Biometrika (2011), **98**, 1, *pp.* 65–80
© 2011 Biometrika Trust
Printed in Great Britain

doi: 10.1093/biomet/asq062
Advance Access publication 4 February 2011

Particle approximations of the score and observed information matrix in state space models with application to parameter estimation

By GEORGE POYIADJIS

BlackRock Investment Management, 33 King William Street, London EC4R 9AS, U.K.
gpoyiadjis@cantab.net

ARNAUD DOUCET

Department of Statistics, University of British Columbia, 333-6356 Agricultural Road, Vancouver, Canada V6T 1Z2
arnaud@stat.ubc.ca

and SUMEETPAL S. SINGH

Department of Engineering, University of Cambridge, Trumpington Street, Cambridge CB2 1PZ, U.K.
sss40@cam.ac.uk

Summary

Particle methods are popular computational tools for Bayesian inference in nonlinear non-Gaussian state space models. For this class of models, we present two particle algorithms to compute the score vector and observed information matrix recursively. The first algorithm is implemented with computational complexity $\mathcal{O}(N)$ and the second with complexity $\mathcal{O}(N^2)$, where N is the number of particles. Although cheaper, the performance of the $\mathcal{O}(N)$ method degrades quickly, as it relies on the approximation of a sequence of probability distributions whose dimension increases linearly with time. In particular, even under strong mixing assumptions, the variance of the estimates computed with the $\mathcal{O}(N)$ method increases at least quadratically in time. The more expensive $\mathcal{O}(N^2)$ method relies on a nonstandard particle implementation and does not suffer from this rapid degradation. It is shown how both methods can be used to perform batch and recursive parameter estimation.

Some key words: Observed information matrix; Particle method; Score; Sequential Monte Carlo simulation; State space model; Stochastic approximation.

1. Introduction

State space models include many nonlinear and non-Gaussian time series models used in statistics, econometrics and information engineering; see Cappé et al. (2005), Durbin & Koopman (2001) and West & Harrison (1997). The following state space model is considered in this paper. Let $\{X_n\}_{n\in\mathbb{N}}$ and $\{Y_n\}_{n\in\mathbb{N}}$ be $\mathcal{X}$- and $\mathcal{Y}$-valued stochastic processes, where $\{Y_n\}_{n\in\mathbb{N}}$ is the observed time series and $\{X_n\}_{n\in\mathbb{N}}$ is the unobserved Markov process with initial density $\mu_\theta(x)$ and Markov transition density $f_\theta(x' \mid x)$:

$$X_1 \sim \mu_\theta(\cdot) \quad \text{and} \quad X_{n+1} \mid (X_n = x) \sim f_\theta(\cdot \mid x) \;\; (n = 1, 2, \ldots). \tag{1}$$

The observation at time n depends on the value of the hidden state at time n only and is drawn from the density $g_\theta(y \mid x)$:

$$Y_n \mid (X_n = x) \sim g_\theta(\cdot \mid x). \tag{2}$$

The variable θ in the above densities represents the particular parameters of the model, where we assume $\theta \in \Theta$, an open subset of $\mathbb{R}^d$. We also assume that $\mu_\theta(x)$, $f_\theta(x \mid x')$ and $g_\theta(y \mid x)$ are densities with respect to suitable dominating measures, such as the Lebesgue measure if $\mathcal{X} \subseteq \mathbb{R}^p$ and $\mathcal{Y} \subseteq \mathbb{R}^q$, denoted generically as dx and dy. These densities are assumed to be twice continuously differentiable with respect to θ.

For any sequence $\{z_k\}$, let $z_{i:j}$ denote $(z_i, z_{i+1}, \ldots, z_j)$. From (1) and (2), the joint density of $(X_{1:n}, Y_{1:n})$ is

$$p_\theta(x_{1:n}, y_{1:n}) = \mu_\theta(x_1) \prod_{k=2}^{n} f_\theta(x_k \mid x_{k-1}) \prod_{k=1}^{n} g_\theta(y_k \mid x_k).$$

Furthermore, the likelihood of the observed process is

$$p_\theta(y_{1:n}) = \int p_\theta(x_{1:n}, y_{1:n})\, dx_{1:n}. \tag{3}$$

We are interested in the problem of computing, recursively in time, the score vector $\nabla \log p_\theta(y_{1:n})$, whose rth component is

$$\{\nabla \log p_\theta(y_{1:n})\}_r = \frac{\partial \log p_\theta(y_{1:n})}{\partial \theta^r},$$

and the observed information matrix $-\nabla^2 \log p_\theta(y_{1:n})$, whose (r, s)th component is

$$\{-\nabla^2 \log p_\theta(y_{1:n})\}_{r,s} = -\frac{\partial^2 \log p_\theta(y_{1:n})}{\partial \theta^r \partial \theta^s} \quad (r, s = 1, \ldots, d).$$

Except for simple models such as the linear Gaussian state space model (Koopman & Shephard, 1992) or when $\mathcal{X}$ is a finite set (Lystig & Hughes, 2002), it is impossible to compute these quantities exactly.

In this paper, we devise sequential Monte Carlo algorithms, henceforth referred to as particle methods, to approximate the score and observed information matrix for models of the form (1) and (2). Particle methods can be used to approximate the sequence of conditional probability distributions of the latent variables $X_{1:n}$, given the observations $y_{1:n}$, i.e. $\{p_\theta(x_{1:n} \mid y_{1:n}) dx_{1:n}\}_{n\in\mathbb{N}}$. A particle approximation of $p_\theta(x_{1:n} \mid y_{1:n}) dx_{1:n}$ is comprised of a set of $N \gg 1$ weighted random samples, termed particles, where

$$\hat{p}_\theta(dx_{1:n} \mid y_{1:n}) = \sum_{i=1}^{N} W_n^{(i)} \delta_{X_{1:n}^{(i)}}(dx_{1:n}), \quad W_n^{(i)} > 0, \quad \sum_{i=1}^{N} W_n^{(i)} = 1,$$

and $\delta_{x_0}(dx)$ denotes the Dirac delta mass located at x_0. From now on, for the sake of brevity, we identify the distributions being approximated using particles by their densities. These particles are propagated in time using importance sampling and resampling steps; see Cappé et al. (2005) and Doucet et al. (2001) for a review of the literature. Using $\{\hat{p}_\theta(dx_{1:n} \mid y_{1:n})\}_{n\in\mathbb{N}}$, it is straightforward to recursively approximate expectations of the form

$$\int \left\{ \sum_{k=1}^{n-1} \varphi_{k+1}(x_k, x_{k+1}) \right\} p_\theta(x_{1:n} \mid y_{1:n}) dx_{1:n},$$

where $\varphi_{k+1} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ (Cappé et al., 2005, §8.3). As detailed in §2·1, computing the score and observed information matrix are instances of this problem. This standard implementation is $\mathcal{O}(N)$ in complexity per time step. However, it is shown in this paper that even under favourable mixing assumptions, the variance of this estimate increases at least quadratically with time n as the particle approximation $\hat{p}_\theta(dx_{1:n} \mid y_{1:n})$ becomes progressively impoverished from the successive resampling steps. That is, the number of distinct particles representing $p_\theta(x_{1:m} \mid y_{1:n})$ for any fixed $m < n$ diminishes as $n - m$ increases. Hence, whatever the number of particles, $p_\theta(x_{1:m} \mid y_{1:n})$ will eventually be approximated by a single unique particle for all sufficiently large n. This makes the method unsuitable for large datasets. This problem is well appreciated in the literature and is known as the path degeneracy problem; see Andrieu et al. (2005, §II.B), Cappé et al. (2005, §8.3) and Olsson et al. (2008) for a discussion of this issue. To overcome it, we propose in §2·2 an original algorithm that relies only on the particle estimate of $\{p_\theta(x_n \mid y_{1:n})\}_{n\in\mathbb{N}}$. This comes at a computational cost of $\mathcal{O}(N^2)$ per time step.

An important application of the proposed particle methods is to infer the parameters of models (1) and (2). Parameter estimates are obtained in §3 by maximizing the likelihood function $p_\theta(y_{1:n})$ with respect to θ using a gradient ascent algorithm, which can be done in both a batch and a recursive setting. An alternative to maximum likelihood is to follow a Bayesian approach. A prior distribution is assigned to θ and the sequence of posteriors $\{p(\theta, x_{1:n} \mid y_{1:n})\}_{n\in\mathbb{N}}$ is estimated recursively using particles; see for example Andrieu et al. (1999), Fearnhead (2002), Storvik (2002) and an unpublished 2010 paper by Lopes, Carvalho, Johannes and Polson. This approach is not general because a recursive implementation is only possible if $p(\theta \mid x_{1:n}, y_{1:n})$ can be summarized by a set of fixed-dimensional sufficient statistics. Additionally, as n increases, these algorithms also suffer from the path degeneracy problem, resulting in unreliable estimates of the posterior $p(\theta, x_{1:n} \mid y_{1:n})$; see Andrieu et al. (2005, §II.C) and Chopin et al. (2010) for some illustrations. A more detailed overview of particle-based methods for parameter estimation is presented in Kantas et al. (2009).

