Ann Inst Stat Math (2010) 62:61–89
DOI 10.1007/s10463-009-0236-2

Smoothing algorithms for state–space models

Mark Briers · Arnaud Doucet · Simon Maskell

Received: 5 June 2008 / Revised: 27 January 2009 / Published online: 9 June 2009

Abstract Two-filter smoothing is a principled approach for performing optimal smoothing in non-linear non-Gaussian state–space models where the smoothing distributions are computed through the combination of 'forward' and 'backward' time filters. The 'forward' filter is the standard Bayesian filter but the 'backward' filter, generally referred to as the backward information filter, is not a probability measure on the space of the hidden Markov process. In cases where the backward information filter can be computed in closed form, this technical point is not important. However, for general state–space models where there is no closed form expression, this prohibits the use of flexible numerical techniques such as Sequential Monte Carlo (SMC) to approximate the two-filter smoothing formula. We propose here a *generalised* two-filter smoothing formula which only requires approximating probability distributions and applies to any state–space model, removing the need to make restrictive assumptions used in previous approaches to this problem. SMC algorithms are developed to implement this generalised recursion and we illustrate their performance on various problems.

Keywords Sequential Monte Carlo · Two-filter smoothing · State–space models · Rao-Blackwellisation · Non-linear diffusion · Parameter estimation

M. Briers
Information Engineering Division, Cambridge University, Trumpington Street,
Cambridge CB2 1PZ, UK

A. Doucet (✉)
The Institute of Statistical Mathematics, 4-6-7 Minami-Azabu, Minato-ku,
Tokyo 106-8569, Japan
e-mail: arnaud@ism.ac.jp

M. Briers · S. Maskell
QinetiQ Ltd, St Andrews Road, Malvern, Worcester WR14 3PS, UK

1 Introduction and motivation

1.1 General state–space models

State–space models are ubiquitous in statistics but also in econometrics, signal processing and robotics; see Doucet et al. (2001) for numerous applications. Formally a state–space model is defined as follows: Let $\{X_t\}_{t\in\mathbb{N}}$ be a discrete-time $\mathcal{X}$-valued Markov process defined by its initial probability density $X_1 \sim \mu(\cdot)$ and for $t \geq 2$

$$X_t|\,(X_{t-1} = x_{t-1}) \sim f(\cdot|x_{t-1}). \tag{1}$$

We do not have direct access to this (hidden) process $\{X_t\}_{t\in\mathbb{N}}$ but only to the $\mathcal{Y}$-valued observation process $\{Y_t\}_{t\in\mathbb{N}}$ which is such that, conditional upon $\{X_t\}_{t\in\mathbb{N}}$, the observations are statistically independent and distributed according to

$$Y_t|\,(X_t = x_t) \sim g(\cdot|x_t). \tag{2}$$

We assume that $\mu(\cdot)$, $f(\cdot|x_{t-1})$ and $g(\cdot|x_t)$ are probability densities with respect to some dominating measure (e.g. Lebesgue).

We are interested in the optimal estimation of the states given a sequence of observations. For any general sequence $\{z_k\}$ we write $z_{i:j} = \left(z_i, z_{i+1}, \ldots, z_j\right)$. Recursively computing (in time) the sequence of posterior densities $\{p(x_t|y_{1:t})\}$ is known as the filtering problem and has generated a huge literature over the past 40 years. A related and important problem addressed within this paper is the fixed-interval smoothing problem, which consists of computing the sequence of posterior densities $\{p(x_t|y_{1:T})\}$ for $t \in \{1, \ldots, T\}$. We provide novel computational methods to solve the fixed-interval smoothing problem, and to provide realizations from the joint smoothing density $p(x_{1:T}|y_{1:T})$.

1.2 Filtering and smoothing recursions

Using (1) and (2) the joint posterior density $p(x_{1:t}|y_{1:t})$ is simply given by

$$p(x_{1:t}|y_{1:t}) \propto \mu(x_1) \prod_{k=2}^{t} f(x_k|x_{k-1}) \prod_{k=1}^{t} g(y_k|x_k) \tag{3}$$

where '$\propto$' stands for 'proportional to'. To obtain the (marginal) filtering density $p(x_t|y_{1:t})$ one can simply marginalize this expression over $x_{1:t-1}$. Similarly, to obtain the (marginal) smoothing density $p(x_t|y_{1:T})$, one can also marginalize $p(x_{1:T}|y_{1:T})$ over $x_{1:t-1}$ and $x_{t+1:T}$, with $p(x_{1:T}|y_{1:T})$ being defined using an expression similar to (3).

However, it is often algorithmically more convenient to consider operations that result in sequential (in time) computations. For example, determination of the filtering density can be performed through the following *prediction-update* recursion

$$p(x_t|y_{1:t-1}) = \int f(x_t|x_{t-1})p(x_{t-1}|y_{1:t-1})dx_{t-1}, \tag{4}$$

$$p(x_t|y_{1:t}) \propto g(y_t|x_t)p(x_t|y_{1:t-1}). \tag{5}$$

Likewise, algorithmically appealing recursions exist for calculating the smoothing density. One such technique is the *forward filtering-backward smoothing recursion* presented in Kitagawa (1987). This recursion shows that, once we have computed all the predicted and filtered densities $\{p(x_t|y_{1:t-1})\}$ and $\{p(x_t|y_{1:t})\}$ over the interval $\{1, \ldots, T\}$, then it is possible to execute a backward recursion to obtain $\{p(x_t|y_{1:T})\}$ using

$$p(x_t|y_{1:T}) = p(x_t|y_{1:t}) \int \frac{p(x_{t+1}|y_{1:T}) f(x_{t+1}|x_t)}{p(x_{t+1}|y_{1:t})} \mathrm{d}x_{t+1}. \tag{6}$$

An alternative approach that allows one to compute $\{p(x_t|y_{1:T})\}$ is through the *two-filter smoothing formula*; see Bresler (1986) or Kitagawa (1994). In this approach, one combines the output of two (independent) filters: the standard (forward) filter given by (4)–(5) and the so-called backward information filter calculating $p(y_{t:T}|x_t)$. This information filter satisfies

$$\begin{aligned} p(y_{t:T}|x_t) &= \int p(y_t, y_{t+1:T}, x_{t+1}|x_t)\mathrm{d}x_{t+1} \\ &= \int p(y_{t+1:T}|x_{t+1}) f(x_{t+1}|x_t) g(y_t|x_t)\mathrm{d}x_{t+1}. \end{aligned} \tag{7}$$

The prediction and the backward information filter are then combined to give the required density $p(x_t|y_{1:T})$ using

$$p(x_t|y_{1:T}) \propto p(x_t|y_{1:t-1}) p(y_{t:T}|x_t). \tag{8}$$

1.3 Motivation

Except for very simple cases (for example, a finite state–space HMM or a linear Gaussian state–space model), it is impossible to compute a closed-form expression for the filtering and smoothing densities. This has seriously limited the use of general state–space models for many years. However, the recent introduction of Sequential Monte Carlo (SMC) methods (also known as particle filtering methods) provide numerical solutions to filtering problems using non-linear and non-Gaussian state–space models; see Doucet et al. (2001) for a review. Broadly speaking, SMC methods are a class of importance sampling and resampling methods to approximate the joint posterior densities $\{p(x_{1:t}|y_{1:t})\}$. This provides approximations of the form

$$\widehat{p}(\mathrm{d}x_{1:t}|y_{1:t}) = \sum_{i=1}^{N} W_t^{(i)} \delta_{X_{1:t}^{(i)}}(\mathrm{d}x_{1:t}) \tag{9}$$

where $\delta_{x_0}(\mathrm{d}x)$ denotes the Dirac delta mass located at x_0, $W_t^{(i)} > 0$, $\sum_{i=1}^N W_t^{(i)} = 1$ and $\{X_{1:t}^{(i)}\}$ are N random samples, named particles. The main advantages of these popular techniques are that they do not rely on any functional approximation of the posterior distributions of interest, and are guaranteed to converge as $N \to \infty$ towards the distributions of interest under minimal assumptions; see Del Moral (2004). At time T, we obtain a Monte Carlo approximation of $p\left(x_{1:T}|y_{1:T}\right)$ of the form (9) from which we can easily approximate the marginals $p\left(x_t|y_{1:T}\right)$. However the performance of this direct method is extremely poor as soon as T is large because of the so-called degeneracy problem discussed in Doucet et al. (2000). Roughly speaking, because of the successive resampling steps of the SMC algorithm, the marginals $p\left(x_t|y_{1:T}\right)$ are approximated by one unique particle as soon as $T-t$ is large; only the marginals for which $T-t$ is 'small', say less than 20–50 for a reasonable number of particles, will be well approximated. Consequently, alternative techniques have been developed to solve the fixed-interval smoothing problem which we review briefly here.

The simplest possible approach proposed in Kitagawa and Sato (2001) relies on the fact that, for hidden Markov models with "good" forgetting properties, we have

$$p\left(x_{1:t}|\,y_{1:T}\right) \approx p\left(x_{1:t}|\,y_{1:\min(t+\Delta,T)}\right) \tag{10}$$

for Δ large enough; that is observations collected at times $k > t + \Delta$ do not bring any additional information about the states $X_{1:t}$. This suggests a very simple scheme —simply do not update the estimate of X_t after time $t+\Delta$. This algorithm is trivial to implement but the main practical problem is that we typically do not know Δ. Hence we need to replace Δ with an estimate of it denoted L. If we select $L < \Delta$, then $p(x_{1:t}|\,y_{1:\min(t+L,T)})$ is a poor approximation of $p\left(x_{1:t}\;|\,y_{1:T}\right)$. If we select a large values of L to ensure that $L \geq \Delta$ then the degeneracy problem remains substantial. Unfortunately, automatic selection of L is difficult (and, of course, for some poorly mixing models Δ is so large that this approach is impractical).

It is also possible to develop a Markov Chain Monte Carlo (MCMC) algorithm to sample from the joint density $p(x_{1:T}|y_{1:T})$ and, hence, from the marginal smoothing densities $\{p(x_t|y_{1:T})\}$. However, the determination of an efficient proposal density for MCMC is difficult for general state–space models where the target posterior density can be highly multi-modal; see Godsill et al. (2004) for such examples. An alternative to the MCMC approach consists of using an SMC implementation of the forward filtering-backward smoothing recursion (6). This has been proposed in Doucet et al. (2000) to compute $\{p(x_t|y_{1:T})\}$ by substituting the marginal in x_t of the SMC approximation (9) into (6) to yield

$$\begin{aligned}\widehat{p}(\mathrm{d}x_t|y_{1:T}) &= \sum_{i=1}^N W_t^{(i)}\left[\sum_{j=1}^N W_{t+1|T}^{(j)} \frac{f\left(X_{t+1}^{(j)}|X_t^{(i)}\right)}{\left[\sum_{l=1}^N W_t^{(l)} f\left(X_{t+1}^{(i)}|X_t^{(l)}\right)\right]}\right]\delta_{X_t^{(i)}}(\mathrm{d}x_t)\\ &= \sum_{i=1}^N W_{t|T}^{(j)}\delta_{X_t^{(i)}}(\mathrm{d}x_t).\end{aligned} \tag{11}$$

A similar idea was extended in Godsill et al. (2004) to sample from $p(x_{1:T}|y_{1:T})$ using the related forward filtering-backward sampling formula.

