An Overview of Sequential Monte Carlo Methods for Parameter Estimation in General State-Space Models

N. Kantas * **A. Doucet** ** **S.S. Singh** * **J.M. Maciejowski** *

* *Cambridge University Engineering Dept., Cambridge CB2 1PZ, UK*
** *The Institute of Statistical Mathematics, Tokyo 106-8569, Japan*

Abstract: Nonlinear non-Gaussian state-space models arise in numerous applications in control and signal processing. Sequential Monte Carlo (SMC) methods, also known as Particle Filters, provide very good numerical approximations to the associated optimal state estimation problems. However, in many scenarios, the state-space model of interest also depends on unknown static parameters that need to be estimated from the data. In this context, standard SMC methods fail and it is necessary to rely on more sophisticated algorithms. The aim of this paper is to present a comprehensive overview of SMC methods that have been proposed to perform static parameter estimation in general state-space models. We discuss the advantages and limitations of these methods.

Keywords: parameter estimation, general state-space models, hidden Markov models, sequential Monte Carlo

1. INTRODUCTION

Let $\{X_n\}_{n\geq 0}$ and $\{Y_n\}_{n\geq 0}$ be $\mathcal{X}$ ($\subseteq \mathbb{R}^{n_x}$) and $\mathcal{Y}$ ($\subseteq \mathbb{R}^{n_y}$)-valued stochastic processes defined on a (measurable) space $(\Omega, \mathcal{F})$. The discrete-time process $\{X_n\}_{n\geq 0}$ is a hidden (or latent) Markov process of initial density $\mu_\theta(x)$ and Markov transition density $f_\theta(x'|x)$. We only have access to the observation process $\{Y_n\}_{n\geq 0}$. The observations $\{Y_n\}_{n\geq 0}$ are assumed conditionally independent given $\{X_n\}_{n\geq 0}$ and are characterized by the conditional marginal density $g_\theta(y|x)$. To summarize, we have

$$\begin{array}{c} X_0 \sim \mu_\theta(\cdot)\,,\ X_n|(X_{n-1}=x_{n-1}) \sim f_\theta(\cdot|x_{n-1}), \\ Y_n|(X_1,\ldots,X_n=x_n,\ldots,X_T) \sim g_\theta(\cdot|x_n). \end{array} \quad (1)$$

The subscript θ on these densities is the parameter of the model. We will assume a parameterization such that $\theta \in \Theta \subset \mathbb{R}^{n_\theta}$ and Θ is open. All densities are taken with respect to appropriate dominating measures, e.g. the Lebesgue measure. This class of models includes many nonlinear and non-Gaussian time series models such as

$$X_{n+1} = \psi_\theta(X_n, V_{n+1})\,,\ Y_n = \phi_\theta(X_n, W_n)\,, \quad (2)$$

where $\{V_n\}_{n\geq 1}$ and $\{W_n\}_{n\geq 0}$ are independent sequences of independent random variables and $(\psi_\theta,\ \phi_\theta)$ are a pair of nonlinear functions. These models are known as general state-space models or hidden Markov models in the literature [8, 17]. They are ubiquitous in applied science and are commonly used in control, signal processing, econometrics, robotics, telecommunications etc.

When the parameter θ is *known*, on-line (resp. off-line) inference about the state process $\{X_n\}$ given the observations $\{Y_n\}$ is a so-called optimal filtering (resp. smoothing) problem. Except for simple models such as the linear Gaussian state-space model, or when $\mathcal{X}$ is a finite set, these optimal state estimation problems cannot be solved exactly. Standard approximation schemes such as the Extended Kalman filter or the Gaussian sum filter can be unreliable, while deterministic integration methods are difficult to implement. SMC methods, also known as *particle methods*, are a class of sequential simulation-based algorithms to approximate the posterior distributions of interest. Their wide spread popularity are due to the fact that they are easy to implement, suitable for parallel implementation and more importantly, have been demonstrated in numerous settings to yield more accurate estimates than the standard alternatives just mentioned [14, 17, 35].

The main objective of this paper is to discuss the scenario where the parameter θ is *unknown* and needs to be estimated from the data either in an on-line or off-line manner. We will assume that the observations are generated from the unknown 'true' model with parameter value θ^*, i.e. $X_n|(X_{n-1}=x_{n-1}) \sim f_{\theta^*}(\cdot|x_{n-1})$ and $Y_n|(X_n=x_n) \sim g_{\theta^*}(\cdot|x_n)$. The static parameter estimation problem has generated a lot of interest over the past few years and many SMC techniques have been proposed to solve it. In this review, we attempt to give insights on the difficulties of this task and provide a comprehensive overview of the literature on the subject. We will present the main features of each method and comment on their pros and cons. No attempt however is made to discuss the intricacies of the specific implementations. For this we refer the reader to the original references.

We have chosen to broadly classify the methods as follows:

- Bayesian or Maximum Likelihood (ML).
- Off-line (batch) or on-line (recursive).

In a Bayesian approach, the unknown parameter is considered random and assigned a suitable prior distribution.

⋆ The first and last authors were supported by the European Commission under project iFly FP6- TREN-037180.

The posterior density of this parameter given the observations is to be characterised. In Maximum Likelihood, the estimate of θ^* is the maximizing argument of the (marginal) likelihood of the observed data. Both these estimation techniques can be implemented off-line or on-line. Specifically, in an off-line framework we infer θ^* by iterating over a fixed observation record $y_{0:T}$. In contrast, on-line methods update the estimate of θ^* sequentially as observations $\{y_n\}_{n\geq 0}$ become available.

The rest of the paper is organized as follows. In Section 2 we review SMC methods for filtering and smoothing when the parameter θ is *known*. In Section 3, we explain how these techniques can be used to perform off-line and on-line ML estimation. In Section 4, we describe algorithms to perform off-line and on-line Bayesian inference. Finally, in Section 5 we discuss the main advantages and drawbacks of the methods presented here.

2. SMC FILTERING AND SMOOTHING

2.1 Preliminaries

Assume for the time being that the parameter θ is *known*. Given observed data $y_{0:n}$ [1], inference about the states $X_{0:n}$ may be based on the following posterior density:

$$p_\theta\left(x_{0:n}|\, y_{0:n}\right) = \frac{p_\theta\left(x_{0:n}, y_{0:n}\right)}{p_\theta\left(y_{0:n}\right)} \tag{3}$$

where

$$p_\theta\left(x_{0:n}, y_{0:n}\right) = \mu_\theta\left(x_0\right)\prod_{k=1}^{n} f_\theta\left(x_k|\, x_{k-1}\right)\prod_{k=0}^{n} g_\theta\left(y_k|\, x_k\right) \tag{4}$$

and the *marginal likelihood*, $p_\theta\left(y_{0:n}\right)$, is given by

$$p_\theta\left(y_{0:n}\right) = \int p_\theta\left(x_{0:n}, y_{0:n}\right) dx_{0:n}. \tag{5}$$

It is easy to check that

$$p_\theta\left(x_{0:n}|\, y_{0:n}\right) = p_\theta\left(x_{0:n-1}|\, y_{0:n-1}\right)\frac{f_\theta\left(x_n|\, x_{n-1}\right) g_\theta\left(y_n|\, x_n\right)}{p_\theta\left(y_n|\, y_{0:n-1}\right)} \tag{6}$$

and

$$p_\theta\left(y_{0:n}\right) = p_\theta\left(y_{0:n-1}\right) p_\theta\left(y_n|\, y_{0:n-1}\right) \tag{7}$$

where

$$p_\theta\left(y_n|\, y_{0:n-1}\right) = \int f_\theta\left(x_n|\, x_{n-1}\right) g_\theta\left(y_n|\, x_n\right) \times p_\theta\left(x_{n-1}|\, y_{0:n-1}\right) dx_{n-1:n}. \tag{8}$$

2.2 SMC filtering

In this section we described a SMC algorithm to numerically approximate the sequence of posterior densities $\{p_\theta\left(x_{0:n}|\, y_{0:n}\right)\}$.

Algorithm In SMC, the distribution of interest is approximated by a large cloud of N ($N >> 1$) random samples termed particles. These particles are propagated over time using an importance sampling and resampling mechanisms. The algorithm relies on the introduction of so-called importance densities: $q_\theta\left(x_0|\, y_0\right)$ at time 0 and $q_\theta\left(x_n|\, y_n, x_{n-1}\right)$ at times $n \geq 1$. A default choice consists of taking $q_\theta\left(x_0|\, y_0\right) = \mu_\theta\left(x_0\right)$ and $q_\theta\left(x_n|\, y_n, x_{n-1}\right) = f_\theta\left(x_n|\, x_{n-1}\right)$. The "optimal" choice is given by $p_\theta\left(x_1|\, y_1\right)$ and $p_\theta\left(x_n|\, y_n, x_{n-1}\right)$ and in practice it is recommended to approximate these distributions if it is not possible to sample from them; see [16, 41] for various approximation strategies. We define the importance weights

$$w_0\left(x_0\right) = \frac{\mu_\theta\left(x_0\right) g_\theta\left(y_0|\, x_0\right)}{q_\theta(\, x_0|\, y_0)}, \tag{9}$$

$$w_n\left(x_{n-1:n}\right) = \frac{f_\theta\left(x_n|\, x_{n-1}\right) g_\theta\left(y_n|\, x_n\right)}{q_\theta\left(x_n|\, y_n, x_{n-1}\right)} \text{ for } n \geq 1. \tag{10}$$

Note that in order to alleviate the notational burden we adopt below the convention that whenever the index i is used we mean 'for all $i \in \{1, ..., N\}$,' and also omit the dependence of the importance weights on θ - we will do so in the remainder of the paper when no confusion is possible. The algorithm can be summarized as follows.

