16th IFAC Symposium on System Identification
The International Federation of Automatic Control
Brussels, Belgium. July 11-13, 2012

On-line Parameter Estimation in General State-Space Models using a Pseudo-Likelihood Approach

Christophe Andrieu[1] **- Arnaud Doucet**[2] **- Vladislav B. Tadić**[1]

[1] *University of Bristol and* [2] *University of Oxford.*

Abstract State-space models are a very general class of time series capable of modeling dependent observations in a natural and interpretable way. While optimal state estimation can now be routinely performed using SMC (sequential Monte Carlo) methods, on-line static parameter estimation largely remains an unsolved problem. In Andrieu and Doucet [2003] it was proposed to use a pseudo-likelihood approach. This pseudo-likelihood can be optimised directly using a stochastic gradient algorithm, but we focus on an on-line Expectation-Maximization (EM). We present here novel simple recursions that allow us to estimate confidence intervals on-line and develope new theoretical results concerning the pseudo-likelihood estimate. More precisely we characterise the loss of efficiency compared to that of the maximum likelihood estimate, and also quantify the bias of the estimate in cases where the pseudo-likelihood needs to be approximated. We show in a tractable situation requiring no Monte Carlo simulation that these theoretical results accurately predict performance, pointing to their practical relevance.

1. INTRODUCTION

This paper is concerned with the on-line estimation of static parameters in non-linear non-Gaussian state-space models. More precisely, we consider models of the following form. For any parameter $\theta \in \Theta$, the hidden/latent state process $\{X_n; n \geq 1\} \subset \mathsf{X}^{\mathbb{N}}$ is a Markov process, characterized by its Markov transition probability density $f_\theta(x'|x)$, *i.e.* $X_1 \sim \nu_\theta$ and for $n \geq 1$,

$$X_{n+1}|(X_n = x) \sim f_\theta(\cdot|x) \ . \tag{1}$$

As indicated by its name $\{X_n\}$ is observed, not directly, but through another process $\{Y_n; n \geq 1\} \subset \mathsf{Y}^{\mathbb{N}}$. The observations are assumed to be conditionally independent given $\{X_n\}$, and their common marginal probability density is of the form $g_\theta(y|x)$; *i.e.* for $1 \leq n \leq m$,

$$Y_n|(X_1, \ldots, X_n = x, \ldots, X_m) \sim g_\theta(\cdot|x) \ . \tag{2}$$

We give here an example used throughout this paper.

Example 1. *SV model* Shephard and Pitt [1997]

$$X_{n+1} = \phi X_n + \sigma_v V_{n+1}, \ V_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$$

$$Y_n = \beta \exp(X_n/2) W_n, \ W_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1),$$

where $\Theta = (-1,1) \times \mathbb{R}^+ \times \mathbb{R}^+$ and $\theta = \left(\phi, \sigma_v^2, \beta^2\right)$ denotes the static parameter vector.

When the static parameter θ is known, sequential inference on the process $\{X_n\}$ is typically based on the sequence of joint posterior densities $\{p_\theta(x_{1:n}|Y_{1:n})\}$ which each summarizes all the information collected about $X_{1:n}$ up to time n. *Optimal filtering* is concerned with the sequential estimation of these densities, for which sequential Monte Carlo (SMC) methods [Doucet et al., 2001, Del Moral, 2004] have shown to be suitable. We focus in this paper on the on-line estimation of the static parameter θ. More precisely, assuming that there is a "true" parameter value θ^* generating the data $\{Y_n\}$ (*i.e.* θ^* is the "best" value of $\theta \in \Theta$ to explain the observations, in a sense made clear later on) and that this value is unknown, our aim is to compute point estimates of θ^* from $\{Y_n\}$ in an on-line manner. This problem appears in numerous applications. First, in most real-world scenarios θ^* is indeed unknown and needs to be estimated. Second, on-line estimation is often the only realistic solution when the amount of data to be processed is large. Although apparently simpler than the optimal filtering problem, the static parameter estimation problem has proved to be much more difficult; no closed form solutions are, in general, available, even for linear Gaussian and finite state-space hidden Markov models. There have already been numerous attempts to solve it in statistics and related fields. The paper is organized as follows. In Section 2, we introduce a family of pseudo-likelihood functions. We establish some novel theoretical properties of the maximum pseudo-likelihood estimator, including the loss of efficiency inherent to the use of a pseudo-likelihood function. The results are expressed in terms of some properties of the model considered and parameters of the pseudo-likelihood. In Section 3, we introduce a simple on-line EM algorithm in order to maximise the pseudo-likelihood function. We also develop novel and computationally efficient recursions that allow us to obtain on-line estimates of confidence intervals for our estimate. Finally, we demonstrate the performance of these methods via computer simulations on an example in Section 4 and show numerically that our theory accurately predicts what is observed in practice.

2. PSEUDO-ML METHODS

We present here an alternative strategy to the static parameter estimation problem, which aims to produce point estimates of θ^* rather than a series of estimates of the posterior densities $\{p(\theta|Y_{1:n})\}$. As a result no particle method is required in the parameter space, and it should also be pointed out that SMC methods in the state-space X are, in general, also not necessary. The most

978-3-902823-06-9/12/$20.00 © 2012 IFAC

10.3182/20120711-3-BE-2027.00422

natural approach to point estimation for the parameter θ^* consists of recursively maximizing the series of likelihoods $\{p(Y_{1:n}|\theta)\}$. We start this section with a discussion in which we highlight the difficulties associated with this type of strategy and this leads us to instead focus on a pseudo-likelihood approach which is, as we shall see, well suited to Monte Carlo approximations (Subsection 2.1). In Subsection 2.2 we study some theoretical aspects of such estimators, which we demonstrate to be of practical relevance on an example for which sophisticated numerical methods are not needed. We postpone the development of efficient algorithm to maximise the pseudo-likelihood to Section 3, where we describe a gradient algorithm in brief, and then focus mainly on on-line EM (Expectation-Maximization) type algorithms which has the advantage of numerical stability.

