SEQUENTIAL MCMC FOR BAYESIAN MODEL SELECTION

Christophe Andrieu - Nando De Freitas - Arnaud Doucet

Department of Engineering, University of Cambridge
Trumpington Street, CB2 1PZ Cambridge, UK
Email: ca226,jfgf,ad2@eng.cam.ac.uk

ABSTRACT

In this paper, we address the problem of sequential Bayesian model selection. This problem does not usually admit any closed-form analytical solution. We propose here an original sequential simulation-based method to solve the associated Bayesian computational problems. This method combines sequential importance sampling, a resampling procedure and reversible jump MCMC moves. We describe a generic algorithm and then apply it to the problem of sequential Bayesian model order estimation of autoregressive (AR) time series observed in additive noise.

1 Introduction

Model selection is a fundamental data analysis task. It has many applications in various fields of science and engineering. Over the past two decades many of these problems have been addressed using information criteria such as AIC or MDL. The widespread use of these criteria is mainly due to their intrinsic simplicity. AIC and MDL are applied by evaluating two terms: a data term which requires the maximization of the likelihood and a penalty term on the complexity of the model. Within a Bayesian framework, model selection appears to be more difficult as it typically involves the evaluation of Bayes factors, which, in turn, require the computation of high-dimensional integrals with no closed-form analytical expression. These computational problems have limited the use of Bayesian model selection, except for the cases for which asymptotic expansions of the Bayes factors are valid [2], [5].

Markov chain Monte Carlo (MCMC) methods are a set of powerful stochastic algorithms that allow us to solve most of these Bayesian computational problems when the data are available in batches [16], [11], [1]. However, we address here the challenging problem of sequential Bayesian model selection, i.e., we are interested in performing Bayesian model selection recursively in time as data arrive. In this scenario, MCMC methods are not applicable. Monte Carlo (MC) methods that perform sequential Bayesian estimation have been recently developed [3], [6], [12] following the seminal article of Gordon *et al.* [10]. These methods are not well adapted to Bayesian model selection. One obvious solution consists of running a sequential MC method for each candidate model. Unfortunately, it is firstly very computationally intensive and secondly the same computational effort is attributed to each model. In fact, some of these models are of no interest in practice because they have a very weak posterior model probability. In this paper, we develop a sequential MC method to perform Bayesian model selection. Our on-line simulation-based estimation scheme combines sequential importance sampling, a selection scheme and reversible jump MCMC moves. It allows us to run a single algorithm which concentrates itself on the model orders of interest.

The article is organized as follows. Section 2 formally presents the signal model and estimation objectives. In Section 3, we propose a generic simulation-based method to perform sequential Bayesian estimation and discuss the related implementation issues. In Section 4, we apply the proposed algorithm to sequential Bayesian model order selection of AR models observed in Gaussian noise. Finally, in Section 5, we discuss the results.

2 Problem Formulation

In this section we describe our signal model and then state the estimation objectives.

2.1 Signal Model

For any sequence $\mathbf{z}_t$, we denote $\mathbf{z}_{i:j} \triangleq \{\mathbf{z}_i, \mathbf{z}_{i+1} \ldots, \mathbf{z}_j\}$ if $i < j$ and $\mathbf{z}_{i:j} \triangleq \{\mathbf{z}_i, \mathbf{z}_{i-1} \ldots, \mathbf{z}_j\}$ if $i > j$. Let us assume that we have a finite set of candidate models $\mathcal{M}_k$, $k = 1, \ldots, k_{\max}$ for the data $\mathbf{y}_{1:t}$. For each candidate model, there is, at time t, a corresponding set of unknown parameters $\boldsymbol{\alpha}_{k,0:t} \in \boldsymbol{\Omega}_{k,0:t}$. We assume that these parameters are random with a given prior distribution. We define the following joint distribution

$$p\left(k, \boldsymbol{\alpha}_{k,0:t}, \mathbf{y}_{1:t}\right) = p\left(\mathbf{y}_{1:t} \mid k, \boldsymbol{\alpha}_{k,0:t}\right) p\left(\boldsymbol{\alpha}_{k,0:t} \mid k\right) p\left(k\right)$$

$p\left(\mathbf{y}_{1:t} \mid k, \boldsymbol{\alpha}_{k,0:t}\right)$ is the likelihood function, $p\left(\boldsymbol{\alpha}_{k,0:t} \mid k\right)$ is the prior parameter distribution and $p\left(k\right)$ is the prior model

Authorship based on alphabetical order.

probability. For the sake of simplicity, we assume that k is fixed over time and that $\boldsymbol{\alpha}_{k,0:t-1} \subset \boldsymbol{\alpha}_{k,0:t}$.

