Iterated Filtering

Edward L. Ionides[134], Anindya Bhadra[1] and Aaron King[23]

February 2, 2009

The University of Michigan, departments of [1]Statistics and [2]Ecology and Evolutionary Biology. [3]Fogarty International Center, National Institutes of Health. [4]Address for correspondence: 1085 South University Ave, Ann Arbor MI 48109-1107, USA; e-mail ionides@umich.edu

Abstract

Inference for partially observed Markov process models has been a longstanding methodological challenge with many scientific and engineering applications. Iterated filtering algorithms maximize the likelihood function for partially observed Markov process models by solving a recursive sequence of filtering problems. We present new theoretical results pertaining to the convergence of iterated filtering algorithms implemented via sequential Monte Carlo filters. This theory complements the growing body of empirical evidence that iterated filtering algorithms provide an effective inference strategy for scientific models that have been intractable via alternative methodologies.

1 Introduction

Partially observed Markov processes are of widespread importance throughout science and engineering. As such, they have been studied under various names including *state space models* (Durbin and Koopman, 2001), *dynamic models* (West and Harrison, 1997) and *hidden Markov models* (Cappé et al., 2005). Applications include ecology (Newman et al., 2008), economics (Fernández-Villaverde and Rubrio-Ramírez, 2007), epidemiology (King et al., 2008), finance (Johannes et al., 2009), meteorology (Anderson and Collins, 2007), neuroscience (Ergun et al., 2007) and target tracking (Godsill et al., 2007). One central and extensively studied issue is the reconstruction of unobserved components of the Markov process from the available observations. Reconstructing the current state of the process (i.e., determining or approximating its conditional distribution given all current and previous observations) is known as filtering (Anderson and Moore, 1979; Arulampalam et al., 2002). Oftentimes one also wishes to draw inferences on unknown model parameters from data; we call these *static parameters* when we wish to distinguish them from the time-varying components of the Markov process.

A successful numerical solution to the filtering problem enables evaluation of the likelihood function and therefore brings one tantalizingly close to efficient estimation of static parameters via likelihood-based approaches, either Bayesian or non-Bayesian. However, numerical instabilities typically arise which have inspired a considerable literature (Kitagawa, 1998; Liu and West, 2001; Storvik, 2002; Ionides et al., 2006; Toni et al., 2008; Polson et al., 2008). As a generalization, the numerical complications derive from difficulties maximizing or numerically integrating a computationally intensive approximation to the likelihood function with the possible additional concern of

Monte Carlo variability. In this article, we investigate approaches which iteratively carry out a filtering procedure to explore the likelihood surface at increasingly local scales in search of a maximum of the likelihood function. Such methodology, called *iterated filtering*, has been shown capable of addressing state-of-the-art inference challenges (Ionides et al., 2006; King et al., 2008; Bretó et al., 2009; He et al., 2009). In Section 2 we develop new theoretical results for iterated filtering. The previous theoretical foundation for iterated filtering, presented by Ionides et al. (2006), did not engage directly in the Monte Carlo issues relating to practical implementation of the methodology. It is relatively easy to check that a (local) maximum has been attained, and therefore one can view the theory of Ionides et al. (2006) as motivation for an algorithm whose capabilities were proven by demonstration. However, the more complete theory presented here gives additional insights into the capabilities, limitations and practical implementation of iterated filtering.

Not all estimation techniques for static parameters are based on solving the filtering problem. Stochastic expectation-maximization and Markov Chain Monte Carlo approaches (Cappé et al., 2005) work instead with the conditional distribution of entire trajectories of the Markov process given the observations. Some estimation methods entirely avoid explicit computation of these conditional distributions (Kendall et al., 1999; Reuman et al., 2006; Toni et al., 2008) but thereby lose the generality and statistical efficiency of likelihood-based inference frameworks. Further discussion of the relative merits of iterated filtering compared to other available methodology is postponed to the discussion in Section 3. Proofs of the theorems stated in Section 2 are given in Section 4.

2 Notation and main results

Let $\{x(t), t \in T\}$ be a Markov process taking values in $\mathbb{R}^{d_x}$ (Rogers and Williams, 1994). The time index set $T \subset \mathbb{R}$ may be an interval or a discrete set, but we are primarily concerned with a finite subset of times $t_1 < t_2 < \cdots < t_N$ at which $x(t)$ is observed, together with some initial time $t_0 < t_1$. We write $x_{0:N} = (x_0, \ldots, x_N) = \big(x(t_0), \ldots, x(t_N)\big)$. We correspondingly denote the observation process by $y_{1:N} = (y_1, \ldots, y_N)$, with y_n taking a value in $\mathbb{R}^{d_y}$. We assume the existence of all required joint and conditional densities for $x_{0:N}$ and $y_{1:N}$. These densities are supposed to depend on an unknown parameter vector θ taking a value in $\mathbb{R}^{d_\theta}$. A partially observed Markov model may then be specified at times $t_{0:N}$ by an initial density $f(x_0 \,|\, \theta)$, conditional transition densities $f(x_n \,|\, x_{n-1}, \theta)$ for $1 \le n \le N$, and the conditional densities of the observation process which have the form $f(y_n \,|\, y_{1:n-1}, x_{1:n}, \theta) = f(y_n \,|\, x_n, \theta)$. This notation overloads $f(\cdot \,|\, \cdot)$ as a generic density which is specified by its arguments, and suppresses the distinction between random variables and their realizations. The generic function f gives a notationally compact representation of a partially observed Markov model, which can be rigorously formalized. Appendix A provides further discussion of generic function notation.

Iterated filtering involves introducing a sequence of approximations to the model f in which a time-varying parameter process $\{\theta_n, 0 \le n \le N\}$ is introduced. Specifically, equations (1–3) define a model g for a Markov process $\{(x_n, \theta_n), 0 \le n \le N\}$ and observation process $y_{1:N}$. Assuming f is continuously parameterized as a function of θ, we see from (1–3) that $g(x_{0:N}, y_{1:N} \,|\, \theta, \sigma, \tau)$ approaches $f(x_{0:N}, y_{1:N} \,|\, \theta)$ as both $\sigma \to 0$ and $\tau \to 0$.

$$g(x_n, \theta_n \,|\, x_{n-1}, \theta_{n-1}, \theta, \sigma, \tau) = f(x_n \,|\, x_{n-1}, \theta_{n-1}) \, \tfrac{1}{\sigma} \, \kappa\Big(\tfrac{\theta_n - \theta_{n-1}}{\sigma}\Big), \tag{1}$$

$$g(y_n \,|\, x_n, \theta_n, \theta, \sigma, \tau) = f(y_n \,|\, x_n, \theta_n), \tag{2}$$

$$g(x_0, \theta_0 \,|\, \theta, \sigma, \tau) = f(x_0 \,|\, \theta_0) \, \tfrac{1}{\tau} \, \kappa\Big(\tfrac{\theta_n - \theta_{n-1}}{\tau}\Big). \tag{3}$$

Here, κ is a probability density function on $\mathbb{R}^{d_\theta}$ which specifies a random walk for θ_n. From (1), the increments of the random walk are independent of the current state of the process x_n. We suppose that the distribution corresponding to κ has mean zero and covariance matrix Σ, so that

$$E[\theta_n \,|\, \theta_{n-1}, \theta, \sigma, \tau] = \theta_{n-1}, \qquad \mathrm{Var}(\theta_n \,|\, \theta_{n-1}, \theta, \sigma, \tau) = \sigma^2\Sigma, \tag{4}$$

$$E[\theta_0 \,|\, \theta, \sigma, \tau] = \theta, \qquad \mathrm{Var}(\theta_0 \,|\, \theta, \sigma, \tau) = \tau^2\Sigma. \tag{5}$$

Reparameterization of f may be required to ensure that θ can take all values in $\mathbb{R}^{d_\theta}$. In practice, κ is typically a multivariate normal density, which must be truncated to meet condition (A4) of Theorem 1. Previous theory for iterated filtering (Ionides et al., 2006) did not require a scale family $g(\theta_n \,|\, \theta_{n-1}, \sigma) = \sigma^{-1}\kappa\big(\sigma^{-1}[\theta_n - \theta_{n-1}]\big)$ for the time-varying distribution of θ_n. However, this specification is natural and will be shown to lead to more transparent convergence conditions. We refer to σ, τ, κ and Σ as *algorithmic parameters* since they play a role in the iterated filtering algorithm but are not part of the statistical model specified by f. The choice of algorithmic parameters may affect the numerical efficiency of iterated filtering algorithms, but does not affect the resulting statistical conclusions.

We define the log likelihood function to be $\ell(\theta) = \log f(y_{1:N} \,|\, \theta)$. We write ∇ for a vector of partial derivatives with respect to each component of θ, and ∇^2 for the Hessian matrix of second partial derivatives. A result underpinning iterated filtering is that $\nabla\ell(\theta)$ can be approximated in terms of moments of the filtering distributions for g. Specifically, the following Theorem 1 relates this derivative to the filtering means and prediction variances for g, defined as

$$\begin{aligned} \theta^F_n &= \theta^F_n(\theta, \sigma, \tau) = E[\theta_n \,|\, y_{1:n}, \theta, \sigma, \tau] \\ V^P_n &= V^P_n(\theta, \sigma, \tau) = \mathrm{Var}(\theta_n \,|\, y_{1:n-1}, \theta, \sigma, \tau) \end{aligned} \tag{6}$$

for $n = 1, \ldots, N$, with $\theta^F_0 = \theta$. We assume the regularity conditions (A1–A4) below, with $|\cdot|$ denoting the absolute value of a vector or the largest absolute eigenvalue of a square matrix.