Equation (11) yields a consistent approximation as $N \to \infty$ under minimal assumptions. This approach is more efficient than the direct SMC appproximation of the smoothing densities outlined at the beginning of this section but its performance can still degrade significantly as T increases. The problem of such methods is that it provides a Monte Carlo approximation of the smoothed distributions which rely on the same random samples $\{X_t^{(i)}\}$ used to approximate the filtered distributions; it only reweights these samples. Hence, if $p(x_t|y_{1:t})$ and $p(x_t|y_{1:T})$ have high probability masses in distinct regions of the space, then the Monte Carlo approximation will have a high variance for reasonable values of N. Moreover, the subsequent approximations of $p(x_k|y_{1:T})$ for $k \leq t$ will also be poor.

In this paper, we propose a generalised version of the two-filter smoothing formula which allows us to propose novel SMC algorithms to perform smoothing. In the generalized two-filter formula, the smoothing density is computed as the combination of two (independent) probability densities on the state–space of the hidden Markov process: the standard forward-time filter and a modified backward-time filter. It is possible to employ two 'standard' SMC algorithms to approximate these filters with good practical and theoretical properties under mild assumptions. Consequently, the resulting approximation does not suffer from poor performance as T increases. This is demonstrated in the simulation section where very significant performance improvements over the forward filtering-backward smoothing formula are reported.

1.4 Organization of the paper and contributions

In Sect. 2, we propose a generalised version of the two-filter smoothing formula where the smoothing densities are computed through a combination of the optimal filter $p(x_t|y_{1:t-1})$ and an (auxiliary) probability density $\widetilde{p}(x_t|y_{t:T})$ in argument x_t, which is computed backward in time. The main advantage of this generalised formulation over the standard formulation involving $p(y_{t:T}|x_t)$ is that it allows us to use standard approximation techniques to approximate $\widetilde{p}(x_t|y_{t:T})$, a probability density by construction. The definition of $\{\widetilde{p}(x_t|y_{1:T})\}$ relies on the introduction of a sequence of artificial probability densities $\{\gamma_t(x_t)\}$ where $t = 1, \ldots, T$. We discuss several appealing choices for these densities.

In Sect. 3, we discuss an SMC approximation of the generalised two-filter formula and an SMC algorithm to sample from $p(x_{1:T}|y_{1:T})$, which is an alternative to the forward filtering-backward sampling formula. In Sect. 4, we focus on the important class of conditionally linear Gaussian state–space models and present an efficient Rao-Blackwellized SMC algorithm to perform optimal smoothing. We also discuss an extension to partially observed linear Gaussian models. Finally, in Sect. 5, we present various applications of these SMC algorithms to perform smoothing for a non-linear time series model and a non-linear diffusion process, parameter estimation using the EM for a stochastic volatility model and blind deconvolution for a seismic signal model.

Note that we have deliberately focused on SMC approximations in this article but it is straightforward to derive functional approximations such as the Extended/Unscented Kalman filter for the generalised-two filter formula; these algorithms are detailed in Briers et al. (2004).

2 Generalized two-filter smoothing recursion

2.1 Artificial distributions

Let us consider a sequence of probability densities $\{\gamma_t(x_t)\}$ where $t = 1, \ldots, T$ which are defined such that

$$\text{if } p(y_{t:T}|x_t) > 0 \quad \text{then } \gamma_t(x_t) > 0. \tag{12}$$

Define for $t = T$

$$\widetilde{p}(x_T|y_T) = \frac{\gamma_T(x_T)g(y_T|x_T)}{\widetilde{p}(y_T)} \tag{13}$$

with

$$\widetilde{p}(y_T) = \int \gamma_T(x_T)g(y_T|x_T)\mathrm{d}x_T, \tag{14}$$

where $\widetilde{p}(x_T|y_T)$ as defined in (13) is a probability density by construction.

Further, let us define the sequence of artificial probability densities for $t \in \{2, \ldots, T-1\}$

$$\widetilde{p}(x_{t:T}|y_{t:T}) = \frac{\gamma_t(x_t)\prod_{k=t+1}^{T} f(x_k|x_{k-1})\prod_{k=t}^{T} g(y_k|x_k)}{\widetilde{p}(y_{t:T})}, \tag{15}$$

where

$$\widetilde{p}(y_{t:T}) = \int \cdots \int \gamma_t(x_t) \prod_{k=t+1}^{T} f(x_k|x_{k-1}) \prod_{k=t}^{T} g(y_k|x_k) dx_{t:T}. \tag{16}$$

Proposition 1 *For any $t \in \{1, \ldots, T\}$ we have*

$$p(y_{t:T}|x_t) = \widetilde{p}(y_{t:T})\frac{\widetilde{p}(x_t|y_{t:T})}{\gamma_t(x_t)} \tag{17}$$

where

$$\widetilde{p}(x_t|y_{t:T}) = \int \cdots \int \widetilde{p}(x_{t:T}|y_{t:T})dx_{t+1:T}. \tag{18}$$

Proof For $t = T$ the result is obvious from (13). For $t \in \{1, \ldots, T-1\}$ we have:

$$\begin{aligned}
p(y_{t:T}|x_t) &= \int \cdots \int p(y_{t:T}, x_{t+1:T}|x_t)\mathrm{d}x_{t+1:T} \\
&= \int \cdots \int p(x_{t+1:T}|x_t)p(y_{t:T}|x_{t:T})\mathrm{d}x_{t+1:T} \\
&= \int \cdots \int \prod_{k=t+1}^{T} f(x_k|x_{k-1}) \prod_{k=t}^{T} g(y_k|x_k)\mathrm{d}x_{t+1:T} \\
&= \int \cdots \int \frac{\gamma_t(x_t)}{\gamma_t(x_t)} \prod_{k=t+1}^{T} f(x_k|x_{k-1}) \prod_{k=t}^{T} g(y_k|x_k)\mathrm{d}x_{t+1:T} \\
&= \widetilde{p}(y_{t:T}) \int \cdots \int \frac{\widetilde{p}(x_{t:T}|y_{t:T})}{\gamma_t(x_t)}\mathrm{d}x_{t+1:T} \\
&= \widetilde{p}(y_{t:T}) \frac{\widetilde{p}(x_t|y_{t:T})}{\gamma_t(x_t)}. \qquad (19)
\end{aligned}$$

□

2.2 Backward recursion and generalized two-filter formula

We now present a backward recursion allowing us to compute $\widetilde{p}(x_t|y_{t:T})$ from $\widetilde{p}(x_{t+1}|y_{t+1:T})$. Although this recursion will not be of any direct practical use when deriving the SMC approximation of $\widetilde{p}(x_t|y_{t:T})$, it emphasises the similarities and differences with the traditional prediction-update recursion given in (4)–(5). Moreover, it is useful when deriving functional approximations using the Extended/Unscented Kalman filter.

Proposition 2 *For any $t \in \{1, \ldots, T-1\}$, the following backward prediction-update recursion holds*

$$\widetilde{p}(x_t|y_{t+1:T}) := \int \widetilde{p}(x_{t+1}|y_{t+1:T}) \frac{f(x_{t+1}|x_t)\gamma_t(x_t)}{\gamma_{t+1}(x_{t+1})}\mathrm{d}x_{t+1}, \qquad (20)$$

$$\widetilde{p}(x_t|y_{t:T}) = \frac{g(y_t|x_t)\widetilde{p}(x_t|y_{t+1:T})}{\int g(y_t|x_t)\widetilde{p}(x_t|y_{t+1:T})\mathrm{d}x_t} \qquad (21)$$

if $\int \widetilde{p}(x_t|y_{t+1:T})\mathrm{d}x_t < \infty$.

Proof The term $\widetilde{p}(x_t|y_{t+1:T})$ is *defined* using (20) so there is nothing to prove. For the update step, we note that

$$
\begin{aligned}
g(y_t|x_t)\widetilde{p}(x_t|y_{t+1:T}) &= g(y_t|x_t)\int \widetilde{p}(x_{t+1}|y_{t+1:T})\frac{f(x_{t+1}|x_t)\gamma_t(x_t)}{\gamma_{t+1}(x_{t+1})}\mathrm{d}x_{t+1}\\
&= g(y_t|x_t)\int \widetilde{p}(x_{t+1:T}|y_{t+1:T})\frac{f(x_{t+1}|x_t)\gamma_t(x_t)}{\gamma_{t+1}(x_{t+1})}\mathrm{d}x_{t+1:T}\\
&= \frac{\widetilde{p}(y_{t+1:T})}{\widetilde{p}(y_{t:T})}\int \widetilde{p}(x_{t+1:T}|y_{t+1:T})\frac{\widetilde{p}(x_{t:T}|y_{t:T})}{\widetilde{p}(x_{t+1:T}|y_{t+1:T})}\mathrm{d}x_{t+1:T}.
\end{aligned}
$$

Hence (21) follows. □

Remark Note that $\widetilde{p}(x_t|y_{t+1:T})$ defined in Proposition 2 is *not* a probability density if $\gamma_{t+1}(x_{t+1}) \neq \int f(x_{t+1}|x_t)\gamma_t(x_t)\mathrm{d}x_t$. A sufficient condition to ensure $\int \widetilde{p}(x_t|y_{t+1:T})\mathrm{d}x_t < \infty$ consists of selecting the artificial densities $\{\gamma_t(x_t)\}$ such that

$$\frac{f(x_{t+1}|x_t)}{\gamma_{t+1}(x_{t+1})} < C < +\infty \tag{22}$$

for any $(x_t, x_{t+1}) \in \mathcal{X} \times \mathcal{X}$; i.e. $\gamma_{t+1}(x_{t+1})$ needs to have thicker tails than $f(x_{t+1}|x_t)$ for any x_t. In the next section we will see that, although the SMC approximation does not directly rely on this backward prediction-update formula and (22) is also required to ensure good performance of this numerical approximation.

Having defined the backward filter $\widetilde{p}(x_t|y_{t:T})$, we are now in position to present the generalised two-filter smoothing formula. Its proof follows directly from the standard two-filter formula (8) and Proposition 1.

Proposition 3 *For any* $t \in \{2, \ldots, T-1\}$, *we have*

$$
\begin{aligned}
p(x_t|y_{1:T}) &\propto \frac{p(x_t|y_{1:t-1})\widetilde{p}(x_t|y_{t:T})}{\gamma_t(x_t)}\\
&\propto \frac{\int f\,(x_t|\,x_{t-1})\,p(x_{t-1}|y_{1:t-1})\mathrm{d}x_{t-1}.\widetilde{p}(x_t|y_{t:T})}{\gamma_t(x_t)}
\end{aligned}
\tag{23}
$$

and for $t=1$

$$p(x_1|y_{1:T}) \propto \frac{\mu\,(x_1)\,\widetilde{p}(x_1|y_{1:T})}{\gamma_1(x_1)}. \tag{24}$$

We also detail the following decomposition of the joint density $p(x_{1:T}|y_{1:T})$, which will be utilized in Sect. 3.

