Afrika Statistika

Afrika Statistika
Vol. 14 (2), 2019, pages 1977 - 1998.
DOI: http://dx.doi.org/10.16929/as/2019.1977.145

ISSN 2316-090X

On some Extensions of the Sequential Monte Carlo methods in high-order Hidden Markov Models

Mouhamad M. Allaya[1*,3], **Alioune Coulibaly** [2], **El Hadji Dème** [3], **Mouhamadou M. Kâ** [4], **and Babacar Sène** [5]

[1]African Institute of Mathematical Sciences, Km2 Joal road, Mbour, BP 1418, Senegal
[2]University Assane Seck , Dept. of Mathematics, Ziguinchor, BP 523, Senegal
[3] LERSTAD, University Gaston Berger Dept. of Mathematics, St-Louis, BP 234, Senegal
[4,5]University Cheikh Anta Diop, Dept. of Economics and Management, Dakar, BP 5005, Senegal

Received on February 25, 2019. Accepted July 27, 2019.

Abstract. We analyze some extensions of the Sequential Monte Carlo (SMC) methods in the context of nonlinear state space models. Namely, we tailor the SMC methods to handle high-order HMM through the customary recursions of posterior distributions. It proceeds on mimicking the two-step procedure that is, the prediction step and the update step, in the derivation of the filter distribution. Once stated, we extend some smoothing recursions as the Forward-Backward algorithm and the Backward smoother to deal with the actual smoothing distributions in high-order HMM. Finally, we give few examples as an application of these extensions.

Key words: Sequential Monte Carlo, high-order HMM, Smoothing, Filtering
AMS 2010 Mathematics Subject Classification : 60J05, 62M05, 65C05

* Corresponding author : Mouhamad M. Allaya (mouhamad.allaya@aims-senegal.org)
Alioune Coulibaly : a.coulibaly5649@zig.univ.sn
El Hadji Dème : elhadjidemeufrsat@gmail.com
Mamadou M. kâ : moustapha.ka@ucad.edu.sn
Babacar Sène : babacar.sene@ucad.edu.sn

Résumé. (Abstract in French) Nous analysons quelques extensions des méthodes de Monte Carlo séquentielles (SMC) dans le contexte des modèles à espace d'états non-linéaires. Précisément,, nous adaptons les méthodes SMC pour traiter les HMM d'ordre supérieur à travers les récursions habituelles des distributions à posteriori. Cela procède par mimer la procédure en deux étapes, c'est-à-dire l'étape de prédiction et l'étape de mise à jour, dans la dérivation de la distribution du filtre. Une fois obtenu, nous étendons certaines récursions de lissage comme l'algorithme Forward-Backward et l'algorithme Backward Smoother pour traiter les distributions de lissage dans les HMM d'ordre supérieur. Enfin, nous donnons quelques exemples de l'application de ces extensions.

1. Introduction

The literature of SMC methods is recent and can be dated from the paper by Gordon *et al.* 1993. Although, several attempts had preceded including the work by Handschin and Mayne 1969, Handschin 1970 among other. The main obstacle to the SMC's growth was the limitation of computing power. Since then, several efforts have been made both in theory and in practice to lay down the foundations of SMC methods. One may consults review articles such as Doucet *et al.* 2000, Cappé *et al.* 2007 or books by Doucet 2001, Del Moral 2004 or Cappé, Moulines and Ryden 2005 which include several theoretical results and a range of rich and varied applications in many areas.

So far, the SMC methods apply to hidden Markov models of order 1, commonly called one-order HMM. Specifically, an $\mathcal{X} \times \mathcal{Y}-$valued bivariate process $\{(X_k, Y_k)\}$, where $\{X_k\}$ is an unobservable dynamic Markov model of order 1. $\{Y_k\}$ represents the observation process used indirectly to quantify the realizations of the process $\{X_k\}$ and satisfying the channel without memory property's. However, it may happen that the signal process $\{X_k\}$ depends on more than one of its lags that is, the memory process of the signal is more persistent. Thus, a direct application of SMC methods still a little tricky.

To overcome this difficulty, one may at first think that a trivial rewriting of the process $\{X_k\}$ according to its lags is enough and may help to fall in the usual case of Markov chain of order 1. However, this formulation is not without causing additional difficulty. In fact, one may face among other the degeneracy problem of the state noise resulting from this state transformation. A new approach is needed. In this perspective, we derive a new approach that helps handling higher-order HMM without any modification of the former kind . To achieve it, we just mimic the different stages in the establishment of the filtering and smoothing equations in non-linear and non-Gaussian state space models. Singularly, we mimic the prediction and the correction steps of the filter distribution in one-order HMM and adapt it to HMM of order strictly greater than one. Once done, we derive analogous recursions to those of the Forward-Backward smoother and the particle smoother by Godsill *et al.* 2004. In the sequel, we show a use of the SMC methods extension in an example a stochastic volatility with an ℓ memory depth. As a final point, we

show the usability of the SMC some parameter inference problems in linear and Gaussian state-space model and in stochastic volatility model.

2. SMC methods in one-order HMM

Particle filter and smoother belong to SMC methods that aim at generating samples realizations from actual and historical state sequences given the whole data set or a part of it. The main idea being that any given measure on a measurable space can be approximated by a sum of empirical measures. Particle filter aims at computing recursively in time, the conditional distribution of the current state given the whole data up to current time k, that is the filtering distribution. Smoothing is more branched, however, most cases can be plugged into the joint smoothing distribution. When classical approaches fail because of lack of analytical solutions or for a non-linear or non Gaussian purpose, the SMC methods can help in a certain way to get rid off most of these limitations. As long as some minimal requirements are met, the SMC methods are set of powerful tools that approximate any function of the state sequences even for a class of unbounded functions (See Hu and Schön 2011 for detail of such unboundedness), given the data up to a given time.

To state the general idea of particle filter and smoother, consider the following one-order HMM:

$$\begin{cases} X_k &= a_k(X_{k-1}, V_k) \\ Y_k &= b_k(X_k, W_k) \end{cases}$$

where $a_k(\cdot)$ and $b_k(\cdot)$ are possibly non-linear functions, $\{X_k\}$ is a $1-$order Markov chain with initial state X_0 distributed according to a diffuse prior distribution $\nu(\cdot)$ and transition kernel M from $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$ to $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$. We assume that M admits a density function m w.r.t a dominating measure λ. $(V_k)_{k\geq 1}$ and $(W_k)_{k\geq 1}$ are i.i.d disturbance noises independent of X_0, respectively the state noise and the measurement noise. We also assume that the observation process $\{Y_k\}$, constructed on the measurable space $(\mathcal{Y}, \mathcal{B}(\mathcal{Y}))$ is conditionally independent given $\{X_k\}$ with a marginal distribution admitting a density function g such that

$$\forall A \in \mathcal{B}(\mathcal{Y}), \quad \mathbb{P}(Y_k \in A | X_k) = \int_A g(X_k, y)\mu(dy),$$

where μ is a $\sigma-$finite measure on $(\mathcal{Y}, \mathcal{B}(\mathcal{Y}))$. To sum up, the model is given by

$$\begin{cases} X_0 \sim \nu(\cdot) \\ X_k | X_{k-1} = x_{k-1} \sim m(\cdot | x_{k-1}) \;\; k \geq 1 \\ Y_k | X_k = x_k \sim g(\cdot, x_k) \end{cases}$$

For the sake of simplicity the data are fixed that is, $Y_k = y_k$ for all time indexes. The Lebesgue measure is used as a reference measure in order to lighten the notations. We also omit the dependence of the so called marginal likelihood function $g(\cdot)$ to

the data by using the shortened notation $g_k(x_k) := g(x_k, y_k)$ and $p(\cdot)$ denotes a generic symbol for densities. The following notations will be used to introduce the quantities of interest. Let $\mathbf{F}_b(\mathcal{X}^{k+1})$ be the set of bounded and measurable functions on $\mathcal{X}^{k+1}$. Given an HMM with initial state X_0 distributed according to $\nu(\cdot)$, define

$$\phi_{\nu,0:k|k}(f) := \mathbb{E}_\nu \left[f(X_{0:k}) | Y_{1:k} \right], \quad k \geq 0, f \in \mathbf{F}_b(\mathcal{X}^{k+1}) \tag{1}$$

as the conditional distribution of $f(X_{0:k})$ given $Y_{1:k}$ with $X_0 \sim \nu(\cdot)$. Whenever f depends only on X_k, it is usual to simplify the notation to:

$$\phi_{\nu,k}(f) := \mathbb{E}_\nu \left[f(X_k) | Y_{1:k} \right], \quad k \geq 1, f \in \mathbf{F}_b(\mathcal{X}) \tag{2}$$

and we refer to this as the filter distribution that is, the conditional distribution of $f(X_k)$ given $Y_{1:k}$. We also introduce the $1-$step predictive distribution

$$\phi_{\nu,k|k-1}(f) := \mathbb{E}_\nu \left[f(X_k) | Y_{1:k-1} \right], \quad k \geq 1, f \in \mathbf{F}_b(\mathcal{X})$$

with the convention $\phi_{\nu,0|-1} := \nu$, where $\mathbb{E}_\nu$ is the expectation taken with the underlying law and emphasizing ν as the initial distribution of X_0. We also denote similarly the corresponding conditional densities of the later distributions as a slight abuse of notation. Their arguments help discriminate between these functions. For example, $\phi_{\nu,k}(x_k)$ is used to denote the filtering density while $\phi_{\nu,k|k-1}(x_k)$ is the $1-$step predictive density.

