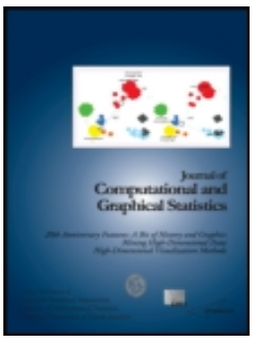

Taylor & Francis
Taylor & Francis Group

Journal of Computational and Graphical Statistics

ISSN: 1061-8600 (Print) 1537-2715 (Online) Journal homepage: https://www.tandfonline.com/loi/ucgs20

# Parameter Estimation in Hidden Markov Models With Intractable Likelihoods Using Sequential Monte Carlo

Sinan Yıldırım, Sumeetpal S. Singh, Thomas Dean & Ajay Jasra

**To cite this article:** Sinan Yıldırım, Sumeetpal S. Singh, Thomas Dean & Ajay Jasra (2015) Parameter Estimation in Hidden Markov Models With Intractable Likelihoods Using Sequential Monte Carlo, Journal of Computational and Graphical Statistics, 24:3, 846-865, DOI: 10.1080/10618600.2014.938811

**To link to this article:** https://doi.org/10.1080/10618600.2014.938811



Published online: 16 Sep 2015.

# Parameter Estimation in Hidden Markov Models With Intractable Likelihoods Using Sequential Monte Carlo

Sinan YILDIRIM, Sumeetpal S. SINGH, Thomas DEAN, and Ajay JASRA

We propose sequential Monte Carlo-based algorithms for maximum likelihood estimation of the static parameters in hidden Markov models with an intractable likelihood using ideas from approximate Bayesian computation. The static parameter estimation algorithms are gradient-based and cover both offline and online estimation. We demonstrate their performance by estimating the parameters of three intractable models, namely the $\alpha$-stable distribution, *g-and-k* distribution, and the stochastic volatility model with $\alpha$-stable returns, using both real and synthetic data.

**Key Words:** Approximate Bayesian computation; Maximum likelihood estimation.

## 1. INTRODUCTION

The hidden Markov model (HMM) is an important statistical model used in many fields including bioinformatics (Durbin et al. 1998), econometrics (Kim et al. 1998), and population genetics (Felsenstein and Churchill 1996); see Cappé, Moulines, and Rydén (2005) for a recent overview. An HMM is comprised of a latent process $\{X_t\}_{t\geq 1}$ and an observed process $\{Y_t\}_{t\geq 1}$. The latent process is a Markov chain with an initial density $\eta_\theta$ and the transition density $f_\theta$, that is,

$$X_t \in \mathcal{X} \subseteq \mathbb{R}^{d_x}, \quad X_1 \sim \eta_\theta(\cdot), \quad X_t|(X_{1:t-1} = x_{1:t-1}) \sim f_\theta(\cdot|x_{t-1}). \quad t \geq 2. \tag{1}$$

It is assumed that $\eta_\theta(x)$ and $f_\theta(x|x')$ are densities on $\mathcal{X}$ with respect to a dominating measure denoted generically as $dx$. The observation at time $t$ is conditionally independent of all other random variables given $X_t = x_t$ and its conditional observation density is $g_\theta(\cdot|x_t)$ on $\mathcal{Y}$ with respect to the dominating measure $dy$, that is,

$$Y_t \in \mathcal{Y} \subseteq \mathbb{R}^{d_y}, \quad Y_t|\{x_i\}_{i\geq 1}, \{y_i\}_{i\geq 1, i\neq t} \sim g_\theta(\cdot|x_t), \quad t \geq 1. \tag{2}$$

Sinan Yıldırım, School of Mathematics, University of Bristol, Bristol BS8 1TH, UK (E-mail: *s.yildirim@bristol.ac.uk*). Sumeetpal S. Singh, Department of Engineering, University of Cambridge, Cambridge CB2 1PZ, UK (E-mail: *sss40@cam.ac.uk*). Thomas Dean is research scientist, Darktrace, Cambridge CB3 0FA, UK (E-mail: *thomas.dean@cantab.net*). Ajay Jasra, Department of Statistics and Applied Probability, National University of Singapore, Singapore 119077 (E-mail: *staja@nus.edu.sg*).



*Journal of Computational and Graphical Statistics, Volume 24, Number 3, Pages 846–865*
*DOI: 10.1080/10618600.2014.938811*
Color versions of one or more of the figures in the article can be found online at *www.tandfonline.com/r/jcgs*.

The law of the HMM is parameterized by a vector $\theta$ taking values in some compact subset $\Theta$ of the Euclidean space $\mathbb{R}^{d_\theta}$.

In this article we focus on HMMs where the probability density $g_\theta(y|x)$ of the observations is *intractable*. By intractable we mean that $g_\theta(y|x)$ cannot be evaluated (or it is computationally prohibitive to calculate). However, we are able to generate samples from $g_\theta(\cdot|x)$ despite its intractability.

We will denote the actual observed random variables of the HMM as $\hat{y}_1, \hat{y}_2, \ldots$ and assume that they are generated by some unknown $\theta^* \in \Theta$ which is to be estimated. The maximum likelihood estimate of $\theta^*$ given $\hat{y}_{1:n}$ is

$$\theta_{\mathrm{ML}} = \arg\max_{\theta\in\Theta} p_\theta(\hat{y}_{1:n}),$$

where $p_\theta(\hat{y}_{1:n})$ is the probability density, or the *likelihood*, of the observations $\hat{y}_{1:n}$, and from (1)–(2), is given by

$$p_\theta(\hat{y}_{1:n}) = \int_{\mathcal{X}^n} \eta_\theta(x_1) g_\theta(\hat{y}_1|x_1) \left[\prod_{t=2}^{n} f_\theta(x_t|x_{t-1}) g_\theta(\hat{y}_t|x_t)\right] dx_{1:n}. \tag{3}$$

Even when $\mathcal{X}$ is a finite set, $p_\theta(\hat{y}_{1:n})$ cannot be evaluated because $g_\theta(y|x)$ is intractable. There is a sizeable literature on the use of sequential Monte Carlo (SMC) methods, also known as particle filters, to evaluate the gradient of $p_\theta(\hat{y}_{1:n})$ with respect to $\theta$, which is subsequently used to compute its maximizer; see for example, the review in Kantas et al. (2009). However, these methods require a *tractable* $g_\theta(y|x)$ and they are not directly applicable when this density is intractable. We thus propose new SMC-based maximum likelihood estimation (MLE) algorithms to fill this void. We handle the intractable $g_\theta(y|x)$ by drawing on ideas from approximate Bayesian computation (ABC), an inference technique initially developed for Bayesian models with an intractable likelihood; see Marin et al. (2012) for a recent review. Our static parameter estimation algorithms are gradient based and cover both offline (or batch) and online estimation.

Recently Ehrlich, Jasra, and Kantas (2013) proposed a gradient-based MLE algorithm for HMMs with an intractable observation density $g_\theta(y|x)$. The authors estimate the gradient, with respect to $\theta$, of the following approximation of the likelihood $p_\theta(\hat{y}_{1:n})$ in (3)

$$\mathbb{E}_\theta\left\{\prod_{t=1}^{n} \mathbb{I}_{B^\epsilon_{\hat{y}_t}}(Y_t)\right\}, \tag{4}$$

where $B^\epsilon_{\hat{y}_i}$ denotes the ball of radius $\epsilon$ centered at $\hat{y}_i$. (See Section 2 for more details on the approximate likelihood (4).) Ehrlich, Jasra, and Kantas (2013) estimated the gradient of (4) using a finite difference approximation where (4) itself, for various values of $\theta$, is calculated using SMC. The major advantage of our method over that of Ehrlich, Jasra, and Kantas (2013) is that we characterize the gradient of (4) directly, by using available information on how the intractable $g_\theta(y|x)$ is simulated from, and subsequently approximate it using SMC, thus avoiding the added error of a finite difference approximation. Our online MLE algorithm is asymptotically unbiased (as our numerical results indicate) as the number of particles increases whereas the same cannot be said for Ehrlich, Jasra, and Kantas (2013) due to the finite difference approximation; their numerical results indicate a significant bias that does not diminish with increasing data, even when $p_\theta(\hat{y}_{1:n})$ can be calculated exactly

as they illustrate for a linear Gaussian state-space model (see Ehrlich, Jasra, and Kantas 2013, Fig. 2). Also, as observed from the results in Ehrlich, Jasra, and Kantas (2013), the variance of the parameter estimates of their recursive MLE algorithm does not diminish with more data while ours does (see the discussion in Section 3.1).

Static parameter estimation for HMMs with intractable state and observation densities have been addressed in a Bayesian context by Campillo and Rossi (2009). Campillo and Rossi (2009) used the so-called convolution particle filter, which uses ideas from kernel density estimation to replace the intractable densities needed for the weight evaluation in the particle filter with their kernel estimates, to sequentially estimate the posterior distribution of $\theta^*$. While an SMC-based Bayesian approach can potentially produce good estimates of $\theta^*$ for short data lengths, at least for tractable models where standard particle methods apply, particle degeneracy does bias the estimation results for long datasets (Andrieu, Doucet, and Tadić 2005; Kantas et al. 2009). In contrast, our methods do give rise to practically consistent estimators as our numerical results indicate.

Finally, we remark that MLE using ABC is studied in the recent work (Rubio and Johansen 2013), but in a non-HMM setting where the likelihood of data $\hat{y}$ given $\theta$ is intractable. The authors form a kernel density estimate of the likelihood from $\theta$ samples drawn from the ABC posterior distribution. They propose maximizing the kernel density estimate as an approximation to MLE. Unlike Rubio and Johansen (2013), we consider the HMM setting and our methods do not need samples of $\theta$.

The remainder of this article is organized as follows. The theory that underpins our MLE methodology is detailed in Section 2, and in Section 3 we describe its SMC implementation. Numerical examples using both simulated and real datasets are given in Section 4. The numerical work covers three intractable models, namely the $\alpha$-stable distribution, *g-and-k* distribution, and the stochastic volatility model with $\alpha$-stable returns. Finally, Section 5 provides a discussion of other possible methods for parameter estimation in HMMs when both state and observation densities are intractable.

