Published in final edited form as:

Stat Comput. 2017 November ; 27(6): 1677–1692. doi:10.1007/s11222-016-9711-9.

A second-order iterated smoothing algorithm

Dao Nguyen¹ and Edward L. Ionides¹

¹ Department of Statistics, University of Michigan, Ann Arbor, MI, USA

Abstract

Simulation-based inference for partially observed stochastic dynamic models is currently receiving much attention due to the fact that direct computation of the likelihood is not possible in many practical situations. Iterated filtering methodologies enable maximization of the likelihood function using simulation-based sequential Monte Carlo filters. Doucet et al. (2013) developed an approximation for the first and second derivatives of the log likelihood via simulation-based sequential Monte Carlo smoothing and proved that the approximation has some attractive theoretical properties. We investigated an iterated smoothing algorithm carrying out likelihood maximization using these derivative approximations. Further, we developed a new iterated smoothing algorithm, using a modification of these derivative estimates, for which we establish both theoretical results and effective practical performance. On benchmark computational challenges, this method beat the first-order iterated filtering algorithm. The method's performance was comparable to a recently developed iterated filtering algorithm based on an iterated Bayes map. Our iterated smoothing algorithm and its theoretical justification provide new directions for future developments in simulation-based inference for latent variable models such as partially observed Markov process models.

Keywords

Iterated smoothing; Sequential Monte Carlo; State space model; Hidden Markov model; Parameter estimation

1 Introduction

During the past three decades, partially observed Markov process (POMP) models (also known as state space models or hidden Markov models) have become ubiquitous tools for modeling and time series data analysis of time series data in many disciplines, including econometrics, ecology, and engineering. However, it can be difficult to make inferences about non-linear or non-Gaussian POMP models owing to the fact that there is no closed form expression for the likelihood function. Linear Gaussian models enable exact likelihood computation, via the Kalman filter, but can lead to unsatisfactory results when the assumptions are violated. In many situations, the transition probability density is intractable

Edward L. Ionides, ionides@umich.edu.
Dao Nguyen, nguyendx@umich.edu

Electronic supplementary material The online version of this article (doi:10.1007/s11222-016-9711-9) contains supplementary material, which is available to authorized users.

or too expensive to evaluate, but easy to sample from (Bretó et al. 2009). Therefore, there has been a surge of interest in simulation-based inference for POMP models (Ionides et al. 2006; Toni et al. 2009; Andrieu et al. 2010; Wood 2010; Chopin et al. 2013; Ionides et al. 2015). Simulation-based methods have also been called plug-and-play (Bretó et al. 2009; He et al. 2010), likelihood-free (Sisson et al. 2007; Yıldırım et al. 2015) or equation-free (Kevrekidis et al. 2004). These methodologies can be categorized into either Bayesian or frequentist approaches and further categorized into full information or partial information approaches. Full information approaches are those which are based on the full likelihood of the data; partial information approaches are those based on summary statistics or quasi-likelihoods, such as approximate Bayesian computing (Toni et al. 2009) or synthetic likelihood (Wood 2010). Here, we are concerned with full information, frequentist, simulation-based inference. The first algorithm developed to carry out such inference was the iterated filtering algorithm of Ionides et al. (2006), which we will call IF1. The theoretical properties of IF1 were studied by Ionides et al. (2011). Doucet et al. (2013) showed that sequential Monte Carlo smoothing can give derivative estimates of the log likelihood that have better asymptotic convergence rates than those used for IF1. One can apply the derivative approximations of Doucet et al. (2013) in a Newton–Raphson type optimization algorithm, and we call this algorithm IS1. However, in practical problems, IS1 has failed to show practical performance living up to its favorable asymptotic theory. This paper develops a modification of the theory of Doucet et al. (2013) giving rise to a new algorithm that we call IS2, which empirically shows clearly enhanced performance over IF1 and IS1 on our benchmarks in Sect. 5. Recently, a new iterated filtering algorithm, which we call IF2, has been developed with a different theoretical justification based on iterated perturbed Bayes maps (Ionides et al. 2015). IS2 shows comparable performance to IF2 on our benchmarks. The substantial differences—both in the theoretical foundations and the resulting algorithms—between IF2 and IS2 indicate that IS2 provides a promising alternative approach to IF2 for future theoretical and methodological developments.

The key contributions of this paper are threefold. First, we demonstrate theoretically that random walk parameter perturbations can be used in place of the white noise perturbations of IS1 while preserving much of the theoretical support provided by Doucet et al. (2013). In particular, IS2 inherits second-order convergence properties (in the sense of using first and second derivative estimates; Spall 2003) from IS1. Second, we discover that the approximation of the observed information matrix using random walk noise is simpler than that using independent white noise. Consequently, IS2 enjoys a computationally cheap estimate of the observed information matrix. Third, IS2 is not only attractive in theory, but we show it also has good numerical performance in practice.

The paper is organized as follows. In Sect. 2, we introduce some notation and discuss some background to the computational challenge we investigate. In Sect. 3, we develop the required theory in the context of latent variable models, which are later extended to the case of partially observed stochastic dynamic systems. In Sect. 4, we state our theorems and present the IS2 algorithm, postponing proofs of our results to the appendix. Section 5 presents a toy problem and a challenging inference problem of fitting a malaria transmission model to time series data, showing empirical results in which IS2 beats IF1 and IS1 while performing comparably to IF2. Section 6 is a concluding discussion.

2 Notation and problem definition

We use capital letters to denote random variables and lower case letters to denote their values. Let $\{X(t), t \in \mathcal{T}\}$ be a Markov process with $X(t)$ taking values in a measurable space $\mathcal{X}$. The time index set, $\mathcal{T}$, may be an interval or a discrete set and contains a finite subset $t_1 < t_2 < \dots < t_N$ at which $X(t)$ is observed, together with some initial time $t_0 < t_1$. We write $X_{0:N} = (X_0, \dots, X_N) = (X(t_0), \dots, X(t_N))$. Hereafter for any generic sequence $\{X_n\}$, we shall use $X_{i:j}$ to denote $(X_i, X_{i+1}, \dots, X_j)$. The distribution of $X_{0:N}$ is characterized by the initial density $X_0 \sim \mu(x_0; \theta)$ and the condition density of X_n given X_{n-1} , written as $f_n(x_n | x_{n-1}; \theta)$ for $1 \leq n \leq N$. Here, θ is an unknown parameter in d . The process $\{X_n\}$ is only observed through another process $\{Y_n, n = 1, \dots, N\}$ taking values in a measurable space $\mathcal{Y}$. The observations are assumed to be conditionally independent given $\{X_n\}$, and their probability density is of the form

$$p_{Y_n | Y_{1:n-1}, X_{0:n}}(y_n | y_{1:n-1}, x_{0:n}; \theta) = g_n(y_n | x_n; \theta),$$

for $1 \leq n \leq N$. We assume that $X_{0:N}$ and $Y_{1:N}$ have a joint density $p_{X_{0:N}, Y_{1:N}}(x_{0:N}, y_{1:N}; \theta)$ on $\mathcal{X}^{N+1} \times \mathcal{Y}^N$. The data are a sequence of observations by $y_{1:N}^* = (y_1^*, \dots, y_N^*) \in \mathcal{Y}^N$, considered as fixed. We write the long likelihood function of the data for the POMP model as $\ell(\theta)$, given by

$$\begin{aligned} \ell(\theta) &= \log p_{Y_{1:N}}(y_{1:N}^*; \theta) \\ &= \log \int \mu(x_0; \theta) \prod_{n=1}^N f_n(x_n | x_{n-1}; \theta) \times g_n(y_n^* | x_n; \theta) dx_{0:N}. \end{aligned}$$

We seek the maximum likelihood estimator, $\hat{\theta} = \arg \max \ell(\theta)$. Maximization of the likelihood function using first-order stochastic approximation (Kushner and Clark 1978) involves a Monte Carlo approximation to a difference equation,

$$\theta_m = \theta_{m-1} + \gamma_m \nabla \ell(\theta_{m-1}),$$

where $\theta_0 \in \Theta$ is an arbitrary initial estimate and $\{\gamma_m\}_{m=1}^\infty$ is a sequence of step sizes with $\sum_{m=1}^\infty \gamma_m = \infty$ and $\sum_{m=1}^\infty \gamma_m^2 < \infty$. Under appropriate regularity conditions, the algorithm converges to a local maximum of $\ell(\theta)$. The term $\nabla \ell(\theta)$ is shorthand for the d -valued vector of partial derivatives,

$$\nabla \ell(\theta) = \frac{\partial \ell(\theta)}{\partial \theta},$$

which is also called the score vector. Stochastic approximation methods can sometimes be improved by exploiting the observed information matrix in a Newton–Raphson approach (Spall 2003). In these second-order methods, the convergence rate is improved by using

$-\{\nabla^2 \ell(\theta)\}^{-1}$ in place of the step size γ_m , where $\nabla^2 \ell(\theta)$ is the $d \times d$ Hessian matrix of $\ell(\theta)$, written with some abuse of notation as

$$\nabla^2 \ell(\theta) = \nabla \nabla' \ell(\theta) = \frac{\partial^2 \ell(\theta)}{\partial \theta^2},$$

where ∇' is a row vector of partial derivatives. The matrix $\nabla^2 \ell(\theta)$ is also known as the observed information at θ . Carrying out a Newton–Raphson approach via a simulation-based algorithm boils down to simulation-based estimation of the score vector and observed information matrix.

Sequential Monte Carlo (SMC) approaches have previously been developed to estimate the score and observed information (Poyiadjis et al. 2011; Nemeth et al. 2013; Dahlin et al. 2015). However, these methods require the ability to evaluate transition densities and sometimes also their derivatives, and so do not have the plug-and-play property of Bretó et al. (2009). One alternative plug-and-play approach which also approximates the score and observed information is the finite difference method. However, this involves carrying out multiple independent filtering operations for a single Monte Carlo gradient estimate, causing significant computational burden in many practical situations. Moreover, it can also result in high variance estimates, which unfortunately are unsolved for plug-and-play setup by current variance reduction techniques in the literature (Doucet et al. 2013). As a plug-and-play alternative, Doucet et al. (2013) used an artificial dynamics approach to estimate the observed information matrix using sequential Monte Carlo smoothing. The approach of Doucet et al. (2013) can be computationally intensive, reducing its practical advantage over the first-order method of Ionides et al. (2011). We propose a computationally less demanding approximation to the score and observed information. Theoretical properties of these approximations are shown in Theorems 3 and 4, and more numerically stable approximations of these quantities are investigated in Theorems 5 and 6. Following the approach of Ionides et al. (2011) and Doucet et al. (2013), we first develop our theory (in Sect. 3) in the context of a latent variable model. Then, in Sect. 4, we extend this to the POMP framework.