2.1. Filtering recursions

Particle filtering goal is to compute recursively in time the joint posterior distribution (1) or some of its features such as (2) a.k.a the filtering distribution. In terms of operator, (1) admits the compact recursive formula

$$\phi_{\nu,0:k|k}(f) = \frac{\phi_{\nu,0:k-1|k-1}(f g_k M)}{\phi_{\nu,0:k-1|k-1}(g_k M)}, \quad \forall f \in \mathbf{F}_b(\mathcal{X}^{k+1}) \tag{3}$$

and (2) satisfies the recursive formulas:

$$\phi_{\nu,k|k-1} = \phi_{\nu,k-1} M \tag{4}$$

and

$$\phi_{\nu,k}(f) = \frac{\phi_{\nu,k|k-1}(f g_k)}{\phi_{\nu,k|k-1}(g_k)}, \quad \forall f \in \mathbf{F}_b(\mathcal{X}). \tag{5}$$

Note that (3) and (5) are obtained via Bayes rule and (4) is a direct application of Kolmogorov equation. A more intuitive interpretation of these relations can be stated in terms of corresponding conditional densities given by:

$$\phi_{\nu,0:k|k}(x_{0:k}) = \frac{\phi_{\nu,0:k-1|k-1}(x_{0:k-1})m(x_{k-1},x_k)g_k(x_k)}{\int_{\mathcal{X}^{k+1}}\phi_{\nu,0:k-1|k-1}(x_{0:k-1})m(x_{k-1},x_k)g_k(x_k)dx_{0:k}} \tag{6}$$

for the joint posterior density and

$$\phi_{\nu,k|k-1}(x_k) = \int_{\mathcal{X}} m(x_{k-1},x_k)\phi_{\nu,k-1|k-1}(x_{k-1})dx_{k-1} \tag{7}$$

$$\phi_{\nu,k}(x_k) = \frac{g_k(x_k)\phi_{\nu,k|k-1}(x_k)}{\int_{\mathcal{X}} g_k(x_k)\phi_{\nu,k|k-1}(x_k)dx_k} \tag{8}$$

for the predictive and the filtering density respectively. So, particle filter is a two-step procedure that uses (4) as a prediction step for the next state and (8) as an update step according to the new observation. Within a Sequential Importance sampling procedure, one can get a PF estimate of (6) :

$$\hat{\phi}_{\nu,0:k|k}(dx_{0:k}) = \Omega_k^{-1}\sum_{i=1}^{N}\omega_k^{(i)}\delta_{\xi_{0:k}^{(i)}}(dx_{0:k})$$

and deduce an estimate of (8) as marginal distribution of the latter :

$$\hat{\phi}_{\nu,k}(dx_k) = \Omega_k^{-1}\sum_{i=1}^{N}\omega_k^{(i)}\delta_{\xi_k^{(i)}}(dx_k) \tag{9}$$

where $\Omega_k := \sum_{i=1}^{N}\omega_k^{(i)}$, $\delta_x(\cdot)$ is the Delta-Dirac mass located at x and $\omega_k^{(i)}$ is the importance weight associated to the particles position $\xi_{0:k}^{(i)}$. The detail derivation of these weights may be found in Doucet 2001, Doucet and Johansen 2011. A summary of particle filter is given bellow.
$q(\cdot)$ is a generic notation for instrumental densities in the Importance Sampling procedure. Note that the resampling step is done only if the degeneracy problem appears, for example when using the effective sample size approximation as a quantifier of this phenomena. Before moving towards, note that one can have an approximation of the joint posterior distribution $p(dx_{0:n}|y_{1:n})$ just on storing the outputs at each time step of the generic particle filter.

2.2. Smoothing recursions

The general idea shared by most of smoothing recursions is the nature of the reversed time of the dynamic model $\{X_k\}$. In fact, $\{X_k\}$ still a Markov chain, backward in time. The following result makes clear that assertion.

Proposition 1. *Given the data, $\{X_k\}$ is a Markov chain backward in time with transition backward kernels from $(\mathcal{X},\mathcal{B}(\mathcal{X}))$ to $(\mathcal{X},\mathcal{B}(\mathcal{X}))$ defined by:*

Algorithm 1 Generic particle filter

1: Initialization : For $i = 1, 2, \ldots, N$ draw $\xi_0^{(i)} \sim q(\cdot)$ and set $\omega_0^{(i)} = 1/N$;
2: Set $k \leftarrow 1$
3: Importance Sampling step: For $i = 1, 2, \ldots, N$
- draw $\bar{\xi}_k^{(i)} \sim q\left(\cdot|\xi_{0:k-1}^{(i)}, y_{1:k}\right)$
- Evaluate and Normalize the importance weights :

$$\omega_k^{(i)} \propto \omega_{k-1}^{(i)} \frac{m\left(\xi_{k-1}^{(i)}, \xi_k^{(i)}\right) g_k\left(\xi_k^{(i)}\right)}{q\left(\xi_k^{(i)}|\xi_{0:k-1}^{(i)}, y_{1:k}\right)}$$

4: Resampling step: (if necessary)
- Multiply/Discard $\bar{\xi}_k^{(i)}$ w.r.t $\omega_k^{(i)}$ to get $\xi_k^{(i)}$ approximately distributed according to $\phi_{\nu,k}$
- For $i = 1, 2, \ldots, N$ Set $\omega_k^{(i)} = 1/N$

5: Set $k \leftarrow k + 1$ and go to the importance sampling step

$$\begin{aligned} B_{k,\nu}(X_{k+1}, f) :&= \mathbb{E}\left[f(X_k)|X_{k+1:n}, Y_{0:n}\right] \\ &= \mathbb{E}\left[f(X_k)|X_{k+1}, Y_{0:k}\right] \end{aligned} \tag{10}$$

for any $f \in \mathbf{F}_b(\mathcal{X})$.

Proof. See Cappé, Moulines and Ryden 2005, p.70. □

Under this backward dynamic, one can make use of smoothing recursions.

2.2.1. Marginal smoothing

The problem in concern is to compute backward and recursively in time the smoothed distribution

$$\phi_{\nu,k|n}(f) := \mathbb{E}\left[f(X_k)|Y_{1:n}\right], \quad k < n$$

for any $f \in \mathbf{F}_b(\mathcal{X})$.

Lemma 1. *For any* $1 \leq k < n$, *the smoothed distribution factorizes as :*

$$\begin{aligned} \phi_{\nu,k|n}(f) &= \int_{\mathcal{X}} f(x_k) \left[\int_{\mathcal{X}} \frac{\phi_{\nu,k}(x_k) m(x_k, x_{k+1})}{\int_{\mathcal{X}} \phi_{\nu,k}(x_k) m(x_k, x_{k+1}) dx_k} \phi_{\nu,k+1|n}(dx_{k+1})\right] dx_k \\ &= \int_{\mathcal{X}^2} f(x_k) B_{\nu,k}(x_{k+1}, dx_k) \phi_{\nu,k+1|n}(dx_{k+1}) \end{aligned}$$

where $B_{k,\nu}(X_{k+1}, .)$ *is the Backward kernel for any function* $f \in \mathbf{F}_b(\mathcal{X})$.

