

**Time series analysis for nonlinear dynamical  
systems with applications to modeling of  
infectious diseases**

by

Anindya Bhadra

A dissertation submitted in partial fulfillment  
of the requirements for the degree of  
Doctor of Philosophy  
(Statistics)  
in The University of Michigan  
2010

Doctoral Committee:

Associate Professor Edward L. Ionides, Chair  
Associate Professor Kerby A. Shedden  
Assistant Professor Yves F. Atchadé  
Assistant Professor Aaron A. King  
Assistant Professor Stilian A. Stoev

To my father, Dr. Asim Kumar Bhadra, for teaching me the value of perseverance by not handing over easy victories in chess, until I earned one, after three years of futile efforts.

## ACKNOWLEDGEMENTS

This dissertation would not have been possible without the mentoring of my advisor, Prof. Ed Ionides, who patiently put up with myriad of seemingly obvious questions I used to ask and still do. I also thank him for critically evaluating everything I write since I do not consider being attentive to details my forté. It has been a joy to work with Prof. Mercedes Pascual and see a truly brilliant mind at work, always in search of new ideas. I thank Dr. Karina Laneri from Prof. Pascual's research group who has been my collaborator for long. Access to the computer clusters of Prof. Pascual and Prof. Aaron King has made the computational work possible. All the co-authors and dissertation committee members have been kind with their time. Finally, my friends and family deserve a special mention for encouraging me throughout.

# TABLE OF CONTENTS

|  |           |
|--|-----------|
| DEDICATION . . . . .   | ii        |
| ACKNOWLEDGEMENTS . . . . .   | iii       |
| LIST OF FIGURES . . . . .  | vi        |
| LIST OF TABLES . . . . .   | ix        |
| ABSTRACT . . . . .   | x         |
| <b>CHAPTER</b>   |           |
| <b>1. Introduction . . . . .</b>   | <b>1</b>  |
| 1.1 Dynamic models . . . . .   | 2         |
| 1.2 Plug and play inference for dynamic models . . . . .   | 3         |
| 1.2.1 Example of a comparative analysis of the state of<br>the art in Bayesian and likelihood based plug and<br>play inference: Particle Markov chain Monte Carlo<br>vs Iterated Filtering . . . . . | 6         |
| 1.3 Sequential Monte Carlo . . . . .   | 10        |
| 1.4 Infectious disease models . . . . .  | 14        |
| <b>2. Iterated Filtering . . . . .</b>   | <b>16</b> |
| 2.1 Introduction . . . . .   | 16        |
| 2.2 Notation and main results . . . . .  | 20        |
| 2.3 Proofs of the main results . . . . .   | 28        |
| 2.3.1 A proof of Theorem 2.2.1 . . . . .   | 28        |
| 2.3.2 Lemmas required for the proof of Theorem 2.2.1 . . . . .   | 30        |
| 2.3.3 A proof of Theorem 2.2.2 . . . . .   | 39        |
| 2.3.4 A proof of Theorem 2.2.3 . . . . .   | 43        |
| 2.4 Discussion of the theory and practice of iterated filtering . . . . .  | 45        |
| 2.4.1 Case studies of iterated filtering . . . . .   | 50        |

|   |           |
|---|-----------|
| <b>3. Malaria in Northwest India: Data analysis via partially observed stochastic differential equation models driven by Lévy noise . . . . .</b> | <b>53</b> |
| 3.1 Introduction . . . . .  | 53        |
| 3.2 Malaria transmission: A statistical model . . . . .   | 58        |
| 3.3 Data analysis . . . . .   | 66        |
| 3.4 Discussion . . . . .  | 77        |
| <b>4. An adaptive particle allocation scheme for off-line sequential importance sampling algorithms . . . . .</b>                                 | <b>79</b> |
| 4.1 Introduction . . . . .  | 79        |
| 4.2 Problem statement and background . . . . .  | 80        |
| 4.3 Applications to likelihood estimation (independent case) . . . . .  | 82        |
| 4.4 Applications to likelihood estimation (dependent case) . . . . .  | 84        |
| 4.4.1 Joint estimation of $q$ and $\phi_k$ . . . . .  | 88        |
| 4.5 Simulation Study . . . . .  | 89        |
| <b>BIBLIOGRAPHY . . . . .</b>   | <b>94</b> |

## LIST OF FIGURES

### Figure

- 1.1 Diagnostic plots for iterated filtering. (a) likelihood at each iteration, evaluated by sequential Monte Carlo; the broken straight line marks the likelihood at the truth. (b) - (e) likelihood surface for each parameter sliced through the maximum; points show parameter values where the likelihoods were evaluated; solid straight lines mark the maximum likelihood estimate; broken straight lines mark the true parameter value. . . . . 9
- 1.2 A pictorial representation of the time evolution of a sequential Monte Carlo scheme adapted from Doucet et al. (2001). The arrows point to the direction of increasing time. Superscripts P and F denote prediction and filtering respectively. The curved solid line denotes the likelihood surface at a given time. The blue and yellow balls represent the different particles in the state space at a given time. The radii of the balls are proportional to their importance weights. . . . . 13