SMC for Filtering

At time $n = 0$

- Sample $X_0^i \sim q_\theta(\, x_0|\, y_0)$.
- Compute the weights $w_0\left(X_0^i\right)$ and set $W_0^i \propto w_1\left(X_0^i\right)$, $\sum_{i=1}^N W_0^i = 1$.
- Resample $\left\{W_0^i, X_0^i\right\}$ to obtain N equally-weighted particles $\left\{\frac{1}{N}, \overline{X}_0^i\right\}$.

At time $n \geq 1$

- Sample $X_n^i \sim q_\theta(\, x_n|\, y_n, \overline{X}_{n-1}^i)$ and set $X_{0:n}^i \leftarrow \left(\overline{X}_{0:n-1}^i, X_n^i\right)$.
- Compute the weights $w_n\left(X_{n-1:n}^i\right)$ defined in (10) and set $W_n^i \propto w_n\left(X_{n-1:n}^i\right)$, $\sum_{i=1}^N W_n^i = 1$.
- Resample $\left\{W_n^i, X_{0:n}^i\right\}$ to obtain N new equally-weighted particles $\left\{\frac{1}{N}, \overline{X}_{0:n}^i\right\}$.

At time n, the approximations of $p_\theta\left(x_{0:n}|\, y_{0:n}\right)$ and $p_\theta\left(y_n|\, y_{0:n-1}\right)$ after the sampling step are

$$\widehat{p}_\theta\left(dx_{0:n}|\, y_{0:n}\right) = \sum_{i=1}^{N} W_n^i \delta_{X_{0:n}^i}\left(dx_{0:n}\right), \tag{11}$$

$$\widehat{p}_\theta\left(y_n|\, y_{0:n-1}\right) = \frac{1}{N}\sum_{i=1}^{N} w_n\left(X_{n-1:n}^i\right). \tag{12}$$

Hence an estimate of the marginal likelihood, by (7), is given by

$$\widehat{p}_\theta\left(y_{0:n}\right) = \widehat{p}_\theta\left(y_0\right)\prod_{k=1}^{n}\widehat{p}_\theta\left(y_k|\, y_{0:k-1}\right). \tag{13}$$

After the resampling step, an alternative approximation of $p_\theta\left(x_{0:n}|\, y_{0:n}\right)$ is

$$\overline{p}_\theta\left(dx_{0:n}|\, y_{0:n}\right) = \frac{1}{N}\sum_{i=1}^{N}\delta_{\overline{X}_{0:n}^i}\left(dx_{0:n}\right). \tag{14}$$

The resampling procedure is introduced to copy (or multiply) particles with high weights and therefore discard-

[1] For any sequence $\{z_n\}$, let $z_{i:j} = (z_i, z_{i+1}, \ldots, z_j)$.

ing particles with low weights. The resampling procedure should satisfy the following unbiasedness property:

$$\mathbb{E}\left[\overline{p}_\theta\left(dx_{0:n}|\, y_{0:n}\right)\right] = \widehat{p}_\theta\left(dx_{0:n}|\, y_{0:n}\right). \tag{15}$$

The resampling procedure serves to focus the computational effort on the "promising" regions of the state-space. The simplest resampling scheme is the following: generate N independent samples from $\widehat{p}_\theta\left(dx_{0:n}|\, y_{0:n}\right)$, i.e., $\overline{X}^i_{0:n} \sim \widehat{p}_\theta\left(dx_{0:n}|\, y_{0:n}\right)$. This is called multinomial resampling. Each original particle $X^i_{0:n}$ inherits N^i_n offspring and the joint distribution of $\left(N^1_n, \ldots, N^N_n\right)$ is a multinomial distribution with parameter $\left(W^1_n, \ldots, W^N_n\right)$. Better resampling schemes that have less variance have been proposed, as have more advanced SMC algorithms with better overall performance. Examples of the latter include the Auxiliary Particle Filter (APF) [41] and the Resample-Move algorithm [23].

Convergence results Many sharp convergence results are available for SMC algorithms; see [12] for a survey for practitioners and [14] for a book length review. A selection of these results that give useful insights on the difficulties of estimating static parameters with SMC are presented below.

Let $\epsilon_{\theta,n}\left(dx_{0:n}\right) = \widehat{p}_\theta\left(dx_{0:n}|\, y_{0:n}\right) - p_\theta\left(dx_{0:n}|\, y_{0:n}\right)$. If $w_0\left(x_0\right)$ and $w_n\left(x_{n-1:n}\right)$ are upper bounded in their arguments then it can be shown that for any bounded test function $\varphi_n : \mathcal{X}^{n+1} \to \mathbb{R}$, there exists constants $C_{\theta,n,p} < \infty$ such that for any $p > 0$,

$$\mathbb{E}\left[\left|\int \varphi_n(x_{0:n})\epsilon_{\theta,n}\left(dx_{0:n}\right)\right|^p\right]^{\frac{1}{p}} \leq \frac{C_{\theta,n,p}\overline{\varphi_n}}{N^{1/2}}, \tag{16}$$

where $\overline{\varphi_n} = \sup_{x_{0:n}\in\mathcal{X}^{n+1}} |\varphi_n(x_{0:n})|$. This result is very weak as typically $C_{\theta,n,p}$ grows exponentially/polynomially with n. Hence to guarantee a fixed precision of the SMC approximation we would need to increase the number of particles as n increases. It appears impossible to establish a result where $C_{\theta,n,p}$ is independent of n. This is intuitively not surprising as the dimension of the target density $p_\theta\left(x_{0:n}|\, y_{0:n}\right)$ we are approximating is increasing with n. Moreover the successive resampling steps lead to a depletion of the particle population; $p_\theta\left(x_{0:m}|\, y_{0:n}\right)$ will eventually be approximated by a single unique particle as $n - m$ increases. This is known in the literature as the **degeneracy** problem. This is a fundamental weakness of SMC: given a fixed number of particles N, it is impossible to approximate $p_\theta\left(x_{0:n}|y_{0:n}\right)$ "well" when n is large. Typically as soon as n is of order N.

Fortunately, it is also possible to establish much more encouraging results. Many state-space models possess the so-called *exponential forgetting* property. This property states that for any $x_0, x'_0 \in \mathcal{X}$ and observation record $y_{0:n}$,

$$\int \left|p_\theta\left(x_n|\, y_{0:n}, x_0\right) - p_\theta\left(x_n|\, y_{0:n}, x'_0\right)\right| dx_n \leq C\lambda^n, \tag{17}$$

where $\lambda \in [0, 1)$ and C is a constant. When exponential forgetting holds, it is possible to establish results of the following form. For an integer $L > 0$ and any bounded test function $\varphi_L : \mathcal{X}^L \to \mathbb{R}$, there exists constants $D_{\theta,L,p} < \infty$ such that for any $p > 0$

$$\mathbb{E}\left[\left|\int \varphi_L(x_{n-L+1:n})\epsilon_{\theta,L}\left(dx_{n-L+1:n}\right)\right|^p\right]^{\frac{1}{p}} \leq \frac{D_{\theta,L,p}\overline{\varphi_L}}{N^{1/2}}, \tag{18}$$

where $\epsilon_{\theta,L}\left(dx_{n-L+1:n}\right) = \int_{\mathcal{X}^{n-L+1}} \epsilon_{\theta,n}\left(dx_{0:n}\right)$. This result explains why SMC is an effective computational tool. If we only look at the most recent marginal density over a fixed horizon $p_\theta\left(x_{n-L+1:n}|\, y_{0:n}\right)$, then there is no accumulation of errors over time if the 'true' optimal filter we are trying to approximate has 'good' mixing properties. Under the same assumption (17), it is possible to derive a central limit theorem for the marginal likelihood estimate $\widehat{p}_\theta\left(y_{0:n}\right)$ and show that its asymptotic relative variance only degrades linearly with the time index n, that is there exists C_θ such that [14]

$$\frac{\mathbb{V}(\widehat{p}_\theta\left(y_{0:n}\right))}{\left(p_\theta\left(y_{0:n}\right)\right)^2} \leq C_\theta\frac{n}{N}. \tag{19}$$

On the contrary, even if (17) holds, then the asymptotic variance of the SMC estimate of the additive functional

$$I^\theta_n = \int \left[\sum_{k=0}^{n} \varphi\left(x_k\right)\right] p_\theta\left(x_{0:n}|\, y_{0:n}\right) dx_{0:n}, \tag{20}$$

which is

$$\widehat{I}^\theta_n = \int \left[\sum_{k=0}^{n} \varphi\left(x_k\right)\right] \widehat{p}_\theta\left(dx_{0:n}|\, y_{0:n}\right), \tag{21}$$

satisfies [44]

$$\mathbb{V}\left(\widehat{I}^\theta_n\right) \geq D_\theta\frac{n^2}{N}. \tag{22}$$

This negative result follows from the particle degeneracy problem.

2.3 SMC smoothing

We have seen previously that SMC methods can provide an approximation of the sequence of densities $\left\{p_\theta\left(x_{0:n}|\, y_{0:n}\right)\right\}$, and hence for $\left\{p_\theta\left(x_k|\, y_{0:n}\right); k = 0, ..., n\right\}$. However these approximations are poor when n is large because of the degeneracy problem. Various alternative schemes which do not suffer from these problems have been proposed in the literature.