Consider the generic particle filter gathering the weighted samples $\left\{\xi_k^{(i)}, \omega_k^{(i)}\right\}_{i=1}^N$ that target the filtering distributions $p(dx_k|y_{1:k})$ in the sense of (9), at time k with $k = 1, 2, \ldots, n$. In addition, assume at time $k+1$ one has weighted sample $\left\{\xi_{k+1}^{(i)}, \omega_{k+1|n}^{(i)}\right\}_{i=1}^N$ targeting the distribution $\phi_{k+1|n}$ in the sense:

$$\hat{\phi}_{\nu,k+1|n}(dx_{k+1}) = \sum_{j=1}^N \omega_{k+1|n}^{(j)} \delta_{\xi_{k+1}^{(j)}}(dx_{k+1}).$$

Combining the former and latter outputs, one can achieve a particle estimate of the smoothed distribution given by:

$$\hat{\phi}_{\nu,k|n}(f) = \sum_{i=1}^N \omega_{k|n}^{(i)} f(\xi_k^{(i)}), \quad \text{for } k < n$$

where the smoothed importance weights are given by :

$$\omega_{k|n}^{(i)} = \omega_k^{(i)} \left(\sum_{j=1}^N \omega_{k+1|n}^{(j)} \frac{m\left(\xi_k^{(i)}, \xi_{k+1}^{(j)}\right)}{\sum_{r=1}^N \omega_k^{(r)} m\left(\xi_k^{(r)}, \xi_{k+1}^{(j)}\right)} \right)$$

The summary of the procedure is given below:

Algorithm 2 Forward-Backward algorithm

1: Forward filtering step : For $k = 0, \ldots, n$
   - run the particles filtering algorithm to get the weighted particles $\left\{\xi_k^{(i)}, \omega_k^{(i)}\right\}_{i=1}^N$.

2: Backward smoothing step
   - For $i = 1, \ldots, N$ set $\omega_{n|n}^{(i)} = \omega_n^{(i)}$
   - For $k = n-1$ down to 0 and $i = 1, \ldots, N$ set

$$\omega_{k|n}^{(i)} = \omega_k^{(i)} \left(\sum_{j=1}^N \omega_{k+1|n}^{(j)} \frac{m\left(\xi_k^{(i)}, \xi_{k+1}^{(j)}\right)}{\sum_{r=1}^N \omega_k^{(r)} m\left(\xi_k^{(r)}, \xi_{k+1}^{(j)}\right)} \right)$$

Remark 1. *As one can notice, the F-B algorithm is nothing but a weigh update since particle positions generated in the forward pass are kept. Moreover, it is an $O(N^2)$ expensive algorithm at each time step.*

2.2.2. Joint smoothing

An extension of the F-B algorithm is reachable for the joint posterior density $p(x_{k:n}|y_{1:n})$. Using similar argument as in the marginal smoothing, one can obtain the following recursions:

Lemma 2. *Under (10), for any $k < n$ the joint smoothed density $p(x_{k:n}|y_{1:n})$ factorizes backward in time as :*

$$p(x_{k:n}|y_{1:n}) = p(x_k|x_{k+1}, y_{1:k})p(x_{k+1:n}|y_{1:n})$$

which iterates to:

$$p(x_{k:n}|y_{1:n}) = p(x_n|y_{1:n}) \prod_{r=k}^{n-1} p(x_r|x_{r+1}, y_{1:r}).$$

From this result, one is able to compute the conditional expectation

$$\begin{aligned} \phi_{k:n|n}(f) &:= \mathbb{E}\left[f(X_{k:n})|Y_{1:n}\right] \\ &= \int_{\mathcal{X}^{n-k+1}} f(x_{k:n})p(x_n|y_{1:n}) \prod_{r=k}^{n-1} p(x_r|x_{r+1}, y_{1:r})dx_{k:n} \end{aligned} \tag{11}$$

for any $f \in \mathbf{F}_b(\mathcal{X}^{n-k+1})$. Note at first that :

$$p(x_r|x_{r+1}, y_{1:r}) \propto p(x_r|y_{1:r})p(x_{r+1}|x_r)$$

From a particle estimate of the density $p(x_r|x_{r+1}, y_{1:r})$:

$$\hat{p}(x_r|x_{r+1}, y_{1:r}) = \sum_{i_r=1}^{N} \kappa_r^{(i_r)} \delta_{\xi_r^{(i_r)}}(x_r)$$

where

$$\kappa_r^{(i_r)} = \frac{\omega_r^{(i_r)} p(\xi_{r+1}^{(i_{r+1})}|\xi_r^{(i_r)})}{\sum_{l=1}^{N} \omega_r^{(l)} p(\xi_{r+1}^{i_{r+1}}|\xi_r^{(l)})}, \quad r = k, k+1, \ldots, n-1$$

one can achieve a particle estimate of (11):

$$\begin{aligned} \hat{\phi}_{\nu,k:n|n}(f) = \sum_{i_k=1}^{N} \sum_{i_{k+1}=1}^{N} \cdots \sum_{i_n=1}^{N} \omega_n^{i_n} \prod_{r=k}^{n-1} \frac{\omega_r^{(i_r)} p(\xi_{r+1}^{(i_{r+1})}|\xi_r^{(i_r)})}{\sum_{l=1}^{N} \omega_r^{(l)} p(\xi_{r+1}^{i_{r+1}}|\xi_r^{(l)})} \\ \times f(\xi_k^{(i_k)}, \xi_{k+1}^{(i_{k+1})}, \ldots, \xi_n^{(i_n)}), \end{aligned} \tag{12}$$

where $\left\{\xi_r^{(i_r)}, \omega_r^{(i_r)}\right\}_{i_r=1}^{N}, r = k, \ldots, n-1$ are sets of weighted particles targeting the filtering distribution $\phi_{\nu,r}$. Note that (12) has not a practical interest since it's complexity is exponential. Nevertheless, it is of great interest in a theoretical perspective. In fact, the deriving marginal smoother estimates inherit the convergence properties of the latter.

2.2.3. Particle smoother

One of the limitations of the F-B algorithm is its computational cost. In fact, it requires O(N^2) operations at each time step to compute the smoothed weights. Following Godsill *et al.* 2004 it is easy to get smoothed distribution estimate with a linear computational effort at each time step under (10). Extending Lemma 2 to whole time indexes one get:

Lemma 3. *The joint posterior density factorizes as:*

$$p(x_{0:n}|y_{1:n}) = p(x_n|y_{1:n}) \prod_{k=0}^{n-1} p(x_k|x_{k+1}, y_{1:k}).$$

Consider a particle estimate of the distribution $p(dx_k|x_{k+1}, y_{1:k})$:

$$\hat{p}(dx_k|x_{k+1}, y_{1:n}) = \sum_{i=1}^{N} \kappa_k^{(i)} \delta_{\xi_k^{(i)}}(dx_k)$$

where

$$\kappa_k^{(i)} = \frac{\omega_k^{(i)} p(x_{k+1}|\xi_k^{(i)})}{\sum_{j=1}^{N} \omega_k^{(j)} p(x_{k+1}|\xi_k^{(j)})}$$

Using the particle revision, one can draw consecutive states backward in time as follows. Assume $\tilde{\xi}_{k+1:n}$ to be a random sample drawn from $p(x_{k+1:n}|y_{1:n})$. Step back in time and draw $\tilde{\xi}_k$ from $p(x_k|\tilde{\xi}_{k+1:n}, y_{1:n})$. The sample $(\tilde{\xi}_k, \tilde{\xi}_{k+1:n})$ is an approximate random realization from $p(x_{k:n}|y_{1:n})$. Iterating the mechanism down to $k = 0$, one get a random sample from the joint smoothing density. The overall algorithm is given bellow.

Algorithm 3 Smoothing algorithm

1: For $i = 1, 2, \ldots, N$ choose $\tilde{\xi}_n = \xi_n^{(i)}$ with probability $\omega_n^{(i)}$
2: For $k = n - 1$ down to 0 and $i = 1, 2, \ldots, N$
   – Evaluate $\kappa_k^{(i)} \propto \omega_k^{(i)} p(\tilde{\xi}_{k+1}|\xi_k^{(i)})$;
   – Choose $\tilde{\xi}_k = \xi_k^{(i)}$ with probability $\kappa_k^{(i)}$;
3: $\tilde{\xi}_{0:n}$ is an approximate random realization from $p(x_{0:n}|y_{0:n})$.

This algorithm is an O(N) expensive at each time step.