2. Particle approximations of the score and observed information matrix

2·1. *The Fisher and Louis identities and their particle approximations*

In this section, algorithms to recursively estimate the score and the observed information matrix for a fixed value of θ are presented. Henceforth, we assume that regularity conditions allowing the interchange of integration and differentiation are satisfied.

Using (3), Fisher's identity for the score is (Cappé et al., 2005, p. 353)

$$\nabla \log p_\theta(y_{1:n}) = \int \nabla \log p_\theta(x_{1:n}, y_{1:n}) p_\theta(x_{1:n} \mid y_{1:n})\, dx_{1:n}. \tag{4}$$

Similarly, the observed information matrix satisfies Louis' identity (Cappé et al., 2005, p. 353)

$$-\nabla^2 \log p_\theta(y_{1:n}) = \nabla \log p_\theta(y_{1:n}) \nabla \log p_\theta(y_{1:n})^{\mathrm{T}} - \frac{\nabla^2 p_\theta(y_{1:n})}{p_\theta(y_{1:n})}, \tag{5}$$

where

$$\frac{\nabla^2 p_\theta(y_{1:n})}{p_\theta(y_{1:n})} = \int \nabla \log p_\theta(x_{1:n}, y_{1:n}) \nabla \log p_\theta(x_{1:n}, y_{1:n})^{\mathrm{T}} p_\theta(x_{1:n} \mid y_{1:n}) dx_{1:n} + \int \nabla^2 \log p_\theta(x_{1:n}, y_{1:n}) p_\theta(x_{1:n} \mid y_{1:n}) dx_{1:n}. \tag{6}$$

Equations (4)–(6) suggest that it is sufficient to obtain a particle approximation of $p_\theta(x_{1:n} \mid y_{1:n})$ to approximate the score and observed information matrix. Many particle algorithms have been proposed in the literature to approximate $\{p_\theta(x_{1:n} \mid y_{1:n})\}_{n\in\mathbb{N}}$. We will focus here on the auxiliary particle filter (Pitt & Shephard, 1999), specifically, on the version of this algorithm presented in Carpenter et al. (1999), Fearnhead et al. (2008) and Papaspiliopoulos (2010), which includes only one resampling step at each time instant. Let

$$q_\theta(x_n, y_n \mid x_{n-1}) = q_\theta(x_n \mid y_n, x_{n-1}) q_\theta(y_n \mid x_{n-1})$$

be a nonnegative function on $\mathcal{X} \times \mathcal{Y}$ whose support includes that of $f_\theta(x_n \mid x_{n-1}) g_\theta(y_n \mid x_n)$. Furthermore, suppose that $q_\theta(x_n \mid y_n, x_{n-1})$ is a probability density function, from which it is easy to sample and that it is possible to evaluate $q_\theta(y_n \mid x_{n-1})$ for any $(x_{n-1}, y_n) \in \mathcal{X} \times \mathcal{Y}$. Pitt & Shephard (1999) suggest choosing $q_\theta(x_n \mid y_n, x_{n-1}) = p_\theta(x_n \mid y_n, x_{n-1})$ and $q_\theta(y_n \mid x_{n-1}) = p_\theta(y_n \mid x_{n-1})$. When this is not possible, an approximation of these quantities can be used. For the choice $q_\theta(x_n \mid y_n, x_{n-1}) = f_\theta(x_n \mid x_{n-1})$ and $q_\theta(y_n \mid x_{n-1}) = h_\theta(y_n)$, where $h_\theta(y_n)$ is an arbitrary strictly positive function, e.g. $h_\theta(y_n) = 1$, the auxiliary particle filter becomes the bootstrap particle filter introduced in the seminal paper of Gordon et al. (1993).

To recursively compute the score and observed information matrix, we use (4)–(6) and the particle approximation of $p_\theta(x_{1:n} \mid y_{1:n})$ as detailed in Algorithm 1. To each particle $X_{1:n}^{(i)}$, we also associate the vector $\alpha_n^{(i)} = \nabla \log p_\theta(X_{1:n}^{(i)}, y_{1:n})$ and the matrix $\beta_n^{(i)} = \nabla^2 \log p_\theta(X_{1:n}^{(i)}, y_{1:n})$. Algorithm 1 proceeds as follows at time $n > 1$.

Algorithm 1. *Particle approximations based on identities* (4) *and* (5).

Step 1. Resample the particle set $\{X_{1:n-1}^{(i)}, \alpha_{n-1}^{(i)}, \beta_{n-1}^{(i)}\}_{i=1}^N$ using the weights $\{W_{n-1}^{(i)} q_\theta(y_n \mid X_{n-1}^{(i)})\}_{i=1}^N$ to obtain a set of N new particles also denoted $\{X_{1:n-1}^{(i)}, \alpha_{n-1}^{(i)}, \beta_{n-1}^{(i)}\}_{i=1}^N$.

Step 2. For $i = 1, \ldots, N$, sample $X_n^{(i)} \sim q_\theta(\cdot \mid y_n, X_{n-1}^{(i)})$ and compute the weights

$$W_n^{(i)} \propto \frac{g_\theta(y_n \mid X_n^{(i)}) f_\theta(X_n^{(i)} \mid X_{n-1}^{(i)})}{q_\theta(X_n^{(i)}, y_n \mid X_{n-1}^{(i)})}.$$

Step 3. Update $\{\alpha_n^{(i)}, \beta_n^{(i)}\}_{i=1}^N$, the score estimate S_n and observed information matrix estimate Σ_n:

$$\alpha_n^{(i)} = \alpha_{n-1}^{(i)} + \nabla \log g_\theta(y_n \mid X_n^{(i)}) + \nabla \log f_\theta(X_n^{(i)} \mid X_{n-1}^{(i)}),$$

$$\beta_n^{(i)} = \beta_{n-1}^{(i)} + \nabla^2 \log g_\theta(y_n \mid X_n^{(i)}) + \nabla^2 \log f_\theta(X_n^{(i)} \mid X_{n-1}^{(i)}),$$

$$S_n = \sum_{i=1}^N W_n^{(i)} \alpha_n^{(i)}, \quad \text{and} \quad \Sigma_n = S_n S_n^{\mathrm{T}} - \sum_{i=1}^N W_n^{(i)} (\alpha_n^{(i)} \alpha_n^{(i)\mathrm{T}} + \beta_n^{(i)}). \tag{7}$$

The estimate S_n of $\nabla \log p_\theta(y_{1:n})$ is obtained by substituting $\hat{p}_\theta(dx_{1:n} \mid y_{1:n})$ for $p_\theta(x_{1:n} \mid y_{1:n})\, dx_{1:n}$ into (4). Similarly, the estimate Σ_n of $-\nabla^2 \log p_\theta(y_{1:n})$ is obtained by substituting

$\hat{p}_\theta(dx_{1:n} \mid y_{1:n})$ for $p_\theta(x_{1:n} \mid y_{1:n})\, dx_{1:n}$ into (6) and then substituting the resulting expression, together with S_n, into (5). Algorithm 1 merely implements this sequentially.

Although there is no need to store the paths $\{X^{(i)}_{1:n}\}$, Algorithm 1 relies on the particle approximation of $p_\theta(x_{1:n} \mid y_{1:n})$ and hence suffers from the path degeneracy problem. Previous particle approximations of related quantities proposed in Cérou et al. (2001) and Doucet & Tadić (2003) suffer from the same problem. Path degeneracy has severe consequences for the particle estimates of the expectations of functions of interest computed with respect to $p_\theta(x_{1:n} \mid y_{1:n})$. Consider

$$I_n = \int \left\{ \sum_{k=1}^{n} \varphi(x_k) \right\} p_\theta(x_{1:n} \mid y_{1:n})\, dx_{1:n}$$

and its particle approximation $\hat{I}_n$ obtained by substituting $\hat{p}_\theta(dx_{1:n} \mid y_{1:n})$ for $p_\theta(x_{1:n} \mid y_{1:n})\, dx_{1:n}$. We show here that the asymptotic variance of $N^{1/2}(\hat{I}_n - I_n)$ increases at least quadratically with n. This complements the result of Del Moral & Doucet (2003) which establishes, under similar assumptions, that the $\mathbb{L}_p$ error $E_\theta(|\hat{I}_n - I_n|^p)^{1/p}$, where the expectation is computed with respect to the law of the particles only, is bounded above by a term of order $\mathcal{O}(N^{-1/2} n^2)$.

