# Laplace Expansions in Markov Chain Monte Carlo Algorithms

Chantal Guihenneuc-Jouyaux, Judith Rousseau

Chantal GUIHENNEUC-JOUYAUX and Judith ROUSSEAU [1]

Summary

Complex hierarchical models lead to a complicated likelihood and then, in a Bayesian analysis, to complicated posterior distributions. To obtain Bayes estimates such as the posterior mean or Bayesian confidence regions, it is therefore necessary to simulate the posterior distribution using a method such as an MCMC algorithm. These algorithms often get slower as the number of observations increases, especially when the latent variables are considered. 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 provide a theoretical study of the impact that such an approximation has on the target posterior distribution. We prove that the distance between the true target distribution and the approximation becomes of order $O(N^{-a})$ with $a \in (0,1)$, $a$ close to 1, as the number of observations $N$ increases. A simulation study illustrates the theoretical results.The approximated MCMC algorithm behaves extremely well on an example which is driven by a study on HIV patients.

*Some key words: Bayesian Hierarchical Model; Gibbs algorithm; Laplace approximation; Latent variable models*

# 1 INTRODUCTION

As the complexity of models considered in statistical inference increases, the need of new computional tools gets increasingly pressing. In this respect, Markov chain Monte Carlo (MCMC) methods have been widely developed in the last decade and have enhanced the use of complex models in different types of applications. Typically

[1]C. Guihenneuc-Jouyaux is Associate Professor, INSERM U170, IFR69, 16 avenue P-V Couturier, 94 807 Villejuif Cedex and Laboratoire MAP5, UMR 81 45, Université Paris 5, France (Email: chantal.guihenneuc@univ-paris5.fr). J. Rousseau is Associate Professor, Laboratoire MAP5, UMR 81 45, Université Paris 5, 45 rue des St-Pères, 75 006 Paris and CREST, Malakoff, France (Email: rousseau@ensae.fr). The authors thank the referees and editors for helpful suggestions that lead to significant improvements. 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, by the French Institute of Health and Medical Research (*program ATC*) and the French Agency for Environmental Health Safety (*"Environnement et Santé, 2003"*).

this has been done using Bayesian inference, (Robert and Casella (1999)). In a Bayesian approach, samples produced by MCMC algorithms are quite appropriate to approximate many aspects of the target posterior distributions using ergodic averages. A typical class of complex Bayesian models are hierarchical models with a large number of parameters.

Suppose we have a hierarchical model with the following joint probability density with respect to some measure $\mu$:

$$f(X|\theta_S, S)\pi(\theta_S|S)p(S|\lambda)h(\lambda),$$

where $X$ is the vector of observations, $S \in \mathcal{S}$ is a vector of parameters - typically latent variables or at least high dimensional - $\lambda \in L$ is the parameter associated with $S$ and $\theta_S \in \Theta$ is the nuisance parameter. Depending on the problem, either $S$, or $\lambda$, or both are the parameters of interest. In this paper we shall therefore consider $(\lambda, S)$ as the parameter of interest, since the marginal distributions of $\lambda$ and $S$ can be obtained from the joint distribution.

Traditionally, there are two ways to calculate posterior quantities of interest: asymptotic expansions or simulations. If the dimension of the parameter $(\lambda, S, \theta_S)$ is too large relative to the number of observations, asymptotic expansions such as Laplace expansions can work quite poorly or even not be valid, see Ghosh (1994, Ch. 5) or Tierney and Kadane (1986). It is then necessary to compute the posterior distribution via an MCMC algorithm. However, it is often the case that 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. We study a way to combine the asymptotic approximations of the target posterior distribution via Laplace, with MCMC.

The most commonly used MCMC methods to sample from the posterior distribution are the Gibbs sampler and the hybrid Gibbs sampler, see for instance Robert and Casella (1999). A Gibbs sampler involves sampling components (possibly vectors) using the full conditional distributions. In a hybrid Gibbs sampler we use Hasting-Metropolis steps when some of the full conditional distributions are too complicated to sample from.

A Gibbs algorithm would run a chain on $(\lambda, S, \theta_S)$ to obtain a sample from the posterior distribution of $(\lambda, S)$ in the following way :

1. $\lambda^t \sim h(\lambda|S^{t-1})$

2. $S^t \sim p(S|X, \lambda^t, \theta_S^{t-1})$

3. $\theta_S^t \sim \pi(\theta_S|X, S^t, \lambda^t)$,

we denote this algorithm by $M_0$.

Instead of running a chain on $(\lambda, S, \theta_S)$ we can run a chain on $(\lambda, S)$ by integrating out the nuisance parameter $\theta_S$, using

$$g(X|S) = \int_\Theta f(X|\theta_S, S)\pi(\theta_S|S)d\theta_S, \tag{1}$$

which leads to the new sampling scheme,

1. $\lambda^t \sim h(\lambda|S^{t-1})$

2. $S^t \sim p(S|X, \lambda^t) \propto g(X|S)p(S|\lambda^t)$.