3 Perturbed parameters and a latent variable model

Consider a parametric model consisting of a density $p_Y(y; \theta)$ with the log likelihood of the data $y^* \in \mathcal{Y}$ given by $\ell(\theta) = \log p_Y(y^*; \theta)$. We define a stochastically perturbed model corresponding to a pair of random variables $(\tilde{\theta}, \tilde{Y})$ having a joint probability density on $d \times \mathcal{Y}$ given by

$$p_{\tilde{\theta}, \tilde{Y}}(\tilde{\vartheta}, y; \theta, \tau) = \tau^{-d} \kappa \left\{ \tau^{-1} (\tilde{\vartheta} - \theta) \right\} p_Y(y; \tilde{\vartheta}).$$

Using a Taylor expansion up to the second order, Ionides et al. (2011) approximated the score function $\nabla \ell(\theta)$ in terms of moments of the conditional distribution of $\tilde{\Theta}$ given $Y = y^*$. Doucet et al. (2013) developed a Taylor expansion to the fourth order and approximated both the score function $\nabla \ell(\theta)$ and the observed information matrix $\nabla^2 \ell(\theta)$. The following results for stochastically perturbed models are restated from Doucet et al. (2013), since they are foundations for our proofs. We denote $|\cdot|$ the L_1 -norm. For any vector $u \in \mathbb{R}^d$, $|u| = \sum_{i=1}^d |u_i|$ and for any matrix $v \in \mathbb{R}^{d \times d}$, $|v| = \sum_{i=1}^d \sum_{j=1}^d |v_{ij}|$. We suppose the following regularity conditions, identical to the assumptions of Doucet et al. (2013).

Assumption 1

There exists $C_1 < \infty$ such that for any integer $k = 1, 1 \leq i_1, \dots, i_k \leq d$ and $\beta_1, \dots, \beta_k = 1$,

$$\int |u_{i_1}^{\beta_1} u_{i_2}^{\beta_2} \dots u_{i_k}^{\beta_k}| \kappa(u) \, du \leq C_1,$$

where κ is a symmetric probability density on $\mathbb{R}^d$ with respect to Lebesgue measure and $\Sigma = (\sigma_{i,j})_{i,j=1}^d$ is the non-singular covariance matrix associated to κ .

Assumption 2

There exist $\gamma, \delta, M > 0$, such that for all $u \in \mathbb{R}^d$,

$$|u| > M \Rightarrow \kappa(u) < e^{-\gamma|u|^\delta}.$$

Assumption 3

ℓ is four times continuously differentiable and δ defined as in Assumption 2. For all $\theta \in \mathbb{R}^d$, there exists $0 < \eta_1 < \delta, \epsilon, D > 0$, such that for all $u \in \mathbb{R}^d$,

$$\mathcal{L}(\theta + u) \leq D e^{\epsilon|u|^{\eta_1}},$$

where $\mathcal{L}: \mathbb{R}^d \rightarrow \mathbb{R}$ is the associated likelihood function $\mathcal{L} = \exp \ell$.

Assumption 4

κ is mesokurtic, meaning that $\int u_i^4 \kappa(u) \, du = 3\sigma_i^4$.

In practice, κ can be chosen by the users and it is often convenient to chose a multivariate normal distribution, so that Assumptions 1, 2, and 4 are satisfied explicitly. In order to prove the approximation of observed information matrix, Doucet et al. (2013) further assumed regularity of the perturbation kernel. Specifically, a non-singular symmetric kernel is assumed, which is consistent with the practical choice of Gaussian perturbations. Parameters with a bounded domain can be transformed so that they are defined on the entire real line, to enable the applicability of multivariate normal perturbations.

Theorem 1

(Doucet et al. 2013) *Suppose Assumption 1, 2, and 3. There exists an η_2 and a constant,*

$$C_2 = C_2(C_1, \gamma, \delta, M, \eta_1, \epsilon, D),$$

such that for every $0 < \tau < \eta_2$,

$$\left| \left(\check{\Theta} - \theta \mid \check{Y} = y^* \right) - \tau^2 \Sigma \nabla \ell(\theta) \right| < C_2 \tau^4. \quad (1)$$

Proof—The theorem statement differs from the statement of Theorem 2 of Doucet et al. (2013) only in that we have explicitly noted the dependence of the constant C_2 on C_1 , γ , δ , M , η_1 , ϵ , and D . This dependence is a direct consequence of the proof provided by Doucet et al. (2013).

We also state a corresponding version of Theorem 3 of Doucet et al. (2013):

Theorem 2

(Doucet et al. 2013) *Suppose Assumption 1, 2, 3, and 4, there exist an η_3 and a constant,*

$$C_3 = C_3(C_1, \gamma, \delta, M, \eta_1, \epsilon, D),$$

such that for every $0 < \tau < \eta_3$,

$$\left| \left[\left(\check{\Theta} - \theta \right) \left(\check{\Theta} - \theta \right)^T \mid \check{Y} = y^* \right] - \tau^2 \Sigma - \tau^4 \Sigma \left(\nabla^2 \ell(\theta) \right) \Sigma \right| < C_3 \tau^6. \quad (2)$$

These approximations are useful for latent variable models, where the log likelihood of the model consists of marginalizing over a latent variable, X ,

$$\ell(\theta) = \log \int p_{X,Y}(x, y^*; \theta) dx.$$

In this case, the expectations in Lemmas 1 and 2 can be approximated by Monte Carlo importance sampling, as proposed by Ionides et al. (2011) and Doucet et al. (2013). Results such as Lemmas 1 and 2 have potential applicability to situations other than POMP models, as discussed by Doucet et al. (2013) and Ionides et al. (2011, 2015). Here, our focus is on POMP models, which is the model class for which these methods have primarily been used. The latent variable setup we consider is identical to that of Doucet et al. (2013), which is also similar to that of Ionides et al. (2011). The three approaches become more distinct in their consequences for the extension from latent variable models to POMP models.

4 An iterated smoothing algorithm

The POMP model is a specific latent variable model with $X = X_{0:N}$ and $Y = Y_{1:N}$. We define a perturbed POMP model having a similar construction to our perturbed latent variable

model with $\check{X} = \check{X}_{0:N}$, $\check{Y} = \check{Y}_{1:N}$, and $\check{\Theta} = \check{\Theta}_{0:N}$. Ionides et al. (2011) perturbed the parameters by setting $\check{\Theta}_{0:N}$ to be a random walk starting at θ , whereas Doucet et al. (2013) took $\check{\Theta}_{0:N}$ to be independent additive white noise perturbations of θ . Our goal is to take advantage of the asymptotic developments of Doucet et al. (2013) while maintaining some practical advantages of random walk perturbations for finite computations. Specifically, we construct $\check{\Theta}_{0:N}$ as follows.

Let $Z_0, \dots, Z_N$ be $N+1$ independent draws from a density ψ . We introduce $N+2$ perturbation parameters, τ and $\tau_0, \dots, \tau_N$, and construct a process $\check{\Theta}_{0:N}$ by setting

$$\check{\Theta}_n = \theta + \tau \sum_{i=0}^n \tau_i Z_i \quad (3)$$

for $0 \leq n \leq N$. We will later consider a limit where $\tau_0:N$ as fixed and the scale factor τ decreases toward zero, and subsequently another limit where τ_0 is fixed but $\tau_{1:N}$ decrease

toward zero together with τ . Let $p_{\check{\Theta}_{0:N}}(\check{\vartheta}_{0:N}; \theta, \tau, \tau_{0:N})$ be the probability density of $\check{\Theta}_{0:N}$.

We define the artificial random variables $\check{\Theta}_{0:N}$ via their density,

$$p_{\check{\Theta}_{0:N}}(\check{\vartheta}_{0:N}; \theta, \tau, \tau_{0:N}) = (\tau \tau_0)^{-d} \psi \left\{ (\tau \tau_0)^{-1} (\check{\vartheta}_0 - \theta) \right\} \times \prod_{n=1}^N (\tau \tau_n)^{-d} \psi \left\{ (\tau \tau_n)^{-1} (\check{\vartheta}_1 - \check{\vartheta}_{n-1}) \right\}.$$

We define the stochastically perturbed model with a Markov process $\{(\check{X}_n, \check{\Theta}_n), 0 \leq n \leq N\}$, observation process $\check{Y}_{1:N}$ and parameter $(\theta, \tau, \tau_{0:N})$ by the factorization of their joint probability density

$$p_{X_{0:N}, Y_{1:N}, \check{\Theta}_{0:N}}(\check{x}_{0:N}, y_{1:N}, \check{\vartheta}_{0:N}; \theta, \tau, \tau_{0:N}) = p_{\check{\Theta}_{0:N}}(\check{\vartheta}_{0:N}; \theta, \tau, \tau_{0:N}) \times p_{X_{0:N}, Y_{1:N} | \check{\Theta}_{0:N}}(\check{x}_{0:N}, y_{1:N} | \check{\vartheta}_{0:N}),$$

where

$$p_{X_{0:N}, Y_{1:N} | \check{\Theta}_{0:N}}(\check{x}_{0:N}, y_{1:N} | \check{\vartheta}_{0:N}; \theta, \tau, \tau_{0:N}) = \mu(\check{x}_0; \check{\vartheta}_0) \prod_{n=1}^N f_n(\check{x}_n | \check{x}_{n-1}; \check{\vartheta}_n) \prod_{n=1}^N g_n(y_n | \check{x}_n; \check{\vartheta}_n).$$

This extended model can be used to define a perturbed parameter log-likelihood function, defined as

$$\check{\ell} \left(\check{\vartheta}_{0:N} \right) = \log p_{Y_{1:N} | \Theta_{0:N}} \left(y_{1:N}^* | \check{\vartheta}_{0:N}; \theta, \tau, \tau_{0:N} \right). \quad (4)$$

Here, we are treating the data as fixed and note that the right hand side does not depend on θ , τ , or $\tau_{0:N}$. We have designed (4) so that, setting

$$\check{\vartheta}^{[N+1]} = (\theta, \theta, \dots, \theta) \in d^{(N+1)},$$

we can write the log likelihood of the unperturbed model as

$$\ell(\theta) = \check{\ell} \left(\check{\vartheta}^{[N+1]} \right).$$

In our POMP framework, $p_{\Theta_{0:N}}$ is analogous to κ in the general latent variable model.

However, to formally match these two frameworks we must bear in mind that $p_{\Theta_{0:N}}$ carries out perturbations in $\Theta^{(N+1)d}$, so Theorems 1 and 2 must be applied in that extended parameter space.

For our perturbed likelihood, we need an extended version of Assumption 3, identical to assumption 5 of Doucet et al. (2013).

