Maximum a Posteriori Parameter Estimation for Hidden Markov Models

By Arnaud Doucet
Signal Processing Group, University of Cambridge
Trumpington St., CB2 1PZ Cambridge, UK
e-mail: ad2@eng.cam.ac.uk

and Christian P. Robert
Laboratoire de Statistique, CREST, INSEE
92241 Malakoff cedex, France
e-mail: robert@ensae.fr

Summary

An iterative stochastic algorithm to perform maximum *a posteriori* parameter estimation of hidden Markov models is proposed. It makes the most of the statistical model by introducing an artificial probability model based on an increasing number of the unobserved Markov chain at each iteration. Under minor regularity assumptions, we provide sufficient conditions to ensure global convergence of this algorithm. It is applied to parameter estimation for finite Gaussian mixtures, Markov-modulated Poisson processes and switching autoregressions with a Markov regime.

Some key words: Bayesian estimation; Data augmentation; Hidden Markov models; Maximum a posteriori; Simulated annealing.

1 Introduction

1.1 Background

Hidden Markov models (hereafter abbreviated as HMM) are specific latent variable models where the completed (or augmented) model is directed by an unobserved Markov chain x_t, that is, conditionally on x_t and on the past observations $y_{1:t-1} \triangleq (y_1, \ldots, y_{t-1})$, the observed quantity y_t is distributed as $y_t \sim f(y_t | y_{1:t-1}, x_t)$ where $f(\cdot)$ is a parameterized density dependent on an unknown parameter θ. The Markov chain x_t is often restricted to be in a finite state space, as in modellings of DNA (Durbin, Krogh & Mitchison, 1998), but extensions to continuous state spaces are also of interest, as in stochastic volatility models (Shephard, 1994). HMM's have a wide ranging number of applications, from Econometrics (Hamilton, 1989; Chib, 1996) to Signal Processing (Rabiner, 1989), as recalled in Robert & Casella (1999, §9.5.1).

While likelihood and pseudo-likelihood methods have been developed around these models, mainly as stochastic extensions of the EM algorithm (Dempster, Laird & Rubin, 1977), like Stochastic EM (Celeux & Diebolt, 1985) or Monte Carlo EM (MCEM) (Chib, 1996; Quian & Titterington, 1991; Wei & Tanner, 1990), Bayesian approaches have been concerned with the derivation of posterior distributions through MCMC algorithms (Chib, 1996; Robert & Casella, 1999, § 9.5.1).

The samples produced by these algorithms are quite appropriate to approximate many aspects of the posterior distribution, in particular to derive minimum mean square estimates via ergodic averages, but they are rather inefficient, *per se*, when considering maximum *a posteriori* (hereafter abbreviated as MAP) estimates of the unknown parameter θ, that is, $\theta_{MAP} = \arg\max p(\theta | y_{1:T})$, where $p(\theta | y_{1:T})$ is the posterior distribution associate with the observations $y_{1:T}$ (rather than the completed posterior). Indeed, MCMC algorithms do not usually focus on high posterior density regions, since they aim at producing acceptable approximations to the whole distribution. Thus a large amount of the computational burden is spent exploring regions of rather low posterior probability, which are of no interest for MAP estimation.

In order to achieve efficient computation of the MAP estimate for parameters of HMM's, it is thus necessary to design a stochastic algorithm which converges asymptotically, under given conditions, towards the set of these estimates.

1.2 Resolution

The algorithm we propose in this paper is a simulated annealing algorithm in the sense that we consider increasing powers of the posterior density $p(\theta | y_{1:T})$ in order to get an almost sure concentration of the simulated $\theta^{(i)}$ around the MAP estimates. However, the standard simulated annealing algorithms can only provide MAP estimates for the joint posterior distribution $p(\theta, x_{1:T} | y_{1:T})$, which includes the missing data (or latent variable) $x_{1:T}$ as well as the parameter θ. It differs from the MAP estimate of the sole parameter θ. The reason for this discrepancy is that simulated annealing algorithms rely on the Metropolis-Hastings (M-H) algorithm (Van Laarhoven & Arts, 1987) or the Gibbs sampler (Geman & Geman, 1984) and therefore require the introduction of the missing data to take the appropriate power of the completed likelihood.

Our resolution of the MAP problem also uses the missing data but in an artificial way which allows for convergence to the marginal MAP, rather than the joint MAP. As in Robert & Titterington (1998), the idea is to duplicate the missing data an increasing number of times as the iterations go by, the number of replicas playing the role of a (statistically significant) cooling schedule.

This *state augmentation for marginal estimation* (SAME) strategy is very general and can be applied to numerous missing data models as shown by Doucet, Godsill & Robert (2000). Here, we restrict ourselves to the study of the resulting SAME algorithm for the important class of HMM models in order to derive the proper cooling rates for the method. The interest of introducing the missing data in the algorithm is twofold. First, from an algorithmic point of view, it allows for a straightforward update in the parameters. Second, from a theoretical point of view, the finite nature of the missing data set ensures, for a suitable cooling schedule, a global convergence result of the algorithm towards the set of MAP parameters, with no restriction on the continuous parameters of the model, which may thus vary in a non-compact set; standard convergence results on simulated annealing algorithms in continuous state space are limited to compact sets (Belisle, 1992; Haario & Sacksman, 1991).

1.3 Plan

The paper is organized as follows. Section 2 formally presents the model and the notations. Section 3 introduces the SAME algorithm and discuss implementation issues. Section 4 gives sufficient conditions to ensure convergence of this scheme towards the set of MAP parameters and recovers the usual logarithmic rate of simulated annealing methods. Section 5 demonstrates the applicability of this algorithm for finite Gaussian mixtures, Markov-modulated Poisson processes, and switching Markov autoregressions examples, illustrated by real datasets.

2 Estimation in HMM's

We consider observations y_t's such that, conditional on the past $y_{1:t-1}$ and a latent variable $x_t \in \{1, \ldots, s\}$,

$$y_t | \left(y_{1:t-1}, x_t\right) \sim f\left(y_t | y_{1:t-1}, \theta_{x_t}\right)$$

The latent variable x_t is distributed from a stationary s-state Markov chain with transition matrix $\Pi = (\pi_{jk})$ where $\pi_{jk} \triangleq \Pr\{x_{t+1} = k | x_t = j\}$, $\Pi_j \triangleq (\pi_{j1}, ..., \pi_{js})$, $j, k \in S \triangleq \{1, 2, ..., s\}$. We assume here that $f(\cdot)$ belongs to an exponential family and denote it by

$$f\left(y_t | y_{1:t-1}, \theta_j\right) = h_t\left(y_{1:t}\right) \exp\left\{\theta_j . \varphi_t\left(y_{1:t}\right) - \psi_t\left(\theta_j\right)\right\}$$

(The dependence of h, φ and ψ on t will be omitted in the sequel to improve the presentation.) The set of parameters to estimate is $\theta \triangleq \{(\theta_j, \Pi_j); j \in S\} \in \Theta \subset \mathbb{R}^{n_\theta}$.

For simplicity's sake, we assume that the parameters $\{(\theta_j, \Pi_j); j \in S\}$ are *a priori* independent, distributed from a conjugate proper prior distribution on θ_j and a Dirichlet prior for Π_j

$$p(\theta_j) \propto \exp\left\{\theta_j.\varphi_j - \lambda_j \psi(\theta_j)\right\}, \Pi_j \sim \mathcal{D}_s(\alpha_j)$$

where $\alpha_j \triangleq (\alpha_{j1}, ..., \alpha_{js})$.