3.1 Flow diagram for a compartment model of malaria transmission. Human classes are  $S_1$  (susceptible),  $S_2$  (partially protected),  $E$  (exposed, carrying a latent infection),  $I_1$  (infected and infectious) and  $I_2$  (asymptomatic, with reduced infectivity). The possibility of transition between class  $X$  and  $Y$  is denoted by a solid arrow, with the corresponding rate written as  $\mu_{XY}$ . The dotted arrows represent interactions between the human and mosquito stages of the parasite. Mosquito dynamics are modeled via the two stages  $\kappa$  (the latent force of infection) and  $\lambda$  (the current force of infection), with  $\tau$  being the mean latency time. The model, which we call  $VS^2EI^2$  with ‘V’ for ‘vector’ followed by a list of the human classes with their multiplicities as superscripts, is formalized by equations (3.1–3.8). We also consider the subcase with  $\mu_{I_2S_2} = \infty$  and  $\mu_{S_2I_2} = \mu_{I_1S_1} = 0$ . The class  $I_2$  can then be eliminated, and so transition directly from  $I_1$  to  $S_2$  becomes possible. Also, individuals in  $S_2$  are fully protected in this case. The remaining classes  $\{S_1, E, I_1, S_2\}$  can then be mapped onto the classes  $\{S, E, I, R\}$  in a standard epidemiological susceptible-exposed-infected-recovered model (Anderson and May, 1991; Keeling and Ross, 2008) with added vector dynamics and waning immunity; we therefore call this special case VSEIR. . . . . 61

3.2 Monthly reported *P. falciparum* malaria cases (solid line) and monthly rainfall from a local weather station (broken line) for Kutch. 67

3.3 Difference between the conditional log likelihood of  $y_n$  given  $y_1, \dots, y_{n-1}$  for the  $VS^2EI^2$  model with rainfall and the VSEIR model with rainfall, plotted against time (bold line). For comparison, reported malaria cases in Kutch are also shown (thin line). . . . . 71

3.4 Profile likelihood plot for the reporting rate ( $\rho$ ) for the  $VS^2EI^2$  model with rainfall (solid line) and the VSEIR model with rainfall (broken line). The profile is estimated via fitting a smooth curve through Monte Carlo evaluations shown as open circles ( $VS^2EI^2$ ) and filled circles (VSEIR). The dashed vertical lines construct approximate 95% confidence intervals (Barndorff-Nielsen and Cox, 1994). . . . . 75

3.5 Profile likelihood plot for the mean development delay time of mosquitoes ( $\tau$ ) for the  $VS^2EI^2$  model with rainfall (solid line) and the VSEIR model with rainfall (broken line). The dashed vertical lines construct approximate 95% confidence intervals. . . . . 76

4.1 Plot of the two dimensional AR(1) process. Left panel shows the state  $(X^1, X^2)$  and observation  $(Y^1, Y^2)$  vectors. Outliers are introduced only in  $Y^1$  at time points 4,51,75 and 90. The vector  $Y^1$  is shown in the right panel, upper part is before introducing outliers (same as that of left panel) and lower part is after introducing outliers. . . . . 91

## LIST OF TABLES

### Table

|     |   |    |
|-----|---|----|
| 3.1 | A likelihood-based comparison of the fitted models. Corresponding point estimates are presented in Table 3.3. The column labeled $p$ corresponds to the number of estimated parameters, including unknown initial conditions. Parameters which were not estimated are documented in Table 3.2. AIC is computed as $AIC = -2\ell + 2p$ .   | 71 |
| 3.2 | List of symbols used in the chapter along with a brief description and units. Some parameters were not estimated as part of the analysis presented in this chapter, and the last column gives their fixed values. Alternative values of these fixed parameters were investigated, but did not affect the conclusions of the analysis. The values of the remaining parameters are give in Table 3.3.   | 72 |
| 3.3 | Estimated model parameters. The columns marked ‘without rain’ correspond to maximum likelihood point estimates under the constraint $\beta = 0$ . The last two columns give the lower and upper bounds for approximate 95% confidence intervals for the VS <sup>2</sup> EI <sup>2</sup> model with rainfall, derived from profile likelihood computations as shown in Figures 3.4 and 3.5; values of 0 and $\infty$ correspond to confidence intervals extending to the boundary of the parameter space. Note: these models coincide with a subset of the models estimated by Laneri et al. (2010). | 73 |
| 4.1 | Comparison of adaptive and non-adaptive particle filters. The gain $\alpha$ as defined in equation (4.10) was found to be 45.862%.  | 93 |

# ABSTRACT

Time series analysis for nonlinear dynamical systems with applications to modeling of infectious diseases

by

Anindya Bhadra

Chair: Edward L. Ionides

Estimation of static (or time constant) parameters in a general class of nonlinear, non-Gaussian, partially observed Markovian state space model is an active area of research that has seen an explosion in the last seventeen years since the formulation of the particle filter and sequential Monte Carlo methods. In this dissertation, we focus on a likelihood based estimation technique known as iterated filtering. The main attractive feature of iterated filtering is we do not need to evaluate the state transition densities in a partially observed Markovian state space model. Instead, we just need to be able to draw samples from those densities, which is typically simpler. This allows great flexibility to the modeler since inference can proceed as long as one is able to write down state transition equations generating trajectories from the model. We discuss some key theoretical properties of iterated filtering. In particular, we prove the consistency of the method and find connections between iterated filtering and well known stochastic approximation methods. We also use the iterated filtering technique to estimate parameters and hence answer scientific questions regarding the effect of climate on malaria transmission in Northwest

India. We conclude by suggesting possible improvements to likelihood estimation techniques via sequential Monte Carlo filters in an off-line setting.

