Ann. Inst. Statist. Math.
Vol. 55, No. 2, 409–422 (2003)
©2003 The Institute of Statistical Mathematics

# PARAMETER ESTIMATION IN GENERAL STATE-SPACE MODELS USING PARTICLE METHODS

ARNAUD DOUCET[1] AND VLADISLAV B. TADIĆ[2]

[1] *Signal Processing Group, Department of Engineering, Cambridge University, Trumpington Street, CB2 1PZ Cambridge, U.K.*, e-mail: ad2@eng.cam.ac.uk
[2] *Department of Electrical and Electronic Engineering, University of Melbourne, Parkville, Victoria 3010, Australia*, e-mail: v.tadic@ee.mu.oz.au

(Received June 7, 2002; revised October 18, 2002)

**Abstract.** Particle filtering techniques are a set of powerful and versatile simulation-based methods to perform optimal state estimation in nonlinear non-Gaussian state-space models. If the model includes fixed parameters, a standard technique to perform parameter estimation consists of extending the state with the parameter to transform the problem into an optimal filtering problem. However, this approach requires the use of special particle filtering techniques which suffer from several drawbacks. We consider here an alternative approach combining particle filtering and gradient algorithms to perform batch and recursive maximum likelihood parameter estimation. An original particle method is presented to implement these approaches and their efficiency is assessed through simulation.

*Key words and phrases*: Optimal filtering, parameter estimation, sequential Monte Carlo, state-space models, stochastic approximation.

## 1. Introduction

### 1.1 *State-space models and problem statement*

Let $(\Omega, \mathcal{F})$ be a measurable space. Let $\{X_n\}_{n\geq 0}$ and $\{Y_n\}_{n\geq 0}$ be $\mathbb{R}^p$ and $\mathbb{R}^q$-valued stochastic processes defined on $(\Omega, \mathcal{F})$ and $\theta \in \Theta$ where $\Theta$ is an open subset of $\mathbb{R}^k$. There exist probability measures $\mathcal{P}_\theta : \mathcal{F} \to [0,1]$, $\mu : \mathcal{B}^p \to [0,1]$ and Borel-measurable functions $p_\theta : \mathbb{R}^p \times \mathbb{R}^p \to [0,\infty)$, $\varepsilon_\theta : \mathbb{R}^p \times \mathbb{R}^q \to [0,\infty)$ such that $\int p_\theta(x,x')dx' = \int \varepsilon_\theta(x,y)dy = 1$, $\mathcal{P}_\theta(X_0 \in B) = \int_B \mu(dx_0)$ and

$$\mathcal{P}_\theta(X_{n+1} \in B \mid X^n, Y^n) = \int_B p_\theta(X_n, x)dx, \qquad \forall B \in \mathcal{B}^p,\ n \geq 0, \tag{1.1}$$

$$\mathcal{P}_\theta(Y_{n+1} \in B \mid X^{n+1}, Y^n) = \int_B \varepsilon_\theta(X_{n+1}, y)dy, \qquad \forall B \in \mathcal{B}^q,\ n \geq 0, \tag{1.2}$$

where we denote by $Z^n \triangleq (Z_0, Z_2, \dots, Z_n)$ the path of a process $\{Z_n\}_{n\geq 0}$ from time 0 to time $n$. This class of models include many nonlinear and non-Gaussian time series models (Kitagawa and Gersch (1996)).

Let us assume that the true value of the parameter $\theta$ is $\theta^*$ and that only the process $\{Y_n\}_{n\geq 0}$ is observed. We are interested in deriving a batch and a recursive algorithm to estimate $\theta^*$. These problems are very complex. Indeed, even if $\theta^*$ were known, the optimal filtering problem, i.e. estimating the conditional distribution of $X_n$ given $Y^n$, does not admit a closed-form solution and one needs to perform numerical approximations.

### 1.2 *A brief literature review*

Recently, there has been a surge of interest in particle filtering methods to perform optimal filtering in general state-space models; see Doucet *et al.* (2001) for a booklength survey and Iba (2001) for alternative applications of particle methods. Although particle filters have proven to be successful in many applications (Doucet *et al.* (2001)), parameter estimation using particle methods still remains a major problem. A standard approach followed in the literature consists of setting a prior distribution on the unknown parameter $\theta$ and then considering the extended state $Z_n \triangleq (X_n, \theta)$. This converts the parameter estimation into an optimal filtering problem. One can then apply, at least theoretically, standard particle filtering techniques. In this approach, if one were to use say the bootstrap filter (Gordon *et al.* (1993); Kitagawa (1996)), then the parameter space is only explored at the initialization of the algorithm. Consequently the algorithm is inefficient; after a few iterations the marginal posterior distribution of the parameter is approximated by a single delta Dirac function. To limit this problem, several authors have proposed to use kernel density estimation methods (Gordon *et al.* (1993); Liu and West (2001)). However, this has the effect of transforming the fixed parameter into a slowly time-varying one. A pragmatic approach consists of introducing explicitly an artificial dynamic on the parameter of interest; see Higuchi (1997) and Kitagawa (1998). In this case, we obtain a so-called self-organizing or self-tuning state-space model (Kitagawa (1998)). To avoid the introduction of an artificial dynamic, a clever approach proposed in Gilks and Berzuini (2001) consists of adding Markov chain Monte Carlo (MCMC) steps so as to add "diversity" among the particles. However, this approach does not really solve the fixed-parameter estimation problem. More precisely, the addition of MCMC steps does not make the dynamic model ergodic. Thus, there is an accumulation of errors over time and the algorithm can diverge as observed by Andrieu *et al.* (1999). Methods based on MCMC steps also require the conditional distribution of the parameter given the data $\{Y_n\}_{n\geq 0}$ and the hidden process $\{X_n\}_{n\geq 0}$ to be in the exponential family so as to be able to express it through a set of sufficient statistics. If this assumption is not satisfied, the computational complexity and memory requirements of these algorithms at time $n$ are of order $\mathcal{O}(nN)$ for $N \gg 1$ particles; this makes these methods useless from a practical point of view.

We propose here two approaches to perform parameter estimation in general state-space models. The first method is a batch gradient algorithm to perform Maximum Likelihood (ML) parameter estimation where one computes at each iteration the gradient of the log-likelihood with respect to the parameters of interest. The second algorithm is a natural recursive version of the first one usually referred to as Recursive ML (RML). Loosely speaking, RML is a stochastic gradient method for maximizing the average log-likelihood. It has been developed originally in automatic control and signal processing where it is widely used; see for example Ljung and Söderström (1987). In these approaches, it is not necessary to set a prior on the unknown static parameter and the conditional distribution of the parameter given the data and the hidden process does not have to be in the exponential family. In the context of general state-space models, these methods require essentially the ability to compute the optimal filter and the derivative of this filter with respect to the parameter of interest. We present here an original particle method to approximate the derivative of the filter and we apply it to perform batch and recursive ML.