THEOREM 1. *Assume there exists a probability density κ on $\mathcal{X}$, positive for all values of $x \in \mathcal{X}$, and constants $0 < \lambda, g_-, g_+ < \infty$ such that for all θ, $(x, x') \in \mathcal{X} \times \mathcal{X}$ and $y \in \mathcal{Y}$,*

$$\lambda^{-1}\, \kappa(x') \leqslant f_\theta(x' \mid x) \leqslant \lambda\, \kappa(x'), \tag{8}$$

$$g_- \leqslant g_\theta(y \mid x) \leqslant g_+. \tag{9}$$

Furthermore, assume the function $\varphi : \mathcal{X} \to \mathbb{R}$ is selected such that it is bounded and

$$\mathrm{var}_\kappa\{\varphi(X)\} = \int \kappa(x)\varphi(x)^2\, dx - \left\{ \int \kappa(x)\varphi(x)\, dx \right\}^2 > 0 \tag{10}$$

and we have

$$q_\theta(x_n, y_n \mid x_{n-1}) = f_\theta(x_n \mid x_{n-1}) h_\theta(y_n), \tag{11}$$

where $h_\theta(\cdot)$ is any strictly positive function on $\mathcal{Y}$. Then there exists a constant $\bar{\lambda} > 1$ and a range of values for λ, which includes the interval $[1, \bar{\lambda})$, such that the asymptotic variance of $N^{1/2}(\hat{I}_n - I_n)$ is bounded below by $c_2 n^2 + c_1 n + c_0$, where (c_0, c_1, c_2) are finite constants with $c_2 > 0$.

Theorem 1 is proved in Appendix. The selection in (11) corresponds to the bootstrap filter (Gordon et al., 1993). The same result holds for other $q_\theta(x_n, y_n \mid x_{n-1})$ and if we substitute $\varphi(x_{k-1}, x_k)$ for $\varphi(x_k)$; the proof would use similar arguments as Theorem 1 but is more complicated. Theorem 1 may be interpreted as follows. Even as the mixing property of the model improves, i.e. as $\lambda \downarrow 1$, the variance of $\hat{I}_n$ will still grow at least quadratically with time n. Even in the case that $\lambda = 1$, which corresponds to a latent process that is independent and identically distributed, the growth of variance is still of order n^2. This rapid growth in variance is also confirmed by simulations in §2·3, in a scenario where these strong mixing assumptions are not satisfied.

2·2. *The marginal Fisher and Louis identities and their particle approximations*

The superlinear growth in the variance of the estimates developed using (4)–(5) is due to their reliance on the particle approximation of $p_\theta(x_{1:n} \mid y_{1:n})$, whose dimension is increasing with time.

This can be circumvented by using versions of the Fisher and Louis identities that are based only on the marginal density $p_\theta(x_n \mid y_{1:n})$. The identity for the score becomes

$$\nabla \log p_\theta(y_{1:n}) = \int \nabla \log p_\theta(x_n, y_{1:n}) p_\theta(x_n \mid y_{1:n})\, dx_n. \tag{12}$$

The observed information matrix satisfies Louis' identity given in (5), but $p_\theta(y_{1:n})^{-1}\nabla^2 p_\theta(y_{1:n})$ is now expressed in terms of $p_\theta(x_n \mid y_{1:n})$,

$$\begin{aligned}\frac{\nabla^2 p_\theta(y_{1:n})}{p_\theta(y_{1:n})} = &\int \nabla \log p_\theta(x_n, y_{1:n}) \nabla \log p_\theta(x_n, y_{1:n})^{\mathrm{T}} p_\theta(x_n \mid y_{1:n})\, dx_n \\ &+ \int \nabla^2 \log p_\theta(x_n, y_{1:n}) p_\theta(x_n \mid y_{1:n})\, dx_n. \end{aligned} \tag{13}$$

Replacing $p_\theta(x_n \mid y_{1:n})\, dx_n$ in the above integrals with its particle approximation will yield the desired approximation to the score and observed information matrix. The same approach was adopted in §2·1. However, unlike the situation there where the first and second derivatives of $\log p_\theta(x_{1:n}, y_{1:n})$ could be computed exactly, there is no analytic expression for the derivatives of $\log p_\theta(x_n, y_{1:n})$. Instead, we recursively compute pointwise approximations of these quantities using particle methods. The details are as follows.

A recursion for $\nabla \log p_\theta(x_n, y_{1:n})$ is obtained by taking the ratio of $\nabla p_\theta(x_n, y_{1:n})$ and $p_\theta(x_n, y_{1:n})$, where

$$\begin{aligned}\nabla p_\theta(x_n, y_{1:n}) = p_\theta(y_{1:n-1}) g_\theta(y_n \mid x_n) &\int f_\theta(x_n \mid x_{n-1}) p_\theta(x_{n-1} \mid y_{1:n-1}) \\ &\times \{\nabla \log g_\theta(y_n \mid x_n) + \nabla \log f_\theta(x_n \mid x_{n-1}) \\ &\quad + \nabla \log p_\theta(x_{n-1}, y_{1:n-1})\}\, dx_{n-1}, \end{aligned} \tag{14}$$

$$p_\theta(x_n, y_{1:n}) = p_\theta(y_{1:n-1}) g_\theta(y_n \mid x_n) \int f_\theta(x_n \mid x_{n-1}) p_\theta(x_{n-1} \mid y_{1:n-1})\, dx_{n-1}. \tag{15}$$

These equations follow from interchanging the order of differentiation and integration. A recursion for $\nabla^2 \log p_\theta(x_n, y_{1:n})$ is established by expressing $\nabla^2 \log p_\theta(x_n, y_{1:n})$ in terms of $\nabla^2 \log p_\theta(x_{n-1}, y_{1:n-1})$:

$$\nabla^2 \log p_\theta(x_n, y_{1:n}) = \frac{\nabla^2 p_\theta(x_n, y_{1:n})}{p_\theta(x_n, y_{1:n})} - \nabla \log p_\theta(x_n, y_{1:n}) \nabla \log p_\theta(x_n, y_{1:n})^{\mathrm{T}} \tag{16}$$

where, by routine differentiation,

$$\begin{aligned}\nabla^2 p_\theta(x_n, y_{1:n}) = p_\theta(y_{1:n-1}) g_\theta(y_n \mid x_n) &\int f_\theta(x_n \mid x_{n-1}) p_\theta(x_{n-1} \mid y_{1:n-1}) \\ &\times [\{\nabla \log g_\theta(y_n \mid x_n) + \nabla \log f_\theta(x_n \mid x_{n-1}) + \nabla \log p_\theta(x_{n-1}, y_{1:n-1})\} \\ &\quad \times \{\nabla \log g_\theta(y_n \mid x_n) + \nabla \log f_\theta(x_n \mid x_{n-1}) + \nabla \log p_\theta(x_{n-1}, y_{1:n-1})\}^{\mathrm{T}} \\ &\quad + \{\nabla^2 \log g_\theta(y_n \mid x_n) + \nabla^2 \log f_\theta(x_n \mid x_{n-1}) \\ &\quad + \nabla^2 \log p_\theta(x_{n-1}, y_{1:n-1})\}]\, dx_{n-1}. \end{aligned} \tag{17}$$

The procedure for approximating the score and observed information matrix using the identities (12)–(13) is summarized in Algorithm 2. At time $n-1$, let the particle approximation be $\hat{p}_\theta(dx_{n-1} \mid y_{1:n-1}) = \sum_{i=1}^N \bar{W}_{n-1}^{(i)} \delta_{X_{n-1}^{(i)}}(dx_{n-1})$. Here, the notation for the normalized weights is different from that used in Algorithm 1, for reasons to become apparent below. Let $\tilde{p}_\theta(x_n, y_{1:n})$ denote the pointwise approximation of $p_\theta(x_n, y_{1:n})$; see (19). For each particle $X_n^{(i)}$, let the vector $\bar{\alpha}_n^{(i)}$ and the matrix $\bar{\beta}_n^{(i)}$ denote the values of the pointwise approximations of $\nabla \log p_\theta(x_n, y_{1:n})$ and $\nabla^2 \log p_\theta(x_n, y_{1:n})$ evaluated at $X_n^{(i)}$, respectively; see (21) and (20). Algorithm 2 proceeds as follows at time $n > 1$.

Algorithm 2. *Particle approximations based on identities* (12) *and* (13).