# CHAPTER 1

## Introduction

Partially observed Markov process models, also known as hidden Markov models, have a long history in many branches of science, engineering and economics (Cappé et al., 2005). This dissertation is broadly concerned with parameter estimation techniques in partially observed Markov process models. We also discuss a practical application of such techniques in estimating parameters in an infectious disease model of malaria transmission in order to better understand interactions of intrinsic and extrinsic factors in disease transmission. Finally, some improvements in estimation techniques are presented. The overall organization of this dissertation is as follows:

- In this chapter, we give necessary background on partially observed Markov process models, sequential Monte Carlo and infectious diseases that are useful to understand the subsequent chapters.
- In chapter 2, we describe some results on iterated filtering, a sequential Monte Carlo based technique for likelihood based parameter estimation for a broad class of partially observed Markov process models.
- In chapter 3, we use the technique described in chapter 2 for a real world application. In particular, we model malaria transmission in India.

- In chapter 4, we discuss some results regarding the improvement in likelihood estimation techniques with sequential Monte Carlo based filters. This is motivated by practical problems encountered with data analysis in chapter 3.

## 1.1 Dynamic models

A dynamical system is a mathematical rule which describes the time evolution of a point's position in space. Examples include the mathematical models that describe the trajectory of an aircraft or the evolution of a disease. These relations are described by a set of differential equations. The equations describing the system can be deterministic or stochastic, giving rise to a deterministic or stochastic dynamical system respectively. Similarly, depending on the nature of the governing differential equations, the dynamical system may be linear or non-linear.

A dynamical system has a state determined by a collection of real numbers, or more generally by a set of points in an appropriate state space. The equations defining the system determines the transitions (which could be deterministic or stochastic) between the different states.

In many situations, only a few of the states of the dynamical system may be observed with some measurement error. Usually, the parameters determining the transition rates between the states are also unknown. Henceforth, we will call such a system a partially observed dynamical system. If the evolution of such a dynamical system follows the Markov property, we will use the term partially observed Markovian dynamical system to denote it. As explained before, such a dynamical system can be deterministic or stochastic.

A common statistical problem in dynamical systems is to estimate the unknown parameters determining the dynamical system. Several techniques, like the least square, Bayesian or maximum likelihood estimation have been developed for this,

as surveyed in Cappé et al. (2005). A particular procedure, developed by Ionides et al. (2006) describes an algorithm that uses sequential Monte Carlo filter (also known as Particle Filter) to recursively compute the maximum likelihood estimate of the model parameters. As will be shown in the subsequent chapters, the procedure is also effective in the estimation of time constant parameters, a traditionally hard problem.

## 1.2 Plug and play inference for dynamic models

In this dissertation, we focus our attention on a specific type of dynamic models, which we call mechanistic models. What this means is the model tries to take into account the evolution of the underlying process from a phenomenological point of view. The main concern in the model building exercise is to conform to the scientific knowledge about the dynamical system under consideration, rather than the ease of statistical modeling. The process of mechanistic modeling thus involves writing down a system of equations describing the time evolution of the process that follows from a scientific understanding of the system (state equations). Equations are also written down to connect the evolution of the state processes to observed quantities (observation equations). These models will often contain non-linearity and non-Gaussianity that arise naturally in many physical phenomena unless one is trying to simplify analysis by making things artificially (in some cases) conform to the illusion of linear Gaussianity.

Once such a model is built, one can consider several questions of statistical as well as scientific nature. Examples of statistical questions will include considerations of model fit, range of plausible values of model parameters as well as their identifiability. The scientific questions will often be concerned about the interpretation of model parameters and will try to connect them meaningfully to questions of practical utility. These two types of questions are of complementary, rather than

of conflicting nature in a mechanistic setting.

Simulation based inference techniques are often popular for mechanistic models. These techniques essentially involve a comparison of sample paths generated from the equations describing the state process of the model to data. Though some ad-hoc metrics can be used for this comparison, especially when computation of likelihood is expensive, e.g., Approximate Bayesian Computation of Marjoram et al. (2003), or gradient matching of Ellner et al. (2002), it is also possible to employ simulation based techniques when one wishes to carry out likelihood based inference (Ionides et al., 2006). We reserve the term “plug and play” for the class of inference techniques that require only simulations from the state transition densities and not its explicit evaluation. Plug and play can still require explicit evaluation of the observation density. A similar terminology used in the literature for this type of inference procedure is “equation free” (Kevrekidis et al., 2004; Xiu et al., 2005).

We now list the properties of the general type of dynamical systems we will treat in this dissertation. These are (a) partially observed, (b) continuous-time, (c) nonlinear, (d) Markovian and (e) stochastic. This combination of properties arises naturally since

(a) In many physical processes it is impossible to observe or measure all the required states in a system.

(b) The underlying process in many naturally arising systems are best described in continuous-time, although observations are usually only available in discrete time points.

(c) The type of systems we focus on come from epidemiology where nonlinearity is very common. Examples can be found in chapter 3. Nonlinearity is found in numerous other physical processes.