### 1.3 *Organization of the paper*

The rest of the paper is organized as follows: In Section 2, we detail the batch and recursive maximum likelihood approaches. Section 3 describes a particle method to implement the algorithms given in the previous section. In Section 4, we apply these algorithms to several parameter estimation problems.

## 2. Batch and recursive maximum likelihood estimation

For ease of presentation, we will consider the case where $\theta \in \mathbb{R}$ in the two following sections; the extension to a multidimensional parameter is straightforward and is described briefly in Subsection 3.2.

The likelihood of the observations $Y^n$ satisfies

$$L_\theta(Y^n) = \int \cdots \int \mu(dx_0) \prod_{k=1}^{n} \varepsilon_\theta(x_k, Y_k) \mathcal{P}_\theta(X_k \in dx_k \mid Y^{k-1}),$$

where $\mathcal{P}_\theta(X_k \in dx_k \mid Y^{k-1})$ is the posterior distribution of the state $X_k$ given the observations $Y^{k-1}$ and $\theta$; the true parameter value being $\theta^*$. The log-likelihood function satisfies

$$l_\theta(Y^n) = \log(L_\theta(Y^n)) = \sum_{k=1}^{n} \log \left( \int \varepsilon_\theta(x_k, Y_k) \mathcal{P}_\theta(X_k \in dx_k \mid Y^{k-1}) \right).$$

We propose here two gradient type algorithms to perform ML estimation. The first method is a batch algorithm designed to maximize $l_\theta(Y^n)$. The second method is a recursive version of the first method maximizing $\lim_{n\to\infty} \frac{1}{n} l_\theta(Y^n)$.

### 2.1 *Batch maximum likelihood*

Let us assume the data $Y^n$ are available. We want to maximize the log-likelihood $l_\theta(Y^n)$ using a gradient algorithm. At iteration $m+1$, the parameter estimate is updated through

$$\theta_{m+1} = \theta_m + \gamma_m (\partial l)_{\theta_m}(Y^n), \tag{2.1}$$

where $\partial$ denotes the derivative evaluated at the point $\theta_m$. The stepsize sequence $\{\gamma_m\}_{m\geq 0}$ is a positive non-increasing sequence typically chosen as $\gamma_m = \gamma_0 \cdot m^{-\alpha}$ with $\gamma_0 > 0$ and $0.5 < \alpha \leq 1$.

The derivative of the log-likelihood satisfies

$$\begin{aligned} &(\partial l)_\theta(Y^n) \\ &= \sum_{k=1}^{n} \partial \log \left( \int \varepsilon_\theta(x_k, Y_k) \mathcal{P}_\theta(X_k \in dx_k \mid Y^{k-1}) \right) \\ &= \sum_{k=1}^{n} \frac{\int (\partial\varepsilon)_\theta(x_k, Y_k) \mathcal{P}_\theta(X_k \in dx_k \mid Y^{k-1}) + \int \varepsilon_\theta(x_k, Y_k)(\partial\mathcal{P})_\theta(X_k \in dx_k \mid Y^{k-1})}{\int \varepsilon_\theta(x_k, Y_k) \mathcal{P}_\theta(X_k \in dx_k \mid Y^{k-1})}. \end{aligned} \tag{2.2}$$

The expression (2.2) involves both the one-step ahead prediction distribution $\mathcal{P}_\theta(X_k \in dx_k \mid Y^{k-1})$ and its derivative $(\partial\mathcal{P})_\theta(X_k \in dx_k \mid Y^{k-1})$. We are interested in computing recursively these quantities. Further on, we use the notation: $\pi^{\theta}_{k|l}(dx_k) \triangleq \mathcal{P}_{\theta}(X_k \in dx_k \mid Y^l)$ and $w^{\theta}_{k|l}(dx_k) = (\partial\mathcal{P})_{\theta}(X_k \in dx_k \mid Y^l)$. It follows that the derivative of the log-likelihood (2.2) can be rewritten as

$$(2.3) \quad (\partial l)_{\theta}(Y^n) = \sum_{k=1}^{n} \frac{\int (\partial\varepsilon)_{\theta}(x_k, Y_k)\pi^{\theta}_{k|k-1}(dx_k) + \int \varepsilon_{\theta}(x_k, Y_k) w^{\theta}_{k|k-1}(dx_k)}{\int \varepsilon_{\theta}(x_k, Y_k)\pi^{\theta}_{k|k-1}(dx_k)}.$$

For any measure $\mu(dx)$, transition kernel $g(x, dx')$ (not necessarily a probability measure or a transition probability kernel) and function $f(x)$, we also use the following standard notation

$$\mu g(dx') = \int \mu(dx)g(x, dx'), \qquad \langle \mu, f\rangle = \int \mu(dx)f(x), \qquad \mu.f(dx) = \mu(dx)f(x).$$

The probability distribution $\pi^{\theta}_{k|k-1}$ and the signed measure $w^{\theta}_{k|k-1}$ satisfy

$$\pi^{\theta}_{k|k-1} = \pi^{\theta}_{k-1|k-1}p_{\theta}, \quad \text{i.e.} \quad \pi^{\theta}_{k|k-1}(dx_k) = \int \pi^{\theta}_{k-1|k-1}(dx_{k-1})p_{\theta}(x_{k-1}, x_k)dx_k,$$

$$\pi^{\theta}_{k|k} = \frac{\pi^{\theta}_{k|k-1}.\varepsilon_{\theta}}{\langle \pi^{\theta}_{k|k-1}, \varepsilon_{\theta}\rangle}, \quad \text{i.e.} \quad \pi^{\theta}_{k|k}(dx_k) = \frac{\pi^{\theta}_{k|k-1}(dx_k)\varepsilon_{\theta}(x_k, Y_k)}{\int \pi^{\theta}_{k|k-1}(dx_k)\varepsilon_{\theta}(x_k, Y_k)},$$

and, using a similar notation,

$$w^{\theta}_{k|k-1} = w^{\theta}_{k-1|k-1}p_{\theta} + \pi^{\theta}_{k-1|k-1}(\partial p)_{\theta},$$

$$w^{\theta}_{k|k} = \frac{w^{\theta}_{k|k-1}.\varepsilon_{\theta} + \pi^{\theta}_{k|k-1}.(\partial\varepsilon)_{\theta}}{\langle \pi^{\theta}_{k|k-1}, \varepsilon_{\theta}\rangle} - \pi^{\theta}_{k|k}\frac{\langle w^{\theta}_{k|k-1}, \varepsilon_{\theta}\rangle + \langle \pi^{\theta}_{k|k-1}, (\partial\varepsilon)_{\theta}\rangle}{\langle \pi^{\theta}_{k|k-1}, \varepsilon_{\theta}\rangle}.$$

As $w^{\theta}_{k|l}(dx_k)$ is the derivative of a probability measure, one can check that under regularity assumptions $\int w^{\theta}_{k|l}(dx_k) = 0$.

