Noname manuscript No.
(will be inserted by the editor)

A second-order iterated smoothing algorithm

Dao Nguyen · **Edward L. Ionides**

Received: date / Accepted: date

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 have been developed to enable inference using simulation-based sequential Monte Carlo filters. An iterated smoothing algorithm was proposed by Doucet et al. (2013) and was found to have favorable theoretical properties. We develop a new iterated smoothing algorithm for which we establish both theoretical results and competitive practical performance. On benchmark computational challenges, our method beats the first-order iterated filtering algorithm. Our method's performance is 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) 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 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) 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) proposed some improvements to this algorithm by further exploiting both the score vector and the observed information matrix to increase convergence rate. The algorithm of Doucet et al. (2013) involves using sequential Monte Carlo methods to carry out iterated smoothing, and we call their algorithm IS1. Doucet

D. Nguyen
Department of Statistics, University of Michigan, Ann Arbor, Michigan, USA
E-mail: nguyenxd@umich.edu

E. L. Ionides
Department of Statistics, University of Michigan, Ann Arbor, Michigan, USA
E-mail: ionides@umich.edu

et al. (2013) showed that IS1 has second order convergence properties. 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 Section 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 theory 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 three-fold. First, we demonstrate theoretically that random walk parameter perturbations can be used in place of the white noise perturbations of IS1 while preserving the theoretical support provided by Doucet et al. (2013). Thus, IS2 inherits second-order convergence properties 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 Section 2, we introduce some notation and discuss some background to the computational challenge we investigate. In Section 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 Section 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 denote their values. Let $\{X(t),\, t\in\mathbb{T}\}$ be a Markov process with $X(t)$ taking values in a measurable space $\mathscr{X}$. The time index set, $\mathbb{T}\subset\mathbb{R}$, may be an interval or a discrete set and contains a finite subset $t_1<t_2<\cdots<t_N$ at which $X(t)$ is observed, together with some initial time $t_0<t_1$. We write $X_{0:N}=(X_0,\,\ldots,\,X_N)=(X(t_0),\,\ldots,\,X(t_N))$. Hereafter for any generic sequence $\{X_n\}$, we shall use $X_{i:j}$ to denote $(X_i,\,X_{i+1},\,\ldots,\,X_j)$. 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 $1\le n\le N$. Here, θ is an unknown parameter in $\mathbb{R}^d$. The process $\{X_n\}$ is only observed through another process $\{Y_n, n=1,\ldots,N\}$ taking values in a measurable space $\mathscr{Y}$. 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 $1\le n\le N$. We assume that $X_{0:N}$ and $\mathrm{Y}_{1:N}$ have a joint density $p_{X_{0:N},\,Y_{1:N}}(x_{0:N},\,y_{1:N};\theta)$ on $\mathscr{X}^{N+1}\times\mathscr{Y}^N$. The data are a sequence of observations by $y^*_{1:N}=(y^*_1,\,\ldots,\,y^*_N)\in\mathscr{Y}^N$, considered as fixed. We write the log likelihood function of the data for POMP model as $\ell(\theta)$, given by

$$\ell(\theta)=\log\int\mu(x_0;\theta)\prod_{n=1}^{N}f_n(x_n|x_{n-1};\theta)\;g_n(y^*_n|x_n;\theta)\;dx_{0:N}.$$

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 $\{\gamma_m\}_{m\ge1}$ is a sequence of step sizes with $\sum_{m\ge1}\gamma_m=\infty$ and $\sum_{m\ge1}\gamma_m^2<\infty$. Under appropriate regularity conditions, the algorithm converges to a local maximum of $\ell(\theta)$. The term $\nabla\ell(\theta)$ is shorthand for the $\mathbb{R}^d$-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 as in a Newton-Raphson approaches (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 a $d\times d$ matrix whose $(r,s)^{\text{th}}$ component $-\nabla^2\ell^{r,s}(\theta)$ are given for $r,s=1,\ldots,d$ and $\theta^r,\ \theta^s$ as rth, sth component of θ by

$$-\nabla^2\ell^{r,s}(\theta)=-\frac{\partial^2\ell(\theta)}{\partial\theta^r\partial\theta^s},$$

which is also known as the observed information matrix. Carrying out this approach via a simulation-based algorithm boils down to simulation-based estimation of the score 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). 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. However, Doucet et al.'s approach 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 approximation are shown in theorems 1 and 2, and more numerically stable approximations of these quantities are investigated in Theorems 3 and 4. Following the approach of Ionides et al. (2011) and Doucet et al. (2013), we first develop our theory (in Section 3) in the context of a latent variable model. Then, in Section 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 likelihood of the data $y^* \in \mathscr{Y}$ given by $\ell(\theta) = f_Y(y^*;\theta)$. We define a stochastically perturbed model corresponding to a pair of random variables $(\check{\Theta},\check{Y})$ having a joint probability density on $\mathbb{R}^d \times \mathscr{Y}$ given by

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

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 $\check{\Theta}$ 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 lemmas for stochastically perturbed models are restated from Doucet et al.'s Theorems 2 and 3, since they are foundations for our proofs. We suppose the following regularity conditions, identical to the assumptions of Doucet et al. (2013):

Assumption 1 *There exists $C < \infty$ such that for any integer $k \geq 1, 1 \leq i_1, \ldots, i_k \leq d$ and $\beta_1, \ldots, \beta_k \geq 1$,*

$$\int \left| u_{i_1}^{\beta_1} u_{i_2}^{\beta_2} \cdots u_{i_k}^{\beta_k} \right| \kappa(u)\, du \leq C,$$

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 *For all $\theta \in \mathbb{R}^d$, there exists $0 < \eta < \delta,\ \varepsilon,\ D > 0$, such that for all $u \in \mathbb{R}^d$,*

$$\mathscr{L}(\theta+u) \leq De^{\varepsilon|u|^\eta},$$

where $\mathscr{L} : \mathbb{R}^d \to \mathbb{R}$ is the likelihood function, $\ell = \log\mathscr{L}$ is four times continuously differentiable and δ defined as in Assumption 2.

Assumption 4 *κ satisfies $\int u_i^4 \kappa(u)du = 3\sigma_i^4$.*

These conditions could be relaxed but at the cost of substantially more complex proofs (Whiteley et al., 2013), so we do not pursue them here.

Lemma 1 *(Doucet et al. Theorem 2) Suppose assumption 1, 2, 3, there exists a constant C such that:*

$$\left|\check{\mathbb{E}}\left(\check{\Theta}-\theta\,\middle|\check{Y}=y\right) - \tau^2\Sigma\nabla\ell(\theta)\right| < C\tau^4. \tag{1}$$

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 was assumed, which is consistent with the practical choice of Gaussian perturbations.