(d) If all quantities required to describe the state transitions are included in the

present state, the future states will be independent of the past given the present, i.e., Markovian.

(e) Stochasticity is the link that explains the difference between the data and the solution to noise-free deterministic equations. We will mainly be concerned with two sources of stochasticity. The first arises from a consideration that our process model is not a perfect representation of the underlying dynamical process and hence a source of error (process noise) is introduced in the equations defining the state transitions of the model. The second source of error connects the hidden states to the observations (measurement noise).

Consideration of only measurement noise reduces our inference problem to non-linear regression, which is well-studied. Perfectly observed dynamical systems, i.e. systems with only process noise but no observation noise, are also amenable to simpler treatments (Basawa and Prakasa Rao, 1980). We will, however, treat both sources of noise simultaneously here, i.e. the more general case.

Considerable work has been done on these models (e.g., Anderson and Moore, 1979; Liu, 2001; Doucet et al., 2001) but methodologies that are applicable to a wide range of models encompassing all the properties (a)-(e) as described above, have been hard to find. Since we are mainly going to focus on plug and play type inference methodology for partially observed dynamical systems, we now give some background on previous works in that.

Kendall et al. (1999) have proposed a method of simulated moments approximating the likelihood. Iterated filtering of Ionides et al. (2006) computes the maximum likelihood estimates in a partially observed Markov model via a plug and play sequential Monte Carlo filter. Approximate Bayesian Sequential Monte Carlo method of Liu and West (2001) is another viable alternative.

A recent addition to the toolbox of Bayesian plug and play approaches is the Particle Markov chain Monte Carlo technique of Andrieu et al. (2010). This ap-

proach designs the proposal distributions in a Markov chain Monte Carlo scheme with the help of sequential Monte Carlo and has potential to be successful in settings where a choice of a good proposal distribution is not obvious. We provide a comparative analysis of Particle Markov chain Monte Carlo and iterated filtering of Ionides et al. (2006) in section 1.2.1 and show that iterated filtering compares favorably to state-of-the-art Bayesian techniques, thus providing some motivation for a detailed study of iterated filtering in chapter 2.

### **1.2.1 Example of a comparative analysis of the state of the art in Bayesian and likelihood based plug and play inference: Particle Markov chain Monte Carlo vs Iterated Filtering**

Before we discuss the properties of iterated filtering, we wish to show a comparative analysis of iterated filtering and Particle Markov chain Monte Carlo of Andrieu et al. (2010), which we believe are the state of the art in likelihood based and Bayesian plug and play techniques for partially observed Markovian dynamical systems. This example is adapted from Bhadra (2010).

Andrieu et al. (2010) present an elegant theory for novel methodology which makes Bayesian inference practical on state space models. We use their example, a sophisticated financial model involving a continuous time stochastic volatility process driven by Lévy noise, to compare their methodology with a state-of-the-art non-Bayesian approach. We applied iterated filtering (Ionides et al., 2006; Ionides et al., 2009) implemented via the `mif` function in the R package `pomp` (King et al., 2009). We describe the model here in brief as the purpose is just to compare the two inference procedures. For a detailed explanation of the models, the reader is referred to Andrieu et al. (2010). The Levy-driven stochastic volatility model

defining the stock prices is

$$\frac{dy(t)}{y(t)} = (\mu + \beta\sigma^2(t))dt + \sigma(t)dB(t) \quad (1.1)$$

$$d\sigma^2(t) = -\lambda\sigma^2(t)dt + dz(\lambda t) \quad (1.2)$$

where  $y(t)$  is the price of a financial asset,  $\mu$  is the drift parameter,  $\beta$  is the risk premium and  $B(t)$  is a Brownian motion. Models like these are developed in Barndorff-Nielsen and Shephard (2001) and have become popular in financial econometrics (Andrieu et al., 2010). Equation (1.2) defines the volatility  $\sigma^2(t)$  as a Levy driven Ornstein-Uhlenbeck process where  $\lambda > 0$  and  $z(t)$  is a purely non-Gaussian Levy process. Now it is easy to see that the state and observation processes defining this dynamical system are  $(\sigma^2(t), z(t))$  and  $y(t)$  respectively. After discretization, we have the following form for the two dimensional state process,

$$(\sigma^2(n\Delta), z(\lambda n\Delta)) = (\exp(-\lambda\Delta)\sigma^2((n-1)\Delta), z(\lambda(n-1)\Delta)) + \eta_n$$

for  $n = 1, \dots, N$ . where,

$$\eta_n = \sum_{i=1}^{\infty} \left\{ \left( \frac{a_i \kappa}{A \lambda \Delta} \right)^{-1/\kappa} \wedge e_i v_i^{1/\kappa} \right\} (\exp(-\lambda \Delta r_i), 1) + \sum_{i=1}^{Z(\lambda \Delta)} c_i (\exp(-\lambda \Delta r_i^*), 1) \quad (1.3)$$