Given the observations $y_{1:T}$, our objective is to obtain the MAP estimate θ_{MAP} of θ, that is,

$$\theta_{MAP} \triangleq \arg\max_{\theta \in \Theta} p(\theta | y_{1:T})$$

Since $p(\theta | y_{1:T})$ involves s^T terms, this estimation problem requires to solve a complex global optimization problem on a continuous non-compact state space.

Note that MAP estimates are much more appropriate in this setting than posterior means, as shown in Celeux *et al.* (2000) for mixtures of distributions. Indeed, if the prior modelling implies that the components are exchangeable, the posterior means are all equal over the possible states, while identifiability constraints induce strong bias in the posterior means.

3 State Augmentation for Marginal Estimation

3.1 Marginalisation by augmentation

The basic idea of our iterative algorithm is based on simulated annealing: we construct an inhomogeneous Markov chain whose invariant distribution at iteration i is proportional to $p^{\gamma(i)}(\theta | y_{1:T})$ where $\gamma(i)$ goes to infinity with i. Under suitable regularity conditions (Hwang, 1980), $p^{\gamma(i)}(\theta | y_{1:T})$ get concentrated on the set of global maxima of $p(\theta | y_{1:T})$. In practice, we define at iteration i an artificial probability model whose marginal distribution is the concentrated distribution by artificially replicating the missing data set in the model.

More specifically, consider the model with $\gamma(i) \in \mathbb{N}^*$ artificial replications of $x_{1:T}$, denoted by $x_{1:T}(1), \ldots, x_{1:T}\{\gamma(i)\}$. Each of these replications is now treated as a distinct and independent missing data set, leading to the joint distribution

$$q_{\gamma(i)}\left[\theta, x_{1:T}(1), \ldots, x_{1:T}\{\gamma(i)\} \mid y_{1:T}\right] \propto \prod_{k=1}^{\gamma(i)} p\left\{\theta, x_{1:T}(k) \mid y_{1:T}\right\} \qquad (1)$$

By construction, the marginal for θ in this distribution is

$$\begin{aligned} q_{\gamma(i)}(\theta \mid y_{1:T}) &\propto \int \prod_{k=1}^{\gamma(i)} p\left\{\theta, x_{1:T}(k) \mid y_{1:T}\right\} dx_{1:T}(1) \ldots dx_{1:T}\{\gamma(i)\} \\ &\propto p^{\gamma(i)}(\theta \mid y_{1:T}) \end{aligned}$$

Thus, if we build a non-homogeneous MCMC algorithm with a transition kernel for which the invariant distribution at iteration i is proportional to (1) then, marginally, $q_{\gamma(i)}(\theta \mid y_{1:T})$ is also the invariant distribution. This so-called SAME strategy can be applied to many missing data problems (Doucet, Godsill & Robert, 2000).

3.2 The SAME Algorithm

The algorithm proceeds as follows.

---

1. Initialization, $i = 0$. Set randomly $\theta^{(0)} \in \Theta$.
2. Iteration i, $i \geq 1$
   - Sample the $\gamma(i)$ missing data sets $\left[x_{1:T}(1), \ldots, x_{1:T}\{\gamma(i)\}\right]$ according to $x_{1:T}^{(i)}(k) \sim p\left\{x_{1:T}(k) \mid y_{1:T}, \theta^{(i-1)}\right\}$ for $k = 1, \ldots, \gamma(i)$.
   - Sample $\theta^{(i)} \sim q_{\gamma(i)}\left[\theta \mid y_{1:T}, x_{1:T}^{(i)}(1), \ldots, x_{1:T}^{(i)}\{\gamma(i)\}\right]$.

---

where $\gamma(i)$ is an increasing sequence of integers.

The different steps of the algorithm are detailed in the following subsections. In order to simplify notation, we drop the superscript $\cdot^{(i)}$ from all variables at iteration i.

3.3 Implementation issues

The algorithm in §3.2 requires to sample from both $p(x_{1:T} \mid y_{1:T}, \theta)$ and

$$q_{\gamma(i)}\left[\theta \mid y_{1:T}, x_{1:T}^{(i)}(1), \ldots, x_{1:T}^{(i)}\{\gamma(i)\}\right].$$

The distribution $p\left(x_{1:T}\right| y_{1:T},\theta)$ is discrete and can be generated using the forward filtering-backward sampling recursion introduced independently by Carter & Kohn (1994) and Chib (1996). The computational cost of this filter is $O\left(s^2T\right)$.

Since

$$q_{\gamma(i)}\left[\theta | \, y_{1:T}, x_{1:T}^{(i)}\left(1\right), \ldots, x_{1:T}^{(i)}\left\{\gamma\left(i\right)\right\}\right] \propto \prod_{k=1}^{\gamma(i)} p\left\{\theta | \, y_{1:T}, x_{1:T}\left(k\right)\right\}$$

and

$$p\left\{\theta | \, y_{1:T}, x_{1:T}\left(k\right)\right\} = \prod_{j=1}^{s} p\left\{\theta_j | \, y_{1:T},x_{1:T}\left(k\right)\right\} p\left\{\Pi_j | \, y_{1:T},x_{1:T}\left(k\right)\right\}$$

we obtain the full conditional distribution of the transition probabilities

$$\begin{aligned}
&\Pi_j | \, y_{1:T}, x_{1:T}\left(1\right), \ldots, x_{1:T}\left\{\gamma\left(i\right)\right\} \\
&\sim \mathcal{D}_s\left\{\gamma\left(i\right)\left(\alpha_{j1}-1\right)+1+\sum_{k=1}^{\gamma(i)} n_{j1}\left(k\right), ..., \gamma\left(i\right)\left(\alpha_{js}-1\right)+1+\sum_{k=1}^{\gamma(i)} n_{js}\left(k\right)\right\}
\end{aligned}$$

where $n_{jl}\left(k\right) = \sum_{t=1}^{T-1}\delta\left\{j, x_t\left(k\right)\right\}\delta\left\{l, x_{t+1}\left(k\right)\right\}$ ($\delta\left(p,q\right)=1$ if $p=q$ and 0 otherwise) is the total number of one-step transitions from state j to state l in $x_{1:T}\left(k\right)$, $j,l \in S$. For the parameters θ_j, we obtain

$$\begin{aligned}
&q_{\gamma(i)}\left[\theta_j | \, y_{1:T}, x_{1:T}\left(1\right), \ldots, x_{1:T}\left\{\gamma\left(i\right)\right\}\right] \\
&\propto \exp\left[\theta_j \cdot \left\{\gamma\left(i\right)\varphi_j + \sum_{k=1}^{\gamma(i)}\sum_{t=1}^{T}\delta_{x_t(k),j}\varphi\left(y_{1:t}\right)\right\}\right. \\
&\qquad \left. - \left\{\gamma\left(i\right)\lambda_j + \sum_{k=1}^{\gamma(i)}\sum_{t=1}^{T}\delta_{x_t(k),j}\right\}\psi\left(\theta_j\right)\right]
\end{aligned} \tag{2}$$

Sampling from (2) is problem dependent and typically relies on classical techniques, since (2) is within an exponential family.