2.1 Likelihood and pseudo-likelihood functions

The log-likelihood function corresponding to model (1)-(2) is given, for $n \geq 1$ observations by

$$\log\ p_\theta\left(Y_{1:n}\right) = \sum_{k=1}^{n} \log\ p_\theta\left(Y_k \middle| Y_{1:k-1}\right)\ , \tag{3}$$

with the convention $Y_{1:0} = \varnothing$. Under ergodicity assumptions discussed in Subsection 2.2, it can be shown that the average log-likelihood is given by

$$\lim_{n\to\infty}\ n^{-1} \log\ p_\theta\left(Y_{1:n}\right) = l\left(\theta\right)\ ,$$

(see e.g. Andrieu and Doucet [2003]). It can be shown that the set of global maxima of this function includes the true value θ^*. Based upon this remark, one can suggest the use of stochastic gradient algorithms to maximize $l\left(\theta\right)$; see Andrieu et al. [2004] for a review. This strategy suffers from two limitations. First, it requires one to estimate the derivative of the optimal filter with respect to θ. Non-standard SMC methods are required to estimate this signed measure and their robust implementation has a computational complexity in $O\left(N^2\right)$, where N is the number of samples used for the SMC [Poyiadjis et al., 2011]. Second it can be difficult to properly scale the gradient components. More elegant and robust algorithms can be proposed that rely on on-line versions of the EM algorithm [Del Moral et al., 2010]. To circumvent these problems it has been proposed in Andrieu and Doucet [2003], Andrieu et al. [2004] to introduce another contrast function which is a pseudo-likelihood function akin to the approach suggested in Rydén [1997] for the particular case of finite state space HMMs, for which no numerical integration is required (the pseudo-likelihood used is refered to in that paper as "split-data likelihood") and for which no efficient on-line algorithm was suggested. It has been shown (see Section 3) that this pseudo-likelihood can be optimised either using a gradient algorithm, or in numerous situations using an efficient on-line EM algorithm. The pseudo-likelihood is defined as follows. Formally, consider for a given time lag $L \geq 1$ and any $k \geq 0$ "blocks" $\mathsf{X}_k = X_{kL+1:(k+1)L}$ and $\mathsf{Y}_k = Y_{kL+1:(k+1)L}$ of $\{X_n\}$ and $\{Y_n\}$. We will assume further on that for any $\theta \in \Theta$ the transition kernel $f_\theta(x'|x)$ admits an invariant density $\pi_\theta\left(x\right)$ and that the initial density of X_1 is given by $\nu_\theta\left(x\right) = \pi_\theta\left(x\right)$; this assumption is satisfied for Example 1 where $\pi_\theta\left(x\right) = \mathcal{N}\left(x; 0, \sigma^2\left(1-\phi^2\right)^{-1}\right)$. Because of this stationarity assumption, the vectors $\{\mathsf{X}_k, \mathsf{Y}_k\}$ are *identically distributed* and their common density is given by

$$p_\theta\left(\mathsf{x}_k, \mathsf{y}_k\right) = \pi_\theta(x_{kL+1}) g_\theta(y_{kL+1}|x_{kL+1}) \prod_{i=kL+2}^{(k+1)L} f_\theta(x_i|x_{i-1}) g_\theta(y_i|x_i)\ . \tag{4}$$

The likelihood of a block Y_k of observations is given by

$$p_\theta\left(\mathsf{Y}_k\right) = \int_{\mathsf{X}^L} p_\theta\left(\mathsf{x}_k, \mathsf{Y}_k\right) d\mathsf{x}_k\ , \tag{5}$$

and we define the log pseudo-likelihood for m blocks of observations by

$$l_L(\theta, \mathsf{Y}_{0:m-1}) := \frac{1}{L}\sum_{k=0}^{m-1} \log\ p_\theta\left(\mathsf{Y}_k\right)\ , \tag{6}$$

which, compared to the true likelihood, essentially ignores the dependence between data blocks. Note it would also be possible to consider overlapping blocks of the form $\left(Y_{1:L}, Y_{2:L+1}, ..., Y_{mL-L+1:mL}\right)$. The developments above parallel those of the classical scenario where the observations $\{\mathsf{Y}_k\}$ are independent and, as a result, the joint pseudo-likelihood is the product of its marginals. The parameter L should be large enough to ensure identifiability. Note also that there will be here an efficiency/computational complexity trade-off associated with L. As L increases, the maximum pseudo-likelihood estimate properties will become comparable to that of the standard ML estimate (see Subsection 2.2) but as we shall see this might result in more complex and computationally intensive algorithms. Under ergodicity assumptions discussed in Subsection 2.2, the average log pseudo–likelihood satisfies

$$\lim_{m\to\infty} \frac{1}{m} l_L(\theta, \mathsf{Y}_{0:m-1}) =: l_L\left(\theta\right)\ , \tag{7}$$

where

$$l_L\left(\theta\right) := \int_{\mathsf{Y}^L} \log\left(p_\theta\left(\mathsf{y}\right)\right) p_{\theta^*}(\mathsf{y}) d\mathsf{y}\ . \tag{8}$$

It can be shown that the set of parameters maximizing $l_L\left(\theta\right)$ includes the true parameter Rydén [1997]. This follows from the fact that maximizing $l_L\left(\theta\right)$ is equivalent to minimizing the following Kullback-Leibler divergence

$$K_L\left(\theta, \theta^*\right) = l_L\left(\theta^*\right) - l_L\left(\theta\right) \geq 0\ . \tag{9}$$