### 2.2 *Recursive maximum likelihood*

Recursive ML is a gradient type approach to maximize the average log-likelihood. This quantity has indeed "good" properties. Under suitable *regularity* conditions (Tadić and Doucet (2002)), one can show that

$$\frac{1}{n} l_{\theta}(Y^n) \to l(\theta)$$

with

$$l(\theta) \triangleq \iint_{\mathbb{R}^q \times \mathcal{P}(\mathbb{R}^p)} \log\left(\int \varepsilon_{\theta}(x, y)\nu(dx)\right) \lambda_{\theta,\theta^*}(dy, d\nu),$$

where $\mathcal{P}(\mathbb{R}^p)$ is the space of probability distributions on $\mathbb{R}^p$ and $\lambda_{\theta,\theta^*}(dy, d\nu)$ is the joint invariant distribution of the couple $(Y_k, \mathcal{P}_{\theta}(X_k \in dx_k \mid Y^{k-1}))$. It is dependent on both $\theta$ and the true parameter $\theta^*$. The following Kullback-Leibler information measure satisfies

$$(2.4) \qquad K(\theta, \theta^*) \triangleq l(\theta^*) - l(\theta) \geq 0$$

and thus

$$\theta^* \subseteq \arg\min_{\theta\in\Theta} K(\theta, \theta^*).$$

To estimate $\theta^*$, RML aims at minimizing $K(\theta, \theta^*)$ through a gradient method. The gradient of $K(\theta, \theta^*)$ does not admit an analytical expression but the gradient of $l_\theta(Y^n)$ satisfies, under regularity assumptions (Tadić and Doucet (2002)),

$$\frac{1}{n}(\partial l)_\theta(Y^n) \to -\partial K(\theta, \theta^*).$$

RML is a stochastic approximation algorithm based on the Robbins-Monro procedure to find the zeros of $-\partial K(\theta, \theta^*)$ (Benveniste *et al.* (1990); Ljung and Söderström (1987)):

$$\begin{aligned}
(2.5)\qquad \theta_{n+1} &= \theta_n + \gamma_n \partial \log\left(\int \varepsilon_{\theta_n}(x_n, Y_n)\pi^{\theta^n}_{n|n-1}(dx_n)\right) \\
(2.6)\qquad &= \theta_n + \gamma_n \left(\frac{\int (\partial\varepsilon)_{\theta_n}(x_n, Y_n)\pi^{\theta^n}_{n|n-1}(dx_n) + \int \varepsilon_{\theta_n}(x_n, Y_n) w^{\theta^n}_{n|n-1}(dx_n)}{\int \varepsilon_{\theta_n}(x_n, Y_n)\pi^{\theta^n}_{n|n-1}(dx_n)}\right)
\end{aligned}$$

where we denote $\pi^{\theta^n}_{n|n-1}(dx_n)$ the probability measure satisfying

$$\begin{aligned}
(2.7)\qquad \pi^{\theta^n}_{n|n-1} &= \pi^{\theta^n}_{n-1|n-1} p_{\theta_n}, \\
(2.8)\qquad \pi^{\theta^{n+1}}_{n|n} &= \frac{\pi^{\theta^n}_{n|n-1}\cdot\varepsilon_{\theta_{n+1}}}{\langle \pi^{\theta^n}_{n|n-1}, \varepsilon_{\theta_{n+1}}\rangle},
\end{aligned}$$

i.e. $\pi^{\theta^{n+1}}_{n|n}(dx_n)$ corresponds to the filter associated to the sequences of parameters values $\theta^{n+1}$, the parameter being updated between the prediction step (see (2.7)) and the updating step (see (2.8)) using (2.5)–(2.6). The derivative of this measure $w^{\theta^n}_{n|n}(dx_n)$ is a signed measure satisfying

$$\begin{aligned}
(2.9)\qquad w^{\theta^n}_{n|n-1} &= w^{\theta^n}_{n-1|n-1} p_{\theta_n} + \pi^{\theta^n}_{n-1|n-1}(\partial p)_{\theta_n}, \\
(2.10)\qquad w^{\theta^{n+1}}_{n|n} &= \frac{w^{\theta^n}_{n|n-1}\cdot\varepsilon_{\theta_{n+1}} + \pi^{\theta^n}_{n|n-1}\cdot(\partial\varepsilon)_{\theta_{n+1}}}{\langle \pi^{\theta^n}_{n|n-1}, \varepsilon_{\theta_{n+1}}\rangle} \\
&\quad - \pi^{\theta^{n+1}}_{n|n}\frac{\langle w^{\theta^n}_{n|n-1}, \varepsilon_{\theta_{n+1}}\rangle + \langle \pi^{\theta^n}_{n|n-1}, (\partial\varepsilon)_{\theta_{n+1}}\rangle}{\langle \pi^{\theta^n}_{n|n-1}, \varepsilon_{\theta_{n+1}}\rangle}.
\end{aligned}$$

The stepsize $\gamma_n$ is a positive non-increasing sequence such that $\sum \gamma_n = \infty$ and $\sum \gamma_n^2 < \infty$; typically one selects $\gamma_n = \gamma_0 . n^{-\alpha}$ where $\gamma_0 > 0$ and $0.5 < \alpha \leq 1$ (Benveniste *et al.* (1990)).

### 2.3 *Convergence and implementation issues*

In the context of finite state-space hidden Markov models, it is possible to compute the optimal filter and its derivative analytically. In this case, under weak assumptions, it can be established that (2.1) converges towards $\{\theta : (\partial l)_\theta(Y^n) = 0\}$ and (2.6) converges towards $\{\theta : \partial K(\theta, \theta^*) = 0\}$ (LeGland and Mevel (1997)). Note that without any additional assumption these gradient methods are not ensured to converge to the true parameter value but only to the set of zeros of the gradient of the cost functions $l_\theta(Y^n)$ and $K(\theta, \theta^*)$ given in (2.4).

Using (Tadić and Doucet (2002)), similar results could also be established for general state-space models. However, in this case, the algorithms defined by (2.1) and (2.6) cannot be implemented as the filter and its derivative cannot be computed exactly and need to be approximated numerically. When these quantities are approximated, results in (Tadić (2000)) suggest that (2.1) and (2.6) converge to a neighborhood of the zeros of the derivative of $l_\theta(Y^n)$ and $K(\theta,\theta^*)$.

Many efficient particle methods have been proposed to approximate the filter (see Doucet *et al.* (2001)) so we focus in this paper on approximating the derivative of the filter. We present in the next section an original particle method to address this problem. An alternative method has been proposed independently in (Cérou *et al.* (2001)) for a specific continuous-time model.

## 3. Parameter estimation using particle methods

In this section, we present a method to compute a particle approximation of $w^\theta_{n|n}$ in Subsection 3.1. This method is based on an importance sampling resampling strategy. A few extensions of the algorithm are presented in Subsection 3.2. Alternative approaches to estimate $w^\theta_{n|n}$ are then discussed in Subsection 3.3.