(A1) There is a constant $C_1(\theta)$ such that $0 < f(\cdot \,|\, \cdot, \theta) < C_1(\theta)$ for all joint and conditional densities of $x_{0:N}$ and $y_{1:N}$. Additionally, $C_1(\theta)$ is bounded on compact subsets of $\mathbb{R}^{d_\theta}$.

(A2) $f(\cdot \,|\, \cdot, \theta)$ is twice continuously differentiable with respect to θ. Further,

$$|\nabla f(y_n \,|\, x_n, \theta)| < \infty, \qquad |\nabla^2 f(y_n \,|\, x_n, \theta)| < \infty,$$

$$\int \big|\nabla f(x_{n+1} \,|\, x_n, \theta)\big|\, dx_{n+1} < \infty, \qquad \int \big|\nabla^2 f(x_{n+1} \,|\, x_n, \theta)\big|\, dx_{n+1} < \infty,$$

$$\int \big|\nabla f(x_0 \,|\, \theta)\big|\, dx_0 < \infty, \qquad \int \big|\nabla^2 f(x_0 \,|\, \theta)\big|\, dx_0 < \infty,$$

with these bounds being uniform over all x_n and y_n with θ ranging over any compact subset of $\mathbb{R}^{d_\theta}$.

(A3) $\kappa(\theta)$ is twice continuously differentiable, with $\int |\nabla^2\kappa(\theta)|\, d\theta < \infty$.

(A4) There is a constant C_2 with $\kappa(\theta) = 0$ for $|\theta| \geq C_2$ and $\kappa(\theta) > 0$ for $|\theta| < C_2$.

The conditions (A1) and (A2) are not restrictive. Conditions (A3) and (A4) can be satisfied by the choice of the algorithmic parameters. The assumption of a spherical support for κ in (A4) is mathematically convenient but we believe this requirement could be relaxed to some more general assumption of compact support.

Theorem 1. *Suppose conditions (A1–A4). Let σ be a function of τ with $\sigma\tau^{-1} \to 0$ as $\tau \to 0$. Using notation from (6),*

$$\lim_{\tau\to 0} \sum_{n=1}^{N} \big(V_n^P\big)^{-1} \big(\theta_n^F - \theta_{n-1}^F\big) = \nabla \ell(\theta). \tag{7}$$

A proof of Theorem 1 is given in Section 4.1, based on a Taylor series expansion of $g(y_n \,|\, y_{1:n-1}, \theta_n, \theta, \sigma, \tau)$ around $\theta_n = \theta_{n-1}^F$. Theorem 1 is based on a result of Ionides et al. (2006), however both the assumptions employed and the details of the proof differ substantially from this previous work.

The quantities θ_n^F and V_n^P in Theorem 1 do not usually have closed form, and so numerical approximations must be made for any practical application of this result. Numerical approximation of moments is generally more convenient than approximating derivatives, and this is the reason that Theorem 1 may be useful. However, on might suspect that there is no "free lunch" and therefore the numerical calculation of the left hand side of (7) should become fragile as σ and τ becomes small. We will see that this is indeed the case, but that iterated filtering methods mitigate the difficulty to some extent by averaging numerical error over subsequent iterations. To be concrete, we suppose henceforth that numerical filtering will be carried out using the basic sequential Monte Carlo (SMC) method presented as Algorithm 1. SMC provides a flexible and widely used class of filtering algorithms, with many variants designed to improve numerical efficiency (Cappé et al., 2007). The relatively simple SMC method in Algorithm 1 is more readily comprehended, analyzed and implemented. It has also been found adequate for previous data analyses using iterated filtering (Ionides et al., 2006; King et al., 2008; Bretó et al., 2009; He et al., 2009). We suspect that the qualitative conclusions obtained here would apply to variations on Algorithm 1.

Input:
parameter vector, ψ
observations $y_{1:N}$ and times $t_{0:N}$
generic density $h(\cdot \,|\, \cdot, \psi)$ for $y_{1:N}$ and an unobserved process $z_{0:N}$
number of particles, J

Procedure:
1 initialize filter particles $Z_{0,j}^F \sim h(z_0 \,|\, \psi)$ for j in $1:J$
2 for n in $1:N$
3 for j in $1:J$ draw prediction particles $Z_{n,j}^P \sim h\big(z_n \,|\, z_{n-1}{=}Z_{n-1,j}^F, \psi\big)$
4 set $w(n,j) = h\big(y_n \,|\, z_n{=}Z_{n,j}^P, \psi\big)$
5 draw $k_1, \ldots, k_J$ such that $\text{Prob}(k_j{=}i) = w(n,i)/\sum_\ell w(n,\ell)$
6 set $Z_{n,j}^F = Z_{n,k_j}^P$
7 end for

Algorithm 1: A basic sequential Monte Carlo procedure for a discrete-time Markov process $\{z_n\}$ with generic density function h. In the current context, z_n will be either x_n or (x_n, θ_n) with h correspondingly set to f or g respectively. The resampling in step 5 is taken to follow a multinomial distribution to build on previous theoretical results making this assumption (Del Moral and Jacod, 2001; Crisan and Doucet, 2002). An alternative is the systematic procedure in Arulampalam et al. (2002, Algorithm 2) which has less Monte Carlo variability; we support the use of systematic sampling in practice and we suppose that all our results would continue to hold in such situations.

To calculate Monte Carlo estimates of the quantities in (6), we apply Algorithm 1 to the

model g with $z_n = (x_n, \theta_n)$, $\psi = (\theta, \sigma, \tau)$ and J particles. We write $Z^F_{n,j} = \big(X^F_{n,j}, \Theta^F_{n,j}\big)$ and $Z^P_{n,j} = \big(X^P_{n,j}, \Theta^P_{n,j}\big)$ for the Monte Carlo samples from the filtering and prediction calculations in Algorithm 1. Then, using x' to denote the transpose of x, we define

$$\begin{aligned} \widetilde{\theta}^F_n &= \widetilde{\theta}^F_n(\theta, \sigma, \tau, J) = \tfrac{1}{J} \textstyle\sum_{j=1}^J \Theta^F_{n,j}, \\ \widetilde{V}^P_n &= \widetilde{V}^P_n(\theta, \sigma, \tau, J) = \tfrac{1}{J-1} \textstyle\sum_{j=1}^J \big(\Theta^P_{n,j} - \widetilde{\theta}^F_{n-1}\big)\big(\Theta^P_{n,j} - \widetilde{\theta}^F_{n-1}\big)'. \end{aligned} \tag{8}$$

We now present an analogue to Theorem 1 in which the filtering means and prediction variances are replaced by their Monte Carlo counterparts. A proof of this result is given in Section 4.3. The stochasticity in Theorem 1 is due to Monte Carlo variability, conditional on the data $y_{1:N}$, and we write $\widetilde{E}$ and $\widetilde{\mathrm{Var}}$ to denote Monte Carlo means and variances. The Monte Carlo random variables required to implement Algorithm 1 are presumed to be drawn independently each time the algorithm is evaluated.

Theorem 2. *Let $\{\sigma_m\}$, $\{\tau_m\}$ and $\{J_m\}$ be positive sequences with $\tau_m \to 0$, $\sigma_m \tau_m^{-1} \to 0$ and $\tau_m J_m \to \infty$. Define $\widetilde{\theta}^F_{n,m} = \widetilde{\theta}^F_n(\theta, \sigma_m, J_m)$ and $\widetilde{V}^P_{n,m} = \widetilde{V}^P_{n,m}(\theta, \sigma_m, J_m)$ via (8). Supposing conditions (A1–A4),*

$$\lim_{m\to\infty} \widetilde{E}\Big[\sum_{n=1}^N \big(\widetilde{V}^P_{n,m}\big)^{-1} \big(\widetilde{\theta}^F_{n,m} - \widetilde{\theta}^F_{n-1,m}\big)\Big] \;=\; \nabla \ell(\theta), \tag{9}$$

$$\lim_{m\to\infty} \tau_m^2 J_m \widetilde{\mathrm{Var}}\Big(\sum_{n=1}^N \big(\widetilde{V}^P_{n,m}\big)^{-1} \big(\widetilde{\theta}^F_{n,m} - \widetilde{\theta}^F_{n-1,m}\big)\Big) \;<\; \infty, \tag{10}$$

with convergence being uniform for θ in compact sets.

Theorem 2 suggests that a Monte Carlo method which leans on Theorem 1 will require a sequence of Monte Carlo sample sizes, J_m, which increases faster than τ_m^{-1}. Otherwise, the Monte Carlo bias in estimating $\theta^F_n - \theta^F_{n-1}$, which is of order τ_m / J_m, will eventually dominate the information in $\theta^F_n - \theta^F_{n-1}$ about $\nabla \ell(\theta)$, which is of order τ_m^2. Even with $\tau_m J_m \to \infty$, we see from (10) that the estimated derivative in (9) may have increasing Monte Carlo variability as $m \to \infty$. This trade-off between bias and variance is to be expected in any Monte Carlo numerical derivative, a classic example being the Kiefer-Wolfowitz algorithm (Kiefer and Wolfowitz, 1952; Spall, 2003). Algorithms which are designed to balance such trade-offs have been extensively studied under the label of *stochastic approximation* (Kushner and Yin, 2003; Spall, 2003).

