PARTICLE FILTERING FOR BAYESIAN PARAMETER ESTIMATION IN A HIGH DIMENSIONAL STATE SPACE MODEL

Joaquín Míguez, Dan Crisan† and Inés P. Mariño‡*

* Department of Signal Theory & Communications, Universidad Carlos III de Madrid.
E-mail: *joaquin.miguez@uc3m.es*. Avda. de la Universidad 30, 28911 Leganés, Madrid (Spain).
† Department of Mathematics, Imperial College London.
E-mail: *d.crisan@imperial.ac.uk*. 180 Queens Gate, London SW7 2BZ (UK).
‡ Dept. of Biology, Geology, Physics & Inorganic Chemistry, Universidad Rey Juan Carlos.
E-mail: *ines.perez@urjc.es*. Tulipán s/n, 28933 Móstoles, Madrid (Spain).

ABSTRACT

Researchers in some of the most active fields of science, including, e.g., geophysics or systems biology, have to deal with very-large-scale stochastic dynamic models of real world phenomena for which conventional prediction and estimation methods are not well suited. In this paper, we investigate the application of a novel nested particle filtering scheme for joint Bayesian parameter estimation and tracking of the dynamic variables in a high dimensional state space model –namely a stochastic version of the two-scale Lorenz 96 chaotic system, commonly used as a benchmark model in meteorology and climate science. We provide theoretical guarantees on the algorithm performance, including uniform convergence rates for the approximation of posterior probability density functions of the fixed model parameters.

***Index Terms*—** Particle filtering, data assimilation, Bayesian parameter estimation, convergence analysis, kernel density estimation.

1. INTRODUCTION

Researchers in some of the most active fields of science have to deal with very large scale stochastic dynamic models of real world phenomena for which conventional prediction and estimation methods are not well suited. In fact, state-of-the-art methods for Bayesian model inference in computational statistics, e.g., the popular particle Markov chain Monte Carlo (pMCMC) [1] and sequential Monte Carlo square (SMC^2) [2] algorithms, are batch techniques and, therefore, they are not well suited to the processing of long observation sequences. Although some recursive algorithms exist [3], they only yield maximum likelihood (point) estimates for the parameters, and hence they are subject to various convergence (and complexity) issues when the likelihood is multimodal, contains singularities or cannot be computed exactly.

In this paper, we investigate a nested particle filtering (PF) scheme for the Bayesian estimation of the dynamic variables and the static parameters of state space models. The method is similar to the SMC^2 algorithm, but enjoys a purely recursive structure that makes it better suited for online estimation and dynamical systems (see [4] for additional details). We apply the new scheme to a stochastic version of the (chaotic) Lorenz 96 system. The latter displays the basic physical features of atmospheric dynamics and, for this reason, the deterministic version of this model is commonly used as a benchmark for data assimilation [5] and parameter estimation techniques [6] in meteorology and climate science. We illustrate the performance of the proposed scheme by means of computer simulations on a stochastic two-scale Lorenz 96 model [6] with 16 slow and 160 fast dynamic variables as well as several unknown parameters.

Besides the numerical results, we establish theoretical guarantees on the performance of the proposed scheme. In particular, we prove that kernel approximations of posterior probability density functions (pdf's) of the model parameters converge uniformly over the parameter space. Furthermore, we obtain explicit convergence rates that link the computational cost of the PF algorithm, the kernel bandwidth and the dimension of the parameter space.

The rest of the paper is organised as follows. In Section 2 we state the Bayesian inference problem to be solved. The proposed nested particle filtering scheme is outlined in Section 3. Kernel density estimators, based on the output of the nested PF algorithm, are investigated in Section 4. Section 5 contains computer simulation results and, finally, Section 6 is devoted to the conclusions.

J. M. acknowledges the financial support of the Spanish Ministry of Economy and Competitiveness (project COMPREHENSION TEC2012-38883-C02-01) and the regional government of Madrid (program CASI-CAM-CM S2013/ICE-2845).

I. P. M. acknowledges the financial support of the Spanish Ministry of Economy and Competitiveness (projects TEC2012-38883-C02-01 and FIS2013-40653-P).

2. PROBLEM STATEMENT

2.1. Notation

We use upper-case letters to denote random variables (r.v.'s), e.g., X, and lower-case letters to indicate their realisations, e.g., $X = x$. Given a r.v. X with pdf $\alpha(x)$, the integral of a real function $f(x)$ with respect to (w.r.t.) the probability measure $\alpha(x)dx$ is denoted $(f,\alpha) \triangleq \int f(x)\alpha(x)dx$. We use r.v. independently of the dimension of X, i.e., X can be a random *vector*. The set of bounded real functions on the set $\mathcal{X}$ is denoted $B(\mathcal{X})$. If $f \in B(\mathcal{X})$, then $\|f\|_\infty \triangleq \sup_{x\in\mathcal{X}} |f(x)| < \infty$.