### 3.1 *Importance sampling*

For simplicity, we only describe here a particle method using the importance density $p_\theta(\cdot,\cdot)$ to approximate $\pi^\theta_{n|n}$ (Gordon *et al.* (1993); Kitagawa (1996)). More elaborate algorithms in (Doucet *et al.* (2000, 2001); Liu and Chen (1998); Pitt and Shephard (1999)) could also be used; see Subsection 3.2.

For two measures $\mu$ and $\nu$ on a common measurable space, we write $\mu \ll \nu$ if and only if $\nu(A) = 0 \Rightarrow \mu(A) = 0$ for all measurable sets $A$. Let us assume that $w^\theta_{n-1|n-1} \ll \pi^\theta_{n-1|n-1}$, then it can be shown using (2.9) and (2.10) that $w^\theta_{n|n-1} \ll \pi^\theta_{n|n-1}$ and $w^\theta_{n|n} \ll \pi^\theta_{n|n}$. This suggests that one can approximate $w^\theta_{n|n}$ by simply computing the importance ratio (i.e. Radon-Nykodym derivative) of $w^\theta_{n|n}$ with $\pi^\theta_{n|n}$. Let us assume that at time $n-1$, we have the following particle approximation of $\pi^\theta_{n-1|n-1}$

$$\widehat{\pi}^\theta_{n-1|n-1}(dx_{n-1}) = \frac{1}{N}\sum_{i=1}^{N}\delta_{\widehat{X}_{n-1,i}}(dx_{n-1}) \tag{3.1}$$

where $\delta_{X_{n-1,i}}(dx_{n-1})$ denotes the delta-Dirac measure at $\widehat{X}_{n-1,i}$ and

$$\widehat{w}^\theta_{n-1|n-1}(dx_{n-1}) = \sum_{i=1}^{N}\beta_{n-1|n-1,i}\delta_{\widehat{X}_{n-1,i}}(dx_{n-1}).$$

Note that in the above expression the coefficients $\beta_{n-1|n-1,i}$ can be negative or positive and do not sum to 1 as $w^\theta_{n-1|n-1}$ is not a probability measure. Using a particle interpretation of (2.9)–(2.10) and rewriting (2.9) as

$$w^\theta_{n|n-1} = w^\theta_{n-1|n-1}p_\theta + \pi^\theta_{n-1|n-1}p_\theta\frac{(\partial p)_\theta}{p_\theta},$$

one can derive a particle method to approximate $w^\theta_{n|n}$. It proceeds as follows at time $n$. We use the notation $\mathbb{I}_A(z) = 1$ if $z \in A$ and 0 otherwise.

Gradient estimation using particle methods

*Sampling step*

- For $i = 1, \dots, N$, sample $\widetilde{X}_{n,i} \sim p_\theta(\widehat{X}_{n-1,i}, \cdot)$.
- $\widehat{\pi}^\theta_{n|n-1}(dx_n) = \frac{1}{N}\sum_{i=1}^N \delta_{\tilde{X}_{n,i}}(dx_n)$ and $\widehat{w}^\theta_{n|n-1}(dx_n) = \sum_{i=1}^N \beta_{n|n-1,i}\delta_{\tilde{X}_{n,i}}(dx_n)$,

where

$$\beta_{n|n-1,i} = \beta_{n-1|n-1,i} + \frac{1}{N}\frac{(\partial p)_\theta(\widehat{X}_{n-1,i}, \widetilde{X}_{n,i})}{p_\theta(\widehat{X}_{n-1,i}, \widetilde{X}_{n,i})}.$$

*Updating step*

- $\widetilde{\pi}^\theta_{n|n}(dx_n) = \sum_{i=1}^N \widetilde{\alpha}_{n|n,i}\delta_{\tilde{X}_{n,i}}(dx_n)$, where $\widetilde{\alpha}_{n|n,i} \propto \varepsilon_\theta(\widetilde{X}_{n,i}, Y_n)$, $\sum_{i=1}^N \widetilde{\alpha}_{n|n,i} = 1$.
- $\widetilde{w}^\theta_{n|n}(dx_n) = \sum_{i=1}^N \widetilde{\beta}_{n|n,i}\delta_{\tilde{X}_{n,i}}(dx_n)$, where

$$\widetilde{\beta}_{n|n,i} = \frac{(\partial\varepsilon)_\theta(\widetilde{X}_{n,i}, Y_n) + N\beta_{n|n-1,i}\varepsilon_\theta(\widetilde{X}_{n,i}, Y_n)}{\sum_{j=1}^N \varepsilon_\theta(\widetilde{X}_{n,j}, Y_n)} - \frac{\sum_{j=1}^N((\partial\varepsilon)_\theta(\widetilde{X}_{n,j}, Y_n) + N\beta_{n|n-1,j}\varepsilon_\theta(\widetilde{X}_{n,j}, Y_n))}{\sum_{j=1}^N \varepsilon_\theta(\widetilde{X}_{n,j}, Y_n)}\widetilde{\alpha}_{n|n,i}.$$

*Resampling step*

- Multiply/Discard particles $\widetilde{X}_{n,i}$ with respect to the high/low weights $\widetilde{\alpha}_{n|n,i}$ to obtain $N$ particles $\widehat{X}_{n,i}$, i.e. $\widehat{X}_{n,i} = \widetilde{X}_{n,\varphi_n(i)}$ where $\varphi_n(i)$ is determined by the resampling mechanism.
- $\widehat{\pi}^\theta_{n|n}(dx_n) = \frac{1}{N}\sum_{i=1}^N \delta_{\hat{X}_{n,i}}(dx_n)$ and $\widehat{w}^\theta_{n|n}(dx_n) = \sum_{i=1}^N \beta_{n|n,i}\delta_{\hat{X}_{n,i}}(dx_n)$ where

$$\beta_{n|n,i} = \frac{\widetilde{\beta}^+_{n|n}}{(\widetilde{\beta}/\widetilde{\alpha})^+_{n|n}}\frac{\widetilde{\beta}_{n|n,\varphi_n(i)}}{\widetilde{\alpha}_{n|n,\varphi_n(i)}}\mathbb{I}_{\mathbb{R}^+}(\widetilde{\beta}_{n|n,\varphi_n(i)}) + \frac{\widetilde{\beta}^-_{n|n}}{(\widetilde{\beta}/\widetilde{\alpha})^-_{n|n}}\frac{\widetilde{\beta}_{n|n,\varphi_n(i)}}{\widetilde{\alpha}_{n|n,\varphi_n(i)}}\mathbb{I}_{\mathbb{R}^-}(\widetilde{\beta}_{n|n,\varphi_n(i)}),$$

with

$$\widetilde{\beta}^+_{n|n} \triangleq \sum_{i=1}^N \widetilde{\beta}_{n|n,i}\mathbb{I}_{\mathbb{R}^+}(\widetilde{\beta}_{n|n,i}),$$

$$(\widetilde{\beta}/\widetilde{\alpha})^+_{n|n} \triangleq \sum_{i=1}^N \widetilde{\beta}_{n|n,\varphi_n(i)}/\widetilde{\alpha}_{n|n,\varphi_n(i)}\mathbb{I}_{\mathbb{R}^+}(\widetilde{\beta}_{n|n,\varphi_n(i)}),$$

$$\widetilde{\beta}^-_{n|n} \triangleq \sum_{i=1}^N \widetilde{\beta}_{n|n,i}\mathbb{I}_{\mathbb{R}^-}(\widetilde{\beta}_{n|n,i}), \quad \text{and}$$

$$(\widetilde{\beta}/\widetilde{\alpha})^-_{n|n} \triangleq \sum_{i=1}^N \widetilde{\beta}_{n|n,\varphi_n(i)}/\widetilde{\alpha}_{n|n,\varphi_n(i)}\mathbb{I}_{\mathbb{R}^-}(\widetilde{\beta}_{n|n,\varphi_n(i)}). \qquad \square$$

The computational complexity and memory requirements for this algorithm are in $\mathcal{O}(N)$ and independent of $n$.