We denote this algorithm by $M_1$.

Liu, Wong and Kong (1994) have proved that, in terms of maximal correlations, Scheme $M_1$ is better than Scheme $M_0$, at least when no Hasting-Metropolis step is included. They suggest that, if there is a strong dependence between two components, integrating one of them out is the best strategy. In our setup, there is often a strong dependence between $S$ and $\theta$, since $S$ can be a parameter of interest, it is natural to integrate $\theta$ out. However, in most cases (1) cannot be obtained analytically but can be approximated using a Laplace expansion. The new sampling scheme is then

1. $\lambda^t \sim h(\lambda|S^{t-1})$

2. $S^t \sim \hat{p}(S|X, \lambda^t) \propto \hat{g}(X|S)p(S|\lambda^t)$,

where $\hat{g}(X|S)$ denotes the Laplace approximation of $g(X|S)$, see for instance Tierney *et al.* (1989).

The stationary distribution of the simulated Markov chain is then modified. The question is how much is it modified? This approach has been used in particular by DiMatteo, Genovese and Kass (2001) in the context of regression function estimation via free-knot splines. In their approach $S = (k, \xi)$ represents the number and the locations of the knots, $\theta_S = (\beta, \sigma)$ is the parameter to be integrated out, although $\beta$

can be a parameter of interest. In other words they simulate directly $\pi(S|X^n)$ without having to simulate $\pi(\beta, \sigma|S, X)$. By considering the number of knots bounded and under a particular prior they prove that the Laplace approximation of $g(X|k, \xi)$ is uniform (in $(k, \xi)$) and thus obtain an approximate posterior density for $S$ close to the true one, in probability.

In this paper we extend this type of result to a more general class of models. Our aim is to measure the impact of the Laplace approximation occurring at each step of the algorithm on the posterior distribution of the parameters of interest, namely $(\lambda, S)$. The first result we give on the error due to the approximated posterior distribution of $(\lambda, S)$ is general. We then focus our attention on latent variable models. These models have been used in many areas as a convenient representation of weakly dependent heterogeneous phenomena. Hidden Markov models (HMM) 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 modeling (Rabiner (1989), Durbin et al (1998), Muri (1998)) and in medicine (Guihenneuc et al (2000), Kirby and Spiegelhalter (1994)). Our work has been motivated by biomedical applications. In medicine, 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 model representation.

In this paper we assume that we have $N$ observations $X = (X_1, ..., X_N)$ and that conditionally on some vector $S$ they are independent with distribution whose density with respect to Lebesgue measure is denoted $f(X_i|\theta_S, S)$, $i = 1, ..., N$. We assume that the distribution of $S$ depends on some parameter $\lambda$ and we denote with $\pi$, $h$ and $p$ any distribution (prior or posterior) on respectively $\theta_S$ (given $S$), $\lambda$ and $S$. In a latent variable model, $S$ would be the latent process.

In Section 2, we present the algorithm based on the Laplace expansion. In Section 2.1 we describe the algorithm and in Section 2.2 we prove that the stationary distribution of the Markov chain generated by the algorithm based on the Laplace

approximation gets close to the true posterior distribution of $(S, \lambda)$ as $N$ goes to infinity. This Section is divided into two parts, first we present a general result and then we focus on latent variable models. In Section 3 we give some simulations that illustrate the good behaviour of the approximated algorithm.

# 2 LAPLACE APPROXIMATION

## 2.1 The Laplace Algorithm

We begin with notation. Let $l_N(\theta_S)$ be the log-likelihood of the completed model conditional on $S$ and let $\hat{\theta}_S$ 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 $\nu$-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. $J = -D_2 l_N(\hat{\theta}_S)$ and let $|J|$ be its determinant. Finally, $||\mu_1 - \mu_2||_{TV}$ denotes the total variation norm of $\mu_1 - \mu_2$ and $\psi = \log \pi(\theta)$.