## 2. THE ABC MLE APPROACH FOR PARAMETER ESTIMATION

The particle filter sequentially approximates the sequence of posterior densities $\{p_\theta(x_{1:t}|Y_{1:t} = \hat{y}_{1:t})\}_{t\geq 1}$ of the HMM $\{X_t, Y_t\}_{t\geq 1}$ using a weighted discrete distribution with $N$ support points for $X_{1:t}$ which are called particles. At each time $t$, the particles are resampled according to their current weights, and then the resampled particles are propagated independently of each other using a proposal transition density $r_\theta(x_{t+1}|x_t)$. The particles are then reweighed to correct for the discrepancy between $p_\theta(x_{1:t+1}|Y_{1:t+1} = \hat{y}_{1:t+1})$ and the law of the proposed particles which is $p_\theta(x_{1:t}|Y_{1:t} = \hat{y}_{1:t})r_\theta(x_{t+1}|x_t)$. This is standard importance sampling and the assumption in the weight correction step is that the law of each resampled particle at time $t$ is $p_\theta(x_{1:t}|Y_{1:t} = \hat{y}_{1:t})$, which is an erroneous but progressively correct as $N$ is increased (Chopin 2002; Crisan and Doucet 2002; Del Moral 2004). In the implementation of the particle filter the normalizing constants of the sequence of target posteriors are not needed but calculating the new weights requires $g_\theta(\hat{y}|x)$ to be tractable. Del Moral (2004) showed that the weights of the particle approximation of $\{p_\theta(x_{1:t}|Y_{1:t} = \hat{y}_{1:t})\}_{t\geq 1}$ can be used to obtain an unbiased estimate of the likelihoods $\{p(Y_{1:t} = \hat{y}_{1:t})\}_{t\geq 1}$. See the Appendix for an example code for a particle filter.

Jasra et al. (2012) considered the problem of constructing an SMC approximation of the *filter* $p_\theta(x_t|Y_{1:t} = \hat{y}_{1:t})$, which is the marginal of the particle approximation for $p_\theta(x_{1:t}|Y_{1:t} = \hat{y}_{1:t})$, for an HMM with an intractable observation density $g_\theta(y|x)$. Since it is not possible to calculate the weights of the particle filter for such an HMM where $g_\theta(y|x)$ is intractable, they proposed a particle filter approximation for the extended HMM $\{(X_t, Y_t), Y_t^\epsilon\}_{t\geq 1}$ where the joint process $\{X_t, Y_t\}_{t\geq 1}$, which is now the latent process of the extended HMM, is defined by (1) and (2) and the new sequence $\{Y_t^\epsilon\}_{t\geq 1}$ is

$$Y_t^\epsilon = Y_t + \epsilon V_t, \quad V_t \sim^{\text{iid}} \text{Unif}\big(B_0^1\big), \quad t \geq 1, \tag{5}$$

where $B_y^r$ denotes the ball of radius $r > 0$ centered at $y \in \mathbb{R}^{d_y}$ and $\text{Unif}(B)$ is the uniform distribution over the set $B$. Then, the density

$$p_{\theta^*}\big(x_t|Y_{1:t}^\epsilon = \hat{y}_{1:t}\big)$$

of the extended HMM is regarded as an approximation for $p_{\theta^*}(x_t|Y_{1:t} = \hat{y}_{1:t})$ where $\epsilon > 0$ reflects the error of the approximation and this error diminishes as $\epsilon \to 0$; see also Calvet and Czellar (2012); Martin et al. (2014) for theoretical results on this approximation. Note that $p_{\theta^*}(x_t|Y_{1:t}^\epsilon = \hat{y}_{1:t})$ does not coincide with $p_{\theta^*}(x_t|Y_{1:t} = \hat{y}_{1:t})$ because $\hat{y}_{1:t}$ obeys the law (1)–(2) and not (5). Jasra et al. (2012) remarked that $p_{\theta^*}(x_t|Y_{1:t}^\epsilon = \hat{y}_{1:t})$ is the ABC approximation for the filter of an HMM. Furthermore, they showed it is straightforward to approximate $p_{\theta^*}(x_t|Y_{1:t}^\epsilon = \hat{y}_{1:t})$ with a bootstrap particle filter.

Consider now the extended HMM $\{(X_t, Y_t), Y_t^\epsilon\}_{t\geq 1}$ specified by (1), (2), and (5) and let $p_\theta(Y_{1:n}^\epsilon = y_{1:n})$ denote the probability density (or likelihood function) of the process $\{Y_t^\epsilon\}_{t\geq 1}$ evaluated at some $y_{1:n} \in (\mathbb{R}^{d_y})^n$. (See (12) for the precise expression of this density.) Dean et al. (2014) studied the theoretical properties of the following maximum likelihood estimate of $\theta^*$:

$$\theta_n^\epsilon = \arg\max_{\theta\in\Theta} p_\theta\big(Y_{1:n}^\epsilon = \hat{y}_{1:n}\big). \tag{6}$$

(We remark that (4) is $p_\theta(Y_{1:n}^\epsilon = \hat{y}_{1:n})$ when the Lebesgue volumes of the balls $B_{\hat{y}_1}^\epsilon, \ldots, B_{\hat{y}_n}^\epsilon$ are omitted from the latter.) Dean et al. (2014) called the procedure (6) ABC MLE. (The use of the acronym ABC is to emphasis that it is the same approximate likelihood which is being maximized here.) The bootstrap particle filter of Jasra et al. (2012) provides an unbiased SMC approximation of the likelihood $p_\theta(Y_{1:n}^\epsilon = \hat{y}_{1:n})$ and this likelihood may be maximized by evaluating the approximation over a grid of values for $\theta$. This, however, is clearly not practical as the dimension of $\theta$ increases, has no straightforward extension for recursive estimation and is not an accurate convergent method.

Dean et al. (2014) showed that the ABC MLE (6) leads to a biased estimate of the parameter vector $\theta^*$ in the sense that as $n \to \infty$, $\theta_n^\epsilon$ will converge to some point $\theta^{*,\epsilon} \neq \theta^* \in \Theta$ and that this bias can be made arbitrarily small, that is, $\theta^{*,\epsilon} \to \theta^*$ as $\epsilon \to 0$. Dean et al. (2014) showed that the bias is $\mathcal{O}(\epsilon)$; Dean and Singh (2011) refined this to $\mathcal{O}(\epsilon^2)$. The bias of ABC MLE is due to the fact that the observed sequence $\hat{y}_1, \hat{y}_2, \ldots$ is the outcome of the law (2) for $\theta = \theta^*$ and not (5). Dean et al. (2014) suggested removing the bias of $\theta_n^\epsilon$ in (6) by adding noise to the real data and then computing the maximum likelihood estimate, that is, let $v_1, \ldots, v_n$ be a realization of iid samples from $\text{Unif}(B_0^1)$ and let

$$y_t^\epsilon = \hat{y}_t + \epsilon v_t, \quad 1 \leq t \leq n. \tag{7}$$

Note that the noisy data $y^{\epsilon}_{1:n}$ now obey the law of $\{Y^{\epsilon}_t\}_{t\geq 1}$ when $\theta = \theta^*$. Therefore, the procedure

$$\theta^{\epsilon}_n = \arg\max_{\theta\in\Theta} p_\theta\left(Y^{\epsilon}_{1:n} = y^{\epsilon}_{1:n}\right), \tag{8}$$

which will be called noisy ABC MLE from now on, can now produce a consistent estimator of the parameter vector $\theta^*$ as $n \to \infty$. This result proved by Dean et al. (2014) can be interpreted as the frequentist equivalence of Wilkinson's observation that the ABC posterior distribution is exact under the assumption of model error (Wilkinson 2013).

Finally, Dean et al. (2014) also remarked that the use of other types of noise in (5) is possible without compromising the asymptotics of noisy ABC MLE, that is,

$$Y^{\epsilon}_t = Y_t + \epsilon V_t, \quad V_t \overset{\text{iid}}{\sim} \kappa, \quad t \geq 1, \tag{9}$$

where $\kappa$ is a smooth centred density. (Accordingly, noisy ABC MLE in (8) is performed with the noise corrupted observations (7) where now $v_i$ are realizations of iid samples from $\kappa$.) As we show, a continuously differentiable $\kappa$ is important for the development of practical gradient-based MLE techniques. In this work we choose $\kappa$ to be the probability density of zero-mean unit-variance Gaussian random variable. Other choices are possible (but not investigated) and our framework would still be applicable.

We remark that although the theoretical basis for ABC MLE was established in Dean et al. (2014), the authors do not propose a practical methodology for implementing ABC MLE in their work; this is indeed an important void to be filled. In this article we demonstrate how, by using ideas from Poyiadjis, Doucet, and Singh (2011), both batch and online versions of noisy ABC MLE can be implemented with SMC.

## 3. IMPLEMENTING ABC MLE WITH SMC

We assume that for all $(x, \theta) \in \mathcal{X} \times \Theta$ there exist a distribution on some auxiliary space $\mathcal{U}$ with a tractable density $\nu_\theta(\cdot|x)$ with respect to $du$ and a function $\tau_\theta : \mathcal{U} \times \mathcal{X} \to \mathcal{Y}$ such that one can sample from $g_\theta(\cdot|x)$ by first sampling $U \in \mathcal{U}$ from $\nu_\theta(\cdot|x)$ and then applying the transformation $U \to \tau_\theta(U, x)$; that is, the law of $\tau_\theta(U, x)$ is $g_\theta(\cdot|x)$. From this it follows that the process $\{Y^{\epsilon}_t\}_{t\geq 1}$ in (9) can be equivalently generated as

$$Y^{\epsilon}_t = \tau_\theta(U_t, X_t) + \epsilon V_t, \quad V_t \overset{\text{iid}}{\sim} \kappa, \quad t \geq 1, \tag{10}$$

where $\{X_t\}_{t\geq 1}$ is the hidden state of the original HMM given by (1) and $U_t \sim \nu_\theta(\cdot|X_t)$ for all $t$. We will implement SMC- based MLE for the following HMM: Let $\{Z_t := (X_t, U_t)\}_{t\geq 1}$ be the latent process and $\left\{Y^{\epsilon}_t\right\}_{t\geq 1}$ in (10) be the observation process. The initial and transition densities for $\{Z_t\}_{t\geq 1}$ (with respect to the dominating measure $dz = dxdu$) and the observation density of $\{Y^{\epsilon}_t\}_{t\geq 1}$ (with respect to the Lebesgue measure on $\mathbb{R}^{d_y}$) are

$$\pi_\theta(z) = \eta_\theta(x)\nu_\theta(u|x), \quad q_\theta(z'|z) = f_\theta(x'|x)\nu_\theta(u'|x'), \quad h^{\epsilon}_\theta(y|z) = \frac{1}{\epsilon}\kappa\left(\frac{y - \tau_\theta(z)}{\epsilon}\right), \tag{11}$$

where $z = (x, u)$ and $z' = (x', u')$. The density of the observed process $Y^{\epsilon}_{1:n}$ of this HMM evaluated at some $y_{1:n}$ is

$$p_\theta(y_{1:n}) := \int_{\mathcal{Z}^n} \pi_\theta(z_1) h^{\epsilon}_\theta(y_1|z_1) \left[\prod_{t=2}^{n} q_\theta(z_t|z_{t-1}) h^{\epsilon}_\theta(y_t|z_t)\right] dz_{1:n}, \tag{12}$$

where $\mathcal{Z} = \mathcal{X} \times \mathcal{U}$. Note that $p_\theta(\cdot)$ in (12) is indeed the likelihood function $p_\theta(Y^{\epsilon}_{1:n} = \cdot)$ to be maximized with respect to $\theta$ in ABC MLE in Section 2; see (6) and (8). Moreover, all the densities declared in (11) are tractable and differentiable functions of $\theta$ (provided that $f_\theta$, $\nu_\theta$, and $\tau_\theta$ are differentiable with respect to $\theta$).

