq n^t\}) + n^a pr(B^c \cap \{\forall s; n_s > n^t\}) \qquad (4)
\end{aligned}
$$

Assumption **[H]** implies that the first term of the right hand side of (4) is bounded by n^{-h}, for all $h > 0$, when n is large enough. We now consider the second term of the right hand side of (4): $n^a E\{pr(B_1|S, \underline{\theta}, \lambda)\}$, where $B_1 = B^c \cap \{n_s > n^t, s = 1, ..., k\}$. Inequality (3) will therefore be satisfied if

$$
pr(B_1|S, \underline{\theta}_0, \lambda_0) \leq \frac{c(\underline{\theta}_0)}{n^{1+a}}, \quad \text{with} \quad \int_\Theta c(\underline{\theta})\pi(\underline{\theta})d\theta < \infty. \tag{5}
$$

Let $B_s = \{X_s; g_s(X_s) = \hat{g}_s(X_s)(1 + n_s^{-a/\beta})\}$, with $X_s = (x_1, ..., x_{n_s})$ are n_s independent and identically distributed random variable distributed according to $f_{\theta_{0,s}}$,

$$g_s(X_s) = \int_{\Theta_s} \prod_{i=1}^{n_s} f_{\theta_s}(x_i)\pi_s(\theta_s)d\theta_s,$$

and $\hat{g}_s$ is its formal Laplace expansion. In other word, B_s is the set on which the Laplace expansion is correct, conditionally on S, in the sub-model s. The conditional independence structure implies that (5) will be obtained if, for all $s \leq k$,

$$P_{\theta_{0,s}}(B_s) \leq \frac{c(\theta_{0,s})}{n_s^{(1+a)/\beta}}, \quad \text{with} \quad \int_{\Theta} c_s(\theta_s)\pi_s(\theta_s)d\theta_s < \infty. \tag{6}$$

We can therefore work in each submodel independently, and drop the s, for simplicity's sake.

In the compact case, i.e. if Θ is compact or equivalently if each Θ_s is compact, Ghosh *et al.* (1982) have obtained conditions on the model, i.e. on f_θ and on π to be able to integrate out the Laplace expansion with respect to π, see also Bickel and Ghosh (1990). Now, if Θ is not compact, as is typically the case in medical studies, no such result exists. Our definition of B and the hypothesis **[A1]**-**[A6]** are defined for such non compact sets. These assumptions can be relaxed slightly in the compact case, see Ghosh *et al.* (1982) and Bickel and Ghosh (1990).

In the general case, dropping the index s, we have :

$$\begin{aligned} P_{\theta_0}(B) \quad \leq \quad & P_{\theta_0}\{\pi(A_n) > n^{-1}\} + P_{\theta_0}\{\inf x'Jx/(x'x) \leq n^{-t}\} \\ & + \sum_{|\nu|=2}^{3} P_{\theta_0}(|D^\nu l_n(\hat{\theta})| \geq n^{1+t}) + P_{\theta_0}\left(\sup_{|\theta-\hat{\theta}|<n^{-c}} |D^4 l_n(\theta)| \geq n^{1+t}\right) \\ & + P_{\theta_0}\left(\sup_{|\theta-\hat{\theta}|<n^{-t}} |D_2\psi(\underline{\theta})| > n^{2t}\right) + P_{\theta_0}\left(D\psi(\hat{\theta}) > n^t\right). \end{aligned} \tag{7}$$

Hypothesis **[A4]** implies that :

$$\int_{\Theta} P_\theta\left(|\hat{\theta} - \theta| > n^{-c}\right)\pi(\theta)d\theta \leq n^{-2}$$

so we only need to work on $\{\theta; |\theta - \theta_0| \leq 2n^{-c}\}$. The last two terms of the right hand side of (7) are bounded by n^{-2} using hypothesis **[A6]**. We now consider the first term of the inequality (7). In Appendix 1, we prove that

$$P_{\theta_0}\left[\int_{|\theta-\theta_0|>n^{-c}} \exp\{l_n(\theta) - l_n(\hat{\theta})\}\pi(\theta)d\theta \geq 2n^{-1}\right] \leq n^{-2}. \tag{8}$$