3. SMC methods in High-order HMM

Consider the following state space model

$$\begin{cases} X_k &= a_k(X_{k-\ell:k-1}, V_k) \\ Y_k &= b_k(X_{k-\ell:k}, W_k) \end{cases}$$

where $a_k(\cdot)$ and $b_k(\cdot)$ are possibly nonlinear functions, $\{X_k\}$ is an $\ell-$order Markov chain with initial state sequences $X_{-\ell:-1}$ distributed according to a diffuse prior distribution $\nu(\cdot)$ and transition kernel M from $(\mathcal{X}^\ell, \mathcal{B}(\mathcal{X})^{\otimes \ell})$ to $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$. We assume that M admits a density function m w.r.t a dominating measure λ. $(V_k)_{k\geq 0}$ and $(W_k)_{k\geq 0}$ are i.i.d disturbance noises possibly correlated with $corr(V_i, W_j) = \rho 1_{i=j}$ and independent of $X_{-\ell:-1}$. We also assume that the observation process $\{Y_k\}$, constructed on the measurable space $(\mathcal{Y}, \mathcal{B}(\mathcal{Y}))$ is conditionally independent given $\{X_k\}$ with a marginal distribution admitting a density function g such that

$$\forall A \in \mathcal{B}(\mathcal{Y}), \quad \mathbb{P}(Y_k \in A | X_{k-\ell:k-1}, X_k) = \int_A g(X_{k-\ell:k-1}, X_k, y)\mu(dy),$$

where μ is a $\sigma-$finite measure on $(\mathcal{Y}, \mathcal{B}(\mathcal{Y}))$. For the sake of simplicity, the data are fixed that is, $Y_k = y_k$ for all time indexes. We also omit the dependence of likelihood function g to the data by using the short hand notation $g_k(\underline{x}_{k-1}, x_k) := g(x_{k-\ell:k-1}, x_k, \cdot)$, with $\underline{x}_{k-1} := x_{k-\ell:k-1}$. To sum up

$$\begin{cases} X_{-\ell:-1} \sim \nu(\cdot) \\ X_k | \underline{X}_{k-1} \sim m(\underline{x}_{k-1}, \cdot) \qquad k \geq 0 \\ Y_k | \underline{X}_{k-1}, X_k \sim g_k(\underline{x}_{k-1}, x_k) \end{cases}$$

3.1. $\ell-$order Filtering recursions

Consider the problem of computing recursively in time the following quantity :

$$\phi_{\nu,k:k+\ell-1|k+\ell-1}(f) := \mathbb{E}_\nu \left[f(X_{k:k+\ell-1}) | Y_{0:k+\ell-1} \right]$$

where $-\ell + 1 \leq k \leq n - \ell$, for any $f \in \mathbf{F}_b(\mathcal{X}^\ell)$. Notice that on taking $\ell = 1$ and $\rho = 0$, we fall in the classical nonlinear filtering problem in $1-$order HMM. Since we deal with state sequences, we shall call it in the sequel an ℓ-filtering problem and the resulting particle solution as an $\ell-$ particle filter to emphasize the overlapping $l-$ size vectors in concern. A common way to approximate such a distribution is to use a cloud of weighted particles $\left\{ \xi^{(i)}_{k:k+\ell-1}, \omega^{(i)}_{k+\ell-1} \right\}_{i=1}^N$ through the estimate :

$$\hat{\phi}_{\nu,k:k+\ell-1|k+\ell-1}(dz_{1:\ell}) = \Omega^{-1}_{k+\ell-1} \sum_{i=1}^N \omega^{(i)}_{k+\ell-1} \delta_{\xi^{(i)}_{k:k+\ell-1}}(dz_{1:\ell})$$

where $\Omega_{k+\ell-1} = \sum_{i=1}^N \omega^{(i)}_{k+\ell-1}$ and $\omega^{(i)}_{k+\ell-1}$ is obtained within an importance sampling procedure. The following result give a way to solve the ℓ-filtering problem recursively in time.

Proposition 2. *For any index* $-\ell \leq k \leq n-\ell$ *and* $f \in \mathbf{F}_b(\mathcal{X}^\ell)$, *the distribution* $\phi_{\nu,k:k+\ell-1|k+\ell-1}$ *satisfies the recursive relation :*

$$\begin{aligned}\phi_{\nu,k:k+\ell-1|k+\ell-1}(f) &\propto \int_{\mathcal{X}^{\ell+1}} f(x_{k:k+\ell-1})\phi_{\nu,k-1:k+\ell-2|k+\ell-2}(\underline{x}_{k+\ell-2}) \\ &\times m(\underline{x}_{k+\ell-2}, x_{k+\ell-1})g_{k+\ell-1}(\underline{x}_{k+\ell-2}, x_{k+\ell-1})dx_{k-1:k+l-1}\end{aligned}$$

with the convention $\phi_{\nu,-\ell:-1|-1} := \nu$.

Proof. It suffices to see that:

$$\begin{aligned}&p(x_{k:k+\ell-1}|y_{0:k+\ell-1}) \\ &= \int_{\mathcal{X}} p(x_{k-1:k+\ell-1}|y_{0:k+\ell-1})dx_{k-1} \propto \int_{\mathcal{X}} p(x_{k-1:k+\ell-1}, y_{0:k+\ell-1})dx_{k-1} \\ &\propto \int_{\mathcal{X}} p(x_{k-1:k+\ell-2}|y_{0:k+\ell-2})m(\underline{x}_{k+\ell-2}; x_{k+\ell-1})g_{k+\ell-1}(\underline{x}_{k+\ell-2}, x_{k+\ell-1})dx_{k-1}\end{aligned}$$

so that,

$$\begin{aligned}&\phi_{\nu,k:k+\ell-1|k+\ell-1}(f) \\ &\propto \int_{\mathcal{X}^\ell} f(x_{k:k+\ell-1}) \left(\int_{\mathcal{X}} p(x_{k-1:k+\ell-1}, y_{0:k+\ell-1})dx_{k-1} \right) dx_{k:k+\ell-1} \\ &\propto \int_{\mathcal{X}^{\ell+1}} f(x_{k:k+\ell-1})\phi_{\nu,k-1:k+\ell-1|k+\ell-2}(x_{k-1:k+\ell-2}) \\ &\times m(\underline{x}_{k+\ell-2}; x_{k+\ell-1})g_{k+\ell-1}(\underline{x}_{k+\ell-2}, x_{k+\ell-1})dx_{k-1:k+\ell-1}.\end{aligned} \tag{13}$$

which leads to the result. □

In order to highlight the two-step procedure mentioned above, the following operator formulation is given :

$$\phi_{\nu,k:k+\ell-1|k+\ell-2} = \phi_{\nu,k-1:k+\ell-2|k+\ell-2}M$$

as the prediction step and

$$\phi_{\nu,k:k+\ell-1|k+\ell-1}(f) = \frac{\phi_{\nu,k:k+\ell-1|k+\ell-2}(fg_{k+\ell-1})}{\phi_{\nu,k:k+\ell-1|k+\ell-2}(g_{k+\ell-1})} \tag{14}$$

as the correction step, for any $f \in \mathbf{F}_b(\mathcal{X}^\ell)$ and $-\ell+1 \leq k \leq n-\ell$. At time $(k+\ell-2)$, assume one has a cloud of weighted sample $\left\{\xi^{(i)}_{k-1:k+\ell-2}, \omega^{(i)}_{k+\ell-2}\right\}_{i=1}^N$ approximating the ℓ-filter distribution $\phi_{\nu,k-1:k+\ell-2|k+\ell-2}$ in the sense

$$\hat{\phi}_{\nu,k-1:k+\ell-2|k+\ell-2}(dx_{k-1:k+\ell-2}) = \Omega_{k+\ell-2}^{-1} \sum_{i=1}^{N} \omega_{k+\ell-2}^{(i)} \delta_{\xi_{k-1:k+\ell-2}^{(i)}}(dx_{k-1:k+\ell-2})$$

where $\Omega_{k+\ell-2} = \sum_{i=1}^{N} \omega_{k+\ell-2}^{(i)}$. A particle estimate at the next time step $(k+\ell-1)$ of the ℓ-filter distribution is achieved by :

$$\begin{aligned}
&\hat{\phi}_{\nu,k:k+\ell-1|k+\ell-1}(f) \\
&\propto \int_{\mathcal{X}} \Omega_{k+\ell-2}^{-1} \sum_{i=1}^{N} f(\xi_{k:k+\ell-2}^{(i)}, x_{k+\ell-1}) \omega_{k+\ell-2}^{(i)} m(\xi_{k-1:k+\ell-2}^{(i)}, x_{k+\ell-1}) \\
&\times g_{k+\ell-1}(\xi_{k-1:k+\ell-2}^{(i)}, x_{k+\ell-1}) dx_{k+\ell-1} \qquad (15) \\
&= \Omega_{k+\ell-2}^{-1} \sum_{i=1}^{N} \omega_{k+\ell-2}^{(i)} \int_{\mathcal{X}} f(\xi_{k:k+\ell-2}^{(i)}, x_{k+\ell-1}) m(\xi_{k-1:k+\ell-2}^{(i)}, x_{k+\ell-1}) \\
&\times g_{k+\ell-1}(\xi_{k-1:k+\ell-2}^{(i)}, x_{k+\ell-1}) dx_{k+\ell-1}.
\end{aligned}$$

where the last integral of (15) can be thought as expectation under either the transition density function or the likelihood density function. Note also that a mixture argument can be considered to evaluate it. Notice that these recursive weights are obtained within a classical bootstrap filter (see Gordon *et al.* 1993) or the general framework of the auxiliary particle filter (see Pitt and Shephard 1999).