Step 1. For $i = 1, \ldots, N$, sample $X_n^{(i)} \sim q_\theta(\cdot \mid y_{1:n})$ where

$$q_\theta(x_n \mid y_{1:n}) \propto \sum_{i=1}^N \bar{W}_{n-1}^{(i)} q_\theta(y_n \mid X_{n-1}^{(i)}) q_\theta(x_n \mid y_n, X_{n-1}^{(i)}),$$

and compute the normalized weights

$$\bar{W}_n^{(i)} \propto \frac{\tilde{p}_\theta(X_n^{(i)}, y_{1:n})}{q_\theta(X_n^{(i)} \mid y_{1:n})}, \tag{18}$$

$$\tilde{p}_\theta(x_n, y_{1:n}) \propto g_\theta(y_n \mid x_n) \sum_{i=1}^N \bar{W}_{n-1}^{(i)} f_\theta(x_n \mid X_{n-1}^{(i)}). \tag{19}$$

Step 2. Update $\{\bar{\alpha}_n^{(i)}, \bar{\beta}_n^{(i)}\}_{i=1}^N$, the score estimate $\bar{S}_n$ and the observed information matrix estimate $\bar{\Sigma}_n$:

$$\bar{\alpha}_n^{(i)} = \frac{\sum_{j=1}^N \bar{W}_{n-1}^{(j)} f_\theta(X_n^{(i)} \mid X_{n-1}^{(j)})}{\sum_{k=1}^N \bar{W}_{n-1}^{(k)} f_\theta(X_n^{(i)} \mid X_{n-1}^{(k)})} \{\nabla \log g_\theta(y_n \mid X_n^{(i)}) + \nabla \log f_\theta(X_n^{(i)} \mid X_{n-1}^{(j)}) + \bar{\alpha}_{n-1}^{(j)}\}, \tag{20}$$

$$\begin{aligned} \bar{\beta}_n^{(i)} = \frac{\sum_{j=1}^N \bar{W}_{n-1}^{(j)} f_\theta(X_n^{(i)} \mid X_{n-1}^{(j)})}{\sum_{k=1}^N \bar{W}_{n-1}^{(k)} f_\theta(X_n^{(i)} \mid X_{n-1}^{(k)})} &[\{\nabla \log g_\theta(y_n \mid X_n^{(i)}) + \nabla \log f_\theta(X_n^{(i)} \mid X_{n-1}^{(j)}) + \bar{\alpha}_{n-1}^{(j)}\} \\ &\times \{\nabla \log g_\theta(y_n \mid X_n^{(i)}) + \nabla \log f_\theta(X_n^{(i)} \mid X_{n-1}^{(j)}) + \bar{\alpha}_{n-1}^{(j)}\}^{\mathrm{T}} \\ &+ \{\nabla^2 \log g_\theta(y_n \mid X_n^{(i)}) + \nabla^2 \log f_\theta(X_n^{(i)} \mid X_{n-1}^{(j)}) + \bar{\beta}_{n-1}^{(j)}\}] - \bar{\alpha}_n^{(i)} \bar{\alpha}_n^{(i)\mathrm{T}}, \end{aligned} \tag{21}$$

$$\bar{S}_n = \sum_{i=1}^N \bar{W}_n^{(i)} \bar{\alpha}_n^{(i)}, \qquad \bar{\Sigma}_n = \bar{S}_n \bar{S}_n^{\mathrm{T}} - \sum_{i=1}^N \bar{W}_n^{(i)} (\bar{\alpha}_n^{(i)} \bar{\alpha}_n^{(i)\mathrm{T}} + \bar{\beta}_n^{(i)}). \tag{22}$$

The approximations (19), (20) and (21) are obtained by substituting $\hat{p}_\theta(dx_{n-1} \mid y_{1:n-1})$ for $p_\theta(x_{n-1} \mid y_{1:n-1}) dx_{n-1}$ into (14), (15) and (17), and using (16).

Algorithm 2 requires $\mathcal{O}(N^2)$ operations instead of $\mathcal{O}(N)$ operations used in Algorithm 1. The benefit of the increased computational complexity of Algorithm 2 is that the score and observed information matrix estimates are based on approximations of integrals of the form $\int \varphi_{\theta,n}(x_n) p_\theta(x_n \mid y_{1:n})\, dx_n$ and do not rely on the particle approximation of the full posterior

$p_\theta(x_{1:n} \mid y_{1:n})$. Uniform convergence in time of the particle approximation of $p_\theta(x_n \mid y_{1:n})$ has been established by Chopin (2004, Thm 5) and Del Moral (2004, Ch. 7). Although these results rely on strong mixing assumptions, uniform convergence has been observed in numerical studies for a wide class of models where these mixing assumptions are not satisfied. Provided the recursion for $\varphi_{\theta,n}$ itself enjoys certain stability properties, we conjecture that it is possible to obtain uniform convergence results for the particle approximation of $\int \varphi_{\theta,n}(x_n) p_\theta(x_n \mid y_{1:n})\, dx_n$ even when the integrand $\varphi_{\theta,n}(x_n)$ is being estimated recursively using the previous particle approximations of the marginals $\{p_\theta(x_k \mid y_{1:k})\}_{k<n}$. This suggests that Algorithm 2 can provide estimates whose variances increase only linearly with the time n, compared to superlinearly for the particle estimates based on the identities (4)–(5). This is what we observed in all the numerical experiments presented in §3.

The auxiliary particle filter in Algorithm 2 requires $\mathcal{O}(N^2)$ operations and can be interpreted as a Rao–Blackwell version of the standard $\mathcal{O}(N)$ auxiliary particle filter of Algorithm 1, since the weights in (18) are evaluated after the auxiliary variables have been integrated out; see Lin et al. (2005) for another example of an $\mathcal{O}(N^2)$ particle filter. Any standard particle filter of complexity $\mathcal{O}(N)$ could be used in Algorithm 2, but the overall complexity will remain $\mathcal{O}(N^2)$.

2·3. *Simulations: comparing the two methods*

We begin with a study of a scalar linear Gaussian state space model, for which we may calculate the score and observed information matrix analytically. We use these exact values as benchmarks for the particle approximations. The model is

$$X_1 \sim \mathcal{N}\left(0, \frac{\sigma_V^2}{1-\phi^2}\right), \quad X_{n+1} = \phi X_n + \sigma_V V_{n+1}, \quad Y_n = X_n + \sigma_W W_n, \tag{23}$$

where $\{V_n\}$ and $\{W_n\}$ are two independent and identically distributed $\mathcal{N}(0, 1)$ sequences, mutually independent of each other and of the initial state X_1. We simulate a single realization of 10 000 observations using the parameters $\theta^* = (\phi^*, \sigma_V^*, \sigma_W^*) = (0{\cdot}8, 0{\cdot}5, 1{\cdot}0)$. We compare the exact value of the score at θ^* with the particle approximations of Algorithms 1 and 2. Comparisons were made after 2500, 5000, 7500 and 10 000 observations to monitor the increase in variance and the experiment was replicated 100 times. Figure 1 shows box plots obtained for parameters ϕ and σ_V; similar box plots were obtained for parameter σ_W. In both algorithms, we used 500 particles, $q_\theta(x_n \mid y_n, x_{n-1}) = p_\theta(x_n \mid y_n, x_{n-1})$ and $q_\theta(y_n \mid x_{n-1}) = p_\theta(y_n \mid x_{n-1})$.

Figure 1 shows that, for a fixed N, the particle estimate of Algorithm 2 significantly outperforms the corresponding particle estimate of Algorithm 1. Similar results not reported here were obtained for the particle estimates of the observed information matrix. A more revealing comparison is presented in Fig. 2. We expect the variance of the score estimate from Algorithm 2 to grow only linearly with the time index compared with a quadratic growth of variance for Algorithm 1. In this example, this indeed appears to be the case. Figure 2 displays the empirical variance of the score estimates as a function of the time index and each plot has been augmented with a best fitting straight line and quadratic curve where appropriate. This trend in the variance growth was also confirmed on the following stochastic volatility model (Pitt & Shephard, 1999)

$$X_{n+1} = \phi X_n + \sigma_V V_{n+1}, \quad Y_n = \beta \exp(X_n/2) W_n, \tag{24}$$

where V_n and W_n are defined as in (23). We simulated 20 000 observations using the parameters $\theta^* = (\phi^*, \sigma_V^*, \beta^*) = (0{\cdot}98, 0{\cdot}2, 0{\cdot}7)$. The previous simulation study was repeated but with Algorithm 1 using 500^2 particles and Algorithm 2 using 500 particles for a fair comparison in