Example. Let us consider the following noisy AR(k) model

$$\mathcal{M}_k: \begin{array}{l} x_t = \mathbf{a}_k^{\mathrm{T}} \mathbf{x}_{t-1:t-k} + \sigma_{v,k} v_t \\ y_t = x_t + \sigma_{w,k} w_t \end{array} \tag{1}$$

where we denote $\mathbf{a}_k = (a_{1,k}, \ldots, a_{k,k})^{\mathrm{T}}$. We assume that $\mathbf{x}_{-k_{\max}:-1} = \mathbf{0}_{k_{\max}\times 1}$, $v_t \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$ and $w_t \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$ are mutually independent sequences. The model order k, the parameters $\boldsymbol{\theta}_k \triangleq (\mathbf{a}_k, \sigma_{v,k}, \sigma_{w,k}, \delta^2)$ and the states of the system $\mathbf{x}_{0:t}$ are unknown. δ is a hyperparameter that controls the variance of $\mathbf{a}_k$. Denoting $\boldsymbol{\alpha}_{k,0:t} \triangleq (\mathbf{x}_{0:t}, \boldsymbol{\theta}_k)$, we have the following joint distribution

$$p(k, \boldsymbol{\alpha}_{k,0:t}, \mathbf{y}_{1:t}) = p(\mathbf{y}_{1:t}|\, k, \mathbf{x}_{0:t})\, p(\mathbf{x}_{0:t}|\, k, \boldsymbol{\theta}_k)\, p(k, \boldsymbol{\theta}_k)$$

We adopt the following hierarchical prior distribution

$$p(k, \boldsymbol{\theta}_k) = p(\mathbf{a}_k|\, \sigma^2_{v,k}, \delta^2)\, p(\sigma^2_{v,k})\, p(\sigma^2_{w,k})\, p(\delta^2)\, p(k) \tag{2}$$

with $k \sim \mathcal{U}_{\{1,\ldots,k_{\max}\}}$, $\mathbf{a}_k|\, \sigma^2_{v,k}, \delta^2 \sim \mathcal{N}\left(\mathbf{0}_{k\times 1}, \sigma^2_{v,k}\delta^2 \mathbf{I}_k\right)$, $\sigma^2_{v,k} \sim \mathcal{IG}\left(\frac{\nu_0}{2}, \frac{\gamma_0}{2}\right)$, $\sigma^2_{w,k} \sim \mathcal{IG}\left(\frac{\nu_0}{2}, \frac{\gamma_0}{2}\right)$ and $\delta^2 \sim \mathcal{IG}\left(\frac{n_0}{2}, \frac{g_0}{2}\right)$. In this case $\boldsymbol{\Omega}_{k,0:t} = \mathbb{R}^{t+1} \times \mathbb{R}^k \times (\mathbb{R}^+)^3$.

2.2 Detection/Estimation objectives

Our aim is to **estimate recursively in time** the joint posterior distribution $p(\boldsymbol{\alpha}_{k,0:t}, k|\, \mathbf{y}_{1:t})$ defined on the space $\boldsymbol{\Omega}_{0:t} = \cup_{k=1}^{k_{\max}} \boldsymbol{\Omega}_{k,0:t} \times \{k\}$. This distribution is equal to

$$\frac{p(\mathbf{y}_t|\mathbf{y}_{1:t-1}, \boldsymbol{\alpha}_{k,0:t}, k)\, p(\boldsymbol{\alpha}_{k,0:t}, k|\mathbf{y}_{1:t-1})}{\sum_{k=1}^{k_{\max}} \int_{\boldsymbol{\Omega}_{k,0:t}} p(\mathbf{y}_t|\mathbf{y}_{1:t-1}, \boldsymbol{\alpha}_{k,0:t}, k)\, p(\boldsymbol{\alpha}_{k,0:t}, k|\mathbf{y}_{1:t-1})\, d\boldsymbol{\alpha}_{k,0:t}}$$

There exist very few cases where it is possible to obtain a closed-form analytical expression for this joint distribution and its features of interest, such as the posterior model probabilities $p(k|\, \mathbf{y}_{1:t})$. It is thus necessary to use computational methods.

3 Sequential Bayesian Computation

Most of the current computational methods to perform sequential Bayesian estimation are based on a combination of sequential Bayesian importance sampling and resampling procedures; see [6], [12] for some recent reviews on the subject. These methods have been developed in a framework where the model is assumed known. We present here the Bayesian importance sampling method in the case of unknown models and introduce reversible jump MCMC steps in the algorithm. In the first subsection, we describe a generic algorithm.Subsequently, we discuss the implementation issues in more details.