Lemma 2 *(Doucet et al. Theorem 3) Suppose assumption 1, 2, 3 and 4, there exists a constant C such that:*

$$\left|\check{\mathbb{E}}\left[\left(\check{\Theta}-\theta\right)\left(\check{\Theta}-\theta\right)^\top\middle|\check{Y}=y\right] - \tau^2\Sigma - \tau^4\Sigma\left(\nabla^2\ell(\theta)\right)\Sigma\right| < C\tau^6. \tag{2}$$

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

$$\breve{\Theta}_n = \theta + \tau \sum_{i=0}^{n} \tau_i Z_i$$

for $0 \le n \le N$. We will later consider a limit where $\tau_{0:N}$ as fixed and the scale factor τ decreases toward zero, and subsequently another a limit where τ_0 is fixed but $\tau_{1:N}$ decrease toward zero together with τ. Let $p_{\breve{\Theta}_{0:N}}(\breve{\vartheta}_{0:N}; \theta,\ \tau,\ \tau_{0:N})$ be the probability density of $\breve{\Theta}_{0:N}$. We define the artificial random variables $\breve{\Theta}_{0:N}$ via their density,

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

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

$$p_{\breve{X}_{0:N}, \breve{Y}_{1:N}, \breve{\Theta}_{0:N}}(x_{0:N}, y_{1:N}, \breve{\vartheta}_{0:N}; \theta,\ \tau,\ \tau_{0:N}) = p_{\breve{\Theta}_{0:N}}(\breve{\vartheta}_{0:N}; \theta,\ \tau,\ \tau_{0:N}) p_{\breve{X}_{0:N}, \breve{Y}_{1:N} | \breve{\Theta}_{0:N}}(x_{0:N},\ y_{1:N} | \breve{\vartheta}_{0:N}),$$

where

$$p_{\breve{X}_{0:N}, \breve{Y}_{1:N} | \breve{\Theta}_{0:N}}(x_{0:N}, y_{1:N} | \breve{\vartheta}_{0:N}; \theta, \tau,\ \tau_{0:N}) = \mu(x_0; \breve{\vartheta}_0) \prod_{n=1}^{N} f_n(x_n | x_{n-1}; \breve{\vartheta}_n) \prod_{n=1}^{N} g_n(y_n | x_n; \breve{\vartheta}_n)$$

and

$$p_{\breve{\Theta}_{0:N}}(\breve{\vartheta}_{0:N}; \theta,\ \tau,\ \tau_{0:N}) = \breve{\psi}(\breve{\vartheta}_{0:N}; \theta, \tau,\ \tau_{0:N})$$

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

$$\breve{\ell}(\breve{\vartheta}_{0:N}) = \log p_{\breve{Y}_{1:N} | \breve{\Theta}_{0:N}}(y^*_{1:N} | \breve{\vartheta}_{0:N}; \theta, \tau, \tau_{0:N}). \tag{3}$$

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 (3) so that, setting

$$\breve{\vartheta}^{[N+1]} = (\theta, \theta, \ldots, \theta) \in \mathbb{R}^{d(N+1)},$$

we can write the likelihood of the unperturbed model as

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

In our POMP framework, $\breve{\psi}$ is analogous to κ in the general latent variable model. However, to formally match these two frameworks we must bear in mind that $\breve{\psi}$ carries out perturbations in $\Theta^{(N+1)d}$, so Lemmas 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 *$\breve{\ell} = \log \breve{\mathscr{L}}$ is four times continuously differentiable. For all $\theta \in \mathbb{R}^d$, there exist $\varepsilon > 0$, $D > 0$ and δ defined as in Assumption 2, such that for all $0 < \eta < \delta$ and $u_{0:N} \in \mathbb{R}^{d(N+1)}$,*

$$\breve{\mathscr{L}}(\breve{\vartheta}^{[N+1]} + u_{0:N}) \le D e^{\varepsilon \sum_{n=1}^{N} |u_n|^{\eta}},$$

where $\breve{\mathscr{L}}(\breve{\vartheta}_{0:N}) = \exp\{\breve{\ell}(\breve{\vartheta}_{0:N})\}$ is the perturbed likelihood.

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

$$p_{\breve{\Theta}_{0:N} | \breve{Y}_{1:N}}(\breve{\vartheta}_{0:N} | y^*_{1:N}; \theta,\ \tau, \tau_{0:N}).$$

To simplify the heavy notation, hereafter, we will use $\breve{\mathbb{E}}$, $\breve{\mathrm{Cov}}$, $\breve{\mathrm{Var}}$ instead of $\breve{\mathbb{E}}_{\theta,\tau,\tau_{0:N}}$, $\breve{\mathrm{Cov}}_{\theta,\tau,\tau_{0:N}}$, $\breve{\mathrm{Var}}_{\theta,\tau,\tau_{0:N}}$ respectively. The following theorems 1 and 2 are our main results, they are similar to theorem 4 and 6 of Doucet et al. (2013) but for random walk noise instead of independent white noise and are much simpler.

Theorem 1 *Suppose assumption 1, 2 and 5, there exists a constant C such that,*

$$\left|\nabla \ell(\theta) - \tau^{-2} \Psi^{-1} \left\{\tau_0^{-2} \breve{\mathbb{E}}\left(\breve{\Theta}_0 - \theta | \breve{Y}_{1:N} = y^*_{1:N}\right)\right\}\right| < C\tau^2,$$

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

Proof See appendix A.1.

We propose to use random walk noise to explore the likelihood surface. Intuitively, at each time point the random walk can take a small step in an appropriate direction on the likelihood surface. This may be more computationally efficient than perturbing one set of parameters at each time point, as in the independent white noise perturbations analyzed by Doucet et al. (2013). We show that random walk perturbations enjoy some 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 2 *Suppose assumptions 1, 2, 4 and 5, the following hold true for random walk noise,*

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

where

$$I_\tau(\theta) = -\tau^{-4} \Psi^{-1} \left\{\tau_0^{-4} \left(\breve{\mathrm{Var}}\left(\breve{\Theta}_0 | \breve{Y}_{1:N} = y^*_{1:N}\right) - \tau_0^2 \Psi\right)\right\} \Psi^{-1}.$$

Proof See appendix A.2.

Theorem 1 and 2 formally allow approximation of $\nabla\ell(\theta)$, $-\nabla^2\ell(\theta)$. However, they rely heavily on the computation of the conditional distribution of $\breve{\Theta}_0$ 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 3 and 4 consider a limit where τ_n is of order τ^2 for each $1 \leq n \leq N$, as $\tau \to 0$. This limit is similar to a limit studied in the context of IF1 by Ionides et al. (2011). This is not an ideal theoretical framework, since it approaches another limit which involves 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. We state two additional theorems as follows.

Theorem 3 *Suppose assumption 1, 2 and 5 hold. In addition, assume that τ_n is $O(\tau^2)$ for every $n = 1 \ldots N$, the following hold true,*

$$\left|\nabla\ell(\theta) - \frac{1}{N+1}\tau^{-2}\tau_0^{-2}\Psi^{-1}\sum_{n=0}^{N}\left\{\breve{\mathbb{E}}\left(\breve{\Theta}_n - \theta | \breve{Y}_{1:N} = y^*_{1:N}\right)\right\}\right| = O(\tau^2). \quad (4)$$

Proof See appendix A.3.

