Particle approximation improvement of the joint smoothing distribution with on-the-fly variance estimation

Cyrille Dubarry

TELECOM SudParis
Département CITI
9 rue Charles Fourrier
Evry, France.
cyrille.dubarry@telecom-sudparis.eu

Randal Douc

TELECOM SudParis
Département CITI
9 rue Charles Fourrier
Evry, France.
randal.douc@telecom-sudparis.eu

Abstract. Particle smoothers are widely used algorithms allowing to approximate the smoothing distribution in hidden Markov models. Existing algorithms often suffer from slow computational time or degeneracy. We propose in this paper a way to improve any of them with a linear complexity in the number of particles. When iteratively applied to the degenerated Filter-Smoother, this method leads to an algorithm which turns out to outperform existing linear particle smoothers for a fixed computational time. Moreover, the associated approximation satisfies a central limit theorem with a close-to-optimal asymptotic variance, which be easily estimated by only one run of the algorithm.

Keywords: Degeneracy, Hidden Markov model, Particle smoothing, Sequential Monte-Carlo, Variance estimation

1. Introduction

A *hidden Markov model* (HMM) is a doubly stochastic process where a Markov chain $\{X_t\}_{t=0}^{\infty}$ is only partially observed through a sequence of observations $\{Y_t\}_{t=0}^{\infty}$. More precisely, let $\mathbb{X}$ and $\mathbb{Y}$ be two spaces equipped with countably generated σ-fields $\mathcal{X}$ and $\mathcal{Y}$, respectively, and denote by M a Markovian transition kernel on $(\mathbb{X}, \mathcal{X})$ and by G a transition kernel from $(\mathbb{X}, \mathcal{X})$ to $(\mathbb{Y}, \mathcal{Y})$. In our setting, the dynamics of the bivariate process $\{(X_k, Y_k)\}_{k=0}^{\infty}$ follows the Markovian transition kernel

$$P\left[(x,y), \mathsf{A}\right] \stackrel{\text{def}}{=} M \otimes G[(x,y), \mathsf{A}] = \iint M(x, \mathrm{d}x')\, G(x', \mathrm{d}y') \mathbf{1}_{\mathsf{A}}(x', y') \;, \tag{1}$$

where $(x, y) \in \mathbb{X} \times \mathbb{Y}$ and $\mathsf{A} \in \mathcal{X} \otimes \mathcal{Y}$.

This work is supported by the Agence Nationale de la Recherche (ANR, 212, rue de Bercy 75012 Paris) through the 2009-2012 project Big MC

We assume that there exist nonnegative σ-finite measures λ on $(\mathbb{X}, \mathcal{X})$ and μ on $(\mathbb{Y}, \mathcal{Y})$ such that for any $x \in \mathbb{X}$, $M(x, \cdot)$ and $G(x, \cdot)$ are dominated by λ and μ, respectively. This implies the existence of kernel densities

$$m(x, x') \stackrel{\text{def}}{=} \frac{\mathrm{d}M(x, \cdot)}{\mathrm{d}\lambda}(x') \quad \text{and} \quad g(x, y) \stackrel{\text{def}}{=} \frac{\mathrm{d}G(x, \cdot)}{\mathrm{d}\mu}(y) \, .$$

In what follows, we simply write $\mathrm{d}x$ for $\lambda(\mathrm{d}x)$.

We are interested here in estimating the expectation of a function of $(X_0, \ldots, X_T)$ conditionally on the observations $Y_0, \ldots, Y_T$ using particle smoothing algorithms. Many different implementations of the particle filters and smoothers have been proposed in the literature with different computational costs; see for example Del Moral (2004); Cappé et al. (2005); Doucet and Johansen (2009). So far, the existing particle smoothers rely on the so-called *Forward-Filter* whose complexity is linear in the number of particles N. In its simplest extension, storing the paths of the Forward-Filter allows to approximate the joint smoothing distribution as seen by Kitagawa (1996). This method known as the *Filter-Smoother* unfortunately suffers from a poor representation of the states corresponding to times $t \ll T$. To circumvent this drawback, the *FFBS (Forward Filtering Backward Smoothing)* algorithm introduced by Doucet et al. (2000) adds a backward pass to the forward filter at the cost of a quadratic complexity when used for approximating the marginal smoothing distributions. However, Godsill et al. (2004) extended it to the *FFBSi (Forward Filtering Backward Simulation)*, an algorithm which can be implemented with a $\mathcal{O}(N)$ computational cost per time step as proposed by Douc et al. (2010) when approximating the whole joint smoothing distribution. If we are interested only in approximations of the marginal smoothing distributions, the *Two-Filter smoother* of Briers et al. (2010) may also be used as an alternative method. This algorithm originally suffers from a quadratic computational cost but has recently been modified in Fearnhead et al. (2010) to get a linear one.

Whereas more and more SMC-based smoothing algorithms are linear in the number of particles, there is a recent surge of interest in mixed strategies (see Andrieu et al. (2010); Olsson and Rydén (2010) or Chopin et al. (2011)) where nice properties of SMC and MCMC algorithms are conjugated to produce better approximations. Whereas these methods are developed mostly in the framework of Bayesian inference for state space models, we focus here on the quality of the approximation of the smoothing distribution associated to a fixed Hidden Markov model. This is a crucial problem to address and the hope is to exhibit the key factors that affects the quality of the estimation. More precisely, fix (once and for all) a set of observations $Y_0, \ldots, Y_T$ and try to approximate the law of $X_0, \ldots, X_T$ conditionally on the observations with a set of particles $(\xi_0^{i,N}, \ldots, \xi_T^{i,N})_{i=1}^N$ associated to equal or unequal weights $(\omega_T^{i,N})_{i=1}^N$. For a fixed CPU time, how to build the best population of particles? Should we use mixed strategies? Can we obtain confidence intervals without additional Monte Carlo passes? These are some of the questions we consider in this work. Since T is fixed, the context of this work does not exactly correspond to the one of Gilks and Berzuini (2001) who propose to sequentially alternate SMC stages and MCMC stages as more and more observations are available. Nevertheless, the MCMC step called the Move stage by these authors is now included in the method proposed in this paper to form an efficient algorithm where some directional update of the components extends sequentially the diversity of the population from high values of t to lower values of t. Despite its simplicity, the resulting algorithm turns out to be more than a strong competitor to existing smoothing samplers.

We propose here to improve any consistent particle approximation of the joint smoothing distribution by moving sequentially the particles according to a Metropolis-within-Gibbs

iteration. Such algorithm has a linear computational cost and can be applied in particular to the Filter-Smoother to reduce the degeneracy without increasing the complexity. The paper is organized as follows: in Section 2, we describe the algorithm. In Section 3, we show that the limiting variance of the algorithm is reduced in comparison with the original SMC-based population with a multinomial resampling stage. One major characteristic of this algorithm is the fact that, by letting the number of iterations of the Markov chains proportional to $\ln N$, the asymptotic variance is close to optimal and can be estimated using the evolution of only one population of particle paths. Up to our knowledge, this feature is totally new in the smoothing literature. Numerical experiments and comparisons with existing linear smoothers are provided in Section 4 for the Linear Gaussian Model (LGM) and the Stochastic Volatility Model (StoVolM).

2. MH-Improvement of a particle path population

Denote for $u \leq s$, $a_{u:s} = (a_u, a_{u+1}, \ldots, a_s)$ and define the smoothing distribution $\Pi_{0:T|T}$ associated to a fixed set of observations $Y_{0:T} = y_{0:T}$ by: for any $\mathsf{A} \in \mathcal{X}^{\otimes(T+1)}$,

$$\Pi_{0:T|T}(\mathsf{A}) \stackrel{\text{def}}{=} \frac{\int \cdots \int \chi(\mathrm{d}x_0) g(x_0, y_0) \left[\prod_{i=1}^{T} m(x_{i-1}, x_i) g(x_i, y_i)\right] \mathbf{1}_{\mathsf{A}}(x_{0:T}) \mathrm{d}x_{1:T}}{\int \cdots \int \chi(\mathrm{d}x_0) g(x_0, y_0) \left[\prod_{i=1}^{T} m(x_{i-1}, x_i) g(x_i, y_i)\right] \mathrm{d}x_{1:T}} ,$$

where χ is a probability measure on $(\mathbb{X}, \mathcal{X})$. The distribution $\Pi_{0:T|T}$ is thus the law of $X_{0:T}$ conditionally to $Y_{0:T} = y_{0:T}$ when X_0 follows the distribution χ. In the sequel, χ is assumed to have a density w.r.t. $\lambda(\mathrm{d}x)$, density which will be denoted by χ by abuse of notation: $\chi(\mathrm{d}x) = \chi(x)\lambda(\mathrm{d}x)$. Then, the density $\pi_{0:T|T}$ of the distribution $\Pi_{0:T|T}$ with respect to $\prod_{t=0}^{T} \lambda(\mathrm{d}x_t)$ writes

$$\pi_{0:T|T}(x_{0:T}) \propto \chi(x_0) g(x_0, y_0) \left[\prod_{i=1}^{T} m(x_{i-1}, x_i) g(x_i, y_i)\right] . \tag{2}$$