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

$$\pi(\lambda, S|X) = \frac{\left\{\int_\Theta \prod_{i=1}^N f(X_i|\theta_s, S)\pi(\theta_S|S)d\theta_s\right\} p(S|\lambda)h(\lambda)}{\int_L \int_{\mathcal{S}} \left\{\int_\Theta \prod_{i=1}^N f(X_i|\theta_s, S)\pi(\theta_S|S)d\theta_s\right\} dp(S|\lambda)h(\lambda)d\lambda}.$$

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

$$\hat{\pi}(\lambda, S|X) = \frac{\left\{\prod_{i=1}^N f(X_i|\hat{\theta}_S)\pi(\hat{\theta_S})J^{-1/2}\right\} p(S|\lambda)h(\lambda)}{\int_L \int_{\mathcal{S}} \left\{\prod_{i=1}^N f(X_i|\hat{\theta}_S)\pi(\hat{\theta_S})J^{-1/2}\right\} dp(S|\lambda)h(\lambda)d\lambda}.$$

Let $g(X|S)$ be the marginal conditional density of $X$ given $S$ and $\hat{g}(X|S)$ its Laplace approximation, i.e.

$$g(X|S) = \int_\Theta e^{l_N(\theta_S)}\pi(\theta_S|S)d\theta_S$$

and

$$\hat{g}(X|S) = (2\pi)^{d/2}\pi(\hat{\theta_S})J^{-1/2}\prod_{i=1}^N f(X_i|\hat{\theta}_S),$$

where $d$ is the dimension of $\theta_S$. We thus would have: $\hat{\pi}(\lambda, S|X) \propto \hat{g}(X|S)p(S|\lambda)h(\lambda)$. However, since this approximation will be used at each iteration of the Gibbs algorithm and since the error depends on $S$, there might be some values of $S$, for which the approximated $\hat{g}(X|S)$ is quite different from $g(X|S)$. Since it is difficult to control the error of the approximation, for such values of $S$, 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 $\mathbb{I}_B$ denotes the indicator function of the set $B$. $a$ will be chosen as close to 1 as possible. Note that, if the error term $|\hat{g}(X|S) - g(X|S)|$ goes to 0 uniformly in $S$, there is no need for the use of $\tilde{g}(X|S)$, and we can use $\hat{g}(X|S)$.

We then have as the limiting target distribution : $\tilde{\pi}(\lambda, S|X) \propto \tilde{g}(X|S)p(S|\lambda)h(\lambda)$. The new algorithm has thus the following structure : at the $t$-th iteration,

1. $\lambda^t \sim h(\lambda|X, S^{t-1})$ which is the true one,
2. $S^t \sim \tilde{p}(S|X, \lambda^t) \propto \tilde{g}(X|S)p(S|\lambda^t)$.

We denote this algorithm by $M_L$ and call it 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.

## 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 another. In Section (2.2.1) we present the result in its most general form. As the general case covers situations that can be very different, for instance latent variable models with discrete or continuous latent variables or curve estimation via free knot splines as in DiMatteao *et al.* (2001), the conditions we give to ensure the validity of the approximation are vague. Therefore, in Section (2.2.2), we deal with the special case of latent variable models with finite state spaces.

### 2.2.1 General Case

In the case of latent variable models, the dimension of $S$ increases with $N$, i.e. $S = (s_1, ..., s_N)$. If the $s_i$'s take their values in a finite space say $\{1, ..., k\}$ then $\theta_S \in (\theta_1, ..., \theta_k)^N$, where $\theta_i \in \mathbb{R}^{p_i}$ is the parameter corresponding to the population for which $s = i$. We can therefore write $\theta_S = (\theta(s_i), ..., \theta(s_N))$. If the $s_i$'s are continuous random variables then $\theta_S$ would be a function of the latent variable $s_i$, for each

observation, parameterised by a finite dimensional parameter $\theta$, i.e. $f(X_i|\theta_S, S) = f(X_i|c_\theta(s_i))$. If $c$ is linear $c_\theta(s_i) = \theta^t s_i$ and the model can be considered as a generalised linear model. In the free knot splines case the number of knots is bounded but must be estimated and $\theta_S$ is also a function in the form $c_\theta(s_i)$. Generally speaking we write the sampling model as $f(X_i|\theta_S, S) = f(X_i|c_\theta(s_i))$ with $\theta \in \Theta \subset \mathbb{R}^d$ and $c_\theta$ a function of $s_i$.

We consider the following assumption.

**[G]** There exist $a > 0$ and $\mathcal{S}_1 \subset \mathcal{S}$, such that $\forall \theta \in \Theta$, $\forall S \in \mathcal{S}_1$

$$P[B^c|\theta, S] \leq M(\theta)/N^{1+a}, \quad \text{with} \int_\Theta M(\theta) d\theta < \infty$$

and

$$\int_L P[\mathcal{S}_1^c|\lambda] h(\lambda) d\lambda < M/N^{1+a},$$

when $N$ is large enough.

We then have the following result,

**Theorem 1** *Under condition* **[G]** *the approximate target distribution is close to the true one in the following sense:*

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

*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 $X$.*

*Proof of Theorem (1)*: The proof is given in Appendix A. The idea is the following, by definition of $B$, $g(X|S) = \tilde{g}(X|S)(1+O(N^{-a}))$ 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 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 $p(B^c|X)$, which is done using assumption **[G]**.