Theorem 4 *Suppose assumptions 1, 2, 4 and 5 hold. In addition, assume that τ_n is $O(\tau^2)$ for every $n = 1 \ldots N$, the following hold true for random walk noise,*

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

where

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

Proof See appendix A.4.

To the best of our knowledge, the approach of Doucet et al. (2013) has not previously been used for data analysis. In part, this could be due to the computational expense of its estimation of the covariance matrix estimation for the perturbed parameters. The computational cost of the full covariance estimation in the method of Doucet et al. (2013), between all pairs of time points, is $\mathscr{O}(N^2)$ at each time point and so $\mathscr{O}(N^3)$ for an entire smoothing computation. As proposed by Doucet et al. (2013), one can omit covariances larger than some lag L, and one can use this same lag L for a fixed-lag particle smoothing algorithm using J particles. Doucet et al. (2013) studied the properties of such an algorithm, under strong mixing assumptions, to derive an algorithm with computational cost $\mathscr{O}(NL^2J)$. Here, we write equivalent results for our algorithm, based on the results proved by Doucet et al. (2013). These result 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 assumptions:

Assumption 6 *(Assumption 6 of Doucet et al.). Define*

$$S(\phi,\phi',\tau) = \left[(\phi,\phi') \in \mathbb{R}^{d\times 2} : \kappa\left\{(\phi-\phi')/\tau\right\} > 0\right].$$

1. *$S(\phi,\phi',\tau)$ is compact.*
2. *for all $n \in \{1,..,N\}$,*

$$\underline{\alpha}_n(\phi') = \inf_{(\phi,\phi',x,x')\in S(\phi,\phi',\tau)\times\mathscr{X}\times\mathscr{X}} f_n(x'|x;\phi') > 0,$$
$$\overline{\alpha}_n(\phi') = \sup_{(\phi,\phi',x,x')\in S(\phi,\phi',\tau)\times\mathscr{X}\times\mathscr{X}} f_n(x'|x;\phi') < \infty,$$
$$\rho_n(\phi') = 1 - \underline{\alpha}_n(\phi')/\overline{\alpha}_n(\phi') > 0.$$

   Let $\rho(\theta) = \max_{n\in\{1,..,N\}}\rho_n(\phi')$.
3. *There exists a probability measure $\lambda(dx)$ on $\mathscr{X}$ such that, for all $y \in \mathscr{Y}$ and for all $n \in \{2,..,N\}$,*

$$\overline{g}_n(y;\phi') = \sup_{(\phi,\phi',x)\in S(\phi,\phi',\tau)\times\mathscr{X}} g_n(y|x;\phi') < \infty,$$
$$\underline{g}_n(y;\phi') = \int g_n(y|x;\phi)\tau^{-d}\kappa\left\{(\phi-\phi')/\tau\right\}d\phi\lambda(dx) > 0$$
$$\overline{g}_1(y;\phi') = \sup_{(\phi,\phi',x)\in S(\phi,\phi',\tau)\times\mathscr{X}} g_1(y|x;\phi') < \infty,$$
$$\underline{g}_1(y;\theta) = \int g_1(y|x;\phi)\tau^{-d}\kappa\{(\phi-\theta)/\tau\}d\phi\mu(x;\theta)\lambda(dx) > 0.$$

Theorem 5 *Suppose assumption 6, the following hold true for random walk noise:*

$$\tau^2\Psi S_{\tau,N}(\theta) = \tau^2\Psi S_{\tau,L,N}(\theta) + O(\rho(\theta)^L),$$
$$\tau^4\Psi\left\{I_{\tau,N}(\theta)\right\}\Psi = \tau^4\Psi I_{\tau,L,N}(\theta)\Psi + O(\rho(\theta)^L),$$

where

$$S_{\tau,N}(\theta) = \frac{1}{N+1}\tau^{-2}\tau_0^{-2}\Psi^{-1}\left\{\sum_{n=0}^{N}\left(\breve{\mathbb{E}}(\breve{\Theta}_n|\breve{Y}_{1:N} = y^*_{1:N}) - \theta\right)\right\},$$

$$S_{\tau,L,N}(\theta) = \frac{1}{N+1}\tau^{-2}\tau_0^{-2}\Psi^{-1} \left\{\sum_{n=0}^{N}\left(\breve{\mathbb{E}}(\breve{\Theta}_n|\breve{Y}_{1:(n+L)\wedge N} = y^*_{1:(n+L)\wedge N}) - \theta\right)\right\},$$

$$I_{\tau,N}(\theta) = -\frac{1}{N+1}\tau^{-4}\tau_0^{-4}\Psi^{-1} \left\{\sum_{n=0}^{N}\left(\breve{\mathrm{Var}}\left(\breve{\Theta}_n|\breve{Y}_{1:N} = y^*_{1:N}\right) - \sum_{k=0}^{n}\tau_k^2\Psi\right)\right\}\Psi^{-1},$$

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

Proof It 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 a theorem of Doucet et al. (2013).

Theorem 6 *(Doucet et al., Proposition 8). Suppose assumption 6, then for all integers $N \geq 1, 0 \leq L \leq N-1$, $J \geq 1$ and for any $p \geq 2$, there exist constants C and C_p, not depending on J, such that:*

$$\tau^2\left|\mathbb{E}\left[\Psi\left\{S^J_{\tau,L,N}(\theta)-S_{\tau,N}(\theta)\right\}\right]\right| \leq \frac{C}{J},$$

$$\tau^4\left|\mathbb{E}\left[\Psi\left\{I^J_{\tau,L,N}(\theta)-I_{\tau,L,N}(\theta)\right\}\Psi\right]\right| \leq \frac{C}{J},$$

and

$$\tau^2\mathbb{E}^{1/p}\left[\left|\Psi\left\{S^J_{\tau,L,N}(\theta)-S_{\tau,L,N}(\theta)\right\}\right|^p\right] \leq \frac{C_p}{\sqrt{J}},$$

$$\tau^4\mathbb{E}^{1/p}\left[\left|\Psi\left\{I^J_{\tau,L,N}(\theta)-I_{\tau,L,N}(\theta)\right\}\Psi\right|^p\right] \leq \frac{C_p}{\sqrt{J}}.$$

Pseudo-code for second order iterated smoothing (IS2) is given in Algorithm 1. The initial value parameters (IVPs) in Algorithm 1 are defined as parameters which are perturbed only at time zero (Bretó et al., 2009). 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). Since IS1 is too computational expensive to apply in real problems, we propose a reduced second order iterated smoothing (RIS1) approach. 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.

The IS2 algorithm, together with IS1 and RIS1 algorithms based on (Doucet et al., 2013), were implemented in an open-source R package *is2* (Nguyen and Ionides, 2015), which is built based on *pomp* package (King et al., 2015b).