Assumption 5

$\check{\ell}$ is four times continuously differentiable. For all $\theta \in d$, there exist $\epsilon > 0$, $D > 0$ and δ defined as in Assumption 2, such that for all $0 < \eta_4 < \delta$ and $u_{0:N} \in d^{(N+1)}$,

$$\check{\mathcal{L}} \left(\check{\vartheta}^{[N+1]} + u_{0:N} \right) \leq D e^{\epsilon \sum_{n=1}^N |u_n|^{\eta_4}},$$

where $\check{\mathcal{L}} \left(\check{\vartheta}_{0:N} \right) = \exp \left\{ \check{\ell} \left(\check{\vartheta}_{0:N} \right) \right\}$ is the perturbed likelihood.

Let $\check{\mathbb{E}}_{\theta, \tau, \tau_{0:N}}$, $\check{\text{Cov}}_{\theta, \tau, \tau_{0:N}}$ denote as the expectation, covariance, and variance with respect to the associated posterior,

$$p_{\Theta_{0:N} | Y_{1:N}} \left(\check{\vartheta}_{0:N} | y_{1:N}^*; \theta, \tau, \tau_{0:N} \right).$$

To simplify the heavy notation, hereafter, we will use $\check{\cdot}$, $\check{\text{Cov}}$, and $\check{\text{Var}}$ instead of $\check{\mathbb{E}}_{\theta, \tau, \tau_{0:N}}$, and $\check{\text{Cov}}_{\theta, \tau, \tau_{0:N}}$, $\check{\text{Var}}_{\theta, \tau_{0:N}}$ respectively. The results. They are similar to Theorems 4 and 6 of

Doucet et al. (2013) but for random walk noise instead of independent white noise. Due to some favorable cancelation that arises in the proof, these new theorems have a relatively simple form.

Theorem 3

Suppose Assumption 1, 2, and 5, there exists a constant C_4 such that,

$$\left| \nabla \ell(\theta) - \tau^{-2} \Psi^{-1} \left\{ \tau_0^{-2} \tilde{\left(\Theta_0 - \theta | \tilde{Y}_{1:N} = y_{1:N}^* \right)} \right\} \right| < C_4 \tau^2,$$

where Ψ is the non-singular covariance matrix associated to ψ .

Proof—See the appendix.

Note that the constant C_4 in Theorem 3 may depend on N , the length of the time series. Here we consider the dataset to be fixed, so N is constant and we do not need to concern ourselves with the issue of how C_4 scales with N .

We propose to use a random walk in parameter space to explore the likelihood surface. These random walk perturbations can investigate the structure of the likelihood surface in a sequence of small steps. This allows some searching of parameter space within a single filtering iteration, which may be more computationally efficient than approaches, such as Doucet et al. (2013), which make white noise perturbations around a parameter value that is fixed for the duration of a filtering operation. However, one might suspect that the random walk perturbations should break the asymptotic guarantees developed by Doucet et al. (2013) for white noise perturbations. We show that this is not the case; random walk perturbations enjoy much of the theoretical support developed by Doucet et al. (2013), while being more computationally efficient empirically. We first state our theorems, leaving proofs to the appendix.

Theorem 4

Suppose Assumptions 1, 2, 4, and 5. There exist a C_5 such that

$$\left| I_\tau(\theta) + \nabla^2 \ell(\theta) \right| < C_5 \tau^2,$$

where

$$I_\tau(\theta) = -\tau^{-4} \Psi^{-1} \left\{ \tau_0^{-4} \tilde{\text{var}} \left(\tilde{\Theta}_0 | \tilde{Y}_{1:N} = y_{1:N}^* \right) - \tau_0^2 \tau^2 \Psi \right\} \Psi^{-1}. \quad (5)$$

Proof—See the appendix.

Theorems 3 and 4 formally allow approximation of $\nabla \ell(\theta), -\nabla^2 \ell(\theta)$. However, they rely heavily on the computation of the conditional distribution of $\check{\Theta}_0$ given $Y_{1:N}$, which is a computationally challenging smoothing problem. We therefore present some alternative variations on these results which lead to more stable Monte Carlo estimation. Our Theorems 5 and 6 consider a limit where τ_n is of order τ^2 for each $1 \leq n \leq N$, as $\tau \rightarrow 0$. This limit is similar to a limit in the context of IF1 by Ionides et al. (2011). The limit approaches numerically difficult smoothing calculations; however, the theorems can still carry out the useful purpose of motivating new algorithms whose finite sample properties are assessed empirically.

When $\tau_{0:N}$ are held constant as τ varies, Theorems 1 and 2 can be applied fairly directly to the perturbed POMP model. When $\tau_{0:N}$ are allowed to vary with τ , the parameter perturbations in Eq. (3) define a set of kernels, $\mathcal{K}$, consisting of densities on $d(N+1)$ parameterized by τ . Specifically, we suppose that $\mathcal{K}$ is comprised the density of

$\tau^{-1} \left(\check{\Theta}_{0:N} - \check{\vartheta}^{[N+1]} \right)$ for each τ . To analyze the convergence in the case of varying $\tau_{0:N}$, we need modification of Theorems 1 and 2 that apply to this set of kernels. We generalize Assumptions 1 and 2 over the set $\mathcal{K}$ as follows.

Assumption 6

$\mathcal{K}$ is a set of symmetric probability density kernels for which each $\kappa \in \mathcal{K}$ is associated with a non-singular and finite covariance matrix, $\Sigma(\kappa)$. There exists $C_6 < \infty$ such that for any integer $k \geq 1$, $1 \leq i_1, \dots, i_k \leq d$ and $\beta_1, \dots, \beta_k \geq 1$ and any $\kappa \in \mathcal{K}$,

$$\int |u_{i_1}^{\beta_1} u_{i_2}^{\beta_2} \dots u_{i_k}^{\beta_k}| \kappa(u) du \leq C_6.$$

Assumption 7

There exist $\gamma, \delta, M > 0$, such that for all $u \in \mathbb{R}^d$ and all $\kappa \in \mathcal{K}$,

$$|u| > M \Rightarrow \kappa(u) < e^{-\gamma|u|^\delta}.$$

Assumptions 6 and 7 are readily checked for Gaussian perturbations, so long as $\tau_{0:N}$ are uniformly bounded as functions of τ . The following first and second derivative approximations are proved in the appendix.

Theorem 5

Suppose Assumptions 4, 5, 6, and 7 hold. In addition, assume that $\tau_n = O(\tau^2)$ for all $n = 1 \dots N$. It follows that

$$\left| \nabla \ell(\theta) - \frac{1}{N+1} \tau^{-2} \tau_0^{-2} \Psi^{-1} \times \sum_{n=0}^N \left\{ \left(\check{\Theta}_n - \theta | \check{Y}_{1:N} = y_{1:N}^* \right) \right\} \right| = O(\tau^2). \quad (6)$$

Theorem 6

Suppose Assumptions 4, 5, 6, and 7 hold. In addition, assume that $\tau_n = O(\tau^2)$ for all $n = 1 \dots N$. It follows that

$$-\nabla^2 \ell(\theta) = I_\tau(\theta) + O(\tau^2),$$

where

$$I_\tau(\theta) = -\frac{1}{N+1} \tau^{-4} \tau_0^{-4} \Psi^{-1} \left\{ \sum_{n=0}^N \left(\tilde{V} \ar \left(\tilde{\Theta}_n | \tilde{Y}_{1:N} = y_{1:N}^* \right) - \sum_{k=0}^n \tau_k^2 \Psi \right) \right\} \Psi^{-1}.$$

To the best of our knowledge, an algorithm such as IS1 has not previously been used for data analysis. However, conceptually, IS1 is a direct application of the derivative estimates of Doucet et al. (2013). Estimation of the full covariance matrix, between all pairs of time points, would have a burdensome $O(N^3)$ computational cost and is typically unnecessary. As proposed by Doucet et al. (2013), one can use mixing assumptions to justify omitting covariances larger than some lag L , and one can use this same lag L for a fixed-lag particle smoothing algorithm using J particles. This gives rise to an $\mathcal{O}(NL^2J)$ algorithm. We develop equivalent results for our algorithm, based on the results proved by Doucet et al. (2013).

These results study the approximation properties of the score function and information matrix estimators for specific values of θ and τ . Full analysis of Algorithm 1 using stochastic approximation theory, as in Ionides et al. (2011), would require some uniformity of this approximation when τ is small and θ is in a neighborhood of the maximum of the likelihood function. Specifically, we make the following assumption essentially identical to Assumption 6 of Doucet et al.:

Assumption 8

For $\tau \in \tau^+$, $\theta \in \mathcal{D}$ and $n \in \{2, \dots, N\}$ define

$$\begin{aligned} S_1(\theta, \tau) &= \left[\phi \in \mathcal{D} : \kappa \{ (\phi - \theta) / \tau \} > 0 \right], \\ S_n(\theta, \tau) &= \left[\phi \in \mathcal{D} : \text{for every } \theta' \in S_{n-1}(\theta, \tau), \kappa \{ (\phi - \theta') / \tau \} > 0 \right]. \end{aligned}$$

There is some $b > 0$ such that, for $B = [0, b]$, the following conditions hold.

1. $S_n(\theta, \tau)$ is compact for all $n \in \{1, \dots, N\}$.
2. for all $n \in \{1, \dots, N\}$,

$$\begin{aligned}\underline{\alpha}_n(\theta) &= \inf_{\tau \in B, \phi \in S_n(\theta, \tau), x \in \mathcal{X}, x' \in \mathcal{X}} f_n(x' | x; \phi) > 0, \\ \overline{\alpha}_n(\theta) &= \sup_{\tau \in B, \phi \in S_n(\theta, \tau), x \in \mathcal{X}, x' \in \mathcal{X}} f_n(x' | x; \phi) < \infty, \\ \rho_n(\theta) &= 1 - \underline{\alpha}_n(\theta) / \overline{\alpha}_n(\theta) > 0.\end{aligned}$$

Let $\rho(\theta) = \max_{n \in \{1, \dots, N\}} \rho_n(\theta)$.