Proposition 4 *For any* $t \in \{2, \ldots, T-1\}$, *we have*

$$p(x_{1:T}|y_{1:T}) = p(x_t|y_{1:T})p(x_{1:t-1}|y_{1:t-1}, x_t)p(x_{t+1:T}|y_{t+1:T}, x_t), \tag{25}$$

where

$$p(x_{1:t-1}|y_{1:t-1}, x_t) = \prod_{k=1}^{t-1} p(x_k|y_{1:k}, x_{k+1}) \tag{26}$$

$$= \prod_{k=1}^{t-1} \frac{f(x_{k+1}|x_k)p(x_k|y_{1:k})}{p(x_{k+1}|y_{1:k})}, \tag{27}$$

and

$$p(x_{t+1:T}|y_{t+1:T}, x_t) = \prod_{k=t+1}^{T} p(x_k|y_{k:T}, x_{k-1}) \tag{28}$$

$$= \prod_{k=t+1}^{T} \frac{f(x_k|x_{k-1})p(y_{k:T}|x_k)}{p(y_{k:T}|x_{k-1})} \tag{29}$$

where

$$\frac{f(x_k|x_{k-1})p(y_{k:T}|x_k)}{p(y_{k:T}|x_{k-1})} \propto \frac{\gamma_{k-1}(x_{k-1})f(x_k|x_{k-1})\widetilde{p}(x_k|y_{k:T})}{\gamma_k(x_k)\widetilde{p}(x_{k-1}|y_{k:T})}. \tag{30}$$

2.3 Choice of artificial distributions

Although theoretically any sequence of artificial densities $\{\gamma_t(x_t)\}$ can be used as long as they satisfy (12), choice of these densities will have a significant impact on the performance of the SMC procedures used to implement the generalised two-filter formula. SMC algorithms approximate $p(x_t|y_{1:t}) \propto p(x_t)p(y_{1:t}|x_t)$ where $p(x_t)$ is the marginal prior density of X_t and $\widetilde{p}(x_t|y_{t:T}) \propto \gamma_t(x_t)p(y_{t:T}|x_t)$ by two clouds of particles located in regions of high probability masses of these densities and these densities are combined through (23) to compute $p(x_t|y_{1:T})$. If these densities have their regions of high probability masses significantly disjoint, then the resulting approximation combining both cannot be expected to perform well. The only degree of freedom being $\gamma_t(x_t)$, it is sensible to select $\gamma_t(x_t)$ as the marginal prior density $p(x_t)$; that is $\gamma_t(x_t)$ is defined recursively through

$$\gamma_t(x_t) := \int \gamma_{t-1}(x_{t-1})f(x_t|x_{t-1})\mathrm{d}x_{t-1} \tag{31}$$

with $\gamma_1(x_1) = \mu(x_1)$. This is the choice adopted in Bresler (1986). In this case $\widetilde{p}(x_t|y_{t+1:T})$ defined in (20) is a probability density as (31) ensures that

$$\gamma(x_t|x_{t+1}) := \frac{f(x_{t+1}|x_t)\gamma_t(x_t)}{\gamma_{t+1}(x_{t+1})} \tag{32}$$

is a (generally time inhomogeneous) backward Markov kernel. Note that (32) is also a backward Markov kernel for any choice of $\gamma_1(x_1)$ if (31) is satisfied. Selecting $\gamma_1(x_1) \neq \mu(x_1)$ can be convenient. Assume, for example, that $f(\cdot|\cdot)$ admits an invariant density $\pi(\cdot)$. Then even if $\mu(x_1) \neq \pi(x_1)$, it is useful to set $\gamma_1(x_1) = \pi(x_1)$ as it ensures that $\gamma_t(x_t) = \pi(x_t)$ for all $t \geq 2$. Moreover, we have $\gamma(x_t|x_{t+1}) = f(x_t|x_{t+1})$ if f is π-reversible.

Although defining $\gamma_t(x_t)$ through (31) appears appealing, this is also *very limiting*. For non-linear non-Gaussian state–space models, it is typically impossible to compute the integrals appearing in (31) in closed-form. Hence, it is impossible to implement the generalised two-filter smoothing formula for this choice of artificial densities. Our approach is clearly more general and does not require $\gamma_t(x_t)$ to satisfy (31) allowing the generalised two-filter smoothing formula to be implemented for any dynamic model.

Nevertheless, a generic sensible choice consists of selecting for $\gamma_t(x_t)$ an analytical approximation of the prior $p(x_t)$. For example in the standard case where one can generate easily a large number P of sample paths $\left\{X_{1:T}^{(i)}\right\}$ from the Markov process $\{X_t\}_{t\geq 1}$, it is possible to fit a mixture of Gaussian densities (or t densities) to the empirical approximations of the priors

$$\widehat{p}(dx_t) = \frac{1}{P}\sum_{i=1}^{P} \delta_{X_t^{(i)}}(\mathrm{d}x_t) \tag{33}$$

to obtain $\gamma_t(x_t)$ or to use

$$\gamma_t(x_t) = \frac{1}{P}\sum_{i=1}^{P} f\left(x_t \middle| X_{t-1}^{(i)}\right). \tag{34}$$

This second approximation might be too computationally intensive for practical applications. If the Markov process $\{X_t\}_{t\geq 1}$ is an ergodic Markov process with an unknown invariant density π, then we propose to select a time-invariant density $\gamma_t(x_t) = \gamma(x_t)$ approximating $\pi(x_t)$. To determine $\gamma(x)$, we can simulate a long sample path $X_{1:P}$, where $P >> T$ and fit a mixture of Gaussian densities (or t densities) to the empirical measure

$$\widehat{\pi}(dx) = \frac{1}{P-P_0}\sum_{t=P_0+1}^{P} \delta_{X_t}(\mathrm{d}x) \tag{35}$$

where P_0 corresponds to the burn-in or use

$$\gamma(x) = \frac{1}{P-P_0}\sum_{t=P_0+1}^{P} f(x|\, X_{t-1}). \tag{36}$$

2.4 Discussion

Several numerical implementations of the standard two-filter smoothing formula (7)–(8) described in the previous Section have already been proposed in the literature; e.g. among others in Helmick et al. (1995) for switching state–space models and in Isard and Blake (1998) for non-linear non-Gaussian state–space models. However, the algorithms proposed in these references rely on strong assumptions on the models and/or provide incorrect results. The main problem is that, although it is tempting to approximate both $p(x_t|y_{1:t})$ and $p(y_{t:T}|x_t)$ using popular approximation schemes, e.g. mixture of Gaussians as in Helmick et al. (1995), Extended/Unscented Kalman filter or SMC as in Isard and Blake (1998), these are not valid approaches in general for the term $p(y_{t:T}|x_t)$ because $p(y_{t:T}|x_t)$ is *not* a probability density in argument x_t and thus its integral over x_t may not be finite. Specifically, techniques such as SMC can only approximate finite measures and so their application to scenarios where $\int p(y_{t:T}|x_t)\mathrm{d}x_t = \infty$ will provide incorrect results. Moreover, even if $\int p(y_{t:T}|x_t)\mathrm{d}x_t < \infty$, current schemes rely on a backward dynamic model which has counter-intuitive properties as illustrated by the following example.

In cases where we have $X_t = \varphi(X_{t-1}, V_t)$ and it is possible to solve this equation in X_{t-1} to obtain

$$X_{t-1} = \zeta(X_t, V_t), \tag{37}$$

several authors in the literature define implicitly a backward Markov kernel (32) through (37) which corresponds to

$$\gamma(x_{t-1}|x_t) = \frac{f(x_t|x_{t-1})}{\int f(x_t|x_{t-1})\mathrm{d}x_{t-1}} \tag{38}$$

but, even if $\int f(x_t|x_{t-1})\mathrm{d}x_{t-1} < \infty$, this typically leads to a backward Markov kernel with undesirable properties. For example, consider a stationary AR(1) process defined by

$$X_t = aX_{t-1} + \sigma V_t, \quad X_1 \sim \mathcal{N}\left(0, \frac{\sigma^2}{1-a^2}\right), V_t \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2) \tag{39}$$

for $|a| < 1$ where $\mathcal{N}(\mu, \upsilon)$ denotes the normal distribution of mean μ and variance υ and $\mathcal{N}(x; \mu, \upsilon)$ the normal density of argument x and similar statistics. It follows from (31) that $\gamma_t(x) = \mathcal{N}\left(x; 0, \frac{\sigma^2}{1-a^2}\right)$ and a backward Markov kernel (32) of the form

$$\gamma(x_{t-1}|x_t) = f(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; ax_t, \sigma^2). \tag{40}$$

On the other hand, $X_{t-1} = a^{-1}(X_t - \sigma V_t)$ leads to

$$\gamma(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; a^{-1}x_t, a^{-2}\sigma^2) \tag{41}$$

which is a non-stationary Markov process.

3 Generalized SMC two-filter smoother

SMC methods are a generic set of methods to sample from any sequence of densities of increasing dimension; see Doucet et al. (2001) for a review. We briefly describe an SMC approximation allowing us to approximate the artificial densities $\{\widetilde{p}\,(x_{t:T}|\, y_{t:T})\}$ hence the marginals $\{\widetilde{p}\,(x_t|\, y_{t:T})\}$. We focus on a simple sequential importance resampling (SIR) strategy although more sophisticated algorithms can be implemented. The SIR algorithm relies on a sequence of importance densities: the density $\widetilde{q}\,(x_T|\, y_T)$ whose support include the support of $\widetilde{p}\,(x_t|\, y_{t:T})$ and $\widetilde{q}\,(x_t|\, y_t, x_{t+1})$ whose support include the support of $g(y_t|x_t)\gamma_t(x_t)\,f\,(x_{t+1}|x_t)$. The algorithm proceeds as follows: see Doucet et al. (2001) for implementation details.

SMC implementation of the generalised backward information filter

1. Initialisation, time $t = T$.
   - For $i = 1, \ldots, N$ sample $\widetilde{X}_T^{(i)} \sim \widetilde{q}\,(\cdot|\, y_T)$; evaluate the weight

$$\widetilde{W}_T^{(i)} \propto \frac{\widetilde{p}\left(\widetilde{X}_T^{(i)}\middle|\, y_T\right)}{\widetilde{q}\left(\widetilde{X}_T^{(i)}\middle|\, y_T\right)} \propto \frac{\gamma_T\left(\widetilde{X}_T^{(i)}\right) g\left(y_T|\,\widetilde{X}_T^{(i)}\right)}{\widetilde{q}\left(\widetilde{X}_T^{(i)}\middle|\, y_T\right)}. \tag{42}$$

   Iterate Steps 2 and 3.
2. Resampling.
   - Normalize the weights $\sum_{i=1}^{N} \widetilde{W}_t^{(i)} = 1$.
   - Resample the particles $\left\{\widetilde{X}_{t:T}^{(i)}\right\}$ and set $\widetilde{W}_t^{(i)} = \frac{1}{N}$.
3. Sampling.
   - Set $t = t - 1$; if $t = 0$ stop.
   - For $i = 1, \ldots, N$ sample $\widetilde{X}_t^{(i)} \sim \widetilde{q}\left(\cdot|\, y_t, \widetilde{X}_{t+1}^{(i)}\right)$; evaluate the weight

$$\begin{aligned}\widetilde{W}_t^{(i)} &\propto \frac{\widetilde{p}\left(\widetilde{X}_{t:t+1}^{(i)}\middle|\, y_{t:T}\right)}{\widetilde{p}\left(\widetilde{X}_{t+1}^{(i)}\middle|\, y_{t+1:T}\right)\widetilde{q}\left(\widetilde{X}_t^{(i)}\middle|\, y_t, \widetilde{X}_{t+1}^{(i)}\right)}\\ &\propto \frac{g\left(y_t|\widetilde{X}_t^{(i)}\right)\gamma_t\left(\widetilde{X}_t^{(i)}\right) f\left(\widetilde{X}_{t+1}^{(i)}\middle|\,\widetilde{X}_t^{(i)}\right)}{\gamma_{t+1}\left(\widetilde{X}_{t+1}^{(i)}\right)\widetilde{q}\left(\widetilde{X}_t^{(i)}\middle|\, y_t, \widetilde{X}_{t+1}^{(i)}\right)}.\end{aligned} \tag{43}$$