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 $g_n(y_n|x_n;\theta)$
$I \subset \{1,..,d\}$ for initial value parameters
data, $y^*_{1:N}$
number of iteration, M
number of particles, J
perturbation scales, $\sigma_{1:d}$, defining a matrix $\Psi = \mathrm{diag}(\sigma^2_{1:d})$
cooling rate, $0 < c < 1$,
lag, L
Output:
Monte Carlo maximum likelihood estimate, θ_M

1: **for** m in $1:M$ **do**
2: $\left[\Theta^F_{0,j}\right]_i \sim \mathscr{N}\left(\left[\theta_{m-1}\right]_i, (c^{m-1}\sigma_i)^2\right)$ for j in $1:J$, i in $1:d$
3: initialize states: simulate $X^F_{0,j} \sim \mu\left(x_0;\Theta^F_{0,j}\right)$ for j in $1:J$
4: **for** n in $1:N$ **do**
5: $\left[\Theta^P_{n,j}\right]_i \sim \mathscr{N}\left(\left[\Theta^F_{n-1,j}\right]_i, (c^{m-1}\sigma_i)^2\right)$ for $i \notin I$, j in $1:J$
6: simulate particles: $X^P_{n,j} \sim f_n\left(x_n|X^F_{n-1,j};\Theta^P_{n,j}\right)$ for j in $1:J$
7: evaluate weights: $w(n,j) = g_n(y^*_n|X^P_{n,j};\Theta^P_{n,j})$ 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: $X^F_{n,j} = X^P_{n,k_j}$ and $\Theta^F_{n,j} = \Theta^P_{n,k_j}$ for j in $1:J$
11: let $a_1(n,k_j) = j$, $a_{l+1}(n,j) = a_1(n-l, a_l(n,j))$ for j in $1:J$, l in $0:L-1$
12: smoothed mean: $\bar{\theta}^L_{n-L} = \sum_{j=1}^J \check{w}(n,j)\Theta^P_{n-L,a_L(n,j)}$ if $n > L$
13: variance:$V^m_{n-L,n-L} = \sum_j \check{w}(n,j)\left(\Theta^P_{n-L,a_L(n,j)} - \bar{\theta}^L_{n-L}\right)\left(\Theta^P_{n-L,a_L(n,j)} - \bar{\theta}^L_{n-L}\right)^\top$ if $n > L$
14: **end for**
15: mean: $\bar{\theta}^L_{n+l-L} = \sum_{j=1}^J \check{w}(N,j)\Theta^P_{n+l-L,a_{L-l}(N,j)}$ for l in $1:L$
16: variance:$V^m_{n+l-L,n+l-L} = \sum_l \check{w}(N,l)\left(\Theta^P_{n+l-L,a_{L-l}(N,l)} - \bar{\theta}^L_{n+l-L}\right)\left(\Theta^P_{n+l-L,a_{L-l}(N,l)} - \bar{\theta}^L_{n+l-L}\right)^\top$ for l in $1:L$
17: update: $S_m = c^{-2(m-1)}\Psi^{-1}\sum_{n=1}^N\left[\left(\bar{\theta}^L_n - \theta_{m-1}\right)\right]$
18: $I_m = -c^{-4(m-1)}\Psi^{-1}\left[\sum_{n=1}^N\left(V^m_{n,n}/(N+1) - c^{2(m-1)}\Psi\right)\right]\Psi^{-1}$
19: update parameters: $\theta_m = \theta_{m-1} + I_m^{-1}S_m$
20: update IVP parameters: $\left[\theta_m\right]_i = \frac{1}{J}\sum_{j=1}^J\left[\Theta^F_{L,j}\right]_i$ for $i \in I$
21: **end for**

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 term 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:

$$X_n|X_{n-1} = x_{n-1} \sim \mathscr{N}(\alpha x_{n-1}, \sigma^\top\sigma),$$
$$Y_n|X_n = x_n \sim \mathscr{N}(x_n, I_2).$$

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 data set 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 starting point $X_0 = (-3,4)$. For this model, we try to estimate parameters α_2 and α_3. We run our experiment with 25 iterations ($M = 25$) and with 1000 particles ($J = 1000$) on a

Table 1 Computation times, in seconds, for the toy example.

	$J = 100$	$J = 1000$	$J = 10000$
IF1	2.206	7.744	71.538
IF2	1.749	6.666	59.450
IS2	4.072	12.971	106.678
IS1	630.120	7516.568	
RIS1	5.358	15.186	102.980

Linux personal computer with a 3.0GHz processor. Our approach, second order iterated smoothing (IS2) is compared against the iterated filtering (IF1) of Ionides et al. (2011), the perturbed Bayes map iterated filtering (IF2) of Ionides et al. (2015), the second-order iterated smoothing (IS1) of Doucet et al. (2013) and the reduced second-order iterated smoothing approach (RIS1)(see supplement S-2). As can be seen from Fig. 1, while MLEs of all approaches touch the true MLE at vertical broken line, the distribution of the estimated MLEs 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 approach climbs up the likelihood surface more efficiently than IF1 approach, similar to IS1 approach but less efficiently than IF2 and IS2 approaches (Fig. 1). Algorithmically, IS2 has similar computational costs with the first order approaches IF1 & IF2 and with the second-order RIS1 approach while the original IS1 of Doucet et al. (2013) takes longer time than any other approaches because of extensive computing covariance between time points. 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 Table1. 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. However, with complex models and large enough number of particles, these overheads become negligible and computational time of IF1, IF2, IS2 and RIS1 are similar. The much longer time of IS1 arise because our implementation of this algorithm computes covariances between all pairs of time points. Algorithms between IS1 and RIS1 that compute covariances only up to some fixed lag would be expected to have computational time and performance intermediate between these extremes.

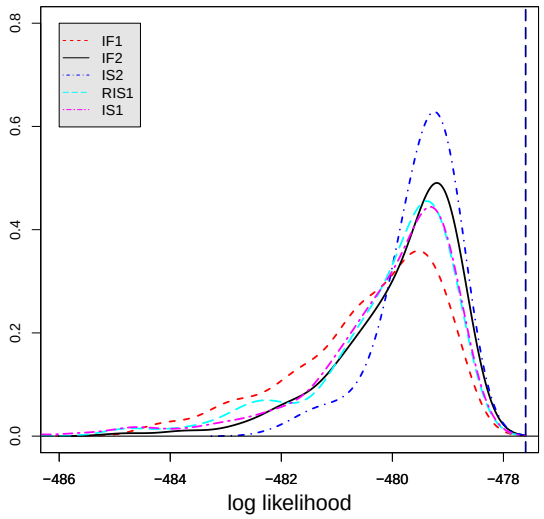

Fig. 1 Comparison of estimators for the linear, Gaussian toy example, showing the densities of the MLEs estimated by the IF1, IF2, IS1, RIS1, and IS2 methods. The parameters α_2 and α_3 were estimated, started from 200 randomly uniform initial values over a large rectangular region $[-1,1]\times[-1,1]$.