3.1 A generic algorithm

Even if k is a fixed parameter over time, we use $k_t^{(i)}$ to denote the value of k simulated at time t and associated to the i^{th} particle. Let $\boldsymbol{\alpha}_{k,-t}$ denote the set of parameters in $\boldsymbol{\alpha}_{k,0:t}$ but not in $\boldsymbol{\alpha}_{k,0:t-1}$ (note this set can be empty for some problems). We introduce an importance distribution for this set of parameters denoted $q(\boldsymbol{\alpha}_{k,-t}|\, \mathbf{y}_{1:t}, k, \boldsymbol{\alpha}_{k,0:t-1})$. It is such that the support of $p(\boldsymbol{\alpha}_{k,0:t}, k|\, \mathbf{y}_{1:t})$ is included in the one of $p(\boldsymbol{\alpha}_{k,0:t-1}, k|\, \mathbf{y}_{1:t-1})\, q(\boldsymbol{\alpha}_{k,-t}|\, \mathbf{y}_{1:t}, k, \boldsymbol{\alpha}_{k,0:t-1})$. Given, at time $t-1$, N particles $\left(\boldsymbol{\alpha}^{(i)}_{k,0:t-1}, k^{(i)}_{t-1}\right)$ distributed approximately according to $p(\boldsymbol{\alpha}_{k,0:t-1}, k|\, \mathbf{y}_{1:t-1})$, the algorithm proceeds as follows as time t to obtain N particles $\left(\boldsymbol{\alpha}^{(i)}_{k,0:t}, k^{(i)}_t\right)$ distributed approximately according to $p(\boldsymbol{\alpha}_{k,0:t}, k|\, \mathbf{y}_{1:t})$.

Monte Carlo Filter for Bayesian Model Selection

Sequential Importance sampling step

• For $i = 1, ..., N$, sample

$$\widetilde{\boldsymbol{\alpha}}^{(i)}_{k,-t} \sim q\left(\boldsymbol{\alpha}_{k,-t}|\, \mathbf{y}_{1:t}, k^{(i)}, \boldsymbol{\alpha}^{(i)}_{k,0:t-1}\right)$$

and $\left(\widetilde{\boldsymbol{\alpha}}^{(i)}_{k,0:t}, \widetilde{k}^{(i)}\right) \triangleq \left(\boldsymbol{\alpha}^{(i)}_{k,0:t-1}, \widetilde{\boldsymbol{\alpha}}^{(i)}_{k,-t}, k^{(i)}\right)$.

• For $i = 1, ..., N$, evaluate the importance weights up to a normalizing constant:

$$\overline{w}^{(i)}_t = \frac{p\left(\widetilde{\boldsymbol{\alpha}}^{(i)}_{k,0:t}, \widetilde{k}^{(i)} \middle| \mathbf{y}_{1:t}\right)}{p\left(\widetilde{\boldsymbol{\alpha}}^{(i)}_{k,0:t-1}, \widetilde{k}^{(i)} \middle| \mathbf{y}_{1:t-1}\right) q\left(\widetilde{\boldsymbol{\alpha}}^{(i)}_{k,-t} \middle| \mathbf{y}_{1:t}, \widetilde{k}^{(i)}, \boldsymbol{\alpha}^{(i)}_{k,0:t-1}\right)} \tag{3}$$

• For $i = 1, ..., N$, normalize the importance weights:

$$\widetilde{w}^{(i)}_t = \left[\sum_{j=1}^{N} \overline{w}^{(j)}_t\right]^{-1} \overline{w}^{(i)}_t \tag{4}$$

Selection step

• Multiply/Suppress samples $\left(\widetilde{\boldsymbol{\alpha}}^{(i)}_{k,0:t}, \widetilde{k}^{(i)}\right)$ respectively with high/low importance weights $\widetilde{w}^{(i)}_t$ to obtain N random samples $\left(\boldsymbol{\alpha}'^{(i)}_{k,0:t}, k'^{(i)}\right)$ approximately distributed according to $p\left(\boldsymbol{\alpha}'^{(i)}_{k,0:t}, k'^{(i)} \middle| \mathbf{y}_{1:t}\right)$.

MCMC step

• Apply a Markov transition kernel of invariant distribution $\prod_{i=1}^{N} p\left(\boldsymbol{\alpha}^{(i)}_{k,0:t}, k^{(i)} \middle| \mathbf{y}_{1:t}\right)$ to obtain $\left(\boldsymbol{\alpha}^{(i)}_{k,0:t}, k^{(i)}\right)$.