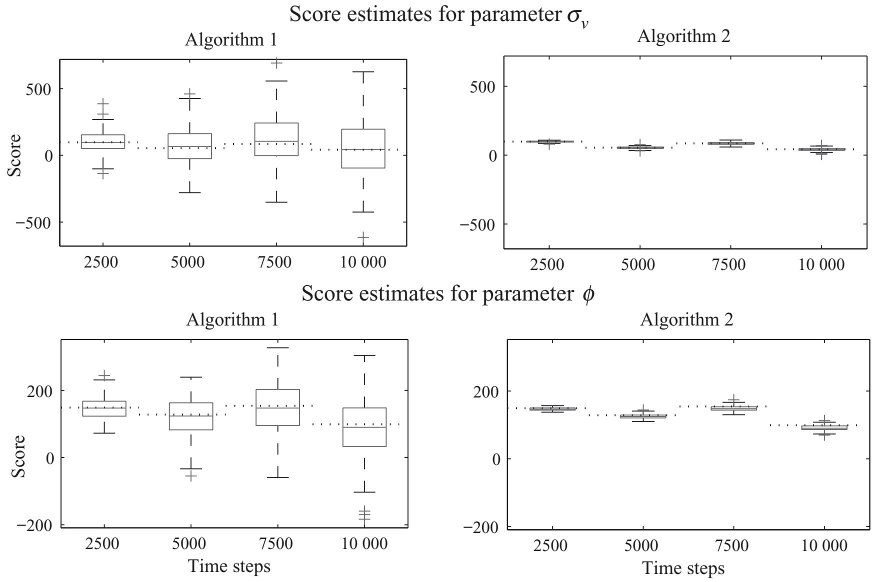

Fig. 1. Box plots of score estimates for parameters σ_V and ϕ of the linear Gaussian state space model in (23). Left column results were based on Algorithm 1 and right column results were based on Algorithm 2. The dotted lines show the true values of the scores.

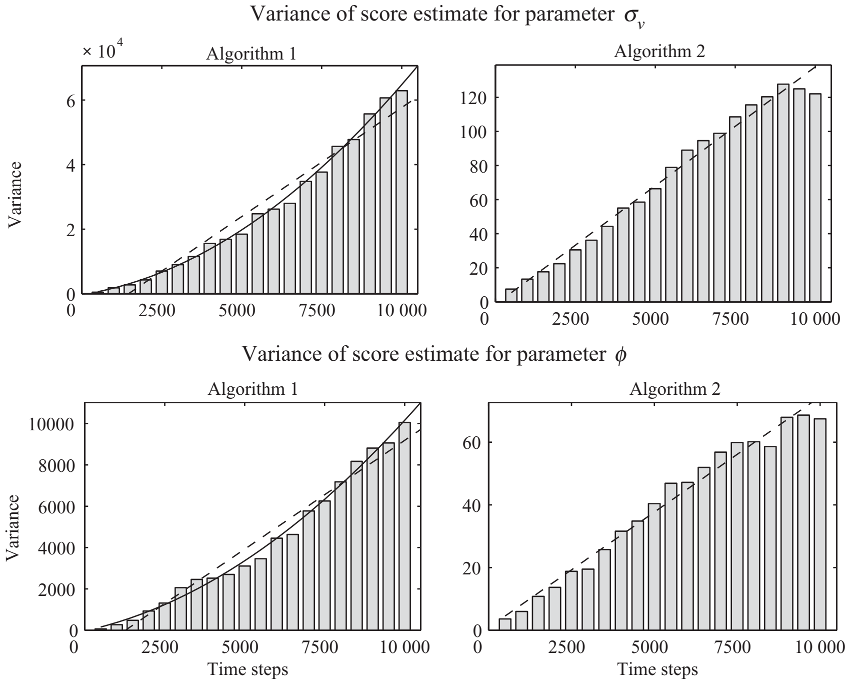

Fig. 2. Comparison of the evolution of the variance of the score estimates (7) of Algorithm 1 (left column) and (22) of Algorithm 2 (right column) at different time steps, for parameters σ_V and ϕ of the linear Gaussian model in (23). In both algorithms, 500 particles were used. The dashed line is the linear fit and the solid line represents the quadratic fit.

terms of computational complexity. As can be seen in Fig. 3, the variance of the score for the σ_V parameter in Algorithm 2 is less than that of Algorithm 1. For parameter ϕ, the quadratic variance growth in Algorithm 1 versus the linear growth in Algorithm 2 will eventually lead to Algorithm 1

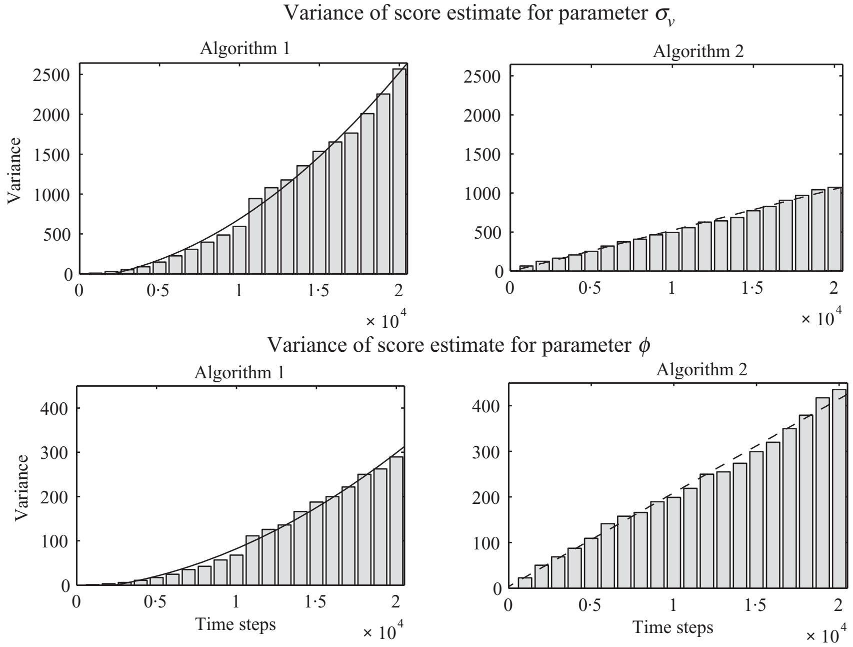

Fig. 3. Comparison of the evolution of the variance of the score estimates (7) of Algorithm 1 with 500^2 particles (left column) and (22) of Algorithm 2 with 500 particles (right column), for parameters σ_V and ϕ of the stochastic volatility model in (24). The dashed line is the linear fit and the solid line represents the quadratic fit.

being outperformed as well. Thus, for the same computational complexity, Algorithm 2 will always outperform Algorithm 1 for large enough observation records. For a small number of observations, however, Algorithm 1 is preferred, as the variance benefit of using Algorithm 2 may be too small to justify the increased computational load. It would be valuable to explore the use of fast multipole methods, dual-trees and the fast Gauss transform (Klaas et al., 2005) to reduce the computational burden of Algorithm 2.

3. APPLICATION TO PARAMETER ESTIMATION

3·1. *Batch parameter estimation*

We show here how the estimates of the score and the observed information matrix presented in §2 can be used to perform parameter estimation. Let the true static parameter generating the sequence of observations be θ^*, which is to be estimated from the observed data $\{y_n\}_{n\in\mathbb{N}}$. Given a batch of observations $y_{1:T}$, the loglikelihood may be maximized with the steepest ascent algorithm,

$$\theta_{k+1} = \theta_k + \gamma_{k+1} \nabla \log p_\theta(y_{1:T})|_{\theta=\theta_k}, \tag{25}$$

where $k = 0, 1, \ldots$ is the iteration number, θ_0 is the initial estimate and $\{\gamma_k\}$ is a sequence of small positive real numbers called the step-size sequence, which should satisfy the constraints $\sum_k \gamma_k = \infty$ and $\sum_n \gamma_k^2 < \infty$. One possible choice would be $\gamma_k = k^{-\alpha}$, $0{\cdot}5 < \alpha < 1$, e.g., $\gamma_k = k^{-2/3}$. It is also possible to include the Hessian by replacing the term multiplying γ_{k+1} with $-\{\nabla^2 \log p_\theta(y_{1:T})|_{\theta=\theta_k}\}^{-1} \nabla \log p_\theta(y_{1:T})|_{\theta=\theta_k}$. In this case, the asymptotic rate of convergence of this Newton–Raphson algorithm is quadratic and thus faster than the EM algorithm. The particle methods described earlier can be used to numerically implement either version of this steepest ascent method. In particular, each iteration of (25) would require Algorithm 1 or Algorithm 2 to be executed for T observations.