4 Convergence properties

In this section, we prove that, if $\gamma\left(i\right)$ is an appropriate increasing sequence (see THEOREM 1), the proposed algorithm converges to the set of MAP parameters,

$$\lim_{i\to+\infty}\left\|\mu_i - q_{\gamma(i)}\right\| = 0 \tag{3}$$

where $\|\cdot\|$ is the total variation norm, μ_i is the distribution of the i^{th} sample $\theta^{(i)}$ of the inhomogeneous Markov chain built from our algorithm and $q_{\gamma(i)}$ $(i \in \mathbb{N})$ is the sequence of normalized versions of $p^{\gamma(i)}(\theta | y_{1:T})$. This sequence converges to a distribution located on the set of MAP parameters.

4.1 Assumptions

Let $\overset{\circ}{\Theta}$ denote the interior set of Θ. We assume that the following assumptions hold for $p(\theta | y_{1:T})$:
Assumption 1. For any $x_{1:T} \in S^T$, $p(\theta | y_{1:T}, x_{1:T})$ is bounded from above.
Assumption 2. The set of global maxima denoted $\Theta_{MAP} = \left\{\theta^1_{MAP}, \ldots, \theta^p_{MAP}\right\}$ is finite and is included in $\overset{\circ}{\Theta}$. The Hessian matrix

$$\left\{-\frac{\partial^2 \log p(\theta | y_{1:T})}{\partial\theta_i \partial\theta_j}\right\}$$

is non singular at any point of Θ_{MAP}.
Assumption 3. $-\int_\Theta p(\theta | y_{1:T}) \log\{p(\theta | y_{1:T})\} d\theta < +\infty$.

4.2 Transition kernels and invariant distributions

By construction, $\left\{\theta^{(i)}; i \in \mathbb{N}\right\}$ is an inhomogeneous Markov chain with transition kernel $K_i(\theta,\theta')$ at iteration i equal to

$$\sum_{x_{1:T}(1),\ldots,x_{1:T}\{\gamma(i)\}} p[\theta' | y_{1:T}, x_{1:T}(1), \ldots, x_{1:T}\{\gamma(i)\}] \\ p[x_{1:T}(1), \ldots, x_{1:T}\{\gamma(i)\} | y_{1:T}, \theta] \quad (4)$$

and $q_{\gamma(i)}(\theta | y_{1:T})$ is its invariant density. For $i \in \mathbb{N}$, it can be shown that this kernel is uniformly ergodic using the *Duality Principle* (Robert, Celeux & Diebolt, 1993). Using Assumption 1, we obtain here a more precise result by showing that there exist constants $0 \leq M < +\infty$, $\delta \in (0,1)$ and $i_0 \in \mathbb{N}$ such that, for all $\theta \in \Theta$ and $i \geq i_0$,

$$K_i(\theta, d\theta') \geq M\delta^{\gamma(i)}\nu_i(d\theta') \quad (5)$$

where ν_i is a probability measure on $(\Theta, \mathcal{B}(\Theta))$ (see Appendix A-1). In other words, (5) establishes that an homogeneous Markov chain with kernel K_i would converge geometrically fast towards $q_{\gamma(i)}$, whatever the initial value.

Assumption 2 ensures that the sequence of target distributions $\left(q_{\gamma(i)}\right)_i$ converges to a probability distribution located on the set of global maxima

Θ_{MAP}

$$q_{\infty}\left(d\theta\right)=\frac{\sum_{l=1}^{p}\alpha\left(\theta_{l}^{MAP}\right)\delta_{\theta_{l}^{MAP}}\left(d\theta\right)}{\sum_{j=1}^{p}\alpha\left(\theta_{l}^{MAP}\right)} \tag{6}$$

where

$$\alpha\left(\theta_{l}^{MAP}\right)\triangleq\left[\det\left\{-\left.\frac{\partial^{2}\log p\left(\theta|\,y_{T}\right)}{\partial\theta_{m}\partial\theta_{n}}\right|_{\theta=\theta_{l}^{MAP}}\right\}\right]^{-1/2}$$

see for instance Hwang (1980).

Our algorithm is based on an inhomogeneous Markov chain. So as to obtain (3), one first needs to bound the total variation norm between μ_i and $q_{\gamma(i)}$. Using (5), we get for any positive integers $i_0 < k < i$ (see Appendix A-2)

$$\left\|\mu_{i}-q_{\gamma(i)}\right\|\leq\prod_{l=k+1}^{i}\left(1-M\delta^{\gamma(l)}\right)+\sum_{l=k}^{i-1}\left\|q_{\gamma(l+1)}-q_{\gamma(l)}\right\| \tag{7}$$

The second term on the right hand side of (7) involves the total variation norm between the successive invariant distributions. Appendix A-3 shows that it can be bounded from above and satisfies

$$\sum_{l=k}^{i-1}\left\|q_{\gamma(l+1)}-q_{\gamma(l)}\right\|\leq 2\log\left[\frac{Z\left\{\gamma\left(k-1\right)\right\}}{Z\left\{\gamma\left(i-1\right)\right\}}\right] \tag{8}$$

where

$$Z\left(\gamma\right)\triangleq\int_{\Theta}\left\{\frac{p\left(\theta|\,y_{1:T}\right)}{p\left(\theta_{MAP}|\,y_{1:T}\right)}\right\}^{\gamma}d\theta=\left(2\pi\gamma\right)^{n_{\theta}/2}\left\{\sum_{l=1}^{p}\alpha\left(\theta_{l}^{MAP}\right)+\varepsilon\left(\gamma\right)\right\} \tag{9}$$

with $\lim_{\gamma\rightarrow+\infty}\varepsilon\left(\gamma\right)=0$.

Our main result is the following theorem, which shows that it is possible to find logarithmic cooling schedules $\gamma\left(i\right)$ to ensure that the right hand side of (7) goes to zero as i increases, see Appendix A-4.

Theorem 1 *For any $\varepsilon\in(0,1)$, if*

$$\gamma\left(i\right)=\max\left[1,\left\lfloor-\left\{\left(1+\varepsilon\right)\log\left(\delta\right)\right\}^{-1}\log\left(i+\gamma_{0}\right)\right\rfloor\right]$$

(where $\gamma_0 > 0$) then, for any initial distribution μ_0 of $\theta^{(0)}$, the Markov chain generated by the SAME algorithm converges towards the set of global maxima in the following sense

$$\lim_{i\rightarrow+\infty}\left\|\mu_{i}-q_{\gamma(i)}\right\|=0 \tag{10}$$

This result implies that for any $\xi > 0$

$$\lim_{i\to+\infty} \Pr\left[\log\left\{\frac{p\left(\theta_{MAP}\right|y_{1:T})}{p\left(\theta^{(i)}\middle|y_{1:T}\right)}\right\} < \xi\right] = 1$$

5 Applications

A logarithmic cooling schedule is required by Theorem 1 to ensure the global convergence of the SAME algorithm towards the set of MAP parameters. In practice, as for classical simulated annealing algorithms, this schedule goes too slowly towards infinity. In all the examples addressed here, we implemented, as usually done in practice (Van Laarhoven & Arts, 1987), $N_0 = 200$ iterations of the SAME algorithm with $\gamma(i) = 1$ for $i = 1, \ldots, 100$ and then we use a *linear* cooling schedule, *i.e.* $\gamma(i) = \lfloor ai + b \rfloor$, satisfying $\gamma(100) = 1$ and $\gamma(N_0) = 200$. We compared through numerical simulations the SAME algorithm with the EM and MCEM algorithms. As in Chib (1996), the MCEM algorithm was first run for 100 iterations using the SEM algorithm then 1000 samples were used at each iteration in the MCEM steps until stabilization of the estimates. In all cases, the algorithms were initialized with the same random parameter $\theta^{(0)}$ and we took as final estimate of θ^{MAP} the parameter $\theta^{(i)}$ maximizing $p(\theta|\, y_{1:T})$.