3.2 Implementation Issues

3.2.1 Bayesian importance sampling step

Bayesian importance sampling has been mainly formulated in scenarios where the distribution from which one wants to sample is defined on a simple set with a unique dominating measure [8]. In our case, this distribution is defined on an union of distinct subspaces, i.e.,

$$\pi\left(\boldsymbol{\alpha}_{k,0:t},k\right)=\sum_{l=1}^{k_{\max}}\pi_l\left(\boldsymbol{\alpha}_{l,0:t},l\right)\mathbb{I}_{\boldsymbol{\Omega}_{l,0:t}\times\{l\}}\left(\boldsymbol{\alpha}_{k,0:t},k\right)$$

Let us introduce the following importance distribution

$$q\left(\boldsymbol{\alpha}_{k,0:t},k\right)=\sum_{l=1}^{k_{\max}}q_l\left(\boldsymbol{\alpha}_{l,0:t},l\right)\mathbb{I}_{\boldsymbol{\Omega}_{l,0:t}\times\{l\}}\left(\boldsymbol{\alpha}_{k,0:t},k\right)$$

where the support of $q_l\left(\boldsymbol{\alpha}_{l,0:t},l\right)$ includes the support of $\pi_l\left(\boldsymbol{\alpha}_{l,0:t},l\right)$ for $l=1,\ldots,k_{\max}$. Then $\pi\left(\boldsymbol{\alpha}_{k,0:t},k\right)$ is equal to

$$\frac{\sum_{l=1}^{k_{\max}}w_l\left(\boldsymbol{\alpha}_{l,0:t},l\right)q_l\left(\boldsymbol{\alpha}_{l,0:t},l\right)\mathbb{I}_{\boldsymbol{\Omega}_{l,0:t}\times\{l\}}\left(\boldsymbol{\alpha}_{k,0:t},k\right)}{\sum_{l=1}^{k_{\max}}\int_{\boldsymbol{\Omega}_{l,0:t}}w_l\left(\boldsymbol{\alpha}_{l,0:t},l\right)q_l\left(\boldsymbol{\alpha}_{l,0:t},l\right)d\boldsymbol{\alpha}_{l,0:t}}$$

where

$$w_l\left(\boldsymbol{\alpha}_{l,0:t},l\right)=\frac{\pi_l\left(\boldsymbol{\alpha}_{l,0:t},l\right)}{q_l\left(\boldsymbol{\alpha}_{l,0:t},l\right)}$$

Let us assume that we have N samples $\left(\boldsymbol{\alpha}_{k,0:t}^{(i)},k^{(i)}\right)$, $i=1,\ldots,N$, distributed according to $q\left(\boldsymbol{\alpha}_{k,0:t},k\right)$, then the Bayesian importance sampling method gives us the following point mass approximation $\widehat{\pi}\left(\boldsymbol{\alpha}_{k,0:t},k\right)$ of $\pi\left(\boldsymbol{\alpha}_{k,0:t},k\right)$

$$\frac{\sum_{l=1}^{k_{\max}}\sum_{I_l^*}w_l\left(\boldsymbol{\alpha}_{l,0:t}^{(i)},l\right)\delta_{\left(\boldsymbol{\alpha}_{k,0:t}^{(i)},k^{(i)}\right)}\left(d\boldsymbol{\alpha}_{k,0:t},k\right)}{\sum_{l=1}^{k_{\max}}\sum_{I_l^*}w_l\left(\boldsymbol{\alpha}_{l,0:t}^{(i)},l\right)}$$

with $I_l^*=\left\{i\in\{1,\ldots,N\}\setminus\left(\boldsymbol{\alpha}_{k,0:t}^{(i)},k^{(i)}\right)\in\boldsymbol{\Omega}_{l,0:t}\times\{l\}\right\}$. Equations (3) and (4) follow.

3.2.2 Selection step

A selection scheme associates to each particle $\left(\widetilde{\boldsymbol{\alpha}}_{k,0:t}^{(i)},\widetilde{k}^{(i)}\right)$ a number of "children", say $N_i\in\mathbb{N}$, such that $\sum_{i=1}^N N_i=N$. We describe some interesting selection schemes proposed in the literature. These schemes satisfy $\mathbb{E}[N_i]=N\widetilde{w}_t^{(i)}$ but their performance varies in terms of $var\left[N_i\right]$.

• *Sampling Importance Resampling (SIR)/Multinomial Sampling procedure*: this procedure, originally introduced by Gordon *et al.* [10], is the most popular one. One samples N times from $\sum_{i=1}^N\widetilde{w}_t^{(i)}\delta_{\left(\widetilde{\boldsymbol{\alpha}}_{k,0:t}^{(i)},\widetilde{k}^{(i)}\right)}\left(d\boldsymbol{\alpha}_{k,0:t}',k'\right)$ to obtain $\left(\left(\boldsymbol{\alpha}_{k,0:t}'^{(i)},k'^{(i)}\right);i=1,\ldots,N\right)$. This is equivalent to drawing jointly $(N_i;i=1,\ldots,N)$ according to a multinomial distribution of parameters N and $\widetilde{w}_t^{(i)}$. This procedure can be implemented in $O\left(N\right)$ iterations [6], [14] according to [15, pp. 96]. However, as pointed out in [4] and [12], it is better to use selection schemes with a limited variance.