2.2 Theoretical properties of the pseudo-ML estimator

Before turning in Section 3 to practical procedures to efficiently optimise $l_L(\theta)$ online, we study here some of the asymptotic properties of the maximum pseudo-likelihood estimate obtained by maximising $l_L\left(\theta\right)$ as L increases. Some results for pseudo-likelihood approaches are already available in the literature but our model is significantly different. First we quantify the loss of efficiency introduced by the use of the pseudo-likelihood $l_L(\theta)$ in place of the "true" log-likelihood $l(\theta)$. More precisely, under assumptions implying that a central limit theorem (CLT) holds for both the sequence of maximum likelihood estimators $\{\hat{\theta}_n^*\}$ and the sequence of pseudo maximum likelihood estimators $\{\hat{\theta}_n^*(L)\}$, *i.e.* there exist covariance matrices Σ and Σ_L such that

$$\sqrt{n}(\hat{\theta}_n^* - \theta^*) \to_{\mathcal{D}} \mathcal{N}(0, \Sigma)\ ,$$
$$\sqrt{n_L}(\hat{\theta}_{n_L}^*(L) - \theta^*) \to_{\mathcal{D}} \mathcal{N}(0, \Sigma_L)\ ,$$

where $n_L \in \{kL, k \geq 1\}$ (expressions for Σ and Σ_L are given in our technical report) we seek to characterise the loss of efficiency of the pseudo-likelihood approach developed in this paper, *i.e.* we compare Σ and Σ_L in terms of the invariant density π_θ of the process $\{X_n\}$, the block length L and constants that characterise the ergodic properties of the process $\{X_k, Y_k\}$ and the associated filtering process (Theorem 1). While it is known that under regularity conditions $\Sigma_L - \Sigma$ is a positive definite matrix, our result provides us with a bound on the rate of convergence of Σ_L to Σ as L increases, which appears to accurately predict what can be observed in simple scenarios where a direct analysis is possible (see Subsection 2.2.2). We stress here on the fact that for brevity we do not prove the validity of the aforementioned CLTs, which however follow from our assumptions, but rather focus on an upper bound on $|\Sigma - \Sigma_L|$. Note also that considering the additional variance introduced by the Monte Carlo nature of the procedures developed in Section 3 is beyond the scope of the present paper. It would furthermore depend on the sampling procedure used within the on-line EM algorithm and require additional technical developments. Second we consider the practically important situation where the invariant density π_θ is not tractable (hence preventing the practical maximisation of $l_L(\theta)$) and replaced in the expression for $l_L(\theta)$ by an approximation μ_θ (see Subsection 3.3 for a discussion of this issue), leading to an approximate pseudo log-likelihood $l_L(\mu_\theta, \theta)$. Denoting $\theta^*(\mu_{\theta^*}, L)$ the assumed maximiser of $l_L(\mu_\theta, \theta)$, a natural question of practical relevance is that of the magnitude of the error $|\theta^*(\mu_{\theta^*}, L) - \theta^*|$ introduced by such an approximation in terms of the quality of the approximation μ_θ and the block length L. Theorem 2 provides an answer to this question, which seems to be of practical relevance as illustrated on an example which lends itself to direct analysis (see Subsection 2.2.2). The proofs of the theorems can be found in our companion paper. We illustrate the relevance of our theoretical results in a situation for which our assumptions are satisfied and where inference does not require (Monte Carlo based) numerical approximation and optimisation techniques developed later on in the paper.

Assumptions and results Note that for simplicity of exposition we will assume that all the probability distributions considered here have a density with respect to the Lebesgue or counting measure on the space concerned; this does not significantly affect the generality of our results.

(**A1**) Conditions on Θ and the likelihood:

(1) Θ is a compact set,
(2) θ^* is a unique strong global maximum of $l(\theta)$ and belongs to the interior of Θ, denoted by $\overset{\circ}{\Theta}$,
(3) $l(\theta)$ is twice continuously differentiable on $\overset{\circ}{\Theta}$ and $H(\theta^*) := -\nabla^2 l(\theta^*)$ is positive definite.

(**A2**) We assume that f_θ and g_θ are twice continuously differentiable and that there exist $\underline{f}_0, \underline{g}_0 > 0$ and $\bar{f}_0, \bar{g}_0, \bar{f}_1, \bar{g}_1, \bar{f}_2, \bar{g}_2 < +\infty$ such that for all $x, x' \in \mathsf{X}$,$y \in \mathsf{Y}$ and $\theta \in \Theta$

$$\underline{f}_0 \leq f_\theta(x' \,|\, x) \leq \bar{f}_0,\ \underline{g}_0 \leq g_\theta(y|x) \leq \bar{g}_0, \qquad (10)$$
$$|\nabla \log f_\theta(x' \,|\, x)| < \bar{f}_1,\ |\nabla \log g_\theta(y|x)| < \bar{g}_1\ ,$$
$$|\nabla^2 \log f_\theta(x' \,|\, x)| < \bar{f}_2 \text{ and } |\nabla^2 \log g_\theta(y|x)| < \bar{g}_2\ .$$

In addition $\nabla^2 \log f_\theta(x' \,|\, x)$ and $\nabla^2 \log g_\theta(y|x)$ are assumed continuous in θ, uniformly in $x, x', y \in \mathsf{X}^2 \times \mathsf{Y}$. We further assume that $X_1 \sim \pi_\theta$ where π_θ is the invariant density of f_θ.