Let $g_\nu(X) = \sup_{|\theta-\theta_0|<n^{-c}} |D^\nu \log f_\theta(X)|$, then

$$\begin{aligned} P_{\theta_0}\left(\sup_{|\theta-\theta_0|<n^{-c}} |D^\nu l_n(\theta)|/n > n^t\right) &\leq P_{\theta_0}\left(\sum_{i=1}^n g_\nu(X_i) > n^{1+t}\right) \\ &\leq n^{-qt} E_{\theta_0}\{g_\nu(X_i)^q\}, \end{aligned}$$

assumption [**A1**] then implies that for all $|\nu| \leq 4$

$$\int_\Theta P_{\theta_0}\left(\sup_{|\theta-\theta_0|<n^{-c}} |D^\nu l_n(\theta)| > n^{1+t}\right) \pi(\theta_0) d\theta_0 \leq n^{-2}. \tag{9}$$

It thus only remains to bound the second term of (7). Let $J_n(\theta) = -n^{-1} D_2 l_n(\theta)$ and $J_n = J_n(\hat{\theta})$, then $J_n = J_n(\theta_0) + (\hat{\theta} - \theta_0) D J_n(\bar{\theta})^T$, with $\bar{\theta} \in (\theta_0, \hat{\theta})$ and where A^T denotes the transpose of A. Let $Z_{n,2}(\theta_0)$ be defined by $J_n(\theta_0) = I(\theta_0) + n^{-1/2} Z_{n,2}(\theta_0)$, then

$$\begin{aligned} P_{\theta_0}\left(|J_n|^{-1} > n^t\right) &\leq P_{\theta_0}\left(|I(\theta_0)|^{-1} > n^t/2\right) + P_{\theta_0}\left(n^{-1/2}|Z_{n,2}(\theta_0)| > 1/4\right) \\ &\quad + P_{\theta_0}\left\{|(\hat{\theta} - \theta_0) D J_n(\bar{\theta})^T| > 1/2\right\}. \end{aligned} \tag{10}$$

Hypothesis [**A5**] implies that the first term of the right hand side of (10) is of the right order. The last term is bounded by

$$a_n = P_{\theta_0}\left(|D J_n(\hat{\theta})| > n^c/2\right) + P_{\theta_0}\left(|\hat{\theta} - \theta| > n^{-c}\right).$$

The first term of a_n is bounded by n^{-2} as previously and the second one also, using hypothesis [**A3**]. We now consider the second term of (10).

$$\begin{aligned} P_{\theta_0}\left(n^{-1/2}|Z_{n,2}(\theta_0)| > 1/4\right) &\leq 4^{q'/2} n^{-q'/2} E_{\theta_0}\left(|Z_{n,2}(\theta_0)|^{q'}\right) \\ &\leq C n^{-q'/2} E_{\theta_0}\left(|D^2 \log f_{\theta_0}(X) + I(\theta_0)|^{q'}\right), \end{aligned}$$

where C is a constant depending only on q'. Inequality (9) implies that there exists $4 \leq q' \leq q$ such that the above expectation is finite and integrable in θ_0. This achieves the proof of Theorem 1. □

The algorithm M_L gives therefore a reasonable answer when the number of individuals is large, in theory. We now present a simulation study, to illustrate this in practice and to compare it to the classical Gibbs algorithm M_0.

3 Simulations

We have simulated a data set in a simple case of HMM, the posterior distribution of the parameters is estimated by the Gibbs sampling M_0 as described in Section (1.2) and by the Gibbs sampling with the Laplace approximation step M_L. The comparison allows us to appreciate the relative performance of each one.

3.1 Simulated Model

The hierarchical model used for the simulations involves 3 states in the Markov process, the third one being the absorbing state, and a gaussian distribution as a link between the observations and the true states. More precisely, if S_{ij} is the state of the individual i, at time t_{ij}, then S_{ij} takes its value in $\{1, 2, 3\}$ and the transitions rates are chosen to be $\lambda_1 = 0.04$, $\lambda_2 = 0.005$ and $\lambda_3 = 0.009$ corresponding, respectively, to the transitions from state 1 to state 2, from state 2 to state 1 and from state 2 to state 3. The third state is supposed to be observed. If X_{ij} is the value of the continuous observed variable, then the conditional distribution of X given S is $\mathcal{L}(X_{ij}|S_{ij} = k) = \mathcal{N}(\mu_k, \sigma_k^2)$, $k = 1, 2$, where $\mu_1 = \log 1100$, $\mu_2 = \log 400$, $\sigma_1^2 = 0.1$, and $\sigma_2^2 = 0.07$. The choice of the parameter's values was inspired by the HIV example where the observed variable X corresponds to the CD4 cell counts in a log scale. 300 individuals were simulated with a number of observations per individual between 10 and 12.