We obtain the following Monte Carlo approximation of $\widetilde{p}\,(x_{t:T}|\, y_{t:T})$

$$\widehat{\widetilde{p}}\,(\mathrm{d}x_{t:T}|\, y_{t:T}) = \sum_{i=1}^{N} \widetilde{W}_t^{(i)} \delta_{\widetilde{X}_{t:T}^{(i)}}\,(\mathrm{d}x_t) \tag{44}$$

hence an approximation of $\widetilde{p}\,(x_t|\,y_{t:T})$

$$\widehat{\widetilde{p}}\,(\mathrm{d}x_t|\,y_{t:T}) = \sum_{i=1}^{N} \widetilde{W}_t^{(i)} \delta_{\widetilde{X}_t^{(i)}}\,(\mathrm{d}x_t).$$

Now combining the particle approximations of $p(x_t|\,y_{1:t-1})$ and $\widetilde{p}(x_t|\,y_{t:T})$ and using (23), we obtain the following approximation of $p\,(x_t|\,y_{1:T})$

$$\widehat{p}(\mathrm{d}x_t|y_{1:T}) = \sum_{j=1}^{N} W_{t|T}^{(j)} \delta_{\widetilde{X}_t^{(j)}}(\mathrm{d}x_t) \tag{45}$$

where

$$W_{t|T}^{(j)} \propto \widetilde{W}_t^{(j)} \sum_{i=1}^{N} W_{t-1}^{(i)} \frac{f\left(\widetilde{X}_t^{(j)}|X_{t-1}^{(i)}\right)}{\gamma_t\left(\widetilde{X}_t^{(j)}\right)}. \tag{46}$$

Note that it is also possible to combine the particle approximations of $p(x_t|y_{1:t})$ and $\widetilde{p}\,(x_t\;|\,y_{t+1:T})$ to obtain an approximation of $p\,(x_t|\,y_{1:T})$. Like the SMC implementation of the forward filtering-backward smoothing formula, the computational complexity of this algorithm is $\mathcal{O}(TN^2)$. However, fast computational methods have been developed to address this problem (Klaas et al. 2006). Moreover, note that if (22) is satisfied, then it is possible to reduce this computational complexity to $\mathcal{O}(TN)$ by using rejection sampling with $\widehat{p}\,(\mathrm{d}x_{t-1}|\,y_{1:t-1})\,\widehat{\widetilde{p}}\,(\mathrm{d}x_t|\,y_{t:T})$ as a proposal. More recently, an important sampling type approach has also been proposed in Fearnhead et al. (2008b) to reduce the computational complexity to $\mathcal{O}(TN)$; see Briers et al. (2005) for a similar idea developed in the context of belief propagation. Fearnhead et al. (2008b) also discusses an extension to address the case where the Radon–Nikodym derivative $f\,(x_t|x_{t-1})\,/\gamma_t\,(x_t)$ is not defined because $f\,(x_t|x_{t-1})$ has a singular component.

3.1 Algorithm settings and convergence results

For the forward filter, it is well-known that a sensible choice for the importance density consists of minimizing the variance of the incremental importance weight $g(\,y_t|\,x_t)\,f(x_t|\,x_{t-1})/q(x_t|\,y_t, x_{t-1})$ conditional upon x_t is given by $q^{\mathrm{opt}}(x_t|\,y_t, x_{t-1}) \propto g\,(\,y_t|\,x_t)\,f\,(x_t|\,x_{t-1})$ and the resulting importance weight is equal to $p\,(\,y_t|\,x_{t-1}) = \int g\,(\,y_t|\,x_t)\,f\,(x_t|\,x_{t-1})\,dx_t$; see Doucet et al. (2000). A reasoning similar to Doucet et al. (2000) leads to the following result: the optimal backward importance density $\widetilde{q}^{\mathrm{opt}}(x_t|\,x_{t+1}, y_t)$ minimizing the variance of the incremental weight

$$\widetilde{w}_t\,(x_t, x_{t+1}) = \frac{g\,(\,y_t|\,x_t)\,\gamma_t\,(x_t)\,f\,(x_{t+1}|\,x_t)}{\gamma_{t+1}\,(x_{t+1})\,\widetilde{q}\,(x_t|\,y_t, x_{t+1})}. \tag{47}$$

given x_{t+1} satisfies

$$\widetilde{q}^{\text{opt}}\left(x_t|\, y_t, x_{t+1}\right) \propto g\left(y_t|\, x_t\right) \frac{\gamma_t\left(x_t\right) f\left(x_{t+1}|\, x_t\right)}{\gamma_{t+1}\left(x_{t+1}\right)} \tag{48}$$

and the associated incremental weight (47) is equal to

$$\widetilde{w}_t^{\text{opt}}\left(x_{t+1}\right) = \int g\left(y_t|\, x_t\right) \frac{\gamma_t\left(x_t\right) f\left(x_{t+1}|\, x_t\right)}{\gamma_{t+1}\left(x_{t+1}\right)} \mathrm{d}x_t. \tag{49}$$

If it is not possible to sample (48) and/or compute (49) then one can design an approximation of (48), for example, using an Extended or Unscented Kalman approximation, computing the resulting incremental importance weight using (47).

From a convergence point of view, the general results on SMC methods presented in Del Moral (2004) such as Lp-convergence, central limit theorem or uniform (in time) convergence can be applied straightforwardly and we do not present them here. We just recall that these results typically require the incremental weight (47) to be upper-bounded on $\mathcal{X} \times \mathcal{X}$.

3.2 SMC sampling from the joint distribution

We now provide detail of the SMC implementation of this algorithm, which is the *generalised* two-filter smoothing analogy of the procedure outlined in Godsill et al. (2004).

SMC procedure to sample approximately from $p\left(x_{1:T}|\, y_{t:T}\right)$

1. Select t to be between 2 and $T-1$, say $t = T/2$ if T is even, and sample $X'_t \sim \widehat{p}(\cdot|y_{1:T})$ (by randomly selecting $\widetilde{X}_t^{(i)}$ with probability $W_{t|T}^{(i)}$).
2. Sample from (26) by sampling each variate recursively backwards in time, $X'_{t-1}, \ldots, X'_1$, by randomly selecting sample X'_k from an approximation of $p(x_k|y_{k:T}, X'_{k+1})$, i.e. by randomly selecting $X_k^{(i)}$ with probability

$$\alpha_k^{(i)} \propto W_k^{(i)} f(X'_{k+1}|X_k^{(i)}).$$

3. Sample from (28) by sampling each variate recursively forward in time, $X'_{t+1}, \ldots, X'_T$, by randomly selecting sample X'_k from an approximation of $p(x_k|y_{k:T}, X'_{k-1})$, i.e. by randomly selecting $\widetilde{X}_k^{(i)}$ with probability

$$\beta_k^{(i)} \propto \widetilde{W}_k^{(i)} \frac{\gamma_{k-1}(X'_{k-1}) f(\widetilde{X}_k^{(i)}|X'_{k-1})}{\gamma_k(\widetilde{X}_k^{(i)})}.$$

It follows from Proposition 4 that $X'_{1:T} = \{X'_1, \ldots, X'_T\}$ is an approximate realisation from $p(x_{1:T}|y_{1:T})$. An important advantage of this algorithm over the forward filtering-backward sampling is that the two sampling operations can be performed in parallel. Moreover, a straightforward generalisation to the case where the time interval $\{1, 2, \ldots, T\}$ is divided into more than two intervals can also be easily derived.

4 Generalized two-filter smoothing for conditionally linear Gaussian models

A popular mechanism for reducing the variance of an estimator is to exploit the Rao–Blackwell theorem. By performing so-called Rao–Blackwellisation, efficient SMC filtering algorithms have appeared in the literature; see Chen and Liu (2000) and Doucet et al. (2000). Informally, Rao–Blackwellisation exploits the fact that analytic substructure appears in the problem under consideration and so one can marginalise such structure to reduce the dimension of the space on which one is performing the Monte Carlo approximation. That is, the space E can be partitioned into two disjoint subsets E_A and E_Z with associated random variables A_t and Z_t. We assume that Z_t can be marginalised analytically.

Consider the following conditionally linear Gaussian model

$$Z_t = H(A_t)Z_{t-1} + J(A_t)V_t \tag{50}$$

$$Y_t = K(A_t)Z_t + L(A_t)W_t \tag{51}$$

where $V_t \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, I_v)$, $W_t \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, I_w)$ and $\{A_t\}_{t\in\mathbb{N}}$ is a latent Markov process. $H(\cdot)$, $J(\cdot)$, $K(\cdot)$ and $L(\cdot)$ are matrices of appropriate dimension. Denote $X_t := (A_t\ Z_t)$. The initial state X_1 is distributed according to $\mu(x_1) = \mu_a(a_1)\mu_z(z_1)$ where $\mu_z(z_1) = \mathcal{N}(z_1; m, \Sigma)$ and the transition kernel satisfies

$$f(x_t|x_{t-1}) = f_z(z_t|z_{t-1}, a_t)f_a(a_t|a_{t-1}) \tag{52}$$

with $f_z(z_t|z_{t-1}, a_t) = \mathcal{N}\left(z_t; H(a_t)z_{t-1}, J(a_t)J(a_t)^T\right)$ and $g(y_t|x_t) = \mathcal{N}\left(y_t; K(a_t)z_t, L(a_t)L(a_t)^T\right)$.

We are interested in estimating the sequence of smoothed densities $\{p(a_t, z_t|y_{1:T})\}$. It is possible to use the methods described in the previous sections directly on the process $\{X_t\}_{t\in\{1,\ldots,T\}}$ to estimate $\{p(a_t, z_t|y_{1:T})\}$ directly. However, this would not take into account the structure of the model (50) and (51). Indeed, conditional upon $\{A_t\}_{t\in\mathbb{N}}$, (50) and (51) define a standard linear Gaussian model and this can be exploited to propose a specific generalised two-filter formula. We are able to reduce the variance of the SMC estimates by performing calculations.

4.1 Generalized backward information filter

It is straightforward to show that

$$p(a_t, z_t|y_{1:T}) \propto \frac{p(a_t, z_t|y_{1:t-1})\widetilde{p}(a_t, z_t|y_{t:T})}{\gamma_t^a(a_t)\gamma_t^z(z_t)} \tag{53}$$

where $\{\gamma_t^a(a_t)\}$ and $\{\gamma_t^z(z_t)\}$ are two sequences of artificial densities with $\{\gamma_t^z(z_t)\}$ selected as $\gamma_t^z(z_t) = \mathcal{N}(z_t; m_t, \Sigma_t)$. As will be shown, $\widetilde{p}(a_t, z_t|y_{t:T})$ is calculated through

$$\widetilde{p}(a_t, z_t|y_{t:T}) = \int \widetilde{p}(a_{t:T}, z_t|y_{t:T}) da_{t+1:T} \tag{54}$$

where the artificial densities $\widetilde{p}(a_{t:T}, z_t|y_{t:T})$ are defined, for $t = T$, as

$$\widetilde{p}(a_T, z_T|y_T) \propto \gamma_T^a(a_T)\gamma_T^z(z_T)g(y_T|a_T, z_T), \tag{55}$$

and for $t \geq 2$

$$\widetilde{p}(a_{t:T}, z_t|y_{t:T}) \propto \gamma_t^a(a_t) \prod_{k=t+1}^{T} f_a(a_k|a_{k-1})\gamma_t^z(z_t)p(y_{t:T}|a_{t:T}, z_t). \tag{56}$$

Equations (55) and (56) will be useful in determining a recursive weight update equation in what follows. Compared to (13) and (15), (55) and (56) rely on the term $p(y_{t:T}|a_{t:T}, z_t)$ which does not have an elegant factorisation as in the standard case. Calculation of this term under the current modelling assumptions is a generalisation of the backward information filter for linear Gaussian models presented in Mayne (1966) and Doucet and Andrieu (2001).