The assumption $X_1 \sim \pi_\theta$ could be suppressed because of the previous ergodicity assumptions but simplifies the proofs. The conditions above will typically only hold in situations where X and Y are compact or finite spaces. These conditions can be weakened in order to consider unbounded spaces; see e.g. Kleptsyna and Veretennikov [2008]. However, this is at the expense of substantial additional technical complications that would distract us here from the essence of the proof: our primary aim here is to keep the level of technicality as low as possible while providing meaningful results that support both intuition and practice. In addition, we do not expect our final results to be significantly modified under such weaker conditions, except when geometric forgetting (see below) is not satisfied, in which case slower rates of convergence to zero (as a function of L) in Theorem 1 and Theorem 2 are to be expected. The loss of efficiency occurring when using $l_L(\theta)$ instead of $l(\theta)$ can be characterised by the following theorem.

Theorem 1. Assume (A1)-(A2). Then there exist $L_0 \geq 0$ and $C \in (0, +\infty)$ such that for any $L \geq L_0$

$$|\Sigma - \Sigma_L| \leq C \left| \left[\nabla^2 l\,(\theta^*)\right]^{-2} \right| \frac{\log(L)^2}{L\ \log(\rho)^2}\ .$$

Here ρ is an upper bound on the forgetting properties of $\{X_k\}$ conditional upon $\{Y_k\}$, that is the ability of the optimal filter to forget its initial condition. The loss of efficiency of the estimator compared to the maximum likelihood estimate vanishes as L increases (at a rate approximately inversely proportional) and depends on the mixing properties of the model. In particular as $\rho \to 0$, we recover the standard textbook independent case while when $\rho \to 1$ the loss of efficiency increases. Note also the dependence on $\nabla^2 l\,(\theta^*)$ which results in a tighter bound for "more informative" models. Again we denote $\theta^*(\mu_{\theta^*}, L)$ a generic maximum of the resulting approximate pseudo-likelihood $l_L(\mu_\theta, \theta)$ where π_θ is replaced by μ_θ. The following result quantifies the bias of this estimate.

Theorem 2. Assume (A1)-(A2). Then there exist $L_0 \geq 1$, $C \in (0, +\infty)$ and $\rho \in [0, 1)$ such that for any $L \geq L_0$ and $\mu_\theta \in \mathcal{P}(\mathsf{X})$ (differentiable with respect to θ)

$$|\theta^*(\mu_{\theta^*}, L) - \theta^*|$$
$$\leq \frac{C}{L(1-\rho)} \left| \left[\nabla^2 l\,(\theta^*)\right]^{-1} \right| (\|\mu_{\theta^*} - \pi_{\theta^*}\| + \|\nabla \mu_{\theta^*} - \nabla \pi_{\theta^*}\|)$$

where $\|\cdot\|$ is the total variation norm.

Again ρ characterises the forgetting properties of $\{X_k\}$ conditional upon $\{Y_k\}$. This result confirms the intuition that the bias introduced when using μ_θ instead of π_θ in the pseudo-likelihood vanishes with L large (and we show that the rate is in fact $1/L$) but also depends on how close μ_{θ^*} is to π_{θ^*} and the ergodicity properties of $\{X_k\}$ given $\{Y_k\}$.

Illustration of the relevance of the theory We illustrate our theoretical results through simple numerical simulations for which none of the Monte Carlo simulation techniques developed later on for more complex and realistic scenarios is required. The example is a finite state-space

hidden Markov model with two underlying states $\{1,2\}$ and four observations states $\{1,2,3,4\}$. Such models, and of course more elaborated versions, are used in bioinformatics where the four observed states correspond to the amino-acids A, G, C, T. We consider two scenarios for which $\pi = (2/3, 1/3)$ and the second largest eigenvalue λ of the transition matrix is either equal to 0.4 or 0.8. The emission matrix is taken to be

$$\begin{bmatrix} 0.1 & 0.3 & 0.4 & 0.2 \\ 0.3 & 0.2 & 0.4 & 0.1 \end{bmatrix} ,$$

and assumed known, *i.e.* the estimation focused on the transition matrix of the underlying chain. We compare the theoretical upper bound in theorem 1 for a manually adjusted constant C to ensure good fit (computing the precise constant C in Theorem 1 is beyond the scope of this paper) to a numerical evaluation of $|\Sigma - \Sigma_L|$ for $L = 2, 4, 8, 16, 32$ and 64, on simulated data. The results, presented in Fig. 1, suggest that our theory (the rate of convergence in L) is relevant given its ability to predict what is observed in practice. Note that it might at first sight seem surprising to find that the curve corresponding to $\lambda = 0.4$ is above that of $\lambda = 0.8$, both in light of our upper bound in Theorem 1 and the expected monotonicity of $\rho(\lambda)$. However one should bear in mind that $\nabla^2 l(\theta^*)$ is itself a function of λ. In fact in the present situation our numerical results show that $\left|\nabla^2 l(\theta^*, \lambda = 0.4)\right| \simeq 10 \times \left|\nabla^2 l(\theta^*, \lambda = 0.8)\right|$.

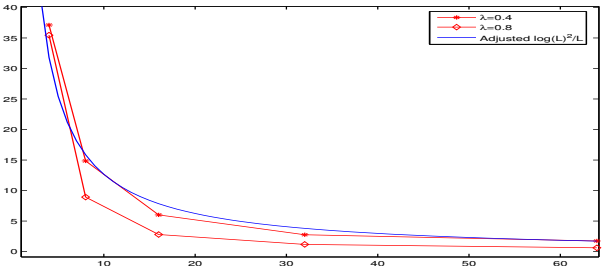

Figure 1. $|\Sigma - \Sigma_L|$ as a function of L for the discrete HMM example.