Fixed-lag approximation The fixed-lag approximation is the simplest approach and it was first proposed in [29]. It relies on the fact that, for state-space models with "good" forgetting properties (e.g. (17)), we have

$$p_\theta\left(x_{0:n}|\, y_{0:T}\right) \approx p_\theta\left(x_{0:n}|\, y_{0:\min(n+\Delta,T)}\right) \tag{23}$$

for Δ large enough; that is observations collected at times $k > n + \Delta$ do not bring any additional information about $X_{0:n}$. This suggests a very simple scheme — simply don't update the estimate of $X_{0:n}$ after time $k = n + \Delta$. Algorithmically, this means do not resample the components $X^i_{0:n}$ of the particles $X^i_{0:k}$ at times $k > n + \Delta$. 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 $p_\theta\left(x_{0:n}|\, y_{0:\min(n+L,T)}\right)$ is a poor approximation of $p_\theta\left(x_{0:n}|\, y_{0:T}\right)$. If we select a large values of L to ensure that $L \geq \Delta$ then the degeneracy problem remains substantial. Moreover, even as $N \to \infty$, this SMC approximation will have a fixed bias since $p_\theta\left(x_{0:n}|\, y_{0:T}\right) \neq p_\theta\left(x_{0:n}|\, y_{0:\min(n+\Delta,T)}\right)$.

Forward filtering-backward smoothing It can be easily established that

$$p_\theta(x_n|y_{0:T}) = \int \frac{f_\theta(x_{n+1}|x_n)}{p_\theta(x_{n+1}|y_{0:n})} p_\theta(x_{n+1}|y_{0:T})\,dx_{n+1} \times p_\theta(x_n|y_{0:n}). \quad (24)$$

So using the particle approximations of $\widehat{p}_\theta(dx_n|y_{0:n}) = \sum_{i=1}^N W_n^i \delta_{X_n^i}(dx_n)$ and plugging them in (24), we obtain a particle approximation of $p_\theta(x_n|y_{0:T})$ of the form $\widehat{p}_\theta(dx_n|y_{0:T}) = \sum_{i=1}^N W_{n|T}^i \delta_{X_n^i}(dx_n)$ where the weights $\left\{W_{n|T}^i\right\}$ satisfy the following backward recursion

$$W_{n|T}^i = W_n^i \left(\sum_{j=1}^N W_{n+1|T}^j \frac{f_\theta(X_{n+1}^j|X_n^i)}{\sum_{l=1}^N W_n^l f_\theta(X_{n+1}^i|X_n^l)} \right) \quad (25)$$

and $W_{T|T}^i = W_T^i$. See [16] for a derivation. This approach is more efficient than the direct SMC approximation of the smoothing densities outlined at the beginning of this section. However the problem of such methods is that it provides a Monte Carlo approximation of the smoothed distributions which relies on the same particles $\left\{X_n^{(i)}\right\}$ used to approximate the filtered distributions; it only re-weights these samples. Hence, if $p_\theta(x_n|y_{0:n})$ and $p_\theta(x_n|y_{0:T})$ have high probability masses in different regions of the space, then the Monte Carlo approximation will have a high variance for reasonable values of N. The forward filtering-backward sampling approach requires also $\mathcal{O}\left(N^2T\right)$ operations to approximate $\{p_\theta(x_n|y_{0:T})\}$ instead of $\mathcal{O}(NT)$ for the direct and fixed-lag methods.

Generalized two-filter smoothing The two-filter formula is a well-established alternative to the forward-filtering backward-smoothing technique to compute $\{p_\theta(x_n|y_{0:T})\}$. It relies on the following identity

$$p_\theta(x_n|y_{0:T}) = \frac{p_\theta(x_n|y_{0:n-1})\, p_\theta(y_{n:T}|x_n)}{p_\theta(y_{n:T}|y_{0:n-1})}, \quad (26)$$

where $p_\theta(y_{n:T}|x_n) = \int p_\theta(y_{n:T}, x_{n+1:T}|x_n)\,dx_{n+1:T}$ is the so-called backward information filter. The backward information filter is not a probability density in argument x_n and it is possible that $\int p_\theta(y_{n:T}|x_n)\,dx_n = \infty$. Although this is not an issue when $p_\theta(y_{n:T}|x_n)$ can be computed exactly, it does preclude the direct use of SMC methods to estimate this integral. To address this problem, a generalized version of the two-filter formula was proposed in [7]. It relies on the introduction of a set of artificial probability densities $\{\widetilde{p}_{\theta,n}(x_n)\}_{n=1}^T$ and the joint densities

$$\widetilde{p}_\theta(x_{n:T}|y_{n:T}) \propto \widetilde{p}_{\theta,n}(x_n)\, p_\theta(y_{n:T}, x_{n+1:T}|x_n), \quad (27)$$

which are constructed such that their marginal densities, $\widetilde{p}_\theta(x_n|y_{n:T}) \propto \widetilde{p}_{\theta,n}(x_n)\, p_\theta(y_{n:T}|x_n)$, are simply "integrable" versions of the backward information filter. It is easy to establish the following generalized two-filter formula:

$$p_\theta(x_n|y_{0:T}) \propto \left[\widetilde{p}_{\theta,n}(x_n)\right]^{-1} \widetilde{p}_\theta(x_n|y_{n:T}) \times \int f_\theta(x_n|x_{n-1})\, p_\theta(x_{n-1}|y_{0:n-1})\,dx_{n-1}. \quad (28)$$

It is possible to use a modified SMC method to perform an approximation of $\{\widetilde{p}_\theta(x_{n:T}|y_{n:T})\}$, say $\widehat{\widetilde{p}}_\theta(dx_{n:T}|y_{n:T}) = \sum_{i=1}^N \widetilde{W}_n^i \delta_{\widetilde{X}_{n:T}^i}(dx_{n:T})$; see [7] for details. Then by using (28), we have $\widehat{p}_\theta(dx_n|y_{0:T}) = \sum_{i=1}^N W_{n|T}^i \delta_{\widetilde{X}_n^i}(dx_n)$ where

$$W_{n|T}^j \propto \widetilde{W}_n^j \sum_{i=1}^N W_{n-1}^i \frac{f_\theta\left(\widetilde{X}_n^j|X_{n-1}^i\right)}{\widetilde{p}_{\theta,n}\left(\widetilde{X}_n^j\right)}. \quad (29)$$

Like the SMC implementation of the forward-backward smoothing algorithm, this approach has a computational complexity $\mathcal{O}(N^2T)$. However fast computational methods have been developed to address this problem [30]. Moreover it is possible to reduce this computational complexity to $\mathcal{O}(NT)$ by using rejection sampling to sample from $\widehat{p}_\theta(dx_n|y_{0:T})$ using $\widehat{p}_\theta(dx_{n-1}|y_{0:n-1})$ and $\widehat{\widetilde{p}}_\theta(dx_n|y_{n:T})$ as proposal distributions if $f_\theta(x'|x)/\widetilde{p}_{\theta,n}(x') < C < \infty$. More recently, an importance sampling type approach has also been proposed in [22] to reduce the computational complexity to $\mathcal{O}(NT)$; see [6] for a similar idea developed in the context of belief propagation. It is demonstrated experimentally in [7] that this procedure outperforms significantly the forward filtering-backward smoothing approach.

3. MAXIMUM LIKELIHOOD PARAMETER ESTIMATION

3.1 Off-line Methods

In Maximum Likelihood, the estimate of θ^* is the maximizing argument of the marginal likelihood of the observed data:

$$\widehat{\theta} = \arg\max_{\theta\in\Theta} \; l_T(\theta) \quad (30)$$

where

$$l_T(\theta) = \log\, p_\theta(y_{0:T}). \quad (31)$$

Likelihood function evaluation For any $\theta \in \Theta$, it is possible to evaluate $l_T(\theta)$ numerically using SMC. It can be shown that $\widehat{p}_\theta(y_{0:T})$ in (13) is an unbiased estimate but $\widehat{l_T}(\theta) = \log \widehat{p}_\theta(y_{0:T})$ is not. A standard bias correction technique can be applied though; see for example [3, 40]. The SMC estimate $\widehat{l_T}(\theta)$ is not a continuous function of θ even when the particle filter is implemented with common random numbers for different values of θ. This is because of the resampling steps. Specifically, in multinomial resampling, a piecewise constant and hence discontinuous cumulative distribution function (cdf) is defined by the weights $\{W_n^i\}_{i=1}^N$ and particles $\{X_n^i\}_{i=1}^N$. A small change in θ will cause a small change in the importance weights $\{W_n^i\}_{i=1}^N$ and this will potentially generate a different set of resampled particles $\{\overline{X}_n^i\}_{i=1}^N$. As a result, the likelihood function estimate will not be continuous in θ. A solution to this problem was proposed in [25] by using an importance sampling method. The proposed method has computational complexity $\mathcal{O}\left(N^2\right)$ and suffers from being only valid in the neighborhood of a suitably preselected parameter value.

When $\mathcal{X} \subseteq \mathbb{R}$, an elegant solution to the discontinuity problem was proposed in [40]. The method uses common random numbers. The resampling operation is made "smooth"by ordering the particles $\{X_n^i\}_{i=1}^N$ and defining a piecewise linear resampling cdf; see [40] for details. This method requires $\mathcal{O}(N \log N)$ operations to sort the particles. The extension to the case when $\mathcal{X} \subseteq \mathbb{R}^{n_x}$ (and $n_x > 1$) is proposed in [33]. Although this method does not generate a continuous estimate of the likelihood function,

it does provide a much "smoother" estimate than the standard approach, which means $\widehat{l_T}(\theta)$ is easier to maximize. The computational complexity is $\mathcal{O}(n_x N \log N)$.