5.1 Finite Gaussian mixtures

Consider the special case of finite Gaussian mixture distributions as in Robert and Mengersen (1999). In this case, we have T independent identically distributed observations $y_1, \ldots, y_T$, $y_t|\, x_t \sim \mathcal{N}\left(m_{x_t}, \sigma^2_{x_t}\right)$ and $\Pr(x_t = j) = \pi_j$, independently of x_{t-1}. We want to estimate $\theta \triangleq \left\{\left(m_j, \sigma_j^2, \pi_j\right); j \in S\right\}$. To complete the Bayesian model, we assume that the $\left(m_j, \sigma_j^2\right)$, $j \in S$, are distributed from the conjugate priors

$$m_j|\, \sigma_j^2 \sim \mathcal{N}\left(\alpha_j, \sigma_j^2/\lambda_j\right), \ \sigma_j^2 \sim \mathcal{IG}\left(\frac{\lambda_j + 3}{2}, \frac{\beta_j}{2}\right)$$

although more involved priors could be considered as well. In the case of independent priors, it is well-known that one must use proper priors to get a proper posterior.

We propose an application of our approach to the benchmark galaxy dataset. First treated by Roeder (1992), this dataset has been analyzed by many authors, including Richardson & Green (1997), and Robert &

Mengersen (1999). It consists of $T = 82$ observations of galaxy velocities and the evaluation of the number of components s for this dataset is controversial since estimates range from 3 for Roeder & Wasserman (1997) to 7 for Escobar & West (1995) and Phillips & Smith (1996). We set $s = 3$, the parameters of the Dirichlet to 1 and $\alpha_j = 0$, $\beta_j = 0.1$ and $\lambda_j = 0.1$, $j \in S$.

Assumption 1 and Assumption 3 are obviously satisfied. Assumption 2 is supposed to be verified: As pointed out in Robert & Titterington (1996), while the likelihood function is not bounded, the inclusion of the prior has a bounding effect and MAP estimates do exist. While the parameter θ is not identifiable, since the mixture distribution is invariant under permutations of the components, it is identifiable under ordering of the mean, variance or weight parameters and thus leads to a finite number of global maxima (for T large enough).

Table 1 gives the mean and standard deviations of these values the log-posterior distribution of the MAP estimates obtained using EM, MCEM and SAME for 50 simulations.

Table 1 about here

In this experiment, the SAME algorithm clearly outperforms EM and MCEM: the former converges to the same mode whatever the initialization point, as shown by the very small variance in Table 1, and the value of the mode obtained by SAME is always higher than the values produced by EM and MCEM. In fact, this example illustrates an advantage of SAME over both EM and MCEM. In about 30% of the simulations, MCEM reaches one stage where no observation is allocated to a given component of the mixture, say component k, so that, at iteration i, in the M-step one obtains $\pi_k^{(i)} = 0$. This is a "trapping state" for MCEM since $\pi_k^{(l)} = 0$ for $l \geq i$. While the EM algorithm cannot give exactly $\pi_k^{(i)} = 0$, it often converges to such values, which correspond to severe local maxima. On the contrary, the SAME algorithm may encounter these values but succeeds to escape from these states. Fig. 1 illustrates such a case on a sequence of weights, where the values of $\pi_k^{(l)}$ come close to 0 and then get away from this value in a few iterations for the SAME sequence. Fig. 2 illustrates the posterior density values against iteration number for EM and SAME. (The graph for MCEM is not represented as it is undistinguishable from EM after a few iterations.)

Figure 1 about here

Figure 2 about here

5.2 Markov-modulated Poisson processes

We address here the case of Markov-modulated Poisson processes (Chib, 1996; Leroux & Puterman, 1992; Robert & Titterington, 1996). We observe counting data which is modeled as a Poisson process whose parameter switches according to a hidden Markov chain. More formally, x_t is an unobserved finite Markov chain with s states, $y_t|\, x_t \sim \mathcal{P}(\lambda_{x_t})$ and we want to estimate $\theta \triangleq \{(\lambda_j, \Pi_j); j \in S\}$. To complete the Bayesian model, we assume that the parameters λ_j, $j \in S$, have Gamma priors $\lambda_j \sim \mathcal{G}(a, b)$.

We consider the foetal movement data analyzed in Chib (1996) and Leroux & Puterman (1992). The data consists of the numbers of movements of a foetal lamb in $T = 240$ consecutive five-seconds intervals. We set $s = 2$, $\alpha_j = (1, \ldots, 1)$, $a = 1$, $b = 0.1$. Assumption 1 as well as Assumption 3 are satisfied. Assumption 2 is supposed to be verified, for the same reasons as above.

Table 2 gives the means and standard deviations of the log-posterior values of the MAP estimates obtained using EM, MCEM and SAME for 50 simulations, initialized from the prior distribution.

Table 2 about here

Similarly to finite Gaussian mixtures, EM and MCEM appear often trapped in configurations where one state of the Markov chain is not visited; besides, they are highly sensitive to the initial value, as shown by the variances in Table 2. The SAME algorithm avoids such trapping states and constantly provides an higher evaluation of the posterior maximum. Fig. 3 illustrates such a case on the sequence of intensities. The EM and MCEM algorithms converge towards severe local maxima whereas SAME converges towards a different point estimate corresponding to the results of Chib (1996). Fig. 4 presents the posterior density values against iteration number for EM and SAME (the same phenomenon occurs for MCEM, namely that it cannot be distinguished from EM after a few iterations).

Figure 3 about here

Figure 4 about here

5.3 Switching Markov autoregressions

Now consider a switching autoregression process with a hidden Markov regime (Hamilton, 1989)

$$y_t = \mu_{x_t} + y'_{t-1:t-p} a_{x_t} + \sigma_{x_t} v_t \qquad t = 1, \ldots, N$$

where $y_{t-1:t-p} \triangleq (y_{t-1}, \ldots, y_{t-p})'$, $y_{0:1-p} = (0, \ldots, 0)'$, $a_j \triangleq (a_{j,1}, \ldots, a_{j,p})'$ for $j \in S$, and v_t is an i.i.d. Gaussian sequence of variance 1. We want to estimate $\theta = \left\{\left(\mu_j, a_j, \sigma_j^2, \Pi_j\right); j \in S\right\}$. We assume that the parameters $\left(a_j, \sigma_j^2\right)$ are distributed according to a normal-inverse Gamma prior distribution

$$\mu_j|\, \sigma_j^2 \sim \mathcal{N}\left(0, \sigma_j^2 \zeta_j^2\right), \ a_j|\, \sigma_j^2 \sim \mathcal{N}\left(0, \sigma_j^2 \Sigma_{0,j}\right), \ \sigma_j^2 \sim \mathcal{IG}\left(\frac{\nu_{0,j}}{2}, \frac{\eta_{0,j}}{2}\right)$$

Assumption 1 and Assumption 3 are satisfied. Assumption 2 is assumed to be verified, given the complexity of the model.