5.2 Application to a malaria transmission model

Many real world dynamic systems are highly nonlinear and partially observed. Further, some combinations of parameters may be weakly identifiable from the available data. To demonstrate the capabilities of iterated smoothing (IS2) for such situations, we consider a model for vivax malaria, a strain of malaria characterized by relapse following initial recovery from symptoms. Malaria transmission is challenging real world system to analyze, and therefore provides a rigorous performance benchmark. Mathematical modeling of malaria has been a foundation for developing malaria control strategies since the work of Ross (1910) and Macdonald (1957). 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, $Q(t)$. Data are a 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

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

where the birth rate for the S class ensures that $S(t)+E(t)+I(t)+Q(t)+\sum_i H_i(t)=P(t)$ 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 and H_3 to Q are specified to be $3\mu_{HI}$. In this model, infected population enters dormancy via $I-$to$-$

H transition at 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

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

and a simple representation of the malaria pathogen reproduction within the mosquito vector is given by

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

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

$$\mu_{SE}(t) = \int_{-\infty}^{t} \gamma(t-s)\lambda(s)\mathrm{d}s, \tag{5}$$

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,

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

where q denotes a reduced infection risk from humans in the Q class and $\{s_i(t), i = 1,\ldots,N_s\}$ is a periodic cubic B-spline basis, with $N_s = 6$. Let the number of new cases in the nth 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 $t_1 < t_2 < \cdots < t_N$ 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 $M_n + M_n^2 \sigma_{\mathrm{obs}}^2$. 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. Unlike the toy example, the second order iterated smoothing with white noise (IS1) was left out as it is too computational demanding for this problem. Our approach is 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 IS2 and RIS1. 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 hyperrectangle (see supplement S-3). The random walk standard deviation is initially set to 0.1 for estimated parameters while the cooling rate c is set to $0.1^{0.02} \approx 0.95$. These corresponding quantities for initial value parameter perturbations are 2 and $0.1^{0.02}$, respectively, but they are applied only at time zero. The standard deviation of independent perturbation for RIS1 is five times as that of other methods. Figure 2 shows the distribution of the MLEs estimated by IF1, IF2, IS2 and RIS1. All distributions touch the global maximum as expected and the higher mean and smaller variance of IF2, 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 Figure 3 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 averaging involved in the parameter update rule for IS2 can be inefficient when the likelihood surface contains nonlinear 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 hours 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.

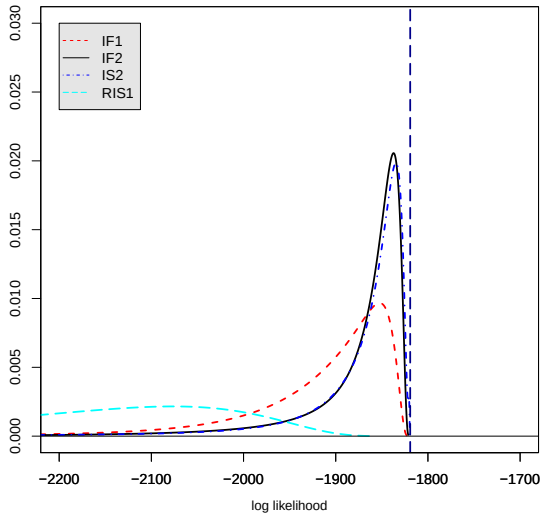

Fig. 2 The density of the maximized log likelihood approximations estimated by IF1, IF2, IS2 and RIS1 for the malaria model when using $J = 1000$ and $M = 50$. The log likelihood at a previously computed MLE is shown as a dashed vertical line.

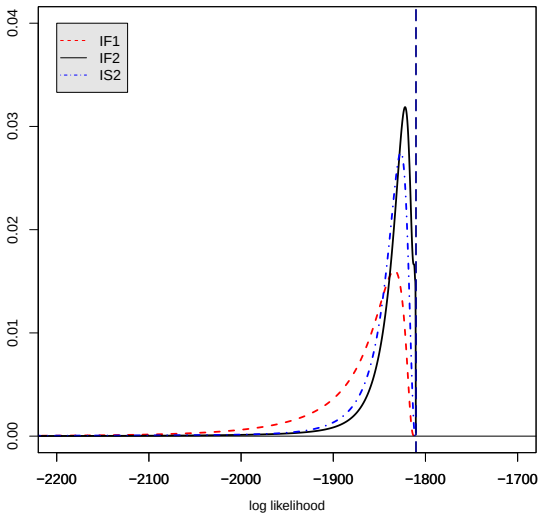

Fig. 3 The density of the maximized log likelihood approximations estimated by IF1, IF2 and IS2 for the malaria model when using $J = 10000$ and $M = 100$

6 Conclusion

In this paper, we presented a novel approach for parameter estimation applicable to a general class of nonlinear, 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 are similar to Doucet et al. (2013). However, we have not been able to develop analogous results to the convergence analysis in their Section 2.4. Nevertheless, our empirical results are stronger. 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; Laneri et al., 2010; He et al., 2010; Bhadra et al., 2011; Camacho et al., 2011; Shrestha et al., 2011; Earn et al., 2012; Lavine and Rohani, 2012; Lavine et al., 2013; He et al., 2013; Roy et al., 2013; Blackwood et al., 2013a,b; Shrestha et al., 2013; Blake et al., 2014; King et al., 2015a; Laneri et al., 2015; Martinez-Bakker et al., 2015; Romero-Severson et al., 2015). However, simulation-based inference for POMP models has potential applicability for statistical inference on nonlinear POMP models arising throughout the biogical, physical and social sciences and in engineering. The theoretical and algorithmic innovations of this paper help to build a new direction for future developments on this frontier.

Acknowledgements 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.

A Proofs

A.1 Proof of Theorem 1

Let

$$R = \begin{bmatrix} \tau_0 I_{d\times d} & 0_{d\times d} & \cdots & 0_{d\times d} \\ \tau_0 I_{d\times d} & \tau_1 I_{d\times d} & \ddots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ \tau_0 I_{d\times d} & \tau_1 I_{d\times d} & \cdots & \tau_N I_{d\times d} \end{bmatrix}, \tag{6}$$

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 $R\tau Z_{0:N}$. Apply Lemma 1 with $\Sigma = \mathrm{Cov}(RZ_{0:N}) = \breve{\Psi}_N$, there exist η and C such that for $0 < \tau < \eta$,

$$\left| \breve{\mathbb{E}}\left(\breve{\Theta}_{0:N} - \theta^{[N+1]} \,|\breve{Y}_{1:N} = y^*_{1:N}\right) - \tau^2 \breve{\Psi}_N \nabla \breve{\ell}\left(\theta^{[N+1]}\right) \right| < C\tau^4,$$

where

$$\breve{\Psi}_N = \begin{bmatrix} \tau_0^2\Psi & \tau_0^2\Psi & \cdots & \tau_0^2\Psi \\ \tau_0^2\Psi & \tau_0^2+\tau_1^2\Psi & \ddots & \tau_0^2+\tau_1^2\Psi \\ \vdots & \vdots & \ddots & \vdots \\ \tau_0^2\Psi & \tau_0^2+\tau_1^2\Psi & \cdots & \sum_{i=1}^N \tau_i^2\Psi \end{bmatrix}.$$