This hypothesis is vague due to the generality of Theorem 1. The idea is that under regularity conditions on the error model $f(X|\theta_S, S)$, Laplace expansions are valid and the set $B$ is large enough to ensure **[G]**. Actually 2 steps are crucial when applying this theorem to a specific model. One is the determination of $B$, which must be defined using quantities that can be evaluated at each iteration. Another is to verify that the marginal probability of $B^c$, $\int_{\Theta \times \mathcal{S}} P[B^c|\theta, S]\pi(\theta|S) d\theta dp(S)$, on which

the Laplace approximation is not controlled, is bounded. When the conditional probabilities of $B^c$ given $(\theta, S)$ are uniformly bounded, as in DiMatteo *et al.* (2001), then Theorem 1 can easily be applied. However, when the dimension of $S$ increases with $N$, as is typically the case with latent variable models, this is not the case and justifying condition [**G**] can be quite involved. In the following section, we therefore study more precisely the case of latent variable models, for which the $s_i$'s belong to some finite state space.

### 2.2.2 Latent Variable Models With Finite State Space

In this Section, we consider the following type of models, based on longitudinal data. The data consist of observed values $X_{ij}$ where $i$ indexes the individual and $j$ the follow-up point, $1 \leq i \leq n$, $1 \leq j \leq n_i$ and $N = \sum_{i=1}^{n} n_i$. We therefore have $n$ individuals, and for each individual $i$, we have a number $n_i$ of observations. The $X_{ij}$'s, conditional on the unobserved random variables $S_{ij} = s$, $s = 1, ..., k$, are independent with distribution $P_{\theta_s}$. The distribution $P_{\theta_s}$ has a density with respect to Lebesgue measure denoted by $f_{\theta_s}(X)$, where $\theta_s \in \Theta_s$, and $\Theta_s$ is an open subset of $\mathbb{R}^{p_s}$. The densities $f_{\theta_s}$ may differ by more than the parameter $\theta_s$ and they can belong to different parametric families. The latent process is defined as follows. The individuals are independent, and for each individual $i$, the $S_{ij}$'s, $j = 1, ..., n_i$, can have a dependence structure (for example Markovian), with a distribution depending on a parameter $\lambda \in L$. In other words, $S_{ij}$ and $S_{i'j'}$ are independent when $i \neq i'$. In this regard, a motivating illustration of such a latent variable model is the HMM used for modeling HIV patients as 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 AIDS status, based on clinical symptoms. This state is therefore perfectly observable. $S$ is modeled by a Markov process on $\{1, ..., 7\}$, for which $\lambda_{ij}$ represents the transition rate to state $j$ starting from state $i$. The conditional distributions for $S$ are therefore given by:

$$p(S_{ij}|\lambda, S_{ij-1}) = (\exp \Lambda dt_{ij})_{S_{ij-1}S_{ij}} \quad \text{and} \quad p(S_{i1} = s) = \delta(s) > 0, \tag{2}$$

where $\Lambda$ is the infinitesimal transition matrix. The error process is Gaussian.

Let $\underline{\theta} = (\theta_1, ...., \theta_k) \in \Theta = \Theta_1 \times \cdots \Theta_k$. We also denote $\pi_s(\theta_s)$ the marginal prior of $\theta_s$, $s = 1, ..., k$. If we want to characterize the progression of the hidden process $S$, $\lambda$ 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 $(\lambda, S)$ 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 $h(\lambda|X)$, the posterior density of $\lambda$. This posterior distribution is obviously not available in closed form and we must simulate it, using an MCMC algorithm. Indeed, the posterior distribution of interest has the following form when $\lambda$ is the parameter of interest:

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

Let $\hat{\underline{\theta}} = (\hat{\theta}_1, ..., \hat{\theta}_k) = \hat{\underline{\theta}}(X, S)$ be the conditional maximum likelihood estimate of $\theta$. 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 submodel, i.e. in each state $s$, $s \in \{1, ..., k\}$. To do so we consider the following assumption on the underlying unobserved process $S$ :

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

$$p(\text{for some} \quad j \leq n_i; S_{ij} = s|\lambda) \geq c_0(\lambda) > 0.$$

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

The results that are stated in this section are written for non compact $\Theta$. Throughout this section we point out how the assumptions [**A1**]-[**A6**] and results would simplify in the compact case. In the following, we denote by $E_\theta\{h(X)\}$ the expectation of $h(X)$ under $P_\theta$.

[**A1**] In each submodel, $s = 1, ..., k$, the log-likelihood, $\log f_{\theta_s}(x)$, is 4 times continuously differentiable in $\theta_s$ 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

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

where $\pi_s$ denotes the marginal prior density of $\theta_s$.