Henceforth, we will work exclusively with the HMM $\{Z_t, Y^{\epsilon}_t\}_{t\geq 1}$ defined in (11). As discussed before, we corrupt the real measurements $\hat{y}_1, \hat{y}_2, \ldots$ with a single realization of independent samples $v_1, v_2 \ldots$ from a $\theta$-independent probability density $\kappa$, that is,

$$y^{\epsilon}_i = \hat{y}_i + \epsilon v_i,$$

to obtain a realization of the observed process of the HMM $\{Z_t, Y^{\epsilon}_t\}_{t\geq 1}$.

### 3.1 GRADIENT ASCENT

One well-known MLE algorithm is the following iterative gradient ascent method which updates the parameter estimate $\theta_j$ using the rule

$$\theta_j = \theta_{j-1} + \gamma_j \nabla \log p_{\theta_{j-1}}(y^{\epsilon}_{1:n}), \tag{13}$$

where $\theta_0 \in \Theta$ is an arbitrary initial estimate. Here $\{\gamma_j\}_{j\geq 1}$ is a sequence of step-sizes satisfying the constraints $\sum_{j\geq 1} \gamma_j = \infty$ and $\sum_{j\geq 1} \gamma_j^2 \leq \infty$ so as to ensure that the algorithm converges to a local maximum of $\log p_\theta(y^{\epsilon}_{1:n})$. The term $\nabla \log p_\theta(y^{\epsilon}_{1:n})$ is shorthand for the $\mathbb{R}^{d_\theta}$-valued vector

$$\nabla \log p_\theta\left(y^{\epsilon}_{1:n}\right) := \frac{\partial \log p_\theta\left(y^{\epsilon}_{1:n}\right)}{\partial \theta},$$

which is also called the score vector, and is given by Fisher's identity (see Cappé, Moulines, and Rydén 2005)

$$\nabla \log p_\theta\left(y^{\epsilon}_{1:n}\right) = \int_{\mathcal{Z}^n} \left[\sum_{t=1}^{n} \nabla \log q_\theta(z_t|z_{t-1}) + \nabla \log h^{\epsilon}_\theta\left(y^{\epsilon}_t|z_t\right)\right] p_\theta\left(z_{1:n}|y^{\epsilon}_{1:n}\right) dz_{1:n} \tag{14}$$

with the convention that $q_\theta(z_1|z_0) = \pi_\theta(z_1) = \eta_\theta(x_1)\nu_\theta(u_1|x_1)$. Note that the method in (13) uses the whole dataset $y^{\epsilon}_{1:n}$ at every parameter update step, which makes it a batch method. An alternative to it is the following online gradient ascent method which updates the parameter estimate every time a new data point is received

$$\theta_n = \theta_{n-1} + \gamma_n \nabla \log p_{\theta_{n-1}}\left(y^{\epsilon}_n|y^{\epsilon}_{1:n-1}\right), \tag{15}$$

where

$$\nabla \log p_{\theta_{n-1}}\left(y^{\epsilon}_n|y^{\epsilon}_{1:n-1}\right) = \nabla \log p_{\theta_{n-1}}\left(y^{\epsilon}_{1:n}\right) - \nabla \log p_{\theta_{n-1}}\left(y^{\epsilon}_{1:n-1}\right). \tag{16}$$

While the subscript $\theta_{n-1}$ indicates that $\nabla \log p_\theta(y^{\epsilon}_n|y^{\epsilon}_{1:n-1})$ is evaluated at $\theta = \theta_{n-1}$, a necessary requirement for a truly online implementation is that the previous values of

$\theta$ estimates (i.e., other than $\theta_{n-1}$) are also used in the evaluation of $\nabla \log p_{\theta_{n-1}}(y_n^\epsilon | y_{1:n-1}^\epsilon)$ (Le Gland and Mevel 1997).

It is important to note that, for both the batch method (13) and the online method (15), we require that the transition density of $\{Z_t\}_{t\geq 1}$ be tractable and differentiable with respect to $\theta$, which is precisely why we propose to work with $\{Z_t, Y_t^\epsilon\}_{t\geq 1}$ rather than $\{(X_t, Y_t), Y_t^\epsilon\}_{t\geq 1}$ whose state transition density contains the intractable $g_\theta$. (We discuss suitable alternatives when the state transition density is intractable in Section 3.3.)

It is apparent from (13) and (15) that an SMC implementation of these MLE algorithms hinges on the availability of a particle approximation of the score in (14). Poyiadjis, Doucet, and Singh (2011) discussed two methods to estimate the score using the SMC approximation of the full posterior $p_\theta(z_{1:n} | y_{1:n}^\epsilon)$. One method is nothing more than the substitution of the law $p_\theta(z_{1:n} | y_{1:n}^\epsilon)$ in (14) with its particle approximation and has a cost, like the particle filter itself, which is $\mathcal{O}(N)$. We will refer to this estimate of the gradient as the $\mathcal{O}(N)$ method (Poyiadjis, Doucet, and Singh 2011, Algorithm 1). Due to resampling step of the particle filter there is a lack of unique samples in the particle approximation of $p_\theta(z_{1:m} | y_{1:n}^\epsilon)$ for $m$ much smaller than $n$, which is called particle degeneracy in the literature. Poyiadjis, Doucet, and Singh (2011) showed that the variance of this $\mathcal{O}(N)$ score estimate, where the variance is computed with respect to the particles being sampled while the observation sequence is held fixed, grows quadratically with time. While this may not be an issue for the batch method in (13), it is not suitable for online estimation (15) since the variance of the resulting estimate of $\nabla \log p_{\theta_{n-1}}(y_n^\epsilon | y_{1:n-1}^\epsilon)$ grows linearly with time $n$.

As an alternative to this standard $\mathcal{O}(N)$ score estimate, Poyiadjis, Doucet, and Singh (2011) propose an $\mathcal{O}(N^2)$ estimate of the score computed using the same particle approximation to $p_\theta(z_{1:n} | y_{1:n}^\epsilon)$ which aims to avoid the particle degeneracy problem mentioned. We will refer to this as the $\mathcal{O}(N^2)$ method (Poyiadjis, Doucet, and Singh 2011, Algorithm 2). The authors experimentally show that the variance of the score estimate now grows linearly in time $n$ while the variance of the resulting estimate of $\nabla \log p_{\theta_{n-1}}(y_n^\epsilon | y_{1:n-1}^\epsilon)$ is time-uniformly bounded (i.e., does not grow); a proof of the latter fact can be found in Del Moral, Doucet, and Singh (2011). Therefore, the SMC implementation of $\nabla \log p_\theta(y_n^\epsilon | y_{1:n-1}^\epsilon)$ we adopt for online estimation (15) is the $\mathcal{O}(N^2)$ method.

Finally, we mention that the score (13) can also be estimated using a fixed-lag method which would have a computational cost which is $\mathcal{O}(N)$ and a variance which grows linearly in time. However, there is the added error introduced by not smoothing beyond a certain lag; see Kantas et al. (2009) for a review of static parameter estimation techniques.

### 3.2 CONTROLLING THE VARIANCE OF THE GRADIENT ESTIMATE

If the Monte Carlo estimates of the gradient terms have high or infinite variances, we expect failure of the gradient ascent methods. We can stabilize the variance by transforming the observed data, but without compromising the identifiability of the model, and then add noise as discussed in noisy ABC. This approach to stabilizing the variance is novel as the issue of infinite variance has not been reported before in the SMC literature.

This issue of the potential for infinite variance (prior to stabilizing by adopting a specific transformation) can be perfectly exemplified by the problem of learning the parameters of a distribution from a sequence of iid random variables which we now discuss. Let $\{Y_t\}_{t\geq 1}$ be

an iid sequence with an intractable probability density $g_\theta(y)$ on $\mathcal{Y}$. For any $\theta$, assume $Y_t$ can be sampled from $g_\theta$ by first generating $U_t \in \mathcal{U}$ from the density $\nu_\theta(u)$ and then followed by the application a certain transformation function $\tau_\theta : \mathcal{U} \to \mathcal{Y}$, that is, the law of $\tau_\theta(U_t)$ is $g_\theta$. (The $\alpha$-stable process is generated precisely in this way; see Example 1) We are given a realization $\hat{y}_1, \hat{y}_2, \ldots$ from $\theta^*$ and the latter is to be estimated. Let $y_t^\epsilon$ be the noise corrupted observed sequence as in (9). In the context of the discussion in Section 3, the aim is to maximize the likelihood of the noisy observations $y_{1:n}^\epsilon$ (generated from the true model $\theta^*$) using the parametric family of HMMs $\{U_t, Y_t^\epsilon\}_{t\geq 1}$. Since $\{U_t\}_{t\geq 1}$ are iid the batch (13) and online (15) update rules become, respectively,

$$\theta_j = \theta_{j-1} + \gamma_j \sum_{t=1}^{n} \nabla \log p_{\theta_{j-1}}(y_t^\epsilon) \quad \text{and} \quad \theta_n = \theta_{n-1} + \gamma_n \nabla \log p_{\theta_{n-1}}\big(y_n^\epsilon\big).$$

$h_\theta^\epsilon$ in (11) becomes $h_\theta^\epsilon(y|u) = \frac{1}{\epsilon}\kappa\left(\frac{y-\tau_\theta(u)}{\epsilon}\right)$ and

$$\nabla \log p_\theta(y^\epsilon) = \int_{\mathcal{Y}} \left[\nabla \log \nu_\theta(u) + \nabla \log h_\theta^\epsilon(y^\epsilon|u)\right] p_\theta(u|y^\epsilon) du, \tag{17}$$

where $p_\theta(u|y^\epsilon) \propto h_\theta^\epsilon(y^\epsilon|u)\nu_\theta(u)$. Therefore, $\nabla \log p_\theta(y_n^\epsilon)$ can be estimated using an $N$-sample Monte Carlo approximation to $p_\theta(u|y_n^\epsilon)$, for example, with either MCMC or importance sampling. One important point to note about this iid case is that the $\mathcal{O}(N^2)$ method becomes $\mathcal{O}(N)$.