where  $A = 2^\kappa \delta \kappa^2 / \Gamma(1 - \kappa)$ . Let  $B = 1/2\gamma^{1/\kappa}$ . Then in equation (1.3),  $a_i, e_i$  and  $v_i$  are independent;  $e_i$  are iid exponential with mean  $1/B$ ;  $\Delta$  is the time step for discretization of the continuous-time model;  $v_i, r_i, r_i^*$  are standard uniform random variables and  $a_1 < a_2 < \dots$  are arrival times of a Poisson process of intensity 1;  $c_i$  are iid  $G(1 - \kappa, 1/B)$ . Here  $G(a, b)$  denotes the Gamma distribution with mean  $ab$  and variance  $ab^2$ .  $Z(\lambda\Delta)$  is a Poisson random variable with mean  $\lambda\Delta\delta\gamma\kappa$ . The

observation process is defined as

$$y_n \sim N(\mu\Delta + \beta\psi_n^2, \psi_n^2)$$

for  $n = 1, \dots, N$ . where,

$$\psi_n^2 = \lambda^{-1} [z(\lambda n\Delta) - \sigma^2(n\Delta) - z(\lambda(n-1)\Delta) + \sigma^2((n-1)\Delta)]$$

In their example Andrieu et al. (2010) use  $\mu = 0, \beta = 0, \Delta = 1, N = 400$  and thus the problem is to estimate  $\delta, \lambda, \kappa$  and  $\gamma$ . Andrieu et al. (2010) use the following parameter values in their simulation study  $(\kappa, \delta, \gamma, \lambda) = (0.5, 1.41, 2.83, 0.10)$ . For their Bayesian analysis, they assume the following priors,  $\kappa \sim B(10, 10), \delta \sim G(1, \sqrt{50}), \gamma \sim G(1, \sqrt{200})$  and  $\lambda \sim G(1, 0.5)$ . Here  $B(\alpha, \beta)$  denotes the Beta distribution with mean  $\alpha/(\alpha + \beta)$  and variance  $\alpha\beta/(\alpha + \beta)^2(\alpha + \beta + 1)$ . A normal random-walk Metropolis-Hastings algorithm is used for joint updating of parameters. The authors use MCMC runs of length 50,000 and report possible lack of identifiability of parameters in the model.

Fig. 1.1 shows some results from applying the iterated filtering algorithm with 1000 particles to the simulation study described by Andrieu et al. (2010, section 3.2) in order to carry out likelihood based inference. If  $\theta$  denotes the parameter vector of interest, the algorithm generates a sequence of parameter estimates  $\hat{\theta}_1, \hat{\theta}_2, \dots$  converging to the maximum likelihood estimate  $\hat{\theta}$ . As a diagnostic, the log-likelihood of  $\hat{\theta}_i$  is plotted against  $i$  (Fig. 1.1(a)). We see the sequence of log-likelihoods rapidly converges. On simulation studies like this, a quick check for successful maximization is to observe that the maximized log-likelihood typically exceeds the log-likelihood at the true parameter value by approximately half the number of estimated parameters (Fig. 1.1(a)). One can also check for successful local maximization by sliced likelihood plots (Fig. 1.1(b-e)), in which the likelihood

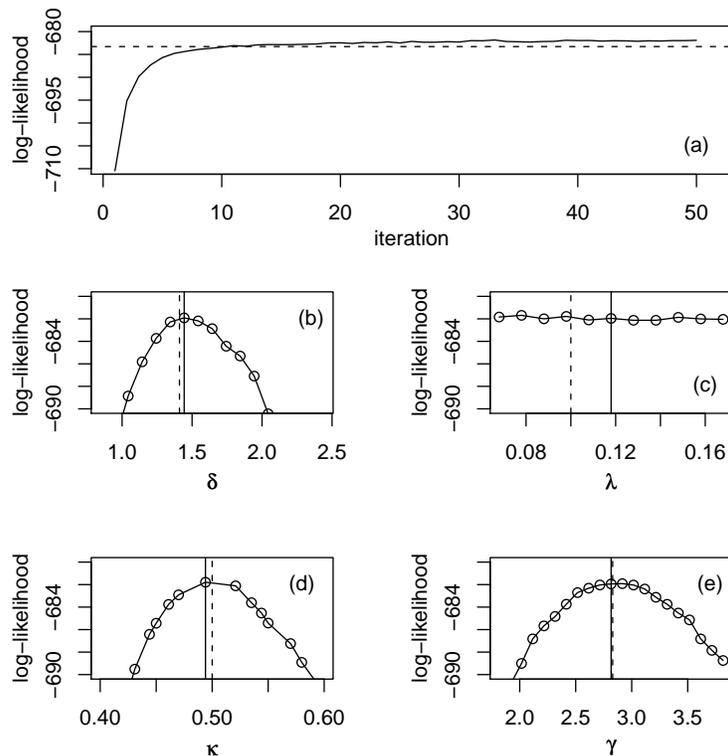


Figure 1.1: Diagnostic plots for iterated filtering. (a) likelihood at each iteration, evaluated by sequential Monte Carlo; the broken straight line marks the likelihood at the truth. (b) - (e) likelihood surface for each parameter sliced through the maximum; points show parameter values where the likelihoods were evaluated; solid straight lines mark the maximum likelihood estimate; broken straight lines mark the true parameter value.

surface is explored along one of the parameters, keeping the other parameters fixed at the estimated local maximum. The likelihood surface is seen to be flat as  $\lambda$  varies, consistent with the authors' observation that some parameter combinations are weakly identified in this model. A profile likelihood analysis could aid the investigation of the identifiability issue. Due to the quick convergence of iterated filtering with a relatively small number of particles, many profile likelihood plots can be generated at the computational expense of, say, one MCMC run of length 50,000.