2.2. State space models

A state space model can be described in terms of two random sequences, $\{X_t\}_{t\geq 0}$ and $\{Y_t\}_{t\geq 1}$, and a r.v., Θ, taking values in the sets $\mathcal{X} \subseteq \mathbb{R}^{d_x}$, $\mathcal{Y} \subseteq \mathbb{R}^{d_y}$ and $\mathsf{S} \subset \mathbb{R}^{d_\theta}$, where d_x, d_y and d_θ are the corresponding dimensions of $\mathcal{X}$, $\mathcal{Y}$ and S, respectively. We refer to the sequence $\{X_t\}_{t\geq 0}$ as the state (or signal) process and we assume that it is a homogeneous Markov chain, with *a priori* pdf $\tau_0(x)$ and Markov transition kernel $\tau_\theta(x_t|x_{t-1})dx_t$, indexed by a realisation $\Theta = \theta$.

The sequence $\{Y_t\}_{t\geq 1}$ is the observation process. Each r.v. Y_t is assumed to be conditionally independent of all other observations given X_t and Θ, namely

$$\mathbb{P}_t\left\{Y_t \in A | x_{0:t}, \theta, \{y_k\}_{k\neq t}\right\} = \mathbb{P}_t\left\{Y_t \in A | x_t, \theta\right\}$$

for any Borel set A, where $\mathbb{P}_t$ denotes the probability measure for the triple $(\{X_n\}_{n\leq t}, \{Y_n\}_{n\leq t}, \Theta)$. We assume that the conditional pdf of Y_t given X_t, denoted $g_\theta(y_t|x_t) \geq 0$, can be evaluated up to a (possibly unknown) proportionality constant. Given $Y_t = y_t$, the function $g_\theta^{y_t}(x_t) \triangleq g_\theta(y_t|x_t)$ on the state space is the likelihood of X_t.

The *a priori* pdf of the parameter r.v. Θ is denoted $\mu_0(\theta)$ (note that X_0 and Θ are *a priori* independent). The set $\{\mu_0, \tau_0, \tau_\theta, \{g_\theta^{y_t}\}_{t\geq 1}\}$ describes a stochastic Markov state-space model in discrete time. Note that the model is indexed by the d_θ-dimensional random parameter $\Theta \in \mathsf{S}$.

2.3. Stochastic Lorenz 96 model

The two-scale Lorenz 96 model is a deterministic system of nonlinear differential equations that displays chaotic dynamics; see, e.g., [6]. The system dimension, i.e., the number of dynamic variables, can be scaled arbitrarily. A stochastic version of the model can be easily obtained by converting each differential equation into a stochastic differential equation driven by an independent and additive Wiener process. Here, for conciseness, we describe the difference equations that result from the application of the Euler-Maruyama integration method to a model with J slow variables, Z_j, $j = 0, \ldots, J-1$, and L fast variables per slow variable, $\bar{Z}_l$, $l = 0, ..., JL-1$. We then obtain

$$\begin{aligned} Z_{j,n} &= Z_{j,n-1} - \Delta Z_{j-1,n-1}(Z_{j-2,n-1} - Z_{j+1,n-1}) \\ &\quad + \Delta\left[\mathsf{f} - Z_{j,n-1} - \frac{\mathsf{hc}}{\mathsf{b}}\sum_{l=(j-1)L}^{Lj-1} \bar{Z}_{l,n-1}\right] \\ &\quad + \sqrt{\Delta}\sigma U_{j,n}, \qquad (1) \\ \bar{Z}_{l,n} &= \bar{Z}_{l,n-1} - \Delta\mathsf{cb}\bar{Z}_{l+1,n-1}(\bar{Z}_{l+2,n-1} - \bar{Z}_{l-1,n-1}) \\ &\quad + \Delta\left[\frac{\mathsf{cf}}{\mathsf{b}} - \mathsf{c}\bar{Z}_{l,n-1} + \frac{\mathsf{hc}}{\mathsf{b}} Z_{\lfloor\frac{l-1}{L}\rfloor,n-1}\right] + \bar{\sigma}\bar{U}_{l,n}, \end{aligned}$$

where $\Delta > 0$ denotes the discretisation period, n represents discrete time, f is a forcing parameter that controls the turbulence of the chaotic flow, c determines the time scale of the fast variables, h controls the strength of the coupling between the fast and slow variables, b determines the amplitude of the fast variables, $\{U_{j,n}, \bar{U}_{l,n}\}_{l,j,n\geq 0}$ are sequences of independent and identically distributed (i.i.d.) standard Gaussian r.v.'s, and $\sigma, \bar{\sigma} > 0$ are scale parameters.