Note that all assumptions 1-3 are automatically satisfied for the multivariate normal like distribution $\breve{\psi}_N$ of random variable $RZ_{0:N}$. As a result, for a random walk noise we have

$$\left| \nabla \breve{\ell}\left(\theta^{[N+1]}\right) - \tau^{-2}\breve{\Psi}_N^{-1} \breve{\mathbb{E}}\left(\breve{\Theta}_{0:N} - \theta^{[N+1]} \,|\breve{Y}_{1:N} = y^*_{1:N}\right) \right| < C\tau^2.$$

An application of the Gaussian-Jordan inverse method gives,

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

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

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

giving rise to an inequality,

$$\left| \sum_{n=0}^{N} \nabla_n \breve{\ell}\left(\theta^{[N+1]}\right) - \tau^{-2} \sum_{n=0}^{N} \left\{ \breve{\Psi}_N^{-1} \breve{\mathbb{E}}\left(\breve{\Theta}_{0:N} - \theta^{[N+1]} | \breve{Y}_{1:N} = y^*_{1:N}\right) \right\}_n \right| < C\tau^2,$$

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

$$\tau^{-2} \sum_{n=0}^{N} \left\{ \breve{\Psi}_N^{-1} \breve{\mathbb{E}}\left(\breve{\Theta}_{0:N} - \theta^{[N+1]} | \breve{Y}_{1:N} = y^*_{1:N}\right) \right\}_n$$

$$= \tau^{-2} \sum_{n=0}^{N} \mathrm{SumCol}_n(\breve{\Psi}_N^{-1}) \breve{\mathbb{E}}\left(\breve{\Theta}_n - \theta | \breve{Y}_{1:N} = y^*_{1:N}\right), \quad (7)$$

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

$$\left| \sum_{n=0}^{N} \nabla_n \breve{\ell}\left(\theta^{[N+1]}\right) - \tau^{-2}\Psi^{-1}\tau_0^{-2} \breve{\mathbb{E}}\left(\breve{\Theta}_0 - \theta | \breve{Y}_{1:N} = y^*_{1:N}\right) \right| < C\tau^2.$$

This can be written as

$$\left| \nabla \ell(\theta) - \tau^{-2}\Psi^{-1}\tau_0^{-2} \breve{\mathbb{E}}\left(\breve{\Theta}_0 - \theta | \breve{Y}_{1:N} = y^*_{1:N}\right) \right| < C\tau^2.$$

A.2 Proof of Theorem 2

Using similar set up as above, let the random walk noise be $R\tau Z_{0:N}$ with R defined as in equation (6). Assumptions 4 is also satisfied as it is kurtosis property of the multivariate normal like distribution $\breve{\psi}_N$ of random variable $RZ_{0:N}$. From Lemma 2, there exist η and C such that for $0 < \tau < \eta$,

$$\left| \nabla^2 \breve{\ell}\left(\theta^{[N+1]}\right) - \tau^{-4} \left[\breve{\Psi}_N^{-1} \left(\breve{\mathrm{Cov}}_{\theta^{[N+1]},\tau}\left(\breve{\Theta}_{0:N} | \breve{Y}_{1:N} = y^*_{1:N}\right) - \tau^2 \breve{\Psi}_N \right) \breve{\Psi}_N^{-1} \right] \right| < C\tau^2. \quad (8)$$

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

$$\nabla^2_{s,n} \breve{\ell}\left(\theta^{[N+1]}\right) = \frac{\partial^2 \breve{\ell}\left(\theta^{[N+1]}\right)}{\partial \theta_s \partial \theta_n}.$$

Applying the chain rule, we have

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

Adding up term in equation (8) we get

$$\left| \sum_{s=0}^{N} \sum_{n=0}^{N} \nabla^2_{s,n} \breve{\ell}(\theta^{[N+1]}) - \tau^{-4} \sum_{s=0}^{N} \sum_{n=0}^{N} \left[\breve{\Psi}_N^{-1} \left(\breve{\mathrm{Cov}}_{\theta^{[N+1]},\tau}\left(\breve{\Theta}_{0:N} | \breve{Y}_{1:N} = y^*_{1:N}\right) - \tau^2 \breve{\Psi}_N \right) \breve{\Psi}_N^{-1} \right]_{s,n} \right| < C\tau^2,$$

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

$$\left| \nabla^2 \ell(\theta) - \tau^{-4} \sum_{s=0}^{N} \sum_{n=0}^{N} \left[\breve{\Psi}_N^{-1} \left(\breve{\mathrm{Cov}}_{\theta^{[N+1]},\tau}\left(\breve{\Theta}_{0:N} | \breve{Y}_{1:N} = y^*_{1:N}\right) - \tau^2 \breve{\Psi}_N \right) \breve{\Psi}_N^{-1} \right]_{s,n} \right| < C\tau^2.$$

Defining SumCol_n as in equation (7), we have

$$\sum_{s=0}^{N} \sum_{n=0}^{N} \left[\breve{\Psi}_N^{-1} \left(\breve{\mathrm{Cov}}_{\theta^{[N+1]},\tau}\left(\breve{\Theta}_{0:N} | \breve{Y}_{1:N} = y^*_{1:N}\right) - \tau^2 \breve{\Psi}_N \right) \breve{\Psi}_N^{-1} \right]_{s,n}$$
$$= \sum_{s=0}^{N} \sum_{n=0}^{N} \mathrm{SumCol}_s(\breve{\Psi}_N^{-1}) \mathrm{SumCol}_n(\breve{\Psi}_N^{-1}) \times \left(\breve{\mathrm{Cov}}_{\theta^{[N+1]},\tau}\left(\breve{\Theta}_s, \breve{\Theta}_n | \breve{Y}_{1:N} = y^*_{1:N}\right) - \sum_{k=0}^{s\wedge n} \tau_k^2 \Psi \right) \Psi^{-1}$$
$$= \left(\breve{\mathrm{Var}}_{\theta^{[N+1]},\tau}\left(\breve{\Theta}_0 | \breve{Y}_{1:N} = y^*_{1:N}\right) - \tau_0^2 \Psi \right).$$

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

$$\left| \nabla^2 \ell(\theta) - \tau^{-4}\Psi^{-1} \left(\breve{\mathrm{Var}}_{\theta^{[N+1]},\tau}\left(\breve{\Theta}_0 | \breve{Y}_{1:N} = y^*_{1:N}\right) - \tau_0^2 \Psi \right) \Psi^{-1} \right| < C\tau^2.$$

A.3 Proof of Theorem 3

From Lemma 1, we have

$$\left| \nabla \breve{\ell}\left(\theta^{[N+1]}\right) - \tau^{-2} \breve{\Psi}_N^{-1} \breve{\mathbb{E}}\left(\breve{\Theta}_{0:N} - \theta^{[N+1]} \left| \breve{Y}_{1:N} = y^*_{1:N}\right.\right) \right| < C\tau^2. \quad (9)$$