*Batch ML algorithm.* At iteration $m$ of the gradient algorithm, the particle method described above is runned from time 0 to $T$ for the current parameter value $\theta_m$. This allows to obtain $\widehat{\pi}^{\theta_m}_{k|k-1}$ and $\widehat{w}^{\theta_m}_{k|k-1}$. These quantities are used to obtain a Monte Carlo estimate of the gradient (2.3) using

$$\widehat{(\partial l)}_{\theta_m}(Y^n) = \sum_{k=1}^{T} \frac{\sum_{i=1}^{N}((\partial\varepsilon)_{\theta_m}(\widetilde{X}_{k,i},Y_k) + N\beta_{k|k-1,i}\varepsilon_{\theta_m}(\widetilde{X}_{k,i},Y_k))}{\sum_{i=1}^{N}\varepsilon_{\theta_m}(\widetilde{X}_{k,i},Y_k)}.$$

This estimate is used to obtain $\theta_{m+1}$ using (2.1).

*Recursive ML algorithm.* At time $n$, we have obtained the sequence of parameters $\theta^n$ and $\widehat{\pi}^{\theta^n}_{n-1|n-1}$ and $\widehat{w}^{\theta^n}_{n-1|n-1}$. The sampling step of the particle method described above is runned with the parameter $\theta_n$. After this step, one obtains $\widehat{\pi}^{\theta^n}_{n|n-1}$ and $\widehat{w}^{\theta^n}_{n|n-1}$. This allows to obtain $\theta_{n+1}$ using

$$\theta_{n+1} = \theta_n + \gamma_n \left( \frac{\sum_{i=1}^{N}((\partial\varepsilon)_{\theta_n}(\widetilde{X}_{n,i},Y_n) + N\beta_{n|n-1,i}\varepsilon_{\theta_n}(\widetilde{X}_{n,i},Y_n))}{\sum_{i=1}^{N}\varepsilon_{\theta_n}(\widetilde{X}_{n,i},Y_n)} \right)$$

which is an approximation of (2.6). Then one runs the updating and resampling steps of the algorithm with the new parameter value to obtain $\widehat{\pi}^{\theta^{n+1}}_{n|n}$ and $\widehat{w}^{\theta^{n+1}}_{n|n}$.

*Remark* 1. *Resampling step.* The prediction and updating steps follow directly from a particle approximation of (2.9)–(2.10). The resampling step requires extra caution. The coefficients $\widetilde{\beta}_{n|n,\varphi_n(i)}$ are divided by $\widetilde{\alpha}_{n|n,\varphi_n(i)}$ so as to correct for the bias introduced by resampling particles according to $\widetilde{\alpha}_{n|n,i}$. It has also been designed so as to ensure that the masses of the positive and negative part of the signed measure $\widetilde{w}^{\theta^{n+1}}_{n|n}(dx_n)$ are preserved. In the algorithm presented above, we resample the particles according to $\widetilde{\alpha}_{n|n,i}$. However, one can expect a high variance for this scheme if the zones of important masses for $\pi^{\theta^{n+1}}_{n|n}$ are distinct from that of $w^{\theta^{n+1}}_{n|n}$. It is possible to consider "hybrid" resampling schemes taking into account both $\widetilde{\alpha}_{n|n,i}$ and $\widetilde{\beta}_{n|n,i}$ so as to perform resampling. Optimizing this resampling scheme is the subject of further research.

*Remark* 2. *Reprojection.* Typically the parameter $\theta = (\theta_1,\ldots,\theta_k)$ belongs to $\Theta$ which is a compact convex subset of $\mathbb{R}^k$. The parameter updating step, see equation (2.5), does not ensure that $\theta_{n+1} \in \Theta$ even if $\theta_n \in \Theta$. A standard approach in stochastic approximation to prevent divergence consists of reprojecting $\theta_{n+1}$ inside $\Theta$ whenever the value obtained through (2.5), say $\widetilde{\theta}_{n+1}$, does not belong to $\Theta$ (Ljung and Söderström (1987)). Typically one has $\Theta = \prod_{i=1}^{k}[\theta_{i,\min},\theta_{i,\max}]$ and the reprojection procedure simply consists of setting $\theta_{i,n+1} = \theta_{i,\min}$ if $\widetilde{\theta}_{i,n+1} < \theta_{i,\min}$ and $\theta_{i,n+1} = \theta_{i,\max}$ if $\widetilde{\theta}_{i,n+1} > \theta_{i,\max}$.

### 3.2 *Extensions*

We discuss here several extensions of this algorithm.

*Multidimensional parameter.* If $\theta = (\theta_1,\ldots,\theta_k)$ where $k > 1$, one needs to propagate $k$ derivatives $\nabla\pi^{\theta}_{n|n-1}(dx_n) = (\partial_{\theta_1}\pi^{\theta}_{n|n-1}(dx_n),\ldots,\partial_{\theta_k}\pi^{\theta}_{n|n-1}(dx_n))^{\mathrm{T}} = w^{\theta}_{n|n-1}(dx_n)$ then the updating step (2.6) of say the RML algorithm becomes

$$\begin{aligned}\theta_{n+1} &= \theta_n + \gamma_n \left( \int \varepsilon_{\theta_n}(x_n,Y_n)\pi^{\theta^n}_{n|n-1}(dx_n) \right)^{-1} \\ &\quad \times \left( \int (\nabla\varepsilon)_{\theta_n}(x_n,Y_n)\pi^{\theta^n}_{n|n-1}(dx_n) + \int \varepsilon_{\theta_n}(x_n,Y_n)w^{\theta^n}_{n|n-1}(dx_n) \right)\end{aligned}$$

where $(\nabla\varepsilon)_\theta(x_n,Y_n)=(\partial_{\theta_1}\varepsilon_\theta(x_n,Y_n),\ldots,\partial_{\theta_k}\varepsilon_\theta(x_n,Y_n))^{\mathrm{T}}$.