3. For $n \in \{2, \dots, N\}$, assume that there exists a probability measure $\lambda(dx)$ on $\mathcal{X}$, define

$$\begin{aligned}h_1(\theta, \phi, \tau) &= \int g_1(y | x; \phi) \tau^{-d} \kappa\{(\phi - \theta) / \tau\} d\phi \mu(x; \theta) \lambda(dx), \\ h_n(\theta, \phi, \tau) &= \int g_n(y | x; \phi) \tau^{-d} \kappa\{(\phi - \theta) / \tau\} d\phi \lambda(dx).\end{aligned}$$

then for all $y \in \mathcal{Y}$,

$$\begin{aligned}\overline{g}_n(y; \theta) &= \sup_{\tau \in B, \phi \in S_n(\theta, \tau), x \in \mathcal{X}} g_n(y | x; \phi) < \infty, \\ \underline{g}_n(y; \theta) &= \inf_{\tau \in B, \phi \in S_1(\theta', \tau) \theta' \in S_{n-1}(\theta, \tau)} h_n(\theta', \phi, \tau) > 0, \\ \overline{g}_1(y; \theta) &= \sup_{\tau \in B, \phi \in S_1(\theta, \tau), x \in \mathcal{X}} g_1(y | x; \phi) < \infty, \\ \underline{g}_1(y; \theta) &= \inf_{\tau \in B, \phi \in S_1(\theta, \tau)} h_1(\theta, \phi, \tau) > 0.\end{aligned}$$

Theorem 7

Suppose Assumption 8, the following hold true for random walk noise:

$$\begin{aligned}\tau^2 \Psi S_{\tau, N}(\theta) &= \tau^2 \Psi S_{\tau, L, N}(\theta) + O(\rho(\theta)^L), \\ \tau^4 \Psi \{I_{\tau, N}(\theta)\} \Psi &= \tau^4 \Psi I_{\tau, L, N}(\theta) \Psi + O(\rho(\theta)^L),\end{aligned}$$

where

$$\begin{aligned}S_{\tau, N}(\theta) &= \frac{1}{N+1} \tau^{-2} \tau_0^{-2} \Psi^{-1} \left\{ \sum_{n=0}^N \left(\check{\Theta}_n | \check{Y}_{1:N} = y_{1:N}^* - \theta \right) \right\}, \\ S_{\tau, L, N}(\theta) &= \frac{1}{N+1} \tau^{-2} \tau_0^{-2} \Psi^{-1} \left\{ \sum_{n=0}^N \left(\check{\Theta}_n | \check{Y}_{1:(n+L) \wedge N} = y_{1:(n+L) \wedge N}^* - \theta \right) \right\}, \\ I_{\tau, N}(\theta) &= - \frac{1}{N+1} \tau^{-4} \tau_0^{-4} \Psi^{-1} \left\{ \sum_{n=0}^N \left(\check{V} \ar \left(\check{\Theta}_n | \check{Y}_{1:N} = y_{1:N}^* \right) - \sum_{k=0}^N \tau_k^2 \Psi \right) \right\} \Psi^{-1}, \\ I_{\tau, L, N}(\theta) &= - \frac{1}{N+1} \tau^{-4} \tau_0^{-4} \Psi^{-1} \left(\sum_{n=0}^N \left(\check{V} \ar \left(\check{\Theta}_n | \check{Y}_{1:(n+L) \wedge N} = y_{1:(n+L) \wedge N}^* \right) - \sum_{k=0}^n \tau_k^2 \Psi \right) \right) \Psi^{-1}.\end{aligned}$$

Proof—The result follows directly from Olsson et al. (2008), as in the proof of proposition 7 of Doucet et al. (2013).

For completeness, we also state a Monte Carlo approximation result which is essentially identical to proposition 8 of Doucet et al. (2013).

Theorem 8

(Doucet et al. (2013), Proposition 8) *Suppose Assumption 8, then for all integers $N \geq 1$, $0 \leq L \leq N-1$, $J \geq 1$ is number of particles and for any $p \geq 2$, there exist constants C_7 and C_p , not depending on J , such that:*

$$\begin{aligned} \tau^2 \left| \left[\Psi \left\{ S_{\tau,L,N}^J(\theta) - S_{\tau,N}(\theta) \right\} \right] \right| &\leq \frac{C_7}{J}, \\ \tau^4 \left| \left[\Psi \left\{ I_{\tau,L,N}^J(\theta) - I_{\tau,L,N}(\theta) \right\} \Psi \right] \right| &\leq \frac{C_7}{J}, \end{aligned}$$

and

$$\begin{aligned} \tau^{21/p} \left[\left| \Psi \left\{ S_{\tau,L,N}^J(\theta) - S_{\tau,L,N}(\theta) \right\} \right|^p \right] &\leq \frac{C_p}{\sqrt{J}}, \\ \tau^{41/p} \left[\left| \Psi \left\{ I_{\tau,L,N}^J(\theta) - I_{\tau,L,N}(\theta) \right\} \Psi \right|^p \right] &\leq \frac{C_p}{\sqrt{J}}, \end{aligned}$$

where the expectations are with respect to the law of the bootstrap filter.

Pseudo-code for second-order iterated smoothing (IS2) is given in Algorithm 1. The initial values of the state variables at time t_0 can be treated as unknown parameters. We call them initial value parameters (IVPs) and denote them in Algorithm 1 by a subset of the parameter indices, $IC \subset \{1, \dots, d\}$. Since IVP affect the dynamics only at time t_0 , there can be no benefit to perturbing the parameters at other times, and so IVPs benefit from some special attention. The perturbations in lines 2 and 5 are taken to follow the normal distribution, though alternative densities with matching mean and variance could be chosen. Pseudo-code for the IS1 algorithm of Doucet et al. (2013) is given in the supplement (Algorithm S-1). We also tried a computationally faster simplification of IS1, which we call reduced iterated smoothing (RIS1). The RIS1 algorithm is the same as Algorithm 1 except that we use white noise to update filter at each time point, in steps 2 and 5. It is discussed further in the supplement (Section S-2).

The IS2 algorithm, together with IS1 and RIS1 algorithms based on (Doucet et al. 2013), were implemented in an open-source R package is 2 (Nguyen 2015), which is built based on *pomp* package (King et al. 2016). For our numerical particle filter implementation, we use systematic resampling (Douc et al. 2005) although our theoretical results are based on the multinomial resampling used for the original bootstrap particle filter of Gordon et al. (1993).

Main features of IF1, IF2, IS1, RIS1, and IS2 are summarized in Table 1.

5 Numerical examples

5.1 Toy example: a linear, Gaussian model

In this section, we evaluate our algorithm, comparing it to existing simulation-based approaches in terms of statistical performance and computational efficiency. We consider a

bivariate discrete time Gaussian autoregressive process, with Gaussian measurement error. This model is chosen so that the Monte Carlo calculations can be verified using a Kalman filter. The model is given by the state space forms:

$$\begin{aligned} X_n|X_{n-1} &= x_{n-1} \sim \mathcal{N}(\alpha x_{n-1}, \sigma^T \sigma), \\ Y_n|X_n &= x_n \sim \mathcal{N}(x_n, I_2). \end{aligned}$$

where α , σ are 2×2 matrices and I_2 is 2×2 identity matrix. Note that the optimal distribution can be derived in closed form. We simulate the dataset with the following parameters:

$$\alpha = \begin{bmatrix} \alpha_1 & \alpha_2 \\ \alpha_3 & \alpha_4 \end{bmatrix} = \begin{bmatrix} 0.8 & -0.5 \\ 0.3 & 0.9 \end{bmatrix}, \sigma = \begin{bmatrix} 3 & 0 \\ -0.5 & 2 \end{bmatrix}.$$

We set the number of time points $N = 100$ and initial value $X_0 = (-3, 4)$. For this model, we estimated the parameters α_2 and α_3 , as well as the initial value parameter (IVP) vector, X_0 . We ran our experiment with 25 iterations ($M = 25$) and with 1000 particles ($J = 1000$) on a workstation=computer with a 2.7 GHz processor. The initial random walk standard deviation of the perturbation should be small enough to not overshoot and big enough to make convergence not too slow. These can be assessed with convergence diagnostic plots (King et al. 2016). We used initial perturbations of 0.02 for regular parameters and 0.2 for IVPs, based on values used in previous work (Ionides et al. 2015). IVPs typically need larger perturbations, since the perturbation is applied only at the start of each filtering iteration. For this simple problem, we can reduce the perturbation intensity fairly quickly and still get successful convergence; we used a geometric scheme with the standard deviation reduced by a factor of 0.95 at each iteration. We use a bigger standard deviation perturbation of 0.23 for regular parameters and 2 for IVPs, for both IS1 and RIS1. Unlike for iterated filtering algorithms and IS2, the independent perturbations for IS1 and RIS1 do not accumulate through time and therefore heuristically should be larger. Preliminary experimentation confirmed this heuristic. First, we investigated the choice of the lag for the fixed-lag smoothing. We varied the lag for the fixed-lag smoothing from 1 to 5 for both method IS1 and IS2. Empirically, the choice of lag made little difference for IS2, and higher lags helped only for earlier iterations for IS1 (Fig. 1). Since increasing the lag also increases the computational demands, we used lag 1 smoothing for the subsequent investigation.

We compared second-order iterated smoothing (IS2) with the iterated filtering (IF1) algorithm of Ionides et al. (2011), the perturbed Bayes map iterated filtering (IF2) of Ionides et al. (2015), the second-order iterated smoothing (IS1) based on the derivative estimates of Doucet et al. (2013), and a reduced second-order iterated smoothing (RIS1) algorithm. Detailed specifications of IS1 and RIS1 are given in the supplement (Section S-2). As can be seen from Fig. 2, while estimated maximized log likelihoods of all approaches touch the true MLE at vertical broken line, the distribution of the estimated maximized log likelihoods using IS2 have higher mean and smaller variance, implying higher empirical convergence rate in this case. In addition, the proposed approach gives results that are reasonably robust

to the starting guesses, since we start at random values uniformly in a large rectangle. We note that, in this example, RIS1 climbs up the likelihood surface more efficiently than IF1, similar to IS1, but less efficiently than IF2 and IS2 approaches (Fig. 2). Additional results demonstrating the performance of IS2 compared to other approaches can be found in the supplement (Section S-1).

Average computational time of ten independent runs of each approach is given in Table 2. Additional overheads for fixed-lag smoothing and estimating score and observed information matrix for this simple model make the computation time of IS2 and RIS1 quite large compared to computational time of IF1 and IF2. The relatively high computational requirement of IS1 arises because this algorithm must estimate the covariance matrix between smoothed particles at all pairs of time points up to some fixed lag. Even for $L = 1$, in this case, we require nearly double the computational time compared to IS2 and RIS1. However, with complex models for which the main computational expense is simulating the dynamics of the latent process, these overheads become negligible and computational time of IF1, IF2, IS1, IS2, and RIS1 are similar.