We assume that observations can only be collected from this system once every n_o discrete time steps. Therefore, the observation process has the form

$$Y_t = [Z_{1,tn_o}, Z_{2,tn_o}, \ldots, Z_{J,tn_o}]^\top + V_t, \qquad (2)$$

where $t = 1, 2, ...$ and $\{V_t\}_{t\geq 1}$ is a sequence of i.i.d. r.v.'s with common pdf $\mathcal{N}(v_t; 0, \sigma_y^2 \mathbf{I}_J)$, which denotes a J-dimensional Gaussian density with 0 mean and covariance matrix $\sigma_y^2\mathbf{I}_J$, where $\mathbf{I}_J$ is the $J \times J$ identity matrix.

In computer experiments, system (1) is often employed to generate both ground-truth values for the slow variables $\{Z_{j,n}\}_{j,n\geq 0}$ and synthetic observations, $\{Y_t\}_{t\geq 1}$. As a forecast model for the slow variables it is common to use the difference equation [6]

$$\begin{aligned} Z_{j,n} &= Z_{j,n-1} - \Delta Z_{j-1,n-1}(Z_{j-2,n-1} - Z_{j+1,n-1}) \\ &+ \Delta\left[\mathsf{f} - Z_{j,n-1} - \ell(Z_{j,n-1}, \mathsf{a})\right] + \sqrt{\Delta}\sigma U_{j,n}, \quad (3) \end{aligned}$$

where $\mathsf{a} = [\mathsf{a}_1, \mathsf{a}_2]^\top \in \mathbb{R}^2$ is a (constant) parameter vector and function $\ell(Z_{j,n-1}, \mathsf{a}) \in \mathbb{R}$ is an ansatz for the coupling term $\frac{\mathsf{hc}}{\mathsf{b}}\sum_{l=(j-1)L}^{Lj-1} \bar{Z}_{l,n-1}$ in (1), to be specified later.

Equations (3) and (2) describe a state space model that can be expressed in terms of the general notation in Section 2.2. The state process at time n is $\tilde{X}_n = [Z_{0,n}, \ldots, Z_{J-1,n}]^\top$ and the transition pdf from time $n-1$ to time n is

$$\tilde{\tau}_\theta(\tilde{x}_n|\tilde{x}_{n-1}) = \mathcal{N}(\tilde{x}_n; \Psi(\tilde{x}_{n-1}, \theta), \sigma_x^2\mathbf{I}_J), \qquad (4)$$

where $\theta = [\mathsf{f}, \mathsf{a}^\top]^\top \in \mathbb{R}^3$, $\sigma_x^2 = \Delta\sigma^2$ and $\Psi(\tilde{x}_{n-1}, \theta) \in \mathbb{R}^J$ is the deterministic transformation that accounts for all the terms on the right hand side of (3) except the noise contribution $\sqrt{\Delta}\sigma U_{j,n}$. Since we only collect observations every $n_o\Delta$ continuous-time units, we need to put the dynamics of the states on the same time scale as the observation process

$$\tau_\theta(x_t|x_{t-1}) = \int \ldots \int \tilde{\tau}_\theta(x_t|\tilde{x}_{tn_o-1}) \prod_{i=1}^{n_o-2} \tilde{\tau}_\theta(\tilde{x}_{tn_o-i}|\tilde{x}_{tn_o-i-1})\tilde{\tau}_\theta(\tilde{x}_{(t-1)n_o+1}|x_{t-1}) \prod_{j=1}^{n_o-1} d\tilde{x}_{tn_o-j}. \quad (5)$$

$\{Y_t\}_{t\geq 1}$ in Eq. (2). If we define $X_t = \tilde{X}_{tn_o} \in \mathcal{X} = \mathbb{R}^J$ then the transition density from X_{t-1} to X_t follows readily from (4), as shown in Eq. (5) at the top of this page. While $\tau_\theta(x_t|x_{t-1})$ cannot be evaluated in closed form, it is straightforward to draw a sample from $X_t|x_{t-1}$ by simply running Eq. (3) n_o times, with starting point x_{t-1}. The likelihood function is

$$g^{y_t}(x_t) \propto \exp\left\{-\frac{1}{2\sigma_y^2}\sum_{r=1}^{J}(y_{r,t} - x_{i_r,t})^2\right\},$$

which follows from (2) and is independent of Θ for this particular model.