We now calculate the variance of the Monte Carlo estimate of (17) at $\theta = \theta^*$ given $N$ iid samples from $p_\theta(u|y_n^\epsilon)$. (Note that in the numerical examples we actually use importance sampling to sample from $p_\theta(u|y_n^\epsilon)$ but the following calculation is done assuming iid samples are available for illustrative purposes.) Dropping the index $t$, given a noise corrupted measurement $Y^\epsilon$ generated from the true model $\theta^*$, and iid samples $U_1, \ldots, U_N \overset{\text{iid}}{\sim} p_{\theta^*}(u|Y^\epsilon)$, an estimate of $\nabla \log p_{\theta^*}(Y^\epsilon)$ is

$$\frac{1}{N}\sum_{i=1}^{N} \frac{1}{\epsilon^2} \nabla \tau_{\theta^*}(U_i)[Y^\epsilon - \tau_{\theta^*}(U_i)] + \nabla \log \nu_{\theta^*}(U_i).$$

We are interested in the variance of this quantity with respect to the law of $(U_{1:N}, Y^\epsilon)$. We consider the case where $\nabla \log \nu_{\theta^*}(U)$ has a finite second moment; for example, see the example to follow. Then, the sum above has a finite second moment if and only if $\nabla \tau_{\theta^*}(U_i)\,[Y^\epsilon - \tau_{\theta^*}(U_i)]$ has a finite second moment with respect to the joint law of $(U_i, Y^\epsilon)$. One can show that

$$\mathbb{E}_{\theta^*}\left[\{\nabla \tau_{\theta^*}(U_i)[Y^\epsilon - \tau_{\theta^*}(U_i)]\}^2\right] = \epsilon^2 \mathbb{E}_{\theta^*}[\{\nabla \tau_{\theta^*}(U_i)\}^2]. \tag{18}$$

If the second moment of $\nabla \tau_{\theta^*}$ is infinite (or very high), we may circumvent this instability problem by transforming the *actual* observed process from $\theta^*$ using a suitable one-to-one function $\psi : \mathcal{Y} \to \mathcal{Y}_s$ prior to adding noise. That is, we replace (9) with the following transformed noise corrupted process

$$Y_t^\epsilon = \psi(Y_t) + \epsilon V_t, \quad V_t \overset{\text{iid}}{\sim} \kappa, \quad t \geq 1. \tag{19}$$

The conditional density $h_\theta^\epsilon(y|u)$ becomes

$$h_\theta^\epsilon(y|u) = \frac{1}{\epsilon}\kappa\left(\frac{y - \psi[\tau_\theta(u)]}{\epsilon}\right)$$

and the right-hand side of (18) now is $\epsilon^2 \mathbb{E}_{\theta^*}[\{\nabla\psi(\tau_{\theta^*}(U_i))\}]$. (Note that this transformation can cause a loss of efficiency due to "squeezing" observations in a smaller interval and rendering the noisy data less informative about the original samples, hence $\theta^*$.) In this article we use $\psi = \tan^{-1}$ throughout, and in the following example we show how (18) is infinite but subsequently stabilized with this transformation.

*Example 1. (The $\alpha$-stable distribution)* Let $\mathcal{A}(\alpha, \beta, \mu, \sigma)$ denote the $\alpha$-stable distribution. The parameters of the distribution,

$$\theta = (\alpha, \beta, \mu, \sigma) \in \Theta = (0, 2] \times [-1, 1] \times \mathbb{R} \times [0, \infty),$$

represent the shape, skewness, location, and scale, respectively. One can generate a random sample from $\mathcal{A}(\alpha, \beta, \mu, \sigma)$ by generating $U = (U_1, U_2)$, where $U_1 \sim \text{Unif}(-\pi/2, \pi/2)$ and $U_2 \sim \text{Exp}(1)$ are independent, and setting

$$Y = \tau_\theta(U) = \sigma\tau_{\alpha,\beta}(U) + \mu.$$

The mapping $\tau_{\alpha,\beta}$ is defined in Chambers, Mallows, and Stuck (1976)

$$\tau_{\alpha,\beta}(U) = \begin{cases} S_{\alpha,\beta}\dfrac{\sin[\alpha(U_1 + B_{\alpha,\beta})]}{[\cos(U_1)]^{1/\alpha}}\left(\dfrac{\cos[U_1 - \alpha(U_1 + B_{\alpha,\beta})]}{U_2}\right)^{(1-\alpha)/\alpha}, & \alpha \neq 1 \\ X = \dfrac{2}{\pi}\left[\left(\dfrac{\pi}{2} + \beta U_1\right)\tan U_1 - \beta\log\left(\dfrac{U_2\cos U_1}{\frac{\pi}{2} + \beta U_1}\right)\right], & \alpha = 1, \end{cases}$$

where

$$B_{\alpha,\beta} = \frac{\tan^{-1}\left(\beta\tan\frac{\pi\alpha}{2}\right)}{\alpha} \quad S_{\alpha,\beta} = \left(1 + \beta^2\tan^2\frac{\pi\alpha}{2}\right)^{1/2\alpha}.$$

Although it is hard to show for $\alpha$ and $\beta$, we can show that

$$\mathbb{E}_\theta\left[\left\{\frac{\partial}{\partial\sigma}\tau_\theta(U)\right\}^2\right] = \mathbb{E}_\theta[\{\tau_{\alpha,\beta}(U)\}^2] = \infty$$

unless $\alpha = 2$. Therefore, it is not desirable to run the gradient ascent method for the process $\{Y_t^\epsilon\}_{t\geq 1}$ with $Y_t^\epsilon = Y_t + \epsilon V_t$ since the variance of the gradient estimate will be infinite. Instead, we use the transformation $\psi = \tan^{-1}$, that is, $Y_t^\epsilon = \tan^{-1}(Y_t) + \epsilon V_t$ to make the gradient ascent method stable. One can indeed check that for the parameter $\sigma$

$$\mathbb{E}_\theta\left[\left\{\frac{\partial}{\partial\sigma}\psi[\tau_\theta(U)]\right\}^2\right] = \mathbb{E}_\theta\left[\left\{\frac{\tau_{\alpha,\beta}(U)}{1 + \tau_\theta(U)^2}\right\}^2\right] < \infty.$$

We also verify numerically in Section 4 that the gradients with respect to the other parameters $\alpha$, $\beta$ are stabilized with $\psi = \tan^{-1}$ (while we can show that $\mathbb{E}_\theta[\{\partial\tau_\theta(U)/\partial\mu\}^2] = 1$).

### 3.3 Other MLE Methods for HMMs With an Intractable Density

Although not as general as the gradient ascent MLE approach, the expectation-maximization (EM) algorithm may be available for some models, at least for a part of the parameters in $\theta$, if the joint density $p_\theta(z_{1:n}, y_{1:n}^\epsilon)$ belongs to an exponential family. Both

$\mathcal{O}(N)$ and $\mathcal{O}(N^2)$ batch and online EM algorithms can be devised using SMC; details of such algorithms can be found in Cappé (2009) and Del Moral, Doucet, and Singh (2009).

There are other gradient MLE methods in the literature that are available for implementing noisy ABC MLE and we have discussed the technique of Ehrlich, Jasra, and Kantas (2013) in the introduction. One advantage of their finite difference method is that it is essentially a gradient-free technique as it bypasses having to calculate the derivatives with respect to $\theta$ of the state transition and observation densities of the HMM and thus can cope, without modification, with an intractable state transition density. Another gradient-based method that uses SMC to approximate the gradient of the log-likelihood without the need to calculate the derivatives of the HMMs densities is the iterated filtering algorithm of Ionides, Bhadra, and King (2011). In particular, one can use iterated filtering for $\{(X_t, Y_t), Y_t^\epsilon\}_{t\geq 1}$ or $\{(X_t, U_t), Y_t^\epsilon\}_{t\geq 1}$ to estimate $\nabla \log p_\theta(y_{1:n}^\epsilon)$. However, the method does not have an extension to online estimation. Another downside is that the algorithm requires an increasing number of particles versus iteration for convergence.

Coquelin, Deguest, and Munos (2009) studied an HMM with a tractable observation density $g_\theta(y|x)$ but an intractable state transition density $f_\theta(x'|x)$. Assume one can generate from $f_\theta(\cdot|x)$ by sampling $U$ from $\mu_\theta(\cdot|x)$ and using a differentiable function $F_\theta : \mathcal{X} \times \mathcal{U} \to \mathcal{X}$ such that $F_\theta(U, x) \sim f_\theta(\cdot|x)$. The gradient of the log-likelihood in such HMMs can be estimated using the infinitesimal perturbation analysis (IPA) approach proposed by Coquelin, Deguest, and Munos (2009), provided that $\mu_\theta(\cdot|x)$, $F_\theta(u, x)$, and $g_\theta(y|x)$ are differentiable with respect to $\theta$ as well as the state variable $x$. We can straightforwardly adopt the IPA approach with our noisy ABC MLE to deal with a fully intractable model, where both the state transition and the observation densities are intractable. However, IPA is a path space method and suffers from particle degeneracy. This will lead to the variance of the estimate of the score in (14) increasing quadratically in time like the $\mathcal{O}(N)$ method in Poyiadjis, Doucet, and Singh (2011). As the authors mentioned, fixed-lag smoothing could be use to control this variance growth but at the cost of a small bias.

## 4. NUMERICAL EXAMPLES

In this section we demonstrate the performance of the gradient ascent methods described in Section 3 on the iid $\alpha$-stable and *g-and-k* models as well as the stochastic volatility model with $\alpha$-stable returns.