The decision about whether one wishes to carry out a Bayesian analysis should

depend on whether one wishes to impose a prior distribution on unknown parameters. Here, we have shown that likelihood-based non-Bayesian methodology provides a computationally viable alternative to a state-of-the-art Bayesian approach for complex dynamic models.

### 1.3 Sequential Monte Carlo

Sequential Monte Carlo filters, starting with the bootstrap filter of Gordon et al. (1993) have inspired a huge body of literature over the last 17 years. It is now possible to use these filters in the state estimation problems in nonlinear dynamical systems where the only viable option used to be the Kalman filter with its often hard to meet linear Gaussian assumption, or its variants using local linearity or Gaussianity, e.g., the extended Kalman filter (Arulampalam et al., 2002).

This inference technique used in chapter 2 uses the basic Sequential Monte Carlo (SMC) filter in order to compute the conditional estimates of the state variables. Here, we describe the basic SMC algorithm that is necessary to understand chapter 2.

The basic SMC algorithm is based on the following identities in a Markovian state space model

$$f_{X_t|Y_{1:t}}(x_t|y_{1:t}; \theta) = \frac{f_{X_t|Y_{1:t-1}}(x_t|y_{1:t-1}; \theta)f_{Y_t|X_t}(y_t|x_t; \theta)}{\int f_{X_t|Y_{1:t-1}; \theta}(x_t|y_{1:t-1}; \theta)f_{Y_t|X_t}(y_t|x_t; \theta)dx_t} \quad (1.4)$$

$$f_{X_{t+1}|Y_{1:t}}(x_{t+1}|y_{1:t}; \theta) = \int f_{X_{t+1}|X_t}(x_{t+1}|x_t; \theta)f_{X_t|Y_{1:t}}(x_t|y_{1:t}; \theta)dx_t \quad (1.5)$$

In this section it is understood that subscripts to the letter  $f$  corresponds to the random variables whose densities are considered. All these conditional densities are assumed to exist. The quantities on the left hand sides of equations (1.4) and (1.5) are known as the filter and prediction densities respectively. Following standard notation in SMC literature,  $x_t$  denotes the unobserved state process,  $y_t$

denotes the observations,  $\theta$  is a vector of parameters and  $y_{1:t} = (y_1, \dots, y_t)$ . It is also assumed that  $x_{1:t}$  and  $y_{1:t}$  are realizations for real valued random variables  $X_{1:t}$  and  $Y_{1:t}$ . This allows us to define the following algorithm

1. Suppose the set of “particles”  $\{X_{t,j}^F, j = 1, \dots, J\}$  be approximately distributed as the conditional density of the state process  $f_{X_t|Y_{1:t}}(x_t|y_{1:t}; \theta)$ .

2. The mutation or prediction step:

Sample  $X_{t+1,j}^P$  from the state transition density  $f_{X_{t+1}|X_t}(x_{t+1}|x_t = X_{t,j}^F; \theta)$ .

Then  $\{X_{t+1,j}^P\}$  has a marginal density  $f_{X_{t+1}|Y_{1:t}}(x_{t+1}|y_{1:t}; \theta)$

3. The selection step:

Resample with replacement  $\{X_{t+1,j}^P\}$  according to their weights  $w_j = f_{Y_t|X_t}(y_t|x_t = X_{t,j}^P; \theta)$ . Note that the weights need to be evaluated. Now  $\{X_{t+1,j}^F\}$  solves the filtering problem at time  $t + 1$ . Go back to step 1 and repeat for  $t = t + 1$ .

4.  $\tilde{E}[x_t|y_{1:t}]$  and  $\tilde{V}[x_t|y_{1:t-1}]$  are the sample mean and sample variance of  $X_{t,j}^F$  and  $X_{t,j}^P$  respectively. These quantities are available as a result of running steps 1 through 3 and will be useful in the iterated filtering algorithm described in chapter 2.

In step 2, note that we draw samples from  $f_{X_{t+1}|X_t}(x_{t+1}|x_t; \theta)$ . This lies at the heart of the plug and play property, since this essentially means we can generate samples from the equations describing the model trajectory, without worrying about the presence of a closed-form state transition density. Step 2 can be modified in various ways (e.g. the auxiliary particle filter of Pitt and Shephard (1999)) that takes advantage of the present observation  $y_t$ . However, that modification comes at the expense of the plug and play property.

Figure 1.2 shows a pictorial representation of the sequential Monte Carlo scheme. Row (a) shows  $J$  particles distributed according to the prediction density at time

$t - 1$ . Each of these  $J$  particles are then weighted according to step 3 of the algorithm described above and the corresponding weights are called  $w_j$  (row (b)). Note how particles whose weights fall below a certain threshold (unfit particles, using evolutionary terms) are eliminated at this stage. Then resampling with replacements according to  $w_j$  takes place among the surviving particles and heavier particles can be resampled more than once (gives rise to more descendants). This is shown in row (c) in yellow and we resample until we reach  $J$  particles (strictly speaking,  $J$  can vary between time points). Each of these particles are then perturbed according to step 2 of the algorithm. Thus, one prediction and one selection step (steps 2 and 3) describe the complete set of operations for one time point. The resultant row (d) shows the progression of the set of particles from time  $t - 1$  to  $t$ . The set of operations is then repeated for the next time point. Note the correspondence between rows (a) and (d), (b) and (e).