2.4. Problem statement

Assume that the parameters in (3) are random, with prior pdf μ_0. Then, the goal is to approximate sequence of conditional pdf's of the parameters $\Theta = [\mathsf{F}, \mathsf{A}_1, \mathsf{A}_2]^\top$ given the available observations at each time step t. We write $\mu_t(\theta)$ to denote the pdf of Θ conditional on $Y_{1:t} = y_{1:t}$. This pdf can be recursively decomposed as $\mu_t(\theta) \propto \lambda_{t,\theta}(y_t)\mu_{t-1}(\theta)$, where $\lambda_{t,\theta}(y_t)$ is the pdf of the r.v. Y_t conditional on $Y_{1:t-1} = y_{1:t-1}$ and $\Theta = \theta$. The latter density, in turn, can be written as an integral, namely $\lambda_{t,\theta}(y_t) = (g^{y_t}, \xi_{t,\theta})$, where $\xi_t(x_{t,\theta})$ is the predictive pdf of the state vector X_t conditional on the observations $Y_{1:t-1} = y_{1:t-1}$ and the parameter $\Theta = \theta$. It is a well known result [1, 2, 4] that the sequence of probability measures $\xi_{t,\theta}(x_t)dx_t$ can be recursively approximated using a standard PF algorithm, and hence the integral $(g^{y_t}, \xi_{t,\theta})$ can be numerically approximated as well.

In next section, we outline a novel PF algorithm that enables the recursive approximation of the pdf's μ_t, $t = 1, 2, ...$, and produces Bayesian estimates of the state variables X_t, $t = 1, 2, ...$, as a by-product.

3. NESTED PARTICLE FILTERING SCHEME

3.1. Standard particle filter

Assume $\Theta = \theta$ are given parameters. The standard particle filter is a recursive Monte Carlo algorithm for the approximation of the sequence of predictive probability measures $\xi_{t,\theta}(x_t)dx_t$ on the state space $\mathcal{X}$, as well as the associated *filtering* probability measures $\phi_{t,\theta}(x_t)dx_t$, where $\phi_{t,\theta}$ is the density of X_t conditional on given observations $Y_{1:t} = y_{1:t}$ and the parameter θ.

At time $t = 0$, we generate M samples (termed *particles*), $x_0^{(i)} \sim \tau_0$, $i = 1, ..., M$, from the prior density τ_0. At every time $t > 0$, we apply the algorithm below, where $\delta_{x'}(dx)$ denotes the unit delta measure located at x'.

Algorithm 1 *We take as inputs the parameter θ, the observation y_t and the particle approximation of $\phi_{t-1,\theta}(x_t)dx_t$ at time $t-1$, $\phi^M_{t-1,\theta}(dx_t) = \frac{1}{M}\sum_{i=1}^M \delta_{x_{t-1}^{(i)}}(dx_t)$.*

Computations:

- *Generate new particles $\bar{x}_t^{(i)} \sim \tau(x_t|x_{t-1}^{(i)})$ and compute normalised importance weights $w_t^{(i)} \propto g^{y_t}(\bar{x}_t^{(i)})$, $i = 1, ..., M$.*
- *Resample: for $i = 1, ..., M$, assign $x_t^{(i)} = \bar{x}_t^{(j)}$ with probability $w_t^{(j)}$, $j \in \{1, ..., M\}$.*

Outputs: *New particle approximations $\xi^M_{t,\theta}(dx_t) = \frac{1}{M}\sum_{i=1}^M \delta_{\bar{x}_t^{(i)}}(dx_t)$, $\phi^M_{t,\theta}(dx_t) = \frac{1}{M}\sum_{i=1}^M \delta_{x_t^{(i)}}(dx_t)$, and $\lambda^M_{t,\theta}(y_t) = (g^{y_t}, \xi^M_{t,\theta}) = \frac{1}{M}\sum_{i=1}^M g^{y_t}(\bar{x}_t^{(i)})$.*

Given any bounded function $f \in B(\mathcal{X})$ and any $p \geq 1$, it can be proved [7] under mild assumptions that

$$\|(f, \phi^M_{t,\theta}) - (f, \phi_{t,\theta})\|_p \leq \frac{c_{t,\theta}}{\sqrt{M}},$$

where $c_{t,\theta}$ is constant w.r.t. M. Similar convergence results hold for $(f, \xi^M_{t,\theta})$ and $\lambda^M_{t,\theta}(y_t)$.