We consider the dataset on quaterly U.S. real GNP analyzed previously in Albert & Chib (1993), Chib (1996) and Hamilton (1989). The variable of interest is the percentage change in the postwar real GNP for the period Feb. 1951 to April 1984. We set $s = 4$, $p = 4$, $\alpha_j = (1, \ldots, 1)$ and $\Sigma_{0,j} = 10^3 I_p$, $\zeta_j^2 = 10^3$, $\nu_{0,j} = \eta_{0,j} = 0.1$, $b_j = 0.1$, $j \in S$.

The values of the log-posterior densities of the MAP estimates obtained using EM, MCEM and SAME for 50 simulations are presented in Table 3, which gives their means and standard deviations. As in the two previous examples, the SAME algorithm appears to give better results with a lower variability in the variation of the log-posterior values.

Table 3 about here

Acknowledgements

We are thankful to Sid Chib for providing us with the datasets of Sections 5.2 and 5.3, to the Department of Statistics, University of Glasgow, for providing support and welcome during a visit of both authors in February 2000, and to the EU TMR network ERB–FMRX–CT96–0095 on *Computational and Statistical Methods for the Analysis of Spatial Data* for supporting several visits in 1999 and 2000 during which this research was conducted.

Appendix

We use the following standard definitions and notations. Let $\{\theta^{(i)}; i \in \mathbb{N}\}$ be an inhomogeneous Markov chain on $\{\Theta, \mathcal{B}(\Theta)\}$ with initial distribution μ_0 and Markov transition kernel K_i at iteration i satisfying, for any $A \in \mathcal{B}(\Theta)$,

$$\Pr\left\{ \theta^{(i)} \in A \middle| \theta^{(i-1)} \right\} = \int_A K_i\left(\theta^{(i-1)}, d\theta\right)$$

We denote, for $m < i$, $K^{(m,i)} \triangleq K_{m+1} K_{m+2} ... K_i$ so that the probability distribution μ_i of $\theta^{(i)}$ is defined by $\mu_i = \mu_{i-1} K_i = \mu_k K^{(k,i)} = \mu_0 K^{(0,i)}$.

A-1: Minorization condition (5)

Let us denote $x_{1:T}\{1 : \gamma(i)\} \triangleq [x_{1:T}(1), \ldots, x_{1:T}\{\gamma(i)\}]$. Instead of obtaining a uniform minorization condition on $K_i\left(\theta, d\theta'\right)$, we establish it on the dual kernel defined as

$$
\begin{aligned}
& K_i\left[x_{1:T}\left\{1:\gamma\left(i\right)\right\},x'_{1:T}\left\{1:\gamma\left(i\right)\right\}\right] \qquad (11)\\
&= \int_{\Theta}\prod_{k=1}^{\gamma(i)} p\left\{\left.x'_{1:T}\left(k\right)\right| y_{1:T},\theta\right\} q_{\gamma(i)}\left[\left.\theta\right| y_{1:T},x_{1:T}\left\{1:\gamma\left(i\right)\right\}\right] d\theta
\end{aligned}
$$

Then the *Duality Principle* (Robert, Celeux & Diebolt, 1993; Robert & Casella, 1999, §7.2.4) ensures that this uniform minorization can be transfered to $K_i\left(\theta,d\theta'\right)$. Using Assumption 1, one obtains

$$
\begin{aligned}
p\left(\left.x'_{1:T}\right| y_{1:T},\theta\right) &= \frac{p\left(\left.\theta\right| y_{1:T},x'_{1:T}\right) p\left(\left.x'_{1:T}\right| y_{1:T}\right)}{\sum_{x_{1:T}\in S^T} p\left(\left.\theta\right| y_{1:T},x_{1:T}\right) p\left(\left.x_{1:T}\right| y_{1:T}\right)}\\
&\geq \frac{p\left(\left.\theta\right| y_{1:T},x'_{1:T}\right) p\left(\left.x'_{1:T}\right| y_{1:T}\right)}{Cs^T}
\end{aligned}
$$

Hence

$$
\begin{aligned}
& K_i\left[x_{1:T}\left\{1:\gamma\left(i\right)\right\},x'_{1:T}\left\{1:\gamma\left(i\right)\right\}\right]\\
&\geq \frac{\prod_{k=1}^{\gamma(i)} p\left\{\left.x'_{1:T}\left(k\right)\right| y_{1:T}\right\}}{\left(Cs^T\right)^{\gamma(i)}} \int \left[\prod_{k=1}^{\gamma(i)} p\left\{\left.\theta\right| y_{1:T},x'_{1:T}\left(k\right)\right\}\right] q_{\gamma(i)}\left[\left.\theta\right| y_{1:T},x_{1:T}\left\{1:\gamma\left(i\right)\right\}\right] d\theta
\end{aligned}
$$

Let

$$
\Phi\left(\varphi_j,\lambda_j\right) \triangleq \int \exp\left\{\theta_j.\varphi_j-\lambda_j\psi\left(\theta_j\right)\right\} d\theta_j
$$

Using elementary calculations, we obtain

$$
\begin{aligned}
& K_i\left[x_{1:T}\left\{1:\gamma\left(i\right)\right\},x'_{1:T}\left\{1:\gamma\left(i\right)\right\}\right] \qquad (12)\\
\geq\ & \frac{\prod_{k=1}^{\gamma(i)} p\left\{\left.x'_{1:T}\left(k\right)\right| y_{1:T}\right\}}{\left(Cs^T\right)^{\gamma(i)}}\\
& \times \prod_{j=1}^{s} \frac{\Phi\left\{n_j^{\gamma(i)}\varphi_j^{\gamma(i)}+n_j'^{\gamma(i)}\varphi_j'^{\gamma(i)}+2\gamma\left(i\right)\varphi_j, n_j^{\gamma(i)}+n_j'^{\gamma(i)}+2\gamma\left(i\right)\lambda_j\right\}}{\Phi\left\{n_j^{\gamma(i)}\varphi_j^{\gamma(i)}+\gamma\left(i\right)\varphi_j, n_j^{\gamma(i)}+\gamma\left(i\right)\lambda_j\right\}\prod_{k=1}^{\gamma(i)}\Phi\left(n_j'^{(k)}\varphi_j'^{(l)}+\varphi_j, n_j'^{(k)}+\lambda_j\right)}\\
& \times \prod_{j=1}^{s}\left[\frac{\Gamma\left\{n_j^{\gamma(i)}+s+\sum_{l=1}^{s}\gamma\left(i\right)\left(\alpha_{jl}-1\right)\right\}}{\prod_{l=1}^{s}\Gamma\left\{n_{jl}^{\gamma(i)}+1+\gamma\left(i\right)\left(\alpha_{jl}-1\right)\right\}}\prod_{k=1}^{\gamma(i)}\frac{\Gamma\left\{n_j'\left(k\right)+\sum_{l=1}^{s}\alpha_{jl}\right\}}{\prod_{l=1}^{s}\Gamma\left\{n_{jl}'\left(k\right)+\alpha_{jl}\right\}}\right.\\
& \left.\times \frac{\prod_{l=1}^{s}\Gamma\left\{n_{jl}^{\gamma(i)}+n_{jl}'^{\gamma(i)}+1+2\gamma\left(i\right)\left(\alpha_{jl}-1\right)\right\}}{\Gamma\left\{n_j^{\gamma(i)}+n_j'^{\gamma(i)}+s+\sum_{l=1}^{s}2\gamma\left(i\right)\left(\alpha_{jl}-1\right)\right\}}\right]
\end{aligned}
$$

where $n_j'^{\gamma(i)}\varphi_j'^{\gamma(i)} = \sum_{k=1}^{\gamma(i)} \sum_{t=1}^{T} \delta\{x_t'(k), j\} \varphi(y_{1:t})$, $n_j'^{(k)} = \sum_{t=1}^{T} \delta\{x_t'(k), j\}$ and $n_j'^{\gamma(i)} = \sum_{k=1}^{\gamma(i)} n_j'^{(k)}$. The quantities $n_j^{\gamma(i)}\varphi_j^{\gamma(i)}$, $n_j^{\gamma(i)}$ and $n_j^{(k)}$ are defined in a similar way.