3·2. *Recursive parameter estimation*

For a long observation sequence, computing the gradient in (25) at each iteration of the algorithm is expensive. A cheaper alternative is a recursive procedure in which the data are run through once sequentially:

$$\theta_{n+1} = \theta_n + \gamma_{n+1} \nabla \log p_\theta(y_n \mid y_{1:n-1})|_{\theta=\theta_n}.$$

Upon receiving y_n, θ_n is updated in the direction of ascent of the conditional density of this new observation. This is not an online algorithm in its present form because computing $\nabla \log p_\theta(y_n \mid y_{1:n-1})$ at the current parameter estimate requires revisiting the entire history of observations. This limitation is removed by utilizing intermediate quantities that facilitate the online evaluation of this gradient (Le Gland & Mevel, 1997). In particular, define

$$p_n(x_n \mid y_{1:n}) = \frac{g_{\theta_n}(y_n \mid x_n) \int f_{\theta_n}(x_n \mid x_{n-1}) p_{n-1}(x_{n-1} \mid y_{1:n-1})\, dx_{n-1}}{\int g_{\theta_n}(y_n \mid x_n) f_{\theta_n}(x_n \mid x_{n-1}) p_{n-1}(x_{n-1} \mid y_{1:n-1})\, dx_{n-1:n}}, \tag{26}$$

$$\begin{aligned}\nabla \log p_n(x_n, y_{1:n}) = {} & \frac{\int f_{\theta_n}(x_n \mid x_{n-1}) p_{n-1}(x_{n-1} \mid y_{1:n-1})}{\int f_{\theta_n}(x_n \mid x_{n-1}) p_{n-1}(x_{n-1} \mid y_{1:n-1})\, dx_{n-1}} \\ & \times \{\nabla \log g_\theta(y_n \mid x_n)|_{\theta=\theta_n} + \nabla \log f_\theta(x_n \mid x_{n-1})|_{\theta=\theta_n} \\ & \quad + \nabla \log p_{n-1}(x_{n-1}, y_{1:n-1})\}\, dx_{n-1}.\end{aligned} \tag{27}$$

Taking the ratio of $\nabla p_\theta(x_n, y_{1:n})$ and $p_\theta(x_n, y_{1:n})$ defined in (14)–(15) will yield a recursion for $\nabla \log p_\theta(x_n, y_{1:n})$; (27) is precisely this recursion for $\nabla \log p_\theta(x_n, y_{1:n})$ but computed using the current estimate θ_n. Thus, $\nabla \log p_n(x_n, y_{1:n})$ and $p_n(x_n \mid y_{1:n})$ are not truly $\nabla \log p_\theta(x_n, y_{1:n})|_{\theta=\theta_n}$ and $p_\theta(x_n \mid y_{1:n})|_{\theta=\theta_n}$ but approximations, as they have been computed using the previous values of the parameter, i.e., $\theta_{1:n-1}$. The update rule is (Le Gland & Mevel, 1997)

$$\begin{aligned}\theta_{n+1} = \theta_n & + \gamma_{n+1} \int \nabla \log p_n(x_n, y_{1:n}) p_n(x_n \mid y_{1:n})\, dx_n \\ & - \gamma_{n+1} \int \nabla \log p_{n-1}(x_{n-1}, y_{1:n-1}) p_{n-1}(x_{n-1} \mid y_{1:n-1})\, dx_{n-1},\end{aligned} \tag{28}$$

where, by (12), the subtraction of the terms on the right-hand side yields the online approximation to $\nabla \log p_\theta(y_n \mid y_{1:n-1})|_{\theta=\theta_n}$. The quantities in (26)–(28) can only be computed exactly when $\mathcal{X}$ is finite and for linear Gaussian state space models. The asymptotic properties of this algorithm have been studied in the case of an independent and identically distributed hidden process by Titterington (1984), and by Le Gland & Mevel (1997) when $\mathcal{X}$ is a finite set. Le Gland & Mevel (1997) show that under regularity conditions this algorithm converges towards a local maximum of the average loglikelihood, and this average loglikelihood is maximized at θ^*.

The particle approximations of the score presented in the previous sections can be used to implement (28); the details are omitted. As convergence of $\{\theta_n\}$ often requires several thousand time steps, it is preferable in this case to implement Algorithm 2 to obtain an online approximation

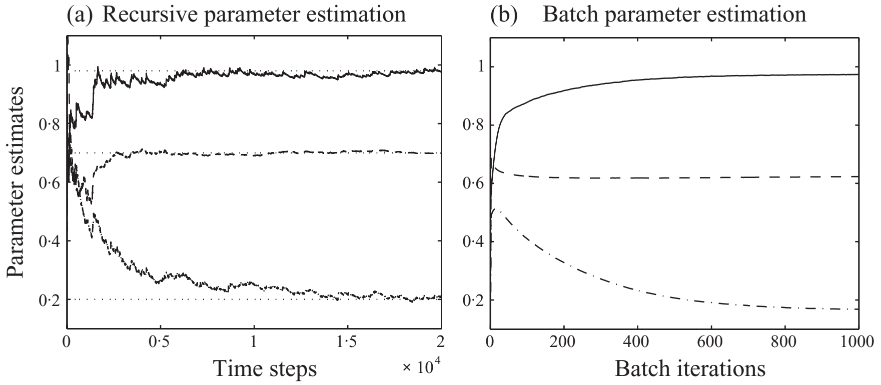

Fig. 4. Sequence of recursive parameter estimates using (28) and batch parameter estimates using (25) for the model in (24). Results are based on Algorithm 2 with 1000 particles. In both panels, estimates from top to bottom are for ϕ (solid), β (dash) and σ_V (dot-dash). True values in (a) are shown by straight dotted lines. The batch example in (b) was based on the real dataset used in Durbin & Koopman (2000).

to $\nabla \log p_\theta(y_n \mid y_{1:n-1})|_{\theta=\theta_n}$ with small variance. Over all the examples considered, we observed experimentally that the variance of Algorithm 2's estimate of $\nabla \log p_\theta(y_n \mid y_{1:n-1})|_{\theta=\theta_n}$ was uniformly bounded over time whereas the variance of Algorithm 1's estimate increased approximately linearly over time. In results not reported here, the parameter estimates for the stochastic volatility model in (24) diverged when (28) was implemented with Algorithm 1 and 10 000 particles. In contrast, Algorithm 2 with as few as 50 particles gave good results.

3·3. *Simulations*

We apply the recursive and batch parameter estimation algorithms to the stochastic volatility model introduced in (24). The model parameters $\theta = (\phi, \sigma_V, \beta)$ are to be estimated. For the recursive case, a long sequence of simulated data with $\theta^* = (0{\cdot}98, 0{\cdot}2, 0{\cdot}7)$ was generated and (28) executed using Algorithm 2 with 1000 particles. As can be seen from the results in Fig. 4(a), the estimates converged to a value in the neighbourhood of the true parameters. Using the same model, the performance of the batch parameter estimation method was assessed on the pound/dollar daily exchange rates analysed in Durbin & Koopman (2000). The steepest ascent algorithm in (25) combined with Algorithm 2 was executed for 1000 iterations with 1000 particles. The results displayed in Fig. 4(b) are consistent with those of Durbin & Koopman (2000). If the batch method is applied to T observations, then each iteration is computationally equivalent to T iterations of the recursive procedure. It is apparent that the batch method should be used when the size of the observation record is too small for the recursive procedure to converge in time. Alternatively, one could use the recursive procedure and run it repeatedly over the fixed record; the final parameter estimate of the previous run could be used as the initialization value of the current run.

We also consider a more elaborate stochastic volatility model that introduces nonlinear dynamics in the state equation. The model is the discretized version of the reparameterized continuous-time Cox–Ingersoll–Ross model discussed in Chib et al. (2006, pp. 16–17), where the volatility follows a square root process,

$$X_{n+1} = \mu + X_n + \phi \exp(-X_n) + \exp(-X_n/2)\, V_{n+1}, \qquad Y_n = \sigma_V \exp(X_n/2)\, W_n \tag{29}$$

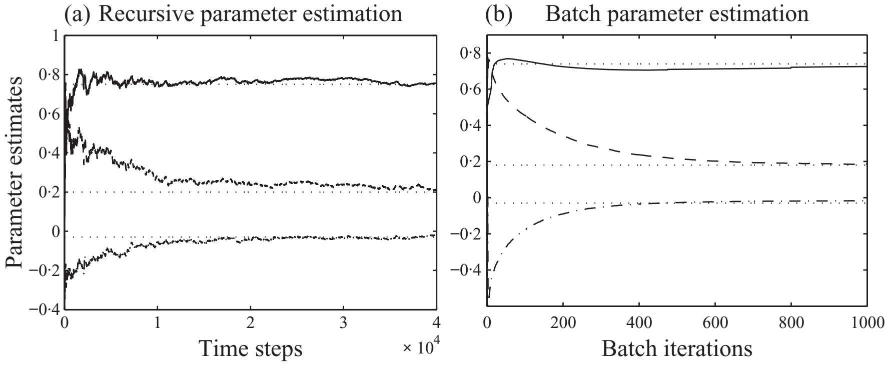

Fig. 5. Sequence of recursive parameter estimates using (28) and batch parameter estimates using (25) for the model in (29). Results are based on Algorithm 2 with 1000 particles. In both panels, estimates from top to bottom are for ϕ (solid), σ_V (dash) and μ (dot-dash). True values are shown by straight dotted lines.

and V_n and W_n are defined as in (23). The parameter $-\mu$ is the speed of mean reversion and σ_V is the volatility term of the square root volatility diffusion. We estimated the model parameters $\theta = (\mu, \phi, \sigma_V)$ in a batch and a recursive fashion using a simulated dataset of 5000 and 40 000 time steps, respectively. In both cases, the true parameters were set to $\theta^* = (-0{\cdot}03, 0{\cdot}75, 0{\cdot}2)$ and Algorithm 2 was used with 1000 particles. The results, displayed in Fig. 5, demonstrate convergence to a neighbourhood of the true parameters.