As noted in Gilks and Berzuini (2001), the smoothing density $\pi_{0:T|T}$ in (2) is known up to a normalizing constant so that approximation of this distribution can be perfectly cast into the general framework of the Metropolis-Hastings algorithm. Given that the resulting Markov chain evolves in the path space $\mathbb{X}^{T+1}$, the candidate at each iteration should be carefully chosen to keep the acceptance rate away from zero which is a delicate task in high dimensional spaces. Considering this, an appealing approach in the MCMC literature is the Gibbs sampler and more generally the Metropolis-within-Gibbs sampler which proposes to update only one component at a time. One could also choose to update components by blocks but as will be seen in Section 4, moving only one component at a time is sufficient for our purpose. A key point for exploring the posterior distribution within a reasonable number of iterations is that the algorithm should be well initialized at least for the first components to be updated. We propose here to achieve this by exploiting approximation of $\Pi_{0:T|T}$ provided by SMC-based algorithms.

More precisely, suppose that we already have an approximation of $\Pi_{0:T|T}$ through a set of (normalized) weighted particle paths, $(\xi_{0:T}^{i,N}, \omega_{0:T}^{i,N})_{i=1}^{N}$ in the sense that

$$\Pi_{0:T|T}(h) \approx \sum_{i=1}^{N} \omega_{0:T}^{i,N} h(\xi_{0:T}^{i,N}) , \quad \sum_{i=1}^{N} \omega_{0:T}^{i,N} = 1 , \tag{3}$$

We intend here to improve this approximation by running N independent Metropolis-within-Gibbs Markov chains $(\xi_{0:T}^{i,N}[k], k \geq 0)$ for $i \in \{1, \ldots, N\}$ starting from each path $\xi_{0:T}^{i,N}$, that is, we set $\xi_{0:T}^{i,N}[0] = \xi_{0:T}^{i,N}$ for $i \in \{1, \ldots, N\}$. The resulting approximation after K iterations of the Markov chains then writes

$$\Pi_{0:T|T}(h) \approx \sum_{i=1}^{N} \omega_{0:T}^{i,N} h(\xi_{0:T}^{i,N}[K]) \; . \tag{4}$$

Let us now detail the transition of $(\xi_{0:T}^{i,N}[k], \; k \geq 0)$. For a simpler exposition, we drop here the dependence on i, N. Now, consider a family of transition kernel densities $(r_t)_{0 \leq t \leq T}$ such that r_0, r_T are transition kernel densities on $(\mathbb{X}, \mathcal{X})$ whereas for $t \in \{1, \ldots, T-1\}$, r_t is a transition kernel density on $(\mathbb{X} \times \mathbb{X}, \mathcal{X})$. For $u, v, w, x \in \mathbb{X}$, set

$$\alpha_0(v, w; x) \stackrel{\text{def}}{=} \frac{\chi(x) g(x, y_0) m(x, w)}{\chi(v) g(v, y_0) m(v, w)} \frac{r_0(w; v)}{r_0(w; x)} \wedge 1 \; , \tag{5}$$

$$\alpha_t(u, v, w; x) \stackrel{\text{def}}{=} \frac{m(u, x) g(x, y_t) m(x, w)}{m(u, v) g(v, y_t) m(v, w)} \frac{r_t(u, w; v)}{r_t(u, w; x)} \wedge 1 \; , \quad 1 \leq t \leq T-1 \; , \tag{6}$$

$$\alpha_T(u, v; x) \stackrel{\text{def}}{=} \frac{m(u, x) g(x, y_T)}{m(u, v) g(v, y_t)} \frac{r_T(u; v)}{r_T(u; x)} \wedge 1 \; . \tag{7}$$

At time k, the new path $\xi_{0:T}[k]$ is obtained by updating backward in time each component $\xi_t[k]$ as follows

(i) Sample a candidate $X \sim r_t(\xi_{t-1}[k-1], \xi_{t+1}[k], \cdot)$,

(ii) Accept $\xi_t[k] = X$ with probability $\alpha_t(\xi_{t-1:t}[k-1], \xi_{t+1}[k]; X)$,

(iii) Otherwise, set $\xi_t[k] = \xi_t[k-1]$.

This procedure is valid for $t \in \{1, \ldots, T-1\}$; we skip the description of the updates for $\xi_0[k]$ and $\xi_T[k]$ since they follow the same lines under very slight modifications. The complete pseudo-code version of the Metropolis-Hastings Improved Particle Smoother (MH-IPS) is given below.

Straightforwardly, for any $t \in \{0, \ldots, T\}$, α_t is the classical Metropolis-Hastings acceptance rate associated to the proposal kernel r_t and the target distribution $\Pi_{0:T|T}$. Due to the specific structure of $\Pi_{0:T|T}$ whose density is a product of quantities involving consecutive components, the acceptance ratios in (5), (6) and (7) do not depend on the path space dimension and are therefore nondegenerated. Of course, it is also possible to update each component from an arbitrary number of neighbors. Nevertheless, in the Gibbs Sampler for which all the acceptance rates are equal to one, the t-th component is updated according to the distribution of X_t conditionally on $X_{0:t-1}, X_{t+1:T}, Y_{0:T}$ which only depends on X_{t-1}, X_{t+1}, Y_t. Such dependence suggests that the candidate in the Metropolis-within-Gibbs algorithm should be proposed according to a distribution which only involves its nearest neighbors.

MH-IPS is based on a first approximation of $\Pi_{0:T|T}$ given in (3) whereas some SMC algorithms like the Filter-Smoother are known to suffer from a poor representation of the states close to 0 but are accurate for states close to T. As a consequence, $(\xi_t^{i,N})_{i=1}^N$ for large values of t are well-distributed and this set of particles is then propagated to the poorer ones by updating the components backward in time. In other words, instead of a random-scan

Algorithm 1 MH-IPS

Initialization
Run an SMC-algorithm targeting $\Pi_{0:T|T}$ and store $(\xi_{0:T}^{i,N}, \omega_{0:T}^{i,N})_{i=1}^{N}$.
Set: $\forall\ 1 \leq i \leq N,\ \xi_{0:T}^{i,N}[0] = \xi_{0:T}^{i,N}$.
K improvement passes
for k from 1 to K **do**
  for i from 1 to N **do**
    Sample $X \sim r_T(\xi_{T-1}^{i,N}[k-1]; \cdot)$,
    Accept $\xi_T^{i,N}[k] = X$ with probability $\alpha_T(\xi_{T-1:T}^{i,N}[k-1], X)$,
    Otherwise, set $\xi_T^{i,N}[k] = \xi_T^{i,N}[k-1]$.
    for t from $T-1$ down to 1 **do**
      Sample $X \sim r_t(\xi_{t-1}^{i,N}[k-1], \xi_{t+1}^{i,N}[k]; \cdot)$,
      Accept $\xi_t^{i,N}[k] = X$ with probability $\alpha_t(\xi_{t-1:t}^{i,N}[k-1], \xi_{t+1}^{i,N}[k], X)$,
      Otherwise, set $\xi_t^{i,N}[k] = \xi_t^{i,N}[k-1]$.
    end for
    Sample $X \sim r_0(\xi_1^{i,N}[k]; \cdot)$,
    Accept $\xi_0^{i,N}[k] = X$ with probability $\alpha_0(\xi_0^{i,N}[k-1], \xi_1^{i,N}[k], X)$,
    Otherwise, set $\xi_0^{i,N}[k] = \xi_0^{i,N}[k-1]$.
  end for
end for

procedure where components are updated at random, this determistic-scan Metropolis-Hastings algorithm extends the diversity of the particle paths to the lower values of t at each backward pass. The fact that MH-IPS uses the SMC-based approximation just once and then, keep the N Metropolis-within-Gibbs Markov chains independent from each other implies that the path degeneracy vanishes as the number of iterations increases. Strong empirical evidences of this phenomenon are provided in Section 4.