3.2. Proposed algorithm

We outline a PF scheme for the recursive approximation of the sequence of probability measures $\mu_t(\theta)d\theta$, $t = 1, 2, ...$. See [4] for full details. It is a nested Monte Carlo scheme, where Algorithm 1 is used to compute importance weights for particles in the parameter space S. We assume that the latter is compact. In particular, we select $\mathsf{S} = [\mathsf{f}^-, \mathsf{f}^+] \times [\mathsf{a}^-, \mathsf{a}^+] \times [\mathsf{a}^-, \mathsf{a}^+] \subset \mathbb{R}^3$ for some known and finite bounds $\mathsf{f}^- < \mathsf{f}^+$ and $\mathsf{a}^- < \mathsf{a}^+$.

Algorithm 2 *At time $t = 0$, draw N i.i.d. particles $\theta_0^{(i)}$ from $\mu_0(\theta)$ and NM i.i.d. particles $x_0^{(i,j)}$ from $\tau_0(x_0)$, $i = 1, ..., N$ and $j = 1, ..., M$. Let $\mu_0^N = \frac{1}{N}\sum_{i=1}^N \delta_{\theta_0^{(i)}}$ and $\phi^M_{0,\theta_0^{(i)}} = \frac{1}{M}\sum_{j=1}^M \delta_{x_0^{(i,j)}}$. Choose a conditional pdf $\kappa_N(\theta|\theta')$ on the parameter space S that satisfies, for at least some $p \geq 1$,*

$$\sup_{\theta' \in \mathsf{S}} \int \|\theta - \theta'\|^p \kappa(\theta|\theta')d\theta \leq \frac{c_\kappa^p}{N^{\frac{p}{2}}}. \quad (6)$$

where c_κ is some constant independent of N.

Computations: *Given $\mu_{t-1}^N = \frac{1}{N}\sum_{i=1}^N \delta_{\theta_{t-1}^{(i)}}$, the N sets $\chi_{t-1}^{i,M} = \{x_{t-1}^{(i,j)}\}_{j=1}^M$ and the new observation y_t, take the following steps at time t.*

- *Draw N new particles $\bar{\theta}_t^{(i)} \sim \kappa_N(\theta|\theta_{t-1}^{(i)})$, $i = 1, ..., N$.*
- *Run Algorithm 1 N times, with inputs $\Theta = \bar{\theta}_t^{(i)}$ and $\chi_{t-1}^{i,M} = \{x_{t-1}^{(i,j)}\}_{j=1}^M$, to obtain*

$$\text{normalised importance weights:} \quad \lambda_t^{(i)} \propto \lambda_{t,\bar{\theta}_t^{(i)}}^M(y_t),$$
$$\text{updated particles in } \mathcal{X}: \quad \bar{\chi}_t^{i,M} = \{x_t^{(i,j)}\}_{j=1}^M,$$

for $i = 1, ..., N$.
- *Resample: for $i = 1, ..., N$, assign $\theta_t^{(i)} = \bar{\theta}_t^{(l)}$ and $\chi_t^{i,M} = \bar{\chi}_t^{l,M}$ with probability $\lambda_t^{(l)}$, $l \in \{1, ..., N\}$.*

Outputs: *The approximations $\mu_t^N(d\theta) = \frac{1}{N}\sum_{i=1}^N \delta_{\theta_t^{(i)}}(d\theta)$ and sets $\chi_t^{i,M} = \{x_t^{(i,j)}\}_{j=1}^M$, for $i = 1, ..., N$.*

Given μ_t^N and the collection of sets $\chi_t^{i,M}$, $i = 1, ..., N$, it is straightforward to obtain posterior estimates of the parameters and the dynamic variables, e.g.,

$$E[\Theta|y_{1:t}] \approx \frac{1}{N}\sum \theta_t^{(i)}, \quad E[X_t|y_{1:t}] \approx \frac{1}{NM}\sum_{i=1,j=1}^{N,M} x_t^{(i,j)}.$$

The inequality (6) simply states that the pdf $\kappa_N(\theta|\theta')$ should have a sufficiently small variance (in all directions). A simple truncated Gaussian kernel with suitable variance, e.g.,

$$\kappa_N(\theta|\theta') \propto I_{\mathsf{S}}(\theta)\mathcal{N}\left(\theta; \theta', \frac{c}{N}\mathbf{I}_3\right),$$

where $I_{\mathsf{S}}(\theta) = 1$ if $\theta \in \mathsf{S}$ and 0 otherwise, suffices to make (6) hold for all $p \geq 1$ (and guarantee convergence [4]).