3.2. ℓ-order smoothing recursions

Before stating ℓ-order smoothing, we precise some smoothing quantities that can be easily handled :

$$\phi_{\nu,k|n}(f) := \mathbb{E}_{\nu}\left[f(X_k)\middle|Y_{0:n}\right], k < n, \qquad (16)$$

$$\phi_{\nu,m,p|n}(g) := \mathbb{E}_{\nu}\left[g(X_p, X_m)\middle|Y_{0:n}\right], \mid p-m \mid \leq \ell+1, \qquad (17)$$

$$\phi_{\nu,-\ell:n|n}(h) := \mathbb{E}_{\nu}\left[h(X_{-\ell:n})\middle|Y_{0:n}\right], \qquad (18)$$

for any $f \in \mathbf{F}_b(\mathcal{X})$, $g \in \mathbf{F}_b(\mathcal{X}^2)$ and $h \in \mathbf{F}_b(\mathcal{X}^{n+\ell+1})$. Since (16) and (17) are particular cases of (18) we do not mention them here. In order to derive similar recursions as in $1-$order HMM, one needs to give the reversed time dynamic of the Markov chain through backward transition kernels. The following result shows that the hidden process still Markovian backward in time given the data.

Proposition 3. *Let ν be an initial distribution on $X_{-\ell:-1}$, $f \in \mathbf{F}_b(\mathcal{X})$, $n > 0$ and $-\ell + 1 \leq p \leq n - \ell$. Then $\{X_{n-k}\}_{k\geq 0}$ is a Markov chain with backward transition kernels defined by:*

$$\begin{aligned} B_{\nu,p+\ell-1}(X_{p+1:p+\ell}, f) &:= \mathbb{E}_\nu \left[f(X_p) \middle| X_{p+1:n}, Y_{0:n} \right] \\ &= \mathbb{E}_\nu \left[f(X_p) \middle| X_{p+1:p+\ell}, Y_{0:p+\ell-1} \right] \end{aligned} \tag{19}$$

Proof. see Appendix A:. □

3.2.1. Joint smoothing

To deal with (18) one needs the following factorization.

Lemma 4. *For any function $f \in \mathbf{F}_b(\mathcal{X}^{n+\ell+1})$, the joint smoothing distribution satisfies the backward kernels decomposition:*

$$\begin{aligned} \phi_{\nu,-\ell:n|n}(f) &= \int_{\mathcal{X}^{n+\ell+1}} f(x_{-\ell:n}) B_{\nu,-1}(x_{-\ell+1:0}, dx_{-\ell}) \phi_{\nu,-\ell+1:n|n}(dx_{-\ell+1:n}) \\ &= \int_{\mathcal{X}^{n+\ell+1}} f(x_{-\ell:n}) \phi_{\nu,n-\ell+1:n|n}(dx_{n-\ell+1:n}) \\ &\times \prod_{p=-\ell}^{n-\ell} B_{\nu,p+\ell-1}(x_{p+1:p+\ell}, dx_p). \end{aligned} \tag{20}$$

To get a particle estimate of (18), one needs to run the following two steps. In the first step, the $\ell-$filter distributions are approximated by

$$\hat{\phi}_{\nu,p:p+\ell-1|p+\ell-1}(dx_{p:p+\ell-1}) = \Omega_{p+\ell-1}^{-1} \sum_{i_p=1}^{N} \omega_{p+\ell-1}^{(i_p)} \delta_{\xi_{p:p+\ell-1}^{(i_p)}}(dx_{p:p+\ell-1}), \tag{21}$$

with $\left\{ \omega_{p+\ell-1}^{(i_p)}, \xi_{p:p+\ell-1}^{(i_p)} \right\}_{i_p=1}^{N}$ being the targeting weighted samples of the $\ell-$filter distributions $\phi_{\nu,p:p+\ell-1|p+\ell-1}(dz_{1:\ell})$, $p = -\ell, -\ell+1, \ldots, n-\ell$. The second step consists in approximating the backward kernels $B_{\nu,p+\ell-1}(x_{p+1:p+\ell}, dx_p)$ by:

$$\hat{B}_{\nu,p+\ell-1}(x_{p+1:p+\ell}, dx_p) = \sum_{i_p=1}^{N} \frac{\omega_{p+\ell-1}^{(i_p)} m(\xi_{p:p+\ell-1}^{(i_p)}, x_{p+\ell})}{\sum_{r=1}^{N} \omega_{p+\ell-1}^{(r)} m(\xi_{p:p+\ell-1}^{(r)}, x_{p+\ell})} \delta_{\xi_p^{(i_p)}}(dx_p) \tag{22}$$

$p = -\ell, -\ell+1, \ldots, n-\ell$.

Plugging (21) and (22) into (20), a particle estimate of (18) is given by:

$$\hat{\phi}_{\nu,-\ell:n|n}(f) = \Omega_n^{-1} \sum_{i_n=1} \left[\sum_{i_{-\ell}=1}^{N} \cdots \sum_{i_{n-\ell}=1}^{N} f(\xi_{-\ell}^{(i_{-\ell})}, \ldots, \xi_{n-\ell}^{(i_{n-\ell})}, \xi_{n-\ell+1:n}^{(i_n)}) \times \prod_{p=-\ell}^{n-\ell} \frac{\omega_{p+\ell-1}^{(i_p)} m(\xi_{p:p+\ell-1}^{(i_p)}, \xi_{p+\ell}^{(i_{p+\ell})})}{\sum_{r=1}^{N} \omega_{p+\ell-1}^{(r)} m(\xi_{p:p+\ell-1}^{(r)}, \xi_{p+\ell}^{(i_{p+\ell})})} \right] \omega_n^{(i_n)}, \tag{23}$$

for $f \in \mathbf{F}_b\left(\mathcal{X}^{n+\ell+1}\right)$. Before moving towards the theoretical properties of this estimator, we give a summary description of the former procedure.

Algorithm 4 Smoothing in $\ell-$order HMM

1: Forward pass: For $p = -\ell, -\ell+1, \ldots, n-\ell+1$ approximate $\phi_{\nu,p:p+\ell-1|p+\ell-1}$ by

$$\hat{\phi}_{\nu,p:p+\ell-1|p+\ell-1}(dx_{p:p+\ell-1}) = \Omega_{p+\ell-1}^{-1} \sum_{i=1}^{N} \omega_{p+\ell-1}^{(i)} \delta_{\xi_{p:p+\ell-1}^{(i)}}(dx_{p:p+\ell-1})$$

2: Backward pass: For $p = n-\ell$ down to $-\ell$ approximate $B_{\nu,p+\ell-1}$ by:

$$\hat{B}_{\nu,p+\ell-1}(x_{p+1:p+\ell}, dx_p) = \sum_{i_p=1}^{N} \frac{\omega_{p+\ell-1}^{(i_p)} m(\xi_{p:p+\ell-1}^{(i_p)}, x_{p+\ell})}{\sum_{r=1}^{N} \omega_{p+\ell-1}^{(r)} m(\xi_{p:p+\ell-1}^{(r)}, x_{p+\ell})} \delta_{\xi_p^{(i_p)}}(dx_p)$$

Once the two passes performed, one can approximate (18) using (23) and deduce approximation for (16) and (17) as marginal of the latter.

3.2.2. Particle smoother

One may also achieve similar particle smoother to those of Godsill *et al.* 2004 using the following identity.