[**A2**] In each submodel $s = 1, ..., k$, the information matrix $I_s(\theta_s)$ is positive definite, 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 $\pi$ 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). It often requires fairly weak conditions on the prior, such as moment conditions, in regular models. We have chosen this general expression because, depending on the model, the appropriate assumptions could be fairly different. For instance, even for very smooth models like the Gaussian $(\mu, \sigma)$ distribution, with $(\mu, \sigma) \in \mathbb{R} \times \mathbb{R}^+$, conditions such as those proposed by Ibragimov and Hasminskii (1981) are not really appropriate.

[**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 $\Theta$ is compact, it is enough to assume that terms are bounded, but when $\Theta$ 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 : $\pi(\sigma > cn/\log n^2) \leq n^{-2}$. 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**], $\beta \in (1/2, 1)$ and let $A_s$ be defined by : $A_s = \{\theta_s; l_s(\theta_s) - l_s(\hat{\theta}_s) > -\log n_s\} \cap \{\theta_s; |\theta_s - \hat{\theta}_s| > n^{-c}\}$, then

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

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

When $B$ is defined as such, $a = \beta(1-3t)$. This definition of $B$ is simpler in the compact case, by dropping the last two constraints on $\psi = \log \pi$.

We now state the main result of this section :

**Theorem 2** *If* [**H**] *is satisfied and if the hypotheses* [**A1**] $-$ [**A6**] *are satisfied, the approximate target distribution is close to the true one in the following sense:*

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

*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.

Note that the error bounds for the posterior distribution of $(\lambda, S)$ are controlled in terms of the number of individuals, $n$, and not by the total number of observations, $N$, as in Theorem 1. However in most applications with longitudinal data, $N$ would typically equal or at least bounded by $n$ times a constant greater than 1, in other words the number of follow ups would be more or less the same for each individual.

*Proof of Theorem 2*: As in the proof of Theorem 1, we obtain

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

Therefore, we just need to prove that

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

The proof of (3) is given in Appendix B. □

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

The algorithm $M_L$, in Section 2.1, 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 $M_0$.

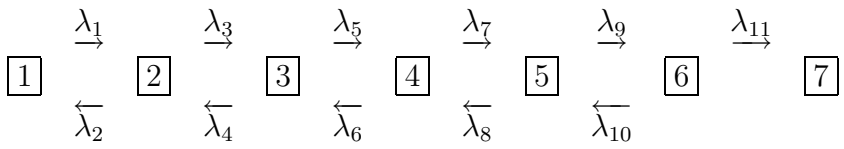
# 3 SIMULATIONS

We have simulated a data set in a case of HMM very close to the HIV example. The posterior distribution of the parameters will be estimated by the classical Gibbs sampling algorithm $M_0$ as described in Section 1 and by the Gibbs sampling algorithm with the Laplace approximation step $M_L$. We will compare the two algorithms in order to appreciate their relative performance.

## 3.1 Simulated Model

The hierarchical model we simulated has 7 states in the Markov process with the seventh an absorbing state that we assume is observed. A Gaussian distribution is

the 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 $\{1, ..., 7\}$. The 11 transitions rates are illustrated as follows:




The transitions rates are chosen to be equal to 0.04 for the forward rates $(\lambda_1, \lambda_3, \lambda_5, \lambda_7, \lambda_9)$ except for the transition to the absorbing state 7, $\lambda_{11} = 0.01$, and equal to 0.005 for the backward rates $(\lambda_2, \lambda_4, \lambda_6, \lambda_8, \lambda_{10})$.

As the seventh state is supposed to be observed and if $X_{ij}$ is the value of the continuous observed variable, then the conditional distribution of $X$ given $S$ is

$$f(X_{ij}|\theta_k, S_{ij} = k) = \mathcal{N}(\mu_k, \sigma_k^2), \quad k = 1, ..., 6,$$

where $\{\mu_1, ..., \mu_6\} = \{\log 1100, \log 800, \log 600, \log 425, \log 275, \log 170\}$ and $\{\sigma_1^2, ..., \sigma_6^2\}=$ $\{0.05, 0.01, 0.01, 0.01, 0.05, 0.05\}$. The choice of the parameters values was inspired by the HIV example where the observed variable $X$ corresponds to the CD4 cell counts on a log scale. We have simulated $n = 300$ individuals with a number of observations $n_i$ per individual between 10 and 12 inclusive.

Figure 1 represents the histogram of the simulated $X$ (on the left) and the values of the $X$'s per state (on the right). Regions where the values of $X$ do not give clear information on the states are highlighted by these graphs.

The goal is to estimate the parameters of interest, $\{\lambda_1, ..., \lambda_{11}\}$.