Figure 1 represents the histogram of the simulated X where the existence of two modes clearly highlights the two states.

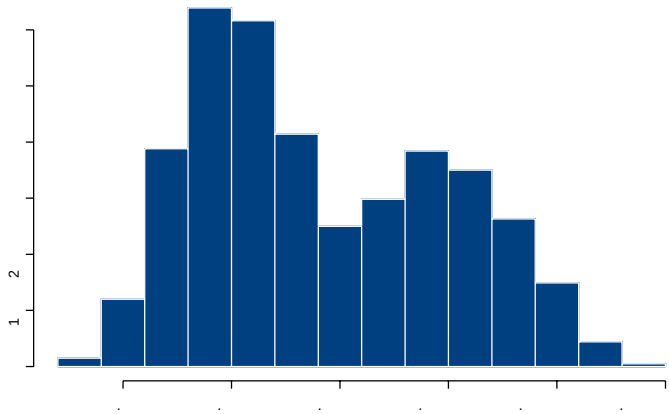

Figure 1: Histogram of simulated data

3.2 Implementation

The nuisance paramaters are globally denoted by $\underline{\theta}$ and the transition rates by λ. We consider two cases. In case 1, the mean parameters, μ_1 and μ_2, are supposed to be known, $\underline{\theta}$ is then simply composed by the variance parameters (σ_1^2, σ_2^2). Then,

it is very easy to obtain an exact analytical expression of $\int_\Theta \pi(\lambda, S, \underline{\theta}|X)d\underline{\theta}$. In this case, we can simulate a Markov chain (λ^t, S^t) whose stationary distribution is the true posterior $\pi(\lambda, S|X)$, without simulating $\underline{\theta}$. This algorithm will be called the Exact algorithm and the posterior distribution of the parameters will be considered as a reference in the comparison with the results obtained by the other two algorithms: the Gibbs Algorithm M_0 and the Laplace algorithm M_L. This provides us a way to evaluate the performance of the Laplace approximation and the effect of the approximation on the posterior distribution. In case 2, we consider the mean parameters as unknown so that the nuisance parameter $\underline{\theta}$ contains in addition μ_1 and μ_2. In this case, no analytical expression of $\int_\Theta \pi(\lambda, S, \underline{\theta}|X)d\underline{\theta}$ exists, excluding the use of the Exact Algorithm.

Recall that, in the Laplace algorithm, we need to control the fact that $(X, S) \in B$, where B is defined in Section 2.2. In our case, since $f(x|\underline{\theta})$ is gaussian, we only need to check that the numbers of observations per state, i.e. the n_s's, is large enough, i.e. greater than $n^{-3/4}$, for instance, and that the $\hat{\sigma}$'s were always neither too large nor too small. It turned out, that these cases never happened.

As in the HIV problem studied by Guihenneuc *et al.* (2000), we consider the following prior distributions. The transition rates are taken to be uniform on $[0, 0.25]$, σ_1 and σ_2 are independent inverse Gamma's with parameters (4,0.1). In the case of unknown mean, i.e. Case 2, e^{μ_2} is uniform random variable on $[100, 1100]$, and μ_1 is fixed as $\log 1100$.