Lemma 5. *Under Prop.3, the joint smoothing density factorizes as*

$$p(x_{-\ell:n}|y_{0:n}) = p(x_{n-\ell+1:n}|y_{0:n}) \prod_{k=-\ell}^{n-\ell} p(x_k|x_{k+1:n}; y_{0:n})$$

where

$$\begin{aligned} p(x_k|x_{k+1:n}, y_{0:n}) &= p(x_k|x_{k+1:k+\ell}, y_{0:k+\ell-1}) \\ &\propto p(x_{k+\ell}|x_{k:k+\ell-1}) p(x_{k:k+\ell-1}|y_{0:k+\ell-1}). \end{aligned}$$

Assume one has run the $\ell-$order filter mentioned previously to get the weighted particles

$$\left\{\omega_{k+\ell-1}^{(i)}, \xi_{k:k+\ell-1}^{(i)}\right\}_{i=1}^{N}, \quad -\ell+1 \leq k \leq n-\ell$$

approximating the $\ell-$filter densities $p(x_{k:k+\ell-1}|y_{0:k+\ell-1})$. Using the previous weighted sample one could get a particle estimate of $p(x_k|x_{k+1:k+\ell}, y_{0:k+\ell-1})$:

$$p(dx_k|x_{k+1:k+\ell}, y_{0:k+\ell-1}) \approx \sum_{i=1}^{N} \kappa_k^{(i)} \delta_{\xi_k^{(i)}}(dx_k)$$

where the modified weights are given by

$$\kappa_k^{(i)} = \frac{\omega_{k+\ell-1}^{(i)} m(\xi_{k:k+\ell-1}^{(i)}, x_{k+\ell})}{\sum_{j=1}^{N} \omega_{k+\ell-1}^{(j)} m(\xi_{k:k+\ell-1}^{(j)}, x_{k+\ell})}$$

With these modified weights, one can simulate consecutive states in the reverse-time as follows. Let $\tilde{x}_{k+1:n}$ be a random sample drawn from $p(x_{k+1:n}|y_{0:n})$, step back in time and draw $\tilde{x}_k$ from $p(x_k|\tilde{x}_{k+1:n}, y_{0:n})$. The pair $(\tilde{x}_k, \tilde{x}_{k+1:n})$ is an approximate random realization of $p(x_{k:n}|y_{0:n})$. Iterating this mechanism backward in time one gets the smoothing algorithm.

Algorithm 5 Particle smoother in $\ell-$order HMM

1: Choose $\tilde{\xi}_{n-\ell+1:n} = \xi_{n-\ell+1:n}^{(i)}$ with probability $\omega_n^{(i)}$
2: For $k = n-\ell$ down to $-\ell$ do
   – Evaluate $\kappa_k^{(i)} \propto \omega_{k+\ell-1}^{(i)} m(\xi_{k:k+\ell-1}^{(i)}, \tilde{\xi}_{k+\ell})$, for $i = 1, \ldots, N$;
   – Choose $\tilde{\xi}_k = \xi_k^{(i)}$ with probability $\kappa_k^{(i)}$
3: EndFor
4: $\tilde{\xi}_{-\ell:n} = (\tilde{\xi}_{-\ell}, \tilde{\xi}_{-\ell+1}, \ldots, \tilde{\xi}_n)$ is an approximate random realization from $p(x_{-\ell:n}|y_{0:n})$.

The computational complexity is $O(N)$ at each time step which compares favorably to the $O(N^2)$ of the marginal smoothing.

4. Parameter estimation

MCEM as a combination of the GEM with SMC is a tool that can be used to estimate HMM when dealing with latent process. We do not fully detail the GEM algorithm since it is well documented (see Dempster, Laird and Rubin 1977 or McLachlan and Krishnan 2008 for a review). However, the main idea is depicted bellow :

Algorithm 6 Generalized EM algorithm

1: Choose an initial guess $\boldsymbol{\theta^{(0)}}$
2: **For** $m = 1, 2, \ldots$ **do**
   1. **E-Step :** Compute $Q(\boldsymbol{\theta}, \boldsymbol{\theta^{(m-1)}})$
   2. **M-Step :** Find $\boldsymbol{\theta^{(m)}}$ s.t $Q(\boldsymbol{\theta^{(m)}}, \boldsymbol{\theta^{(m-1)}}) \geq Q(\boldsymbol{\theta^{(m-1)}}, \boldsymbol{\theta^{(m-1)}})$
3: **EndFor.**

The E-step consists in computing the *intermediate quantity*, that is the conditional expectation of the logarithm of the complete data likelihood given the data and the current value of the parameter vector $\theta^{(m-1)}$:

$$Q(\theta^{(m)}, \theta^{(m-1)}) = \mathbb{E}_{\theta^{(m-1)}} \left[\log p_{\theta^{(m)}} \left(X_{-2:n}, Y_{0:n} \right) | Y_{0:n} \right]$$

where $(n+1)$ is the sample size of the data indexed from 0 to n, $p_{\theta^{(m)}}$ a generic notation for densities depending on parameter $\theta^{(m)}$ and X is a hidden signal initialized to a diffuse prior distribution ν on $X_{-2:-1}$ and Y the observation process. As a first illustration, consider the following toy example

$$\begin{cases} X_k = \pi_1 X_{k-1} + \pi_2 X_{k-2} + \sigma_W W_k \\ Y_k = X_k + \sigma_V V_k, \end{cases} \tag{24}$$

We assume that $(V_k, W_k)_{k \geq 0}$ are i.i.d and independent of $X_{-2:-1}$ with $(V_k, W_k) \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \right)$,
where $|\pi_1 \pm \pi_2| < 1$ and $|\pi_2| < 1$ ensuring the stationarity of X. At iteration m of the GEM, the parameters are updated through the recursive scheme

$$\begin{cases} \pi_1^{(m)} & = \dfrac{\sum_{k=2}^{n} \mathbb{E}_{\theta^{(m-1)}} \left[\left(X_{k-1} X_k - \pi_2^{(m)} X_{k-1} X_{k-2} \right) \middle| Y'_{0:n} \right]}{\sum_{k=1}^{n} \mathbb{E}_{\theta^{(m-1)}} \left[X_{k-1}^2 | Y'_{0:n} \right]} \\ \pi_2^{(m)} & = \dfrac{\sum_{k=2}^{n} \mathbb{E}_{\theta^{(m-1)}} \left[\left(X_{k-2} X_k - \pi_1^{(m)} X_{k-1} X_{k-2} \right) \middle| Y'_{0:n} \right]}{\sum_{k=2}^{n} \mathbb{E}_{\theta^{(m-1)}} \left[X_{k-2}^2 | Y'_{0:n} \right]} \\ \left[\sigma_V^{(m)} \right]^2 & = \frac{1}{n} \sum_{k=2}^{n} \mathbb{E}_{\theta^{(m-1)}} \left[(Y_k - X_k)^2 \middle| Y_{0:n} \right] \\ \left[\sigma_W^{(m)} \right]^2 & = \frac{1}{n} \sum_{k=2}^{n} \mathbb{E}_{\theta^{(m-1)}} \left[\left(X_k - \pi_1^{(m)} X_{k-1} - \pi_2^{(m)} X_{k-2} \right)^2 \middle| Y_{0:n} \right] \end{cases}$$

As a second illustration, consider the following discrete stochastic volatility model

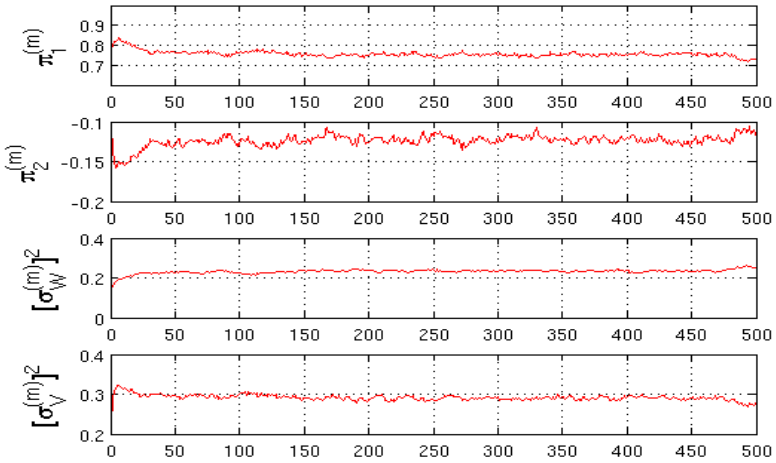

Fig. 1. MCEM iterations for (24) with true parameter $\theta^* = (0.7, -0.15, 0.2, 0.3)$

$$\begin{cases} X_k = \pi_1 X_{k-1} + \pi_2 X_{k-2} + \sigma W_k \\ Y_k = \beta \exp(X_k/2)\, V_k, \end{cases} \tag{25}$$

under the same assumptions as in the former model. Since (V_k) and (W_k) are independent and Gaussian it's common to use a linearized version of (25) given by:

$$\begin{cases} X_{k+1} = \pi_1 X_k + \pi_2 X_{k-1} + \sigma W_{k+1} \\ Y'_{k+1} = \alpha + X_{k+1} + \eta_{k+1} - \zeta \end{cases} \tag{26}$$