5.2 Application to a malaria transmission model

Many real-world dynamic systems are highly non-linear and partially observed. Further, some combinations of parameters may be weakly identifiable from the available data. To demonstrate the capabilities of iterated smoothing for such situations, we consider a model for vivax malaria, a strain of malaria characterized by relapse following initial recovery from symptoms. Mathematical modeling of malaria has been a foundation for developing malaria control strategies since the work of Ross (1910) and Macdonald (1957). Malaria transmission is a challenging real-world system to investigate and leads to complex partially observed stochastic dynamic models. Inference for malaria models therefore provides a rigorous performance benchmark. Statistical inference for models of this kind was considered an unsolved scientific problem until recently (Bjørnstad and Grenfell 2001). This suggests that previously available techniques, such as employing sequential Monte Carlo likelihood estimates in generic stochastic optimization algorithms, without taking advantage of the POMP structure, are not readily applicable.

We consider the $SEIH^3QS$ model of Roy et al. (2013) which splits up the study population of size $P(t)$ into seven classes: susceptible individuals, $S(t)$, exposure $E(t)$, infected individuals, $I(t)$, dormant classes $H_1(t)$, $H_2(t)$, $H_3(t)$, and recovered individuals, $Q(t)$. Note that since infection with malaria results in incomplete and waning immunity, the last S in the model name indicates the possibility that a recovered person can return to the class of susceptible individuals. Data are sequence of monthly reported malaria morbidity, denoted by $y_{1:N}^*$. The latent force of infection $\lambda(t)$ passes through a delay stage, $\kappa(t)$ and the contributes to the current force of infection, $\mu_{SE}(t)$, with mean latency time τ_D . The state process is

$$(S(t), E(t), I(t), Q(t), H_1(t), H_2(t), H_3(t), \kappa(t), \mu_{SE}(t)),$$

where the birth rate for the S class ensures that $S(t) + E(t) + I(t) + Q(t) + \sum_j H_j(t) = P(t)$ while $P(t)$ is assumed known from the census data. The transition rates from stage H_1 to H_2 , H_2 to H_3 to Q are specified to be $3\mu_{HI}$. In this model, infected population enters dormancy via $I \rightarrow H$ transition are rate μ_{IH} , and the treated humans join non-relapsing infected in moving to the Q class. We suppose that $\{X(t), t \geq t_0\}$ follows a stochastic differential equation in which the human stage of the malaria pathogen lifecycle is modeled by

$$\begin{aligned} dS/dt &= \delta P + dP/dt + \mu_{IS}I + \mu_{QS}Q - m + a\mu_{IH}I + b\mu_{EI}E - \mu_{SE}(t)S - \delta S, \\ dE/dt &= \mu_{SE}(t)S - \mu_{EI}E - \delta E, \\ dI/dt &= (1-b)\mu_{EI}E + 3\mu_{HI}H_1 - (\mu_{IH} + \mu_{IS} + \mu_{IQ})I - \delta I, \\ dH_1/dt &= (1-a)\mu_{IH}I - n\mu_{HI}H_1 - \delta H_1, \\ dH_i/dt &= 3\mu_{HI}H_{i-1} - 3\mu_{HI}H_i - \delta H_i \quad \text{for } i \in [2, 3], \\ dQ/dt &= \mu_{IQ}I - \mu_{QS}Q - \delta Q, \end{aligned}$$

where δ represent mortality rate as defined in the supplementary table (Section S-3). When it goes with a compartment class, it represents the average number of deceased people in that class per time unit. A simple representation of the malaria pathogen reproduction within the mosquito vector is given by

$$\begin{aligned} d\kappa/dt &= [\lambda(t) - \kappa(t)]/\tau_D, \\ d\mu_{SE}/dt &= [\kappa(t) - \mu_{SE}(t)]/\tau_D. \end{aligned}$$

The Gamma-distributed delay imposed on $\lambda(t)$ by $\kappa(t)$ and $\mu_{SE}(t)$ can also be written as

$$\mu_{SE}(t) = \int_{-\infty}^t \gamma(t-s) \lambda(s) ds, \quad (7)$$

with $\gamma(s) = \frac{(2/\tau_D)^2 s^{2-1}}{(2-1)!} \exp(-2s/\tau_D)$, a gamma distribution with shape parameter 2. The latent force of infection contains a rainfall covariate $R(t)$, as described by Roy et al. (2013), and a Gamma white noise term,

$$\lambda(t) = \left(\frac{I+qQ}{P} \right) \times \exp \left\{ \sum_{i=1}^{N_s} b_i s_i(t) + b_r R(t) \right\} \times \left[\frac{d\Gamma(t)}{dt} \right],$$

where q denotes a reduced infection risk from humans in the Q class and $\{s_i(t), i = 1, \dots, N_s\}$ is a periodic cubic B-spline basis, with $N_s = 6$. We approximated the solution to this coupled system of stochastic differential equations using an Euler-Maruyama scheme (Kloeden and Platen 1999) with a time step of 1/20 month. Let the number of new cases in the n th interval be $M_n = \rho \int_{t_{n-1}}^{t_n} [\mu_{EI}E(s) + 3\mu_{HI}H_3(s)] ds$ where the times of the N observations are $t_1 < t_2 < \dots < t_N$ and the system is initialized at a time $t_0 = t_1 - 1/12$. The measurement model for Y_n given M_n is a negative binomial distribution with mean M_n and variance $M_n + M_n^2 \sigma_{obs}^2$. A table of parameter definitions and units is provided in the supplement (Section S-3).

We carried out inference for this model on data obtained from National Institutes of Malaria Research by Roy et al. (2013) using IF1, IF2, IS2, and RIS1. We ran our experiment on a Linux cluster, with $M = 50$ iterations and $J = 10^3$ particles. IS2 comparable to the recently developed algorithm IF2 (Ionides et al. 2015) for this example. Ionides et al. (2015) compared IF2 against IF1 on a benchmark problem in epidemiological dynamics, and we use this approach to test our methodology. In the presence of possible multi-modality, weak identifiability, and considerable Monte Carlo error of this model, we start 200 random searches in a large hyper-rectangle (see supplement S-3). The random walk standard deviation was initially set to 0.1 for regular parameters while the cooling rate c was set to $0.1^{0.02} \approx 0.95$. These corresponding quantities for initial value parameter perturbations were 2 and $0.1^{0.02}$, respectively. The standard deviations used for independent perturbations (for IS1 and RIS1) was five times that used for the random walk perturbation methods. Figure 3 shows the distribution of the maximized log likelihoods estimated by IF1, IF2, IS2, and RIS1. All distributions touch the global maximum as expected and the higher mean and smaller variance of IF2 and IS2 estimation clearly demonstrate that they are considerably more effective than IF1 and IS1. Experimentation with more extensive computation ($M = 100$ and $J = 10^4$) in Fig. 4 suggests that the performance improvement of IS2 over IF2 occurs primarily in simpler models, such as the toy example, or during earlier stages of optimization on complex models. We have had similar experiences with other complex models (results not shown). Our interpretation is that the parameter averaging involved in the parameter update rule for IS2 (lines 18 and 19 of Algorithm 1) can be inefficient when the likelihood surface contains non-linear ridges, whereas the IF2 algorithm does not carry out any averaging in parameter space. The computational times for IF1, IF2, and IS2 were 12.70, 12.34 and 14.56 h respectively, confirming that the computational complexities are similar for all three methods. In this computational challenge, we see that both IS2 and IF2 offer substantial improvement over IF1.

6 Conclusion

In this paper, we presented a novel approach for parameter estimation applicable to a general class of non-linear, non-Gaussian POMP models. We used artificial dynamics to estimate simultaneously the parameters and the states of the latent process of the POMP model. We were also able to approximate the score vector and the observed information matrix to accelerate the convergence rate of the inference. Previous approaches for POMP models involving an estimated information matrix have either excluded the plug-and-play property or experienced heavy computational costs that made practical implementation for real-world problems infeasible.

When the length of the time series goes to infinity, the parameter updating rule in our Algorithm 1 (IS2) approaches the time average of the smoothed perturbed parameters. It may be surprising that this simple updating rule has second-order convergence properties, at least in some asymptotic sense.

We have shown that the iterated smoothing theory of Doucet et al. (2013) can be adapted to apply with random walk perturbations. In other words, we have analyzed separately the two ways in which Doucet et al. (2013) modified Ionides et al. (2011): smoothing versus

filtering, and white noise perturbations versus random walk perturbations. Our theoretical results for random walk perturbations are similar to the results for white noise perturbations in Doucet et al. (2013), though we have not been able to develop analogous results to the convergence analysis in their Section 2.4. However, our empirical results for the random walk perturbation algorithm, IS2, are superior to those we obtained for IS1. In principle, different simulation-based inference methods can readily be hybridized to build on the strongest features of multiple algorithms. Our results could also be applied to develop other plug-and-play methodologies which can take advantage of estimators of the derivatives of the likelihood. For example, it may be possible to use our approach to help design efficient proposal distributions for particle Markov chain Monte Carlo algorithms, taking into account the local geometry of the target distribution.

Iterated filtering methodology has been applied to study epidemiological dynamics in various situations (King et al. 2008, 2015; Laneri et al. 2010, 2015; He et al. 2010, 2013; Bhadra et al. 2011; Camacho et al. 2011; Shrestha et al. 2011, 2013; Earn et al. 2012; Lavine and Rohani 2012; Lavine et al. 2013; Roy et al. 2013; Blackwood et al. 2013a,b; Blake et al. 2014; Martinez-Bakker et al. 2015; Romero-Severson et al. 2015). However, simulation-based inference for POMP models has potential applicability for statistical inference on non-linear POMP models arising throughout the biological, physical, and social sciences and in engineering. The theoretical and algorithmic innovations of this paper help build a new direction for future developments on this frontier.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This research was funded in part by National Science Foundation Grant DMS-1308919 and National Institutes of Health Grants 1-U54-GM111274 and 1-U01-GM110712.