3.3 Results

The results are obtained on the basis of 50000 iterations of each algorithm excluding 1000 iterations for the burn-in. Figure 2 represents the estimated posterior distributions of λ_1 when the mean parameters are considered as known for the three algorithms (case 1). We notice that the Exact algorithm and the Laplace algorithm give very similar results in terms of posterior means and credibility intervals. This remark shows in this example a very good performance of the Laplace approximation. The third algorithm, i.e. the Gibbs Algorithm, leads to a posterior mean close to the other two means and coherent with the true value of λ but the estimated credibility intervals are smaller than those obtained by the first two algorithms. The Gibbs algorithm therefore seems to underestimate the tails of the posterior density of λ. This point could be explained by the strong correlation between S and θ leading to poor mixing properties in the Gibbs sampling as suggested in Section 1.2. The

classical diagnostic tools such as those provided by CODA give no real indication of divergence for the three algorithms. This is probably due to the well known fact, that those tools fail to diagnostic when the Markov chain never visits parts of the support of the target density. Moreover, this feature of the Gibbs algorithm, i.e. it doesn't explore well the tails of the target density when the number of parameters is very large is well known. What is rather surprising, is that the algorithm improves so much, by reducing the number of parameters by a small amount.

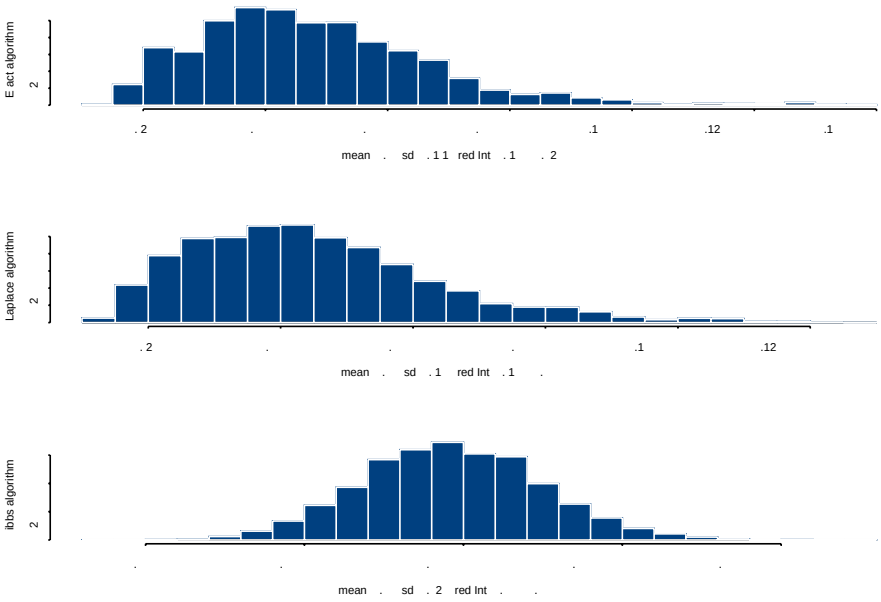

Figure 2: Posterior distributions of λ_1 for the case 1 obtained by, from top to bottom: Exact, Laplace, Gibbs algorithm

A new parameter of interest which can be evaluated at each iteration is the waiting times $T_{i\to j}$ of passage into state j starting from state i. Figure 3 gives the estimated posterior distributions of $T_{1\to 3}$ by the three algorithms. This parameter can be considered as a summary of the global trajectory of the individuals, its computation involves the value of the three transition rates. The three algorithms give similar results in terms of posterior mean and credibility intervals, suggesting that the differences previously noticed are counterbalanced. We observe the same phenomenon in the second case when the mean parameters are unknown. Figures 4 and 5 represent the estimated posterior distribution of respectively λ_1 and $T_{1\to 3}$ by the Laplace algorithm (top) and the Gibbs Algorithm (bottom). Remember that, in this case, the Exact algorithm can not be implemented. We observe again