Thus far the methods discussed only provide estimates of the likelihood function. When θ is high-dimensional, the optimization over the parameter space may be made more efficient with estimates of the gradient as well. This is the subject of the next section.

Gradient approach When gradient methods are considered we will assume all functions are regular enough so that the change of order of integration and differentiation is permitted. The log-likelihood may be maximized with the following steepest ascent algorithm:

$$\theta_{k+1} = \theta_k + \gamma_{k+1} \nabla_\theta l_T(\theta)|_{\theta=\theta_k}. \tag{32}$$

The gradient (denoted by ∇) is taken w.r.t. the parameter θ, as indicated by the subscript of the gradient operator. $\{\gamma_k\}$ is a sequence of small positive real numbers, called the step-size sequence, that should satisfy the following constraints: $\sum_k \gamma_k = \infty$ and $\sum_k \gamma_k^2 < \infty$. One possible choice would be $\gamma_k = k^{-\alpha}$, $0.5 < \alpha < 1$ (e.g. $\gamma_k = k^{-2/3}$).

To obtain the gradient of the log likelihood, which is so-called *score*, we can use Fisher's identity:

$$\begin{aligned}&\nabla_\theta l(\theta) = \int \nabla_\theta \log p_\theta(x_{0:T}, y_{0:T})\, p_\theta(x_{0:T}|\, y_{0:T})\, dx_{0:T} \\ &= \int (\nabla_\theta \log \mu_\theta(x_0) + \nabla_\theta \log g_\theta(y_0|\, x_0))\, p_\theta(x_0|\, y_{0:T})\, dx_0 \\ &+ \sum_{n=1}^{T} \int (\nabla_\theta \log f_\theta(x_n|\, x_{n-1}) + \nabla_\theta \log g_\theta(y_n|\, x_n)) \\ &p_\theta(x_{n-1:n}|\, y_{0:T})\, dx_{n-1:n}.\end{aligned} \tag{33}$$

The simplest method to estimate the score numerically would be to use the SMC approximation of $p_\theta(x_{0:T}|\, y_{0:T})$ in (11) and re-weight the samples using $\nabla_\theta \log p_\theta(x_{0:T}, y_{0:T})$ [3]. However, the variance of this estimate increases typically quadratically with T [44]. To improve over this direct method, we can use the fixed-lag approximation (which however introduces a bias). These two approaches admit on-line implementations. Alternatively, we can use the forward filtering backward smoothing and the generalized two filter smoothing methods which can be extended straightforwardly to approximate the marginals $\{p_\theta(x_{n-1:n}|\, y_{0:T})\}$. These methods do not admit on-line implementations. An alternative to Fisher's identity to compute the score is a method based on Infinitesimal Perturbation Analysis which has been recently proposed [10]. This method is also estimating the expectation with respect to $p_\theta(x_{0:T}|\, y_{0:T})$ of an additive functional of the form $s(x_0, y_0) + \sum_{n=1}^{T} s(x_{n-1:n}, y_n)$ so all the SMC smoothing techniques described earlier can also be applied to estimate this expectation.

Expectation-Maximization The Expectation Maximization (EM) algorithm for maximizing $l_T(\theta)$ is a two step procedure, [15]. The first step, the expectation or E-step, computes

$$\begin{aligned}&Q(\theta_k, \theta) = \int \log p_\theta(x_{0:T}, y_{0:T})\; p_{\theta_k}(x_{0:T}|y_{0:T}) dx_{0:T} \\ &= \int (\log \mu_\theta(x_0) + \log g_\theta(y_0|\, x_0))\, p_{\theta_k}(x_0|\, y_{0:T})\, dx_0 \\ &+ \sum_{n=1}^{T} \int (\log f_\theta(x_n|\, x_{n-1}) + \log g_\theta(y_n|\, x_n)) \\ &p_{\theta_k}(x_{n-1:n}|\, y_{0:T})\, dx_{n-1:n}.\end{aligned} \tag{34}$$

The second step is the maximization or M-step that updates the parameter θ_k,

$$\theta_{k+1} = \arg\max_\theta \; Q(\theta_k, \theta) \tag{35}$$

The sequence $\{l_T(\theta_k)\}_k$ generated by the EM is non-decreasing, i.e. $l_T(\theta_{k+1}) \geq l_T(\theta_k)$.

All the SMC smoothing techniques presented earlier can also be used to approximate $Q(\theta_k, \theta)$ numerically; see [3] for the direct method, [39] for the fixed-lag approximation, [48] for the forward filtering-backward smoothing and [7] for generalized two-filter smoothing. When $p_\theta(x_{0:T}, y_{0:T})$ is in the exponential family, $p_\theta(x_{0:T}, y_{0:T})$ depends on $(x_{0:T}, y_{0:T})$ only through a set of fixed dimensional sufficient statistics of the form $\sum_{n=1}^{T} s_\theta(x_{n-1:n}, y_n)$. In this case, the M step can typically be performed explicitly. Since $Q(\theta_k, \theta)$ is being approximated numerically, it cannot be guaranteed that log-likelihood will increase monotonically. Having said that, these algorithms display good performance when the number of particles is large enough.

Discussion In some cases, one might prefer an steepest ascent procedure over the Expectation-Maximization (EM) algorithm for a number of reasons. Firstly, if the step-size sequence γ_{k+1} were to be replaced by $-\gamma_{k+1}\Gamma_k^{-1}$ where Γ_k is Hessian of $l_T(\theta)$ evaluated at θ_k (which can be computed using SMC techniques, see [43], [44]) then the rate of convergence is quadratic and thus faster than the EM which converges linearly. The second reason is that the gradient algorithm can be implemented even when the M-step of the EM cannot be solved in closed-form (see [32] for several examples of this.). On the other hand one might prefer an EM approach if the M-step can be computed analytically. Scaling the gradients for a large n_θ might be quite hard. In addition, the EM is numerically more stable and typically computationally cheaper for high dimensional parameter θ. Finally, both methods are locally optimal. They are thus sensitive to initialization and might get trapped in a local maximum. It is recommended that multiple runs of the algorithms from different initial conditions are used.

Iterated filtering An alternative approach for off-line ML estimation has been proposed in [26]. At iteration k, let θ_k be estimate of the true parameter θ^*. The key idea is that $\nabla l_T(\theta_k)$ can be approximated by the *posterior* moments of an artificial state-space model with latent Markov process $\{Z_n = (X_n, \widetilde{\theta}_n)\}_{n=0}^T$,

$$\widetilde{\theta}_{n+1} = \widetilde{\theta}_n + \varepsilon_{n+1}, \quad X_{n+1} \sim f_{\widetilde{\theta}_n}(\cdot|x_n) \tag{36}$$

and observed process $Y_n \sim g_{\widetilde{\theta}_n}(\cdot|x_n)$. $\{\varepsilon_n\}_{n\geq 1}$ is a zero-mean white noise sequence with variance $\sigma^2\Sigma$, $\mathbb{E}(\widetilde{\theta}_0) = \theta_k$, $\mathbb{V}(\widetilde{\theta}_0) = \tau^2\Sigma$. It is shown in [26] that $\nabla l_T(\theta_k)$ can be approximated using $\{\mathbb{E}(\widetilde{\theta}_n|y_{0:n}), \mathbb{V}(\widetilde{\theta}_n|y_{0:n})\}_{n=0}^T$ and this approximation improves as $\sigma^2, \tau^2 \to 0$. Having estimated $\{\mathbb{E}(\widetilde{\theta}_n|y_{0:n}), \mathbb{V}(\widetilde{\theta}_n|y_{0:n})\}_{n=0}^T$ using SMC, the estimate θ_k can be improved using a gradient ascent method.

Clearly as the variance of the additive noise $\{\varepsilon_n\}$ decreases, it will be necessary to use more particles as the mixing properties of the artificial dynamic model deteriorates. [26] provides some conditions ensuring that this iterative procedure converges towards a local maximum of the likelihood function. An advantage of this procedure over standard gradient and EM techniques is that it only

requires being able to sample from $f_\theta(x'|x)$ and there is no explicit calculations of the derivative. However, it might require a bit of tuning when the parameter is high-dimensional.

3.2 On-line Methods

For a long observation sequence the computation of the gradient of $l_T(\theta)$ can be prohibitive, moreover we might have real-time constraints. An alternative would be a recursive procedure in which the data is run through once sequentially. If θ_n is the estimate of the model parameter after $n-1$ observations, a recursive method would update the estimate to θ_{n+1} after receiving the new data y_n. Several recursive estimation procedures are now described.