A last but striking particularity of MH-IPS when compared to classical MH algorithms is the fact that the approximation (4) only involves the states at iteration K of the N Markov chains instead of using all the history of these Markov chains. Indeed, since only one component is updated at a time, the consecutive paths are highly positively correlated so that including them into (4) is detrimental to the quality of the approximation. Another advantage of considering only states at iteration K is that the CLT of the approximation (4) which is quite easy to establish when $K \propto \ln N$ includes a very simple and close-to-optimal expression of the asymptotic variance. The estimation of this variance can be performed using the evolution of only one population of sample paths. Therefore, on the contrary to all the smoothing algorithms proposed in the literature so far, confidence intervals can be obtained without additional Monte Carlo passes.

3. Properties of the algorithm

In this section, since the number of observations is fixed, T is dropped for simplicity from the notation. For example, we set $\Pi = \Pi_{0:T|T}$, $\boldsymbol{\xi}^{i,N} = \xi_{0:T|T}^{i,N}$, $\omega^{i,N} = \omega_{0:T|T}^{i,N}$ and so on.

The general procedure induced by MH-IPS can be described as follows. Let Q be a Markov transition kernel on $(\mathbb{X}^{T+1}, \mathcal{X}^{\otimes(T+1)})$ with invariant distribution Π. Consider a

set of normalized weighted particles $(\boldsymbol{\xi}^{i,N}, \omega^{i,N})_{i=1}^N$ and move the particles *independently* according to the kernel Q. To be specific, define N *independent* Markov chains $(\boldsymbol{\xi}^{i,N}[k], k \geq 0)_{i=1}^N$ such that:

$$\boldsymbol{\xi}^{i,N}[0] = \boldsymbol{\xi}^{i,N} \ , \tag{8}$$

$$\boldsymbol{\xi}^{i,N}[k+1] \sim Q(\boldsymbol{\xi}^{i,N}[k], \cdot) \ , \quad k \geq 0 \ . \tag{9}$$

According to (4), Πh is approximated after k iterations of the Markov chains by:

$$\Pi h \approx \sum_{i=1}^N \omega^{i,N} h(\boldsymbol{\xi}^{i,N}[k]), \quad \sum_{i=1}^N \omega^{i,N} = 1 \ . \tag{10}$$

3.1. A resampling step in the initialization

Let us first consider the impact of the weights on the quality of the approximation. A resampling step in the initialization consists in replacing the weighted particles $(\boldsymbol{\xi}^{i,N}, \omega^{i,N})_{i=1}^N$ by the unweighted particles $(\tilde{\boldsymbol{\xi}}^{i,N}, 1/N)_{i=1}^N$ such that some unbiasedness condition is fulfilled. Whereas many resampling strategies have been developed in the literature (Liu and Chen (1998), Kitagawa (1998), Carpenter et al. (1999); see also Douc et al. (2005) for a brief review of their different properties), we only focus here on the most simple one, the multinomial resampling:

(i) $(\tilde{\boldsymbol{\xi}}^{j,N})_{j=1}^N$ are independent conditionally on $(\boldsymbol{\xi}^{i,N}, \omega^{i,N})_{i=1}^N$,

(ii) for all $i, j \in \{1, \ldots, N\}$, $\mathbb{P}\left[\tilde{\boldsymbol{\xi}}^{j,N} = \boldsymbol{\xi}^{i,N}\right] = \omega^{i,N}$.

A straightforward calculation yields:

$$\mathbb{V}\mathrm{ar}\left(\sum_{i=1}^N \omega^{i,N} h(\boldsymbol{\xi}^{i,N})\right) \leq \mathbb{V}\mathrm{ar}\left(\sum_{i=1}^N h(\tilde{\boldsymbol{\xi}}^{i,N})/N\right) \ ,$$

showing that at time 0, the particle system with equal weights is less efficient than the one with original weights. Despite this, the resampling stage discards particles with small weights and duplicates "informative" particles (with high weights). As in the particle filtering theory, our hope is that the resampling stage increases the number of Markov chains starting from interesting regions with respect to the target distribution.

Denote by $\|\cdot\|_{\mathrm{TV}}$ the total variation norm: $\|\mu\|_{\mathrm{TV}} \stackrel{\mathrm{def}}{=} \sup_{|f|_\infty \leq 1} |\mu(f)|$ where $|f|_\infty \stackrel{\mathrm{def}}{=} \sup_{x \in \mathbb{X}} |f(x)|$ and assume that

(**A1**) For any $x \in \mathbb{X}^{T+1}$, $\lim_{k \to \infty} \left\|Q^k(x, \cdot) - \Pi\right\|_{\mathrm{TV}} = 0$.

Under this assumption, it is straightforward that for any bounded measurable function h, $\sum_{i=1}^N \omega^{i,N} h(\boldsymbol{\xi}^{i,N}[k])$ is asymptotically unbiased whatever the weights are, provided their sum is equal to one. To go further, consider the effect of the weights on the second order approximation. The following proposition shows that as the iterations of the Markov chains goes to infinity, the quadratic error tends to a limit which is minimal when all the weights are equal to $1/N$. This advocates for a particle system with equal weights in the initialization as provided by a resampling step before letting evolve the N Markov chains.

PROPOSITION 1. *Assume (**A1**). Then, for any bounded measurable function h,*

$$\lim_{k\to\infty} \mathbb{E}\left[\left(\sum_{i=1}^{N} \omega^{i,N} h(\boldsymbol{\xi}^{i,N}[k]) - \Pi h\right)^2\right] = \mathbb{V}\mathrm{ar}_{\Pi}(h)\ \mathbb{E}\left[\sum_{i=1}^{N} \left(\omega^{i,N}\right)^2\right]$$

where $\mathbb{V}\mathrm{ar}_{\Pi}(h) = \Pi h^2 - (\Pi h)^2$. *Moreover, the previous limit is minimized when all the weights are equal:* $\omega^{i,N} = 1/N$ *for all* $i \in \{1,\ldots,N\}$.

PROOF. Proof is given the Appendix. □

As a consequence of this proposition, it is assumed in the sequel that the *multinomial resampling stage* has been performed in the initialization, i.e. (8), (9) and (10) are replaced by

$$\tilde{\boldsymbol{\xi}}^{i,N}[0] = \tilde{\boldsymbol{\xi}}^{i,N}\ , \tag{11}$$

$$\tilde{\boldsymbol{\xi}}^{i,N}[k+1] \sim Q(\tilde{\boldsymbol{\xi}}^{i,N}[k], \cdot)\ , \quad k \geq 0\ , \tag{12}$$

$$\Pi h \approx \sum_{i=1}^{N} h(\tilde{\boldsymbol{\xi}}^{i,N}[k])/N\ , \tag{13}$$

Then, according to Proposition 1,

$$\lim_{k\to\infty} \mathbb{E}\left[\left(\sum_{i=1}^{N} h(\tilde{\boldsymbol{\xi}}^{i,N}[k])/N - \Pi h\right)^2\right] = \mathbb{V}\mathrm{ar}_{\Pi}(h)/N\ . \tag{14}$$

Thus, when N is fixed and k goes to infinity, (14) shows that the approximation cannot be better than having N independent draws from the distribution Π. A natural question is now to properly tune the number of iterations k of the Markov chains to the number N of initial points so that the unweighted particles $(\boldsymbol{\xi}^{i,N}[k], 1/N)_{i=1}^{N}$ have properties close to iid draws according to Π *without* letting k go to infinity. Before treating this question, let us examine some non-asymptotic result with respect to the approximation.

3.2. Deviation Inequality

Noting that $(\tilde{\boldsymbol{\xi}}^{i,N}[k])_{i=1}^{N}$ are i.i.d conditionally to $\tilde{\mathcal{F}}_0^N \stackrel{\mathrm{def}}{=} \sigma\left\{\tilde{\boldsymbol{\xi}}^{i,N}, i \in \{1,\ldots,N\}\right\}$ and that $\mathbb{E}\left[h(\tilde{\boldsymbol{\xi}}^{i,N}[k])|\tilde{\mathcal{F}}_0^N\right] = Q^k h(\tilde{\boldsymbol{\xi}}^{i,N})$, the conditional Hoeffding inequality directly yields:

PROPOSITION 2. *For any bounded measurable function h, any $k \in \mathbb{N}$ and any $\epsilon > 0$,*

$$\mathbb{P}\left[\left|\sum_{i=1}^{N} h(\tilde{\boldsymbol{\xi}}^{i,N}[k])/N - \Pi h\right| > \epsilon\right] \leq 2\exp\left(-\frac{N\epsilon^2}{2\left(\mathrm{osc}(h)\right)^2}\right) + \mathbb{P}\left[\left|\sum_{i=1}^{N} Q^k h(\tilde{\boldsymbol{\xi}}^{i,N})/N - \Pi h\right| > \epsilon/2\right]\ , \tag{15}$$

where $\mathrm{osc}(h) = \sup_{u,v\in\mathbb{X}} |h(u) - h(v)|$.