4.1.1 SMC implementation

An SMC algorithm approximates the density $\widetilde{p}(a_{t:T}|y_{t:T})$ (and so $\widetilde{p}(a_t|y_{t:T})$ by marginalisation) through a set of weighted samples. Calculation of the weights is performed by marginalisation of Z_t. That is, at time $t = T$, the weight is given by

$$\widetilde{W}_T^{(i)} = \frac{\widetilde{p}(\widetilde{A}_T^{(i)}|y_T)}{\widetilde{q}(\widetilde{A}_T^{(i)}|y_T)} = \frac{\int \widetilde{p}(\widetilde{A}_T^{(i)}, z_T|y_T)\mathrm{d}z_T}{\widetilde{q}(\widetilde{A}_T^{(i)}|y_T)}. \tag{57}$$

Direct substitution of Eq. (55) into (57) yields

$$\widetilde{W}_T^{(i)} \propto \frac{\gamma_T^a(\widetilde{A}_T^{(i)}) \int \gamma_T^z(z_T)g(y_T|\widetilde{A}_T^{(i)}, z_T)\mathrm{d}z_T}{\widetilde{q}(\widetilde{A}_T^{(i)}|y_T)}. \tag{58}$$

Similarly, the weight at time $t \geq 2$ is given by

$$\begin{aligned}\widetilde{W}_t^{(i)} &\propto \widetilde{W}_{t+1}^{(i)} \frac{\widetilde{p}(\widetilde{A}_{t:T}^{(i)}|y_{t:T})}{\widetilde{p}(\widetilde{A}_{t+1:T}^{(i)}|y_{t+1:T})\widetilde{q}(\widetilde{A}_t^{(i)}|y_{t:T}, \widetilde{A}_{t+1:T}^{(i)})} \\ &\propto \widetilde{W}_{t+1}^{(i)} \frac{\int \widetilde{p}(\widetilde{A}_{t:T}^{(i)}, z_t|y_{t:T})\mathrm{d}z_t}{\int \widetilde{p}(\widetilde{A}_{t+1:T}^{(i)}, z_{t+1}|y_{t+1:T})\mathrm{d}z_{t+1}\widetilde{q}(\widetilde{A}_t^{(i)}|y_{t:T}, \widetilde{A}_{t+1:T}^{(i)})}. \end{aligned} \tag{59}$$

Direct substitution of (56) into (59) yields

$$\widetilde{W}_T^{(i)} \propto \widetilde{W}_{t+1}^{(i)} \frac{\gamma_t^a\left(\widetilde{A}_t^{(i)}\right) f_a\left(\widetilde{A}_{t+1}^{(i)} | \widetilde{A}_t^{(i)}\right) \int \gamma_t^z(z_t) p(y_{t:T} | \widetilde{A}_{t:T}^{(i)}, z_t) \mathrm{d}z_t}{\gamma_{t+1}^a\left(\widetilde{A}_{t+1}^{(i)}\right) \widetilde{q}\left(\widetilde{A}_t^{(i)} | y_{t:T}, \widetilde{A}_{t+1:T}^{(i)}\right) \int \gamma_{t+1}^z(z_{t+1}) p(y_{t+1:T} | \widetilde{A}_{t+1:T}^{(i)}, z_{t+1}) \mathrm{d}z_{t+1}}. \tag{60}$$

As is apparent, to implement this backward filter, it is necessary to be able to compute $\int \gamma_t^z(z_t) p(y_{t:T} | a_{t:T}, z_t) \mathrm{d}z_t$ pointwise up to a normalising constant. This integral, constructed through the following proposition, justifies the necessary introduction of the artificial densities $\{\gamma_t^z\}$, and the selection of a Gaussian density (or more generally a Gaussian mixture density) for the variables $\{Z_t\}$.

Proposition 5 *Assume that $p(y_{t:T} | a_{t:T}, z_t)$ is parameterised by its information matrix, $\widetilde{P}_{t|t}^{-1}(a_{t:T})$, and information vector, $\widetilde{\nu}_{t|t}(a_{t:T})$, with constant term $\widetilde{c}_{t|t}(a_{t:T})$, for each $t \in \{1, \ldots, T\}$. Then, for any $t \in \{1, \ldots, T\}$ we have*

$$\int \gamma_t^z(z_t) p(y_{t:T} | a_{t:T}, z_t) \mathrm{d}z_t \propto |\widetilde{\Omega}_{t|t}^{-1}(a_{t:T})|^{-1/2} \exp\left\{-\frac{1}{2}\left(\widetilde{c}_{t|t}(a_{t:T}) - \widetilde{z}_{t|t}(a_{t:T})^T \widetilde{\Omega}_{t|t}^{-1}(a_{t:T}) \widetilde{z}_{t|t}(a_{t:T})\right)\right\} \tag{61}$$

where

$$\widetilde{\Omega}_{t|t}^{-1}(a_{t:T}) = \widetilde{P}_{t|t}^{-1}(a_{t:T}) + \Sigma_t^{-1} \tag{62}$$

$$\widetilde{z}_{t|t}(a_{t:T}) = \widetilde{\Omega}_{t|t}(a_{t:T})\left(\widetilde{\nu}_{t|t}(a_{t:T}) + \Sigma_t^{-1} m_t\right). \tag{63}$$

The resulting SMC algorithm proceeds as follows.

SMC implementation of the generalised backward information filter for a conditionally Gaussian linear model

1. Initialise at time $t = T$.
   - For $i = 1, \ldots, N$, sample $\widetilde{A}_T^{(i)} \sim \widetilde{q}(\cdot | y_T)$.
   - For $i = 1, \ldots, N$, compute and normalise the importance weights:

$$\widetilde{W}_T^{(i)} \propto \frac{\gamma_T^a\left(\widetilde{A}_T^{(i)}\right) \int \gamma_T^z(z_T) p\left(y_T | \widetilde{A}_T^{(i)}, z_T\right) \mathrm{d}z_T}{\widetilde{q}\left(\widetilde{A}_T^{(i)} | y_T\right)}. \tag{64}$$

   Iterate Steps 2 and 3.
2. Resampling.
   - Normalize the weights $\sum_{i=1}^N \widetilde{W}_t^{(i)} = 1$.
   - Resample the particles $\left\{\widetilde{A}_{t:T}^{(i)}\right\}$ and set $\widetilde{W}_t^{(i)} = \frac{1}{N}$.

3. Sampling.
   - Set $t = t - 1$; if $t = 0$ stop.
   - For $i = 1, \ldots, N$, sample $\widetilde{A}_t^{(i)} \sim \widetilde{q}(\cdot|y_t, \widetilde{A}_{t+1:T}^{(i)})$; evaluate the weight

$$\widetilde{W}_t^{(i)} \propto \widetilde{W}_{t+1}^{(i)} \frac{\gamma_t^a\left(\widetilde{A}_t^{(i)}\right) f_a\left(\widetilde{A}_{t+1}^{(i)}|\widetilde{A}_t^{(i)}\right) \int \gamma_t^z(z_t) p(y_{t:T}|\widetilde{A}_{t:T}^{(i)}, z_t) \mathrm{d}z_t}{\gamma_{t+1}^a\left(\widetilde{A}_{t+1}^{(i)}\right) \widetilde{q}\left(\widetilde{A}_t^{(i)}|y_{t:T}, \widetilde{A}_{t+1:T}^{(i)}\right) \int \gamma_{t+1}^z(z_{t+1}) p(y_{t+1:T}|\widetilde{A}_{t+1:T}^{(i)}, z_{t+1}) \mathrm{d}z_{t+1}}. \tag{65}$$

At first glance, this algorithm seems to require the complete path $\{\widetilde{A}_{t:T}^{(i)}\}$ at time t. However, like the Rao-Blackwellised particle filters described in Chen and Liu (2000) and Doucet et al. (2000) the weight update (60) depends on $a_{t:T}$ only through the set of sufficient statistics of the backward information filter for linear Gaussian state–space models, which will now be derived. Proof of this proposition can be found in Briers et al. (2004).

Proposition 6 *For each $t \in \{1, \ldots, T\}$ we have:*

$$p(y_{t:T}|a_{t:T}, z_t) \propto \exp\left\{-\frac{1}{2}\widetilde{c}_{t|t}(a_{t:T}) - \frac{1}{2}z_t^T \widetilde{P}_{t|t}^{-1}(a_{t:T}) + \widetilde{\nu}_{t|t}(a_{t:T})\right\} \tag{66}$$

where the constant term, information matrix and information vector ($\widetilde{c}_{t|t}(a_{t:T})$, $\widetilde{P}_{t|t}^{-1}(a_{t:T})$ and $\widetilde{\nu}_{t|t}(a_{t:T})$ respectively) satisfy the following recursion

$$\widetilde{c}_{t|t}(a_{t:T}) = \widetilde{c}_{t|t+1}(a_{t+1:T}) - \log\left(\left|\left(L(a_t)L(a_t)^T\right)^{-1}\right|\right) + y_t^T\left(L(a_t)L(a_t)^T\right)y_t \tag{67}$$

$$\widetilde{P}_{t|t}^{-1}(a_{t:T}) = \widetilde{P}_{t|t+1}^{-1}(a_{t+1:T}) + K(a_t)^T\left(L(a_t)L(a_t)^T\right)^{-1}K(a_t) \tag{68}$$

$$\widetilde{\nu}_{t|t}(a_{t:T}) = \widetilde{\nu}_{t|t+1}(a_{t+1:T}) + K(a_t)^T\left(L(a_t)L(a_t)^T\right)^{-1}y_t \tag{69}$$

with the intermediate terms $\widetilde{c}_{t|t+1}(a_{t+1:T})$, $\widetilde{P}_{t|t+1}^{-1}(a_{t+1:T})$ and $\widetilde{\nu}_{t|t+1}(a_{t+1:T})$ parameterising
$p(y_{t+1:T}|a_{t+1:T}, z_t)$, given as

$$\begin{aligned}\widetilde{c}_{t|t+1}(a_{t+1:T}) = {} & \widetilde{c}_{t+1|t+1}(a_{t+1:T}) + \log\left(\left|(J(a_{t+1})J(a_{t+1})^T)\right|\right) \\ & - \log\left(\left|J(a_{t+1})\Delta_{t+1}(a_{t+1:T})J(a_{t+1})^T\right|\right) \\ & - \widetilde{\nu}_{t+1|t+1}(a_{t+1:T})^T J(a_{t+1})J(a_{t+1})^T \Delta_{t+1}(a_{t+1:T})\widetilde{\nu}_{t+1|t+1}(a_{t+1:T})\end{aligned} \tag{70}$$

$$\begin{aligned}\widetilde{P}_{t|t+1}^{-1}(a_{t+1:T}) = {} & H(a_{t+1})^T\left[I - \widetilde{P}_{t+1|t+1}^{-1}(a_{t+1:T})J(a_{t+1})\Delta_{t+1}(a_{t+1:T})J(a_{t+1})^T\right] \\ & \times \widetilde{P}_{t+1|t+1}^{-1}(a_{t+1:T})H(a_{t+1})\end{aligned} \tag{71}$$

$$\begin{aligned}\widetilde{\nu}_{t|t+1}(a_{t+1:T}) = {} & H(a_{t+1})^T\left[I - \widetilde{P}_{t+1|t+1}^{-1}(a_{t+1:T})J(a_{t+1})\Delta_{t+1}(a_{t+1:T})J(a_{t+1})^T\right] \\ & \times \widetilde{\nu}_{t+1|t+1}(a_{t+1:T})^T,\end{aligned} \tag{72}$$

where:

$$\Delta_{t+1}(a_{t+1:T}) = \left(I + J(a_{t+1})^T \widetilde{P}^{-1}_{t+1|t+1}(a_{t+1:T}) J(a_{t+1})\right)^{-1}.$$

The boundary conditions parameterising $p(y_T|a_T, z_T)$ are as follows

$$\widetilde{c}_{T|T}(a_T) = -\log\left(\left|\left(L(a_T)L(a_T)^T\right)^{-1}\right|\right) + y_T^T\left(L(a_T)L(a_T)^T\right)^{-1} y_T \tag{73}$$

$$\widetilde{P}^{-1}_{T|T}(a_T) = K(a_T)^T\left(L(a_T)L(a_T)^T\right)^{-1} K(a_T) \tag{74}$$

$$\widetilde{\nu}_{T|T}(a_T) = K(a_T)^T\left(L(a_T)L(a_T)^T\right)^{-1} y_T. \tag{75}$$

It is important that one calculates the constant terms in the above recursion since they are dependent upon $a_{t:T}$ and so contribute to the SMC weight update equation. This is in contrast to the traditional two-filter formulation for Gaussian state–space models, in which calculation of this constant term is unnecessary.