Appendix: Proofs

Proof of Theorem 3

Let

$$R = \begin{bmatrix} \tau_0 I_{d \times d} & 0_{d \times d} & \cdots & 0_{d \times d} \\ \tau_0 I_{d \times d} & \tau_1 I_{d \times d} & \ddots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ \tau_0 I_{d \times d} & \tau_1 I_{d \times d} & \cdots & \tau_N I_{d \times d} \end{bmatrix}, \quad (8)$$

where $I_{d \times d}$ is identity matrix of dimension d and $0_{d \times d}$ is zero matrix of dimension d , then a random walk noise will be $R\tau Z_{0:N}$. From Assumption 5, $\check{\ell}$ is four times continuously differentiable for $\mathcal{D}^{N+1}$. Since N is fixed, we can apply Theorem 1, with

$\Sigma = Cov(RZ_{0:N}) = \check{\Psi}_N$, to obtain the existence of an η and a C_8 such that for every $\tau < \eta$ we have,

$$\left| \left(\check{\Theta}_{0:N} - \theta^{[N+1]} | \check{Y}_{1:N} = y_{1:N}^* \right) - \tau^2 \check{\Psi}_N \nabla \check{\ell} \left(\theta^{[N+1]} \right) \right| < C_8 \tau^4, \quad (9)$$

where

$$\check{\Psi}_N = \begin{bmatrix} \tau_0^2 \Psi & \tau_0^2 \Psi & \cdots & \tau_0^2 \Psi \\ \tau_0^2 \Psi & \tau_0^2 + \tau_1^2 \Psi & \ddots & \tau_0^2 + \tau_1^2 \Psi \\ \vdots & \vdots & \ddots & \vdots \\ \tau_0^2 \Psi & \tau_0^2 + \tau_1^2 \Psi & \cdots & \sum_{i=0}^N \tau_i^2 \Psi \end{bmatrix}. \quad (10)$$

Note that Assumptions 1 and 2 are automatically satisfied for the multivariate normal distribution with mean zero and variance $\check{\Psi}_N$, corresponding to the random variable $RZ_{0:N}$. As a result, for fixed $\tau_0, \dots, \tau_N$ and for a random walk noise, we have

$$\left| \nabla \check{\ell} \left(\theta^{[N+1]} \right) - \tau^{-2} \check{\Psi}_N^{-1} \left(\check{\Theta}_{0:N} - \theta^{[N+1]} | \check{Y}_{1:N} = y_{1:N}^* \right) \right| < C_9 \tau^2.$$

An application of the Gaussian-Jordan inverse method gives

$$\check{\Psi}_N^{-1} = \begin{bmatrix} (\tau_0^{-2} + \tau_1^{-2}) \Psi^{-1} & -\tau_1^{-2} \Psi^{-1} & \cdots & 0 \\ -\tau_1^{-2} \Psi^{-1} & (\tau_1^{-2} + \tau_2^{-2}) \Psi^{-1} & \cdots & \vdots \\ 0 & -\tau_2^{-2} \Psi^{-1} & \cdots & \vdots \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & (\tau_{N-1}^{-2} + \tau_N^{-2}) \Psi^{-1} & -\tau_N^{-2} \Psi^{-1} \\ 0 & 0 & -\tau_N^{-2} \Psi^{-1} & \tau_N^{-2} \Psi^{-1} \end{bmatrix}.$$

We write $\nabla_n \check{\ell} \left(\theta^{[N+1]} \right)$ for the d -dimensional vector of partial derivation of $\check{\ell} \left(\theta^{[N+1]} \right)$ with respect to each of the d components of θ_n . An application of the chain rule gives the identity

$$\nabla \ell(\theta) = \sum_{n=0}^N \nabla_n \check{\ell} \left(\theta^{[N+1]} \right),$$

giving rise to an inequality,

$$\left| \nabla \ell(\theta) - \tau^{-2} \sum_{n=0}^N \left\{ \Psi_N^{-1} \mathbb{E} \left(\Theta_{0:N} - \theta^{[N+1]} | Y_{1:N} = y_{1:N}^* \right) \right\}_n \right| < C_{10} \tau^2,$$

where $\{s\}_n$ is the entries $d n + 1, \dots, d(n + 1)$ of a vector $s \in \mathbb{R}^{d(N+1)}$. Decomposing the matrix multiplication by Ψ_N^{-1} into $d \times d$ blocks, we have

$$\tau^{-2} \sum_{n=0}^N \left\{ \Psi_N^{-1} \mathbb{E} \left(\Theta_{0:N} - \theta^{[N+1]} | Y_{1:N} = y_{1:N}^* \right) \right\}_n = \tau^{-2} \sum_{n=0}^N \text{SumCol}_n \left(\Psi_N^{-1} \right) \left(\Theta_n - \theta | Y_{1:N} = y_{1:N}^* \right), \quad (11)$$

where SumCol_n is the sum of the n th column in the $d \times d$ block construction of Ψ_N^{-1} . Every column of Ψ_N^{-1} except the first sums to 0, and this special structure of Ψ_N^{-1} gives a simple form,

$$\left| \sum_{n=0}^N \nabla_n \ell \left(\theta^{[N+1]} \right) - \tau^{-2} \Psi^{-1} \tau_0^{-2} \left(\Theta_0 - \theta | Y_{1:N} = y_{1:N}^* \right) \right| < C_{11} \tau^2.$$

This can be written as

$$\left| \nabla \ell(\theta) - \tau^{-2} \Psi^{-1} \tau_0^{-2} \left(\Theta_0 - \theta | Y_{1:N} = y_{1:N}^* \right) \right| < C_{11} \tau^2.$$

Proof of Theorem 4

Using similar set up as above, let the random walk noise be $R\tau Z_{0:N}$ with R defined as in Eq. (8). By selecting $p_{\Theta_{0:N}}$ to follow a multivariate normal distribution, Assumption 4 is also satisfied. From Theorem 2, for fixed $\tau_0, \dots, \tau_N$, there exist η and C_{12} such that for $0 < \tau < \eta$,

$$\left| \nabla^2 \ell \left(\theta^{[N+1]} \right) - \tau^{-4} \left[\Psi_N^{-1} \left(\text{Cov}_{\theta^{[N+1]}, \tau} \left(\Theta_{0:N} | Y_{1:N} = y_{1:N}^* \right) - \tau^2 \Psi_N \right) \Psi_N^{-1} \right] \right| < C_{12} \tau^2. \quad (12)$$

Define $\nabla_{s,n}^2 \ell \left(\theta^{[N+1]} \right)$ as

$$\nabla_{s,n}^2 \check{\ell}(\theta^{[N+1]}) = \frac{\partial^2 \check{\ell}(\theta^{[N+1]})}{\partial \theta_s \partial \theta_n}.$$

Applying the chain rule, we have

$$\nabla^2 \ell(\theta) = \sum_{s=0}^N \sum_{n=0}^N \nabla_{s,n}^2 \check{\ell}(\theta^{[N+1]}).$$

Adding up term in Eq. (12), we get

$$\left| \sum_{s=0}^N \sum_{n=0}^N \nabla_{s,n}^2 \check{\ell}(\theta^{[N+1]}) - \tau^{-4} \sum_{s=0}^N \sum_{n=0}^N \left[\check{\Psi}_{N-1} \left(\check{C} \text{ov}_{\theta^{[N+1]}, \tau} \left(\check{\Theta}_{0:N} | \check{Y}_{1:N} = y_{1:N}^* \right) - \tau^2 \check{\Psi}_N \right) \check{\Psi}_N^{-1} \right]_{s,n} \right| < C_{14} \tau^2.$$

where $\{A\}_{s,n}$ is the entries of rows $\{ds + 1, \dots, d(s + 1)\}$ and of columns $\{dn + 1, \dots, d(n + 1)\}$ of a matrix $A \in \mathbb{R}^{d(N+1) \times d(N+1)}$. Therefore,

$$\left| \nabla^2 \ell(\theta) - \tau^{-4} \sum_{s=0}^N \sum_{n=0}^N \left[\check{\Psi}_N^{-1} \left(\check{C} \text{ov}_{\theta^{[N+1]}, \tau} \left(\check{\Theta}_{0:N} | \check{Y}_{1:N} = y_{1:N}^* \right) - \tau^2 \check{\Psi}_N \right) \check{\Psi}_N \right]_{s,n} \right| < C_{14} \tau^2.$$

Defining SumCol_n as in Eq. (11), we have

$$\begin{aligned} & \sum_{s=0}^N \sum_{n=0}^N \left[\check{\Psi}_N^{-1} \left(\check{C} \text{ov}_{\theta^{[N+1]}, \tau} \left(\check{\Theta}_{0:N} | \check{Y}_{1:N} = y_{1:N}^* \right) - \tau^2 \check{\Psi}_N \right) \check{\Psi}_N \right]_{s,n} \\ &= \sum_{s=0}^N \sum_{n=0}^N \text{SumCol}_s \left(\check{\Psi}_N^{-1} \right) \text{SumCol}_n \left(\check{\Psi}_N \right) \\ & \times \left(\check{C} \text{ov}_{\theta^{[N+1]}, \tau} \left(\check{\Theta}_s, \check{\Theta}_n | \check{Y}_{1:N} = y_{1:N}^* \right) - \sum_{k=0}^{s \wedge n} \tau_k^2 \tau^2 \check{\Psi} \right) \check{\Psi}^{-1} = \left(\check{V} \text{ar}_{\theta^{[N+1]}, \tau} \left(\check{\Theta}_0 | \check{Y}_{1:N} = y_{1:N}^* \right) - \tau_0^2 \tau^2 \check{\Psi} \right). \end{aligned}$$

The last equality follows since $\check{\Psi}_N^{-1}$ is symmetric matrix with block of $d \times d$ for which each column except the first sums to 0. Thus we obtain

$$\left| \nabla^2 \ell(\theta) - \tau^{-4} \check{\Psi}^{-1} \left(\check{V} \text{ar}_{\theta^{[N+1]}, \tau} \left(\check{\Theta}_0 | \check{Y}_{1:N} = y_{1:N}^* \right) - \tau_0^2 \tau^2 \check{\Psi} \right) \check{\Psi} \right| < C_{15} \tau^2.$$

Proof of Theorem 5

In order to prove Theorem 5, we use the following corollary to Theorem 1.

Corollary 1

Suppose the perturbation kernel takes a value $\kappa \in \mathcal{K}$ satisfying Assumptions 6 and 7. Suppose also assumption 3. There exists an η and a constant C_{16} such that for every $0 < \tau < \eta$ and every $\kappa \in \mathcal{K}$,

$$\left| \left(\check{\Theta} - \theta | \check{Y} = y^* \right) - \tau^2 \Sigma \nabla \ell(\theta) \right| < C_{16} \tau^4. \quad (13)$$

Corollary 1 follows directly from applying Theorem 1, noting that Assumptions 6 and 7 imply a uniform bound on C_2 in Theorem 1. Applying Corollary 1, we obtain the existence of an η and C_{18} such that for every $\tau < \eta$ we have

$$\left| \tau^2 \check{\Psi}_N \nabla \check{\ell}(\theta^{[N+1]}) - \check{\left(\check{\Theta}_{0:N} - \theta^{[N+1]} | \check{Y}_{1:N} = y_{1:N}^* \right)} \right| < C_{18} \tau^4. \quad (14)$$

For compactness of notation, we write $E_n = \check{\left(\check{\Theta}_n - \theta | \check{Y}_{1:N} = y_{1:N}^* \right)}$ and

$D_n = \check{\Psi}_N \nabla \check{\ell}(\theta^{[N+1]})$, with $\check{\Psi}_N$ as in Eq. (10), writing out terms of the vector equation in (14) gives

$$E_0 = \tau^2 \tau_0^2 \sum_{n=0}^N D_n + O(\tau^4), \quad (15)$$

$$E_1 = \tau^2 \tau_0^2 \sum_{n=0}^N D_n + \tau^2 \tau_1^2 \sum_{n=1}^N D_n + O(\tau^4), \quad (16)$$

$$\vdots \quad (17)$$

$$E_{N-1} = \tau^2 \tau_0^2 \sum_{n=0}^N D_n + \cdots + \tau^2 \tau_{N-1}^2 \sum_{n=N-1}^N D_n + O(\tau^4), \quad (18)$$

$$E_N = \tau^2 \tau_0^2 \sum_{n=0}^N D_n + \cdots + \tau^2 \tau_N^2 D_N + O(\tau^4). \quad (19)$$

Using our assumption that for all $n = 1 \dots N$, $\tau_n = O(\tau^2)$, we get that $E_n = E_0 + O(\tau^4)$, from which we can conclude that

$$\frac{1}{N+1} \sum_{n=0}^N E_n = E_0 + O(\tau^4).$$

From Eq. (14) and the chain rule, for fixed τ_0 we have

$$\left| \nabla \ell(\theta) - \tau^{-2} \Psi^{-1} \tau_0^{-2} \left(\check{\Theta}_0 - \theta | \check{Y}_{1:N} = y_{1:N}^* \right) \right| < C_{10} \tau^2,$$

which then completes the proof.