*Alternative importance densities.* Performance of particle filters can be improved significantly if one samples the particles according to an importance density $q_\theta(x_{n-1},Y_n,\cdot)$ (i.e. $\widetilde{X}_{n,i}\sim q_\theta(\widehat{X}_{n-1,i},Y_n,\cdot)$) different from the prior density $p_\theta(x_{n-1},\cdot)$ (with $p_\theta(x_{n-1},\cdot)\ll q_\theta(x_{n-1},Y_n,\cdot)$). In this case, to perform parameter estimation, one uses

$$\begin{aligned}&\log\left(\int\varepsilon_\theta(x_n,Y_n)\pi^\theta_{n|n-1}(dx_n)\right)\\&\quad=\log\left(\iint\frac{\varepsilon_\theta(x_n,Y_n)p_\theta(x_{n-1},x_n)}{q_\theta(x_{n-1},Y_n,x_n)}q_\theta(x_{n-1},Y_n,x_n)dx_n\pi^\theta_{n-1|n-1}(dx_{n-1})\right),\end{aligned}$$

where $\rho_\theta(x_{n-1},Y_n,x_n)=\frac{\varepsilon_\theta(x_n,Y_n)p_\theta(x_{n-1},x_n)}{q_\theta(x_{n-1},Y_n,x_n)}$ corresponds to the incremental unnormalized importance weight (Doucet *et al.* (2000); Liu and Chen (1998)). It follows that in the parameter updating step of the algorithm, see (2.5), one now uses

$$\begin{aligned}&\partial\log\left(\int\varepsilon_\theta(x_n,Y_n)\pi^\theta_{n|n-1}(dx_n)\right)\\&\quad=\partial\log\left(\iint\rho_\theta(x_{n-1},Y_n,x_n)q_\theta(x_{n-1},Y_n,x_n)dx_n\pi^\theta_{n-1|n-1}(dx_{n-1})\right)\\&\quad=\left(\iint\rho_\theta(x_{n-1},Y_n,x_n)q_\theta(x_{n-1},Y_n,x_n)dx_n\pi^\theta_{n-1|n-1}(dx_{n-1})\right)^{-1}\\&\qquad\times\left(\iint(\partial\rho)_\theta(x_{n-1},Y_n,x_n)q_\theta(x_{n-1},Y_n,x_n)dx_n\pi^\theta_{n-1|n-1}(dx_{n-1})\right.\\&\qquad\quad\left.+\iint\rho_\theta(x_{n-1},Y_n,x_n)\partial(q_\theta(x_{n-1},Y_n,x_n)dx_n\pi^\theta_{n-1|n-1}(dx_{n-1}))\right)\end{aligned}$$

where

$$\begin{aligned}\partial(q_\theta(x_{n-1},Y_n,x_n)dx_n\pi^\theta_{n-1|n-1}(dx_{n-1}))&=(\partial q)_\theta(x_{n-1},Y_n,x_n)dx_n\pi^\theta_{n-1|n-1}(dx_{n-1})\\&\quad+q_\theta(x_{n-1},Y_n,x_n)dx_nw^\theta_{n-1|n-1}(dx_{n-1}).\end{aligned}$$

The particle method to approximate these equations follows straightforwardly.

*Tracking a time-varying parameter.* If the parameter $\theta$ is actually time-varying but one does not have a dynamic model for its evolution, a standard approach in control and signal processing to "track" this parameter consists of using the recursive algorithm presented before using a fixed-step size $\gamma$ instead of a decreasing sequence $\gamma_n$. Bounds on the tracking errors can be established (Benveniste *et al.* (1990)). Selecting the step size is a difficult problem. If $\gamma$ is too large, the statistical fluctuations around the parameter are too large. If $\gamma$ is too small, the algorithm loses its tracking ability. In the context of adaptive filtering, adaptive step size schemes have been developed where the step size adapts itself automatically given the observations. It would be of interest to develop similar schemes in the context of general state-space models.

## 3.3 *Alternative approaches*

### 3.3.1 *Gradient methods*

We have investigated several alternative approaches to estimate $w^\theta_{n|n}$ using particle methods. The most natural one consists of using two distinct set of particles for $\pi^\theta_{n|n}$

and $w^{\theta}_{n|n}$. In this approach, one reinterprets (2.9) in a probabilistic way by using the fact that $(\partial p)_\theta$ can be rewritten as

$$(\partial p)_\theta = c_{p,\theta}(p'_\theta - p''_\theta) \tag{3.2}$$

where $c_{p,\theta} \in \mathbb{R}^+$, $p'_\theta$ and $p''_\theta$ are two probability transition kernels. There is actually an infinity of such decompositions; the most popular being the Hahn-Jordan decomposition where $p'_\theta \propto 0 \vee (\partial p)_\theta$ and $p''_\theta \propto -(0 \wedge (\partial p)_\theta)$. Another approach consists of rewriting

$$w^{\theta}_{n|n} = c_{n|n}(w'^{\theta}_{n|n} - w''^{\theta}_{n|n})$$

where $c_{n|n} \in \mathbb{R}^+$, $w'^{\theta}_{n|n}$ and $w''^{\theta}_{n|n}$ are two probability measures. Using the fact that

$$(\partial \varepsilon)_\theta = c_{\varepsilon,\theta}(\varepsilon'_\theta - \varepsilon''_\theta)$$

where $c_{\varepsilon,\theta} \in \mathbb{R}^+$, $\varepsilon'_\theta$ and $\varepsilon''_\theta$ are two probability densities and the expressions (2.9)–(2.10)–(3.2), it is trivial to obtain recursions for $c_{n|n}$, $w'^{\theta}_{n|n}$ and $w''^{\theta}_{n|n}$. One can then propagate $w'^{\theta}_{n|n}$ and $w''^{\theta}_{n|n}$ using particle methods. Though these methods sound attractive, they are more computationally intensive than the importance sampling approach. Moreover, the algorithms we proposed to approximate these measures appear numerically unstable. This approach deserves however further study.

### 3.3.2 *Non-gradient methods*

As discussed briefly in Section 1, several non-gradient approaches for parameter estimation using particle methods have been proposed in the literature. A simple approach consists of transforming fixed parameters into slowly time-varying ones (Higuchi (1997); Kitagawa (1998)) or keeping the original model and combining the particle filter with MCMC steps (Gilks and Berzuini (2001)). The first method modifies the original problem whereas the second one suffers from an accumulation of errors over time and eventually diverges if the dataset is large. Another approach consists of discretizing the parameter space and evaluating at each point the likelihood (Higuchi (2001)). This method is computationally expensive so one can only perform a coarse discretization of the parameter space. On the other hand, the main problem with gradient algorithms is their sensitivity to initial conditions. A pragmatic approach might consist of using one of the methods mentioned above to initialize a gradient method.

## 4. Applications

In all examples, we apply a simple standard particle method using the prior density as importance density. As discussed in Subsection 3.2, more elaborate algorithms could be developed using "tailored" importance densities. When Monte Carlo simulations are presented, new data have been simulated for each realization. We adopted $\gamma_n = \gamma_0 \cdot n^{-0.8}$ for the stepsize sequence.