Nevertheless, when reading the inequality in Proposition 2, the question of knowing whether MH-IPS improves or does not improve the approximation is far from being obvious. We now answer this question in terms of the Central Limit Theorem.

3.3. Central limit theorem

MH-IPS is based on a first approximation of Πh by a family of normalized weighted particles $(\boldsymbol{\xi}^{i,N}, \omega^{i,N})_{i=1}^N$. For various versions of SMC methods, the asymptotic normality of $(\boldsymbol{\xi}^{i,N}, \omega^{i,N})_{i=1}^N$ have already been obtained under different techniques (see for example Del Moral and Guionnet (1999), Künsch (2000), Chopin (2004) or Douc and Moulines (2008)). The following proposition now focus on the effect of the multinomial resampling on the central limit theorem: whatever SMC method is chosen, if $(\boldsymbol{\xi}^{i,N}, \omega^{i,N})_{i=1}^N$ are asymptotically normal, then $(\tilde{\boldsymbol{\xi}}^{i,N}, 1/N)_{i=1}^N$ are also asymptotically normal with $\mathbb{V}\mathrm{ar}_\Pi(h)$ as an *additional* term in the variance.

PROPOSITION 3. *Assume that $(\boldsymbol{\xi}^{i,N}, \omega^{i,N})_{i=1}^N$ are asymptotically normal, in the sense that for any bounded measurable function h, there exists $0 < \sigma^2(h) < \infty$ such that*

$$N^{1/2}\left[\sum_{i=1}^N \omega^{i,N} h(\boldsymbol{\xi}^{i,N}) - \Pi h\right] \xrightarrow{\mathcal{D}} \mathcal{N}(0, \sigma^2(h)) \; .$$

Then, for any bounded measurable function h,

$$N^{1/2}\left[\sum_{i=1}^N h(\tilde{\boldsymbol{\xi}}^{i,N})/N - \Pi h\right] \xrightarrow{\mathcal{D}} \mathcal{N}(0, \mathbb{V}\mathrm{ar}_\Pi(h) + \sigma^2(h)) \; .$$

The proof follows closely the lines of (Chopin, 2004, Theorem 1) or (Douc and Moulines, 2008, Theorem 4) and is omitted for the sake of brevity.

Proposition 3 shows the asymptotic normality of $(\tilde{\boldsymbol{\xi}}^{i,N}[k], 1/N)_{i=1}^N$ for $k = 0$. The Markov chains are then run independently according to the transition kernel Q and we now consider the impact on the approximation given in (13) for $k = k_N$. To be specific, the following theorem shows that under the assumption that the kernel Q is V-geometrically ergodic, for $k_N \propto \ln N$, the unweighted particles $(\tilde{\boldsymbol{\xi}}^{i,N}[k_N], 1/N)_{i=1}^N$ are asymptotically normal with a reduced asymptotic variance. Define the following set of assumptions:

(**A2**) There exists a measurable function $V : \mathbb{X}^{T+1} \to [1, \infty)$ such that

(i) $\Pi V < \infty$ and for any $x \in \mathbb{X}$ and any $k \in \mathbb{N}$, $Q^k V(x) < \infty$,

(ii) there exists $\beta \in (0,1)$ such that for any $h \in \mathcal{C}_V \stackrel{\text{def}}{=} \{h; |h/V|_\infty < \infty\}$ and any $x \in \mathbb{X}$,

$$|Q^k h(x) - \Pi h| \leq \beta^k V(x) \; ,$$

(iii) the sequence $\{N^{-1}\sum_{i=1}^N V^2(\tilde{\boldsymbol{\xi}}^{i,N})\}_{N \geq 1}$ of random variables is bounded in probability.

(**A2**)-(i) ensures that the quantities appearing in (**A2**)-(ii) are well defined. (**A2**)-(ii) shows that Q is V-geometrically ergodic. (**A2**)-(iii) is a weak assumption concerning the initial unweighted particles $(\tilde{\boldsymbol{\xi}}^{i,N}, 1/N)_{i=1}^N$. If for example, $(\tilde{\boldsymbol{\xi}}^{i,N}, 1/N)_{i=1}^N$ is consistent with respect to the function V^2 in the sense that $\sum_{i=1}^N V^2(\tilde{\boldsymbol{\xi}}^{i,N})/N$ converges in probability to ΠV^2, then (**A2**)-(iii) holds. Condition under which such convergence results hold for possibly unbounded functions may be found for example in Douc and Moulines (2008).

THEOREM 1. *Assume (**A2**). Let $(k_N)_{N\geq 0}$ be a sequence of integers such that*

$$\lim_{N\to\infty} k_N + \ln N/(2\ln\beta) = \infty \ . \tag{16}$$

Then, for any h such that $h^2 \in \mathcal{C}_V$, the following central limit theorem holds:

$$N^{-1/2}\sum_{i=1}^{N}\left[h(\tilde{\boldsymbol{\xi}}^{i,N}[k_N]) - \Pi h\right] \xrightarrow{\mathcal{D}} \mathcal{N}(0, \mathbb{V}\mathrm{ar}_{\Pi}(h)) \ .$$

PROOF. Proof is given in the Appendix. □

Theorem 1 and Proposition 3 show that k_N iterations of the Markov chains reduce the asymptotic variance when compared to a sample obtained by multinomial resampling of a population issued from any SMC method. The asymptotic variance $\mathbb{V}\mathrm{ar}_{\Pi}(h)$ in Theorem 1 is close to optimal since it is the same as for i.i.d. draws with distribution Π. Moreover, the expression of $\sigma^2(h)$ in Proposition 3 is usually quite involved and for obtaining confidence intervals, the estimation of the asymptotic variance in Proposition 3 is classically obtained by adding some Monte Carlo passes. This is not at all the case in Theorem 1 since estimation of $\mathbb{V}\mathrm{ar}_{\Pi}(h)$ can be performed directly via $(\tilde{\boldsymbol{\xi}}^{i,N}[k_N], 1/N)_{i=1}^{N}$. Finally, by adding typically $k_N = -\ln N/\ln\beta$ iterations of a transition kernel to a SMC-based population of particles, we obtain a sample with a reduced and close-to-optimal variance which can be easily approximated without additional simulations.

The fact that the CLT holds for $k_N \propto \ln N$ suggests that a good approximation of the target distribution may be achieved with only a few number of iterations of the parallel Markov chains. This will be confirmed empirically in the next section.

4. Experiments

The *Filter-smoother* is known to be quite easy to implement and efficient in terms of CPU time, but suffers dramatically from the degeneracy of the ancestors. We now see how only a few iterations of MH-IPS reduce the degeneracy and turn the *Filter-smoother* to a strong competitor to the existing smoother algorithms. In the sequel, denote by the *Metropolis-Hastings Improved Filter-Smoother* (MH-IFS), Algorithm 1 initialized with the Filter-Smoother. The performance of this algorithm is now compared to the other linear-in-N particle smoothers (Filter-Smoother, FFBSi, Two-Filter). In order to be as computationally fair as possible, all these algorithms are implemented in the same way as their common base, the Forward-Filter.

4.1. Linear Gaussian Model

We first consider the LGM defined by:

$$X_{t+1} = \phi X_t + \sigma_u U_t \ , \quad Y_t = X_t + \sigma_v V_t \ ,$$

where $X_0 \sim \mathcal{N}\left(0, \frac{\sigma_u^2}{1-\phi^2}\right)$, $\{U_t\}_{t\geq 1}$ and $\{V_t\}_{t\geq 1}$ are independent sequences of i.i.d. standard gaussian random variables (independent of X_1). $T+1 = 101$ observations were generated using the model with $\phi = 0.9$, $\sigma_u = 0.6$ and $\sigma_v = 1$. Furthermore, in this model, the fully-adapted filters are explicitly computable when needed and the Gibbs sampler may be implemented.