where $Y'_{k+1} := \log Y_{k+1}^2$, $\eta_k := \log V_k^2$ are i.i.d noises independent of (W_k) with a $\log \chi^2(1)$ distribution, $\zeta := \mathbb{E}(\log V_k^2) = -1.27049$, $\alpha := \log \beta^2 + \zeta$ and $\theta := (\pi_1, \pi_2, \sigma, \alpha)$ is the parameter vector. On taking the derivatives of $Q(\theta^{(m)}, \theta^{(m-1)})$ with respect to each parameter on gets the recursive following parameter update :

$$\begin{cases} \alpha^{(m)} &= \log\left[\frac{1}{n}\sum_{k=0}^{n}\mathbb{E}_{\theta^{(m-1)}}\left[\exp(Y_k - X_k + \zeta)|Y'_{0:n}\right]\right] \\ \pi_1^{(m)} &= \dfrac{\sum_{k=2}^{n}\mathbb{E}_{\theta^{(m-1)}}\left[\left(X_{k-1}X_k - \pi_2^{(m)}X_{k-1}X_{k-2}\right)\middle|Y'_{0:n}\right]}{\sum_{k=1}^{n}\mathbb{E}_{\theta^{(m-1)}}\left[X_{k-1}^2|Y'_{0:n}\right]} \\ \pi_2^{(m)} &= \dfrac{\sum_{k=2}^{n}\mathbb{E}_{\theta^{(m-1)}}\left[\left(X_{k-2}X_k - \pi_1^{(m)}X_{k-1}X_{k-2}\right)\middle|Y'_{0:n}\right]}{\sum_{k=2}^{n}\mathbb{E}_{\theta^{(m-1)}}\left[X_{k-2}^2|Y'_{0:n}\right]} \\ \sigma^{(m)} &= \sqrt{\frac{1}{n}\sum_{k=2}^{n}\mathbb{E}_{\theta^{(m-1)}}\left[\left(X_k - \pi_1^{(m)}X_{k-1} - \pi_2^{(m)}X_{k-2}\right)^2\middle|Y'_{0:n}\right]} \end{cases}$$

As a synthetic example, Fig.(2) is generated using the true parameter vector $(\pi_1^* = 0.8, \pi_2^* = 0.10, \sigma^* = \sqrt{0.3}, \log[\beta^*]^2 = -0.8612)$.

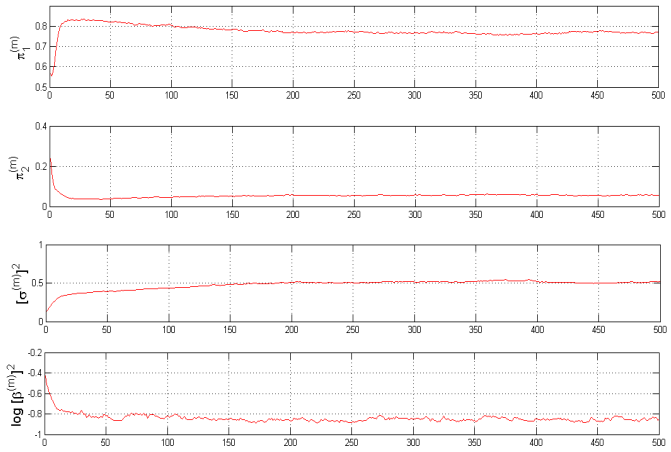

Fig. 2. MCEM iterations for (26) with $\theta^* = (0.8, 0.10, 0.3, -0.8612)$

As point perspective, the adjustment of the smoothing weights is required. Indeed, one can notice that the estimates are not entirely satisfactory for some parameters in this example. More attention is needed to correct this shortcoming.

5. Conclusion

In this paper, we were interested in extending classical sequential Monte Carlo Methods in high-Order hidden Markov models in a methodological perspective. We

have shown that it is possible to have similar recursive solutions when dealing with posterior distributions in HMM whether for smoothing or filtering purposes. We also illustrate some applications via parameters inference of linear Gaussian model and stochastic volatility model using EM algorithm. This work is far from over. We did not discuss the convergence of smoothing and filtering quantities. However, there is a good chance of being able to adapt certain convergence results existing in the literature of the SMC methods, notably those in Del Moral 2004, Olsson *et al.* 2008 or Jasra 2015 among others.

Appendix A: Proof of Proposition 3

One may use the following intermediate result.

Lemma 6. *For any function* $f \in \mathbf{F}_b(\mathcal{X}^\ell)$ *and index* $k \geq 1-\ell$,

$$\phi_{\nu,k:k+\ell-1|k+\ell-1}(f)L_{k+\ell-1} = \int_{\mathcal{X}^{k+2\ell}} f(x_{k:k+\ell-1})\nu(x_{-\ell:-1}) \prod_{i=0}^{k+\ell-1} m(\underline{x}_{i-1}; x_i) g_i(\underline{x}_{i-1}, x_i) dx_{-\ell:k+\ell-1}$$

where $L_{k+\ell-1}$ *denotes the likelihood density of* $y_{0:k+\ell-1}$.

Proof. It suffices to see that:

$$\begin{aligned} p(x_{k:k+\ell-1}|y_{0:k+\ell-1}) &= \int_{\mathcal{X}^{k+\ell}} p(x_{-\ell:k+\ell-1}|y_{0:k+\ell-1})dx_{-\ell:k-1} \\ &= \int_{\mathcal{X}^{k+\ell}} \frac{p(x_{-\ell:k+\ell-1}, y_{0:k+\ell-1})}{p(y_{0:k+\ell-1})} dx_{-\ell:k-1} \\ &= L_{k+\ell-1}^{-1} \int_{\mathcal{X}^{k+\ell}} p(x_{-\ell:k+\ell-1}, y_{0:k+\ell-1})dx_{-\ell:k-1} \end{aligned} \tag{A1}$$

Using (A1), the expectation of $f(X_{k:k+\ell-1})$ conditional on $Y_{0:k+\ell-1}$ is given by :

$$\begin{aligned} &\phi_{\nu,k:k+\ell-1|k+\ell-1}(f) \\ &= \int_{\mathcal{X}^\ell} f(x_{k:k+\ell-1}) \left(L_{k+\ell-1}^{-1} \int_{\mathcal{X}^{k+\ell}} p(x_{-\ell:k+\ell-1}, y_{0:k+\ell-1})dx_{-\ell:k-1} \right) dx_{k:k+\ell-1} \\ &= L_{k+\ell-1}^{-1} \int_{\mathcal{X}^{k+2\ell}} f(x_{k:k+\ell-1}) p(x_{-\ell:k+\ell-1}, y_{0:k+\ell-1})dx_{-\ell:k+\ell-1} \\ &= L_{k+\ell-1}^{-1} \int_{\mathcal{X}^{k+2\ell}} f(x_{k:k+\ell-1})\nu(x_{-\ell:-1}) \prod_{i=0}^{k+\ell-1} m(\underline{x}_{i-1}, x_i) g_i(\underline{x}_{i-1}, x_i) dx_{-\ell:k+\ell-1} \end{aligned}$$

which leads to the identity. □

Note that this identity is extensible up to the final time index n :

$$\phi_{\nu,k:n|n}(f)L_n = \int_{\mathcal{X}^{n+\ell+1}} f(x_{k:n})\nu(x_{-\ell:-1}) \prod_{i=0}^{n} m(\underline{x}_{i-1}; x_i) g_i(\underline{x}_{i-1}, x_i) dx_{-\ell:n}$$

for any function $f \in \mathbf{F}_b(\mathcal{X}^{n-k+1})$.