• *Residual Resampling*: this procedure performs as follows, see [12] for further details. Set $\widetilde{N}_i=\left\lfloor N\widetilde{w}_t^{(i)}\right\rfloor$ then perform a SIR procedure to select the remaining $\overline{N}_t=N-\sum_{i=1}^N\widetilde{N}_i$ samples with the new weights $w_t'^{(i)}=\overline{N}_t^{-1}\left(\widetilde{w}_t^{(i)}N-\widetilde{N}_i\right)$, finally add the results to the current $\widetilde{N}_i$. For this scheme $var\left[N_i\right]$ is smaller than the one given by the SIR scheme. Moreover, this procedure is computationally cheaper.

• *Systematic Sampling* [4]*:* this method introduces a variance on N_i even smaller than the residual resampling scheme. However, its computational complexity is $O\left(N\right)$.

Actually, the restriction $\mathbb{E}[N_i]=N\widetilde{w}_t^{(i)}$ is unecessary to obtain convergence results. So it is possible to use biased but very quick selection schemes.

3.2.3 MCMC steps

After the selection scheme at time t, we obtain N particles distributed marginally approximately according to $p\left(\boldsymbol{\alpha}_{k,0:t},k\middle|\mathbf{y}_{1:t}\right)$. In practice, a skewed importance weights distribution implies that many particles have had no children, i.e. $N_i=0$, whereas others have had a large number of children, the extreme case being $N_i=N$. In these cases, there is a severe depletion of samples. A strategy for improving the results consists of introducing MCMC steps of invariant distribution $p\left(\boldsymbol{\alpha}_{k,0:t},k\middle|\mathbf{y}_{1:t}\right)$ on each particle [4], [7], [9], [13]. The basic idea is that, by applying a Markov transition kernel, the total variation of the current distribution with respect to the invariant distribution can only decrease. Note, however, that we do not require this kernel to be ergodic. Actually it is possible to generalize this idea and to introduce MCMC steps on the product space with an invariant distribution $\prod_{i=1}^N p\left(\boldsymbol{\alpha}_{k,0:t},k\middle|\mathbf{y}_{1:t}\right)$, that is apply MCMC steps on the *whole population* of particles. In doing so, it allows us to introduce the algorithms developed in parallel MCMC computation. In this paper, we limit ourselves to the use of simple MCMC transitions steps on each particle. In particular, we propose to use reversible jump MCMC steps [11], [1] so as to allow the particles to move from one subspace to the others.

3.2.4 Remark

At first sight, the generic algorithm described above can appear intractable as it seems that it requires to keep in memory at time t the simulated trajectories from time 0 to time t. This is not necessary in many cases of interest, where only a set of sufficient statistics of highly reduced and fixed dimension has to be kept in memory.

4 Application to sequential AR model order selection

Let us consider the model specified by Eq. (1) to (2). We want to estimate recursively in time the joint distribution

$p\left(\mathbf{x}_{0:t}, \boldsymbol{\theta}_k, k \middle| \mathbf{y}_{0:t}\right)$. This is a complex optimal estimation problem.

4.1 Implementation issues

To improve readibility, we add a subscript t for fixed-parameters simulated at time t.

4.1.1 Sampling step/Selection step

In this case $\boldsymbol{\alpha}_{k,-t} = x_t$. We select as importance distribution $\widetilde{x}_t^{(i)} \sim p\left(x_t \middle| y_t, \widetilde{\mathbf{x}}_{0:t-1}^{(i)}, \widetilde{k}_t^{(i)}, \widetilde{\boldsymbol{\theta}}_{k,t}^{(i)}\right)$, i.e.

$$\widetilde{x}_t \left| \left(y_t, \widetilde{\mathbf{x}}_{0:t-1}^{(i)}, \widetilde{k}_t^{(i)}, \widetilde{\boldsymbol{\theta}}_{k,t}^{(i)}\right) \right. \sim \mathcal{N}\left(\widetilde{m}_t^{(i)}, \widetilde{\sigma}_t^{2(i)}\right)$$

where $\widetilde{\sigma}_t^{-2(i)} = \widetilde{\sigma}_{w,t}^{-2(i)} + \widetilde{\sigma}_{v,t}^{-2(i)}$ and

$$\widetilde{m}_t^{(i)} = \widetilde{\sigma}_t^{2(i)} \left(\frac{\widetilde{\mathbf{a}}_{k,t}^{\mathrm{T}(i)} \widetilde{\mathbf{x}}_{t-1:t-k}^{(i)}}{\widetilde{\sigma}_{v,t}^{2(i)}} + \frac{y_t}{\widetilde{\sigma}_{w,t}^{2(i)}} \right)$$