Gradient approach A standard approach to perform on-line parameter estimation is the following gradient method:

$$\theta_{n+1} = \theta_n + \gamma_{n+1} \nabla_\theta \log p_{\theta_{0:n}}(y_n|y_{0:n-1}). \tag{37}$$

where $\{\gamma_n\}_{n\geq 1}$ is the step-size sequence as in Section 3. Upon receiving y_n, θ_n is updated in the direction of ascent of the predictive density of this new observation. A necessary requirement for on-line implementation is that the previous values of estimates, other than θ_n, are also used in the evaluation of $\nabla_\theta \log p_\theta(y_n|y_{0:n-1})$ at $\theta = \theta_n$. This is indicated in the notation $\nabla_\theta \log p_{\theta_{0:n}}(y_n|y_{0:n-1})$. (Not doing so would require browsing through the entire history of observations.) This approach has previously appeared in [34, 11] for the finite state-space case. This algorithm maximizes the average log-likelihood,

$$\overline{l}(\theta) = \lim_{n\to\infty} \frac{1}{n} l_n(\theta), \tag{38}$$

where

$$\overline{l}(\theta) = \int_{\mathcal{Y}\times\mathcal{P}(\mathcal{X})} \log\left(\int g_\theta(y|x)\mu(x)dx\right) \lambda_{\theta,\theta^*}(dy, d\mu). \tag{39}$$

Here $\mathcal{P}(\mathcal{X})$ is the space of probability distributions on $\mathcal{X}$, and $\lambda_{\theta,\theta^*}(dy, d\mu)$ is the marginal of the invariant distribution the Markov chain $\{X_n, Y_n, p_\theta(x_n|Y_{0:n-1})\}_{n\geq 0}$. (See [34] for conditions ensuring the existence of $\overline{l}(\theta)$.) It is easy to check that the set of global maxima of $\overline{l}(\theta)$ includes indeed θ^*. The asymptotic properties of this algorithm (i.e. the behavior of θ_n in the limit as n goes to infinity) have been studied in the case of an i.i.d. hidden process by [47] and for a HMM with a finite state-space in [34].

For general state-space models, it is impossible to compute $\nabla_\theta \log p_\theta(y_n|y_{0:n-1})$ exactly, which motivates the use of SMC methods. Noting that $\nabla_\theta \log p_\theta(y_n|y_{0:n-1})$ is equal to

$$\nabla_\theta \log p_\theta(y_{0:n}) - \nabla_\theta \log p_\theta(y_{0:n-1}),$$

it would be possible to use Fisher's identity (33) with the particle approximation of $p_\theta(x_{0:n}|y_{0:n})$ in (11) to compute this gradient on-line. However, the variance of the estimate of $\nabla_\theta \log p_\theta(y_{0:n})$ would increase at least quadratically with n; see Section 2 and [44]. A high and increasing variance estimate of $\nabla_\theta \log p_\theta(y_n|y_{0:n-1})$ would render the stochastic optimization algorithm unreliable. The fixed-lag approximation could be used, but it introduces a bias which is difficult to control.

An alternative approach to compute $\nabla_\theta \log p_\theta(y_{0:n})$ has been proposed in [43]. It relies on the "marginal" Fisher identity

$$\nabla_\theta \log p_\theta(y_{0:n}) = \int \nabla_\theta \log p_\theta(x_n, y_{0:n}) \; p_\theta(x_n|y_{0:n})\, dx_n. \tag{40}$$

The advantage of this identity over (33) is that it only relies on the SMC approximation of the marginal $p_\theta(x_n|y_{0:n})$. However it requires approximating $\nabla_\theta \log p_\theta(x_n, y_{0:n})$ which is given by the ratio of the following two equations:

$$\begin{aligned} &p_\theta(y_{0:n-1})^{-1} \nabla_\theta p_\theta(x_n, y_{0:n}) = g_\theta(y_n|x_n) \times \\ &\int [\nabla_\theta \log p_\theta(x_{n-1}, y_{0:n-1}) + \nabla_\theta \log f_\theta(x_n|x_{n-1}) \\ &+\nabla_\theta \log g_\theta(y_n|x_n)] f_\theta(x_n|x_{n-1})\, p_\theta(x_{n-1}|y_{0:n-1})\, dx_{n-1} \end{aligned} \tag{41}$$

and

$$\begin{aligned} &p_\theta(y_{0:n-1})^{-1} p_\theta(x_n, y_{0:n}) = g_\theta(y_n|x_n) \times \\ &\int f_\theta(x_n|x_{n-1})\, p_\theta(x_{n-1}|y_{0:n-1})\, dx_{n-1}. \end{aligned} \tag{42}$$

Given an SMC approximation $\frac{1}{N}\sum_{i=1}^N \delta_{\overline{X}^i_{n-1}}(dx_{n-1})$ of $p_\theta(x_{n-1}|y_{0:n-1})$ after the resampling step, we obtain the following pointwise approximation of $\nabla_\theta \log p_\theta(x_n, y_{0:n})$

$$\widetilde{\nabla_\theta \log p_\theta}(x_n, y_{0:n}) = \frac{\widetilde{\nabla_\theta p_\theta}(x_n, y_{0:n})}{\widetilde{p_\theta}(x_n, y_{0:n})} \tag{43}$$

where

$$p_\theta(y_{0:n-1})^{-1} \widetilde{p_\theta}(x_n, y_{0:n}) = \frac{g_\theta(y_n|x_n)}{N} \sum_{i=1}^N f_\theta(x_n|\overline{X}^i_{n-1}), \tag{44}$$

$$\begin{aligned} p_\theta(y_{0:n-1})^{-1} \widetilde{\nabla_\theta p_\theta}(x_n, y_{0:n}) = \frac{g_\theta(y_n|x_n)}{N} \sum_{i=1}^N f_\theta(x_n|\overline{X}^i_{n-1}) \\ \times [\widetilde{\nabla_\theta \log p_\theta}\left(\overline{X}^i_{n-1}, y_{0:n-1}\right) + \nabla_\theta \log f_\theta(x_n|\overline{X}^i_{n-1}) \\ +\nabla_\theta \log g_\theta(y_n|x_n)]. \end{aligned} \tag{45}$$

Using $\widehat{p}_\theta(dx_n|y_{0:n}) = \sum_{i=1}^N W^i_n \delta_{X^i_n}(dx_n)$, it follows that an alternative SMC estimate of $\nabla_\theta \log p_\theta(y_{0:n})$ is given by

$$\widehat{\nabla_\theta \log p_\theta}(y_{0:n}) = \sum_{i=1}^N W^i_n \widetilde{\nabla_\theta \log p_\theta}\left(X^i_n, y_{0:n}\right) \tag{46}$$

Experiments have confirmed that the variance of this estimate of the score only increases linearly with n and the variance of the estimate of $\nabla_\theta \log p_\theta(y_n|y_{0:n-1})$ is uniformly bounded (in time) [44]. Hence this procedure appears not to suffer from the degeneracy problem faced by standard techniques. The price to pay is that its computational cost is $\mathcal{O}\left(N^2\right)$ per time step.

It is straightforward to modify this $\mathcal{O}\left(N^2\right)$ algorithm to compute the gradient recursively as required in (37) and we refer the reader to [43, 44] for details. Finally, this algorithm was used to successfully perform high-dimensional parameter estimation in a robotics application [38].

Expectation-Maximization A potential criticism of the on-line gradient approach discussed above is that it can be difficult to properly scale the gradient components especially when n_θ is large. As an alternative we can use an on-line version of the EM algorithm. We assume that

$p_\theta(x_{0:n}, y_{0:n})$ is in the exponential family. In the off-line approach, we needed to compute, at iteration k,

$$Q(\theta_k, \theta) = \int \log p_\theta(x_{0:T}, y_{0:T})\ p_{\theta_k}(x_{0:T}|y_{0:T})dx_{0:T} \quad (47)$$

and then maximize this function (with respect to θ) to obtain θ_{k+1}, i.e.

$$\theta_{k+1} = \Lambda\left(\mathcal{S}_k\right) \quad (48)$$

where Λ is a deterministic mapping and $\mathcal{S}_k$ is a set of sufficient statistics of the form

$$\mathcal{S}_k = \frac{1}{T}\int \left(\sum_{n=1}^{T} s\left(x_{n-1:n}, y_n\right)\right) p_{\theta_k}(x_{0:T}|y_{0:T})dx_{0:T}. \quad (49)$$

In the on-line approach, we simply use

$$\theta_{n+1} = \Lambda\left(\overline{\mathcal{S}}_n\right) \quad (50)$$

where

$$\begin{aligned}\overline{\mathcal{S}}_n = \gamma_{n+1} &\int s\left(x_{n-1:n}, y_n\right) p_{\theta_{0:n}}(x_{n-1:n}|y_{0:n})dx_{n-1:n} \\ &+ (1-\gamma_{n+1}) \int \left(\sum_{k=1}^{n-1} s\left(x_{k-1:k}, y_k\right)\right) \\ &\times p_{\theta_{0:n}}(x_{0:n-1}|y_{0:n})dx_{0:n-1} \end{aligned} \quad (51)$$

We could use the direct SMC approach to approximate $\overline{\mathcal{S}}_n$ on-line but the variance of the estimate would increase over time. The fixed-lag approximation could also be used but it introduces a bias which is difficult to control. In [18], the authors propose an alternative $\mathcal{O}\left(N^2\right)$ per time step method to compute $\overline{\mathcal{S}}_n$. Experimentally, it appears that the variance of this estimate is uniformly bounded over time. A theoretical analysis of this algorithm is currently being undertaken.

Discussion Experimentally the $\mathcal{O}(N^2)$ algorithms provide estimates of $\nabla_\theta \log p_\theta(y_n|y_{0:n-1})$ (gradient [44]) or $\overline{\mathcal{S}}_n$ (EM [18]) whose variance is uniformly bounded in time for numerous models. The key is that they only rely on the SMC approximation of the marginal $p_\theta\left(x_n|\, y_{0:n}\right)$. Hence they are much more robust than algorithms that rely on the SMC approximation of $p_\theta\left(x_{0:n}|\, y_{0:n}\right)$ which degenerates over time. Alternatively, the fixed-lag approximation could have been used but it introduces a bias which can be difficult to quantify.