## 3.2 Implementation

The nuisance parameters are globally denoted by $\underline{\theta}$ and the parameters of interest, the transition rates, by $\lambda$. We consider two cases. In case 1, the mean parameters, $\{\mu_1, ..., \mu_6\}$, are supposed to be known, $\underline{\theta}$ is then simply composed by the variance parameters $\{\sigma_1^2, ..., \sigma_6^2\}$. Then, it is very easy to obtain an exact analytical expression of $\pi(\lambda, S|X)$. In this case, we can simulate a Markov chain $(\lambda^t, S^t)$ 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 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.

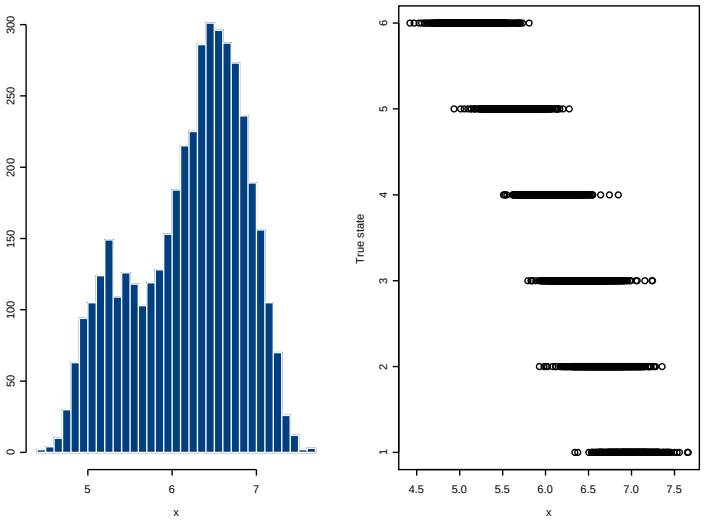



Figure 1: Histogram of simulated data $X$ (left) and values of the $X$'s per state (right)

In case 2, we consider the mean parameters as unknown so that the nuisance parameter $\underline{\theta}$ is now composed by $\{(\mu_1, \sigma_1^2), \ldots, (\mu_6, \sigma_6^2)\}$. In this case, no analytical expression of $\pi(\lambda, S|X)$ exists which excludes the use of the Exact Algorithm.

Recall that, in the Laplace algorithm, we need to ensure that $(X, S) \in B$, where $B$ is defined in Section 2.2. In this example, $f(x|\underline{\theta}, S)$ is Gaussian, so we only need to check that the number of observations per state is large enough, i.e. greater than $n^{3/4}$ 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 $[0, 0.25]$, $\{\sigma_1^2, \ldots, \sigma_6^2\}$ are independent inverse Gamma random variables with parameters (0.01,0.01). In Case 2, $\{e^{\mu_2}, ..., e^{\mu_6}\}$ are generated from order statistics on $[100, 1100]$ with a mean spacing equal to 200 in order to reasonably separate their values, and $\mu_1$ is fixed at $\log 1100$.

## 3.3 Results

The results are obtained on the basis of 20000 iterations of each algorithm excluding 1000 iterations for the burn-in. A new parameter of interest which can be evaluated at each iteration is the waiting time $T_{i \to j}$ of passage into state $j$ starting from state $i$. This parameter is directly deduced from the transition rates and is easier to

interpret, it is thus often estimated in practice.

An illustration of the results is given by Figure 2 which represents the estimated marginal posterior distributions of $T_{3\to4}$ and $T_{3\to5}$ when the mean parameters are considered as known for the three algorithms (case 1). Note that this figure is representative of the estimated marginal posterior distributions of the other parameters. The three algorithms give very similar results, although the Gibbs algorithm tends to overestimate the tails of the posterior density. To be more precise, the Exact algorithm and the Laplace algorithm give equivalent estimated posterior means and 95% credible intervals which are, respectively for $T_{3\to4}$ and $T_{3\to5}$, equal to 34.3 ($[26.4, 44.7]$) and 56.1 ($[45.9, 67.9]$). This remark shows a very good performance of the Laplace approximation. The Gibbs algorithm leads to posterior means close to those obtained from the other two algorithms but the estimated credible intervals are slightly larger (34.0, $[25.8, 44.9]$ for $T_{3\to4}$ and 56.1, $[46.4, 69.0]$ for $T_{3\to5}$). However classical diagnostic tools such as those provided by CODA software (*Convergence Diagnostic and Output Analysis*) give no real indication of divergence for the three algorithms.