We deduce easily that the importance weight satisfies

$$\widetilde{w}_t^{(i)} \propto \frac{\exp\left(-0.5\left(\widetilde{\sigma}_{v,t}^{2(i)} + \widetilde{\sigma}_{w,t}^{2(i)}\right)^{-1}\left(y_t - \widetilde{\mathbf{a}}_{k,t}^{(i)\mathrm{T}} \widetilde{\mathbf{x}}_{t-1:t-k}^{(i)}\right)^2\right)}{\left(\widetilde{\sigma}_{v,t}^{2(i)} + \widetilde{\sigma}_{w,t}^{2(i)}\right)^{1/2}}$$

The selection step is done according to the residual resampling scheme[1].

4.1.2 MCMC step

We update each particle independently using reversible jump MCMC moves. The transition kernel is a mixture of different transition kernels. At each iteration, one of the candidate moves, birth, death or update is randomly chosen. The probabilities for choosing these moves are b_k, d_k and u_k respectively, such that $b_k + d_k + u_k = 1$ for all $1 \leq k \leq k_{\max}$. A move is performed if the algorithm accepts it. For $k = 1$ the death move is impossible, so that $d_0 \triangleq 0$. For $k = k_{\max}$ the birth move is impossible and thus $b_{k_{\max}} \triangleq 0$. Otherwise, $b_k = d_k = u_k$.

Update move. In this move, the dimension is kept fixed, i.e. $k^{(i)} = k'^{(i)}$. To update the parameters, we perform a partial Gibbs sampling algorithm where the last state x_t and the fixed hyperparameters are updated according to their full conditional distributions, i.e., we sample successively from

$$x_t^{(i)} \sim \mathcal{N}\left(m_t^{(i)}, \sigma_t^{2(i)}\right) \tag{5}$$

where $\sigma_t^{-2(i)} = \sigma_{w,t}^{-2(i)} + \sigma_{v,t}^{-2(i)}$ and

$$m_t^{(i)} = \sigma_t^{2(i)} \left(\frac{\mathbf{a}_{k,t}^{\prime(i)\mathrm{T}} \mathbf{x}_{t-1:t-k}^{(i)}}{\sigma_{v,t}^{\prime 2(i)}} + \frac{y_t}{\sigma_{w,t}^{\prime 2(i)}} \right) \tag{6}$$

and set $\mathbf{x}_{0:t}^{(i)} = \left(\mathbf{x}_{0:t-1}^{\prime(i)}, x_t^{(i)}\right)$. Then

$$\begin{array}{l} \sigma_{w,t}^{2(i)} \sim \mathcal{IG}\left(\frac{\nu_0+t}{2}, \frac{\gamma_0 + \sum_{l=1}^t \left(y_l - x_l^{(i)}\right)^2}{2}\right) \\ \sigma_{v,t}^{2(i)} \sim \mathcal{IG}\left(\frac{\nu_0+t}{2}, \frac{\gamma_0 + \mathbf{x}_{1:t}^{(i)\mathrm{T}} \mathbf{P}_{\mathbf{a},t}^{(i)} \mathbf{x}_{1:t}^{(i)}}{2}\right) \\ \mathbf{a}_{k,t}^{(i)} \sim \mathcal{N}\left(\mathbf{m}_{\mathbf{a},t}^{(i)}, \sigma_{v,t}^{2(i)} \boldsymbol{\Sigma}_{\mathbf{a},t}^{(i)}\right) \end{array} \tag{7}$$

where

$$\begin{array}{l} \boldsymbol{\Sigma}_{\mathbf{a},t}^{-1(i)} = \sum_{l=1}^t \mathbf{x}_{l-1:l-k}^{(i)} \mathbf{x}_{l-1:l-k}^{\mathrm{T}(i)} + \delta_t^{\prime -2(i)} \mathbf{I}_k \\ \mathbf{m}_{\mathbf{a},t}^{(i)} = \boldsymbol{\Sigma}_{\mathbf{a},t}^{(i)} \left(\sum_{l=1}^t \mathbf{x}_{l-1:l-k}^{(i)} x_l^{(i)}\right) \\ \mathbf{x}_{1:t}^{\mathrm{T}(i)} \mathbf{P}_{\mathbf{a},t}^{(i)} \mathbf{x}_{1:t}^{(i)} = \mathbf{x}_{1:t}^{\mathrm{T}(i)} \mathbf{x}_{1:t}^{(i)} - \mathbf{m}_{\mathbf{a},t}^{\mathrm{T}(i)} \boldsymbol{\Sigma}_{\mathbf{a},t}^{-1(i)} \mathbf{m}_{\mathbf{a},t}^{(i)} \end{array} \tag{8}$$