### 4.1 MLE FOR IID $\alpha$-STABLE RANDOM VARIABLES

We first consider the problem of estimating the parameters of an $\alpha$-stable distribution $\mathcal{A}(\alpha, \beta, \mu, \sigma)$ (developed in Example 1) from a sequence of iid samples. Several methods for estimating parameter values for stable distributions have been proposed, including a Bayesian approach based on ABC; see Peters, Sisson, and Fan (2011). In this example we consider estimating these parameters using the online gradient ascent method to implement noisy ABC MLE. Since the only discontinuity in the transformation function $\tau_\theta$ for generating an $\alpha$-stable random variable is at $\alpha = 1$, we can safely use the gradient ascent method for estimating $\theta^*$ with $\alpha^*$ being not in the close vicinity of 1.

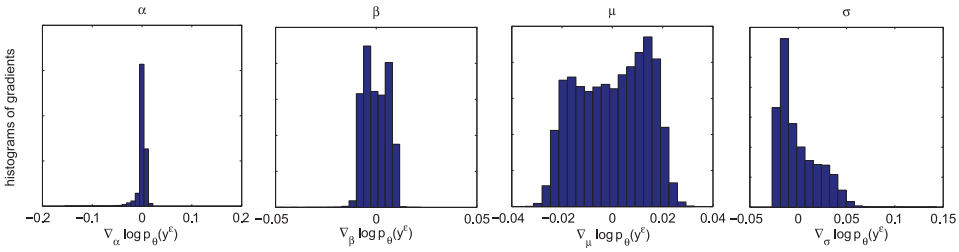


Figure 1. Histograms of estimates of $\nabla \log p_\theta(y_i^\epsilon)$, $1 \le i \le 10^5$ computed at $\theta = (1.5, 0.2, 0, 0.5)$ where $y_i^\epsilon = \tan^{-1}(\hat{y}_i) + 0.1 v_i$, $\hat{y}_i \overset{\text{iid}}{\sim} \mathcal{A}(1.5, 0.2, 0, 0.5)$, $v_i \overset{\text{iid}}{\sim} \mathcal{N}(0, 1)$.

As recommended in Example 1, we transform the observations using $\psi = \tan^{-1}$ for stability. To check, numerically, whether the transformation in (19) with $\psi = \tan^{-1}$ stabilizes the gradients, we can look at the empirical distribution of the Monte Carlo estimates of $\nabla \log p_\theta(Y_i^\epsilon)$ after transforming the observations $Y_i$. For this purpose, we generate $10^5$ samples $\hat{y}_i$ from $\mathcal{A}(1.5, 0.2, 0, 0.5)$ and $v_i$ from $\kappa$ for $i = 1, \ldots, 10^5$, and for each sample we estimate $\nabla \log p_\theta(y_i^\epsilon)$, where $y_i^\epsilon = \tan^{-1}(\hat{y}_i) + \epsilon v_i$, with $\epsilon = 0.1$, using self-normalized importance sampling with $N = 1000$ samples generated from $\nu_\theta$. Figure 1 shows the histograms of the Monte Carlo estimates of $\nabla \log p_\theta(y_i^\epsilon)$ which confirms that the transformation does stabilize the gradients.

The outcome of online gradient ascent method to implement noisy ABC MLE for the same dataset is shown in Figure 2. A trace plot of the sequence of gradient estimates (as $\theta$ is adjusted) is also shown as further confirmation of the stability of the estimated gradients.

The next experiment contrasts the ABC MLE and noisy ABC MLE solutions for the same dataset. The results in Figure 3 compare the online $\theta^*$ estimates averaged over 50

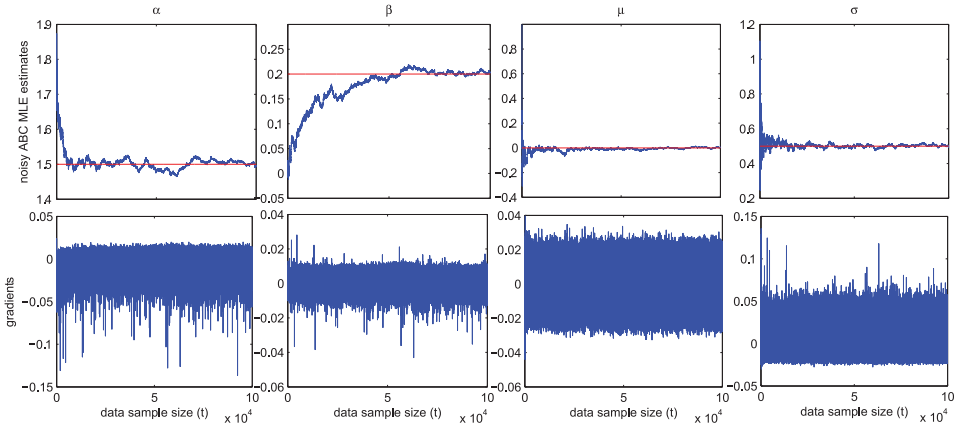


Figure 2. Online estimation of $\alpha$-stable parameters (top figure) from a sequence of iid random variables using online gradient ascent MLE and the corresponding online gradient estimates of the incremental likelihood (bottom figure). $\theta^* = (\alpha^*, \beta^*, \mu^*, \sigma^*) = (1.5, 0.2, 0, 0.5)$ is indicated with a horizontal line. At the bottom: Gradient of incremental likelihood for the $\alpha$-stable parameters.

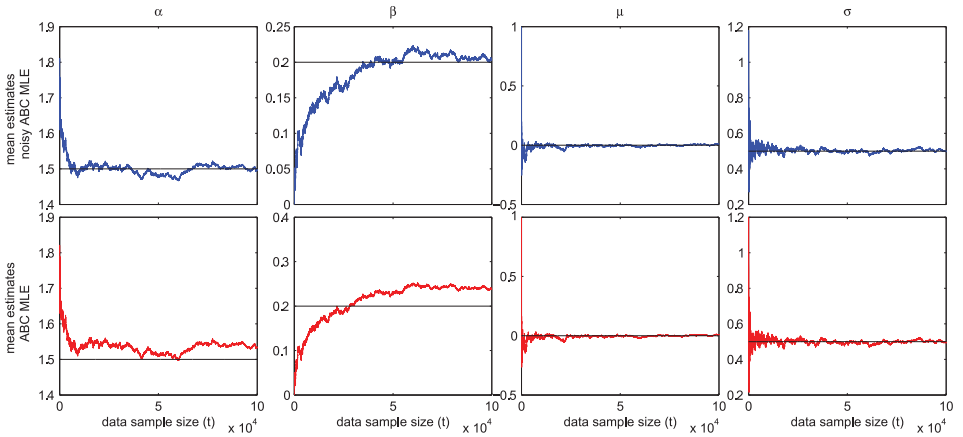


Figure 3. ABC MLE and noisy ABC MLE estimates of the parameters of the $\alpha$-stable distribution (averaged over 50 runs) using the online gradient ascent algorithm for the same dataset. For noisy ABC MLE, a different noisy data sequence obtained from the original dataset is used in each run. $\theta^* = (\alpha^*, \beta^*, \mu^*, \sigma^*) = (1.5, 0.2, 0, 0.5)$ is indicated with a horizontal line.

independent runs for both algorithms. Each run used the same dataset but a new realization of particles. The outcome of this comparison is that ABC MLE yields biased estimates for the shape and skewness parameters $\alpha$ and $\beta$ whereas the bias is not present in noisy ABC MLE.

### 4.2 MLE FOR *g*-AND-*k* DISTRIBUTION

The *g-and-k* distribution is defined by the following parameterized quantile (or inverse distribution) function $Q_\theta$

$$Q_\theta(u) = F_\theta^{-1}(u) = A + B\left[1 + c\frac{1 - e^{-g\phi(u)}}{1 + e^{-g\phi(u)}}\right](1 + \phi(u)^2)^k\phi(u), \quad u \in (0, 1), \tag{20}$$

where $\phi(u)$ is the $u$th standard normal quantile. The parameters

$$\theta = (g, k, A, B) \in \Theta = \mathbb{R} \times (-0.5, \infty) \times \mathbb{R} \times [0, \infty)$$

are the skewness, kurtosis, location, and scale, and $c$ is usually fixed to 0.8. Therefore, one can generate from the *g-and-k* distribution by first sampling $U \sim \text{Unif}(0, 1)$ and then returning $\tau_\theta(U) = Q_\theta(U)$ (Rayner and MacGillivray 2002).

Bayesian parameter estimation for the *g-and-k* distribution using ABC was recently proposed by Fearnhead and Prangle (2012). We consider online MLE for $\theta$ using the noisy ABC likelihood. Note that $Q_\theta$ in (20) is differentiable with respect to $\theta$ and so the gradient ascent method is applicable. To avoid gradients with very high variances resulting from the factor $\left(1 + \phi(u)^2\right)^k$ in $Q_\theta$, similar to the case of $\alpha$-stable distribution, we transform the actual observations using $\psi = \tan^{-1}$ and add noise with $\epsilon = 0.1$. In our experiments it was noticed that our method performs better when the location parameter $A$ is closer to 0, which must be a result of the nonlinear behavior of the transformation function $\tan^{-1}$. Therefore, whenever possible, it is suggested to estimate $A$ using some (possibly

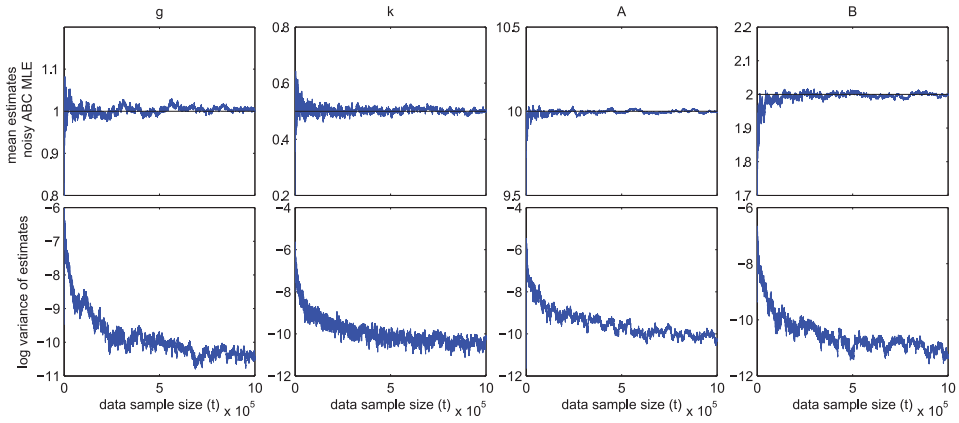


Figure 4. Mean and the variance (over 50 runs) of noisy ABC MLE estimates using the online gradient ascent algorithm. Same noisy data sequence used in each run. $\theta^* = (g^*, k^*, A^*, B^*) = (1, 0.5, 10, 2)$ indicated by horizontal lines.

heuristic) method (such as using the mean of the first few samples) as a preprocessing step, subtract the heuristically estimated value $\hat{A}$ of $A$ from the samples, perform MLE on the (approximately) centred data, and then add back $\hat{A}$ to the estimated location obtained by the MLE algorithm.