In order to illustrate our theoretical result in Theorem 2 we chose $\mu = (1/3, 2/3)$. In Fig 2 we present $|\theta(\mu, L) - \theta^*|$ obtained with the wrong initial distribution μ for $L = 4, 8, 16, 32$ and 64 (maximisation was performed using the first version of the on-line EM algorithm presented in Section 3, but any good method would do). In order to ease comparison with our theory, we have superimposed the function C/L suggested by Theorem 2 : it is clear that our theory is relevant and seems to even provide interesting bounds for even reasonably low values of L. Note finally that this result can be helpful in practice in order to select L *e.g.* if $\theta(\mu, L)$ is left virtually unchanged when L is changed to, say, $2L$, then we might have good reasons to think that $|\theta(\mu, L) - \theta^*|$ is small.

3. ON-LINE ALGORITHMS

In Andrieu and Doucet [2003] it has been proposed to maximise $l_L(\theta)$ recursively using on-line EM techniques. Whereas the maximization of the true average log-likelihood function requires complex SMC methods in order to either evaluate the filter derivative or estimate expectations with respect to distributions defined on X^n at time n, the key advantage (detailed further) of the average

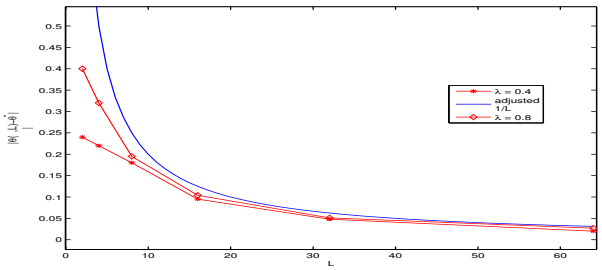

Figure 2. $|\theta(\mu, L) - \theta^*|$ as a function of L for the discrete HMM example.

log pseudo-likelihood function is that it only requires the estimation of expectations with respect to distributions defined on X^L. For the purpose of illustration we briefly discuss a direct steepest descent algorithm to minimize the Kullback-Leibler divergence $K_L(\theta, \theta^*)$ given in (9). Under regularity assumptions, the gradient with respect to θ of the cost function is given by

$$\nabla l_L(\theta) = \int_{\mathsf{Y}^L} \nabla \log\left(p_\theta(\mathsf{y})\right) \; p_{\theta^*}(\mathsf{y}) d\mathsf{y} \; . \tag{11}$$

An analytic expression for this gradient is rarely available, and we can resort to a stochastic approximation technique, *i.e.* replace (11) with an (possibly asymptotically) unbiased estimate of this gradient. This can be achieved by noting two key points. First, Fisher's identity yields for any $k \geq 0$,

$$\nabla \log p_\theta(\mathsf{Y}_k) = \mathbb{E}_\theta\left[\nabla \log p_\theta(\mathsf{X}_k, \mathsf{Y}_k) \middle| \mathsf{Y}_k\right] \; ,$$

where the expectation is with respect to $p_\theta(\mathsf{x}_k | \mathsf{Y}_k)$ defined in (4). When this expectation cannot be computed in closed-form we can resort to Monte Carlo methods (see developments later). Second, the observations $\{\mathsf{Y}_k\}$ are distributed according to $p_{\theta^*}(\mathsf{y})$ and can therefore be used as Monte Carlo samples to compute the integral in (11). We will not detail this approach any further here, but will focus on on-line EM type algorithms as they are more numerically stable and widely applicable to models used in practice.

3.1 On-line EM algorithm

To introduce the on-line EM, we first present an "ideal" batch EM algorithm to minimize $K_L(\theta, \theta^*)$ with respect to θ or equivalently to maximize $l_L(\theta)$ (more details are provided in Andrieu and Doucet [2003]). At iteration $k+1$, given an estimate θ_k of θ^*, we maximize for θ_{k+1}

$$Q(\theta, \theta_k) = \int_{\mathsf{X}^L \times \mathsf{Y}^L} \log\left(p_\theta(\mathsf{x}, \mathsf{y})\right) p_{\theta_k}(\mathsf{x}|\mathsf{y}) \, p_{\theta^*}(\mathsf{y}) d\mathsf{x} d\mathsf{y} \; . \tag{12}$$

Now for any $\theta \in \Theta$

$$Q(\theta_{k+1}, \theta_k) - Q(\theta_k, \theta_k) = K_L(\theta_k, \theta^*) - K_L(\theta_{k+1}, \theta^*) + \int_{\mathsf{X}^L \times \mathsf{Y}^L} \log\left(\frac{p_{\theta_{k+1}}(\mathsf{x}|\mathsf{y})}{p_{\theta_k}(\mathsf{x}|\mathsf{y})}\right) p_{\theta_k}(\mathsf{x}|\mathsf{y}) \, p_{\theta^*}(\mathsf{y}) d\mathsf{x} d\mathsf{y}$$

and since the second term on the rhs is negative by Jensen's inequality, we see that an iteration of this "ideal" EM algorithm decreases the value of $K_L(\theta, \theta^*)$, and the stationary points correspond to the zeros of $K_L(\theta, \theta^*)$. In practice for the models which we will consider, it is necessary to compute a set of sufficient statistics $\Phi(\theta_k, \theta^*)$ at time k in order to compute Q. Given $\Phi(\theta_{k-1}, \theta^*)$, it is possible to maximize $Q(\theta, \theta_{k-1})$ analytically when $L \geq 2$