### 4.1 *A nonlinear time series*

Let us consider the following model

$$\begin{aligned} X_{n+1} &= \cos(2\pi\phi X_n) + \sigma_v V_{n+1}, \qquad X_0 \sim \mathcal{N}(0,2), \\ Y_n &= X_n + \sigma_w W_n \end{aligned}$$

where $V_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$ and $W_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$ are two mutually independent sequences of independent identically distributed (i.i.d.) Gaussian random variables. We are interested in estimating the parameter $\theta \triangleq (\phi, \sigma_v, \sigma_w)$ where $\Theta = (0,1) \times (0,M) \times (0,M)$ where $M = 100$. Methods combining MCMC steps with particle filters cannot be used in this case as the conditional distribution of $\phi$ given the data and the hidden process is not in the exponential family.

The true parameter values are $\theta^* = (0.5, 1.0, 1.0)$. The RML algorithm was implemented using $N = 1000$ and $N = 10000$ particles and the initial parameter $\theta_0$ was randomly selected. In the updating step of the algorithm, the parameter was reprojected inside $\Theta$ whenever necessary. In our 50 simulations, the algorithm appeared trapped in a local maximum located around $(0.71, 0.99, 1.10)$ in 12 simulations. For the remaining 38 simulations, the algorithm converges towards a value located around $\theta^*$. In this case, as $N$ increases, $\|\widehat{\theta} - \theta^*\|$ decreases; see Fig. 1 for an example.

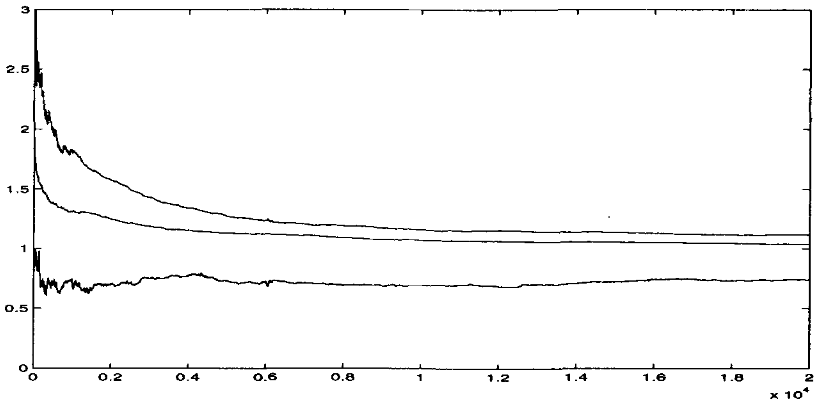

Fig. 1. Sequence of parameter estimates $\theta_n = (\sigma_{v,n}, \sigma_{w,n}, \phi_n)$ for $N = 10000$. From top to bottom: $\sigma_{v,n}$, $\sigma_{w,n}$ and $\phi_n$.

### 4.2 *Another nonlinear time series*

We consider the following nonlinear model (Kitagawa (1996, 1998))

$$X_{n+1} = \theta_3 X_n + \frac{\theta_4 X_n}{1 + X_n^2} + \theta_5 \cos(1.2(n+1)) + \theta_1 V_{n+1}, \qquad X_0 \sim \mathcal{N}(0,2),$$
$$Y_n = \theta_6 X_n^2 + \theta_2 W_n$$

where $V_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$ and $W_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$ are two mutually independent sequences of i.i.d. Gaussian random variables. We are interested in estimating $\theta \triangleq (\theta_1, \dots, \theta_6)$.

The true parameter values are $\theta^* = (1, \sqrt{10}, 0.5, 25, 8, 0.05)$. This problem is extremely complex. The likelihood function is highly multimodal. It is important in this example to initialize the algorithm properly. The batch ML algorithm was implemented using $N = 1000$ particles. With $\theta_0 = (0.5, 2, 1, 15, 4, 0.1)$, the algorithm converges towards to the neighborhood of the true parameter. We present in Fig. 2 the sequence of parameter estimates $(\theta_4, \theta_5)$. Clearly, there is a significant variance for the gradient estimate associated to $\theta_4$.

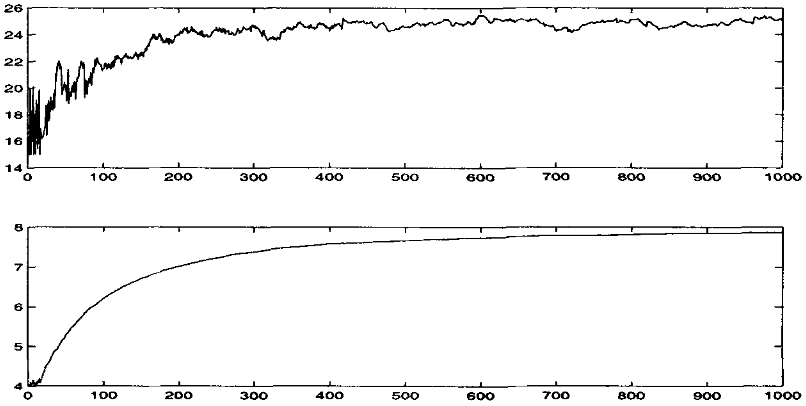

Fig. 2. Sequence of parameter estimates for batch ML $\theta_{4,m}$ (top) and $\theta_{5,m}$ (bottom) for $N = 1000$.

### 4.3 *Stochastic volatility model*

Let us consider the following model (Pitt and Shephard (1999))

$$X_{n+1} = \phi X_n + \sigma V_{n+1}, \qquad X_0 \sim \mathcal{N}\left(0, \frac{\sigma^2}{1-\phi^2}\right)$$
$$Y_n = \beta \exp(X_n/2) W_n$$

where $V_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$ and $W_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$ are two mutually independent sequences, independent of the initial state $X_0$. We are interested in estimating the parameter $\theta \triangleq (\phi, \sigma, \beta)$ where $\Theta = (-1,1) \times (0, M) \times (0, M)$ where $M = 100$.

The true parameter values are $\theta^* = (0.8, 0.5, 1.0)$. We first implemented the RML algorithm using $N = 10000$ particles. As in the previous example, the parameter was reprojected inside $\Theta$ whenever necessary. In our 50 simulations, our algorithm did not appear sensitive to the initialization $\theta_0$. It converged to a value $\widehat{\theta}$ in the neighborhood of $\theta^*$; the larger $N$ the smaller $\|\widehat{\theta} - \theta^*\|$; see Fig. 3 for an example.

We then apply our batch ML method to the pound/dollar daily exchange rates; see (Durbin and Koopman (2000)) and (Harvey *et al.* (1994)). This time series consists of 945 data. We apply the batch ML algorithm for $M = 1000$ iterations with $N = 10000$ particles; see Fig. 4. Our results for the ML estimate are consistent with recent results obtained in (Durbin and Koopman (2000)). We obtain $\widehat{\theta}_{ML} = (0.968, 0.188, 0.638)$ whereas the estimate in (Durbin and Koopman (2000)) is $\widehat{\theta}_{ML} = (0.973, 0.173, 0.634)$.