4.1.2 Algorithm settings

In this context the optimal importance function for the generalised backward information filter is given by

$$\widetilde{q}^{\,\text{opt}}\left(a_t|y_{t:T}, a_{t+1:T}\right) \propto \frac{\gamma_t^a\left(a_t\right) f_a\left(a_{t+1}|a_t\right) \int \gamma_t^z(z_t) p(y_{t:T}|a_{t:T}, z_t) \mathrm{d}z_t}{\gamma_{t+1}^a\left(a_{t+1}\right) \int \gamma_{t+1}^z(z_{t+1}) p(y_{t+1:T}|a_{t+1:T}, z_{t+1}) \mathrm{d}z_{t+1}}. \tag{76}$$

The optimal importance density only depends on $a_{t:T}$ through $a_{t:t+1}$ and the the sufficient statistics for the backward information filter derived above. Hence, when using this importance density there is no need to store the complete path $\left\{\widetilde{A}^{(i)}_{t:T}\right\}$.

4.1.3 Combination step

Using equation (53), it is possible to combine the approximation of the (generalised) backward information filter with an approximation based on a standard Rao-Blackwellised SMC filter presented in Chen and Liu (2000) and Doucet et al. (2000)

$$\begin{aligned}\widehat{p}\left(\mathrm{d}a_t, z_t|y_{1:t-1}\right) \\ = \sum_{i=1}^{N} W^{(i)}_{t-1} f\left(a_t|A^{(i)}_{t-1}\right)\mathcal{N}\left(z_t;\, P^{-1}_{t|t-1}\left(A^{(i)}_{1:t}\right)\nu_{t|t-1}(a_{1:t}),\, P_{t|t-1}\left(A^{(i)}_{1:t}\right)\right),\end{aligned} \tag{77}$$

where the terms $P^{-1}_{t|t-1}\left(A^{(i)}_{1:t}\right)$ and $\nu_{t|t-1}(a_{1:t})$ are the (predicted) information matrix and vector computed in the forward filtering operation. All that remains to be specified is the approximation to the backward information filtering quantity $\widetilde{p}(a_t, z_t|y_{t:T})$. By (54) we can write the following

$$\widetilde{p}(a_t, z_t|y_{t:T}) = \gamma_t^z(z_t) \int \frac{\widetilde{p}(a_{t:T}|y_{t:T}) p(y_{t:T}|a_{t:T}, z_t)}{\int \gamma_t^z(z_t') p(y_{t:T}|a_{t:T}, z_t') \mathrm{d}z_t'} \mathrm{d}a_{t+1:T}, \tag{78}$$

and so

$$
\begin{aligned}
&\widehat{\widehat{p}}(a_t, z_t|y_{t:T})\\
&\quad= \gamma_t^z(z_t)\sum_{j=1}^{N}\widetilde{W}_t^{(j)}\delta_{A_t^{(j)}}(a_t)\exp\left\{-\frac{1}{2}\widetilde{z}_{t|t}(A_{t:T}^{(j)})^T\widetilde{\Omega}_{t|t}^{-1}(A_{t:T}^{(j)})z_{t|t}(A_{t:T}^{(j)})\right.\\
&\qquad\left.-\frac{1}{2}z_t^T\widetilde{P}_{t|t}^{-1}(A_{t:T}^{(j)})z_t+\widetilde{\nu}_{t|t}(A_{t:T}^{(j)})\right\}. \qquad (79)
\end{aligned}
$$

Through straightforward algebraic manipulations, one obtains the following approximation of the desired density

$$
\begin{aligned}
\widehat{p}(\mathrm{d}a_t, z_t|y_{1:T}) = &\sum_{j=1}^{N}\sum_{i=1}^{N}W_{t|T}^{(i,j)}\delta_{\widetilde{A}_t^{(j)}}(\mathrm{d}a_t)\mathcal{N}\\
&\times\left(z_t;\, P_{t|T}^{-1}(A_{1:t-1}^{(i)},\widetilde{A}_{t:T}^{(i)})\nu_{t|T}(A_{1:t-1}^{(i)},\widetilde{A}_{t:T}^{(i)}),\,P_{t|T}(A_{1:t-1}^{(i)},\widetilde{A}_{t:T}^{(i)})\right),
\end{aligned}
$$

where

$$
\begin{aligned}
P_{t|T}^{-1}(a_{1:T}) &= P_{t|t-1}^{-1}(a_{1:t})+\widetilde{P}_{t|t}^{-1}(a_{t:T}) &(80)\\
\nu_{t|T}(a_{1:T}) &= P_{t|T}(a_{1:T})\left[\nu_{t|t-1}(a_{1:t})+\widetilde{\nu}_{t|t}(a_{t:T})\right] &(81)
\end{aligned}
$$

and the weight equation being given as

$$
\begin{aligned}
W_{t|T}^{(i,j)} \propto\ &\widetilde{W}_t^{(j)}W_{t-1}^{(i)}\frac{f_a\left(\widetilde{A}_t^{(j)}|A_{t-1}^{(i)}\right)}{\gamma_t^a(\widetilde{A}_t^{(j)})}\exp\left\{-\frac{1}{2}\left(\log(|\Omega_{t|T}^{-1}(A_{1:t-1}^{(i)},\widetilde{A}_{t:T}^{(j)})|)\right.\right.\\
&+\log(|P_{t|t-1}(A_{1:t-1}^{(i)})|)-\log(|P_{t|T}(A_{1:t-1}^{(i)},A_{t:T}^{(j)})|)\\
&-\widetilde{z}_{t|t}(\widetilde{A}_{t:T}^{(j)})^T\widetilde{\Omega}_{t|t}(\widetilde{A}_{t:T}^{(j)})^{-1}\widetilde{z}_{t|t}(\widetilde{A}_{t:T}^{(j)})^T\\
&+\nu_{t|t-1}(A_{1:t-1}^{(i)})P_{t|t-1}^{-1}(A_{1:t-1}^{(i)})\nu_{t|t-1}(A_{1:t-1}^{(i)})\\
&\left.\left.-\nu_{t|T}(A_{1:t-1}^{(i)},\widetilde{A}_{t:T}^{(j)})P_{t|T}^{-1}(A_{1:t-1}^{(i)},\widetilde{A}_{t:T}^{(j)})\nu_{t|T}(A_{1:t-1}^{(i)},\widetilde{A}_{t:T}^{(j)})\right)\right\}. \qquad (82)
\end{aligned}
$$

4.2 Extensions

A similar Rao-Blackwellised SMC idea can be applied to the class of partially observed linear Gaussian models defined by

$$
\begin{aligned}
Z_t &= HZ_{t-1}+JV_t, &(83)\\
A_t &= KZ_t+LW_t, &(84)\\
Y_t|\left(Z_t=y_t, A_t=a_t\right) &\sim g\left(\cdot|\,a_t\right), &(85)
\end{aligned}
$$

where $\{Z_t\}_{t\in\mathbb{N}}$ and $\{A_t\}_{t\in\mathbb{N}}$ are *unobserved* processes, $V_t \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, I_v)$, $W_t \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, I_w)$ and H, J, K and L are matrices of appropriate dimension; see Andrieu and Doucet (2002) and de Jong (1997) for applications. Let us denote $X_t := (A_t, Z_t)$. The initial state X_1 is distributed according to $\mu(x_1) = \mathcal{N}(z_1; m, \Sigma)\, f_a(a_t|\, z_t)$ and the transition kernel satisfies

$$f(x_t|\, x_{t-1}) = f_{a,z}(a_t|\, a_{t-1}, z_t)\, f_z(z_t|\, z_{t-1}) \tag{86}$$

where $f_{a,z}\left(a_t \middle| a_{t-1}, z_t\right) = \mathcal{N}\left(a_t; K z_t, L L^{\mathrm{T}}\right)$ and $f_z\left(z_t|\, z_{t-1}\right) = \mathcal{N}\left(z_t; H z_{t-1}, J J^{\mathrm{T}}\right)$. We are interested in computing the sequence of smoothed densities $\{p(a_t|\, y_{1:T})\}$. In this case, we can exploit the structure of the model to integrate out z_t through Kalman filtering techniques as in Andrieu and Doucet (2002).

The generalised two-filter smoothing formula is given by

$$p(a_t|\, y_{1:T}) \propto \frac{p(a_t|\, y_{1:t-1})\, \widetilde{p}(a_t|\, y_{t:T})}{\gamma_t(a_t)} \tag{87}$$

where

$$\widetilde{p}(a_{t:T}|\, y_{t:T}) \propto \gamma_t(a_t)\, p(a_{t+1:T}|\, a_t) \prod_{k=t}^{T} g(y_k|\, a_k)\,. \tag{88}$$

In an SMC implementation, the forward filter $\{p(a_t|\, y_{1:t})\}$ will be approximated by a Rao-Blackwellised particle filter which is a random sum of Kalman filters described in Andrieu and Doucet (2002) whereas $\{\widetilde{p}(a_{t:T}|\, y_{t:T})\}$ can be implemented in a similar manner. It relies on computing

$$\begin{aligned}\frac{\widetilde{p}(a_{t:T}|\, y_{t:T})}{\widetilde{p}(a_{t+1:T}|\, y_{t+1:T})\, \widetilde{q}(a_t|\, y_{t:T}, a_{t+1:T})} &\propto \frac{\gamma_t(a_t)\, p(a_{t+1:T}|\, a_t)\, g(y_t|\, a_t)}{\gamma_{t+1}(a_{t+1})\, p(a_{t+2:T}|\, a_{t+1})} \\ &\propto \frac{\gamma_t(a_t)\, p(a_t|\, a_{t+1:T})\, p(a_{t+1})\, g(y_t|\, a_t)}{\gamma_{t+1}(a_{t+1})\, p(a_t)}.\end{aligned} \tag{89}$$

We can use here $\gamma_t(a_t) = p(a_t)$ as it can easily be computed analytically using (83)–(84). All these calculations can be done efficiently using a generalised backward information filter similar to the one developed in the previous section but associated here to (83)–(84) as

$$p(a_t|\, a_{t+1:T}) = \frac{p(a_{t+1:T}|\, a_t)\, p(a_t)}{p(a_{t+1:T})} \tag{90}$$

where

$$p\left(a_{t+1:T}\,|\,a_t\right) = \int p\left(a_{t+1:T}\,|\,z_t\right) p\left(z_t\,|\,a_t\right) \mathrm{d}z_t, \tag{91}$$

$$p\left(a_{t+1:T}\right) = \int p\left(a_{t+1:T}\,|\,z_t\right) p\left(z_t\right) \mathrm{d}z_t. \tag{92}$$

5 Applications

In this section we present four applications of the generalised two-filter smoothing algorithm.