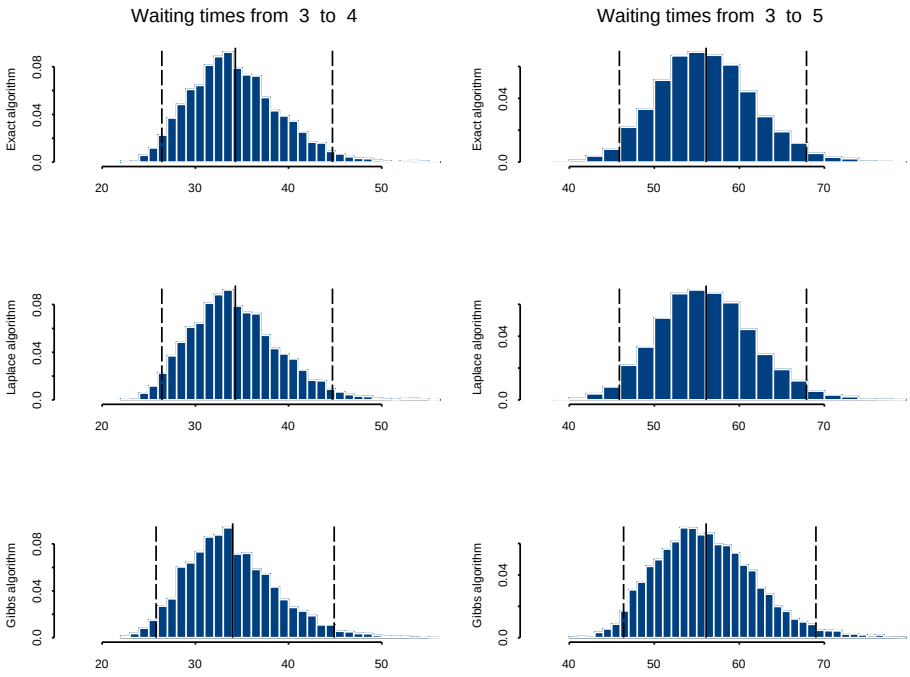



Figure 2: Marginal posterior distributions of $T_{3\to4}$ and $T_{3\to5}$ in case 1 obtained by, from top to bottom: Exact, Laplace, Gibbs algorithm. Solid and dashed lines represent, respectively, estimated posterior means and 95% credible intervals.

Table 1: Posterior mean and 95% credible interval of $T_{3\to4}$ and $T_{3\to5}$ in case 2

| | Laplace algorithm on 20000 iterations | Gibbs algorithm on 20000 iterations | Gibbs algorithm on 40000 iterations |
|---|---|---|---|
| $T_{3\to4}$ | 36.1 [26.4,47.2] | 35.1 [26.7,45.3] | 35.6 [26.7,46.8] |
| $T_{3\to5}$ | 58.6 [47.7,71.8] | 57.9 [48.1,69.8] | 57.9 [47.8,70.4] |

We observe a different phenomenon in the second case when the mean parameters are unknown. The first two rows of Figure 3 represent the estimated marginal posterior distribution of $T_{3\to4}$ and $T_{3\to5}$ by the Laplace algorithm (top) and the Gibbs Algorithm (second row) on the basis of 20000 iterations. The estimated posterior means and 95% credible intervals are given in table 1 (first two columns). Remember that, in this case, the Exact algorithm cannot be implemented. Both the posterior means and the credible intervals differ between the Laplace algorithm and the Gibbs algorithm. In particular the Gibbs algorithm does not seem to visit correctly, within 20000 iterations, the tails of the distribution. This point suggests the need of a greater number of iterations for the Gibbs algorithm. We have thus simulated 40000 iterations of the Gibbs algorithm (third row of Figure 3 and third column of Table 1). The estimated marginal posterior distributions get closer to those obtained from the Laplace algorithm (with 20000 iterations). Therefore, not only does the Laplace algorithm give a good approximation of true posterior distributions, but it also converges more quickly.

# 4 CONCLUSIONS

In this paper, we propose an algorithm, that simulates from 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