Let us denote $\psi^+(\theta_j) \triangleq \max\{0, \psi(\theta_j)\}$ and $\psi^-(\theta_j) \triangleq \min\{0, \psi(\theta_j)\}$. For each component of the vector φ_j, there exist lower and upper bounds which are independent of $\gamma(i)$ on $\gamma^{-1}(i)\left(n_j^{\gamma(i)}\varphi_j^{\gamma(i)} + n_j'^{\gamma(i)}\varphi_j'^{\gamma(i)}\right)$. Thus, there exists $\overline{\varphi}_{j,1}$ such that

$$\begin{aligned}
&\Phi\left\{n_j^{\gamma(i)}\varphi_j^{\gamma(i)} + n_j'^{\gamma(i)}\varphi_j'^{\gamma(i)} + 2\gamma(i)\varphi_j, n_j^{\gamma(i)} + n_j'^{\gamma(i)} + 2\gamma(i)\lambda_j\right\} \qquad (13)\\
&\geq \int \exp\left(\gamma(i)\left[\theta_j.\left\{\frac{n_j^{\gamma(i)}\varphi_j^{\gamma(i)} + n_j'^{\gamma(i)}\varphi_j'^{\gamma(i)}}{\gamma(i)} + 2\varphi_j\right\} - 2\lambda_j\psi(\theta_j) - 2T\psi^+(\theta_j)\right]\right) d\theta_j \\
&\geq \int \exp\left[\gamma(i)\left\{\theta_j.\left(\overline{\varphi}_{j,1} + 2\varphi_j\right) - 2\lambda_j\psi(\theta_j) - 2T\psi^+(\theta_j)\right\}\right] d\theta_j
\end{aligned}$$

Let $h_1(\theta_j) = \theta_j.\left(\overline{\varphi}_j + 2\varphi_{j,1}\right) - 2\lambda_j\psi(\theta_j) - 2T\psi^+(\theta_j)$, and $\widehat{\theta}_{j,1} = \arg\max h_1(\theta_j)$. Using Laplace's approximation (Robert & Casella, 1999, §3.5), one can obtain the following asymptotic equivalent of the lower bound of (13)

$$\exp\left\{\gamma(i) h_1\left(\widehat{\theta}_{j,1}\right)\right\}\left(\det\left[-\frac{2\pi}{\gamma(i)}\left\{h_1''\left(\widehat{\theta}_{j,1}\right)\right\}^{-1}\right]\right)^{1/2} \qquad (14)$$

Similarly there exists $\overline{\varphi}_{j,2}$ independent of $\gamma(i)$ such that

$$\begin{aligned}
&\Phi\left\{n_j^{\gamma(i)}\varphi_j^{\gamma(i)} + \gamma(i)\varphi_j, n_j^{\gamma(i)} + \gamma(i)\lambda_j\right\} \\
&\leq \int \exp\left[\gamma(i)\left\{\theta_j.\left(\overline{\varphi}_{j,2} + \varphi_j\right) - T\psi^-(\theta_j) - \lambda_j\psi(\theta_j)\right\}\right] d\theta_j \\
&\simeq \exp\left\{\gamma(i) h_2\left(\widehat{\theta}_{j,2}\right)\right\}\left(\det\left[-\frac{2\pi}{\gamma(i)}\left\{h_2''\left(\widehat{\theta}_{j,1}\right)\right\}^{-1}\right]\right)^{1/2} \qquad (15)
\end{aligned}$$

with $h_1(\theta_j) = \theta_j.\left(\overline{\varphi}_{j,2} + \varphi_j\right) - T\psi^-(\theta_j) - \lambda_j\psi(\theta_j)$, and $\widehat{\theta}_{j,2} = \arg\max h_2(\theta_j)$. Finally there exists $\overline{\varphi}_{j,3}$ independent of $\gamma(i)$ such that

$$\begin{aligned}
&\prod_{k=1}^{\gamma(i)} \Phi\left(n_j'^{(k)}\varphi_j'^{(l)} + \varphi_j, n_j'^{(k)} + \lambda_j\right) \\
&= \prod_{k=1}^{\gamma(i)} \int \exp\left\{\theta_j.\left(n_j'^{(k)}\varphi_j'^{(l)} + \varphi_j\right) - \left(n_j'^{(k)} + \lambda_j\right)\psi(\theta_j)\right\} d\theta_j \\
&\leq \left[\int \exp\left\{\theta_j.\left(\overline{\varphi}_{j,3} + \varphi_j\right) - T\psi^-(\theta_j) - \lambda_j\psi(\theta_j)\right\} d\theta_j\right]^{\gamma(i)} \qquad (16)
\end{aligned}$$

One can obtain, through similar methods, exponential bounds in $\gamma(i)$ for the Gamma terms involved in (12) and (5) follows.

A-2: Total variation bound (7)

We have

$$\begin{aligned}\left\|\mu_i - q_{\gamma(i)}\right\| &= \left\|\mu_k K^{(k,i)} - q_{\gamma(i)}\right\| \\ &\leq \left\|\left(\mu_k - q_{\gamma(k)}\right) K^{(k,i)}\right\| + \left\|q_{\gamma(k)} K^{(k,i)} - q_{\gamma(i)}\right\|\end{aligned}$$

From (5), one obtains for $i_0 \leq k$

$$\left\|\left(\mu_k - q_{\gamma(k)}\right) K^{(k,i)}\right\| \leq \prod_{l=k+1}^{i} \left(1 - M\delta^{\gamma(l)}\right)$$

and one can check that

$$q_{\gamma(k)} K^{(k,i)} - q_{\gamma(i)} = \sum_{l=k}^{i-1} \left(q_{\gamma(l)} - q_{\gamma(l+1)}\right) K^{(l,i)}$$

Thus

$$\begin{aligned}\left\|q_{\gamma(k)} K^{(k,i)} - q_{\gamma(i)}\right\| &\leq \sum_{l=k}^{i-1} \left\|\left(q_{\gamma(l)} - q_{\gamma(l+1)}\right) K^{(l,i)}\right\| \\ &\leq \sum_{l=k}^{i-1} \left\|\left(q_{\gamma(l)} - q_{\gamma(l+1)}\right)\right\|\end{aligned}$$

A-3: Derivation of the bound (8)