Figure 4 shows the results of online gradient ascent method (15) to implement noisy ABC MLE for estimating $\theta^* = (1, 0.5, 10, 2)$. In the figure we observe the mean and log-variance of 50 runs on the *same* noisy transformed data sequence. (Therefore, the accuracy and the variance of the estimates correspond to the performance of the Monte Carlo approximation of the gradients $\nabla \log p_\theta(y_i^\epsilon)$.) Self-normalized importance sampling is used with $N = 1000$ samples generated from $\nu_\theta$. From the results in Figure 4, we can see that the bias introduced by the finite number of particles is negligible for $N = 1000$ and that the variance of the algorithm reduces in time suggesting the convergence of the estimates in each run to essentially the true parameter values.

The next experiment shows how the noisy ABC MLE can be implemented with the batch gradient ascent method (13) when the dataset is too small for the online method to converge. A detailed study of MLE for *g-and-k* distribution can be found in Rayner and MacGillivray (2002) where MLE methods based on numerical approximation of the likelihood itself are investigated. We generated 500 datasets of size $n = 1000$ from the same *g-and-k* distribution with $\theta^* = (2, 0.5, 10, 2)$ and executed the batch gradient ascent method with $\epsilon = 0.1$ on each dataset. Again, self-normalized importance sampling is used with $N = 1000$ samples. The upper half of Figure 5 shows the estimation results with noisy ABC MLE versus number of iterations for a single dataset. Note that for short datasets, $\theta^*$ is usually not the true maximum likelihood solution. The lower half of Figure 5 shows the distributions (histograms over 20 bins) of the converged maximum likelihood solution for $\theta^*$. The mean and variance of the estimates for $(g, k, A, B)$ are $(2.004, 0.503, 9.995, 1.996)$ and $(0.0151, 0.0021, 0.0052, 0.0213)$, respectively. Comparable values for these moments at this particular $\theta^*$ and data size $n$ were also obtained in Rayner and MacGillivray (2002, Table 3).

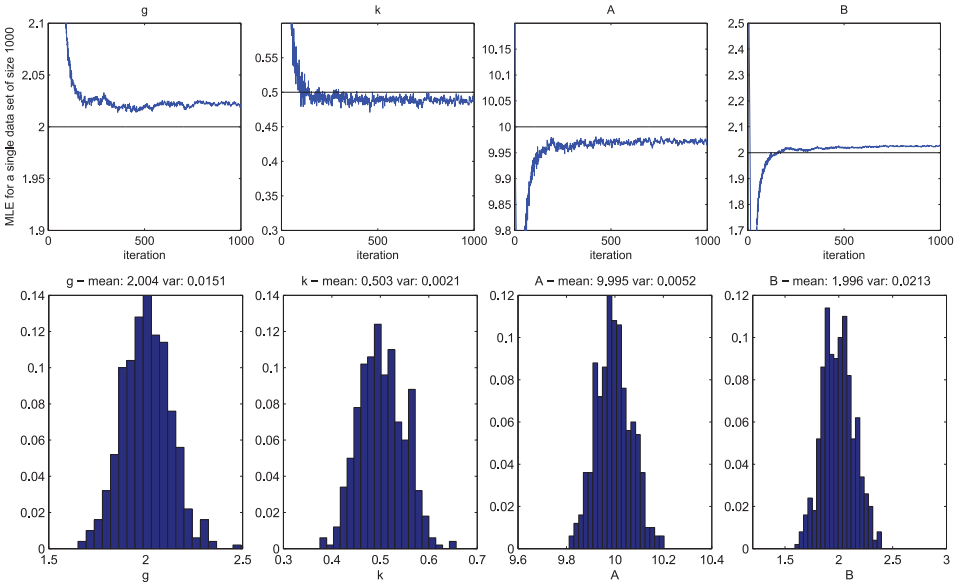


Figure 5. Top: Noisy ABC MLE estimates of *g-and-k* parameters from a sequence of 1000 iid random variables computed using the batch gradient ascent algorithm. $\theta^* = (g^*, k^*, A^*, B^*) = (2, 0.5, 10, 2)$ indicated by horizontal lines. Bottom: Empirical distributions of the estimates over 500 datasets.

### 4.3 THE STOCHASTIC VOLATILITY MODEL WITH SYMMETRIC $\alpha$-STABLE RETURNS

The stochastic volatility model with $\alpha$-stable returns (SV$\alpha$R) is a financial data model (Lombardi and Calzolari 2009). The hidden process $\{X_t\}_{t\geq 1}$ represents the log-volatility in time whereas the observation process $\{Y_t\}_{t\geq 1}$ is the log return values. The model for $\{X_t, Y_t\}_{t\geq 1}$ with parameters $\theta = (\alpha, \phi, \sigma_x^2)$ is

$$X_t = \phi X_{t-1} + S_t, \quad S_t \sim \mathcal{N}\big(0, \sigma_x^2\big), \quad Y_t = \exp(X_t/2)W_t, \quad W_t \sim \mathcal{A}(\alpha, 0, 0, 1).$$

This model is an alternative to the stochastic volatility model with Gaussian returns to account for an observed series which is heavy-tailed and displays outliers. For more discussion on the model as well as a review of methods for estimating the static parameters of such models, see Lombardi and Calzolari (2009) and the references therein. These existing methods for parameter estimation in SV$\alpha$R are batch and suitable for only short data sequences. We simulated a scenario where a very long data sequence generated from this model with $\theta^* = (1.9, 0.9, 0.1)$ is being received sequentially. We used online gradient ascent method (15) to find the noisy ABC MLE solution for this data sequence, where the $\mathcal{O}(N^2)$ method (Poyiadjis, Doucet, and Singh 2011, Algorithm 2) with $N = 500$ particles was used to estimate (16). Again, we transform the actual observations with the function $\psi = \tan^{-1}$ and then add noise. Figure 6 shows the online estimates of $\theta^*$ for $2 \times 10^6$ data samples. The estimates seem to converge after around $5 \times 10^5$ samples and are accurate.

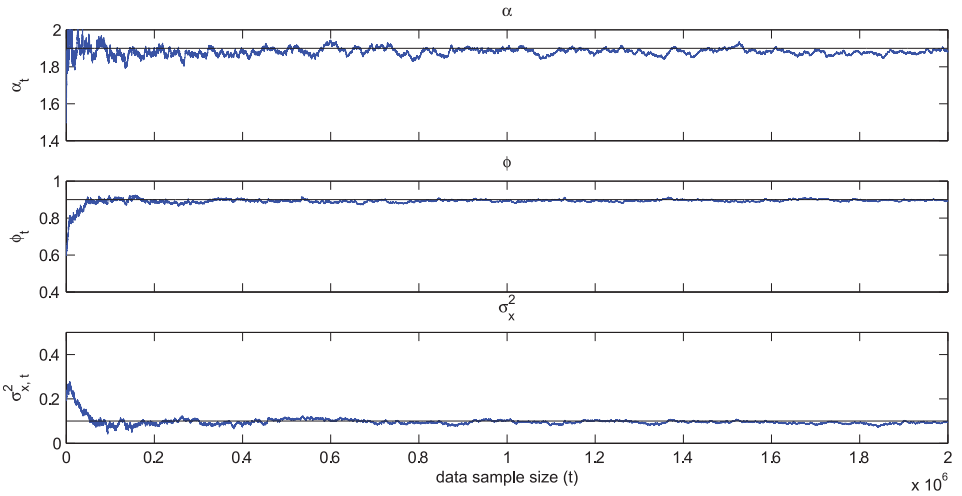


Figure 6. Online estimation of SV$\alpha$R parameters using online gradient ascent algorithm to implement noisy ABC MLE. $\theta^* = (\alpha^*, \phi^*, \sigma_x^{2,*}) = (1.9, 0.9, 0.1)$ indicated with horizontal lines.

### 4.4 Offline Noisy ABC MLE for Real Data

We now consider a real data experiment, where the data are the daily GBP-DEM exchange rates between 01.01.1987 to 31.12.1995 containing 3287 samples $o_1, \ldots, o_{3287}$; these data are considered in Lombardi and Calzolari (2009). Log-returns $r_{1:3286}$ are obtained by $r_t = 100 \log(o_{t+1}/o_t)$, $1 \le t \le 3286$. The observations, $\hat{y}_{1:3285}$, are the residuals of the AR(1) process that is fitted to $r_{1:3286}$. (We used the same model and dataset as Lombardi and Calzolari (2009) to compare our results with theirs.) The SV$\alpha$R model above is assumed for $\hat{y}_{1:n}$, where the hidden process has an extra parameter $\delta$:

$$X_t = \phi X_{t-1} + \delta + S_t, \quad S_t \sim \mathcal{N}\big(0, \sigma_x^2\big),$$

hence, $\theta = (\alpha, \phi, \sigma_x^2, \delta)$.