The diversity of the particle population at each time step for each algorithm is measured by an estimate of the *effective sample size* $N_{\text{eff}}^{\text{algo}}(t)$ as defined in Fearnhead et al. (2010). Motivated by the fact that $\mathbb{E}\left[\left(\bar{X}_N - \mu\right)^2/\sigma^2\right] = 1/N$, when $X^{(1)}, \ldots, X^{(N)}$ are i.i.d. with $\mathbb{E}[X^{(1)}] = \mu$, $\mathbb{V}\text{ar}(X^{(1)}) = \sigma^2$ and $\bar{X}_N$ is their sample mean, we set

$$N_{\text{eff}}^{\text{algo}}(t) \stackrel{\text{def}}{=} \mathbb{E}\left[\left(\frac{\pi_{t|T}^{\text{algo},N}(\text{Id}) - \mu_t}{\sigma_t}\right)^2\right]^{-1}, \tag{17}$$

where Id is the identity function on $\mathbb{R}$, μ_t and σ_t^2 are the exact mean and variance of X_t conditionally to $Y_{0:T}$ obtained from the Kalman smoother. In some sense, the weighted sample produced by a given algorithm is as accurate at estimating X_t as an "independent" sample of size $N_{\text{eff}}^{\text{algo}}(t)$. The expression of $N_{\text{eff}}^{\text{algo}}(t)$ given in (17) shows that it is inversely proportional to the quadratic error associated to a normalized estimator of $\mathbb{E}(X_t|Y_{0:T})$. To estimate the expectation in (17) we use the mean value from 250 repetitions of each algorithm with a number of particles chosen such that the computation time of each of them is the same.

Figure 1.a shows that when the number of improvements increases, the degeneracy of the particle population for small values of t decreases and for $K = 8$ all the time steps have the same diversity.

Figure 1.b displays the effective sample size of the four linear smoothing algorithms. As expected, the Filter-Smoother is highly degenerated for small values of t as opposed to the other algorithms. Furthermore, the MH-IFS clearly outperforms all others within a fixed computational time. In order to check that this efficiency is not due to the fact that the LGM allows to easily implement the Gibbs sampler, we now turn to a model where a rejection sampling is required.

4.2. Stochastic Volatility Model

StoVolM have been introduced in financial time series modeling to capture more realistic features than ARCH/GARCH models (Hull and White (1987)). Despite its apparent simplicity, the following equations do not allow to directly simulate according to $r_t(u, w; \cdot) \propto m(u, \cdot)g(\cdot, y_t)m(\cdot, w)$:

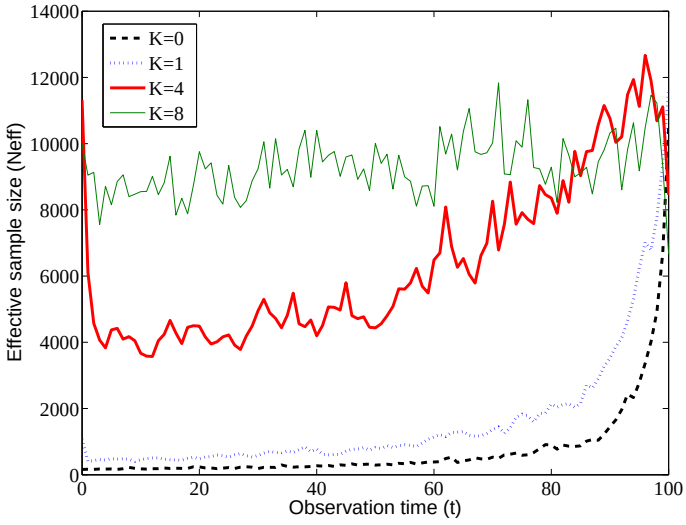
$$X_{t+1} = \alpha X_t + \sigma U_{t+1}\,, \quad Y_t = \beta \mathrm{e}^{\frac{X_t}{2}} V_t\,,$$

where $X_0 \sim \mathcal{N}\left(0, \frac{\sigma^2}{1-\alpha^2}\right)$, U_t and V_t are independent standard gaussian random variables. $T + 1 = 101$ observations were generated using the model with $\alpha = 0.3$, $\sigma = 0.5$ and $\beta = 1$ in order to estimate the effective sample size defined in (17). The true values of μ_t and σ_t cannot be computed explicitly so they are estimated by running the MH-IFS with $N = 650000$.

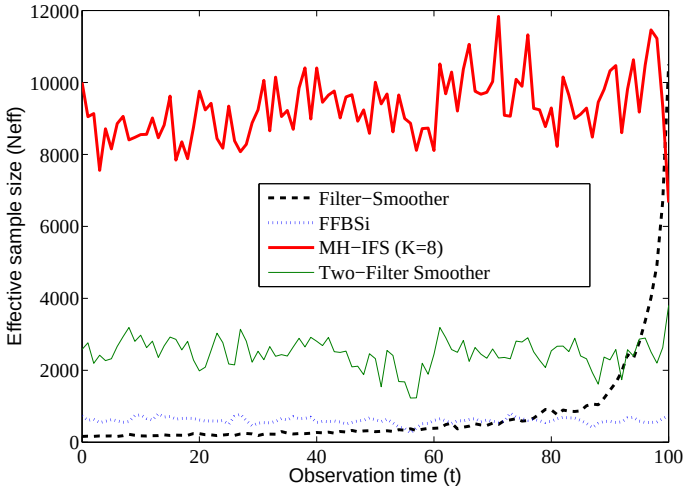
4.2.1. Gibbs sampler

In the StoVolM, the Gibbs sampler requires to sample exactly from

$$r_t(u, w; x) \propto \exp\left\{-\frac{\mathrm{e}^{-x}}{2\beta^2}y_t^2 - \frac{1+\alpha^2}{2\sigma^2}\left[x - \left(\frac{\alpha}{1+\alpha^2}(u+w) - \frac{\sigma^2/2}{1+\alpha^2}\right)\right]^2\right\}, \tag{18}$$



(a) Influence of the number of improvements K



(b) Comparison of four linear smoothing algorithms

Figure 1: Average effective sample size for each of the 100 time steps of the LGM using different smoothing algorithms for a fixed CPU time.

for $1 \leq t \leq T-1$ (the cases $t = 0$ and $t = T$ are dealt with in a similar way) which does not correspond to a classical distribution. However, we propose here to implement a rejection sampling. The first idea is to sample the proposal candidate $X = x$ according to the *a priori* distribution of X_t conditionally to $X_{t-1} = u$ and $X_{t+1} = w$. The corresponding ratio of acceptance is then given by $(|y_t|/\beta)\exp\left\{-(x-1)/2 - \mathrm{e}^{-x}y_t^2/(2\beta^2)\right\}$ and will obviously lead

to poor results for small values of y_t. To counterbalance the effect of y_t in the acceptance rate, the proposal distribution should also take the value of y_t into account; we then rewrite (18) for any $\gamma_t \geq 0$ (possibly depending on y_t):

$$r_t(u,w;x) \propto \mathrm{e}^{-\frac{\gamma_t}{2}x - \frac{\mathrm{e}^{-x}}{2\beta^2}y_t^2} \times \exp\left\{-\frac{1+\alpha^2}{2\sigma^2}\left[x - \left(\frac{\alpha}{1+\alpha^2}(u+w) - \frac{\sigma^2/2}{1+\alpha^2}(1-\gamma_t)\right)\right]^2\right\}, \tag{19}$$

which suggests to propose x according to $\mathcal{N}\left(\frac{\alpha}{1+\alpha^2}(u+w) - \frac{\sigma^2/2}{1+\alpha^2}(1-\gamma_t), \frac{\sigma^2}{1+\alpha^2}\right)$ and to accept it with a probability given by:

$$\left(\frac{|y_t|}{\gamma_t^{1/2}\beta}\right)^{\gamma_t} \exp\left\{-\frac{\gamma_t}{2}(x-1) - \frac{\mathrm{e}^{-x}}{2\beta^2}y_t^2\right\}. \tag{20}$$

An optimal choice for γ_t would consist in maximizing the smoothed expectation of (20) but this quantity is intractable. An intuitive choice for γ_t is then:

$$\gamma_t = \begin{cases} (|y_t|/\beta)^2, & \text{if } |y_t| \leq \beta, \\ |y_t|/\beta, & \text{if } |y_t| > \beta. \end{cases} \tag{21}$$

Indeed, for small values of y_t, (20) is then close to one and for bigger values, the exponential becomes very small but the first term remains non-neglectable.

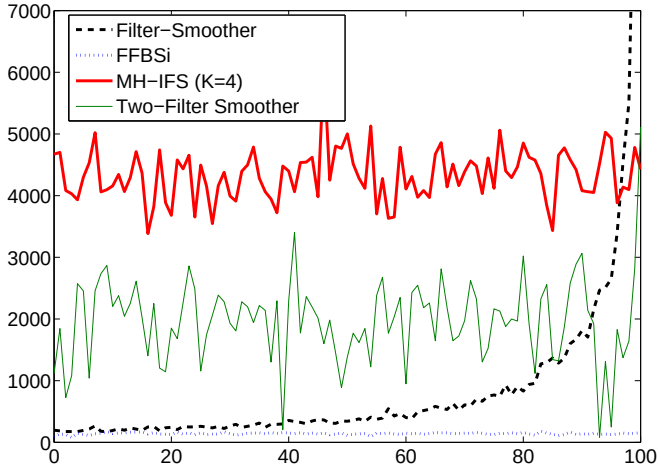


Figure 2: Average effective sample size for each of the 100 time steps of the StoVolM using different smoothing algorithms for a fixed CPU time.