Acknowledgement

Doucet's research is funded by the Natural Sciences and Engineering Research Council of Canada. Singh's research is funded by the Engineering and Physical Sciences Research Council of the United Kingdom.

Appendix

The proof of Theorem 1 holds for any fixed θ, which is omitted from the notation. We commence by stating several auxiliary results.

It follows from assumptions (8)–(9) that the following forgetting properties hold (Cappé et al., 2005, Ch. 4; Del Moral, 2004, Ch. 4): for $t > m$,

$$\|\mathrm{pr}(X_m \in \cdot \mid y_{1:t-1}, x_t) - \mathrm{pr}(X_m \in \cdot \mid y_{1:t-1}, x_t')\|_{\mathrm{TV}} \leqslant \rho^{t-m}, \tag{A1}$$

for any (x_t, x_t') and $y_{1:t-1}$ where $\|\cdot\|_{\mathrm{TV}}$ is the total variation norm. The constant

$$\rho = 1 - \lambda^{-2} \in (0, 1) \tag{A2}$$

where λ was defined in (8). Under assumptions (8)–(9), it is also true that

$$\|\mathrm{pr}(X_n \in \cdot \mid y_{k+1:t}, x_k) - \mathrm{pr}(X_n \in \cdot \mid y_{k+1:t}, x_k')\|_{\mathrm{TV}} \leqslant \rho^{n-k}, \tag{A3}$$

for any $n > k$, (x_k, x_k') and $y_{k+1:t}$. In the literature, (A1) and (A3) are referred to as the backward and forward forgetting properties of the smoother, respectively.

The following bounds called on in the proof are a consequence of (8), (9), (10) and (11). First, there exist finite positive constants δ and Δ such that for all $i \leqslant k < t$ and $y_{1:t}$

$$\delta \leqslant \mathrm{var}_{p(x_{1:k}\mid y_{1:k-1})}\left[\frac{p(X_{1:k} \mid y_{1:t})}{p(X_{1:k} \mid y_{1:k-1})}\left\{\varphi(X_i) - \int \varphi(x_i)p(x_i \mid y_{1:t})dx_i\right\}\right] \leqslant \Delta. \tag{A4}$$

Secondly, there exist finite positive constants δ_B and Δ_B such that for all $1 < k \leqslant t$ and $y_{1:t}$,

$$\delta_B \leqslant \frac{p(x_{k-1}, x_k \mid y_{1:t})}{p(x_{k-1}, x_k \mid y_{1:k-1})} \leqslant \Delta_B. \tag{A5}$$

Rough estimates of these constants, by standard calculations, are

$$\delta_B = \lambda^{-2}\frac{g_-}{g_+}, \quad \Delta_B = \lambda^2\frac{g_+}{g_-}, \quad \Delta = \Delta_B\|\varphi\|^2, \quad \delta = \delta_B\lambda^{-3}\frac{g_-}{g_+}\mathrm{var}_\kappa\{\varphi(X)\} \tag{A6}$$

where $\|\varphi\| = \sup_{x\in\mathcal{X}} |\varphi(x)|$.

Proof of Theorem 1. We outline the main steps of the proof and omit some calculations. The expression for asymptotic variance of the particle estimate of $N^{1/2}\sum_{i=1}^t \int \varphi(x_i)p(x_i \mid y_{1:t})\,dx_i$ is (Chopin, 2004; Del Moral, 2004):

$$\begin{aligned}
&\int \frac{p(x_1 \mid y_{1:t})^2}{\mu(x_1)}\left\{\int S_t(x_{1:t})p(x_{2:t} \mid x_1, y_{2:t})\,dx_{2:t} - \bar{S}_t\right\}^2 dx_1, && \text{(term A)}\\
&+\sum_{k=2}^{t-1}\int \frac{p(x_{1:k} \mid y_{1:t})^2}{p(x_{1:k} \mid y_{1:k-1})}\left\{\int S_t(x_{1:t})p(x_{k+1:t} \mid x_k, y_{k+1:t})dx_{k+1:t} - \bar{S}_t\right\}^2 dx_{1:k}, && \text{(term B)}\\
&+\int \frac{p(x_{1:t} \mid y_{1:t})^2}{p(x_{1:t} \mid y_{1:t-1})}\{S_t(x_{1:t}) - \bar{S}_t\}^2\,dx_{1:t}, && \text{(term C)}
\end{aligned} \tag{A7}$$

where $S_t(x_{1:t}) = \sum_{i=1}^t \varphi(x_i)$, and $\bar{S}_t = \int S_t(x_{1:t})p(x_{1:t} \mid y_{1:t})\,dx_{1:t}$. The focus is on term B only and it will be established that it is bounded below by a term that grows quadratically with t.

Using the forward equation (A3) and backward equation (A1) forgetting property of the smoother, each term in the sum that defines term B can be bounded by:

$$\begin{aligned}
&\int \frac{p(x_{1:k} \mid y_{1:t})^2}{p(x_{1:k} \mid y_{1:k-1})}\left\{\int S_t(x_{1:t})p(x_{k+1:t} \mid x_k, y_{k+1:t})dx_{k+1:t} - \bar{S}_t\right\}^2 dx_{1:k}\\
&\geqslant \int \frac{p(x_{1:k} \mid y_{1:t})^2}{p(x_{1:k} \mid y_{1:k-1})}\left\{\sum_{i=1}^k \varphi(x_i) - \bar{\varphi}_{i,t}\right\}^2 dx_{1:k}\\
&\quad - 2\|\varphi\|^2\left(\frac{2}{1-\rho}+2\right)\left(\frac{\rho}{1-\rho}\right)\int \frac{p(x_{k-1}, x_k \mid y_{1:t})}{p(x_{k-1}, x_k \mid y_{1:k-1})}p(x_{k-1}, x_k \mid y_{1:t})\,dx_{k-1:k},
\end{aligned}$$

where $\bar{\varphi}_{i,t} = \int \varphi(x_i)p(x_i \mid y_{1:t})\,dx_i$. Details are routine calculations and are omitted. Combining this bound with (A5) yields:

$$\begin{aligned}
\text{term B} \geqslant &\sum_{k=2}^{t-1}\underbrace{\int \frac{p(x_{1:k} \mid y_{1:t})^2}{p(x_{1:k} \mid y_{1:k-1})}\left\{\sum_{i=1}^k \varphi(x_i) - \bar{\varphi}_{i,t}\right\}^2 dx_{1:k}}_{\mathrm{B}_{k,t}}\\
&- 2(t-2)\|\varphi\|^2\left(\frac{2}{1-\rho}+2\right)\left(\frac{\rho}{1-\rho}\right)\Delta_B.
\end{aligned} \tag{A8}$$

The second term on the right grows linearly with t and the first term will be shown to grow quadratically with t.