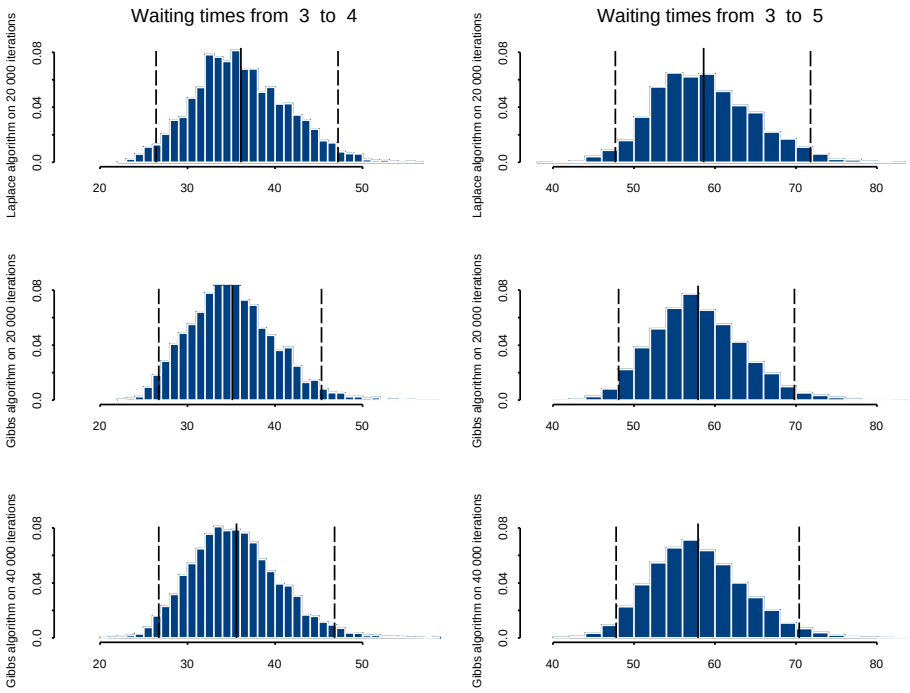



Figure 3: Marginal Posterior Distributions of $T_{3\to4}$ and $T_{3\to5}$ in case 2 obtained by, from top to bottom: Laplace algorithm on 20000 iterations, Gibbs algorithm on 20000 iterations, Gibbs algorithm on 40000 iterations. Solid and dashed lines represent, respectively, estimated posterior means and 95% credible intervals.

suggested by Kass, Tierney and Kadane (1989).

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. In our simulated example, the computational time per iteration was exactly the same for the Laplace algorithm and the Gibbs algorithm. However, as the Laplace algorithm converges more rapidly to the stationary distribution (40000 iterations of the Gibbs algorithm lead to approximately the same result as 20000 iterations of the Laplace algorithm), the global computational time was a lot better in the case of the Laplace algorithm. In the case of DiMatteo *et al.* (2001), the computational time per iteration was even better for the Laplace algorithm. In their paper, DiMatteo *et al.* (2001) used the same kind of approximation to avoid the simulation of an extra parameter. They also proved that this approximation is good, as the number of observations goes to infinity. However, their context is very special, since the dimension of the whole parameter, in their case $(k, \xi, \beta, \sigma)$, is finite and does not depend on $n$. In this paper, we present a very general result (Theorem 1) and a more practical one (Theorem 2). In both cases, we allow the dimension of the parameter to increase with $n$, or even to be infinite in Theorem 1. This makes a lot of difference when dealing with Laplace approximations, since uniform approximations cannot be obtained.

In some cases, especially when the dimension of the nuisance parameter $\theta$ is large, adding a second order term in the Laplace expansion (which would lead to an approximation of order $N^{-2}$) improves greatly the approximation at each iteration. By adding a few more assumptions on the model (typically on the prior), this should also improve the global error on the posterior distribution of $(\lambda, S)$. This would require some technical work but the basic idea would be the same.

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

# APPENDIX A: PROOF OF THEOREM 1

We have, for any Borel set $A$ on $L \times \mathcal{S}$ :

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

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,

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

We use the Markov inequality,

$$
\begin{aligned}
P_{m(X)}\{p(B^c|X) > N^{-a}\} &\leq N^a p(B^c) \\
&\leq MN^{-1},
\end{aligned}
$$

where the last inequality comes from hypothesis [**G**].

# APPENDIX B: PROOF OF INEQUALITY (3)

We use Markov's inequality :

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

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): $n^a E\{p(B_1|S,\underline{\theta},\lambda)\}$, where $B_1 = B^c \cap \{n_s > n^t, s = 1, ..., k\}$. Inequality (3) will therefore be satisfied if

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

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 variables 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 $s \leq k$,

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

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

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

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

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 C , 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

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

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

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

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 $\theta_0$. This achieves the proof of Theorem (2). □

# APPENDIX C : PROOF OF INEQUALITY (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 $\theta$. Then

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

where $Z_n(\theta) = n^{-1/2}\{l_n(\theta) - l_n(\theta_0) + nK(\theta_0,\theta)\}$. 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 $\theta \in \tilde{A}_n$,

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

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

Dimatteo I, Genovese C.R. & Kass R.E. (2001). Bayesian curve fitting with free-knot splines. *Biometrika*, **88**, 1055-1071.

Ghosh, J.K. (1994). Higher order asymptotics. *NSF-CBMS Regional Conference Series in Probability and Statistics*, **4**, ed. IMS. ASA.

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.

Liu, J.S, Wong, W.H. & Kong, A. (1994). Covariance structure of the Gibbs sampler with applications to the comparisons of estimators and augmentation schemes. *Biometrika*, **81**, 27-40.

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. & Kadane, J.B. (1986). : Accurate Approximations For Posterior Moments and Marginal Densities. *J. Amer. Statist. Assoc.*, **81**, 82-86.

Tierney, L., Kass, R.E. & Kadane, J.B. (1989) : Approximate marginal densities of nonlinear functions (Corr. vol. 78, 233-234), *Biometrika*, **76**, 425-433.

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