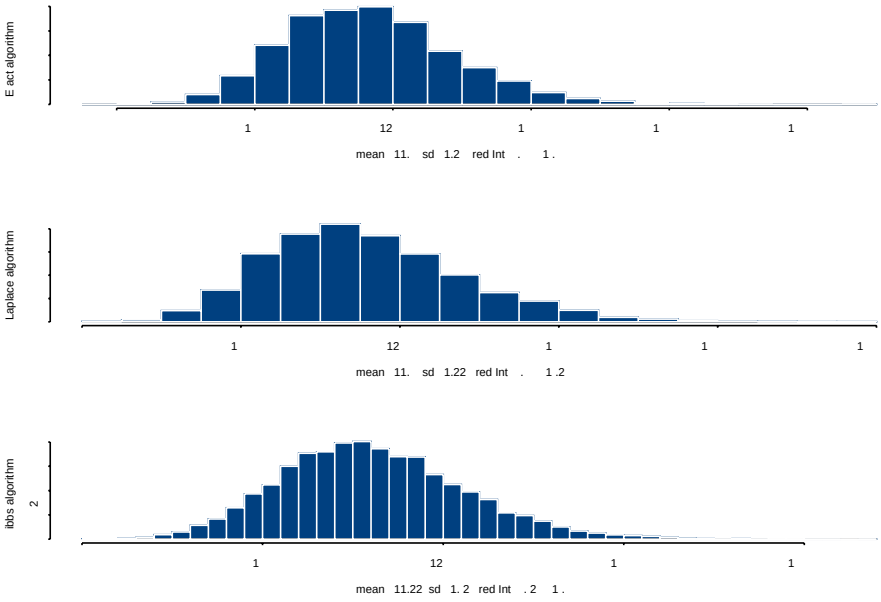

Figure 3: Posterior distributions of $T_{1\to 3}$ for the case 1 obtained by, from top to bottom: Exact, Laplace, Gibbs algorithm

a coherence between the posterior means but not between the ranges of λ_1. As previously, the Gibbs Algorithm seems to have some difficulties to cover the support of posterior density of this parameter. Note also that the distributions of $T_{1\to 3}$ obtained from both algorithms are again very similar.

We do not present here the results for λ_2 and λ_3 since they are very similar to those obtained for λ_1.

A good estimation of the transition rates is very important because it characterizes the trajectories before the absorbing state. It is specially relevant, when a covariate effect, such as a treatment, is studied. The covariate effect, say the treatment, can be measured, for instance, by the ratio between the rates associated to treated patients and the rates associated to non treated patients. The treatment would have an effect if 1 does not belong to HPD regions for this ratio. Guihenneuc *et al.* (2000) show that the covariate effect can be highlighted on a part of trajectories but is diluted in global measures like waiting times, $T_{1\to 3}$.

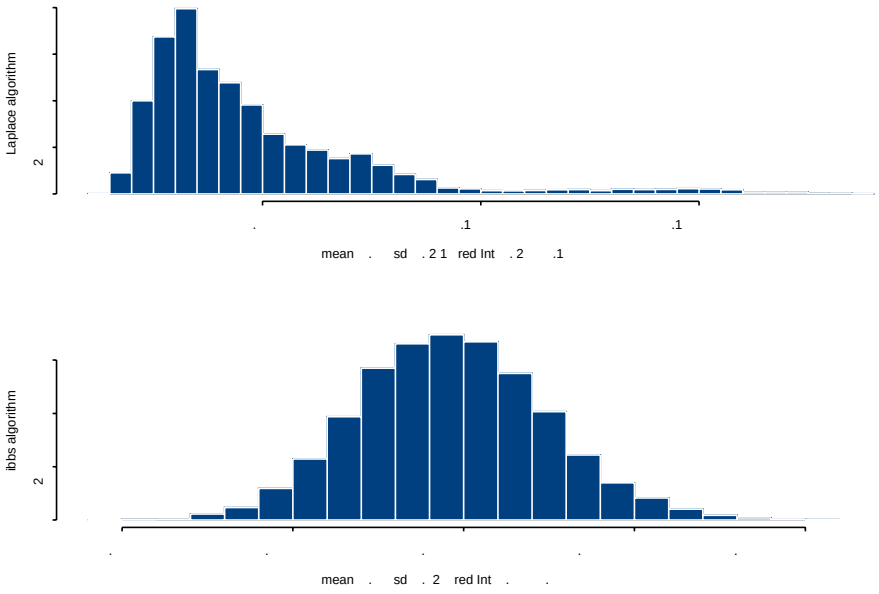

Figure 4: Posterior distributions of λ_1 for the case 2 obtained by, from top to bottom: Laplace, Gibbs algorithm