4. KERNEL DENSITY ESTIMATORS

We aim at approximating the posterior pdf of the parameters F and $\mathsf{A} \in \mathbb{R}^2$. If $\Theta = [\mathsf{F}, \mathsf{A}^\top]^\top$, then we denote $\mu_t(\theta) = \mu_t(\mathsf{f}, \mathsf{a})$ and calculate the posterior marginal densities as

$$\mu_{t,\mathsf{F}}(\mathsf{f}) = \int \mu_t(\mathsf{f}, \mathsf{a})d\mathsf{a}, \quad \mu_{t,\mathsf{A}}(\mathsf{a}) = \int \mu_t(\mathsf{f}, \mathsf{a})d\mathsf{f}.$$

Similary, the particle approximation $\mu_t^N(d\theta) = \mu_t^N(d\mathsf{f} \times d\mathsf{a})$ yields the approximate marginal probability measures

$$\mu_{t,\mathsf{F}}^N(d\mathsf{f}) = \int_{[\mathsf{a}^-,\mathsf{a}^+]^2} \mu_t^N(d\mathsf{f} \times d\mathsf{A}) = \frac{1}{N}\sum_{i=1}^N \delta_{\mathsf{f}_t^{(i)}}(d\mathsf{f}),$$
$$\mu_{t,\mathsf{A}}^N(d\mathsf{a}) = \int_{[\mathsf{f}^-,\mathsf{f}^+]} \mu_t^N(d\mathsf{f} \times d\mathsf{A}) = \frac{1}{N}\sum_{i=1}^N \delta_{\mathsf{a}_t^{(i)}}(d\mathsf{a}),$$

where we have used the obvious notation $\theta_t^{(i)} = [\mathsf{f}_t^{(i)}, \mathsf{a}_t^{(i)\top}]^\top$.

Let $\phi : \mathbb{R} \to \mathbb{R}$ and $\psi : \mathbb{R}^2 \to \mathbb{R}$ be continuous and bounded pdf's, both with finite second order moments, and, for any positive bandwidths $h, \tilde{h} > 0$, let

$$\phi_h^u(u') = h^{-1}\phi\left(\frac{u - u'}{h}\right) \text{ and } \psi_{\tilde{h}}^v(v') = \tilde{h}^{-2}\psi\left(\frac{v - v'}{\tilde{h}}\right).$$

Then, we build estimators of $\mu_{t,\mathsf{F}}(\mathsf{f})$ and $\mu_{t,\mathsf{A}}(\mathsf{a})$ as [7]

$$\hat{\mu}_{t,\mathsf{F}}^N(\mathsf{f}) = (\phi_h^{\mathsf{f}}, \mu_{t,\mathsf{f}}^N) = \frac{1}{N}\sum_{i=1}^N \phi_h(\mathsf{f} - \mathsf{f}_t^{(i)}),$$
$$\hat{\mu}_{t,\mathsf{A}}^N(\mathsf{a}) = (\psi_{\tilde{h}}^{\mathsf{a}}, \mu_{t,\mathsf{A}}^N) = \frac{1}{N}\sum_{i=1}^N \psi_{\tilde{h}}(\mathsf{a} - \mathsf{a}_t^{(i)}).$$

Both estimators converge uniformly over the support sets of the r.v.'s F and A, provided that the jittering kernel κ_N in Algorithm 2 and the bandwidths h and $\tilde{h}$ are suitably chosen. This is formally given below.

Theorem 1 *Choose κ_N such that the inequality* (6) *holds for every $p \geq 1$ and select $h \geq N^{-\frac{1}{6}}$ and $\tilde{h} \geq N^{-\frac{1}{10}}$. If $\mu_0(\theta)$ is Lipschitz, then there exist almost surely (a.s.) finite r.v.'s U^ϵ and W^ϵ such that*

$$\sup_{\mathsf{f}\in[\mathsf{f}^-,\mathsf{f}^+]} |\hat{\mu}_{t,\mathsf{F}}^N(\mathsf{f}) - \mu_{t,\mathsf{F}}(\mathsf{f})| \leq N^{-\frac{1-\epsilon}{6}}U^\epsilon, \tag{7}$$
$$\sup_{\mathsf{a}\in[\mathsf{a}^-,\mathsf{a}^+]^2} |\hat{\mu}_{t,\mathsf{A}}^N(\mathsf{a}) - \mu_{t,\mathsf{A}}(\mathsf{a})| \leq N^{-\frac{1-\epsilon}{10}}W^\epsilon, \tag{8}$$

for any, arbitrarily small, $\epsilon > 0$. In particular,

$$\lim_{N\to\infty} |\hat{\mu}_{t,\mathsf{F}}^N(\mathsf{f}) - \mu_{t,\mathsf{F}}(\mathsf{f})| = \lim_{N\to\infty} |\hat{\mu}_{t,\mathsf{A}}^N(\mathsf{a}) - \mu_{t,\mathsf{A}}(\mathsf{a})| = 0$$

a.s. and uniformly over the parameter support.