5.1 A non-linear time series

Consider the time series problem

$$X_t = \frac{1}{2}X_{t-1} + 25\frac{X_{t-1}}{1+X_{t-1}^2} + 8\cos(1.2\,(t-1)) + V_t \tag{93}$$

$$Y_t = \frac{X_t^2}{20} + W_t, \tag{94}$$

where $X_1 \sim \mathcal{N}(0,5),.V_t \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,15)$ and $W_t \overset{\text{i.i.d.}}{\sim} \mathcal{N}\,(0, 0.01)$.

To enable a comparison between the two (forward–backward and two-filter) smoothing methodologies, 100 Monte Carlo simulations were performed. We chose to use an unscented approximation to the optimal importance function for both the forward and backward passes. We compared the two-filter smoothing algorithm using a time-invariant artificial density given by an approximation of (33) (t-f(1)) and (35) (t-f(2)) by a mixture of three Gaussians together with the forward–backward smoothing algorithm f-b. Note that the same set of samples constructed in the forward filter was used by all three algorithms. Here $\{X_t\}$ does not admit an invariant density because of the cyclical component in (93) but it is still possible to select (35).

In Table 1, we display the effective sample size (ESS) for varying numbers of samples averaged over 100 Monte Carlo simulations for t-f(1) and f-b. The ESS provides a measure of the degeneracy of the particle set and a larger ESS value is desired; see Liu and Chen (1998) for example. The SMC t-f(1) algorithm provides significantly higher ESS and is thus expected to yield lower-variance estimate that the SMC f-b algorithm; similar results were obtained for t-f(2). This is confirmed by evaluating the average RMS error values for the minimum mean square estimate (MMSE) for these simulations are displayed in Table 2. It is clear to see that the two-filter SMC algorithms converges to the MMSE estimate faster than the forward–backward SMC algorithm. Moreover, the use of either approximation for γ using the transition density appears to provide a reasonable artificial density in this particular case.

Table 1 Average ESS values for 100 Monte Carlo runs of 50 time epochs

N	f-b	t-f(1)
50	34.8	47.2
100	67.7	94.3
500	327.9	472.2
1000	645.2	940.2

Table 2 Average RMS values for 100 Monte Carlo runs of 50 time epochs

N	f-b	t-f(1)	t-f(2)
50	90.14	41.34	41.03
100	86.40	41.66	40.38
500	81.12	43.56	45.14
1000	83.36	39.73	39.99

5.2 A non-linear diffusion process

Consider a scalar diffusion process

$$\mathrm{d}X_t = \alpha(X_t)\mathrm{d}t + \mathrm{d}B_t; \quad t \in [0, T] \tag{95}$$

where $\mathrm{d}B_t$ is a Brownian motion. As in the previous sections, we assume that we observe the process at discrete times $t_1, t_2, \ldots$ in some additive white Gaussian noise. In Fearnhead et al. (2008a) it is shown that is possible to write the transition density of the diffusion process as

$$f(x_{t_{k+1}}|x_{t_k}) = \mathcal{N}(x_{t_{k+1}} - x_{t_k}; 0, \Delta t_{k+1} I_d) \exp\left\{A(x_{t_{k+1}}) - A(x_{t_k})\right\} a(x_{t_k}, x_{t_{k+1}}), \tag{96}$$

where $A(u) = \int_u \alpha(z)\mathrm{d}z$ is the anti-derivative of α, $\Delta t_{k+1} = t_{k+1} - t_k$, and

$$a(x_{t_k}, x_{t_{k+1}}) = \mathbb{E}_{\mathbb{W}^{(x_{t_k}, x_{t_{k+1}})}}\left[\exp\left\{-\frac{1}{2}\int_{t_k}^{t_{k+1}} (\alpha^2 + \alpha')(\omega_s)\mathrm{d}s\right\}\right]. \tag{97}$$

Here $\mathbb{W}^{(x_{t_k}, x_{t_{k+1}})}$ denotes a Brownian bridge starting at x_{t_k} and finishing at $x_{t_{k+1}}$. It is not possible to estimate $f(x_{t_{k+1}}|x_{t_k})$ exactly but it is shown in Fearnhead et al. (2008a) how to obtain a positive unbiased estimate of this quantity. Hence, it is possible to obtain an unbiased estimate of the weights given by (46) appearing in the generalised two-filter smoothing algorithm

$$\begin{aligned} \widehat{W}^{(j)}_{t_{k+1}|T} \propto \widetilde{W}^{(j)}_{t_{k+1}} \times \sum_{i=1}^{N} \frac{W^{(i)}_{t_k}}{\gamma_t\left(\widetilde{X}^{(j)}_{t_{k+1}}\right)} \mathcal{N}(\widetilde{X}^{(j)}_{t_{k+1}} - X^{(i)}_{t_k}; 0, \Delta t_{k+1} I_d) \\ \times \exp\left\{A(\widetilde{X}^{(j)}_{t_{k+1}}) - A(X^{(i)}_{t_k}) - l\Delta t\right\} r(x_{t_k}, x_{t_{k+1}}) \end{aligned} \tag{98}$$

where $r(x_{t_k}, x_{t_{k+1}})$ is the realisation of a random variable with mean $a(x_{t_k}, x_{t_{k+1}})$; see Fearnhead et al. (2008a) for details. Obtaining an unbiased estimate of the importance weights is sufficient to ensure that our smoothing estimates are (asymptotically) consistent.

Note that using the forward filtering-backward smoothing recursion in this context would be much more complex as the transition density appears both at the numerator and the denominator of (11); hence we cannot obtain easily an unbiased estimate of (11). This means that the generalised two-filter smoothing algorithm is the only practical algorithm for use within this class of models.

We apply this algorithm to the following diffusion process considered in Fearnhead et al. (2008a)

$$\mathrm{d}X_t = \sin(X_t)\mathrm{d}t + \mathrm{d}B_t. \tag{99}$$

Since the diffusion is a Langevin diffusion, we are able to take the invariant density as the artificial density in the backward filter. We assume that this diffusion is observed in some additive white Gaussian noise of variance $\sigma^2 = 1$. The simulated diffusion, data, and results (in the form of residuals), based on $N = 1{,}000$ samples, can be found in Fig. 1. In this example we use the prior density as the proposal densities in the forward and backward filters, respectively. Clearly, we should be able to improve performance through the the incorporation of the measurement information into the proposal density as done in Fearnhead et al. (2008a).

5.3 Parameter estimation through EM for stochastic volatility

In many cases of interest, the state–space model (1)–(2) depends on (additional) unknown parameters $\theta \in \Theta$, i.e.

$$X_t | X_{t-1} = x_{t-1} \sim f_\theta(\cdot | x_{t-1}), \tag{100}$$

$$Y_t | X_t = x_t \sim g_\theta(\cdot | x_t), \tag{101}$$

with $X_1 \sim \mu_\theta$. To estimate θ given $y_{1:T}$, we propose to maximise the log-likelihood

$$\log p_\theta\,(y_{1:T}) = \log p_\theta(y_1) + \sum_{t=2}^{T} \log p_\theta(y_t | y_{1:t-1})\,. \tag{102}$$

A direct maximization of the likelihood is typically complex and so we utilize the standard Expectation-Maximization algorithm instead. This iterative algorithm proceeds as follows: given a current estimate $\theta^{(i-1)}$ of θ then

$$\theta^{(i)} = \underset{\theta \in \Theta}{\arg\max}\; Q\big(\theta^{(i-1)}, \theta\big)$$

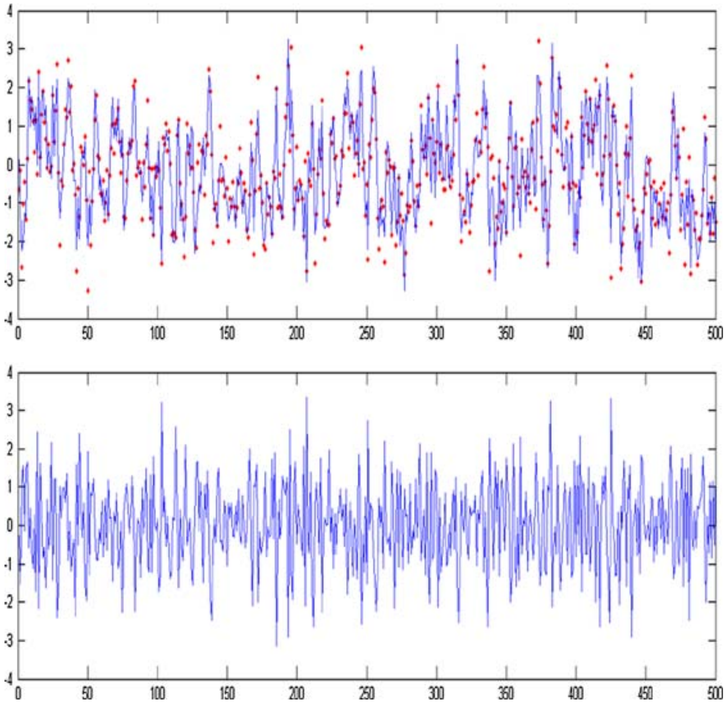

Fig. 1 *Top* Smoothed MAP estimate (*solid line*) and observations Y_t (*dots*). *Bottom* Residuals

where

$$Q\left(\theta^{(i-1)}, \theta\right) = \int \log\left(p_\theta\left(x_{1:T}, y_{1:T}\right)\right) p_{\theta^{(i-1)}}\left(x_{1:T} \mid y_{1:T}\right) \mathrm{d}x_{1:T}. \tag{103}$$

The EM algorithm guarantees that $Q\left(\theta^{(i)}, \theta\right) \geq Q\left(\theta^{(i-1)}, \theta\right)$ making it a popular technique. Note than when $Q\left(\theta^{(i-1)}, \theta\right)$ is approximated numerically using SMC methods then we cannot ensure this property. When the complete data density is from the exponential family, then computing $Q\left(\theta^{(i-1)}, \theta\right)$ only requires evaluating expectations of the form $\mathbb{E}_{\theta^{(i-1)}}\left[\varphi_1\left(x_t\right) \mid y_{1:T}\right], \mathbb{E}_{\theta^{(i-1)}}\left[\varphi_2\left(x_{t-1}, x_t\right) \mid y_{1:T}\right]$ and $\mathbb{E}_{\theta^{(i-1)}}\left[\varphi_3\left(x_t, y_t\right) \mid y_{1:T}\right]$. This can be done using any smoothing technique approximating the marginal densities $p(x_t|y_{1:T})$ and $p\left(x_{t-1:t} \mid y_{1:T}\right)$. We can compute $p\left(x_{t-1:t} \mid y_{1:T}\right)$ using a straightforward generalization of the two-filter formula which we omit here.