Theorem 3 gives an example of a stochastic approximation procedure, defined by the recursive sequence $\hat{\theta}_m$ in (11). Because each step of this recursion involves an application of the filtering procedure in Algorithm 1, we call (11) an iterated filtering algorithm. To prove the convergence of this algorithm to a value $\hat{\theta}$ maximizing the log likelihood function $\ell(\theta)$ we make the following assumptions, which are standard sufficient conditions for stochastic approximation methods.

(B1) Define $Z(t)$ to be a solution to $dZ/dt = \nabla \ell(Z(t))$. Suppose that $\hat{\theta}$ is an *asymptotically stable equilibrium point*, meaning that (i) for every $\eta > 0$ there exists a $\delta(\eta)$ such that $|Z(t) - \hat{\theta}| \le \eta$ for all $t > 0$ whenever $|Z(0) - \hat{\theta}| \le \delta$, and (ii) there exists a δ_0 such that $Z(t) \to \hat{\theta}$ as $t \to \infty$ whenever $|Z(0) - \hat{\theta}| \le \delta_0$.

(B2) With probability one, $\sup_m |\hat{\theta}_m| < \infty$. Further, $\hat{\theta}_m$ falls infinitely often into a compact subset of $\{Z(0) : \lim_{t\to\infty} Z(t) = \hat{\theta}\}$.

Although neither (B1–B2) nor alternative sufficient conditions (Spall, 2003, Chapter 4) are easy to verify, stochastic approximation methods have nevertheless been found effective in many situations. Condition (B2) is most readily satisfied if $\hat{\theta}_m$ is constrained to a neighborhood in which $\hat{\theta}$ is a unique local maximum, which gives a guarantee of local rather than global convergence. Global convergence results have been obtained for related stochastic approximation procedures (Maryak and Chin, 2008) but are beyond the scope of this current paper. Practical implementation issues are discussed in Section 3 below.

Theorem 3. *Let $\{a_m\}$, $\{\sigma_m\}$, $\{\tau_m\}$ and $\{J_m\}$ be positive sequences with $\tau_m \to 0$, $\sigma_m \tau_m^{-1} \to 0$, $J_m \tau_m \to \infty$, $a_m \to 0$, $\sum_m a_m = \infty$ and $\sum_m a_m^2 J_m^{-1} \tau_m^{-2} < \infty$. Specify a recursive sequence of parameter estimates $\{\hat{\theta}_m\}$ by*

$$\hat{\theta}_{m+1} = \hat{\theta}_m + a_m \sum_{n=1}^{N} \big(\widetilde{V}_{n,m}^P\big)^{-1} \big(\widetilde{\theta}_{n,m}^F - \widetilde{\theta}_{n-1,m}^F\big), \tag{11}$$

where $\widetilde{\theta}_{n,m}^F = \widetilde{\theta}_n^F(\hat{\theta}_m, \sigma_m, J_m)$ and $\widetilde{V}_{n,m}^P = \widetilde{V}_{n,m}^P(\hat{\theta}_m, \sigma_m, J_m)$ are defined in (8) via an application of Algorithm 1. Assuming conditions (A1–A4) and (B1–B2), $\lim_{m\to\infty} \hat{\theta}_m = \hat{\theta}$ with probability one.

The proof of Theorem 3, given in Section 4.5, is based on applying Theorem 2 in the context of existing results on stochastic approximation. The rate assumptions in Theorem 3 are satisfied, for example, by $a_m = m^{-1}$, $\tau_m^2 = m^{-1}$, $\sigma_m^2 = m^{-(1+\delta)}$ and $J_m = m^{(\delta+1/2)}$ for $\delta > 0$.

3 Discussion of the theory and practice of iterated filtering

One of the attractive features of the iterated filtering algorithm in Theorem 3 is that it depends on the unobserved Markov process only through generation of sample paths. In particular, this procedure can be implemented in situations where an expression for transition densities is unavailable. This property has been called plug-and-play (Bretó et al., 2009; He et al., 2009) since it permits simulation code to be simply plugged into the inference procedure, enabling scientists to analyze multiple alternative models with only minor changes to the computations involved. The iterated filtering algorithm in Theorem 3 inherits the plug-and-play property from Algorithm 1. Alternative implementations of iterated filtering, for example making use of computationally efficient variations on sequential Monte Carlo (Cappé et al., 2007), do not generally possess the plug-and-play property.

Other plug-and-play methods proposed for partially observed Markov models include artificial parameter evolution method (Liu and West, 2001), an approximate Bayesian computation (ABC) approach (Toni et al., 2008), and simulation-based forecasting (Kendall et al., 1999). In all these methods, statistical efficiency is sacrificed at the altar of computational convenience. For example, the stochasticity added for the artificial parameter evolution of Liu and West (2001) dilutes the influence of the earlier observations in the time series; ABC methods work only with a low-dimensional summary statistic of the data; non-likelihood-based methods such as least square prediction minimization (Kendall et al., 1999) are also generally statistically efficient. Iterated filtering, on the other hand, has the efficiency guarantee of maximum likelihood estimation. Of course, the resulting estimates are still subject to Monte Carlo variability due to the infeasibility of attaining the infinite limit $m \to \infty$. Ultimately, the value of all such asymptotic theory is dependent on its finite sample relevance.

For challenging numerical computations, there is often a gap between available theorems and practical techniques. A classic example of this is optimization by simulated annealing, a popular

stochastic optimization technique (Kirkpatrick et al., 1983; Spall, 2003) which draws on physical insights from statistical mechanics and mathematical foundations from Markov chain theory. Theoretically motivated convergence rates for simulated annealing are often too slow for practical implementation while variations on the algorithm with less attractive theory have been found to be effective (Ingber, 1993). Although there are substantial differences between simulated annealing and iterated filtering (e.g. global versus local theory, exact versus stochastic objective functions), the similarities between these two stochastic search algorithms nevertheless provide a worthwhile comparison. To relate simulated annealing and iterated filtering, it is helpful to adopt from simulated annealing an analogy whereby σ_m and τ_m are thought of as temperatures which approaching freezing as $\sigma_m \to 0$ and $\tau \to 0$. If the temperature cools sufficiently slowly, iterated filtering and simulated annealing theoretically approach the maximum of their respective target functions. In practice, quicker cooling schedules are used for simulated annealing, in which case it is more properly called simulated quenching (Ingber, 1993). Periodically increasing the the temperature, by chaining together quenched searches, is known as simulated tempering and can lead to a reasonable trade-off between investigating fine scale and larger scale structure of the objective function. It is generally possible to confirm the success of an optimization procedure by running it from multiple widely separated starting points, which makes possible post-hoc validation of the search strategy. Our experience suggests that tempered searches are an effective technique for iterated filtering. In addition, the rounds of quenching provide a sequence of parameter estimates which are useful for learning about the structure of the likelihood surface.

Likelihood maximization provides not just point estimates of unknown parameters but also confidence intervals, either through profile likelihood calculations or Hessian approximations, and likelihood ratio tests of competing hypotheses. The interested reader is referred elsewhere for case studies demonstrating practical implementations of iterated filtering (King et al., 2008; Bretó et al., 2009; He et al., 2009; Ionides et al., 2006). These practical implementations did not employ the increasing Monte Carlo sample size suggested by Theorem 3 and used a constant ratio $\sigma_m \tau_m^{-1}$ rather than a sequence tending to zero. Nevertheless, they were shown to be capable of maximizing complex likelihood surfaces to an adequate level of accuracy. Since SMC can provide an unbiased estimate of the likelihood function (see Corollary 1 in Section 4.3) it is relatively straightforward to confirm whether the likelihood has indeed been successfully maximized.

The incorporation of iterated filtering into the framework of stochastic approximation, which underlies the proof of Theorem 3, suggests several avenues for further investigation. Existing modifications of stochastic approximation techniques (Spall, 2003) include: (i) averaging parameter estimates across iterations; (ii) breaking down high-dimensional problems into a sequence of randomly selected lower dimensional problems; (iii) making use of a plug-and-play estimate of second partial derivatives. Iterated filtering, unlike generic stochastic approximation algorithms, uses the Monte Carlo variability in Algorithm 1 to explore the parameter space $\mathbb{R}^{d_\theta}$ while simultaneously evaluating an approximation to the likelihood function and its derivative. This can lead to high numerical efficiency which is essential in situations which stretch available computational resources. For example, iterated filtering has been able to adequately identify a maximum likelihood estimate in a 13-dimensional space based on 50 iterations (Ionides et al., 2006), a comparable computational burden to 50 evaluations of the likelihood function. There is undoubtedly potential to construct hybrid procedures which combine the strength of iterated filtering—making efficient use of few filtering operations to approach the maximum of the likelihood function—with the strengths of other methodologies. For example, a basic Kiefer-Wolfowitz algorithm (Spall, 2003) applied to an

unbiased SMC estimate of the likelihood function would provide a sequence of estimators which converges to the maximum likelihood estimate with probability one, for a fixed Monte Carlo sample size (i.e., without the requirement $J_m \to \infty$ in Theorem 3).