The importance weight simplifies to

$$\frac{p(x_{1:k}\mid y_{1:t})}{p(x_{1:k}\mid y_{1:k-1})} = w_t(x_k) = \frac{p(x_k\mid y_{1:t})}{p(x_k\mid y_{1:k-1})}, \quad k \leqslant t.$$

Consider the constituent terms that define the sum of the lower bound of term B in (A8):

$$\begin{aligned}
\mathrm{B}_{k,t} = E_{p(x_{1:k}|y_{1:k-1})}&\left[w_t(X_k)^2\left\{\sum_{i=1}^{k-1}\varphi(X_i)-\bar\varphi_{i,t}\right\}^2\right]\\
&+E_{p(x_{1:k}|y_{1:k-1})}[w_t(X_k)^2\{\varphi(X_k)-\bar\varphi_{k,t}\}^2]\\
&+2\sum_{j=1}^{k-1}E_{p(x_{1:k}|y_{1:k-1})}[w_t(X_k)^2\{\varphi(X_k)-\bar\varphi_{k,t}\}\{\varphi(X_j)-\bar\varphi_{j,t}\}].
\end{aligned}$$

The following bound is needed for the cross terms:

$$\begin{aligned}
&E_{p(x_{1:k}|y_{1:k-1})}[w_t(X_k)^2\{\varphi(X_k)-\bar\varphi_{k,t}\}\{\varphi(X_j)-\bar\varphi_{j,t}\}]\\
&\quad\leqslant 2\|\varphi\|\rho^{k-j}\left\{\int|\varphi(x_k)-\bar\varphi_{k,t}|w_t(x_k)p(x_k\mid y_{1:t})dx_k\right\}\\
&\quad\leqslant 2\|\varphi\|\rho^{k-j}\Delta^{1/2}\Delta_B^{1/2},
\end{aligned}$$

where the first bound is arrived at using (A1) and the second using (A4), (A5) and Cauchy's inequality. Thus,

$$\begin{aligned}
\mathrm{B}_{k,t} \geqslant E_{p(x_{1:k}|y_{1:k-1})}&\left[w_t(X_k)^2\left\{\sum_{i=1}^{k-1}\varphi(X_i)-\bar\varphi_{i,t}\right\}^2\right]\\
&+E_{p(x_{1:k}|y_{1:k-1})}[w_t(X_k)^2\{\varphi(X_k)-\bar\varphi_{k,t}\}^2]-4\|\varphi\|\Delta^{1/2}\Delta_B^{1/2}\frac{\rho}{1-\rho}\\
\geqslant E_{p(x_{1:k}|y_{1:k-1})}&\left[w_t(X_k)^2\left\{\sum_{i=1}^{k-1}\varphi(X_i)-\bar\varphi_{i,t}\right\}^2\right]+\delta-4\|\varphi\|\Delta^{1/2}\frac{\rho}{1-\rho}\Delta_B^{1/2}\\
\geqslant k&\left(\delta-4\|\varphi\|\Delta^{1/2}\frac{\rho}{1-\rho}\Delta_B^{1/2}\right).
\end{aligned}$$

The second inequality follows from (A4) and the last by repeating the derivation of the bound a remaining $k-1$ times. Combining all bounds leads to the following lower bound for (A7):

$$\begin{aligned}
\text{term B} &\geqslant \left(\sum_{k=2}^{t-1}\mathrm{B}_{k,t}\right)-2(t-2)\|\varphi\|^2\left(\frac{1}{1-\rho}+2\right)\left(\frac{\rho}{1-\rho}\right)\Delta_B\\
&\geqslant\left(\delta-4\|\varphi\|\Delta^{1/2}\frac{\rho}{1-\rho}\Delta_B^{1/2}\right)\frac{1}{2}(t+1)(t-2)-2(t-2)\|\varphi\|^2\left(\frac{2}{1-\rho}+2\right)\left(\frac{\rho}{1-\rho}\right)\Delta_B.
\end{aligned}$$

Thus, by (A2) and (A6), there exists a $\bar\lambda > 1$ such that for all $\lambda \in [1,\bar\lambda)$, $\delta > 4\|\varphi\|\Delta^{1/2}\rho(1-\rho)^{-1}\Delta_B^{1/2}$. □

References

Andrieu, C., De Freitas, N. & Doucet, A. (1999). Sequential Markov chain Monte Carlo for Bayesian model selection. In *IEEE Sig. Proces. Workshop on Higher-Order Statist.*, pp. 130–4, Caesarea, Israel: Institute of Electrical and Electronics Engineers.

Andrieu, C., Doucet, A. & Tadić, V. B. (2005). On-line parameter estimation in general state-space models. In *44th IEEE Conf. on Decis. Contr.*, pp. 332–7. Institute of Electrical and Electronics Engineers.

Cappé, O., Moulines, E. & Rydén, T. (2005). *Inference in Hidden Markov Models*. New York: Springer.

Carpenter, J., Clifford, P. & Fearnhead, P. (1999). An improved particle filter for nonlinear problems. *IEE Proc. Radar Sonar Navig.* **146**, 2–7.

Cérou, F., Le Gland, F. & Newton, N. J. (2001). Stochastic particle methods for linear tangent filtering equations. In *Optimal Control and PDE's – Innovations and Applications*, pp. 231–40. Amsterdam: IOS Press.

Chib, S., Pitt, M. K. & Shephard, N. (2006). Likelihood based inference for diffusion driven state space models. Working paper. Oxford: Nuffield College.

Chopin, N. (2004). Central limit theorem for sequential Monte Carlo and its application to Bayesian inference. *Ann. Statist.* **32**, 2385–411.

Chopin, N., Iacobucci, A., Marin, J., Mengersen, K., Robert, C. P., Ryder, R. & Schäfer, C. (2010). On particle learning. ArXiv e-prints URL: http://arxiv.org/abs/1006.0554.

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

Del Moral, P. & Doucet, A. (2003). *On a Class of Genealogical and Interacting Metropolis Models*, 415–46. Lecture Notes in Mathematics 1832. Berlin: Springer.

Doucet, A., de Freitas, N. & Gordon, N., Eds. (2001). *Sequential Monte Carlo Methods in Practice*. New York: Springer.

Doucet, A. & Tadić, V. B. (2003). Parameter estimation in general state-space models using particle methods. *Ann. Inst. Statist. Math.* **55**, 409–22.

Durbin, J. & Koopman, S. J. (2000). Time series analysis of non-Gaussian observations based on state space models from both classical and Bayesian perspectives (with Discussion). *J. R. Statist. Soc.* B **62**, 3–56.

Durbin, J. & Koopman, S. J. (2001). *Time Series Analysis by State-Space Methods*. Oxford: Oxford University Press.

Fearnhead, P. (2002). MCMC, sufficient statistics and particle filters. *J. Comp. Graph. Statist.* **11**, 848–62.

Fearnhead, P., Papaspiliopoulos, O. & Roberts, G. O. (2008). Particle filters for partially-observed diffusions. *J. R. Statist. Soc. B* **70**, 755–77.

Gordon, N. J., Salmond, D. & Smith, A. F. M. (1993). Novel approach to nonlinear/non-Gaussian Bayesian state estimation. *IEE Proc. Radar Sig. Proces.* **140**, 107–13.

Kantas, N., Doucet, A., Singh, S. S. & Maciejowski, J. M. (2009). An overview of sequential Monte Carlo methods for parameter estimation on general state space models. In E. Walter, ed. *15th IFAC Symp. Syst. Identification* **15**. Saint Malo, France: International Federation of Automatic Control.

Klaas, M., de Freitas, N. & Doucet, A. (2005). Toward practical N^2 Monte Carlo: the marginal particle filter. In *Proc. 21st Conf. Uncertainty in Artif. Intel*. Edinburgh: AUAI Press.

Koopman, S. & Shephard, N. (1992). Exact score for time series models in state space form. *Biometrika* **79**, 823–6.

Le Gland, F. & Mevel, M. (1997). Recursive estimation in hidden Markov models. In *Proc. 36th IEEE Conf. Decis. Contr.* **4**, 10–2. San Diego, CA: Institute of Electrical and Electronics Engineers.

Lin, M. T., Zhang, J. L., Cheng, Q. & Chen, R. (2005). Independent particle filters. *J. Am. Statist. Assoc.* **100**, 1412–21.

Lystig, T. C. & Hughes, J. P. (2002). Exact computation of the observed information matrix for hidden Markov models. *J. Comp. Graph. Statist.* **11**, 678–89.

Olsson, J., Douc, R., Cappé, O. & Moulines, E. (2008). Sequential Monte Carlo smoothing with application to parameter estimation in nonlinear state space models. *Bernoulli* **14**, 155–79.

Papaspiliopoulos, O. (2010). A methodological framework for Monte Carlo probabilistic inference for diffusion processes. In *Inference and Learning in Dynamic Models*, Ed. D. Barber, A. T. Cemgil & S. Chiappa, pp. 95–120. Cambridge: Cambridge University Press.

Pitt, M. K. & Shephard, N. (1999). Filtering via simulation: auxiliary particle filter. *J. Am. Statist. Assoc.* **94**, 590–9.

Storvik, G. (2002). Particle filters in state space models with the presence of unknown static parameters. *IEEE Trans. Sig. Proc.* **50**, 281–9.

Titterington, D. M. (1984). Recursive parameter estimation using incomplete data. *J. R. Statist. Soc.* B **46**, 257–67.

West, M. & Harrison, P. J. (1997). *Bayesian Forecasting and Dynamic Models*. New York: Springer.

[*Received May* 2009. *Revised July* 2010]