The resampling step of the algorithm (step 3) is crucial. In absence of it after only a few time steps, all the weights are concentrated on only one particle and the rest of the particles are eliminated (Gordon et al., 1993). Thus, with only one surviving particle, the algorithm fails to describe the required densities. However, the resampling stage is not without its share of problems. Most importantly, it results in the so called “sample impoverishment”. Heavier particles are resampled many times, introducing a lack of diversity in the resultant sample. As is easy to see in the transitions from row (c) to (d) in figure 1.2, the only source of diversity is then the randomness introduced by the independent draws from  $f_{X_{t+1}|X_t}(x_{t+1}|x_t; \theta)$ , i.e., the process noise. Hence this problem is most severe in the case of small process noise and the particles collapse to a single point within a few iterations (Arulampalam et al., 2002). Also, this lack of diversity among the particles explain why smoothers based on particles’ paths degenerate (Arulampalam et al., 2002). Strategies to counter sample impoverishment is an active area of research in it-

self (Gilks and Berzuini, 2001). We will not focus on these problems any further in this dissertation.

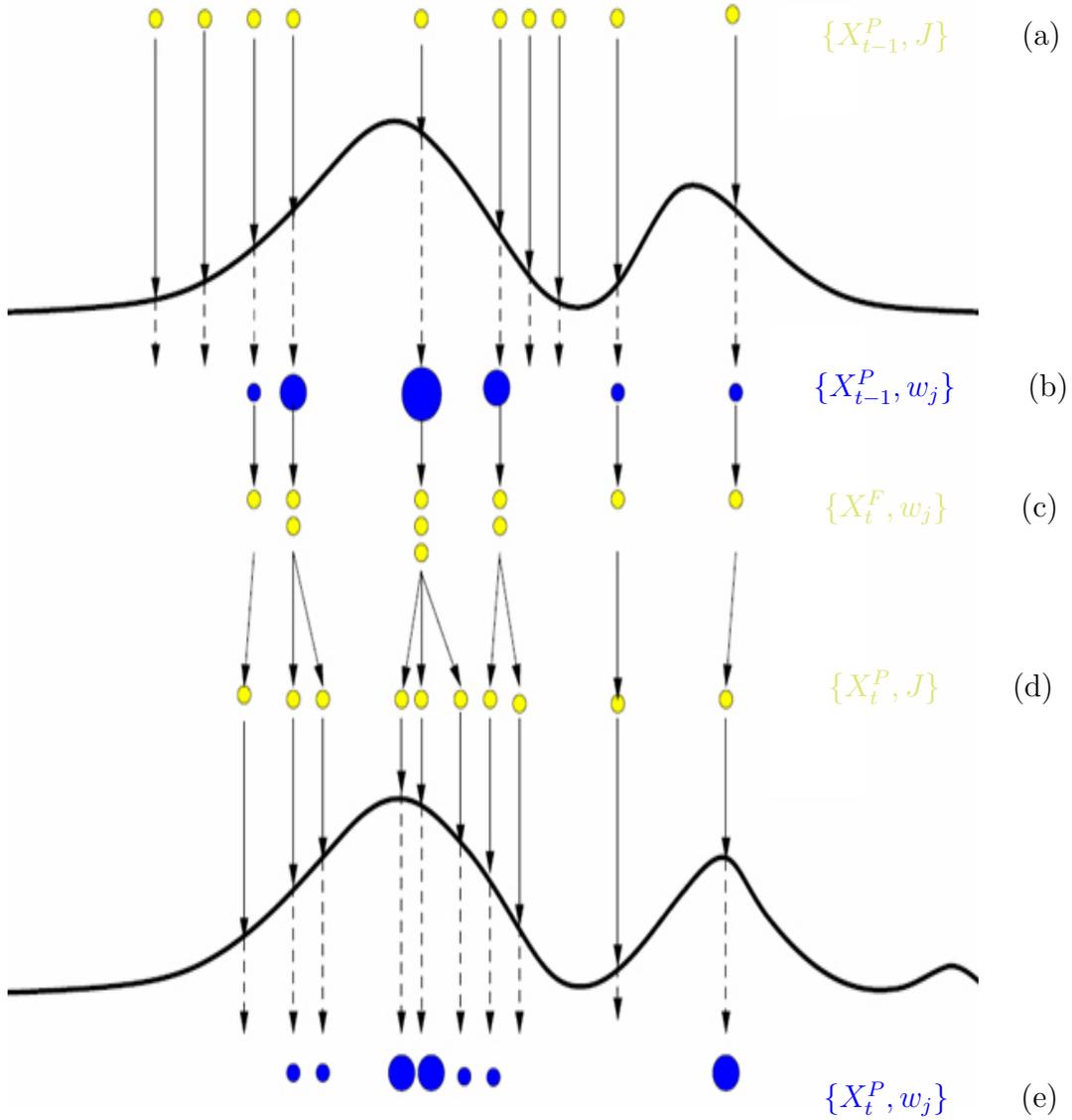


Figure 1.2: A pictorial representation of the time evolution of a sequential Monte Carlo scheme adapted from Doucet et al. (2001). The arrows point to the direction of increasing time. Superscripts P and F denote prediction and filtering respectively. The curved solid line denotes the likelihood surface at a given time. The blue and yellow balls represent the different particles in the state space at a given time. The radii of the balls are proportional to their importance weights.