The Improved Filter-Smoother used to generate Figure 2 performs simulations using the Gibbs sampler with the previous rejection sampling. We can see that this algorithm still leads to better results than the other ones within an equivalent computational time.

In many instances (for example Expectation-Maximization algorithm, score computation), it is necessary to estimate smoothed additive functionals such as $\Pi_{0:T|T}(H)$ where

for all $x_{0:T} \in \mathbb{X}^{T+1}$, $H(x_{0:T}) = \sum_{t=0}^{T} x_t$. In order to assess the smoothing algorithms on this matter, $T+1 = 1001$ observations were generated. As seen before, the computational cost of the MH-IFS is linear in N which is verified by numerical experiments in Figure 3.

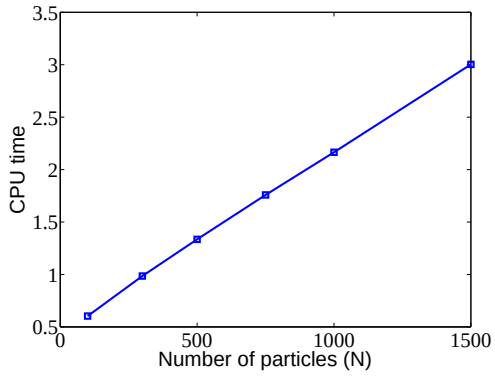


Figure 3: Average CPU time for computing a smoothed additive functional with the MH-IFS as a function of the number of particles.

Figure 4.a shows that the variance vanishes quickly with the number of improvement passes and only 4 iterations of the Markov chains are sufficient to get an efficient estimator. Then, the variances displayed in Figure 4.b allow again to draw the conclusion that for a fixed CPU time, the MH-IFS is more efficient than the Two-Filter. Finally, one improvement pass has been applied to the particle paths given by the FFBSi. The variance reduction is again significant as shown in Figure 4.c.

4.2.2. Metropolis-within-Gibbs and confidence interval

In order to assess Algorithm 1 in the case where the Gibbs sampler could not be implemented, we now turn to the Metropolis-within-Gibbs sampler which is implemented by using again the proposal distribution:

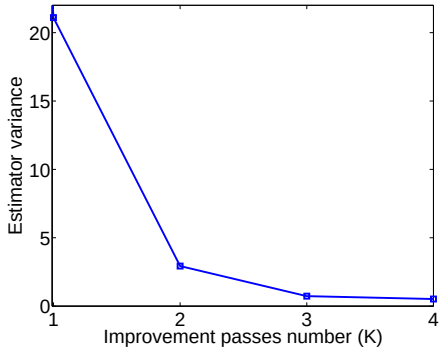
$$r_t(u, w; \cdot) \sim \mathcal{N}\left(\frac{\alpha}{1+\alpha^2}(u+w) - \frac{\sigma^2/2}{1+\alpha^2}(1-\gamma_t), \frac{\sigma^2}{1+\alpha^2}\right) ,$$

where γ_t is defined in (21), and the associated acceptance rate is now given by:

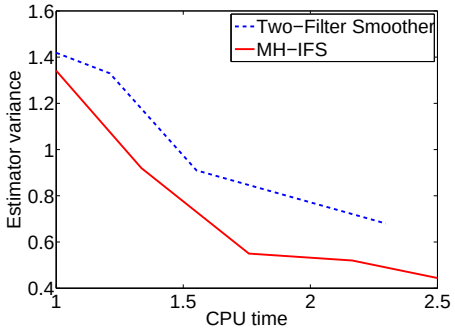
$$\alpha_t(u, v, w; x) = \exp\left\{-\frac{\gamma_t}{2}(x-v) - \frac{\mathrm{e}^{-x} - \mathrm{e}^{-v}}{2\beta^2} y_t^2\right\} \wedge 1 .$$

Figure 5 compares the empirical variance of the Gibbs and Metropolis-within-Gibbs samplers of the smoothed additive functional conditionally to the $T+1 = 1001$ observations used previously. The efficiency of both algorithms is equivalent, showing that Algorithm 1 remains a great performer even when exact *a posteriori* simulation is not possible.

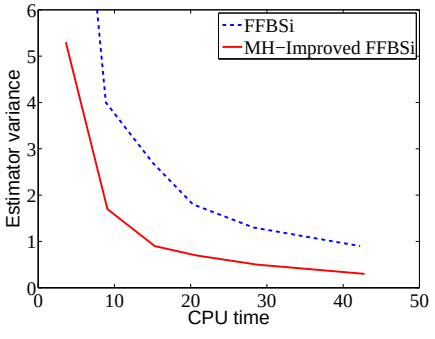
Finally, Theorem 1 is assessed in Figure 6. The empirical variance of the estimator given by Algorithm 1 run with $K_N \propto \ln N$ has been computed over 250 runs using the Gibbs and the Metropolis-within-Gibbs samplers for different number of particles N and compared to the asymptotic variance $\mathbb{V}\mathrm{ar}_{\Pi}(h)/N$ estimated through only one population



(a) Variance of the Improved Filter-Smoother according to the number of improvement passes K



(b) Variance of the Two-Filter Smoother and the Improved Filter-Smoother according to the CPU time



(c) Variance of the FFBSi and its improved version according to the CPU time

Figure 4: Variance of different smoothed additive functional particle estimators in the StoVolM.

of particles. The results show that it is possible in practice to get a confidence interval for the approximation with only one run of Algorithm 1 of complexity $\mathcal{O}(N \ln N)$.

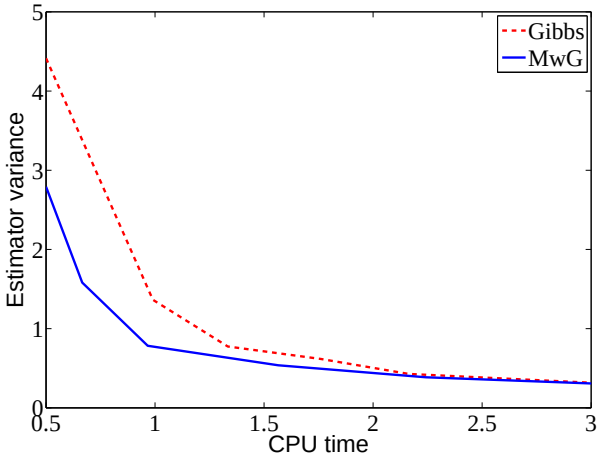


Figure 5: Variance of the Gibbs and Metropolis-within-Gibbs samplers according to the CPU time.

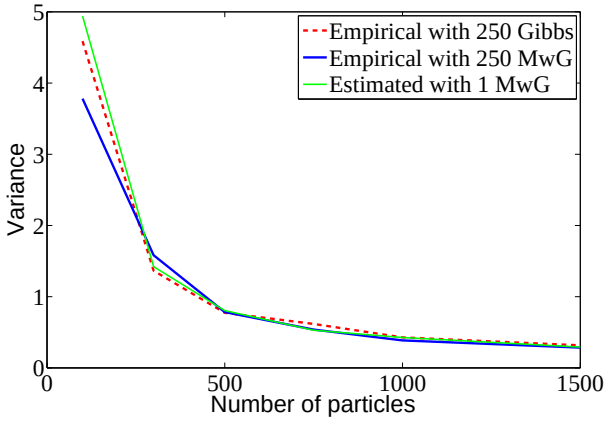


Figure 6: Algorithm 1 variance according to the number of observations.

5. Conclusion

At first sight, one could fear that the MH-IPS is too slow since the updates concern only one component at a time. The various comparisons performed for a *fixed* CPU time in the previous section show that this *is not* the case at all. Roughly speaking, a backward pass in the MH-IPS proposes to sequentially modify each component of the N parallel Markov chains. This can be seen as one run of N particles through $T+1$ observations which is computationally equivalent to one pass of the bootstrap filter. By empirical evidences, we have seen that only a few backward passes ($K = 4$ or 8 in the examples) of the MH-IFS sweep out the degeneracy of the ancestors by extending backward in time the diversity of the particles.