Proof. From previous lemma, for any functions $e \in \mathbf{F}_b(\mathcal{X}^{\ell-1})$, $f \in \mathbf{F}_b(\mathcal{X})$ and $h \in \mathbf{F}_b(\mathcal{X}^{n-k-\ell+1})$,

$$\begin{aligned}
&\mathbb{E}\left[f(X_k)e(X_{k+1:k+\ell-1})h(X_{k+\ell:n})\middle|Y_{0:n}\right] \\
&= \int_{\mathcal{X}^{n-k+1}} f(x_k)e(x_{k+1:k+\ell-1})h(x_{k+\ell:n})\phi_{\nu,k:n|n}(dx_{k:n}) \\
&= L_n^{-1} \int_{\mathcal{X}^{k+2\ell+1}} f(x_k)e(x_{k+1:k+\ell-1})\nu(x_{-\ell:-1}) \prod_{i=0}^{k+\ell-1} m(\underline{x}_{i-1}, x_i) g_i(\underline{x}_{i-1}, x_i) \\
&\times m(\underline{x}_{k+\ell-1}, x_{k+\ell}) g_{k+\ell}(\underline{x}_{k+\ell-1}, x_{k+\ell}) \\
&\times \left[\int_{\mathcal{X}^{n-k-\ell}} h(x_{k+\ell:n}) \prod_{i=k+\ell+1}^{n} m(\underline{x}_{i-1}, x_i) g_i(\underline{x}_{i-1}, x_i) dx_{k+\ell+1:n}\right] dx_{-\ell:k+\ell} \\
&= \frac{L_{k-\ell+1}}{L_n} \int_{\mathcal{X}^{\ell+1}} f(x_k)e(x_{k+1:k+\ell-1})\phi_{\nu,k:k+\ell-1|k+\ell-1}(dx_{k:k+\ell-1}) m(\underline{x}_{k+\ell-1}; x_{k+\ell}) \\
&\times g_{k+\ell}(\underline{x}_{k+\ell-1}, x_{k+\ell}) \left[\int_{\mathcal{X}^{n-k-\ell}} h(x_{k+\ell:n}) \prod_{i=k+\ell+1}^{n} m(\underline{x}_{i-1}; x_i) g_i(\underline{x}_{i-1}, x_i) dx_{k+\ell+1:n}\right] dx_{k+\ell}
\end{aligned}$$

using the implicit definition of the backward kernel (19) applied to the function

$$\begin{aligned}
r(x_{k:k+\ell-1}, x_{k+\ell}) &= f(x_k)e(x_{k+1:k+\ell-1})g_{k+\ell}(\underline{x}_{k+\ell-1}, x_{k+\ell}) \\
&\times \left[\int_{\mathcal{X}^{n-k-\ell}} h(x_{k+\ell:n}) \prod_{i=k+\ell+1}^{n} m(\underline{x}_{i-1}; x_i) g_i(\underline{x}_{i-1}, x_i) dx_{k+\ell+1:n}\right] dx_{k+\ell}
\end{aligned}$$

one could get

$$\begin{aligned}
&\mathbb{E}\left[f(X_k)e(X_{k+1:k+\ell-1})h(X_{k+\ell:n})\middle|Y_{0:n}\right] \\
&= \frac{L_{k-\ell+1}}{L_n}\int_{\mathcal{X}^{\ell+1}} B_{\nu,k+\ell-1}(x_{k+1:k+\ell}, dx_k)f(x_k)e(x_{k+1:k+\ell-1}) \\
&\times \phi_{\nu,k+1:k+\ell|k+\ell-1}(dx_{k+1:k+\ell})g_{k+\ell}(\underline{x}_{k+\ell-1}, x_{k+\ell}) \\
&\times \left[\int_{\mathcal{X}^{n-k-\ell}} h(x_{k+\ell:n}) \prod_{i=k+\ell+1}^{n} m(\underline{x}_{i-1}; x_i)g_i(\underline{x}_{i-1}, x_i)dx_{k+\ell+1:n}\right]
\end{aligned} \tag{A2}$$

taking $f \equiv 1$, for any functions $h' \in \mathbf{F}_b(\mathcal{X}^{n-k-\ell+1})$ and $e' \in \mathbf{F}_b(\mathcal{X}^{\ell-1})$

$$\begin{aligned}
&\mathbb{E}\left[e'(X_{k+1:k+\ell-1})h'(X_{k+\ell:n})\middle|Y_{0:n}\right] \\
&= \frac{L_{k-\ell+1}}{L_n}\int_{\mathcal{X}^{\ell}} e'(x_{k+1:k+\ell-1})\phi_{\nu,k+1:k+\ell|k+\ell-1}(dx_{k+1:k+\ell})g_{k+\ell}(\underline{x}_{k+\ell-1}, x_{k+\ell}) \\
&\times \left[\int_{\mathcal{X}^{n-k-\ell}} h'(x_{k+\ell:n}) \prod_{i=k+\ell+1}^{n} m(\underline{x}_{i-1}, x_i)g_i(\underline{x}_{i-1}, x_i)dx_{k+\ell+1:n}\right]
\end{aligned}$$

Identifying $e'h'$ with $e(x_{k+1:k+\ell-1})h(x_{k+\ell:n})\int_{\mathcal{X}} B_{\nu,k+\ell-1}(x_{k+1:k+\ell}, x)f(x)dx$, (A2) may be rewritten as

$$\begin{aligned}
&\mathbb{E}\left[f(X_k)e(X_{k+1:k+\ell-1})h(X_{k+\ell:n})\middle|Y_{0:n}\right] \\
&= \mathbb{E}\left[e(X_{k+1:k+\ell-1})h(X_{k+\ell:n})\int_{\mathcal{X}} B_{\nu,k+\ell-1}(x_{k+1:k+\ell}, x)f(x)dx\middle|Y_{0:n}\right]
\end{aligned}$$

which leads to the result. □

References

Allaya, M. M. (2013). Méthodes de Monte Carlo EM et Approximations particulaires: Application à la Calibration d'un modèle de volatilité. *PhD thesis*, Université Paris 1 Panthéon-Sorbonne
et al.

Cappé, O., Moulines, E. and Ryden T. (2005). Inference in Hidden Markov Models, Springer.

Cappé, O., Godsill, S., J. and Moulines, E. (2007). An overview of existing methods and recent advances in sequential Monte Carlo, *Proceedings of the IEEE*, Vol. 95, (5) p. 899-924

Chan, K. S. and Ledolter, J. (1995). Monte Carlo EM Estimation for Times Series Models Involving Counts. *Journal of American Statistical Association*, Vol. 90, (429), pp. 242-252.

Del Moral, P. (2004). Feynman-Kac Formulae. Genealogical and Interacting Particle Systems with Applications, *Springer*

Del Moral, P. and Guionnet, A. (1998). On the stability of measure valued processes. Applications to nonlinear filtering and interacting particle systems, *Publication du laboratoire de Statistique et Probabilités* 3-98, Université Paul Sabatier, Toulouse,

Dempster, A. P. and Laird, N. M. and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society Series B*, Vol 39, (1)

Doucet, A. and de Freitas, N. and Gordon, N., J. (2001). Sequential Monte Carlo Methods in Practice : Statistics for Engineering and Information Science, Springer-Verlag, New York,

Doucet, A. and Johansen, A., M. (2011). *A tutorial on particle filtering and smoothing: Fifteen years later*, The Oxford Handbook of Nonlinear Filtering.

Doucet, A. and Godsill, S., J. and Andrieu, C. (2000). On Sequential Monte Carlo Sampling methods for Bayesian filtering. *Statistics and Computing*, Vol. 10, (3), pp. 197-208

Engle, R. F. (1982). Autoregressive Conditional Heteroscedasticity with Estimates of the Variance of United Kingdom Inflation. *Econometrica*. Vol.50, (4) pp. 987-1007.

Godsill, S., J. and Doucet, A. and West, M. (2004). Monte Carlo Smoothing for Nonlinear Time Series *Journal of the American Statistical Association*, Vol. 99, (465), p. 156-168

Gordon, N. J., Salmond, D. J., and Smith, A. F. M. (1993). Novel approach to non-linear non-Gaussian Bayesian state estimation. *Radar and Signal Processing, IEE Proceedings F*, Vol. 140, (2), p. 107-113

Handschin, J., (1970). Monte Carlo techniques for prediction and filtering of non-linear stochastic processes, *Automatica*, p. 555-563, Vol.6,

Handschin, J. and Mayne, D. (1969). Monte Carlo techniques to estimate the conditional expectation in multi-stage non-linear filtering, *Int. J. Control*, Vol. 9, pp. 547-559

Heston, S. L. (1993). A Closed-Form Solution for Options with Stochastic Volatility with Applications to Bond and Currency Options. *The Review of Financial Studies*, Vol. 6, (6), pp 327-343

Hu, X. and Schön, T., B. (2011). A General Convergence Result for Particle Filtering, *IEEE Transactions on Signal Processing*, Vol. 59, (7), p. 3424-3429

Jasra A. (2015) On the behaviour of the backward Feynman-Kac formula., *Journal of Appl. Probab.*, Vol. 52, (2), p. 339-359

Kitagawa, G. (1996). Monte Carlo filter and smoother for non-Gaussian nonlinear state space models, Journal of Computational and Graphical Statistics, Vol. 5, (1), pp. 1-25

McLachlan, G. J. and Krishnan, T. (2008) The EM Algorithm and Extensions *Wiley Series in Probability and Statistics*,

Olsson, J. and Cappé, O. and Douc, R. and Moulines, E. (2008). *Sequential Monte Carlo smoothing with application to parameter estimation in nonlinear state space models*, Bernoulli 14, (1), p. 155-179,

Pitt, M., K. and Shephard, N. (1999). Filtering via Simulation: Auxiliary Particle Filters. Journal of the American Statistical Association, Vol. 94, (446), p. 590-599