and find $\theta_k = \Lambda(\Phi(\theta_{k-1}, \theta^*))$ where $\Lambda(\cdot)$ is problem dependent. We notice here that the minimal value L required will in many situations be given by the smallest L such that $\Lambda(\cdot)$ is defined unambiguously. In practice, $Q(\theta, \theta_{k-1})$ cannot be computed as the expectations appearing in the expression for $\Phi(\theta_k, \theta^*)$ are with respect to a measure dependent on the unknown parameter value θ^*. However, this ideal batch algorithm can be approximated using the following on-line scheme. Indeed, thanks to the ergodicity and stationarity assumptions, the observations $\{\mathsf{Y}_k\}$ provide us with samples from $p_{\theta^*}(\mathsf{y})$ which can be used for the purpose of Monte Carlo integration. More precisely we recursively approximate the sufficient statistics $\Phi(\theta_k, \theta^*)$ with the following update, given here at time k,

$$\hat{\Phi}_k = (1-\gamma_k)\,\hat{\Phi}_{k-1} + \gamma_k \mathbb{E}_{\theta_{k-1}}(\Psi(\mathsf{X}_k, \mathsf{Y}_k)|\,\mathsf{Y}_k)\ , \quad (13)$$

where $\mathbb{E}_{\theta_{k-1}}(\phi(\mathsf{X}_k)|\,\mathsf{Y}_k)$ denotes the expectation of ϕ with respect to $p_{\theta_{k-1}}(\mathsf{x}_k|\mathsf{Y}_k)$. We then substitute $\hat{\Phi}_k$ for $\Phi(\theta_k, \theta^*)$ and obtain $\theta_k = \Lambda(\hat{\Phi}_k)$. If θ_k was constant and $\gamma_k = k^{-1}$ then $\hat{\Phi}_k$ would simply compute the arithmetic average of $\{\mathbb{E}_{\theta_{k-1}}(\Psi(\mathsf{X}_k, \mathsf{Y}_k)|\,\mathsf{Y}_k)\}$, and converge towards $\Phi(\theta_k, \theta^*)$ by ergodicity. In fact, under mild appropriate conditions, convergence is in general ensured for any non-increasing positive stepsize sequence $\{\gamma_k\}$ such that $\sum \gamma_k = \infty$ and $\sum \gamma_k^2 < \infty$; we can select $\gamma_k = C.k^{-\alpha}$ where $C > 0$ and $\alpha \in \left(\frac{1}{2}, 1\right]$ thanks to the theory of stochastic approximation *e.g.* Benveniste et al. [1990]. To summarize, the vector of sufficient statistics $\hat{\Phi}_{-1}$ is arbitrarily initialized and the on-line EM algorithm proceeds as follows for the data block indexed by $k \geq 0$.

E-step $\hat{\Phi}_k = (1-\gamma_k)\hat{\Phi}_{k-1} + \gamma_k \mathbb{E}_{\theta_{k-1}}(\Psi(\mathsf{X}_k, \mathsf{Y}_k)|\,\mathsf{Y}_k)$.

M-step $\theta_k = \Lambda(\hat{\Phi}_k)$.

In scenarios where $\mathbb{E}_{\theta_k}(\Psi(\mathsf{X}_k, \mathsf{Y}_k)|\,\mathsf{Y}_k)$ does not admit an analytical expression, a further Monte Carlo approximation can be used. Assume that a good approximation $q_{\theta_{k-1}}(\mathsf{x}_k|\mathsf{Y}_k)$ of $p_{\theta_{k-1}}(\mathsf{x}_k|\mathsf{Y}_k)$ is available and that it is easy to sample from $q_{\theta_{k-1}}(\mathsf{x}_k|\mathsf{Y}_k)$. In this case the E-step of the algorithm presented above can be altered as follows.

E-step $\mathsf{X}_k^{(i)} \sim q_{\theta_{k-1}}(\cdot|\mathsf{Y}_k)$ *compute import. weights,* $W_k^{(i)}$.

$$\hat{\Phi}_k = (1-\gamma_k)\hat{\Phi}_{k-1} + \gamma_k \textstyle\sum_{i=1}^N W_k^{(i)} \Psi(\mathsf{X}_k^{(i)}, \mathsf{Y}_k)\ ,$$

As N increases the importance sampling approximation converges towards the true expectation. Note that as such the algorithm above leads to asymptotically biased estimates, but that this can be easily corrected by considering instead the following recursion for the estimation of the conditional expectation $\hat{F}_k = (1-\gamma_k)\hat{F}_{k-1} + \gamma_k \frac{1}{N}\sum_{i=1}^N W_k^{(i)} \Psi(\mathsf{X}_k^{(i)}, \mathsf{Y}_k)$ and $\hat{N}_k = (1-\gamma_k)\hat{N}_{k-1} + \gamma_k \frac{1}{N}\sum_{i=1}^N W_k^{(i)}$ and let $\hat{\Phi}_k = \hat{F}_k/\hat{N}_k$. As an alternative to importance sampling, we can use SMC techniques to approximate this expectation. We stress here on the fact that in this context, the path degeneracy issue is easily dealt with since L is fixed, and very often of small dimension. Observe also that it might be possible to sample exactly from $p_{\theta_{k-1}}(\mathsf{x}_k|\mathsf{Y}_k)$ using rejection sampling. In this case, it is not necessary to use a large number N of samples and a single sample is sufficient. Indeed it is only necessary to produce unbiased estimates of $\mathbb{E}_{\theta_{k-1}}(\Psi(\mathsf{X}_k, \mathsf{Y}_k)|\,\mathsf{Y}_k)$. In applications where the number of data is limited, the on-line EM may not have 'time' to converge. In such scenarios, it is possible to pass through the data repeatedly until convergence is observed. Assuming we have access to m blocks of L data, the resulting algorithm maximizes $l_L(\theta, \mathsf{Y}_{0:m-1})$ given by (6). We will demonstrate in Section 4 that this approach can be an attractive alternative to MCMC as it is computationally typically much cheaper.