This method is linear in N and outperforms other existing algorithms as the FFBSi or the Two-Filter within a fixed CPU time. These performance results may be explained by

the fact that in the FFBSi algorithm, the points are sampled in the forward pass once and for all; the backward pass in the FFBSi only modifies the weights of the particles without moving them. On the contrary, the MH-IPS allows in the backward pass to move the particles and thus to explore interesting regions of the posterior distribution. In the Two-Filter sampler, two populations (the "forward" population and the "backward" population) evolve *independently*. At time t, a particle is sampled after choosing a couple of particles at time $t-1$ and $t+1$. The two components of these couples belong to independent populations and it is likely that even if their weights are respectively high, associating these independent particles could be detrimental to the approximation. On the contrary, in the MH-IPS, even if the Markov chains are independent, the proposed modification of the component is sampled with respect to its two neighbors which both belong to the *same* Markov chain. Note that we did not compare this algorithm to the Population Monte Carlo by Markov chains (PMCMC) samplers introduced by Andrieu et al. (2010) since the framework here is not the Bayesian inference of parameterized Markov chains.

Another major advantage here is the fact that a CLT can be obtained with a very simple asymptotic variance which can be estimated with only one run of the Algorithm and a complexity in $\mathcal{O}(N \ln N)$. This is totally new in comparison to all the smoothing algorithms proposed in the literature so far, where the asymptotic variances are usually particularly involved. Thus, for a fixed CPU time and only one run, this algorithm is able to produce both approximations of the smoothing distributions and confidence intervals.

Finally, we only focus here on the MH-IFS since it is efficient enough for our purpose. Of course, many other variants with different SMC-based approximations in the initialization step may be performed. In the context of the paper, the MH-IPS only uses the SMC-based approximation once before starting independent MCMC Markov chains. The empirical performances of this algorithm, namely with respect to the diversity of the population and the precision of the approximation, seem to us convincing enough to let the Markov chains evolve independently without trying to interact them again. Of course, as previously noted in Gilks and Berzuini (2001), in some different contexts, where for example, the observations are available sequentially whereas approximations of the smoothing distributions are needed at each time, some variants with SMC steps mixed with MCMC steps can also be elaborated. Nevertheless, in the framework of this paper, the number T of the observations is fixed and we only focus here on how the independent MCMC steps drastically improve the first approximation obtained by SMC algorithms. In this context, there is no need to interact again the Markov chains; this allows to keep the diversity of the population while approximations and confidence intervals are obtained without effort.

Appendix

A. Proof of Proposition 1

For all $k \geq 0$, the bias plus variance decomposition writes

$$\mathbb{E}\left[\left(\sum_{i=1}^{N} \omega^{i,N} h(\boldsymbol{\xi}^{i,N}[k]) - \Pi h\right)^2\right]$$
$$= \left\{\mathbb{E}\left[\sum_{i=1}^{N} \omega^{i,N} h(\boldsymbol{\xi}^{i,N}[k])\right] - \Pi h\right\}^2 + \mathbb{V}\mathrm{ar}\left(\sum_{i=1}^{N} \omega^{i,N} h(\boldsymbol{\xi}^{i,N}[k])\right)$$

$$= \left\{ \mathbb{E}\left[\sum_{i=1}^{N} \omega^{i,N} h(\boldsymbol{\xi}^{i,N}[k]) \right] - \Pi h \right\}^2 + \mathbb{V}\mathrm{ar}\left(\mathbb{E}\left[\sum_{i=1}^{N} \omega^{i,N} h(\boldsymbol{\xi}^{i,N}[k]) \middle| \mathcal{F}_0^N \right] \right) + \mathbb{E}\left[\mathbb{V}\mathrm{ar}\left(\sum_{i=1}^{N} \omega^{i,N} h(\boldsymbol{\xi}^{i,N}[k]) \middle| \mathcal{F}_0^N \right) \right] , \quad (22)$$

where $\mathcal{F}_0^N = \sigma\left\{ \boldsymbol{\xi}^{i,N}, \omega^{i,N},\ i \in \{1,\dots,N\} \right\}$. Now, by definition of $\boldsymbol{\xi}^{i,N}[k],\ i \in \{1,\dots,N\}$,

$$\mathbb{E}\left[\sum_{i=1}^{N} \omega^{i,N} h(\boldsymbol{\xi}^{i,N}[k]) \middle| \mathcal{F}_0^N \right] = \sum_{i=1}^{N} \omega^{i,N} Q^k h(\boldsymbol{\xi}^{i,N}) ,$$

and the first term of the RHS of (22) is bounded by

$$\left| \mathbb{E}\left[\sum_{i=1}^{N} \omega^{i,N} h(\boldsymbol{\xi}^{i,N}[k]) \right] - \Pi h \right| \leq \mathbb{E}\left[\sum_{i=1}^{N} \omega^{i,N} \left| Q^k h(\boldsymbol{\xi}^{i,N}) - \Pi h \right| \right] .$$

The RHS goes to 0 as k tends to infinity by the Lebesgue convergence theorem since h is bounded. The same argument holds to handle the second term of the RHS of (22):

$$\lim_{k\to\infty} \mathbb{V}\mathrm{ar}\left(\mathbb{E}\left[\sum_{i=1}^{N} \omega^{i,N} h(\boldsymbol{\xi}^{i,N}[k]) \middle| \mathcal{F}_0^N \right] \right) = \lim_{k\to\infty} \mathbb{V}\mathrm{ar}\left(\sum_{i=1}^{N} \omega^{i,N} Q^k h(\boldsymbol{\xi}^{i,N}) \right) = \mathbb{V}\mathrm{ar}\left(\sum_{i=1}^{N} \omega^{i,N} \Pi h \right) = \mathbb{V}\mathrm{ar}\left(\Pi h\right) = 0 .$$

Finally, conditionally to $\mathcal{F}_0^N$, the random variables $(\boldsymbol{\xi}^{i,N}[k])_{i=1}^N$ are independent and

$$\mathbb{V}\mathrm{ar}\left(\sum_{i=1}^{N} \omega^{i,N} h(\boldsymbol{\xi}^{i,N}[k]) \middle| \mathcal{F}_0^N \right) = \sum_{i=1}^{N} (\omega^{i,N})^2 \mathbb{V}\mathrm{ar}\left(h(\boldsymbol{\xi}^{i,N}[k]) \middle| \mathcal{F}_0^N \right) = \sum_{i=1}^{N} (\omega^{i,N})^2 \left[Q^k h^2(\boldsymbol{\xi}^{i,N}) - \left(Q^k h(\boldsymbol{\xi}^{i,N}) \right)^2 \right] ,$$

leading to

$$\lim_{k\to\infty} \mathbb{E}\left[\mathbb{V}\mathrm{ar}\left(\sum_{i=1}^{N} \omega^{i,N} h(\boldsymbol{\xi}^{i,N}[k]) \middle| \mathcal{F}_0^N \right) \right] = \left[\Pi h^2 - (\Pi h)^2 \right] \mathbb{E}\left[\sum_{i=1}^{N} (\omega^{i,N})^2 \right] .$$

This shows the first part of the proposition. Now, by the Cauchy-Schwartz inequality:

$$1 = \sum_{i=1}^{N} \omega^{i,N} \leq \left(\sum_{i=1}^{N} (\omega^{i,N})^2 \right)^{1/2} N^{1/2} ,$$

i.e. $\sum_{i=1}^{N} (\omega^{i,N})^2 \geq 1/N$ with equality only for $\omega^{i,N} = 1/N$ for all i. The proof is completed.

B. Proof of Theorem 1

Let $\gamma_N = k_N + \ln N/(2\ln\beta)$. Under the assumptions of Theorem 1, $\lim_{N\to\infty}\gamma_N = \infty$. Now, write

$$N^{-1/2}\sum_{i=1}^{N}\left[h(\tilde{\boldsymbol{\xi}}^{i,N}[k_N]) - \Pi h\right] = N^{-1/2}\sum_{i=1}^{N}\left[Q^{k_N}h(\tilde{\boldsymbol{\xi}}^{i,N}) - \Pi h\right] + N^{-1/2}\sum_{i=1}^{N}\left[h(\tilde{\boldsymbol{\xi}}^{i,N}[k_N]) - Q^{k_N}h(\tilde{\boldsymbol{\xi}}^{i,N})\right] . \quad (23)$$