Proof of Theorem 6

As for Theorem 5, we need a corollary to Theorem 2 over the kernel set $\mathcal{K}$, making use of Assumptions 6 and 7.

Corollary 2

Suppose Assumptions 3, 6, and 7. Suppose also that every kernel in $\mathcal{K}$ satisfies the mesokurtic condition in Assumptions 3. There exists an η and a constant C_{20} such that for every $0 < \tau < \eta$ and every $\kappa \in \mathcal{K}$,

$$\left| \left[\left(\check{\Theta} - \theta \right) \left(\check{\Theta} - \theta \right)^\top \mid \check{Y} = y^* \right] - \tau^2 \Sigma - \tau^4 \Sigma \left(\nabla^2 \ell(\theta) \right) \Sigma \right| < C_{20} \tau^6.$$

Applying Corollary 2, we have

$$\left| \tau^4 \check{\Psi}_N \nabla^2 \check{\ell} \left(\theta^{[N+1]} \right) \check{\Psi}_N - \left(\check{C} \text{ov}_{\theta^{[N+1]}, \tau} \left(\check{\Theta}_{0:N} \mid \check{Y}_{1:N} = y_{1:N}^* \right) - \tau^2 \check{\Psi}_N \right) \right| < C_{21} \tau^6.$$

For compact notation, we write

$$\check{C} \text{ov}_{s,n} = \check{C} \text{ov} \left(\check{\Theta}_s, \check{\Theta}_n \mid \check{Y}_{1:N} = y_{1:N}^* \right) - \sum_{k=0}^{s \wedge n} \tau_k^2 \tau^2 \Psi$$

and

$$\nabla_{s,n}^2 \check{\ell} = \nabla_{s,n}^2 \check{\ell} \left(\theta^{[N+1]} \right).$$

From the diagonal terms of the above matrix norm inequality, we derive $N+1$ equations,

$$\check{C} ov_{0,0} = \tau^4 \left[\check{\Psi}_N \nabla^2 \check{\ell} \check{\Psi}_N \right]_{0,0} + O(\tau^6), \quad (20)$$

$$\check{C} ov_{1,1} = \tau^4 \left[\check{\Psi}_N \nabla^2 \check{\ell} \check{\Psi}_N \right]_{1,1} + O(\tau^6), \quad (21)$$

$$\vdots \quad (22)$$

$$\check{C} ov_{N-1,N-1} = \tau^4 \left[\check{\Psi}_N \nabla^2 \check{\ell} \check{\Psi}_N \right]_{N-1,N-1} + O(\tau^6), \quad (23)$$

$$\check{C} ov_{N,N} = \tau^4 \left[\check{\Psi}_N \nabla^2 \check{\ell} \check{\Psi}_N \right]_{N,N} + O(\tau^6). \quad (24)$$

Using (20) through (24), and expanding out a matrix multiplication, we get

$$\begin{aligned} & \left[\check{\Psi}_N \nabla^2 \check{\ell} \check{\Psi}_N \right]_{n,n} \\ &= \Psi^2 \sum_{j=0}^n \left(\sum_{k=0}^j \tau_k^2 \right) \left[\sum_{i=0}^n \left(\sum_{k=0}^i \tau_k^2 \right) \nabla_{i,j}^2 \check{\ell} + \sum_{i=n+1}^N \left(\sum_{k=0}^n \tau_k^2 \right) \nabla_{i,j}^2 \check{\ell} \right] \\ &+ \Psi^2 \sum_{j=n+1}^N \left(\sum_{k=0}^n \tau_k^2 \right) \left[\sum_{i=0}^n \left(\sum_{k=0}^i \tau_k^2 \right) \nabla_{i,j}^2 \check{\ell} + \sum_{i=n+1}^N \left(\sum_{k=0}^n \tau_k^2 \right) \nabla_{i,j}^2 \check{\ell} \right]. \end{aligned}$$

Using our assumption that for all $n = 1 \dots N$, $\tau_n = O(\tau^2)$, we get that

$$\check{C} ov_{n,n} = \check{C} ov_{0,0} + O(\tau^6),$$

from which we can conclude that

$$\frac{1}{N+1} \sum_{n=0}^N \check{C} ov_{n,n} = \check{C} ov_{0,0} + O(\tau^6).$$

For $n = 0$, we have

$$\left[\tilde{\Psi}_N \nabla^2 \tilde{\ell} \tilde{\Psi}_N \right]_{0,0} = \Psi^2 \tau_0^4 \sum_{j=0}^N \sum_{i=0}^N \nabla_{i,j}^2 \tilde{\ell}, \quad (25)$$

from which, applying the chain rule completes the proof.

References

- Andrieu C, Doucet A, Holenstein R. Particle Markov chain Monte Carlo methods. *J. R. Stat. Soc. Ser. B.* 2010; 72(3):269–342.
- Bhadra A, Ionides EL, Laneri K, Pascual M, Bouma M, Dhiman RC. Malaria in Northwest India: data analysis via partially observed stochastic differential equation models driven by Lévy noise. *J. Am. Stat. Assoc.* 2011; 106:440–451.
- Bjørnstad ON, Grenfell BT. Noisy clockwork: time series analysis of population fluctuations in animals. *Science.* 2001; 293:638–643. [PubMed: 11474099]
- Blackwood JC, Cummings DAT, Broutin H, Iamsirithaworn S, Rohani P. Deciphering the impacts of vaccination and immunity on pertussis epidemiology in Thailand. *Proc. Natl. Acad. Sci. USA.* 2013a; 110:9595–9600. [PubMed: 23690587]
- Blackwood JC, Streicker DG, Altizer S, Rohani P. Resolving the roles of immunity, pathogenesis, and immigration for rabies persistence in vampire bats. *Proc. Natl. Acad. Sci. USA.* 2013b; 110:2083720842.
- Blake IM, Martin R, Goel A, Khetsuriani N, Everts J, Wolff C, Wassilak S, Aylward RB, Grassly NC. The role of older children and adults in wild poliovirus transmission. *Proc. Natl. Acad. Sci. USA.* 2014; 111(29):10604–10609. [PubMed: 25002465]
- Bretó C, He D, Ionides EL, King AA. Time series analysis via mechanistic models. *Ann. Appl. Stat.* 2009; 3:319–348.
- Camacho A, Ballesteros S, Graham AL, Carrat F, Ratmann O, Cazelles B. Explaining rapid reinfections in multiple-wave influenza outbreaks: Tristan da Cunha 1971 epidemic as a case study. *Proc. R. Soc. Lond. Ser. B.* 2011; 278(1725):3635–3643.
- Chopin N, Jacob PE, Papaspiliopoulos O. SMC²: an efficient algorithm for sequential analysis of state space models. *J. R. Stat. Soc. Ser. B.* 2013; 75(3):397–426.
- Dahlin J, Lindsten F, Schön TB. Particle Metropolis-Hastings using gradient and Hessian information. *Stat. Comput.* 2015; 25(1):81–92.
- Douc, R., Cappé, O., Moulines, E. Proceedings of the 4th International Symposium on Image and Signal Processing and Analysis. IEEE; New York: 2005. Comparison of resampling schemes for particle filtering.; p. 64–69.2005
- Doucet A, Jacob PE, Rubenthaler S. Derivative-free estimation of the score vector and observed information matrix with application to state-space models (version 2). 2013 arXiv:1304.5768v2.
- Earn DJ, He D, Loeb MB, Fonseca K, Lee BE, Dushoff J. Effects of school closure on incidence of pandemic influenza in Alberta. *Ann. Int. Med.* 2012; 156(3):173–181. [PubMed: 22312137]
- Gordon NJ, Salmond DJ, Smith AF. Novel approach to nonlinear/non-Gaussian Bayesian state estimation. *IEE Proc. F. Radar Signal Process.* 1993; 140:107–113.
- He D, Dushoff J, Day T, Ma J, Earn DJD. Inferring the causes of the three waves of the 1918 influenza pandemic in England and Wales. *Proc. R. Soc. Lond. Ser. B.* 2013; 280(1766):20131345.
- He D, Ionides EL, King AA. Plug-and-play inference for disease dynamics: measles in large and small populations as a case study. *J. R. Soc. Interface.* 2010; 7(43):271–283. [PubMed: 19535416]
- Ionides EL, Bhadra A, Atchadé Y, King A. Iterated filtering. *Ann. Stat.* 2011; 39:1776–1802.
- Ionides EL, Bretó C, King AA. Inference for nonlinear dynamical systems. *Proc. Natl. Acad. Sci. USA.* 2006; 103:18438–18443. [PubMed: 17121996]
- Ionides EL, Nguyen D, Atchadé Y, Stoev S, King AA. Inference for dynamic and latent variable models via iterated, perturbed Bayes maps. *P. Natl. Acad. Sci. USA.* 2015; 112(3):719–724.