The major challenge for likelihood-based inference in complex models is to identify a neighborhood containing those models which are plausibly consistent with the data. Once such a region has been identified, one then seeks to describe the likelihood surface in this neighborhood via construction of point estimates, confidence intervals and profile likelihood computations. A theoretical basis for this philosophy is Le Cam's quadratic estimation (Le Cam and Yang, 2000), in which the likelihood surface is approximated in a neighborhood of a $\sqrt{n}$-consistent estimator. Le Cam's ideas can be extended from quadratic approximation of the log likelihood surface to more practically attractive smooth local likelihood approximations (Ionides, 2005). These theoretical results highlight the statistical importance of correctly capturing the features of the likelihood on the scale of the uncertainty in the parameters. Smaller scale features in the likelihood surface, which may be a feature of the model or arise due to numerical considerations, are a distraction from effective inference. From this perspective, the efficient identification of plausible models which is the main strength of iterated filtering techniques is also the key step in model-based data analysis.

4 Proofs of the theorems stated in Section 2

The following proofs rely heavily on the generic probability density functions defined in Section 2. The reader is directed to Appendix A for a discussion on the formal use of this notation. Additionally, $\phi(\tau) = O\big(\psi(\tau)\big)$ will mean that ϕ/ψ is bounded, and $\phi(\tau) = o\big(\psi(\tau)\big)$ will mean that $\lim_{\tau\to 0} \phi/\psi = 0$. In Sections 4.1 and 4.2, we write ∇_θ and ∇_{θ_n} for vectors of partial derivatives with respect to the components of θ and θ_n respectively.

4.1 A proof of Theorem 1

Suppose inductively that $|V_n^P| = O(\tau^2)$ and $|\theta_{n-1}^F - \theta| = O(\tau^2)$, which holds for $n = 1$ by construction. We now employ Bayes' formula, suppressing the dependence of g on θ, σ and τ to give

$$\frac{g(\theta_n \,|\, y_{1:n})}{g(\theta_n \,|\, y_{1:n-1})} = \frac{g(y_n \,|\, y_{1:n-1}, \theta_n)}{\int g(y_n \,|\, y_{1:n-1}, \theta_n)\, g(\theta_n \,|\, y_{1:n-1})\, d\theta_n} \tag{12}$$

$$= \frac{g(y_n|y_{1:n-1}, \theta_n{=}\theta_{n-1}^F) + (\theta_n - \theta_{n-1}^F)'\, \nabla_{\theta_n} g(y_n|y_{1:n-1}, \theta_n{=}\theta_{n-1}^F) + R_1}{g(y_n \,|\, y_{1:n-1}, \theta_n{=}\theta_{n-1}^F) + O(\tau^2)} \tag{13}$$

$$= \Big\{ 1 + (\theta_n - \theta_{n-1}^F)' \nabla_{\theta_n} \log g(y_n \,|\, y_{1:n-1}, \theta_n{=}\theta_{n-1}^F) + \frac{R_1}{g(y_n \,|\, y_{1:n-1}, \theta_n{=}\theta_{n-1}^F)} \Big\} \times \big(1 + O(\tau^2)\big). \tag{14}$$

The numerator in (13) comes from a Taylor series expansion of $g(y_n \,|\, y_{1:n-1}, \theta_n)$ about $\theta_n = \theta_{n-1}^F$. From (A1), (A2) and (A3) there is a constant C_3 such that $|R_1|$ is bounded by $C_3|\theta_n - \theta_{n-1}^F|^2/2$. This assertion is reasoned formally as Lemma 3 in Section 4.2. The denominator of (13) then follows from applying this expansion to the integral in (12) by observing that $E[\theta_n \,|\, y_{1:n-1}] = \theta_{n-1}^F$

and, from the induction hypothesis, $E\big[|\theta_n - \theta^F_{n-1}|^2 \,|\, y_{1:n-1}\big] = O(\tau^2)$. We now calculate

$$\begin{aligned}
\theta^F_n - \theta^F_{n-1} &= E[\theta_n - \theta^F_{n-1} \,|\, y_{1:n}] \\
&= \int (\theta_n - \theta^F_{n-1})\, g(\theta_n \,|\, y_{1:n})\, d\theta_n && (15)\\
&= V^P_n\, \nabla_{\theta_n} \log g(y_n \,|\, y_{1:n-1}, \theta_n = \theta^F_{n-1}) + o(\tau^2) && (16)\\
&= V^P_n\, \nabla_{\theta} \log f(y_n \,|\, y_{1:n-1}, \theta) + o(\tau^2). && (17)
\end{aligned}$$

Equation (16) follows from (15) using (14) and (A4). Equation (17) follows from (16) via Lemma 1 in Section 4.2 below. The inductive assumption on θ^F_n is then justified by (17). A similar argument gives

$$\begin{aligned}
V^P_{n+1} &= \mathrm{Var}(\theta_{n+1} \,|\, y_{1:n}) = \mathrm{Var}(\theta_n \,|\, y_{1:n}) + \sigma^2\Sigma \\
&= E[(\theta_n - \theta^F_n)(\theta_n - \theta^F_n)' \,|\, y_{1:n}] + \sigma^2\Sigma \\
&= E[(\theta_n - \theta^F_{n-1})(\theta_n - \theta^F_{n-1})' \,|\, y_{1:n}] - (\theta^F_n - \theta^F_{n-1})(\theta^F_n - \theta^F_{n-1})' + \sigma^2\Sigma && (18)\\
&= V^P_n + \sigma^2\Sigma + o(\tau^2), && (19)
\end{aligned}$$

where (19) follows from (18) via (14) and (17) together with the induction hypothesis on V^P_n. Equation (19) in turn justifies this hypothesis. Multiplying (17) through by $(V^P_n)^{-1}$ and then summing over n gives (7), completing the proof of Theorem 1.

4.2 Additional details on the proof of Theorem 1

The passage from (16) to (17) may appear natural, given the smoothly parameterized sequence of approximations by which g approaches f. However, there is in fact some subtlety which explains the necessity of the two approximation parameters σ and τ with $\sigma\tau^{-1} \to 0$. If the variability of $g(\theta_{1:n} \,|\, \theta_0, \sigma, \tau)$ is small compared to the variability of $g(\theta_0 \,|\, \theta, \sigma, \tau)$ then, heuristically, one expects $g(\theta_{0:n-1} \,|\, y_{1:n}, \theta_n, \theta, \sigma, \tau)$ to be concentrated around θ_n in the limit as $\tau \to 0$. Lemma 1 takes advantage of a formalization of this limit. However, the issue may be of minor relevance in practice because one expects that $g(\theta_{n-k:n-1} \,|\, y_{1:n}, \theta_n)$ will indeed be concentrated around θ_n when $k \ll n$ even if σ is not small compared to τ. Under typical mixing conditions, the distribution of y_n given $y_{1:n-1}, \theta_{0:n}, \theta, \sigma$ depends only weakly on $\theta_{0:(n-k-1)}$ unless k is small. Introducing mixing conditions typically improves the theoretical properties of filtering procedures (e.g., Crisan and Doucet, 2002). We conjecture that one could achieve a result similar to Lemma 1 for a constant ratio $\sigma\tau^{-1}$ in a limit with some appropriate mixing properties, though investigating such scenarios is outside the scope of this article.

Lemma 1. *Suppose the conditions (A2) and (A4). In the limit as $\tau \to 0$,*

$$\nabla_{\theta_n} \log g(y_n \,|\, y_{1:n-1}, \theta_n = \theta + c\tau, \theta, \sigma, \tau) = \nabla_{\theta} \log f(y_n \,|\, y_{1:n-1}, \theta) + o(1) \qquad (20)$$

uniformly over $0 \le c < C_2 - \delta$ for any $\delta > 0$. In particular, (20) holds for $\theta_n = \theta^F_{n-1}$.

Proof. We suppress the dependence of g on θ, σ and τ. Adopting the notation that $\{\theta_{0:n} = \psi\}$ means $\{\theta_k = \psi, 0 \le k \le n\}$, it follows from (A2) and (A4) that

$$\begin{aligned}
\log g(y_n \,|\, y_{1:n-1}, \theta_n = \theta + c\tau) &= \log g(y_n \,|\, y_{1:n-1}, \theta_{0:n} = \theta + c\tau) + O(\sigma) \\
&= \log f(y_n \,|\, y_{1:n-1}, \theta = \theta + c\tau) + O(\sigma), && (21)
\end{aligned}$$

with the term $O(\sigma)$ being uniform over sets of values of c bounded away from C_2. Further details on this step are provided in Appendix C. We then calculate

$$\begin{aligned}
&\nabla_{\theta_n} \log g(y_n \,|\, \theta_n{=}\theta + c\tau, y_{1:n-1}) \\
&= \lim_{\delta \to 0} \delta^{-1} \big\{ \log g(y_n \,|\, \theta_n{=}\theta + c\tau + \delta, y_{1:n-1}) - \log g(y_n \,|\, \theta_n{=}\theta + c\tau, y_{1:n-1}) \big\} \\
&= \lim_{\delta \to 0} \delta^{-1} \big\{ \log f(y_n \,|\, y_{1:n-1}, \theta + c\tau + \delta) - \log f(y_n \,|\, y_{1:n-1}, \theta + c\tau) + R_2 \big\}
\end{aligned}$$

where R_2 is $O(\sigma)$, uniformly in δ as long as δ is small compared to τ. Setting $\delta = \delta(\tau)$ with $\delta\tau^{-1} \to 0$ and $\sigma\delta^{-1} \to 0$, it follows that

$$\begin{aligned}
&\nabla_{\theta_n} \log g(y_n \,|\, y_{1:n-1}, \theta_n{=}\theta + c\tau) \\
&\qquad = \nabla_\theta \log f(y_n \,|\, y_{1:n-1}, \theta{=}\theta + c\tau) + O(\sigma/\delta) + o(1) \qquad (22) \\
&\qquad = \nabla_\theta \log f(y_n \,|\, y_{1:n-1}, \theta) + o(1). \qquad (23)
\end{aligned}$$

To justify (23) it is necessary to notice that the term $o(1)$ in (22) depends only on δ and not on σ or τ. The uniformity of (23) over c allows one to choose $c = c(\tau) = (\theta_n^F - \theta)/\tau$, completing the proof of Lemma 1. □

We now proceed to derive the uniform bound on $\nabla^2_{\theta_n} g(y_n | y_{1:n-1}, \theta_n)$ required to bound R_1 in (13). To prove this result, stated as Lemma 3, we present a bound on $\nabla_{\theta_n} g(y_n | y_{1:n-1}, \theta_n)$ and then argue that the method can be extended to the second derivative at the expense of additional routine algebra. By contrast, a bound on $\nabla_{\theta_n} g(y_n | y_{1:n-1}, \theta_n)$ can be obtained more directly from Lemma 1, but this approach does not generalize to the second derivative unless $\sigma = o(\tau^2)$.