Proof. A full proof is too long to be given here, hence we just sketch the argument. The key is to prove that the optimal filters $\phi_{t,\theta}(x_t)$, $t \geq 0$, are Lipschitz functions of θ. It is straightforward to prove that $\tau_\theta(x_t|x_{t-1})$ is Lipschitz. Since the likelihood function is, in our case, independent of θ, i.e., $g_\theta^{y_t} = g^{y_t}$, then it can be shown, by an induction argument, that both $\xi_{t,\theta}$ and $\phi_{t,\theta}$ are Lipschitz for every t.

Since the pdf $\xi_{t,\theta}$ is a Lipschitz function of θ, then

- it can be shown, again by induction, that $\mu_t(\theta) \propto \mu_{t-1}(\theta)\int g^{y_t}(x_t)\xi_{t,\theta}(x_t)dx_t$ is Lipschitz and
- it is possible to apply Theorem 2 in [4] to show that

$$\|(q, \mu_t^N) - (q, \mu_t)\|_p \leq \frac{c_t(q)}{\sqrt{N}} \tag{9}$$

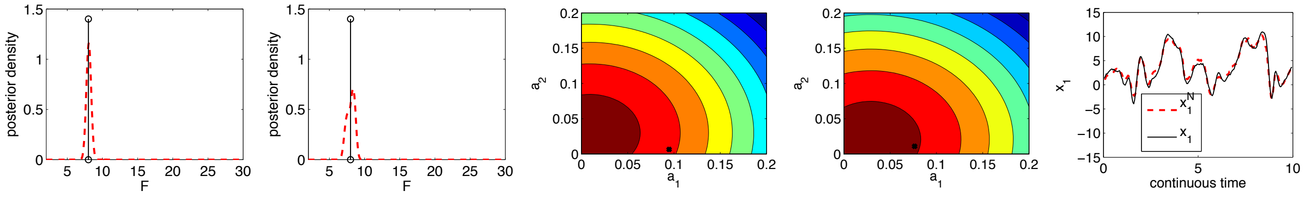
for any $q \in B(\mathsf{S})$ and a constant $c_t(q)$ independent of N.

The inequality (9) establishes the consistency of Algorithm 2 and, together with the Lipschitz continuity of $\mu_t(\theta)$, enables us to apply Theorem 4.2 in [7] on the convergence of kernel density estimators, that yields the inequalities (7) and (8). ■

5. NUMERICAL RESULTS

We have run the two-scale Lorenz 96 model in (1), with parameters $J = 16$, $L = 160$,

$$\{\mathsf{f}, \mathsf{c}, \mathsf{b}, \mathsf{h}, \Delta, \sigma^2, \bar{\sigma}^2\} = \left\{8, 10, 15, 0.75, 2 \times 10^{-4}, \frac{1}{\sqrt{2}}, \frac{1}{\sqrt{8}}\right\},$$



(a) (b) (c) (d) (e)

Fig. 1: (a)-(b) Estimates of $\mu_{t,\mathsf{F}}(\mathsf{f})$ over the interval $[2, 30]$. (c)-(d) Estimates of $\mu_{t,\mathsf{A}}(\mathsf{a})$ over the set $[0, 0.2]^2$. All pdf estimates have been computed using a Gaussian kernel ϕ. (e) Posterior-mean estimates of $X_{1,t}$ (axis in continuous time units).

prior distributions $Z_j \sim U(0,1)$ and $\bar{Z}_l \sim U(-1/2\mathsf{cb}, +1/2\mathsf{cb})$, $j = 1, ..., J$ and $l = 1, ..., L$, to generate synthetic observations of the form

$$Y_t = [Z_{1,tn_o}, \ldots, Z_{J,tn_o}]^\top + V_t,$$

where $n_o = 250$ (i.e., one observation vector is collected every 250 discrete time steps) and $V_t \sim \mathcal{N}(v_t; 0, 4\mathbf{I}_J)$. The simulations are run for 10 continuous time units, which amount to 5×10^4 discrete time steps.