## 5. Discussion

In this paper, we have proposed two original gradient type algorithms to perform batch and recursive parameter estimation in general state-space models. These gradient algorithms require being able to estimate the derivatives of the filtering distribution with respect to the unknown parameters. We have proposed a simple method based on importance sampling resampling to estimate this quantity and have demonstrated the utility of our algorithms for nonlinear state-space models.

In the context of recursive parameter estimation, our algorithm requires a substantial amount of data to converge. While this is not a problem in most engineering

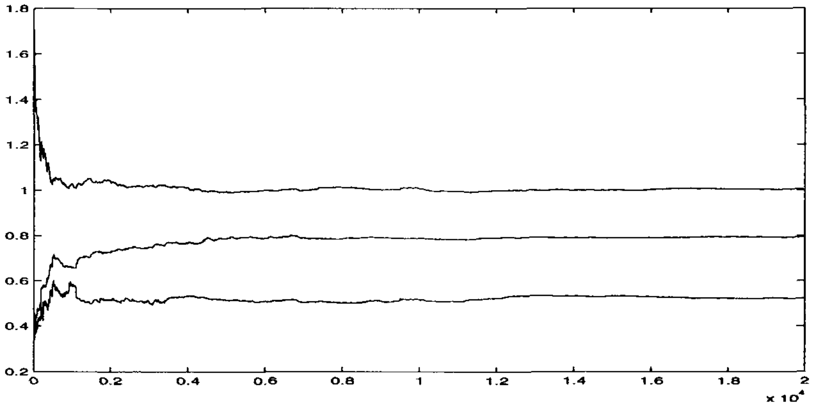

Fig. 3. Sequence of parameter estimates for RML $\theta_n = (\beta_n, \phi_n, \sigma_n)$ for $N = 10000$. From top to bottom: $\beta_n$, $\phi_n$ and $\sigma_n$.

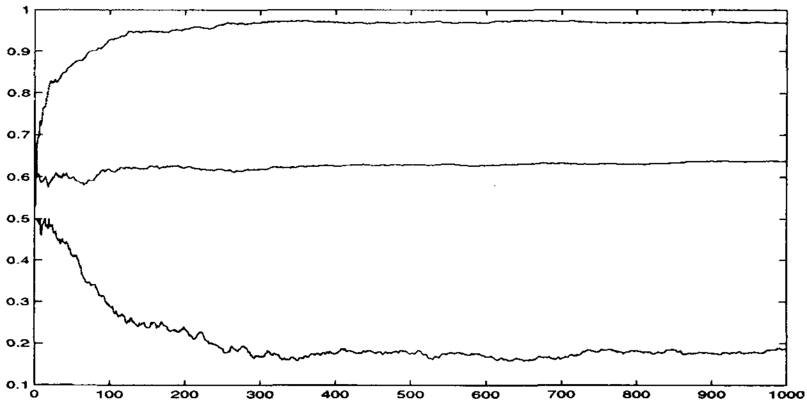

Fig. 4. Sequence of parameter estimates for batch ML $\theta_m = (\beta_m, \phi_m, \sigma_m)$ for $N = 10000$. From top to bottom: $\beta_m$, $\phi_m$ and $\sigma_m$.

applications, this could be a severe limitation in other contexts. To overcome this problem, it is necessary to develop alternative gradient estimation and variance reduction methods. Finally, from a theoretical viewpoint, non standard stochastic approximation results presented by Tadić (2000) need to be used to prove convergence of the algorithms.

## Acknowledgements

The authors are grateful to the reviewers for their comments which allow to significantly improve this paper.

## References

Andrieu, C., De Freitas, J. F. G. and Doucet, A. (1999). Sequential MCMC for Bayesian model selection, *Proceedings of the IEEE Workshop on Higher Order Statistics*, 130–134.

Benveniste, A., Métivier, M. and Priouret, P. (1990). *Adaptive Algorithms and Stochastic Approximation*, Springer, New York.

Cérou, F., LeGland, F. and Newton, N. J. (2001). Stochastic particle methods for linear tangent equations, *Optimal Control and PDE's - Innovations and Applications* (eds. J. Menaldi, E. Rofman and A. Sulem), 231–240, IOS Press, Amsterdam.

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

Doucet, A., de Freitas, J. F. G. and Gordon, N. J. (eds.) (2001). *Sequential Monte Carlo Methods in Practice*, Springer, New York.

Durbin, J. and 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. Roy. Statist. Soc. Ser. B*, **62**, 3–56.

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

Gordon, N. J., Salmond, D. J. and Smith, A. F. M. (1993). Novel approach to nonlinear non-Gaussian Bayesian state estimation, *IEE Proceedings F*, **140**, 107–113.

Harvey, A. C., Ruiz, E. and Shephard, N. (1994). Multivariate stochastic variance models, *Rev. Econom. Stud.*, **61**, 247–264.

Higuchi, T. (1997). Monte Carlo filter using the genetic algorithm operators, *J. Statist. Comput. Simulation*, **59**, 1–23.

Higuchi, T. (2001). Evolutionary time series model with parallel computing, *Proceedings of the 3rd Japan-US Seminar on Statistical Time Series Analysis*, 183–190.

Iba, Y. (2001). Population Monte Carlo algorithms, *Transactions of the Japanese Society for Artificial Intelligence*, **16**, 279–286.

Kitagawa, G. (1996). Monte Carlo filter and smoother for non-Gaussian nonlinear state space models, *J. Comput. Graph. Statist.*, **5**, 1–25.

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

Kitagawa, G. and Gersch, W. (1996). *Smoothness Priors Analysis of Time Series*, Lecture Notes in Statist., Vol. 116, Springer, New York.

LeGland, F. and Mevel, L. (1997). Recursive identification in hidden Markov models, *Proceedings of 36th IEEE Conference on Decision and Control*, 3468–3473.

Liu, J. S. and Chen, R. (1998). Sequential Monte Carlo methods for dynamic systems, *J. Amer. Statist. Assoc.*, **93**, 1032–1043.

Liu, J. and West, M. (2001). Combined parameter and state estimation in simulation-based filtering, *Sequential Monte Carlo Methods in Practice* (eds. A. Doucet, J. F. G. de Freitas and N. J. Gordon), 197–223, Springer, New York.

Ljung, L. and Söderström, T. (1987). *Theory and Practice of Recursive Identification*, MIT Press, Cambridge, Massachusetts.

Pitt, M. K. and Shephard, N. (1999). Filtering via simulation: Auxiliary particle filter, *J. Amer. Statist. Assoc.*, **94**, 590–599.

Tadić, V. B. (2000). Asymptotic analysis of stochastic approximation algorithms under violated Kushner-Clark conditions, *Proceedings of 39th IEEE Conference on Decision and Control*, 2875–2880.

Tadić, V. B. and Doucet, A. (2002). Exponential forgetting and geometric ergodicity in state space models, *Proceedings of 41th IEEE Conference on Decision and Control*, 2231–2235.