We generalize here the proof of Haario & Sacksman (1991). To take the derivative of $Z(\gamma)$ with respect to γ, see (9), Haario & Sacksman (1991) assume that $-\log p(\theta|\, y_{1:T})$ is bounded above on Θ. This is clearly not the case for non-compact state-spaces. Here, we use instead Assumption 3. For any $\gamma(i) \in \mathbb{N}$ and all $\theta \in \Theta$,

$$\begin{aligned}0 &\geq \log\left\{\frac{p(\theta|\, y_{1:T})}{p(\theta_{MAP}|\, y_{1:T})}\right\} \exp\left[\gamma(i) \log\left\{\frac{p(\theta|\, y_{1:T})}{p(\theta_{MAP}|\, y_{1:T})}\right\}\right] \\ &\geq \log\left\{\frac{p(\theta|\, y_{1:T})}{p(\theta_{MAP}|\, y_{1:T})}\right\} \exp\left[\log\left\{\frac{p(\theta|\, y_{1:T})}{p(\theta_{MAP}|\, y_{1:T})}\right\}\right]\end{aligned}$$

We can thus apply the Lebesgue dominated convergence theorem and write

$$-\frac{dZ(\gamma)}{d\gamma} = -\int_{\Theta} \log\left\{\frac{p(\theta|\, y_{1:T})}{p(\theta_{MAP}|\, y_{1:T})}\right\} \exp\left[\gamma \log\left\{\frac{p(\theta|\, y_{1:T})}{p(\theta_{MAP}|\, y_{1:T})}\right\}\right] d\theta$$

This is the starting point of the proof of Theorem 3.2. in Haario & Sacksman (1991), which can then be reproduced in our setting and leads to (8).

A-4: Proof of Theorem 1

We want to prove that (10) is satisfied. Conditions (7) and (8) hold for any $i_0 < k < i$. Let choose a sequence k dependent on i, that is $k_i = \lfloor i - i^{1-\varepsilon/2} \rfloor$ where $\varepsilon \in (0,1)$. We show that, for this sequence k_i and the specified cooling schedule $\gamma(i)$, the two terms on the right hand side of (7) converge towards zero.
First term in (7):

$$\lim_{i\to\infty} \prod_{l=k_i+1}^{i} \left(1 - M\delta^{\gamma(l)}\right) = 0$$

is equivalent to $\lim_{i\to\infty} \sum_{l=k_i+1}^{i} \delta^{\gamma(l)} = 0$. Let us denote $a = (1+\varepsilon)^{-1}$; $\log(i+\gamma_0)$ being a monotonic increasing function, there exists i_1 such that for any $i \geq i_1$,

$$2 \leq \gamma(i) \leq -\{(1+\varepsilon)\log(\delta)\}^{-1} \log(i+\gamma_0)$$

Then, for any $k_i \geq i_1$

$$\begin{aligned}
\sum_{l=k_i+1}^{i} \delta^{\gamma(l)} &= \sum_{l=k_i+1}^{i} \exp\{\gamma(l)\log(\delta)\} \\
&\geq \sum_{l=k_i+1}^{i} \exp\left\{-(1+\varepsilon)^{-1}\log(l+\gamma_0)\right\} = \sum_{l=k_i+1}^{i} (l+\gamma_0)^{-a} \\
&\geq \int_{k_i+1}^{i} (x+\gamma_0)^{-a}\,dx = \frac{1}{a\varepsilon}\{(i+\gamma_0)^{a\varepsilon} - (k_i+1+\gamma_0)^{a\varepsilon}\} \\
&\geq \frac{1}{a\varepsilon}(i+\gamma_0)^{a\varepsilon}\left\{1 - \left(1 + \frac{\gamma_0 - (i+1)^{1-\varepsilon/2}}{i+\gamma_0}\right)^{a\varepsilon}\right\} \qquad (17)
\end{aligned}$$

which converges to infinity with i.
Second term in (7): From (9)

$$\lim_{i\to+\infty} \log\left[\frac{Z\{\gamma(k_i-1)\}}{Z\{\gamma(i-1)\}}\right] = \lim_{i\to+\infty} \frac{n_\theta}{2}\log\left\{\frac{\log\gamma(k_i-1)}{\log\gamma(i-1)}\right\} = 0 \qquad (18)$$

Now (10) follows straightforwardly from (17) and (18).

References

[1] Albert, J. & Chib, S. (1993). Bayes inference via Gibbs sampling of autoregressive time series subject to Markov mean and variance shifts. *J. Business Economic Statist.* **11**, 1-15.

[2] Belisle, C.J.P. (1992). Convergence theorems for a class of simulated annealing algorithms on $\mathbb{R}^d$. *J. Applied Prob.* **29**, 885-95.

[3] Carter, C.K. & Kohn, R. (1994). On Gibbs sampling for state space models. *Biometrika* **81**, 541-53.

[4] Celeux, G. & Diebolt, J. (1985). The SEM algorithm: a probabilistic teacher algorithm derived from the EM algorithm for the mixture problem. *Comp. Statist. Quaterly* **2**, 73-82

[5] Celeux, G., Hurn, M. & Robert, C.P.(2000). Computational and inferential difficulties with mixtures posterior distribution. *J. Am. Statist. Assoc.* **95** (to appear).

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

[7] Dempster, A.P., Laird, N.M. & Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. R. Statist. Soc. B* **39**, 1-38.

[8] Doucet, A., Godsill, S.J. & Robert, C.P. (2000). Marginal maximum a posteriori estimation using MCMC. Technical report CUED/F/INFENG TR 375.

[9] Durbin, R., Krogh, A. & Mitchison, G. (1998). *Biological Sequence Analysis.* Cambridge University Press.

[10] Escobar, M.D. & West, M. (1995) Bayesian density estimation and inference using mixtures. *J. Am. Statist. Assoc.* **90**, 577–588

[11] Geman, S. & Geman, D. (1984). Stochastic relaxation, Gibbs distribution and the Bayesian restoration of images. *IEEE Trans. Patt. Ana. Mac. Int.* **6**, 721-41.

[12] Haario, H. & Sacksman, E. (1991). Simulated annealing in general state space. *Adv. Appl. Prob.* **23**, 866-93.

[13] Hamilton, J.D. (1989). *Time Series Analysis.* Princeton University Press.

[14] Hwang, C. (1980). Laplace's method revisited: weak convergence of probability measures. *Ann. Prob.* **8**, 1177-82.

[15] Leroux, B.G., & Puterman, M.L. (1992). Maximum-penalized-likelihood estimation for independent and Markov-dependent mixture models. *Biometrics* **48**, 545-58.

[16] PHILLIPS, D.B., & SMITH, A.F.M. (1996) Bayesian model comparison via jump diffusions. In *Markov chain Monte Carlo in Practice*, W.R. Gilks, S.T. Richardson and D.J. Spiegelhalter (Eds.), 215-240. Chapman and Hall: London.

[17] QUIAN, W. & TITTERINGTON, D.M. (1991). Estimation of parameters in hidden Markov models. *Phil. Trans. R. Soc. London* **A 337**, 407-28.

[18] RABINER, L.R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proc. IEEE* **77**, 257-86.

[19] RICHARDSON, S. & GREEN, P.J. (1997) On Bayesian analysis of mixtures with an unknown number of components (with discussion). *J. R. Statist. Soc.* **B 59**, 731-792.

[20] ROBERT, C.P. & CASELLA, G. (1999). *Monte Carlo Statistical Methods.* Springer-Verlag: New-York.

[21] ROBERT, C.P., CELEUX, G. & DIEBOLT, J. (1993). Bayesian estimation of hidden Markov models: a stochastic implementation. *Stat. Prob. Lett.* **16**, 77-83.

[22] ROBERT, C.P. & MENGERSEN, K.L. (1999) Reparametrization issues in mixture estimation and their bearings on the Gibbs sampler. *Comput. Statist. Data Ana.* **29**, 325-343.