3.2 On-line confidence intervals estimation

We present here novel and simple recursions that allow us to estimate the asymptotic covariance matrix of the estimate of θ^*, in an on-line manner. We can show that $\Sigma_L = H_L^{-1}(\theta^*)\, G_L(\theta^*)\, H_L^{-1}(\theta^*)$ where $H_L(\theta^*)$ and $G_L(\theta^*) - H_L(\theta^*)$ can be rewritten as

$$\frac{1}{L}\mathbb{E}\left[\mathbb{E}\left[\nabla \log p_{\theta^*}(\mathsf{X}_0, \mathsf{Y}_0)|\mathsf{Y}_0\right]\ \mathbb{E}\left[\nabla \log p_{\theta^*}(\mathsf{X}_0, \mathsf{Y}_0)|\mathsf{Y}_0\right]^{\mathsf{T}}\right]$$

and for any $n \geq 1$, with $E_n := \mathbb{E}\left[\nabla \log p_{\theta^*}(\mathsf{X}_n, \mathsf{Y}_n)|\mathsf{Y}_n\right]$

$$\frac{2}{L}\mathbb{E}\left[E_n \sum_{k=1}^{+\infty} \mathbb{E}\left[\nabla \log p_{\theta^*}(\mathsf{X}_{n-k}, \mathsf{Y}_{n-k})|\mathsf{Y}_{n-k}\right]\right]\ .$$

We first focus on the recursive estimation of the second term on the rhs in the expression for $G_L(\theta^*)$ above. Let Δ_k denote an estimator of $\mathbb{E}\left[\nabla \log p_{\theta^*}(\mathsf{X}_k, \mathsf{Y}_k)|\mathsf{Y}_k\right]$. The expression for the second part of the expression for $G_L(\theta^*)$ suggests the following recursions to estimate $G_L(\theta^*)$ on-line

$$\bar{G}_k = (1-\gamma_k)\bar{G}_{k-1} + \gamma_k \frac{1}{2}\left(\Delta_k \bar{\Delta}_{k-1}^{\mathsf{T}} + \bar{\Delta}_{k-1}\Delta_k^{\mathsf{T}}\right)$$
$$\bar{\Delta}_k = \Delta_k + \gamma \bar{\Delta}_{k-1}\ , \quad (14)$$

with $\bar{G}_0$ and $\bar{\Delta}_0$ arbitrarily initialised ($\bar{G}_0$ positive definite though) and $\gamma \in (0,1)$ (but close to 1). The sequence $\{\bar{G}_k\}$ is a sequence of asymptotically biased estimators of $G_L(\theta^*)$, but of finite variance. The bias vanishes as $\gamma \to 1$, but at the expense of increased variance. Estimating $H_L(\theta^*)$ is straightforward and follows from the recursion

$$\bar{H}_k = (1-\gamma_k)\bar{H}_{k-1} + \gamma_k \Delta_k \Delta_k^{\mathsf{T}}\ .$$

However in light of the expression for Σ_L and in order to reduce computational complexity it is preferable to directly estimate the inverse of $H_L(\theta^*)$. This can still be achieved in an on-line manner using the matrix inversion lemma, which yields the simple recursion

$$\bar{H}_k^{-1} = (1-\gamma_k)^{-1}\bar{H}_{k-1}^{-1} - \frac{\gamma_k}{(1-\gamma_k)^2}\frac{\bar{H}_{k-1}^{-1}\Delta_k\Delta_k^{\mathsf{T}}\bar{H}_{k-1}^{-1}}{1 + \frac{\gamma_k}{1-\gamma_k}\Delta_k^{\mathsf{T}}\bar{H}_{k-1}^{-1}\Delta_k}\ . \quad (15)$$

It is possible to simplify this expression further by considering first order terms in γ_k only, for $\gamma_k << 1$. The output of recursions (14) and (15) can be combined to produce estimates of Σ_L. In practice θ^* needs to be replaced with an estimator in order to estimate Δ_k: possible natural choices include θ_k, $\theta_{k+\tau}$ for some $\tau > 0$ for example.

3.3 Discussion

Although the procedure described sofar covers a large class of problems (in particular it covers the commonly used linear Gaussian models with non-linear observation equation Shephard and Pitt [1997]), in situations where π_θ is not known analytically the "exact" on-line EM algorithm described above cannot be exactly implemented. We outline here simple solutions, motivated by the result of Theorem 2 which suggests that the bias can be reduced by increasing

L or minimising the discrepancy between π_θ and μ_θ. A straightforward solution might consist of approximating π_θ with a parametric family $\{\mu_\beta\}$ for some parameter β. This distribution can be fitted to approximate samples from π_θ obtained by simulation of a long Markov chain $\{X_k\}$ with transition probability f_θ. A particularly interesting choice of $\{\mu_\beta\}$ relies on the identity $\int_{\mathsf{X}} \pi_\theta(x')f_\theta(x|\,x')dx' = \pi_\theta(x)$ which suggests the estimator, known as the look-ahead estimator, $\mu_\theta(x) = M^{-1}\sum_{k=1}^{M} f_\theta(x|\,X_k)$, again for $\{X_k\}$ sampled from the transition f_θ and some $M > 0$. This estimator is however random. An original solution we much prefer consists of using

$$\mu_\theta(x_1) := \int_{\mathsf{X}^M} \mu(x_{-M}) \prod_{k=-M+1}^{1} f_\theta(x_k|x_{k-1})$$

for some $\mu \in \mathcal{P}(\mathsf{X})$ and $M > 0$. This choice is attractive in two respects : (a) the parameter M allows one to modulate the bias introduced by the approximation (b) the approach is computationally particularly interesting since the resulting joint likelihood $p_\theta(\mathsf{x}_0, \mathsf{y}_0, \mu_\theta)$ is the marginal of $p_\theta(x_{-M:1}, \mathsf{x}_0, \mathsf{y}_0, \mu_\theta)$. As a result the EM algorithm or gradient algorithm described earlier can be used on this extended latent variable model, leading to the optimisation of $l_L(\mu_\theta, \theta)$, for μ_θ arbitrarily close to π_θ for an appropriate choice of M. In a similar vein we can suggest replacing π_θ with the filtering distribution calculated sofar. In this case the "asymptotic" criterion implicitly optimised is of the above and corresponds precisely to the maximum likelihood estimator for $L = 1$. Unfortunately in practice, the combined estimation and maximisation of the Q function (12) involved in the implementation of the EM algorithm might be too difficult. A pragmatic approach can consist of ignoring the expression for the filter in the calculations required.