For compactness of notation, we write $E_n = \breve{\mathbb{E}}\left(\breve{\Theta}_n - \theta \left| \breve{Y}_{1:N} = y^*_{1:N}\right.\right)$ and $D_n = \nabla_n \breve{\ell}\left(\theta^{[N+1]}\right)$. Writing out terms of the vector equation in (9) gives

$$(\tau_0^{-2}+\tau_1^{-2})E_0 - \tau_1^{-2}E_1 = \tau^2 D_0 + O(\tau^4), \quad (10)$$
$$-\tau_1^{-2}E_0 + (\tau_1^{-2}+\tau_2^{-2})E_1 - \tau_2^{-2}E_2 = \tau^2 D_1 + O(\tau^4), \quad (11)$$
$$\vdots \quad (12)$$
$$-\tau_{N-1}^{-2}E_{N-2} + (\tau_{N-1}^{-2}+\tau_N^{-2})E_{N-1} - \tau_N^{-2}E_N = \tau^2 D_{N-1} + O(\tau^4), \quad (13)$$
$$-\tau_N^{-2}E_{N-1} + \tau_N^{-2}E_N = \tau^2 D_N + O(\tau^4). \quad (14)$$

Summing up (10) through (14) gives $\tau_0^{-2}E_0 = \tau^2 \breve{\nabla}\ell + O(\tau^4)$, as in Theorem 1. Substituting back into each row of (10) through (14), we

get a set of equations,

$$\begin{aligned}
\tau_0^{-2} E_0 &= \tau^2 \textstyle\sum_{n=0}^{N} D_n + O(\tau^4),\\
\tau_1^{-2}(E_1 - E_0)) &= \tau^2 \textstyle\sum_{n=1}^{N} D_n + O(\tau^4),\\
&\vdots\\
\tau_{N-1}^{-2}(E_{N-1} - E_{N-2}) &= \tau^2 \textstyle\sum_{n=N-1}^{N} D_n + O(\tau^4),\\
\tau_N^{-2}(E_N - E_{N-1}) &= \tau^2 D_N + O(\tau^4).
\end{aligned}$$

Solving for E_n we get

$$\begin{aligned}
E_0 &= \tau^2 \tau_0^2 \textstyle\sum_{n=0}^{N} D_n + O(\tau^4),\\
E_1 &= \tau^2 \left(\tau_0^2 \textstyle\sum_{n=0}^{N} D_n + \tau_1^2 \sum_{n=1}^{N} D_n\right) + O(\tau^4),\\
&\vdots\\
E_{N-1} &= \tau^2 \left(\tau_0^2 \textstyle\sum_{n=0}^{N} D_n + \tau_1^2 \sum_{n=1}^{N} D_n + \ldots + \tau_{N-1}^2 \sum_{n=N-1}^{N} D_n\right) + O(\tau^4),\\
E_N &= \tau^2 \left(\tau_0^2 \textstyle\sum_{n=0}^{N} D_n + \tau_1^2 \sum_{n=1}^{N} D_n + \ldots + \tau_N^2 D_N\right) + O(\tau^4).
\end{aligned}$$

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

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

Application of Theorem 1 then completes the proof.

A.4 Proof of Theorem 4

From Lemma 2, we have

$$\left| \nabla^2 \check{\ell}\left(\theta^{[N+1]}\right) - \tau^{-4}\left[\check{\Psi}_N^{-1}\left(\check{\mathrm{Cov}}_{\theta^{[N+1]},\tau}\left(\check{\Theta}_{0:N} | \check{Y}_{1:N} = y^*_{1:N}\right) - \tau^2 \check{\Psi}_N\right)\check{\Psi}_N^{-1}\right] \right| < C\tau^2.$$

For compact notation, we write

$$\check{Cov}_{s,n} = \check{\mathrm{Cov}}\left(\check{\Theta}_s, \check{\Theta}_n | \check{Y}_{1:N} = y^*_{1:N}\right) - \tau_s \tau_n \tau^2 \Psi$$

and

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

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

$$\check{\mathrm{Cov}}_{0,0} = \tau^4 \left[\check{\Psi}_N \nabla^2 \check{\ell}\left(\theta^{[N+1]}\right)\check{\Psi}_N\right]_{0,0} + O(\tau^6), \tag{15}$$

$$\check{\mathrm{Cov}}_{1,1} = \tau^4 \left[\check{\Psi}_N \nabla^2 \check{\ell}\left(\theta^{[N+1]}\right)\check{\Psi}_N\right]_{1,1} + O(\tau^6), \tag{16}$$

$$\vdots \tag{17}$$

$$\check{\mathrm{Cov}}_{N-1,N-1} = \tau^4 \left[\check{\Psi}_N \nabla^2 \check{\ell}\left(\theta^{[N+1]}\right)\check{\Psi}_N\right]_{N-1,N-1} + O(\tau^6), \tag{18}$$

$$\check{\mathrm{Cov}}_{N,N} = \tau^4 \left[\check{\Psi}_N \nabla^2 \check{\ell}\left(\theta^{[N+1]}\right)\check{\Psi}_N\right]_{N,N} + O(\tau^6). \tag{19}$$

Using (15) through (19), and expanding out a matrix multiplication, we get

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

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

$$\check{\mathrm{Cov}}_{n,n} = \check{\mathrm{Cov}}_{0,0} + O(\tau^6),$$

from which we can conclude that

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

An application of Theorem 2 then completes the proof.

References

Andrieu, C., Doucet, A., and Holenstein, R. (2010). Particle Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(3):269–342.

Bhadra, A., Ionides, E. L., Laneri, K., Pascual, M., Bouma, M., and Dhiman, R. C. (2011). Malaria in Northwest India: Data analysis via partially observed stochastic differential equation models driven by Lévy noise. *Journal of the American Statistical Association*, 106:440–451.

Blackwood, J. C., Cummings, D. A. T., Broutin, H., Iamsirithaworn, S., and Rohani, P. (2013a). Deciphering the impacts of vaccination and immunity on pertussis epidemiology in Thailand. *Proceedings of the National Academy of Sciences of the USA*, 110:9595–9600.

Blackwood, J. C., Streicker, D. G., Altizer, S., and Rohani, P. (2013b). Resolving the roles of immunity, pathogenesis, and immigration for rabies persistence in vampire bats. *Proceedings of the National Academy of Sciences of the USA*, 110:2083720842.

Blake, I. M., Martin, R., Goel, A., Khetsuriani, N., Everts, J., Wolff, C., Wassilak, S., Aylward, R. B., and Grassly, N. C. (2014). The role of older children and adults in wild poliovirus transmission. *Proceedings of the National Academy of Sciences of the USA*, 111(29):10604–10609.

Bretó, C., He, D., Ionides, E. L., and King, A. A. (2009). Time series analysis via mechanistic models. *Annals of Applied Statistics*, 3:319–348.

Camacho, A., Ballesteros, S., Graham, A. L., Carrat, F., Ratmann, O., and Cazelles, B. (2011). Explaining rapid reinfections in multiple-wave influenza outbreaks: Tristan da Cunha 1971 epidemic as a case study. *Proceedings of the Royal Society of London, Series B*, 278(1725):3635–3643.