Lemma 2. *(A1–A4) implies* $\nabla_{\theta_n} g(y_{1:n}, \theta_n \,|\, \theta, \sigma, \tau) = \sum_{i=1}^n \big(U_i + V_i\big) + V_0 + W_0$ *where*

$$\begin{aligned}
U_i &= \int \big[\nabla_\theta f(y_i \,|\, x_i, \theta_i)\big] g(y_{1:i-1}, y_{i+1:n}, x_{0:n}, \theta_{0:n} \,|\, \theta, \sigma, \tau)\, dx_{0:n}\, d\theta_{0:n-1}, \\
V_i &= \int \big[\nabla_\theta f(x_i \,|\, x_{i-1}, \theta_{i-1})\big] g(y_{i:n}, x_{i+1:n} \,|\, x_i, \theta_{0:n}, \theta, \sigma, \tau) \\
&\qquad\qquad g(y_{1:i-1}, x_{1:i-1} \,|\, \theta_{0:n}, \theta, \sigma, \tau) g(\theta_{0:n} \,|\, \theta, \sigma, \tau)\, dx_{0:n}\, d\theta_{0:n-1} \\
V_0 &= \int \big[\nabla_\theta f(x_0 \,|\, \theta_0)\big] g(y_{1:n}, x_{1:n} \,|\, x_0, \theta_{0:n}, \theta, \sigma, \tau) g(\theta_{0:n} \,|\, \theta, \sigma, \tau)\, dx_{0:n}\, d\theta_{0:n-1}, \\
W_0 &= \nabla_\theta\, g(y_{1:n}, \theta_n \,|\, \theta, \sigma, \tau).
\end{aligned}$$

Proof. Integrating $g(y_{1:n}, x_{0:n}, \theta_{0:n} \,|\, \theta, \sigma, \tau)$ over $x_{0:n}$ and $\theta_{0:n-1}$ and passing ∇_{θ_n} through the resulting integral gives $\nabla_{\theta_n} g(y_{1:n}, \theta_n \,|\, \theta, \sigma, \tau) = U_n + T_n$ for

$$\begin{aligned}
T_i = \int \prod_{j=1}^n f(y_j \,|\, x_j, \theta_j) f(x_j \,|\, x_{j-1}, \theta_{j-1}) f(x_0 \,|\, \theta_0) \times \Big[\nabla_{\theta_i} \kappa\left(\frac{\theta_i - \theta_{i-1}}{\sigma}\right)\Big] \\
\times \frac{1}{\sigma^n} \prod_{j \neq i} \kappa\left(\frac{\theta_j - \theta_{j-1}}{\sigma}\right) \times \frac{1}{\tau} \kappa\left(\frac{\theta_0 - \theta}{\tau}\right) dx_{0:n}\, d\theta_{0:n-1}. \qquad (24)
\end{aligned}$$

Noticing that $\nabla_{\theta_i} \kappa\big([\theta_i - \theta_{i-1}]/\sigma\big) = -\nabla_{\theta_{i-1}} \kappa\big([\theta_i - \theta_{i-1}]/\sigma\big)$ and applying integration by parts to (24) one finds that $T_i = V_i + U_{i-1} + T_{i-1}$ for $2 \le i \le n$. A very similar calculation gives $T_1 = V_1 + V_0 + W_0$, completing the proof of Lemma 2. □

Lemma 3. *Supposing (A1–A4), it follows that $\nabla^2_{\theta_n} g(y_n \,|\, y_{1:n-1}, \theta_n, \theta, \sigma, \tau)$ is uniformly bounded for σ and τ in a neighborhood of zero, over compact regions of θ with $|\theta_n - \theta| < c\tau$ for $0 < c < C_2$.*

Proof. From (A1–A4) it follows that $U_1, \ldots, U_n$, $V_0, \ldots, V_n$ and W_0 in Lemma 2 are $O(\tau^{-1})$ and therefore that $\nabla_{\theta_n} g(y_{1:n}, \theta_n \,|\, \theta, \sigma, \tau) = O(\tau^{-1})$. Suppressing dependence on θ, σ and τ to write

$$\nabla_{\theta_n} g(y_n \,|\, y_{1:n-1}, \theta_n) = \frac{g(y_{1:n-1}, \theta_n) \nabla_{\theta_n} g(y_{1:n}, \theta_n) - g(y_{1:n}, \theta_n) \nabla_{\theta_n} g(y_{1:n-1}, \theta_n)}{[g(y_{1:n-1}, \theta_n)]^2}$$

and observing that $C_4 \tau^{-1} < g(y_{1:n}, \theta_n \,|\, \theta, \sigma, \tau) < C_5 \tau^{-1}$ for some positive constants C_4 and C_5, we find $\nabla_{\theta_n} g(y_n \,|\, y_{1:n-1}, \theta_n, \theta, \sigma, \tau) = O(1)$. This argument can be routinely extended to the second derivative, completing the proof of Lemma 3. □

4.3 A proof of Theorem 2

Let $u_{n,m} = (\theta^F_{n,m} - \theta^F_{n-1,m})/\tau_m$ and $v_{n,m} = V^P_{n,m}/\tau^2_m$. The corresponding Monte Carlo estimates of these quantities are $\widetilde{u}_{n,m} = (\widetilde{\theta}^F_{n,m} - \widetilde{\theta}^F_{n-1,m})/\tau_m$ and $\widetilde{v}_{n,m} = \widetilde{V}^P_{n,m}/\tau^2_m$. We claim that there are constants $C_6, \ldots, C_9$ with

$$\big|\, \widetilde{E}\big[\widetilde{u}_{n,m} - u_{n,m}\big] \,\big| \le C_6/J_m \qquad \big|\, \widetilde{E}\big[\widetilde{v}_{n,m} - v_{n,m}\big] \,\big| \le C_7/J_m \tag{25}$$

$$\widetilde{E}\big[|\widetilde{u}_{n,m} - u_{n,m}|^2\big] \le C_8/J_m \qquad \widetilde{E}\big[|\widetilde{v}_{n,m} - v_{n,m}|^2\big] \le C_9/J_m \tag{26}$$

uniformly for θ in any compact set. Previous bounds similar to (25,26) have been given for a fixed model as the Monte Carlo sample size J_m increases, for example by Del Moral and Jacod (2001); Del Moral (2004, Section 11.8.4); Crisan and Doucet (2002). The complication in (25,26) is that the model is varying with σ_m and τ_m. However, the uniform bound on $\widetilde{u}_{n,m}$ and $\widetilde{v}_{n,m}$, together with the continuity of $g(\cdot \,|\, \cdot, \sigma, \tau)$ as a function of σ and τ, is enough to show that a convergence result for fixed models (Crisan and Doucet, 2002, Theorem 2) applies uniformly in this context. Further details of this argument are deferred to Section 4.4. Carrying out a Taylor series expansion, we find

$$\begin{aligned} \widetilde{v}^{-1}_{n,m} \widetilde{u}_{n,m} \;&=\; v^{-1}_{n,m} u_{n,m} + v^{-1}_{n,m}(\widetilde{u}_{n,m} - u_{n,m}) \\ &\quad - v^{-1}_{n,m}(\widetilde{v}_{n,m} - v_{n,m}) v^{-1}_{n,m} \widetilde{u}_{n,m} + R_3 \end{aligned} \tag{27}$$

where $|R_3| < C_{10}(|\widetilde{u}_{n,m} - u_{n,m}|^2 + |\widetilde{v}_{n,m} - v_{n,m}|^2)$ for some constant C_{10}. The existence of such a C_{10} is guaranteed since the determinant of $v_{n,m}$ is bounded away from zero. Taking expectations of both sides of (27) and applying (25,26) gives

$$\big|\widetilde{E}[\widetilde{v}^{-1}_{n,m} \widetilde{u}_{n,m}] - v^{-1}_{n,m} u_{n,m}\big| \le C_{11}/J_m. \tag{28}$$

Another Taylor series expansion,

$$\widetilde{v}^{-1}_{n,m} \widetilde{u}_{n,m} = v^{-1}_{n,m} u_{n,m} + R_4 \tag{29}$$

with $|R_4| < C_{12}(|\widetilde{u}_{n,m} - u_{n,m}| + |\widetilde{v}_{n,m} - v_{n,m}|)$ implies

$$\widetilde{\mathrm{Var}}(\widetilde{v}^{-1}_{n,m} \widetilde{u}_{n,m}) \le C_{13}/J_m. \tag{30}$$

Putting together (28) and (30), we see that

$$\begin{aligned} \widetilde{E}\big[\big(\widetilde{V}^P_{n,m}\big)^{-1}(\widetilde{\theta}^F_{n,m} - \widetilde{\theta}^F_{n-1,m})\big] \;&=\; \big(V^P_{n,m}\big)^{-1}(\theta^F_{n,m} - \theta^F_{n-1,m}) + O\big(1/(\sigma_m J_m)\big) \\ \widetilde{\mathrm{Var}}\big[\big(\widetilde{V}^P_{n,m}\big)^{-1}(\widetilde{\theta}^F_{n,m} - \widetilde{\theta}^F_{n-1,m})\big] \;&=\; O\big(1/(\sigma^2_m J_m)\big). \end{aligned}$$

Theorem 2 then follows via the assumed continuity with respect to θ.