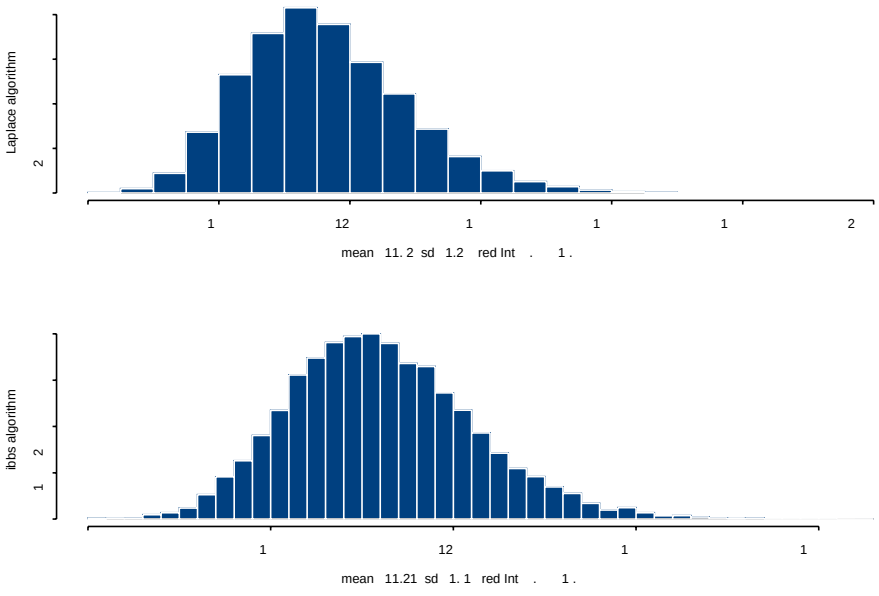

Figure 5: Posterior distributions of $T_{1\to 3}$ for the case 2 obtained by, from top to bottom: Laplace, Gibbs algorithm

4 Conclusion

In this paper, we therefore propose an algorithm, which simulates an approximated posterior density, by using a Laplace approximation at each iteration of a Gibbs algorithm. We have proved that the new target density gets close to the true one, as the number of observations increases. In the simulations we have carried out, we observed that, even with a reasonable number of individuals (300), the posterior distribution was very well approximated by the Laplace algorithm. The surprisingly good behaviour of the Laplace approximation, might be due to the fact that Laplace approximations of posterior quantities are actually correct to the order $n^{-3/2}$ instead of n^{-1}, as was suggested by Tierney, see Kass, Tierney and Kadane (1989).

The types of models considered here, i.e. hidden Markov models with conditional independence, are of great interest in many fields of applications. This algorithm could therefore be used in many applied studies where the large computation time is a real problem, as an improvement of the classical Gibbs algorithm. It seems that, not only it reduces the computational time, but also that the tails of the posterior distribution are better estimated, leading to more reliable confidence regions.

It would be interesting to study the behaviour of the Laplace algorithm on real data, where we are, in addition, often faced to the misspecification of the model.

An other extension of this work, could be to consider more complex dependence structure, as is encountered, for instance, in DNA problems.

Acknowledgement

The authors thank C.P. Robert for helpful comments on this work. This work was partially supported by the Training and Mobility of Researchers (TMR) network.

5 Appendix 1 : Proof of (8)

We recall that $A_n = \{\theta; |\theta - \theta_0| > n^{-c}; l_n(\theta) - l_n(\hat{\theta}) > -\log n\}$. In this proof, for clarity's sake, we denote $\pi(B)$ the probability of B under the prior distribution of

θ. Then

$$\begin{aligned} P_{\theta_0}\left\{\pi(A_n) > n^{-1}\right\} &\leq nE_{\theta_0}\{\pi(A_n)\} \\ &= n\int_{|\theta-\theta_0|>n^{-c}} P_{\theta_0}\left\{l_n(\theta) - l_n(\theta_0) > -\log n\right\}\pi(\theta)d\theta \\ &\leq n\int_{|\theta-\theta_0|>n^{-c}} P_{\theta_0}\left\{Z_n(\theta) > \sqrt{n}K(\theta_0,\theta) - \log n/\sqrt{n}\right\}\pi(\theta)d\theta, \end{aligned}$$

where $Z_n(\theta) = n^{-1/2}\left\{l_n(\theta) - l_n(\theta_0) + nK(\theta_0,\theta)\right\}$. Let

$$\tilde{A}_n = \{\theta; |\theta-\theta_0| > n^{-c}, K^2(\theta_0,\theta) \geq (2+a/\beta)\log n/nM_2(\theta_0,\theta)\},$$