Finally

$$\delta_t^{2(i)} \sim \mathcal{IG}\left(\frac{n_0 + k^{(i)}}{2}, \frac{g_0 + \sigma_{v,t}^{-2(i)} \mathbf{a}_{k,t}^{\mathrm{T}(i)} \mathbf{a}_{k,t}^{(i)}}{2}\right) \tag{9}$$

Birth/Death moves. These moves perform dimension changes respectively from $k'^{(i)}$ to $k^{(i)} = k'^{(i)} + 1$ and to $k^{(i)} = k'^{(i)} - 1$. When we propose to move from dimension $k'^{(i)}$ to dimension k^*, each move is accepted with the following acceptance probability

$$\alpha_{k'^{(i)} \to k^*}^{(i)} = \min\left\{1, \frac{p\left(k^*, \delta'^{2(i)} \middle| \mathbf{x}_{0:t}^{\prime(i)}\right)}{p\left(k'^{(i)}, \delta'^{2(i)} \middle| \mathbf{x}_{0:t}^{\prime(i)}\right)}\right\}$$

where

$$p\left(k, \delta^2 \middle| \mathbf{x}_{0:t}\right) \propto \exp\left(-\frac{g_0}{2\delta^2}\right) \frac{\left[\gamma_0 + \mathbf{x}_{1:t}\mathbf{P}_{\mathbf{a},t}\mathbf{x}_{1:t}\right]^{-(t+\nu_0)/2}}{\left|\boldsymbol{\Sigma}_{\mathbf{a},t}\right|^{-1/2} \left(\delta^2\right)^{(n_0+k)/2-1}}$$

If the move is accepted then $k^{(i)} = k^*$ and the parameters are updated according to Eq. (5) to (9). Otherwise, there is no modification.

It is not necessary to keep in memory $\mathbf{x}_{0:t}^{(i)}$ in this example. One only needs to store the following set of sufficient statistics: $\sum_{l=1}^t \mathbf{x}_{l-1:l-k_{\max}}^{(i)} \mathbf{x}_{l-1:l-k_{\max}}^{\mathrm{T}(i)}$, $\sum_{l=1}^t x_l^{2(i)}$, $\sum_{l=1}^t \mathbf{x}_{l-1:l-k_{\max}}^{(i)} x_l^{(i)}$ and $\sum_{l=1}^t \left(y_l - x_l^{(i)}\right)^2$.

4.2 Simulation

We simulated $T = 150$ observations of an $AR(2)$ model with a pair of conjuguate poles $0.9e^{\pm j0.3}$ (i.e. $a_1 = -1.72$ and $a_2 = 0.81$), $\sigma_v = 0.2$ and $\sigma_w = 0.1$. For the parameters of the signal, we set $k_{\max} = 5$, $\nu_0 = 2$, $\gamma_0 = 0.01$, $n_0 = 0.1$ and $g_0 = 0.1$. We ran the algorithm with $N = 50000$ particles. In Fig. 1, we present the estimated probability distribution $p\left(k \middle| \mathbf{y}_{1:t}\right)$. It clearly appears that $k = 2$ is the selected model order. We present respectively in Fig. 2 and in Fig. 3 the MMSE estimates $\mathbb{E}\left[a_1, a_2 \middle| \mathbf{y}_{1:t}, k = 2\right]$ and the MMSE estimates $\mathbb{E}\left[\sigma_v \middle| \mathbf{y}_{1:t}, k = 2\right]$, $\mathbb{E}\left[\sigma_w \middle| \mathbf{y}_{1:t}, k = 2\right]$.

[1] In this case, the importance weight being independent of x_t, it is possible (and better) to reverse the order of the sampling and selection steps. However, this is impossible in the general case.

5 Discussion

In this paper, we have proposed a simulation-based approach to perform sequential Bayesian model selection. Our method combines sequential Monte Carlo methods and reversible jump MCMC methods. It can be interpreted as an extension of [4], [9] and [13] to model selection. Although it is computationally intensive, it can be straightforwardly implemented on parallel computers. In a **simple case**, we obtained satisfactory results. However, we think that several major problems still have to be addressed to make such a methodology efficient for more complex cases. Even in the fixed-dimension case, we observed a drift of the estimated fixed-parameters for long simulations, i.e. $T \gg 1$. This phenomenon has not been reported previously in the literature where shorter time series were processed. In our opinion, this problem is related to the fact that there is an accumulation over time of the Monte Carlo errors. It is likely that one cannot establish a uniform (in time) convergence result for such an algorithm; the system does not forget its past contrary to [7] for example. A practical, but not entirely satisfactory, way of solving this problem consists of introducing an artificial time-varying model. Other approaches are under study.