Chopin, N., Jacob, P. E., and Papaspiliopoulos, O. (2013). SMC^2: an efficient algorithm for sequential analysis of state space models. *Journal of the Royal Statistical Society, Series B (Statistical Methodology)*, 75(3):397–426.

Dahlin, J., Lindsten, F., and Schön, T. B. (2015). Particle Metropolis-Hastings using gradient and Hessian information. *Statistics and Computing*, 25(1):81–92.

Doucet, A., Jacob, P. E., and Rubenthaler, S. (2013). Derivative-free estimation of the score vector and observed information matrix with application to state-space models. *ArXiv:1304.5768*.

Doucet, A. and Johansen, A. M. (2009). A tutorial on particle filtering and smoothing: Fifteen years later. In Crisan, D. and Rozovskii, B., editors, *Handbook of Nonlinear Filtering*, pages 656–704. Oxford University Press, Oxford.

Earn, D. J., He, D., Loeb, M. B., Fonseca, K., Lee, B. E., and Dushoff, J. (2012). Effects of school closure on incidence of pandemic influenza in Alberta, Canada. *Annals of Internal Medicine*, 156(3):173–181.

He, D., Dushoff, J., Day, T., Ma, J., and Earn, D. J. D. (2013). Inferring the causes of the three waves of the 1918 influenza pandemic in England and Wales. *Proceedings of the Royal Society of London, Series B*, 280(1766):20131345.

He, D., Ionides, E. L., and King, A. A. (2010). Plug-and-play inference for disease dynamics: Measles in large and small populations as a case study. *Journal of the Royal Society Interface*, 7(43):271–283.

Ionides, E. L., Bhadra, A., Atchadé, Y., and King, A. (2011). Iterated filtering. *Annals of Statistics*, 39:1776–1802.

Ionides, E. L., Bretó, C., and King, A. A. (2006). Inference for nonlinear dynamical systems. *Proceedings of the National Academy of Sciences of the USA*, 103:18438–18443.

Ionides, E. L., Nguyen, D., Atchadé, Y., Stoev, S., and King, A. A. (2015). Inference for dynamic and latent variable models via iterated, perturbed Bayes maps. *Proceedings of the National Academy of Sciences of the USA*, 112(3):719–724.

Kevrekidis, I. G., Gear, C. W., and Hummer, G. (2004). Equation-free: The computer-assisted analysis of complex, multiscale systems. *American Institute of Chemical Engineers Journal*, 50:1346–1354.

King, A. A., Domenech de Celle, M., Magpantay, F. M. G., and Rohani, P. (2015a). Avoidable errors in the modelling of outbreaks of emerging pathogens, with special reference to Ebola. *Proceedings of the Royal Society of London, Series B*, 282:20150347.

King, A. A., Ionides, E. L., Pascual, M., and Bouma, M. J. (2008). Inapparent infections and cholera dynamics. *Nature*, 454:877–880.

King, A. A., Nguyen, D., and Ionides, E. L. (2015b). Statistical inference for partially observed Markov processes via the R package pomp. *Journal of Statistical Software*, to appear.

Kushner, H. J. and Clark, D. S. (1978). *Stochastic Approximation Methods for Constrained and Unconstrained Systems*. Springer-Verlag, New York.

Laneri, K., Bhadra, A., Ionides, E. L., Bouma, M., Dhiman, R. C., Yadav, R. S., and Pascual, M. (2010). Forcing versus feedback: Epidemic malaria and monsoon rains in Northwest India. *PLoS Computational Biology*, 6(9):e1000898.

Laneri, K., Paul, R. E., Tall, A., Faye, J., Diene-Sarr, F., Sokhna, C., Trape, J.-F., and Rodó, X. (2015). Dynamical malaria models reveal how immunity buffers effect of climate variability. *Proceedings of the National Academy of Sciences of the USA*, 112(28):8786–8791.

Lavine, J. S., King, A. A., Andreasen, V., and Bjrnstad, O. N. (2013). Immune boosting explains regime-shifts in prevaccine-era pertussis dynamics. *PLoS ONE*, 8(8):e72086.

Lavine, J. S. and Rohani, P. (2012). Resolving pertussis immunity and vaccine effectiveness using incidence time series. *Expert Review of Vaccines*, 11:1319–1329.

Macdonald, G. (1957). *The Epidemiology and Control of Malaria*. Oxford University Press, Oxford.

Martinez-Bakker, M., King, A. A., and Rohani, P. (2015). Unraveling the transmission ecology of polio. *PLoS Biology*, 13(6):e1002172.

Nemeth, C., Fearnhead, P., and Mihaylova, L. (2013). Particle approximations of the score and observed information matrix for parameter estimation in state space models with linear computational cost. *ArXiv:1306.0735*.

Nguyen, D. and Ionides, E. L. (2015). Iterated smoothing for partially observed Markov processes via the R package is2. *R Packages*.

Olsson, J., Cappé, O., Douc, R., and Moulines, E. (2008). Sequential Monte Carlo smoothing with application to parameter estimation in nonlinear state space models. *Bernoulli*, 14(1):155–179.

Poyiadjis, G., Doucet, A., and Singh, S. S. (2011). Particle approximations of the score and observed information matrix in state space models with application to parameter estimation. *Biometrika*, 98(1):65–80.

Romero-Severson, E., Volz, E., Koopman, J., Leitner, T., and Ionides, E. (2015). Dynamic variation in sexual contact rates in a cohort of HIV-negative gay men. *American Journal of Epidemiology*, page kwv044.

Ross, R. (1910). *The Prevention of Malaria*. Dutton, Boston.

Roy, M., Bouma, M. J., Ionides, E. L., Dhiman, R. C., and Pascual, M. (2013). The potential elimination of plasmodium vivax malaria by relapse treatment: Insights from a transmission model and surveillance data from NW India. *PLoS Neglected Tropical Diseases*, 7(1):e1979.

Shrestha, S., Foxman, B., Weinberger, D. M., Steiner, C., Viboud, C., and Rohani, P. (2013). Identifying the interaction between influenza and pneumococcal pneumonia using incidence data. *Science Translational Medicine*, 5(191):191ra84.

Shrestha, S., King, A. A., and Rohani, P. (2011). Statistical inference for multi-pathogen systems. *PLoS Computational Biology*, 7(8):e1002135.

Sisson, S. A., Fan, Y., and Tanaka, M. M. (2007). Sequential Monte Carlo without likelihoods. *Proceedings of the National Academy of Sciences of the USA*, 104(6):1760–1765.

Spall, J. C. (2003). *Introduction to Stochastic Search and Optimization*. Wiley, Hoboken.

Toni, T., Welch, D., Strelkowa, N., Ipsen, A., and Stumpf, M. P. (2009). Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems. *Journal of the Royal Society Interface*, 6:187–202.

Whiteley, N. et al. (2013). Stability properties of some particle filters. *The Annals of Applied Probability*, 23(6):2500–2537.

Wood, S. N. (2010). Statistical inference for noisy nonlinear ecological dynamic systems. *Nature*, 466(7310):1102–1104.