It is important to note that the on-line gradient and on-line EM algorithms can be used in off-line applications where a large dataset is available. For a sequence of observations $y_{0:T}$, provided T is large, "convergence" will typically occur before T, i.e., $\theta_n \approx \overline{\theta}$ where $\overline{\theta}$ is a local maximum of the likelihood for $n < T$. Of course this will depend on the choice of step-size, how well the log-likelihood discriminates between the competing choice of models as well as the starting point θ_0 of the algorithm. One advantage of these on-line algorithms for a finite dataset of length T is the computational savings of not having to browse through the whole dataset repeatedly either to compute the gradient of the likelihood or the Q function. For smaller datasets, these algorithms can be used by going through the data say K times. Typically this method is cheaper than iterating (32) K times the off-line algorithms and can yield comparable parameter estimates.

In practice, it can be beneficial to start with a constant but small step-size $\gamma_n = \gamma$. If the step-size decreases too quickly in the first time steps, these algorithms might get stuck at an early stage and fail to come close to a local maximum of the likelihood.

Online Pseudo-Likelihood Estimation The previous on-line gradient and on-line EM algorithms have a computational load of $\mathcal{O}(N^2)$ per time step. To bypass this problem, [2] modify the function to be minimized. In particular, they suggest the use of a pseudo likelihood function which had been proposed earlier in [45] for finite state-space HMM.

Assume that the state process $\{X_n\}$ defined by (1) is *stationary* with a *known* invariant distribution ν_θ. The data set is divided into blocks, each containing L observations, i.e. $y_{0:L-1}, y_{L:2L-1}, \ldots, y_{(p-1)L:pL-1}$. Let $\mathsf{y}_p = y_{(p-1)L:pL-1}$. Similarly for the latent process, let $\mathsf{X}_p = X_{(p-1)L:pL-1}$. Due to the stationarity assumption the joint process $\{\mathsf{X}_p, \mathsf{y}_p\}_{p\geq 1}$ is identically distributed with the common distribution given by

$$\begin{aligned} p_\theta\left(\mathsf{x}_p, \mathsf{y}_p\right) = \nu_\theta\left(x_{(p-1)L}\right) g_\theta\left(y_{(p-1)L}\middle|\, x_{(p-1)L}\right) \\ \times \prod_{l=(p-1)L+1}^{pL-1} f_\theta\left(x_l|\, x_{l-1}\right) g_\theta\left(y_l|\, x_l\right). \end{aligned} \quad (52)$$

Define the marginal pseudo likelihood (SDL) of p blocks of observations as [2]:

$$p_\theta\left(\mathsf{y}_{1:p}\right) = \prod_{k=1}^{p} p_\theta\left(\mathsf{y}_k\right) \quad (53)$$

where

$$p_\theta\left(\mathsf{y}_p\right) \triangleq \int_{\mathcal{X}^L} p_\theta\left(\mathsf{y}_p, \mathsf{x}_p\right) d\mathsf{x}_p. \quad (54)$$

An on-line EM is proposed in [2] to maximize this pseudo-likelihood. The main advantage of this approach is that it only requires an approximation of the fixed-dimensional distributions $p_\theta\left(\mathsf{x}_k|\, \mathsf{y}_k\right)$. SMC methods can be used and do not suffer the degeneracy problem as long as L is not large, say L in the range from 5 to 30. The computational cost will then be $\mathcal{O}(LN)$ per on-line EM step. This scheme however requires knowledge of the stationary distribution ν_θ. When it is not available, [2] propose an alternative approach based on indirect inference.

4. BAYESIAN PARAMETER ESTIMATION

In the Bayesian setting, we choose a suitable prior density $p\left(\theta\right)$ for θ and compute the joint posterior density $p\left(x_{0:T}, \theta|\, y_{0:T}\right)$ in the off-line case, or the sequence of posterior densities $\{p\left(x_{0:n}, \theta|\, y_{0:n}\right)\}$ in the on-line setting.

4.1 Off-line Methods

Particle Markov chain Monte Carlo A standard approach in statistics to approximate $p\left(x_{0:T}, \theta|\, y_{0:T}\right)$ is to use MCMC. Unfortunately it is quite difficult to design efficient MCMC sampling algorithms for non-linear non-Gaussian state-space models. Particle MCMC (PMCMC) are a new class of MCMC techniques which rely on SMC methods to build efficient high dimensional proposal distributions. Although this sounds a natural idea, technically it is a difficult task to formulate and it has only appeared very recently [1]. We limit ourselves here to the

presentation of the Particle Marginal Metropolis-Hastings (PMMH) sampler, which is an approximation of an ideal MMH sampler for sampling from $p(x_{0:T}, \theta | y_{0:T})$. The ideal MMH sampler would utilize the following proposal density:

$$q((x'_{0:T}, \theta') | (x_{0:T}, \theta)) = q(\theta' | \theta) p_{\theta'}(x'_{0:T} | y_{0:T}) \quad (55)$$

where $q(\theta' | \theta)$ is a proposal density to obtain a candidate θ' when we are at location θ. The acceptance probability of the associated sampler is

$$1 \wedge \frac{p(x'_{0:T}, \theta' | y_{0:T}) q((x_{0:T}, \theta) | (x'_{0:T}, \theta'))}{p(x_{0:T}, \theta | y_{0:T}) q((x'_{0:T}, \theta') | (x_{0:T}, \theta))} \quad (56)$$

$$= 1 \wedge \frac{p_{\theta'}(y_{0:T}) p(\theta') q(\theta | \theta')}{p_{\theta}(y_{0:T}) p(\theta) q(\theta' | \theta)}. \quad (57)$$

Unfortunately this ideal algorithm cannot be implemented as we cannot sample exactly from $p_{\theta'}(x'_{0:T} | y_{0:T})$ and we cannot compute the terms $p_\theta(y_{0:T})$ and $p_{\theta'}(y_{0:T})$ appearing in the acceptance probability. The PMMH sampler is an approximation of this ideal MMH sampler which relies on the SMC approximations of the unknown terms. It proceeds as follows.

PMMH Sampler

At iteration $k = 0$,

- Set $\theta(0)$ arbitrarily.
- Run an SMC algorithm targeting $p_{\theta(0)}(x_{0:T} | y_{0:T})$, sample $X_{0:T}(0) \sim \widehat{p}_{\theta(0)}(dx_{0:T} | y_{0:T})$, and compute the marginal likelihood estimate $\widehat{p}_{\theta(0)}(y_{0:T})$

At iteration $k \geq 1$

- Sample a proposal $\theta' \sim q(\theta | \theta(k-1))$.
- Run an SMC algorithm targeting $p_{\theta'}(x_{0:T} | y_{0:T})$, sample $X'_{0:T} \sim \widehat{p}_{\theta'}(dx_{0:T} | y_{0:T})$, and compute the marginal likelihood estimate $\widehat{p}_{\theta'}(y_{0:T})$.
- Set $\theta(k) = \theta'$, $X_{0:T}(k) = X'_{0:T}$, $\widehat{p}_{\theta(k)}(y_{0:T}) = \widehat{p}_{\theta'}(y_{0:T})$ with probability

$$1 \wedge \frac{\widehat{p}_{\theta'}(y_{0:T}) p(\theta') q(\theta(k-1) | \theta')}{\widehat{p}_{\theta(k-1)}(y_{0:T}) p(\theta(k-1)) q(\theta' | \theta(k-1))}, \quad (58)$$

otherwise set $\theta(k) = \theta(k-1)$, $X_{0:T}(k) = X_{0:T}(k-1)$, $\widehat{p}_{\theta(k)}(y_{0:T}) = \widehat{p}_{\theta(k-1)}(y_{0:T})$.

The remarkable feature of this algorithm is that the invariant distribution of the Markov chain $\{X_{0:T}(k), \theta(k)\}$ is $p(x_{0:T}, \theta | y_{0:T})$ whatever being the number of particles N used in our SMC approximation; i.e. using an SMC approximation does not introduce any bias. However, obviously, the higher N the better the mixing properties of the algorithm. Under favorable mixing assumptions, (19) suggests that the variance of the acceptance rate of the PMMH sampler is proportional to T/N so N should roughly increase linearly with T. This was confirmed experimentally in [1]. In [1], for some difficult inference problems, the authors show that PMMH performs well with minimal tuning, even when the transition density is used as the proposal in the SMC algorithm. This is a very attractive feature since, unlike standard MCMC methods, the user does not need to design sophisticated proposals for the state variables $X_{0:T}$ but only for the parameter θ. A particle version of the Gibbs sampler has also been developed [1].

4.2 On-line Methods

Deficiency of standard SMC At first sight, it seems that estimating the sequence of posterior densities $\{p(x_{0:n}, \theta | y_{0:n})\}_{n \geq 0}$ can be easily achieved using standard SMC methods by merely introducing the extended state $Z_n = (X_n, \theta_n)$ with initial density $p(\theta_0) \mu_{\theta_0}(x_0)$ and transition density $f_{\theta_n}(x_n | x_{n-1}) \delta_{\theta_{n-1}}(\theta_n)$; i.e. $\theta_n = \theta_{n-1}$. Applying a standard SMC algorithm to the Markov process $\{Z_n\}_{n \geq 0}$ means that the parameter space would only be explored at the initialization of the algorithm. As a result of the successive resampling steps, after a certain time n, the approximation $\widehat{p}(d\theta | y_{0:n})$ will only contain a single unique value for θ. This is clearly a bad estimator. Although we can establish results of the form

$$\mathbb{E}\left[\left|\int_\Theta \varphi(\theta) \epsilon_n(d\theta)\right|^p\right]^{\frac{1}{p}} \leq \frac{C_n \overline{\varphi}}{N^{1/2}}, \quad (59)$$

where $\varphi : \Theta \to \mathbb{R}$, $\overline{\varphi} = \sup_{\theta \in \Theta} |\varphi(\theta)|$, $\epsilon_n(d\theta) = \widehat{p}(d\theta | y_{0:n}) - p(d\theta | y_{0:n})$, it is still unsatisfactory as C_n typically grows exponentially/polynomially with n.