Since $V \geq 1$, (**A2**)-(iii) implies that $\{N^{-1}\sum_{i=1}^{N} V(\tilde{\boldsymbol{\xi}}^{i,N})\}_{N\geq 1}$ is bounded in probability. Combining this with

$$\left|N^{-1/2}\sum_{i=1}^{N}\left[Q^{k_N}h(\tilde{\boldsymbol{\xi}}^{i,N}) - \Pi h\right]\right| \leq N^{-1/2}\beta^{k_N}\sum_{i=1}^{N}V(\tilde{\boldsymbol{\xi}}^{i,N}) = \beta^{\gamma_N}\times N^{-1}\sum_{i=1}^{N}V(\tilde{\boldsymbol{\xi}}^{i,N}) ,$$

shows that the first term of the RHS of (23) converges in probability to 0. Now, the second term of the RHS of (23) writes

$$N^{-1/2}\sum_{i=1}^{N}\left[h(\tilde{\boldsymbol{\xi}}^{i,N}[k_N]) - Q^{k_N}h(\tilde{\boldsymbol{\xi}}^{i,N})\right] = \sum_{i=1}^{N}\left\{U_{N,i} - \mathbb{E}\left[U_{N,i}|\mathcal{F}_{N,i-1}\right]\right\} ,$$

where

$$U_{N,i} = N^{-1/2}h\left(\tilde{\boldsymbol{\xi}}^{i,N}[k_N]\right) ,$$
$$\mathcal{F}_{N,i} = \sigma\left\{\tilde{\boldsymbol{\xi}}^{\ell,N}, \tilde{\boldsymbol{\xi}}^{j,N}[k_N], (\ell,j)\in\{1,\ldots,i\}^2\right\} .$$

To apply (Douc and Moulines, 2008, Theorem A3) with $M_N = N$ and $\sigma^2 = \mathbb{V}\mathrm{ar}_{\Pi}(h)$, we need to check that

$$\sum_{i=1}^{N}\mathbb{V}\mathrm{ar}\left(U_{N,i}|\mathcal{F}_{N,i-1}\right) \xrightarrow{\mathbb{P}} \sigma^2 , \quad (24)$$

$$\sum_{i=1}^{N}\mathbb{E}\left[U_{N,i}^2\mathbf{1}_{\{|U_{N,i}|\geq\varepsilon\}}\middle|\mathcal{F}_{N,i-1}\right] \xrightarrow{\mathbb{P}} 0 , \quad \text{for any } \epsilon > 0 . \quad (25)$$

We start with (24). Write

$$\left|\sum_{i=1}^{N}\mathbb{V}\mathrm{ar}\left(U_{N,i}|\mathcal{F}_{N,i-1}\right) - \sigma^2\right| \leq N^{-1}\sum_{j=1}^{N}\left|Q^{k_N}h^2(\tilde{\boldsymbol{\xi}}^{j,N}) - \Pi h^2\right| + N^{-1}\sum_{j=1}^{N}\left|\left[Q^{k_N}h(\tilde{\boldsymbol{\xi}}^{j,N})\right]^2 - (\Pi h)^2\right| . \quad (26)$$

As $h^2 \in \mathcal{C}_V$, the first term of the RHS is upper-bounded by

$$\beta^{k_N}\times N^{-1}\sum_{i=1}^{N}V(\tilde{\boldsymbol{\xi}}^{i,N}) ,$$

which converges in probability to 0. Now, note that the functions h^2 and V are in $\mathcal{C}_V$ and $|h| \leq \max(h^2, 1) \leq \max(h^2, V)$ so that $h \in \mathcal{C}_V$. By applying $|a^2 - b^2| \leq |a-b|^2 + 2|b||a-b|$, the second term of (26) is then upper-bounded by

$$\beta^{2k_N} \times N^{-1} \sum_{i=1}^{N} \left[V(\tilde{\boldsymbol{\xi}}^{i,N}) \right]^2 + 2|\Pi h| \beta^{k_N} \times N^{-1} \sum_{i=1}^{N} V(\tilde{\boldsymbol{\xi}}^{i,N}) \; ,$$

which again converges in probability to 0. This proves (24). Now, let $\varepsilon > 0$,

$$\begin{aligned}
&\sum_{i=1}^{N} \mathbb{E}\left[U_{N,i}^2 \mathbf{1}_{\{|U_{N,i}| \geq \varepsilon\}} \middle| \mathcal{F}_{N,i-1} \right] \\
&\quad \leq \Pi\left[h^2 \mathbf{1}_{\{h^2 \geq \varepsilon^2 N\}} \right] + N^{-1} \sum_{i=1}^{N} \left| Q^{k_N}\left[h^2(\tilde{\boldsymbol{\xi}}^{i,N}) \mathbf{1}_{\{h^2(\tilde{\boldsymbol{\xi}}^{i,N}) \geq \varepsilon^2 N\}} \right] - \Pi\left[h^2 \mathbf{1}_{\{h^2 \geq \varepsilon^2 N\}} \right] \right| \\
&\quad \leq \Pi\left[h^2 \mathbf{1}_{\{h^2 \geq \varepsilon^2 N\}} \right] + \beta^{k_N} \times N^{-1} \sum_{i=1}^{N} V(\tilde{\boldsymbol{\xi}}^{i,N}) \; , \qquad (27)
\end{aligned}$$

where $h^2 \mathbf{1}_{\{h^2 \geq \varepsilon^2 N\}} \in \mathcal{C}_V$. Since $h^2 \in \mathcal{C}_V$,(**A2**)-(i) implies that $\Pi h^2 < \infty$. Then, the RHS of (27) converges in probability to 0, showing (25). The proof is completed.

References

Andrieu, C., A. Doucet, and R. Holenstein (2010). Particle markov chain monte carlo methods. *J. Roy. Statist. Soc. B 72*(Part 3), 269–342.

Briers, M., A. Doucet, and S. Maskell (2010). Smoothing algorithms for state-space models. *Annals Institute Statistical Mathematics 62*(1), 61–89.

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

Carpenter, J., P. Clifford, and P. Fearnhead (1999). An improved particle filter for non-linear problems. *IEE Proc., Radar Sonar Navigation 146*, 2–7.

Chopin, N. (2004). Central limit theorem for sequential Monte Carlo methods and its application to Bayesian inference. *Ann. Statist. 32*(6), 2385–2411.

Chopin, N., P. Jacob, and O. Papaspiliopoulos (2011). *smc*^{2}: A sequential monte carlo algorithm with particle markov chain monte carlo updates. Preprint, arXiv:1011.1528v2.

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

Del Moral, P. and A. Guionnet (1999). Central limit theorem for nonlinear filtering and interacting particle systems. *Ann. Appl. Probab. 9*(2), 275–297.

Douc, R., O. Cappé, and E. Moulines (2005, September). Comparison of resampling schemes for particle filtering. In *4th International Symposium on Image and Signal Processing and Analysis (ISPA)*, Zagreb, Croatia. arXiv: cs.CE/0507025.

Douc, R., A. Garivier, E. Moulines, and J. Olsson (2010, 4). Sequential Monte Carlo smoothing for general state space hidden Markov models. *To appear in Ann. Appl. Probab.*.

Douc, R. and E. Moulines (2008). Limit theorems for weighted samples with applications to sequential Monte Carlo methods. *Ann. Statist. 36*(5), 2344–2376.

Doucet, A., S. Godsill, and C. Andrieu (2000). On sequential Monte-Carlo sampling methods for Bayesian filtering. *Stat. Comput. 10*, 197–208.

Doucet, A. and A. Johansen (2009). A tutorial on particle filtering and smoothing: fifteen years later. *Oxford handbook of nonlinear filtering*.

Fearnhead, P., D. Wyncoll, and J. Tawn (2010). A sequential smoothing algorithm with linear computational cost. *Biometrika 97*(2), 447–464.

Gilks, W. R. and C. Berzuini (2001). Following a moving target—Monte Carlo inference for dynamic Bayesian models. *J. Roy. Statist. Soc. B 63*(1), 127–146.

Godsill, S. J., A. Doucet, and M. West (2004). Monte Carlo smoothing for non-linear time series. *J. Am. Statist. Assoc. 99*, 156–168.

Hull, J. and A. White (1987). The pricing of options on assets with stochastic volatilities. *J. Finance 42*, 281–300.

Kitagawa, G. (1996). Monte-Carlo filter and smoother for non-Gaussian nonlinear state space models. *J. Comput. Graph. Statist. 1*, 1–25.

Kitagawa, G. (1998). A self-organizing state-space model. *J. Am. Statist. Assoc. 93*(443), 1203–1215.

Künsch, H. R. (2000). State space and hidden Markov models. In O. E. Barndorff-Nielsen, D. R. Cox, and C. Kluppelberg (Eds.), *Complex Stochastic Systems*. CRC Press.

Liu, J. and R. Chen (1998). Sequential Monte-Carlo methods for dynamic systems. *J. Am. Statist. Assoc. 93*(443), 1032–1044.

Olsson, J. and T. Rydén (2010). Metropolising forward particle filtering backward sampling and rao-blackwellisation of metropolised particle smoothers. Preprint, arXiv:1011.2153v1.