4.4 Justification of the sequential Monte Carlo bounds in Section 4.3

We draw on the general theory of sequential Monte Carlo by Crisan and Doucet (2002) and Del Moral and Jacod (2001). For completeness, and to help the reader translate these results into the current context and notation, we have included the following two theorem statements.

Theorem 4. (Crisan and Doucet, 2002) *Let $h(\cdot \,|\, \cdot, \theta)$ be a generic density for an unobserved Markov process $z_{0:N}$ with observations $y_{1:N}$ and parameter vector θ. Define $Z^F_{n,j}$ via applying Algorithm 1 to $h(\cdot \,|\, \cdot, \theta)$ with J particles. Assume that $h(y_n \,|\, z_n, \theta)$ is bounded as a function of z_n. For any $\varphi : \mathbb{R}^{d_z} \to \mathbb{R}$, denote the filtered mean of $\varphi(z_n)$ and its Monte Carlo estimate by*

$$\varphi^F_n = \int \varphi(z_n) h(z_n \,|\, y_{1:n}, \theta)\, dz_n, \qquad\qquad \widetilde{\varphi}^F_n = \frac{1}{J} \sum_{j=1}^{J} \varphi\big(Z^F_{n,j}\big). \tag{31}$$

There is a C_{14} independent of J such that

$$\widetilde{E}\left[(\widetilde{\varphi}^F_n - \varphi^F_n)^2\right] \leq \frac{C_{14} \sup_x |\varphi(x)|^2}{J}. \tag{32}$$

Specifically, C_{14} can be written as a linear function of 1 and $\eta_{n,1}, \ldots, \eta_{n,n}$ defined as

$$\eta_{n,i} = \prod_{k=n-i+1}^{n} \left(\frac{\sup_{z_k} h(y_k \,|\, z_k, \theta)}{h(y_k \,|\, y_{1:k-1}, \theta)} \right)^2. \tag{33}$$

Theorem 5. (Del Moral and Jacod, 2001) *As in Theorem 4, let $h(\cdot \,|\, \cdot, \theta)$ be a generic density for an unobserved Markov process $z_{0:N}$ with observations $y_{1:N}$ and parameter vector θ. Let $\varphi : \mathbb{R}^{d_z} \to \mathbb{R}$ be a bounded function, with φ^F_n and $\widetilde{\varphi}^F_n$ specified in (31). Define the un-normalized filtered mean φ^U_n and its Monte Carlo estimate $\widetilde{\varphi}^U_n$ by*

$$\varphi^U_n = \varphi^F_n \prod_{k=1}^{n} h(y_k \,|\, y_{1:k-1}, \theta), \qquad\qquad \widetilde{\varphi}^U_n = \widetilde{\varphi}^F_n \prod_{k=1}^{n} \frac{1}{J} \sum_{j=1}^{J} w(k, j). \tag{34}$$

where $w(k, j)$ is computed in Step 4 of Algorithm 1 when evaluating φ^F_n. Then

$$\widetilde{E}[\widetilde{\varphi}^U_n] \;=\; \varphi^U_n, \tag{35}$$

$$\widetilde{E}\Big[(\widetilde{\varphi}^U_n - \varphi^U_n)^2\Big] \;\leq\; \frac{(n+1) \sup_x |\varphi(x)|^2}{J} \prod_{k=1}^{n} \sup_{z_k} h(y_k \,|\, z_k, \theta)^2. \tag{36}$$

Corollary 1. *Setting $\varphi(z) = 1$ in (34) we see that $1^U_N = \prod_{k=1}^{N} h(y_k \,|\, y_{1:k-1}, \theta)$. Suppressing the dependence on $y_{1:N}$, we write $1^U_N = L(\theta)$, the likelihood function for h. Correspondingly, one can write $\widetilde{1}^U_N = \widetilde{L}(\theta)$. It follows from (35) that the Monte Carlo likelihood $\widetilde{L}(\theta)$ is an unbiased estimate of $L(\theta)$.*

Some intuition arising from Theorem 5 is that the bias in using the Monte Carlo estimate $\widetilde{\varphi}^F_n$ for φ^F_n arises due to the nonlinearity of the normalization procedure. From (35), we see that $\widetilde{\varphi}^U_n$ is an unbiased estimate of φ^U_n. Defining the unit function $1(x) = 1$, it also follows that $\widetilde{1}^U_n$ is an unbiased estimate of 1^U_n. However, $\widetilde{\varphi}^F_n = \widetilde{\varphi}^U_n / \widetilde{1}^U_n$ is generally a biased estimate of $\varphi^F_n = \varphi^U_n / 1^U_n$. The un-normalized filtered mean in Theorem 5 is not usually a quantity of direct interest (Corollary 1 being an exception). However, Theorem 5 is necessary to justify (37) and (38) below.

The bound in terms of (33) was not explicitly mentioned by Crisan and Doucet (2002) but is a direct consequence of their proof (see Appendix B). The uniform bound in (26) then follows from the observation that η_k in (33) is continuous as a function of σ in the context of Theorem 2. To show that (25) follows from (26) we follow the approach of Del Moral and Jacod (2001, Equation 3.3.14). Noting that $\varphi_n^F = \varphi_n^U / 1_n^U$ and $\widetilde{\varphi}_n^F = \widetilde{\varphi}_n^U / \widetilde{1}_n^U$, Theorem 5 implies the identity

$$\widetilde{E}[\widetilde{\varphi}_n^F - \varphi_n^F] = \widetilde{E}\Big[(\widetilde{\varphi}_n^F - \varphi_n^F)\Big(1 - \frac{\widetilde{1}_n^U(1)}{1_n^U(1)}\Big)\Big]. \tag{37}$$

Applying the Cauchy-Schwarz inequality, together with (32) and (36), gives

$$\big|\, \widetilde{E}[\widetilde{\varphi}_n^F - \varphi_n^F] \,\big| \leq C_{15} \frac{\sup_x |\varphi(x)|}{J}. \tag{38}$$

The uniform bound in (25) follows from the observation that the bounding constants in (32) and (36) are continuous as a function of σ in the context of Theorem 2.

4.5 A proof of Theorem 3

Theorem 3 follows directly from a general stochastic approximation result, presented as Theorem 6 below. In the context of Theorem 3, conditions (B5) and (B6) of Theorem 6 hold from Theorem 2 and the remaining assumptions of Theorem 6 hold by hypothesis.

Theorem 6. *Let $\ell(\theta)$ be a continuously differentiable function $\mathbb{R}^{d_\theta} \to \mathbb{R}$ and let $\{D_m(\theta), m \geq 1\}$ be a sequence of independent Monte Carlo estimators of the vector of partial derivatives $\nabla\ell(\theta)$. Define a sequence $\{\hat{\theta}_m\}$ recursively by $\hat{\theta}_{m+1} = \hat{\theta}_m + a_m D_m(\hat{\theta}_m)$. Assume (B1–B2) of Section 2 together with the following conditions:*

(B4) $a_m > 0$, $a_m \to 0$, $\sum_m a_m = \infty$.

(B5) $\sum_m a_m^2 \sup_{|\theta|<r} \widetilde{\mathrm{Var}}\big(D_m(\theta)\big) < \infty$ *for every* $r > 0$.

(B6) $\lim_{m\to\infty} \sup_{|\theta|<r} \big|\widetilde{E}[D_m(\theta)] - \nabla\ell(\theta)\big| = 0$ *for every* $r > 0$.

Then $\hat{\theta}_m$ converges to $\hat{\theta} = \arg\max \ell(\theta)$ with probability one.