- Kevrekidis IG, Gear CW, Hummer G. Equation-free: the computer-assisted analysis of complex, multiscale systems. *Am. Inst. Chem. Eng. J.* 2004; 50:1346–1354.
- King AA, Domenech de Celle M, Magpantay FMG, Rohani P. Avoidable errors in the modelling of outbreaks of emerging pathogens, with special reference to Ebola. *Proc. R. Soc. Lond. Ser. B.* 2015; 282:20150347.
- King AA, Ionides EL, Pascual M, Bouma MJ. Inapparent infections and cholera dynamics. *Nature.* 2008; 454:877–880. [PubMed: 18704085]
- King AA, Nguyen D, Ionides EL. Statistical inference for partially observed Markov processes via the R package pomp. *J. Stat. Softw.* 2016; 69:1–43.
- Kloeden, PE., Platen, E. Numerical Solution of Stochastic Differential Equations. 3rd edn.. Springer; New York: 1999.
- Kushner, HJ., Clark, DS. Stochastic Approximation Methods for Constrained and Unconstrained Systems. Springer; New York: 1978.
- Laneri K, Bhadra A, Ionides EL, Bouma M, Dhiman RC, Yadav RS, Pascual M. Forcing versus feedback: epidemic malaria and monsoon rains in Northwest India. *PLoS Comput. Biol.* 2010; 6(9):e1000898. [PubMed: 20824122]
- Laneri K, Paul RE, Tall A, Faye J, Diene-Sarr F, Sokhna C, Trape J-F, Rodó X. Dynamical malaria models reveal how immunity buffers effect of climate variability. *Proc. Natl. Acad. Sci. USA.* 2015; 112(28):8786–8791. [PubMed: 26124134]
- Lavine JS, King AA, Andreasen V, Bjørnstad ON. Immune boosting explains regime-shifts in prevaccine-era pertussis dynamics. *PLoS ONE.* 2013; 8(8):e72086. [PubMed: 23991047]
- Lavine JS, Rohani P. Resolving pertussis immunity and vaccine effectiveness using incidence time series. *Expert Rev. Vaccines.* 2012; 11:1319–1329. [PubMed: 23249232]
- Macdonald, G. The Epidemiology and Control of Malaria. Oxford University Press; Oxford: 1957.
- Martinez-Bakker M, King AA, Rohani P. Unraveling the transmission ecology of polio. *PLoS Biol.* 2015; 13(6):e1002172. [PubMed: 26090784]
- Nemeth C, Fearnhead P, Mihaylova L. Particle approximations of the score and observed information matrix for parameter estimation in state space models with linear computational cost. 2013 arXiv: 1306.0735.
- Nguyen, D. Iterated smoothing r package, is2. 2015. <https://r-forge.r-project.org/projects/is2>
- Olsson J, Cappé O, Douc R, Moulines E. Sequential Monte Carlo smoothing with application to parameter estimation in nonlinear state space models. *Bernoulli.* 2008; 14(1):155–179.
- Poyiadjis G, Doucet A, Singh SS. Particle approximations of the score and observed information matrix in state space models with application to parameter estimation. *Biometrika.* 2011; 98(1): 65–80.
- Romero-Severson E, Volz E, Koopman J, Leitner T, Ionides E. Dynamic variation in sexual contact rates in a cohort of HIV-negative gay men. *Am. J. Epidemiol.* 2015; 182:255–262. [PubMed: 25995288]
- Ross, R. The Prevention of Malaria. Dutton; Boston: 1910.
- Roy M, Bouma MJ, Ionides EL, Dhiman RC, Pascual M. The potential elimination of plasmodium vivax malaria by relapse treatment: Insights from a transmission model and surveillance data from NW India. *PLoS Negl. Trop. Dis.* 2013; 7:e1979. [PubMed: 23326611]
- Shrestha S, Foxman B, Weinberger DM, Steiner C, Viboud C, Rohani P. Identifying the interaction between influenza and pneumococcal pneumonia using incidence data. *Sci. Transl. Med.* 2013; 5(191):191ra84.
- Shrestha S, King AA, Rohani P. Statistical inference for multi-pathogen systems. *PLoS Comput. Biol.* 2011; 7(8):e1002135. [PubMed: 21876665]
- Sisson SA, Fan Y, Tanaka MM. Sequential Monte Carlo without likelihoods. *Proc. Natl. Acad. Sci. USA.* 2007; 104(6):1760–1765. [PubMed: 17264216]
- Spall, JC. Introduction to Stochastic Search and Optimization. Wiley; Hoboken: 2003.
- Toni T, Welch D, Strelkowa N, Ipsen A, Stumpf MP. Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems. *J. R. Soc. Interface.* 2009; 6:187–202. [PubMed: 19205079]

- Wood SN. Statistical inference for noisy nonlinear ecological dynamic systems. *Nature*. 2010; 466(7310):1102–1104. [PubMed: 20703226]
- Yıldırım S, Singh SS, Dean T, Jasra A. Parameter estimation in hidden Markov models with intractable likelihoods using sequential Monte Carlo. *J. Comput. Graph. Stat.* 2015; 24:846–865.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Fig. 1. Comparison of mean shortfall in maximized log likelihood from the true maximum likelihood estimate (MLE) for IF1, IS1 lag 1–5, and IS2 lag 1–5, based on 100 Monte Carlo replications. The *vertical bar* is the standard error of the mean shortfall. The True MLE and true log likelihood of estimated parameters were computed from Kalman filter

Fig. 2.

Comparison of estimators for the linear, Gaussian toy example, showing the densities of the maximized log likelihoods estimated by the IF1, IF2, IS1, RIS1, and IS2 methods. The parameters α_2 and α_3 were estimated, started from 200 randomly uniform starting values over the square region $[-1, 1] \times [-1, 1]$. The IVP vector, X_0 , was also estimated, starting from uniform draws on $[-5, 5] \times [-5, 5]$. The maximized log likelihood, computed by numerical maximization of the exact likelihood, is shown as a *dashed vertical line*

Fig. 3. The density of the maximized log likelihood approximations estimated by IF1, IF2, IS2, and RIS1 for the malaria model when using $J = 1000$ and $M = 50$. A previously computed maximized log likelihood is shown as a *dashed vertical line*

Fig. 4.
The density of the maximized log likelihood approximations estimated by IF1, IF2, and IS2 for the malaria model when using $J = 10,000$ and $M = 100$

Algorithm 1

Iterating smoothing (IS2)

Input:

starting parameter, θ_0
 simulator for $\mu(x_0; \theta)$
 simulator for $f_n(x_n | x_{n-1}; \theta)$
 evaluator for $g_n(y_n | x_n; \theta)$
 $I \subset \{1, \dots, d\}$ for initial value parameters
 data, $y_{1:N}^*$
 number of iteration, M
 number of particles, J
 perturbation scales, $\sigma_{1:d}$ defining a matrix $\Psi = \text{diag}(\sigma_{1:d}^2)$
 cooling rate, $0 < c < 1$,
 lag, L

Output:

Monte Carlo maximum likelihood estimate, θ_M

- 1: **for** m in $1 : M$ **do**
- 2: $[\theta_{0,j}^F]_i \sim \mathcal{N}([\theta_{m-1}]_i, (c^{m-1} \sigma_i)^2)$ for j in $1 : J$, i in $1 : d$
- 3: initialize states: simulate $X_{0,j}^F \sim \mu(x_0; \theta_{0,j}^F)$ for j in $1 : J$
- 4: **for** n in $1 : N$ **do**
 FILTERING:
- 5: $[\theta_{n,j}^P]_i \sim \mathcal{N}([\theta_{n-1,j}^F]_i, (c^{m-1} \sigma_i)^2)$ for $i \notin I$, j in $1 : J$
- 6: simulate particles: $X_{n,j}^P \sim f_n(x_n | X_{n-1,j}^F; \theta_{n,j}^P)$ for j in $1 : J$
- 7: evaluate weights: $w(n, j) = g_n(y_n^* | X_{n,j}^P; \theta_{n,j}^P)$ for j in $1 : J$
- 8: normalize weights: $\check{w}(n, j) = w(n, j) / \sum_{u=1}^J w(n, u)$
- 9: re-sample to select indices $k_{1:J}$ with $P\{k_u = j\} = \check{w}(n, j)$
- 10: re-sample: $X_{n,j}^F = X_{n,k_j}^P$ and $\theta_{n,j}^F = \theta_{n,k_j}^P$ for j in $1 : J$
- FIXED-LAG SMOOTHING FOR n in $1 : N - L$
- 11: let $a_1(n, k) = j$, $a_{l+1}(n, j) = a_l(n - L, a_l(n, j))$ for j in $1 : J$, l in $1 : L - 1$
- 12: smoothed mean: $\bar{\theta}_{n-L}^L = \sum_{j=1}^J \check{w}(n, j) \theta_{n-L, a_L(n, j)}^P$ if $n > L$
- 13: variance: $V_{n-L, n-L}^m = \sum_j \check{w}(n, j) \left(\theta_{n-L, a_L(n, j)}^P - \bar{\theta}_{n-L}^L \right) \left(\theta_{n-L, a_L(n, j)}^P - \bar{\theta}_{n-L}^L \right)^T$ if $n > L$
- 14: **end for**
- FIXED-LAG SMOOTHING FOR n in $N - L + 1 : N$

15: mean: $\bar{\theta}_{N+l-L}^L = \sum_{j=1}^J \check{w}(N, j) \theta_{N+l-L, a_{L-l}(N, j)}^P$ for l in $1 : L$

16: $V_{N+l-L, N+l-L}^m = \sum_j \check{w}(N, j) \left(\theta_{N+l-L, a_{L-l}(N, j)}^P - \bar{\theta}_{N+l-L}^L \right) \left(\theta_{N+l-L, a_{L-l}(N, j)}^P - \bar{\theta}_{N+l-L}^L \right)^T$ for l
 in $1 : L$

 PARAMETER UPDATE:

17: update: $S_m = c^{-2(m-1)} \Psi^{-1} \sum_{n=1}^N \left[\left(\bar{\theta}_n^L - \theta_{m-1} \right) \right]$

18: $I_m = -c^{-4(m-1)} \Psi^{-1} \left[\sum_{n=1}^N \left(V_{n,n}^m / (N+1) - c^{2(m-1)} \Psi \right) \right] \Psi^{-1}$

19: update parameters: $\theta_m = \theta_{m-1} + I_m^{-1} S_m$

20: update IVP parameters: $[\theta_m]_i = \frac{1}{J} \sum_{j=1}^J [\theta_{L,j}^F]_i$

21: **end for**

Table 1

Summary of algorithms in iterated filtering/smoothing class

Method	Perturbation	Filter type	Complexity
IF1	Random walk	Filtering	$O(NJ)$
IF2	Random walk	Filtering	$O(NJ)$
IS2	Random walk	Smoothing	$O(NLJ)$
IS1	White noise	Smoothing	$O(NL^2J)$
RIS1	White noise	Smoothing	$O(NLJ)$

Table 2

Computation times, in seconds, for the toy example

	$J = 100$	$J = 1000$	$J = 10,000$
IF1	4.201	14.820	140.889
IF2	4.141	14.055	125.624
IS2	6.858	27.014	281.736
IS1	10.156	47.901	466.182
RIS1	6.851	27.152	242.832