We implemented noisy ABC MLE using batch gradient ascent (13) with the $\mathcal{O}(N)$ method (Poyiadjis, Doucet, and Singh 2011, Algorithm 1) with $N = 2000$ particles to approximate (14). To measure the variability of the estimates as a function of the realization of added noise and the $\epsilon$ value, we repeated the estimation with $\epsilon = 0.05$, $\epsilon = 0.1$, and $\epsilon = 0.2$, separately, where for each $\epsilon$ we ran the method with 10 different added noise realizations. For all runs, we terminated the batch gradient ascent algorithm after 20, 000 iterations. $N = 2000$ particles were used to evaluate the gradients at each iteration. Figure 7 (top) shows the estimates versus number of iterations, where the trajectories for different noisy datasets for the same value of $\epsilon$ are superimposed. Also, the bottom part of Figure 7 shows the boxplots of the estimates of $\theta^*$ for different $\epsilon$ values, where the boxplots for each $\epsilon$ were created from the converged estimates of $\theta^*$ (the average of the estimates at the last 1000 iterations) obtained from 10 different noisy datasets generated using that value of $\epsilon$. For the ease of explanation, we will denote them as

$$\theta_{0.05}^{(1)}, \ldots, \theta_{0.05}^{(10)}; \quad \theta_{0.1}^{(1)}, \ldots, \theta_{0.1}^{(10)}; \quad \theta_{0.2}^{(1)}, \ldots, \theta_{0.2}^{(10)}, \tag{21}$$

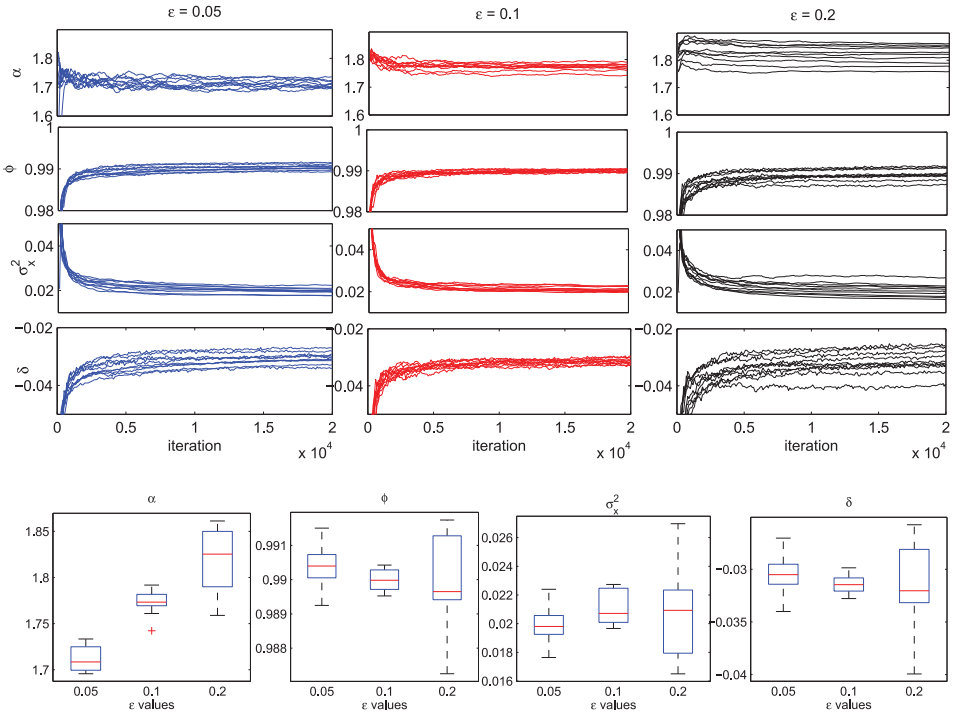


Figure 7. Top: results for noisy ABC MLE implemented by the $\mathcal{O}(N)$ batch gradient MLE algorithm for three different values of $\epsilon$. Estimates vs number of iterations for different noisy datasets are superimposed for the same value of $\epsilon$. Bottom: boxplots of the batch ABC MLE estimates vs $\epsilon$. The boxplot for each $\epsilon$ was created from $\theta_\epsilon^{(1)}, \ldots, \theta_\epsilon^{(10)}$, the converged estimates obtained from the trace plots at the top.

where $\theta_\epsilon^{(i)}$ is the converged estimate obtained from the $i$th noisy dataset that was generated using $\epsilon$.

Figures 7 suggests a tradeoff between accuracy in the estimates and computational efficiency in the following sense. A smaller value of $\epsilon$ is expected yield less biased estimates (with respect to the maximizer of the true likelihood of the real data) with less variance (with respect to the added noise) provided that the maximization $\arg\max_{\theta\in\Theta} p_\theta(Y_{1:n}^\epsilon = y_{1:n}^\epsilon)$ is performed exactly, that is, with infinitely many $N$ and infinitely many number of parameter updates. On the other hand, smaller $\epsilon$ results in the decrease of the effective sample size in the SMC algorithm and hence increases the variance of the SMC estimate of the gradient of the log-likelihood. The effect of this on our results is the larger variance in the estimates obtained with $\epsilon = 0.05$ compared to those obtained with $\epsilon = 0.1$ (which would eventually be smaller if the maximization were performed exactly). In conclusion, for a fixed batch data size and a given amount of computational resource, one must optimize the tradeoff between the (average) accuracy and the variability in the estimates, for which the effective sample size of the particles could be used as a rule of thumb.

Lombardi and Calzolari (2009) fit the same model to the same dataset using the indirect estimation method and their estimates of $\theta^*$ was $\theta_{\text{ind}} := (1.7963, 0.9938, 0.0940^2, -0.0076)$, which is slightly different to our results. Both methods (theirs and ours) aim for the maximum likelihood solution, which suggests that it would

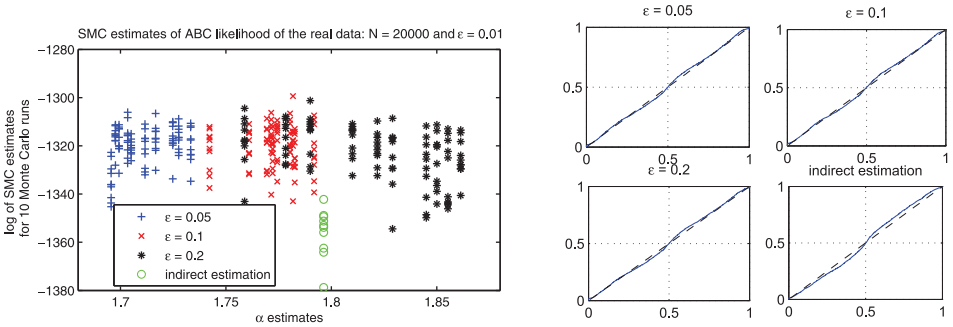


Figure 8. Left: Logarithm of the 10 different SMC estimates (with $N = 20{,}000$) of $p_\theta(Y^\epsilon_{1:n} = \hat{y}_{1:n})$, $\epsilon = 0.01$. Each color represents a different $\epsilon$ value which was used to obtain the noisy datasets and the ABC MLE estimates from them. For the blue, red, and black points, their horizontal axis locations correspond to the $\alpha$ components of $\theta^{(1)}_\epsilon, \ldots, \theta^{(10)}_\epsilon$ for $\epsilon = 0.05$, $\epsilon = 0.1$, and $\epsilon = 0.2$, respectively. Similarly, the horizontal axis location of the black points is the $\alpha$ component of the estimate of $\theta^*$ obtained using the indirect estimation method. Right: Empirical cumulative distribution plots for model checking: for each $\epsilon$ value, $\theta$ is taken to be the mean of $\theta^{(1)}_\epsilon, \ldots, \theta^{(10)}_\epsilon$.

be sensible to compare the likelihood of the true data sequence for the estimates of $\theta^*$ obtained from both methods. However, this is not possible since neither $p_\theta(Y_{1:n} = \hat{y}_{1:n})$ nor an unbiased Monte Carlo estimator of it is available. Instead, we compared the unbiased SMC estimates of the ABC likelihoods $p_\theta(Y^\epsilon_{1:n} = \hat{y}_{1:n})$ using an $\epsilon$ small enough to make the effect of model mismatch negligible (see the discussion of model mismatch error in Section 2) for comparison and $N$ large enough to ensure that the variability of the SMC estimate of the likelihood across the particle realizations is not too much; for these reasons we chose $\epsilon = 0.01$ and $N = 20{,}000$. (See the Appendix for the details of the implementation.) The left-hand side of Figure 8 shows the logarithms of the 10 independent SMC estimates of $p_\theta(Y^\epsilon_{1:n} = \hat{y}_{1:n})$ calculated at the value of each estimate in (21). For comparison, the results are shown with 10 independent SMC estimates of $p_\theta(Y^\epsilon_{1:n} = \hat{y}_{1:n})$ at $\theta = \theta_{\text{ind}}$. The figure shows that noisy ABC MLE has improved the results of Lombardi and Calzolari (2009) for all values of $\epsilon$ that we used, in the sense that almost all the estimates resulting from the ABC MLE method yields a higher likelihood of the dataset to which the model is fitted.

Finally, we perform a simple model check for by considering the conditional cumulative distribution functions

$$F_{\theta,t}(\hat{y}_t; \hat{y}_{1:t-1}) := P_\theta(Y_t \le \hat{y}_t | Y_{1:t-1} = \hat{y}_{1:t-1}), \quad t = 1, \ldots, n$$

at the values of $\theta^*$ estimated using noisy ABC MLE and indirect estimation in Lombardi and Calzolari (2009). Since $\{F_{\theta,t}(Y_t; Y_{1:t-1})\}_{1 \le t \le n}$ are iid uniform random variables on [0, 1] (Diebold, Gunther, and Tay 1998), we expect the probability plot (for the uniform distribution) of the population $\{F_{\theta,t}(\hat{y}_t; \hat{y}_{1:t-1})\}_{1 \le t \le n}$ to approximate the $y = x$ line under the hypothesis that $\hat{y}_{1:n}$ is generated from the SV$\alpha$R model $\{X_t, Y_t\}_{t \ge 1}$. However, we are unable to perform these calculations for the original HMM due to the intractability of $g_\theta(y|x)$. Instead, we use the modified HMM $\{(X_t, Y_t), Y^\epsilon_t\}_{t \ge 1}$ but with $\epsilon$ small enough for one to neglect the difference between the two models (as in the previous experiment). The probability plots at the right-hand side of Figure 8 were generated from the SMC estimates

of

$$F_{\epsilon,\theta,t}(\hat{y}_t; \hat{y}_{1:t-1}) := P_\theta(Y_t^\epsilon \leq \hat{y}_t | Y_{1:t-1}^\epsilon = \hat{y}_{1:t-1}), \quad t = 1, \ldots, n,$$

(see the Appendix for details), with $\epsilon = 0.01$ and $N = 20{,}000$, for four different values of $\theta$: the first three are the means of $\theta_\epsilon^{(1)}, \ldots, \theta_\epsilon^{(10)}$ for $\epsilon = 0.05$, $\epsilon = 0.1$, and $\epsilon = 0.2$, respectively, and the fourth one is $\theta_{\text{ind}}$. The probability plots are all close to the $y = x$ line which justifies the SV$\alpha$R model; they also indicate that there is more agreement between the SV$\alpha$R model and the data when $\theta$ is the noisy ABC MLE solution than when it is the maximum likelihood solution of the indirect estimation method.