Hypothesis [A4] implies that

$$P_{\theta_0}\{\pi(A_n) > n^{-1}\} \leq n\int_{\tilde{A}_n} P_{\theta_0}\left(Z_n(\theta) > \sqrt{n}K(\theta_0,\theta)/2\right)\pi(\theta)d\theta + n^{-1-(a/\beta)}.$$

Let $\theta \in \tilde{A}_n$,

$$\begin{aligned} P_{\theta_0}\left(Z_n(\theta) > \sqrt{n}K(\theta_0,\theta)/2\right) &\leq e^{-t\sqrt{n}K(\theta_0,\theta)/2}\left[E_{\theta_0}\left\{e^{t(\log f_\theta - \log f_{\theta_0})/\sqrt{n}}\right\}e^{tK(\theta_0,\theta)/\sqrt{n}}\right]^n \\ &= e^{-t\sqrt{n}K(\theta_0,\theta)/2}\left[1 + \frac{t^2}{2n}\int_0^1 E_{\theta_0}\left\{(l(\theta_0)-l(\theta))^2 e^{ut(l(\theta)-l(\theta_0))/\sqrt{n}}\right\}du\right]^n \\ &\leq e^{-t\sqrt{n}K(\theta_0,\theta)/2}\left\{1 + \frac{t^2M_2(\theta_0,\theta)}{n}\right\}^n \\ &\leq e^{-t\sqrt{n}K(\theta_0,\theta)/2 + t^2M_2(\theta,\theta_0)/2}. \end{aligned}$$

Let $t = 2\sqrt{n}K/M_2$, then $P_{\theta_0}\left(Z_n(\theta) > \sqrt{n}K(\theta_0,\theta)/2\right) \leq e^{-nK(\theta_0,\theta)^2/M_2(\theta,\theta_0)}$, we thus obtain $\int_{\tilde{A}_n} e^{-nK(\theta_0,\theta)^2/M_2(\theta,\theta_0)} \leq n^{-2-(a/\beta)}$, which achieves the proof of (8).

References

Bickel, P.J. & Ghosh, J.K. (1990). A decomposition for the likelihood ratio-statistic and the Bartlett correction- A bayesian argument. *Ann. Statist.*, **18**, 1070-90.

Chib, S. (1996). Calculating posterior distributions and modal estimates in Markov mixture models. *J. Economet.* **75**, 79-97

Ghosh, J.K., Sinha, B. & Joshi, S.N. (1982). Expansions for posterior probability and integrated Bayes risk. *Statistical Decision Theory and Related topics 3* (S.S. Gupta and J.O. Berger, Eds.) **1** 403-456. Academic, New York.

Guihenneuc-Jouyaux, C., Richardson, S. & Longini, I.M. (2000). Modeling Markers of disease progression by a hidden Markov process. *Biometrics*, **56**, 3, 733-741.

Hamilton, J.D. (1989). *Time Series Analysis*. Princeton university Press

Hodgson, M.E.A. (1999). A bayesian restoration of an ion channel signal. **61**, 95-114.

Ibragimov; I. & Has'minskii, R. (1981). *Statistical estimation*. Springer, New-York.

Kass, R., Tierney, L. & Kadane, J.B. (1989). Approximate methods for assessing influence and sensitivity in Bayesian analysis. *Biometrika*, **76**, 663-74.

Kirby, A.J. & Spiegelhalter D.J. (1994). Statistical Modeling for the Precursors of Cervical Cancer. In *Case Studies in Biometry*, N. Lange (Ed.), John Wiley, New-York.

Muri, F. (1998). Modeling bacterial genomes using hidden Markov models. In *Compstat'98, Proceedings in Computational Statistics* (Eds R. Payne and P. Green), 89-100, Heidelberg: Physica-Verlag.

Rabiner, L.R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proc. IEEE*, **77**, 257-286.

Robert, C.P. & Casella, G. (1999). *Monte Carlo Statistical Methods*. Springer-Verlag, New-York.

Shephard, N. & Pitt, M. K. (1997). : Likelihood analysis of non-Gaussian time series. *Biometrika*, **84**, 653-667.

Tierney, L. & Mira, A. (1999). Some adaptative Monte Carlo methods for Bayesian inference. *Statist. in Med.*, **18**, 2507-2515.