Theorem 6 is a special case of Theorem 2.3.1 of Kushner and Clark (1978). The most laborious step in deducing Theorem 6 from Kushner and Clark (1978) is to check that (B1–B6) imply that, for all $\epsilon > 0$,

$$\lim_{n\to\infty} P\Big[\sup_{j\geq 1}\Big|\sum_{m=n}^{n+j} a_m\big\{D_m(\hat{\theta}_m) - \widetilde{E}[D_m(\hat{\theta}_m)\,|\,\hat{\theta}_m]\big\}\Big| \geq \epsilon\Big] = 0, \tag{39}$$

which in turn implies condition A2.2.4 of Kushner and Clark (1978). To show (39), we define $\xi_m = D_m(\hat{\theta}_m) - \widetilde{E}[D_m(\hat{\theta}_m)\,|\,\hat{\theta}_m]$ and

$$\xi_m^k = \begin{cases} \xi_m & \text{if } |\hat{\theta}_m| \leq k \\ 0 & \text{if } |\hat{\theta}_m| > k \end{cases}. \tag{40}$$

Define processes $\{M_j^n = \sum_{m=n}^{n+j} a_m \xi_m, j \geq 0\}$ and $\{M_j^{n,k} = \sum_{m=n}^{n+j} a_m \xi_m^k, j \geq 0\}$ for each k and n. These processes are martingales with respect to the filtration defined by the Monte Carlo stochasticity. From the Doob-Kolmogorov martingale inequality (e.g., Grimmett and Stirzaker, 1992),

$$P\big[\sup_j |M_j^{n,k}| \geq \epsilon\big] \leq \frac{1}{\epsilon^2} \sum_{m=n}^{\infty} a_m^2 \sup_{|\theta|<k} \widetilde{\mathrm{Var}}\big(D_m(\theta)\big). \tag{41}$$

Define events $F_n = \{\sup_j |M_j^n| \geq \epsilon\}$ and $F_{n,k} = \{\sup_j |M_j^{n,k}| \geq \epsilon\}$. It follows from (B5) and (41) that $\lim_{n\to\infty} P\{F_{n,k}\} = 0$ for each k. In light of the non-divergence assumed in (B2), this implies $\lim_{n\to\infty} P\{F_n\} = 0$ which is exactly (39).

To expand on this final assertion, let $\Omega = \{\sup_m |\hat{\theta}_m| < \infty\}$ and $\Omega_k = \{\sup_m |\hat{\theta}_m| < k\}$. Assumption (B2) implies that $P(\Omega) = 1$. Since the sequence of events $\{\Omega_k\}$ is increasing up to Ω, we have $\lim_{k\to\infty} P(\Omega_k) = P(\Omega) = 1$. Now observe that $\Omega_k \cap F_{n,j} = \Omega_k \cap F_n$ for all $j \geq k$, as there is no truncation of the sequence $\{\xi_m^j, m = 1, 2, \dots\}$ for outcomes in Ω_k when $j \geq k$. Then,

$$\begin{aligned}
\lim_{n\to\infty} P[F_n] &\leq \lim_{n\to\infty} P[F_n \cap \Omega_k] + 1 - P[\Omega_k] \\
&= \lim_{n\to\infty} P[F_{n,k} \cap \Omega_k] + 1 - P[\Omega_k] \\
&\leq \lim_{n\to\infty} P[F_{n,k}] + 1 - P[\Omega_k] \\
&= 1 - P[\Omega_k].
\end{aligned}$$

Since k can be chosen to make $1 - P[\Omega_k]$ arbitrarily small, it follows that $\lim_{n\to\infty} P[F_n] = 0$.

A Formalizing the use of generic functions via overloading

A partially observed Markov model was specified by a generic probability density functions $g(\cdot \,|\, \cdot)$. This was used to describe all joint and conditional distributions, with the argument to $g(\cdot \,|\, \cdot)$ specifying the density in question. For example,

$$g(y_n \,|\, y_{1:n-1}, \theta_n) \tag{42}$$

gives the distribution of y_n given $y_{1:n-1}$ and θ_n. This notation has advantages that it is self-documenting and one does not have to define additional notation for the density of every new quantity that is brought into consideration. Ensuring that the notation results in correct mathematics requires some amount of care for the following two reasons: (i) we do not distinguish between random variables and their realizations; (ii) the meaning of a function should not usually depend on the name of the argument supplied, so $f(x)$ and $f(y)$ should correspond to the same function f evaluated at x or y respectively.

An equally adaptable notation, which is more explicit but cumbersome to the point of being unusable, involves rewriting (42) as

$$g_{Y_n|Y_{1:n-1},\Theta_n}(y_n \,|\, y_{1:n-1}, \theta_n). \tag{43}$$

The map from (42) onto (43) is an instance of function overloading, which for the current purposes is synonymous with the technical term *function polymorphism* in computer languages (Strachey, 2000). To define a polymorphic function, which takes different functional forms depending on the arguments, one must label the arguments with *types*. We suppose that y_n has type Y_n, and similarly θ_n has type Θ_n, etcetera. The overloaded function $g(\cdot \,|\, \cdot)$ in (42) then looks to the type of its arguments when it is evaluated via (43). Suppose that we wish to evaluate (42) at $\theta_n^* = \theta_n + \epsilon_n$. We can achieve this by writing $g(y_n \,|\, y_{1:n-1}, \theta_n^*)$ if it is considered clear that θ_n^* should possess the same type as θ_n, namely Θ_n. We also interpret $g(y_n \,|\, y_{1:n-1}, \theta_n{=}\theta_n^*)$ as an explicit instruction to evaluate $g(y_n \,|\, y_{1:n-1}, \theta_n^*)$ with θ_n^* being assigned the type of θ_n.

The arguments for and against overloading in (42) are essentially the same as those in computer languages. Overloading has a potential for conceptual clarity and conciseness which must be weighed against the potential cost in terms of errors arising from incorrect applications of the

typing rules. Overloading is fundamental to object oriented computer programming, and notation such as (42) has similarly been found useful for working with dynamic models. The formalization of (42) in terms of function overloading enables more confident use of this convenient notation.

B Some additional details on Theorem 4

Here, we seek to support our assertion in the statement of Theorem 4 that the proof in Crisan and Doucet (2002) implies the constant C_{14} in equation (32) can be written as a linear function of 1 and $\eta_{n,1}, \ldots, \eta_{n,n}$ defined as

$$\eta_{n,i} = \prod_{k=n-i+1}^{n} \left(\frac{\sup_{z_k} h(y_k \mid z_k, \theta)}{h(y_k \mid y_{1:k-1}, \theta)} \right)^2 .$$

Below, we use the notation of Crisan and Doucet (2002) and a reader wishing to follow our argument is advised to have a copy of this article at hand. Crisan and Doucet (2002, Section V) introduced the following recursion:

$$c_{n|n} = \left(\sqrt{C} + \sqrt{\tilde{c}_{n|n}} \right)^2 \tag{44}$$

$$\tilde{c}_{n|n} = 4 c_{n|n-1} \left(\frac{||h||_n}{h(y_n | y_{1:n-1}, \theta)} \right)^2 \tag{45}$$

$$c_{n|n-1} = \left(1 + \sqrt{c_{n-1|n-1}} \right)^2 \tag{46}$$

where $||h||_n = \sup_{z_n} h(y_n | z_n, \theta)$. Here, C is a constant that depends on the resampling procedure but not on the number of particles J. The constant C_{14} in equation (32) corresponds to $c_{n|n}$. Now, (44–46) can be reformulated by routine algebra as

$$c_{n|n} \leq K_1 + K_2 \, \tilde{c}_{n|n} \tag{47}$$

$$\tilde{c}_{n|n} \leq K_3 \, q_n \, c_{n|n-1} \tag{48}$$

$$c_{n|n-1} \leq K_4 + K_5 \, c_{n-1|n-1} \tag{49}$$

where $q_n = ||h||_n^2 \left[h(y_n | y_{1:n-1}, \theta) \right]^{-2}$ and $K_1, \ldots, K_5$ are constants which do not depend on h, θ, $y_{1:N}$ or J. Putting (48) and (49) into (47),

$$\begin{aligned} c_{n|n} &\leq K_1 + K_2 K_3 q_n c_{n|n-1} \\ &\leq K_1 + K_2 K_3 K_4 q_n + K_2 K_3 K_5 q_n c_{n-1|n-1}. \end{aligned} \tag{50}$$

Since $\eta_{n,i} = q_n \eta_{n-1,i}$ for $i < n$, and $\eta_{n,n} = q_n$, the required assertion follows from (50).

C Some additional details on the proof of Lemma 1

We wish to show the validity of the assertion that

$$\log g(y_n \mid y_{1:n-1}, \theta_n = \theta + c\tau) = \log g(y_n \mid y_{1:n-1}, \theta_{0:n} = \theta + c\tau) + O(\sigma). \tag{51}$$

We will instead show that

$$g(y_n \mid y_{1:n-1}, \theta_n = \theta + c\tau) = g(y_n \mid y_{1:n-1}, \theta_{0:n} = \theta + c\tau) + O(\sigma), \tag{52}$$

from which (51) follows since g is bounded away from zero on compact sets. Begin by defining

$$\begin{aligned} I &= g(y_n \,|\, y_{1:n-1}, \theta_n{=}\theta + c\tau) \\ &= \int g(y_n \,|\, y_{1:n-1}, \theta_{0:n-1}, \theta_n{=}\theta + c\tau)\, g(\theta_{0:n-1} \,|\, \theta_n{=}\theta + c\tau, y_{1:n})\, d\theta_{0:n-1}. \end{aligned} \tag{53}$$

Now a Taylor expansion gives

$$g(y_n \,|\, y_{1:n-1}, \theta_{0:n-1}, \theta_n{=}\theta + c\tau) = A + B, \tag{54}$$

where

$$\begin{aligned} A &= g(y_n \,|\, y_{1:n-1}, \theta_{0:n}{=}\theta + c\tau) \\ B &= (\theta_{0:n-1} - \theta - c\tau)' \nabla_{\theta_{0:n-1}} g(y_n \,|\, y_{1:n-1}, \theta_n{=}\theta + c\tau, \theta^*_{0:n-1}) \end{aligned}$$

for some $\theta^*_{0:n-1}$ satisfying

$$\theta_{0:n-1} \le \theta^*_{0:n-1} \le \theta + c\tau. \tag{55}$$

Here, it is understood that $\theta^*_{0:n-1} \le \theta + c\tau$ means $\theta^*_k \le \theta + c\tau$ componentwise for each $0 \le k \le n-1$. Putting (54) into (53),

$$I = \int (A + B)\, g(\theta_{0:n-1} \,|\, \theta_n{=}\theta + c\tau, y_{1:n})\, d\theta_{0:n-1}.$$