6 REFERENCES

[1] C. Andrieu, P.M. Djurić and A. Doucet, "Model selection by MCMC computation", to appear *Signal Processing*, 1999.

[2] J.M. Bernardo and A.F.M. Smith, *Bayesian Theory*, Wiley, 1994.

[3] C. Berzuini, N. Best, W.R. Gilks and C. Larizza, "Dynamic conditional independence models and Markov chain Monte Carlo methods", *J. Amer. Stat. Assoc.*, vol. 92, 1997, 1403-1412.

[4] J. Carpenter, P. Clifford and P. Fearnhead, "Building robust simulation-based filters for evolving data sets", technical report, University of Oxford, Dept. of Statistics, 1999.

[5] P.M. Djurić, "Asymptotic MAP criteria for model selection", *IEEE Trans. Signal Processing*, vol. 46, pp. 2726-2735, 1998.

[6] A. Doucet, "On sequential simulation-based methods for Bayesian filtering", technical report, CUED/F-INFENG/TR 310, 1998.

[7] A. Doucet and N.J. Gordon, "Simulation-based optimal filter for maneuvring target tracking", in *Proc. SPIE Signal and Data Processing of Small Targets*, July 1999.

[8] J. Geweke, "Bayesian inference in econometrics models using Monte Carlo integration", *Econometrica*, vol. 57, 1989, pp. 1317-1339.

[9] W.R. Gilks and C. Berzuini, "Following a moving target - Monte Carlo inference for dynamic Bayesian models", technical report, Cambridge University, MRC, 1998.

[10] N.J. Gordon, D.J. Salmond and A.F.M. Smith, "Novel approach to nonlinear/non-Gaussian Bayesian state estimation", *IEE-Proceedings-F*, vol. 140, no. 2, 1993, pp. 107-113.

[11] P.J. Green, "Reversible jump MCMC computation and Bayesian model determination", *Biometrika*, vol. 82, pp. 711-732, 1995.

[12] J.S. Liu and R. Chen, "Sequential Monte Carlo methods for dynamic systems", *J. Amer. Stat. Assoc.*, vol. 93, pp. 1-22-1031, 1998.

[13] S.N. MacEachern, M. Clyde and J.S. Liu, "Sequential importance sampling for nonparametric Bayes models: the next generation", forthcoming *Can. J. Stat.*, 1999.

[14] M.K. Pitt and N. Shephard, "Filtering via simulation:auxiliary particle filters", forthcoming *J. Amer. Stat. Assoc.,* 1999.

[15] B.D. Ripley, *Stochastic Simulation*, Wiley, New York, 1987.

[16] L. Tierney, "Markov chains for exploring posterior distributions", *Ann. Stat.*, pp. 1701-1728, 1994.

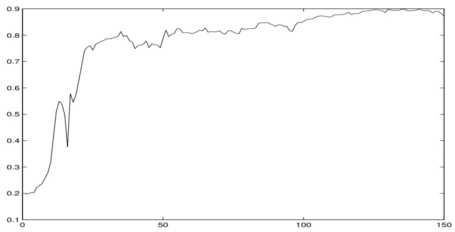

Figure 1: Estimated probability $p\left(k = 2 \middle| \mathbf{y_{1:t}}\right)$

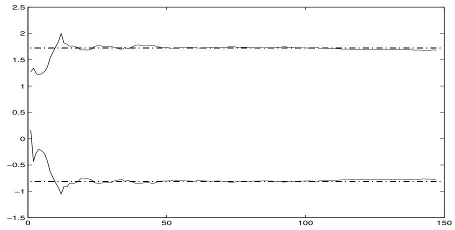

Figure 2: MMSE Estimates of the AR coefficients; the dashed lines correspond to the true values.

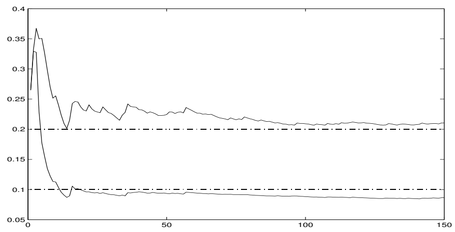

Figure 3: MMSE Estimates of the noises standard deviations; the dashed lines correspond to the true values.