[23] ROBERT, C.P. & TITTERINGTON, D.M. (1998). Reparameterisation strategies for hidden Markov models and Bayesian approaches to maximum-likelihood estimation. *Stat. Comp.* **8**, 145-58.

[24] ROEDER, K. (1992) Density estimation with confidence sets exemplified by superclusters and voids in galaxies. *J. Am. Statist. Assoc.* **85**, 617-624.

[25] ROEDER, K. & WASSERMAN, L. (1997) Practical Bayesian density estimation using mixtures of normals. *J. Am. Statist. Assoc.* **92**, 894-902.

[26] SHEPHARD, N. (1994). Partial non-Gaussian times series models. *Biometrika* **81**, 115-31.

[27] VAN LAARHOVEN, P.J. & ARTS, E.H.L. (1987). *Simulated Annealing: Theory and Applications*, Reidel: Amsterdam.

[28] WEI, G.C.G. & TANNER, M.A. (1990). A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithm. *J. Am. Statist. Assoc.* **85**, 699-704.

Tables and Figures

Algorithm	EM	MCEM	SAME
Mean of the log-posterior values	65.47	60.73	66.22
Standard deviation of the log-posterior values	2.31	4.48	0.02

Table 1: Performances of the EM, MCEM and SAME algorithms for finite Gaussian mixtures and the galaxy dataset, obtained over 50 replications

Algorithm	EM	MCEM	SAME
Mean of the log-posterior values	-152.77	-164.20	-151.70
Standard deviation of the log-posterior values	1.83	9.20	0.01

Table 2: Performances of the EM, MCEM and SAME algorithms for Markov-modulated Poisson processes and the foetal lamb dataset, obtained over 50 replications

Algorithm	EM	MCEM	SAME
Mean of the log-posterior values	-201.57	-203.85	-198.93
Standard deviation of the log-posterior values	3.11	5.32	2.17

Table 3: Performances of the EM, MCEM and SAME algorithms for switching autoregressions and the U.S. GNP dataset, obtained over 50 replications

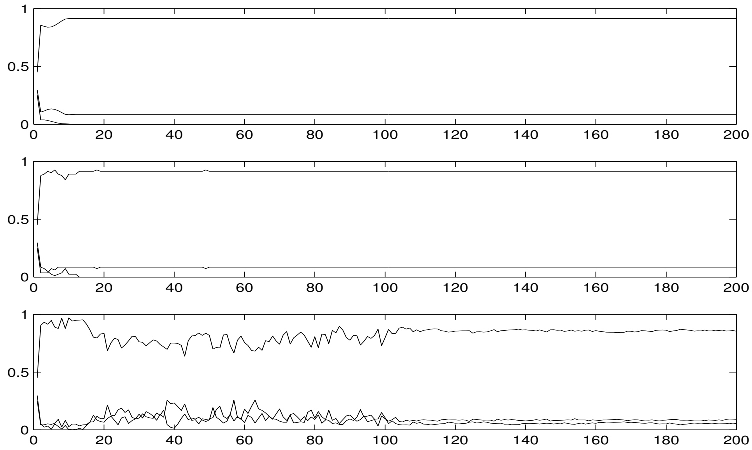

Figure 1: Evolution of the mixing weights π_j against iteration number for finite Gaussian mixtures and the galaxy dataset. *Top:* EM algorithm, *Middle:* SEM algorithm, *Bottom:* SAME algorithm.

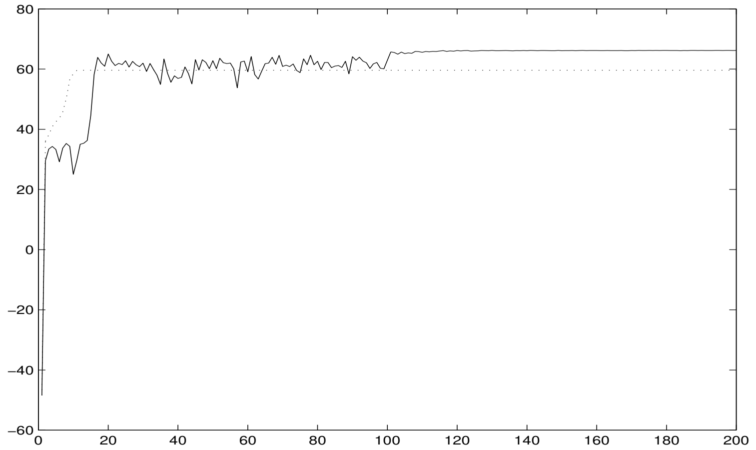

Figure 2: Posterior density values against iteration number for the EM *(dotted line)* and the SAME *(solid line)* for finite Gaussian mixtures and the galaxy dataset.

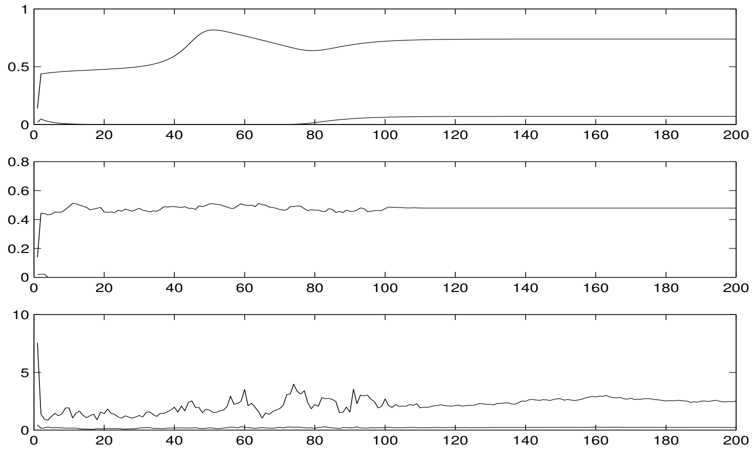

Figure 3: Evolution of the intensities λ_j against iteration number i (for $i \geq 2$) for Markov-modulated Poisson processes and the foetal lamb dataset. *Top:* EM algorithm, *Middle:* SEM algorithm, *Bottom:* SAME algorithm.

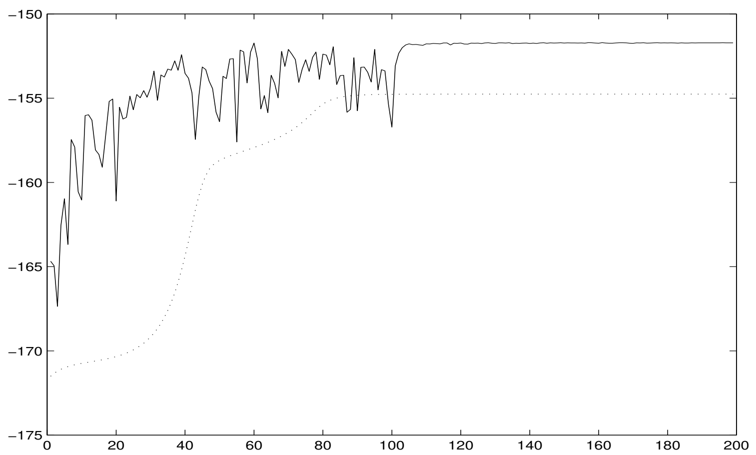

Figure 4: Posterior density values against iteration number ($i \geq 2$) for Markov-modulated Poisson processes and the foetal lamb dataset for the EM *(dotted line)* and the SAME *(solid line)*.