Since A does not depend on $\theta_{0:n-1}$, and $g(.|.)$ is a density, we find that

$$I = A + \int B\, g(\theta_{0:n-1} \,|\, \theta_n{=}\theta + c\tau, y_{1:n})\, d\theta_{0:n-1}. \tag{56}$$

To show (51), it therefore suffices to argue that the second term of (56) is $O(\sigma)$. For $\theta^*_{0:n-1}$ restricted to a compact subset of $\mathbb{R}^{d_\theta}$, (A2) guarantees the existence of a C_{16} such that

$$|\nabla_{\theta_{0:n-1}} g(y_n | y_{1:n-1}, \theta_n{=}\theta + c\tau, \theta^*_{0:n-1}, x_{1:n})| < C_{16}.$$

Therefore,

$$\begin{aligned} &\nabla_{\theta_{0:n-1}} g(y_n | y_{1:n-1}, \theta_n = \theta + c\tau, \theta^*_{0:n-1}) \\ &= \int \nabla_{\theta_{0:n-1}} g(y_n \,|\, y_{1:n-1}, \theta_n{=}\theta{+}c\tau, \theta^*_{0:n-1}, x_{1:n})\, g(x_{1:n} \,|\, y_{1:n-1}, \theta_n{=}\theta{+}c\tau, \theta^*_{0:n-1})\, dx_{1:n} \\ &\le C_{16} \end{aligned} \tag{57}$$

since $g(.|.)$ integrates to 1. Using (57), we expand the second term of (56) to give

$$\begin{aligned} &\int B\, g(\theta_{0:n-1} \,|\, \theta_n{=}\theta + c\tau, y_{1:n})\, d\theta_{0:n-1} \\ &= \int (\theta_{0:n-1} - \theta - c\tau)' \nabla_{\theta_{0:n-1}} g(y_n \,|\, y_{1:n-1}, \theta_n{=}\theta{+}c\tau, \theta^*_{0:n-1}) g(\theta_{0:n-1} \,|\, \theta_n{=}\theta{+}c\tau, y_{1:n})\, d\theta_{0:n-1} \\ &\le C_{16} \int (\theta_{0:n-1} - \theta - c\tau)' g(\theta_{0:n-1} | \theta_n = \theta + c\tau, y_{1:n})\, d\theta_{0:n-1} \\ &= C_{16} \big\{ E[\theta_{0:n-1} | \theta_n = \theta + c\tau] - \theta - c\tau \big\} \qquad (58) \\ &= O(\sigma). \qquad (59) \end{aligned}$$

(59) follows from (58) by (A4), completing our demonstration of (51).

Acknowledgments

The authors acknowledge the financial support of the National Science Foundation (Grants DMS-0805533 and EF-0430120), the Graham Environmental Sustainability Institute, the RAPIDD program of the Science & Technology Directorate, Department of Homeland Security, and the Fogarty International Center, National Institutes of Health. This work was conducted as a part of the Inference for Mechanistic Models Working Group supported by the National Center for Ecological Analysis and Synthesis, a Center funded by NSF (Grant DEB-0553768), the University of California, Santa Barbara, and the State of California.

References

Anderson, B. D. and J. B. Moore (1979). *Optimal Filtering*. New Jersey: Prentice-Hall.

Anderson, J. L. and N. Collins (2007). Scalable implementations of ensemble filter algorithms for data assimilation. *Journal of Atmospheric and Oceanic Technology 24*, 1452–1463.

Arulampalam, M. S., S. Maskell, N. Gordon, and T. Clapp (2002). A tutorial on particle filters for online nonlinear, non-Gaussian Bayesian tracking. *IEEE Transactions on Signal Processing 50*, 174 – 188.

Bretó, C., D. He, E. L. Ionides, and A. A. King (2009). Time series analysis via mechanistic models. *Annals of Applied Statistics*, To appear.

Cappé, O., S. Godsill, and E. Moulines (2007, May). An overview of existing methods and recent advances in sequential Monte Carlo. *Proceedings of the IEEE 95*(5), 899–924.

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

Crisan, D. and A. Doucet (2002). A survey of convergence results on particle filtering methods for practitioners. *IEEE Transactions on Signal Processing 50*(3), 736–746.

Del Moral, P. (2004). *Feynman-Kac Formulae: Genealogical and Interacting Particle Systems with Applications*. New York: Springer.

Del Moral, P. and J. Jacod (2001). Interacting particle filtering with discrete observations. In A. Doucet, N. de Freitas, and N. J. Gordon (Eds.), *Sequential Monte Carlo Methods in Practice*, pp. 43–75. New York: Springer.

Durbin, J. and S. J. Koopman (2001). *Time Series Analysis by State Space Methods*. Oxford: Oxford University Press.

Ergun, A., R. Barbieri, U. Eden, M. Wilson, and E. Brown (2007). Construction of point process adaptive filter algorithms for neural systems using sequential Monte Carlo methods. *IEEE Transactions on Biomedical Engineering 54*(3), 419–428.

Fernández-Villaverde, J. and J. F. Rubrio-Ramírez (2007). Estimating macroeconomic models: A likelihood approach. *Review of Economic Studies 74*, 1059–1087.

Godsill, S., J. Vermaak, W. Ng, and J. Li (2007). Models and algorithms for tracking of maneuvering objects using variable rate particle filters. *Proceedings of the IEEE 95*(5), 925–952.

Grimmett, G. R. and D. R. Stirzaker (1992). *Probability and Random Processes*. Oxford: Oxford University Press.

He, D., E. L. Ionides, and A. A. King (2009). Plug-and-play inference for disease dynamics: Measles in large and small towns as a case study. *Submitted for publication*.

Ingber, L. (1993). Simulated annealing: Practice versus theory. *Mathematical and Computer Modelling 18*, 29–57.

Ionides, E. L. (2005). Maximum smoothed likelihood estimation. *Statistica Sinica 15*, 1003–1014.

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

Johannes, M., N. Polson, and J. Stroud (2009). Learning about jumps and stochastic volatility: Filtering stochastic differential equations with jumps. *Review of Financial Studies*, To appear.

Kendall, B. E., C. J. Briggs, W. W. Murdoch, P. Turchin, S. P. Ellner, E. McCauley, R. M. Nisbet, and S. N. Wood (1999). Why do populations cycle? A synthesis of statistical and mechanistic modeling approaches. *Ecology 80*, 1789–1805.

Kiefer, J. and J. Wolfowitz (1952). Stochastic estimation of the maximum of a regression function. *Annals of Mathematical Statistics 23*, 462–466.

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

Kirkpatrick, S., J. Gelatt, C. D., and M. P. Vecchi (1983). Optimization by simulated annealing. *Science 220*(4598), 671–680.

Kitagawa, G. (1998). A self-organising state-space model. *Journal of the American Statistical Association 93*, 1203–1215.

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

Kushner, J. and G. G. Yin (2003). *Stochastic Approximation and Recursive Algorithms and Applications*. New York: Springer.

Le Cam, L. and G. L. Yang (2000). *Asymptotics in Statistics* (2nd edition ed.). New York: Springer.

Liu, J. and M. West (2001). Combining parameter and state estimation in simulation-based filtering. In A. Doucet, N. de Freitas, and N. J. Gordon (Eds.), *Sequential Monte Carlo Methods in Practice*, pp. 197–224. New York: Springer.

Maryak, J. and D. Chin (2008). Global random optimization by simultaneous perturbation stochastic approximation. *IEEE Transactions on Automatic Control 53*(3), 780–783.

Newman, K. B., C. Fernandez, L. Thomas, and S. T. Buckland (2008). Monte Carlo inference for state-space models of wild animal populations. *Biometrics*, Pre–published online.

Polson, N. G., J. R. Stroud, and P. Muller (2008). Practical filtering with sequential parameter learning. *Journal of the Royal Statistical Society, Series B (Statistical Methodology) 70*(2), 413–428.

Reuman, D. C., R. A. Desharnais, R. F. Costantino, O. S. Ahmad, and J. E. Cohen (2006). Power spectra reveal the influence of stochasticity on nonlinear population dynamics. *Proceedings of the National Academy of Sciences of the USA 103*, 18860–18865.

Rogers, L. C. G. and D. Williams (1994). *Diffusions, Markov Processes, and Martingales. Volume 1: Foundations* (2nd ed.). New York: Wiley.

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

Storvik, G. (2002). Particle filters for state-space models with the presence of unknown static parameters. *Signal Processing, IEEE Transactions on 50*(2), 281–289.

Strachey, C. (2000). Fundamental concepts in programming languages. *Higher-Order and Symbolic Computation 13*(1), 11–49.

Toni, T., D. Welch, N. Strelkowa, A. Ipsen, and M. P. Stumpf (2008). Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems. *Journal of the Royal Society Interface*, Pre–published online.

West, M. and J. Harrison (1997). *Bayesian forecasting and dynamic models*. New York: Springer-Verlag.