4. APPLICATION TO STOCHASTIC VOLATILITY

We apply our algorithm to a stochastic volatility model in Example 1. The importance sampling density q_θ is chosen to be a Gaussian approximation of $\overline{p}_\theta$ as described in [Shephard and Pitt, 1997]. We first demonstrate the performance of our algorithm on a large simulated data set, $T = 2,500,000$ data points, with parameters $\phi = 0.8$, $\sigma_v^2 = 0.1$ and $\beta^2 = 1$. The algorithm was ran with $L = 10$, $N = 10$ and $\gamma_k = 0.01$ for $k \leq 5000$ and $\gamma_k = 1/(k - 5000)^{\cdot 75}$ for $k > 5000$ and the results presented in below. We used the Polyak-Ruppert averaging procedure in order to reduce the variance of our estimates and correspond to the smoothed (in time) estimates. Confidence intervals can be estimated on-line, using the recursions (14) and (15) in Subsection 3.2. We then turned to a real dataset, the returns for pound sterling/US dollar used in Shephard and Pitt [1997]. We ran the algorithm through the dataset 250 times with $L = 10$ and obtained the following values: $\hat{\phi} = 0.968 \pm 0.016$, $\hat{\sigma}^2 = 0.035 \pm 0.005$ and $\hat{\beta}^2 = 0.144 \pm 0.009$ (*i.e.* $\hat{\beta} = 0.38 \pm 0.3$) which are consistent, especially for ϕ and σ^2, with the posterior *means* obtained by Shephard and Pitt [1997] in a Bayesian setup, where the parameters were assigned priors.

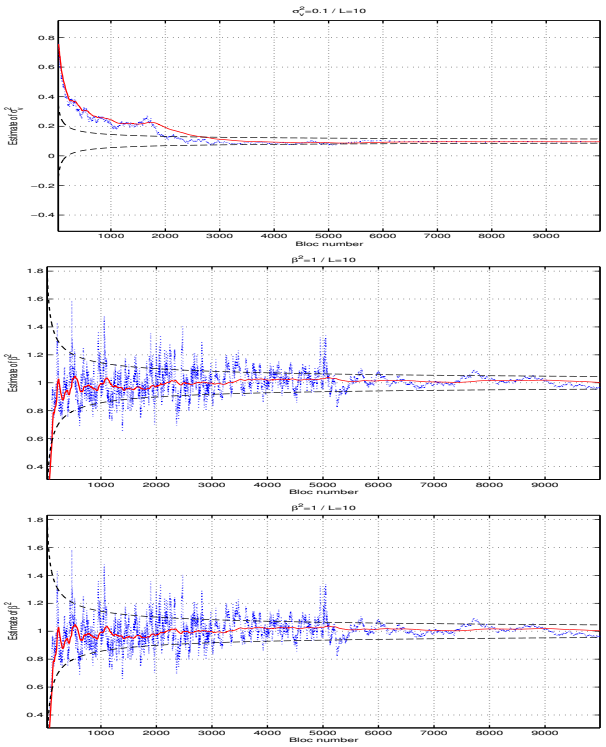

Figure 3. Blue dashed line: convergence of the estimate of ϕ, β^2, σ^2 as a function of the number of datapoints. Red solid line: estimate obtained after applying the Polyak-Ruppert averaging procedure.

REFERENCES

C. Andrieu and A. Doucet. Online expectation-maximization type algorithms for parameter estimation in general state space models. In *Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03). 2003 IEEE International Conference on*, volume 6, pages VI–69. IEEE, 2003.

C. Andrieu, A. Doucet, S.S. Singh, and V.B. Tadic. Particle methods for change detection, system identification, and control. *Proceedings of the IEEE*, 92(3):423–438, 2004.

A. Benveniste, P. Priouret, and M. Métivier. *Adaptive algorithms and stochastic approximations*. Springer-Verlag New York, Inc., 1990.

P. Del Moral. *Feynman-Kac formulae: genealogical and interacting particle systems with applications*. Springer Verlag, 2004.

P. Del Moral, A. Doucet, and S. Singh. Forward smoothing using sequential monte carlo. *Arxiv preprint arXiv:1012.5390*, 2010.

A. Doucet, N. De Freitas, and N. Gordon. *Sequential Monte Carlo methods in practice*. Springer Verlag, 2001.

ML Kleptsyna and A.Y. Veretennikov. On discrete time ergodic filters with wrong initial data. *Probability Theory and Related Fields*, 141(3):411–444, 2008.

G. Poyiadjis, A. Doucet, and S.S. Singh. Particle approximations of the score and observed information matrix in state space models with application to parameter estimation. *Biometrika*, 98(1):65, 2011.

T. Rydén. On recursive estimation for hidden markov models. *Stochastic Processes and Their Applications*, 66(1):79–96, 1997.

N. Shephard and M.K. Pitt. Likelihood analysis of non-gaussian measurement time series. *Biometrika*, 84(3): 653–667, 1997.