As a result on-line Bayesian parameter estimation is typically implemented using advanced SMC algorithms or standard SMC (see Section 2.2) but with introduced artificial dynamics for the fixed parameter. However, we believe that none of the methods presented in this section are satisfactory. This should not come as a surprise. Even for a *fixed* θ, the SMC estimate of $p_\theta(y_{0:n})$ has a relative variance that increases linearly with n under favorable mixing assumptions; see (19). The methods in this section try to approximate $p(\theta | y_{0:n})$ which is given by

$$p(\theta | y_{0:n}) \propto p_\theta(y_{0:n}) p(\theta). \quad (60)$$

This is a hard problem as it implicitly requires having to approximate $p_{\theta^{(i)}}(y_{0:n})$ for all the particles $\{\theta^{(i)}\}$ approximating $p(\theta | y_{0:n})$. Hence we expect all on-line Bayesian methods to provide estimates whose variance will increase at least linearly with n; that is these methods will provide poorer approximations as n increases.

Artificial dynamics A pragmatic approach that might be useful in some applications is to introduce artificial dynamics for the parameter θ [24], [28]:

$$\theta_{n+1} = \theta_n + \varepsilon_{n+1} \quad (61)$$

where $\{\varepsilon_n\}_{n \geq 0}$ is a small (and decreasing with n) artificial dynamic noise. SMC can now be applied to approximate $\{p(x_{0:n}, \theta | y_{0:n})\}$. A related kernel density estimation method also appeared in [37], which proposes to use a kernel density estimate of the target

$$\widehat{p}(\theta | y_{0:n}) = \frac{1}{N} \sum_{i=1}^{N} M_h\left(\theta - \theta_n^{(i)}\right) \quad (62)$$

where M is a convolution kernel, e.g. Gaussian or Epanechnikov, and h the smoothing parameter or width. Then at time $n+1$ samples from (62) to obtain a new set of particles. As before the static parameter is transformed to a slowly time-varying one, whose dynamics is related to the width of the kernel M_h. Modifying the original

problem means that it is hard to quantify how much bias is introduced in the resulting estimates. Also, both these methods require a significant amount of tuning, e.g choosing of the kernel width or the variance of the artificial dynamic noise.

Practical filtering In [42], the authors rely on the following *key* fixed-lag approximation:

$$p\left(x_{0:n-L}, \theta \middle| y_{0:n-1}\right) \approx p\left(x_{0:n-L}, \theta \middle| y_{0:n}\right) \tag{63}$$

for L large enough; that is $x_{0:n-L}$ has very little influence on observations coming after n. To sample approximately from $p\left(\theta \middle| y_{0:n}\right)$, they propose using several MCMC chains in parallel to sample from

$$p\left(x_{n-L+1:n}, \theta \middle| y_{0:n}, X^{(i)}_{0:n-L}\right) \tag{64}$$

$$= p\left(x_{n-L+1:n}, \theta \middle| y_{n-L+1:n}, X^{(i)}_{n-L}\right) \tag{65}$$

which is an approximation to $p\left(x_{n-L+1:n}, \theta \middle| y_{0:n}\right)$. Then they collect the last simulated sample for each chain $X^{(i)}_{n-L+1:n}$, increment the time index and run several MCMC chains in parallel to sample from

$$p\left(x_{n-L+2:n+1}, \theta \middle| y_{n-L+2:n+1}, X^{(i)}_{0:n-L+1}\right) \tag{66}$$

and so on. Like all methods based on fixed-lag approximation, the choice of the lag L is difficult and there is a non-vanishing bias which is difficult to quantify. However, the method seems to perform well on the examples presented by the authors.

Using MCMC steps within SMC algorithms To avoid the introduction of an artificial dynamic model or of a fixed-lag approximation, an approach originally proposed in [23] consists of adding MCMC steps to re-introduce "diversity" among the particles. More precisely, assume that at time n that we have access to an SMC approximation of $p\left(x_{0:n}, \theta \middle| y_{0:n}\right)$ of the form

$$\overline{p}_\theta\left(d\left(x_{0:n}, \theta\right) \middle| y_{0:n}\right) = \frac{1}{N}\sum_{i=1}^{N} \delta_{\left(\overline{X}^i_{0:n}, \overline{\theta}^i_n\right)}\left(d\left(x_{0:n}, \theta\right)\right).$$

To add diversity in this population of particles, we simply use an MCMC kernel with invariant density $p\left(x_{0:n}, \theta \middle| y_{0:n}\right)$, i.e.

$$\left(X^{(i)}_{0:n}, \theta^{(i)}_n\right) \sim K_n\left(\cdot, \cdot \middle| \overline{X}^i_{0:n}, \overline{\theta}^i_n\right)$$

where by construction K_n satisfies

$$p\left(x'_{0:n}, \theta' \middle| y_{0:n}\right) = \int p\left(x_{0:n}, \theta \middle| y_{0:n}\right) K_n\left(x'_{0:n}, \theta' \middle| x_{0:n}, \theta\right) d\left(x_{0:n}, \theta\right).$$

Contrary to standard applications of MCMC, the kernel does not have to be ergodic. It is actually never ergodic in practice as ensuring ergodicity would require one to sample an increasing number of variables over time – the algorithm would not be sequential anymore. In practice one therefore sets $X^{(i)}_{0:n-L} = \overline{X}^i_{0:n-L}$ for some integer $L \geq 1$ and only sample $\theta^{(i)}_n$ and possibly $X^{(i)}_{n-L+1:n}$. Note that the memory requirements for this method does not increase over time if $p\left(y_{0:n} \middle| \theta, x_{0:n}\right)$ can be summarized by a set of fixed dimensional sufficient statistics.

This method was first used in an on-line Bayesian parameter estimation context in [4]. In this paper the authors were using

$$K_n\left(x'_{0:n}, \theta' \middle| x_{0:n}, \theta\right) = \delta_{x_{0:n}}\left(x'_{0:n}\right) p(\theta' | x_{0:n}, y_{0:n}),$$

that is a Gibbs step to update the parameter values. It was used in a context where $p\left(y_{0:n} \middle| \theta, x_{0:n}\right) = p(\theta | s_n(x_{0:n}, y_{0:n}))$, where $s_n(x_{0:n}, y_{0:n})$ is a fixed- dimensional vector of sufficient statistics. In this case, the algorithm is particularly elegant as the memory requirements do not increase over time. Similar strategies were adopted in [21] and [46]. In cases where it is possible to sample from $p_\theta(x_n | y_n, x_{n-1})$ and compute $p_\theta(y_n | x_{n-1})$, it is beneficial to swap the order of the sampling and resampling steps; this is just a particular case of the APF of [41]. This strategy is also beneficial in an on-line Bayesian parameter estimation context and has been discussed in [4, See footnote, page 4] and at length in [31].

Degeneracy issues As opposed to the methods relying on kernel or artificial dynamics, these sufficient statistics/MCMC-based approaches have the advantage of adding diversity to the particles approximating $p\left(\theta | y_{0:n}\right)$ without perturbing the target distribution. Unfortunately, these elegant algorithms are *not* robust. This was noticed in [4]. The reason for this is that these algorithms rely implicitly on the SMC approximation of $p\left(x_{0:n} \middle| y_{0:n}\right)$ whose dimension increases with n. Hence they suffer from the standard degeneracy problem. This poor approximation of $p\left(x_{0:n} \middle| y_{0:n}\right)$ induces a poor approximation of $p(s_n(x_{0:n}, y_{0:n}) | y_{0:n})$ and the errors do built up over time. Results such as (22) suggest that the approximation error increases at least quadratically with n. A convincing example of this degeneracy problem for sufficient statistics is given in [2]. However, these methods cannot completely be ruled out. For small time horizons, a low dimensional parameter space (typically not more than 5-10), an informative prior and a large number of particles, they can perform reasonably well.

5. CONCLUSION

We have reviewed the various SMC algorithms that have been developed over the past ten years for estimating the static parameters of a general state-space model. The degeneracy of the particle population was identified as the main cause of the poor performance of some of these SMC algorithms. Several new algorithms that avoided the particle degeneracy problem and yielded estimates with less variance were proposed at an increased computational cost. A summary of most of the algorithms discussed in this paper, including their advantages/disadvantages and computational cost, is presented in the following table.

REFERENCES

[1] Andrieu C., Doucet, A. , and Holenstein R. (2007) Particle Markov chain Monte Carlo. Technical report, Department of Mathematics, Bristol University.

[2] Andrieu C., Doucet A. and Tadic V.B. (2005) Online parameter estimation in general state-space models, In Proc. IEEE CDC/ECC.

[3] Andrieu C., Doucet A., Singh S.S., Tadić V., (2004) Particle methods for change detection, identification and control, Proceedings of the IEEE, vol. 92, pp. 423 - 438.

[4] Andrieu, C., De Freitas, J.F.G. and Doucet, A. (1999). Sequential MCMC for Bayesian model selection. In Proc. IEEE Workshop on Higher Order Statistics.