We propose here an application of this parameter estimation algorithm to a stochastic volatility model which can be written as follows

$$X_t = \theta_1 X_{t-1} + \theta_2 V_t, \; X_1 \sim \left(0, \frac{\theta_2^2}{1-\theta_1^2}\right) \tag{104}$$

$$Y_t = \theta_3 \exp\left(X_t/2\right) W_t, \tag{105}$$

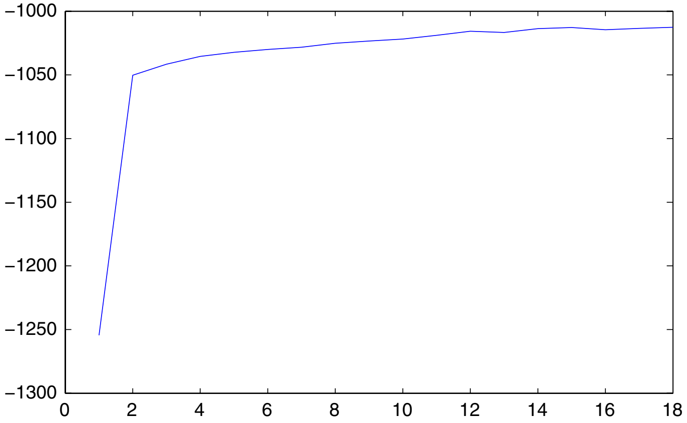

Fig. 2 Sequence of log-likelihood values with $N = 500$; Iteration number (x-axis) and log-likelihood (y-axis)

where $V_t \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, 1)$ and $W_t \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, 1)$. This process is stationary and we use for the artificial densities the invariant density. As X_1 follows the invariant density of the process, we cannot maximise Q in closed form but the EM is still a valid maximization procedure if we just obtain $\theta^{(i)}$ such that $Q\left(\theta^{(i-1)}, \theta^{(i)}\right) \geq Q\left(\theta^{(i-1)}, \theta\right)$. We used for $\theta^{(i)}$

$$\theta_1^{(i)} = \frac{\sum_{t=2}^{T} \mathbb{E}_{\theta^{(i-1)}}\left[X_{t-1}X_t \middle| y_{1:T}\right]}{\sum_{t=1}^{T-1} \mathbb{E}_{\theta^{(i-1)}}\left[X_t^2 \middle| y_{1:T}\right]}, \tag{106}$$

$$\begin{aligned}\theta_2^{(i)} = &\left((T-1)^{-1}\left(\sum_{t=2}^{T} \mathbb{E}_{\theta^{(i-1)}}[X_t^2|y_{1:T}] + \theta_1^{(i)2} \sum_{t=2}^{T} \mathbb{E}_{\theta^{(i-1)}}[X_{t-1}^2|y_{1:T}]\right.\right.\\ &\left.\left. - 2\theta_1^{(i)} \sum_{t=2}^{T} \mathbb{E}_{\theta^{(i-1)}}[X_{t-1}X_t|y_{1:T}]\right)\right)^{1/2},\end{aligned} \tag{107}$$

$$\theta_3^{(i)} = \left(T^{-1} \sum_{t=1}^{T} y_t^2 \mathbb{E}_{\theta^{(i-1)}}\left[\exp\left(-X_t\right) \middle| y_{1:T}\right]\right)^{1/2} \tag{108}$$

which corresponds to the maximum of a modified Q function where the initial state is discarded and we checked that $Q\left(\theta^{(i-1)}, \theta^{(i)}\right) \geq Q\left(\theta^{(i-1)}, \theta\right)$. We first tested the algorithm on a simulated dataset of length 100 and used $N = 500$ particles. In Fig. 2, we display the log-likelihood against iteration number averaged over 100 Monte Carlo runs.

5.4 Blind deconvolution of seismic signals

In several problems related to seismic signal processing and nuclear science, the signal of interest can be modelled as the output of a linear filter excited by a

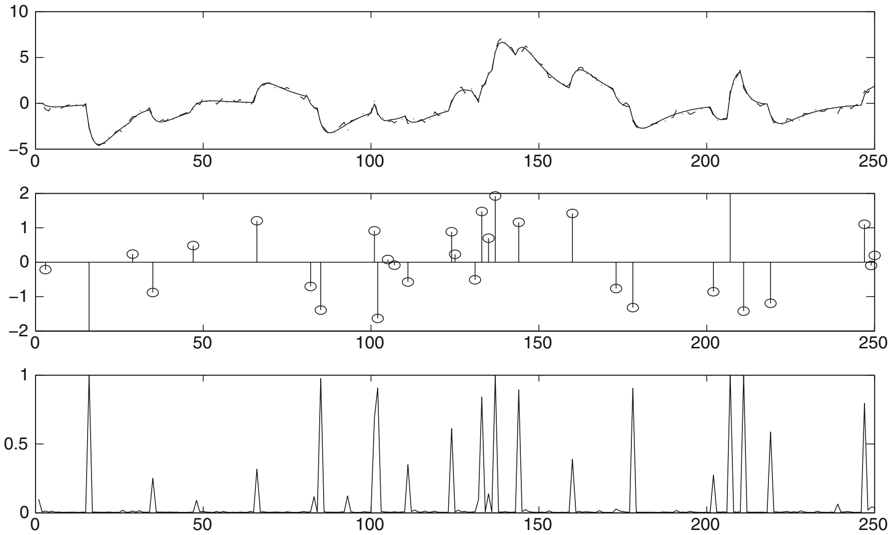

Fig. 3 *Top* Simulated signal X_t (*solid line*) and observations Y_t (*dotted line*). *Middle* Simulated sequence v'_t. *Bottom* Smoothed posterior estimates $p(A_t = 1|y_{1:T})$

Bernoulli–Gaussian (BG) process and observed in white Gaussian noise; see Cappé et al. (1999). This section provides an example of the detection of a Bernoulli–Gauss process related to simulated data from the aforementioned application domain using the conditionally Linear Gaussian smoothing algorithm framework.

The noise process on the input sequence is distributed according to $\lambda\mathcal{N}(0, \sigma_v^2) + (1-\lambda)\delta_0,\ 0 < \lambda < 1$, with δ_0 the delta-Dirac measure in 0. The observation noise is distributed according to $\mathcal{N}(0, \sigma_w^2)$. It is algorithmically convenient to introduce a hidden Bernoulli process $A_t \in \{1, 2\}$ such that $P(A_t = 1) = \lambda$. By modelling the linear filter using an AR(2) model, the signal admits the following conditionally Gaussian state–space model representation, the parameters of (50) and (51) are as follows

$$H = \begin{pmatrix} c_1 & c_2 \\ 1 & 0 \end{pmatrix}, \quad K = (1\ \ 0), \quad L = \sigma_w^2. \tag{109}$$

since H, K, and L are not dependent upon the value of the latent discrete process, and

$$J(a_t = 1) = (\sigma_v^2\ \ 0)^T, \quad J(a_t = 2) = (0\ \ 0)^T. \tag{110}$$

In the following simulations, we set the parameters to $c_1 = 1.51$, $c_2 = -0.55$, $\sigma_w = 0.25$, and $\sigma_v = 0.50$. $T = 250$ observations are generated and the exemplar data set is shown in Fig. 3. At the bottom of this figure is the smoothed posterior estimates of $p(A_t = 1|y_{1:T})$ for $N = 50$. It is clear to see that the algorithm is able to detect all significant changepoints. Increasing the number of particles did not appear to improve the results.

6 Discussion

SMC approximations to generalised two-filter smoothing provide a non-iterative alternative to MCMC to perform Bayesian inference in non-linear non-Gaussian state–space models. This approach performs experimentally significantly better than the alternative non-iterative forward filtering-backward smoothing approach. It is also more widely applicable as it allows us to deal easily with scenarios where only unbiased estimates of the target densities are available such as for partially observed diffusions. Compared to MCMC, the main advantage of this approach will be for multimodal situations where SMC methods provide typically a better exploration of the space than MCMC thanks to the exploratory abilities of a large number of interacting particles.

Acknowledgments The authors would like to thank the Royal Commission for the Exhibition of 1851. This research was supported by the UK MOD through work funded by the Defence Technology Centre for Data and Information Fusion, and the Engineering and Physical Sciences Research Council (EPSRC), UK. The authors are also grateful to Omiros Papaspiliopoulos for having given us implementation details about the SMC algorithm discussed in Fearnhead et al. (2008a).

References

Andrieu, C., Doucet, A. (2002). Particle filtering for partially observed Gaussian state–space models. *Journal of the Royal Statistical Society B, 64*, 4, 827–836.

Bresler, Y. (1986). Two-filter formula for discrete-time non-linear Bayesian smoothing. *International Journal of Control, 43*, 2, 629–641.

Briers, M., Doucet, A., Maskell, S. (2004). Smoothing algorithms for state–space models, Technical report, Cambridge University CUED/F-INFENG.TR. 498.

Briers, M., Doucet, A., Singh, S. S. (2005). Sequential auxiliary particle belief propagation. In *Eighth international conference of information fusion*.

Cappé O., Doucet, A., Moulines, E., Lavielle, M. (1999). Simulation-based methods for blind maximum-likelihood filter identification. *Signal Processing, 73*, 1, 3–25.

Chen, R., Liu, J. S. (2000). Mixture Kalman filters. *Journal of the Royal Statistical Society B, 62*, 493–508.

de Jong, P. (1997). The scan sampler. *Biometrika, 84*, 929–937.

Del Moral, P. (2004). *Feynman-Kac formulae: Genealogical and interacting particle systems with applications, series probability and applications*. New York: Springer.

Doucet, A., Andrieu, C. (2001). Iterative algorithms for state estimation of jump Markov linear systems. *IEEE Transactions on Signal Processing, 49*, 6, 1216–1227

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

Doucet, A., de Freitas, J. F. G., Gordon, N. J. (Eds.) (2001). *Sequential Monte Carlo methods in practice*. New York: Springer.

Fearnhead, P., Papaspiliopoulos, O., Roberts, G. O. (2008a). Particle filters for partially-observed diffusions. *Journal of the Royal Statistical Society B, 70*, 4, 755–777.

Fearnhead, P., Wyncoll, D., Tawn, J. (2008b). A sequential smoothing algorithm with linear computational cost. Technical report, Department of Statistics, Lancaster University.

Godsill, S. J., Doucet, A., West, M. (2004). Monte Carlo smoothing for non-linear time series. *Journal of the American Statistical Association, 99*, 156–168.

Helmick, R. E., Blair, W. D., Hoffman, S. A. (1995). Fixed-interval smoothing for Markovian switching systems. *IEEE Transactions on Information Theory, 41*, 6, 1845–1855.

Isard, M., Blake, A. (1998). A smoothing filter for Condensation. In *Fifth European conference on computer vision* (pp. 767–781).

Kitagawa, G. (1987). Non-Gaussian state–space modeling of nonstationary time series. *Journal of the American Statistical Association, 82*, 1032–1063.

Kitagawa, G. (1994). The two-filter formula for smoothing and an implementation of the Gaussian-sum smoother. *Annals of the Institute of Statistical Mathematics, 46*, 4, 605–623.

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

Kitagawa, G., Sato, S. (2001). Monte Carlo smoothing and self-organizing state–space model. In A. Doucet, J. F. G. de Freitas, N. J. Gordon (Eds.), *Sequential Monte Carlo methods in practice* (pp. 177–195). New York: Springer.

Klaas, M., Briers, M., De Freitas, N., Doucet, A., Maskell, S., Lang, D. (2006). Fast particle smoothing: If I had a million particles. In *International conference on machine learning*.

Liu, J. S., Chen, R. (1998). Sequential Monte Carlo methods for dynamic systems. *Journal of the American Statistical Association, 93*, 1032–1044.

Mayne, D. Q. (1966). A solution of the smoothing problem for linear dynamic systems, *Automatica, 4*, 73–92.