For Algorithm 2, the prior pdf's on the state and the parameter spaces are both uniform, namely, $\tau_0(x_t) = \prod_{j=1}^J U(0,1)$ and $\mu_0(\mathsf{f}, \mathsf{a}) = U(\mathsf{f}^-, \mathsf{f}^+) \times U([\mathsf{a}^-, \mathsf{a}^+]^2)$, with $\mathsf{f}^- = 2$, $\mathsf{f}^+ = 30$, $\mathsf{a}^- = 0$ and $\mathsf{a}^+ = 0.2$. We have carried out 20 independent simulations with $N = 200$ particles in the outer filter and $M = 600$ particles in the inner filters used to compute the importance weights of the outer filter. The jittering kernel is Gaussian, with the form

$$\kappa_N(\mathsf{f}, \mathsf{a}) = \mathcal{N}(\mathsf{f}; 0, 20 \times N^{-\frac{3}{2}}) \times \mathcal{N}(\mathsf{a}; 0, 0.04 \times N^{-\frac{3}{2}}\mathbf{I}_2).$$

The values of the parameters a in the ansatz functions $\ell(Z_{j,t}, \mathsf{a})$ cannot be known exactly. For reference, we approximate them in each simulation run using the least squares estimator $\hat{\mathsf{a}}_{LS} = \arg\min_{\mathsf{a} \in \mathbb{R}^2} \sum_{j,n} \left(\ell(Z_{j,n-1}, \mathsf{a}) - \frac{\mathsf{hc}}{\mathsf{b}} \sum_{l=(j-1)L}^{Lj-1} \bar{Z}_{l,n-1} \right)^2$ where $\ell(Z, \mathsf{a}) = Z(\mathsf{a}_1 + \mathsf{a}_2 Z)$. This genie-aided estimator is only used for numerical comparison, not for running Algorithm 2.

Figures 1a and 1b display two sample estimates of the pdf $\mu_{t,\mathsf{F}}(\mathsf{f})$ for two independent simulations, with $t = 200$. These are the best (a) and worst (b) outcomes of the 20 simulation experiments ($\mathsf{f} = 8$ is indicated with a vertical line).

Two sample realisations of the kernel estimates $\hat{\mu}_{t,\mathsf{A}}^N(\mathsf{a})$ are displayed in Figs. 1c and 1d. The LS estimate $\hat{\mathsf{a}}_{LS}$ is indicated by a black cross in each case. It lies within the highest density region in the example of Fig. 1d, and a bit shifted in the example of Fig. 1c. Recall that a_1 and a_2 are the parameters of a simple ansatz for the contribution of the fast variables, so it is not surprising that the resulting pdf's are flatter than in the case of $\hat{\mu}_{t,\mathsf{F}}(\mathsf{f})$.

Finally, Fig. 1e shows the posterior-mean estimates of the first dynamic variable, $X_{1,t} = Z_{1,tn_o}$, in a single simulation run. We observed how the actual value is closely tracked. This is the case for the remaining dynamic variables, $X_{2:J,t}$ (not shown). The mean square error of the dynamic variable estimates over the 20 independent simulations (normalised w.r.t. the power of the signals) was ≈ 0.0313.

6. CONCLUSIONS

We have proposed a nested PF scheme to jointly track the dynamic variables and approximate the posterior pdf of the fixed parameters in state space models. We have proved a.s. convergence of the pdf approximations and displayed numerical results for a stochastic two-scale Lorenz 96 system.

REFERENCES

[1] C. Andrieu, A. Doucet, and R. Holenstein, “Particle Markov chain Monte Carlo methods,” *Journal of the Royal Statistical Society B*, vol. 72, pp. 269–342, 2010.

[2] N. Chopin, P. E. Jacob, and O. Papaspiliopoulos, “SMC2: an efficient algorithm for sequential analysis of state space models,” *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 2012.

[3] C. Andrieu, A. Doucet, S. S. Singh, and V. B. Tadić, “Particle methods for change detection, system identification and control,” *Proceedings of the IEEE*, vol. 92, no. 3, pp. 423–438, March 2004.

[4] D. Crisan and J. Miguez, “Nested particle filters for online parameter estimation in discrete-time state-space markov models,” *arXiv: 1308.1883v1 [stat.CO]*, 2013.

[5] P. J. Van Leeuwen, “Nonlinear data assimilation in geosciences: an extremely efficient particle filter,” *Quarterly Journal of the Royal Meteorological Society*, vol. 136, no. 653, pp. 1991–1999, 2010.

[6] J. Hakkarainen, A. Ilin, A. Solonen, M. Laine, H. Haario, J. Tamminen, E. Oja, and H. Järvinen, “On closure parameter estimation in chaotic systems,” *Nonlinear Proc. Geoph.*, vol. 19, no. 1, pp. 127–143, 2012.

[7] D. Crisan and J. Miguez, “Particle-kernel estimation of the filter density in state-space models,” *Bernoulli*, vol. 20, no. 4, pp. 1879–1929, 2014.