Name	Type	References	Pros	Cons	Comp. Cost
Artificial Dynamics	On-line Bayesian	[24, 28, 37]	Standard SMC applicable	Target distribution altered Difficult to tune introduced dynamics	$\mathcal{O}(NT)$
Resample-Move	On-line Bayesian	[4, 46, 21, 31]	Elegant Target distribution unaltered	Restricted model class Degeneracy problem Informative priors n_θ small	$\mathcal{O}(NT)$
Particle MCMC	Off-line Bayesian	[1]	Standard SMC applicable, only requires design of a proposal for θ	Expensive	$\mathcal{O}(NT)$ per MCMC step
Smooth Likelihood Evaluation	Off-line MLE	[40]	Simple to implement Can use standard optimization packages	Requires $n_x = 1$	$\mathcal{O}(N \log N\ T)$ per evaluation
ML (with gradients)	Off-line MLE	[10, 43]	Generally applicable Standard SMC smoothing algorithms applicable	Difficult to tune scaling (especially if n_θ large) Locally optimal	$\mathcal{O}(NT)$ or $\mathcal{O}(N^2T)$ per parameter update
EM	Off-line MLE	[7, 3, 39, 48]	Standard SMC smoothing algorithms applicable	Restricted model class Locally optimal	$\mathcal{O}(NT)$ or $\mathcal{O}(N^2T)$ per parameter update
Iterated Filtering	Off-line MLE	[26]	Standard SMC applicable	Difficult to tune scaling (especially if n_θ large) Locally optimal	$\mathcal{O}(NT)$ per iteration
On-line gradient	On-line MLE	[43, 44]	Asymptotically efficient Generally applicable	Expensive Difficult to tune scaling (especially if n_θ large) Locally optimal	$\mathcal{O}(N^2)$ per parameter update
On-line EM	On-line MLE	[18]	Asymptotically efficient	Restricted model class Expensive Locally optimal	$\mathcal{O}(N^2)$ per parameter update
On-line EM pseudo	On-line pseudo MLE	[2]	Minimal tuning No degeneracy for small L	Requires stationary distribution Loss of asymptotic efficiency	$\mathcal{O}(NL)$ per parameter update

Table 1. Table summarizing the main advantages and disadvantages of the methods described in this paper.

[5] Benveniste, A., Métivier, M. and Priouret, P. (1990). Adaptive Algorithms and Stochastic Approximation. New York: Springer-Verlag.
[6] Briers, M., Doucet, A. and Singh, S.S. (2005) Sequential auxiliary particle belief propagation. In Proc. Conf. Fusion.
[7] Briers M., Doucet A. and Maskell S. R. (2009) Smoothing algorithms for state-space models, Ann. Inst. Stat. Math., to appear.
[8] Cappé, O., Moulines, E. and Rydén, T. (2005) Inference in Hidden Markov Models. New York: Springer Verlag.
[9] Cérou F., LeGland F. and Newton N.J. (2001) Stochastic particle methods for linear tangent equations. In Optimal Control and PDE's - Innovations and Applications (eds. J. Menaldi, E. Rofman & A. Sulem), pp. 231-240, IOS Press, Amsterdam.
[10] Coquelin, A., Deguest, R. and Munos, R. (2008) Perturbation analysis for parameter estimation in continuous space HMMs. Technical report, INRIA-Sequel.
[11] Collings, I.B. and Ryden, T. (1998) A new maximum likelihood gradient algorithm for on-line hidden Markov model identification. In Proc. IEEE ICASSP, pp. 2261-2264.
[12] Crisan D. and Doucet A. (2002) A survey of convergence results on particle filtering for practitioners, IEEE Trans. Signal Processing, vol. 50, pp. 736-746.
[13] De Jong, N. et al. (2009) Efficient likelihood evaluation of state-space representations, Technical report, Department of Economics, Pittsburgh University.
[14] Del Moral P. (2004) Feynman-Kac formulae: genealogical and interacting particle systems with applications. New York: Springer Verlag.
[15] Dempster N.P., Laird N.M. and Rubin D.B. (1977) Maximum likelihood from incomplete data via the EM algorithm, J. R. Stat. Soc. B, vol. 39, pp. 1-38.
[16] Doucet, A., Godsill, S.J. and Andrieu, C. (2000). On sequential Monte Carlo sampling methods for Bayesian filtering. Statist. Comput., vol. 10, pp. 197-208.
[17] Doucet, A., De Freitas, J.F.G. and Gordon N.J. (eds.) (2001). Sequential Monte Carlo Methods in Practice. New York: Springer-Verlag.
[18] Doucet, A. and S.S. Singh (2009) Sequential Monte Carlo methods for additive functionals with applications to recursive parameter estimation. In preparation.
[19] Doucet A., Johansen A. M. (2009) A tutorial on particle filtering and smoothing: Fifteen years later. In Oxford Handbook of Nonlinear Filtering, D. Crisan and B. Rozovsky (eds.). Oxford University Press.
[20] Doucet, A. and Tadić, V.B. (2003). Parameter estimation in general state-space models using particle methods. Ann. Inst. Stat. Math., vol. 55, pp. 409-422.
[21] Fearnhead P. (2002) MCMC, sufficient statistics and particle filter. J. Comp. Graph. Stat., vol. 11, pp. 848-862.
[22] Fearnhead, P., Wyncoll, D. and Tawn, J. (2008) A sequential smoothing algorithm with linear computational cost. Technical report, Department of Mathematics and Statistics, Lancaster University.
[23] Gilks, W. R. and Berzuini, C. (2001) Following a moving target - Monte Carlo inference for dynamic Bayesian models. J.R. Statist. Soc. B, vol. 63, pp. 127-146.
[24] Higuchi, T. (2001) Self-organizing time series model. In [17].
[25] Hürzeler, M. and Künsch, H.R. (2001) Approximation and maximising the likelihood for a general state-space model. In [17].
[26] Ionides, E. L., Bhadra, A. and King, A. A. (2009). Iterated filtering. Technical report, Department of Statistics, University of Michigan.
[27] Kitagawa G. (1996) Monte Carlo filter and smoother for non-Gaussian nonlinear state-space models. J. Comput. Graph. Statist., vol. 5, pp. 1-25.
[28] Kitagawa, G. (1998). A self-organizing state-space model. J. Am. Statist. Ass., vol. 93, pp. 1203-1215.
[29] Kitagawa, G. and Sato, S. (2001) Monte Carlo smoothing and self-organizing state-space model. In [17].
[30] Klaas, M. et al. (2005) Fast particle smoothing: if I had a million particles. In Proc. International Conference on Machine Learning.
[31] Johannes, M. and Polson, N.G. (2007) Particle filtering and parameter learning. Technical report, Chicago business school.
[32] Lange, K. (1995) A gradient algorithm locally equivalent to the EM algorithm, J. R. Statist. Soc. B, vol. 57, pp. 425–437.
[33] Lee, A. (2008) Towards smoother multivariate particle filters. M.Sc. Computer Science, University of British Columbia.
[34] Le Gland F. and Mevel L. (1997) Recursive identification in hidden Markov models, In Proc. 36th IEEE Conf. Decision and Control, pp. 3468-3473.
[35] Liu, J. S. (2001). Monte Carlo Strategies in Scientific Computing. New York: Springer-Verlag.
[36] Liu J.S. and Chen R. (1998) Sequential Monte Carlo methods for dynamic systems. J. Am. Statist. Ass., vol. 93, pp. 1032-1044.
[37] Liu J. and West M. (2001) Combined parameter and state estimation in simulation-based filtering, In [17].
[38] Martinez-Cantin, R., Castellanos, J. and de Freitas, N. (2007) Analysis of particle methods for simultaneous robot localization and mapping and a mew algorithm: Marginal-SLAM. In Proc. Int. Conf. Robotics and Automation.
[39] Olsson J., Douc R., Cappé O., Moulines É. (2008), Sequential Monte Carlo smoothing with application to parameter estimation in nonlinear state-space models. Bernoulli, vol. 14, pp. 155-179.
[40] Pitt, M. K. (2002) Smooth particle filters for likelihood evaluation and maximisation. Warwick Economic Research Papers, No. 651.
[41] Pitt, M. K. and Shephard, N. (1999). Filtering via simulation: auxiliary particle filter. J. Amer. Statist. Assoc., vol. 94, pp. 590-599.
[42] Polson, N.G., Stroud J.R. and Müller P. (2008). Pratical filtering with sequential parameter learning. J. R. Statist. Soc. B, vol. 70, pp. 413-428.
[43] Poyadjis G., Doucet A. and Singh S.S., (2005) Maximum likelihood parameter estimation using particle methods, In Proc. Joint Statistical Meeting.
[44] Poyadjis G., Doucet A. and Singh S.S., (2009) Se-

quential Monte Carlo for computing the score and observed information matrix in state-space models with applications to parameter estimation. Technical report CUED/F-INFENG/TR.628, Cambridge University.

[45] Rydén, T. (1997) On recursive estimation for hidden Markov models. Stoch. Proc. and Appl, vol. 66, pp. 79-96.

[46] Storvik G. (2002), Particle filters in state-space models with the presence of unknown static parameters, IEEE Trans. Signal Processing, vol. 50, pp. 281–289.

[47] Titterington D. M. (1984) Recursive parameter estimation using incomplete data. J. R. Statist. Soc. B, vol. 46, pp. 257-267.

[48] Wills, A., Schon, T. and Ninness, B. (2008) Parameter Estimation for Discrete-Time Nonlinear Systems Using EM. In Proc. 17th IFAC World Congress.