## 5. DISCUSSION

In this article, we have presented SMC implementations of MLE for HMMs with an intractable observation density. We showed how SMC versions of both batch and online gradient ascent algorithms can be used to implement ABC MLE and noisy ABC MLE and how a further transformation of the data can stabilize the variance of the SMC gradient estimate. We have shown that SMC implementations of the methodology in Dean et al. (2014) is practical and yields convergent and accurate estimates of $\theta^*$ even when the exact procedures in Dean et al. (2014) are replaced by their SMC counterparts.

We have implemented noisy ABC MLE in our experiments. In general though there is a choice to be made between ABC MLE or noisy ABC MLE, which may be resolved by taking into account the mean squared error (MSE) of the estimate of $\theta^*$. For long datasets, noisy ABC MLE may be more appropriate since it removes the asymptotic bias of ABC MLE (which is $O(\epsilon^2)$ (Dean and Singh 2011)) so that the MSE is dominated roughly by $1/n$ times the variance of the central limit theorem (CLT). Noisy ABC MLE suffers from an $O(\epsilon^2)$ increase in CLT variance over the ideal MLE procedure (Dean et al. 2014). For shorter datasets, ABC MLE would be more appropriate owing to the fact that the bias of ABC MLE is $O(\epsilon^2)$ and we may end up introducing more error to the estimate of $\theta^*$ (in the MSE sense) by adding noise to the data.

## APPENDIX

*Algorithm A.1. SMC for estimating* $p_\theta(Y_{1:n}^\epsilon = \hat{y}_{1:n})$ *and* $\{F_{\epsilon,\theta,t}(\hat{y}_t|\hat{y}_{1:t-1})\}_{1\leq t\leq n}$
Begin with $p_\theta(\hat{y}_0) = 1$. For $t = 1, \ldots, n$,

- *Prediction:* for $i = 1, \ldots, N$, sample $z_t^{(i)} = (x_t^{(i)}, u_t^{(i)})$ as follows:
  - If $t = 1$, sample $x_1^{(i)} \sim \eta_\theta(\cdot)$, $u_1^{(i)} \sim \nu_\theta(\cdot|x_1^{(i)})$
  - If $t > 1$, sample $x_t^{(i)} \sim f_\theta(\cdot|\bar{x}_{t-1}^{(i)})$, $u_t^{(i)} \sim \nu_\theta(\cdot|x_t^{(i)})$.
- *Weighting:* for $i = 1, \ldots, N$, calculate the unnormalized weights $w_t^{(i)} = h^\epsilon(\hat{y}_t|z_t^{(i)})$
- *Likelihood estimate:* Update the likelihood estimate by $p_\theta^N(\hat{y}_{1:t}) = p_\theta^N(\hat{y}_{1:t-1})\frac{1}{N}\sum_{i=1}^N w_t^{(i)}$.

- *Conditional cumulative distribution function:* Calculate

$$F^N_{\epsilon,\theta,t}(\hat{y}_t; \hat{y}_{1:t-1}) = \frac{1}{N}\sum_{i=1}^{N} w_t^{(i)} \int_{-\infty}^{\hat{y}_t} h^{\epsilon}(y|z_t^{(i)})dy$$

- *Resampling:* Sample $\{\bar{x}_t^{(i)}\}_{1\le i\le N}$ from $\{x_t^{(i)}\}_{1\le i\le N}$ using the weights $\{w_t^{(i)}\}_{i=1,\ldots,N}$.

## ACKNOWLEDGMENTS

S.S. Singh and T. Dean's research was funded by the Engineering and Physical Sciences Research Council (EP/G037590/1), whose support is gratefully acknowledged. A. Jasra was supported by an MOE Singapore grant (R-155-000-119-133) and is also affiliated with the Risk Management Institute at the National University of Singapore.

*[Received November 2013. Revised April 2014.]*

## REFERENCES

Andrieu, C., Doucet, A., and Tadić, V. B. (2005), "On-Line Parameter Estimation in General State-Space Models," in *Proceedings of the 44th IEEE Conference on Decision and Control*, pp. 332–337. [848]

Calvet, C., and Czellar, V. (2012), "Tracking Beliefs: Accurate Methods for Approximate Bayesian Computation Filtering," Technical Report 236, HEC Paris. [849]

Campillo, F., and Rossi, V. (2009), "Convolution Particle Filter for Parameter Estimation in General State-Space Models," *IEEE Transactions on Aerospace and Electronic Systems*, 45, 1063–1072. [848]

Cappé, O. (2009), "Online Sequential Monte Carlo EM Algorithm," in *Proceedings of the IEEE Workshop on Statistical Signal Processing*. [855]

Cappé, O., Moulines, E., and Rydén, T. (2005), *Inference in Hidden Markov Models*, New York: Springer. [846,851]

Chambers, J. M., Mallows, C. L., and Stuck, B. W. (1976), "Method for Simulating Stable Random Variables," *Journal of the American Statistical Association*, 71, 340–344. [854]

Chopin, N. (2002), "A Sequential Particle Filter Method for Static Models," *Biometrica*, 89, 539–551. [848]

Coquelin, P., Deguest, R., and Munos, R. (2009), "Sensitivity Analysis in HMMs With Application to Likelihood Maximization," in *Advances in Neural Information Processing Systems 22*, eds. Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, pp. 387–395. [855]

Crisan, D., and Doucet, A. (2002), "A Survey of Convergence Results on Particle Filtering Methods for Practitioners," *IEEE Transactions on Signal Processing*, 50, 736–746. [848]

Dean, T., and Singh, S. (2011), "Asymptotic Behaviour of Approximate Bayesian Estimators," Technical Report 1105.3655, arXiv.org. [849,863]

Dean, T., Singh, S., Jasra, A., and Peters, G. (2014), "Parameter Estimation for Hidden Markov Models With Intractable Likelihoods," *Scandinavian Journal of Statistics*. Available at arXiv:1103.5399. [849,850,863]

Del Moral, P. (2004), *Feynman-Kac Formulae: Genealogical and Interacting Particle Systems With Applications*, New York: Springer-Verlag. [848]

Del Moral, P., Doucet, A., and Singh, S. (2009), "Forward Smoothing Using Sequential Monte Carlo," Technical Report 638, Engineering Department, Cambridge University. [855]

——— (2011), "Uniform Stability of a Particle Approximation of the Optimal Filter Derivative," Technical Report CUED/F-INFENG/TR 668, Engineering Department, Cambridge University. [852]

Diebold, F. X., Gunther, T., and Tay, A. (1998), "Evaluating Density Forecasts, With Applications to Financial Risk Management," *International Economic Review*, 39, 863–883. [862]

Durbin, R., Eddy, S., Krogh, A., and Mitchison, G. (1998), *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*, Cambridge: Cambridge University Press. [846]

Ehrlich, E., Jasra, A., and Kantas, N. (2013), "Gradient Free Parameter Estimation for Hidden Markov Models With Intractable Likelihoods," *Methodology and Computing in Applied Probability (to appear)*. Available at *http://link.springer.com/article/10.1007%2Fs11009-013-9357-4#*. [847,855]

Fearnhead, P., and Prangle, D. (2012), "Constructing Summary Statistics for Approximate Bayesian Computation: Semi-Automatic Approximate Bayesian computation," *Journal of the Royal Statistical Society*, Series B, 74, 419–474. [857]

Felsenstein, J., and Churchill, G. (1996), "A Hidden Markov Model Approach to Variation Among Sites in Rate of Evolution," *Molecular Biology and Evolution*, 13, 93–104. [846]

Ionides, E. L., Bhadra, A., and King, A. (2011), "Iterated Filtering," *The Annals of Statistics*, 39, 1776–1802. [855]

Jasra, A., Singh, S., Martin, J., and McCoy, E. (2012), "Filtering via Approximate Bayesian Computation," *Statistics and Computing*, 22, 1223–1237. [849]

Kantas, N., Doucet, A., Singh, S. S., and Maciejowski, J. M. (2009), "An Overview of Sequential Monte Carlo Methods for Parameter Estimation in General State-Space Models," in *Proceedings of the IFAC System Identification (SysId) Meeting*, pp. 774–785. [847,848,852]

Kim, S., Shephard, N., and Chib, S. (1998), "Stochastic Volatility: Likelihood Inference and Comparison With ARCH Models," *The Review of Economic Studies*, 65, 361–393. [846]

Le Gland, F., and Mevel, L. (1997), "Recursive Estimation in Hidden Markov Models," in *Decision and Control, 1997, Proceedings of the 36th IEEE Conference on* (vol. 4), pp. 3468–3473. [852]

Lombardi, M. J., and Calzolari, G. (2009), "Indirect Estimation of $\alpha$-Stable Stochastic Volatility Models," *Computational Statistics & Data Analysis*, 53, 2298–2308. [859,861,862]

Marin, J.-M., Pudlo, P., Robert, C. P., and Ryder, R. J. (2012), "Approximate Bayesian Computational Methods," *Statistics and Computing*, 22, 1167–1180. [847]

Martin, J., Jasra, A., Singh, S. S., Whiteley, N., McCoy, E., and Del Moral, P. (2014), "ABC Smoothing," *Stochastic Analysis (to appear)*. Available at *http://www.tandfonline.com/doi/full/10.1080/07362994.2013.879262#.VFvVDVOsUnM*. [849]

Peters, G., Sisson, S., and Fan, Y. (2011), "Likelihood-Free Bayesian Inference for Alpha Stable Models," *Computational Statistics and Data Analysis*, 56, 3743–3756. [855]

Poyiadjis, G., Doucet, A., and Singh, S. S. (2011), "Particle Approximations of the Score and Observed Information Matrix in State Space Models With Application to Parameter Estimation," *Biometrika*, 98, 65–80. [850,852,855,859,860]

Rayner, G. D., and MacGillivray, H. L. (2002), "Numerical Maximum Likelihood Estimation for the g-and-k and Generalized g-and-h Distributions," *Statistics and Computing*, 12, 57–75. [857,858]

Rubio, D. B., and Johansen, A. M. (2013), "A Simple Approach to Maximum Intractable Likelihood Estimation," *Electronic Journal of Statistics*, 7, 1632–1654. [848]

Wilkinson, R. (2013), "Approximate Bayesian Computation (ABC) Gives Exact Results Under the Assumption of Model Error," *Statistical Approaches in Genetics and Molecular Biology*. Available at arXiv:0811.3355v2. [850]