## 1.4 Infectious disease models

It is useful to model the evolution of an infectious disease as a dynamical system. Information needed to build a dynamic model is often available from the biological considerations of the model and time series data of disease cases are available to fit and test the models. Generally, the states of the system are defined keeping in mind the separate stages in the evolution of the disease e.g., individuals who are susceptible to the disease, recovered from the disease or currently infected etc. Usually, not all the states can be observed, i.e. the number of people currently undergoing treatment for the disease at a hospital might be counted but there is no way to exactly count the people who are susceptible to the disease at any given instant. It is therefore an example of a partially observed dynamical system. Usually the transition rates are unknown, and have to be estimated from the data, though an initial guess is often available based on prior scientific knowledge.

There are several illustrative examples of the use of dynamical system in the modeling of an infectious disease. We will focus our attention to the compartment model, where each compartment denotes a state in the system. The classic SIR model by Kermack and McKendrick (1927) groups  $N_t$  individuals as susceptible ( $S_t$ ), infected ( $I_t$ ) and recovered ( $R_t$ ). Inclusion of exposed classes, age-structured classes are some natural extensions. In chapter 3, we will use one such compartment model for malaria transmission. Population models can be stochastic or deterministic, may be in discrete or continuous time and may take discrete or continuous values. Real world processes are continuous time, discrete values and stochastic. The stochasticity could be demographic (arising due to uncertainty in individual outcomes) or environmental (arising due to factors affecting the whole population, e.g. climate or economy). It is sensible to model the variance of demographic stochasticity as linear in population size and that environmental stochasticity as quadratic in population size.

Mechanistic modeling of dynamical systems offers a way to integrate ecological parameters into disease models and to study the relationship between the two. One of the main objectives of this dissertation is to describe such a partially observed stochastic dynamic system model of the evolution of malaria developed by Laneri et al. (2010). We will use the iterated filtering procedure developed by Ionides et al. (2006) to estimate the parameters for this system. A central scientific question for epidemic malaria has been whether the interannual cycles observed in the epidemics are driven by climate or are instead generated by the intrinsic dynamics of the disease itself or a combination of the two. The particular climate driver we are interested in this case is rainfall. Determining the role of climate on disease transmission is also central in building an early warning system for epidemics. We discuss a mechanistic model of malaria transmission to explicitly test the hypothesis of climate forcing versus intrinsic dynamics in chapter 3.

## CHAPTER 2

# Iterated Filtering

### 2.1 Introduction

Partially observed Markov process (POMP) models 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 Rubio-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).

This chapter investigates theoretical properties of a technique for estimating unknown parameters of POMPs, called *iterated filtering*, which was proposed by Ionides et al. (2006). Iterated filtering methodology provides simple algorithms to compute maximum likelihood estimates for a very general class of partially observed Markov process models. At each subsequent filtering iteration, a class of approximating models converges toward the POMP under investigation. The theoretical foundation of iterated filtering is a novel relationship between certain conditional moments of the approximating models and the derivative of the log likelihood, presented as Theorem 2.2.1 below. In several case-studies, iterated fil-

tering algorithms have been shown capable of addressing scientific challenges in the study of infectious disease transmission in large populations (Ionides et al., 2006; King et al., 2008; Bretó et al., 2009; He et al., 2010). The partially observed nonlinear stochastic systems arising in the study of disease transmission and related ecological systems are a challenging environment to test statistical methodology (Bjørnstad and Grenfell, 2001), and many statistical methodologies have been tested on these systems in the past fifty years (e.g., Cauchemez and Ferguson, 2008; Toni et al., 2008; Keeling and Ross, 2008; Ferrari et al., 2008; Morton and Finkenstadt, 2005; Grenfell et al., 2002; Kendall et al., 1999; Bartlett, 1960; Bailey, 1955). Since iterated filtering has already demonstrated potential for generating state-of-the-art analyses on a major class of scientific models, we are motivated to investigate its theoretical properties. It is beyond the scope of this chapter to investigate whether alternative methodologies which are untested, as yet, on this class of scientific problems could compete as effective inference tools. Instead, the goal of this chapter is to present the first formal and complete mathematical analysis of an iterated filtering algorithm implemented via sequential Monte Carlo. There are obvious potential extensions, both theoretical and methodological, of the iterated filtering algorithm studied here. Similar procedures could be developed for multiple time series (longitudinal data analysis), and for more general situations where the likelihood is most readily calculated by sequential Monte Carlo or importance sampling algorithms. This chapter provides a theoretical foundation for future work on iterated filtering techniques.

One extensively studied issue for POMPs 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. Iterated filtering algorithms repeatedly carry out a filtering procedure to explore the likelihood surface at increasingly local scales in search of a maximum of the likelihood function.

Basic iterated filtering algorithms, such as the one studied in this chapter, have an attractive practical property that the dynamic model enters the algorithm only through the requirement that sample paths can be generated at arbitrary parameter values. This property has been called *plug-and-play* (Bretó et al., 2009; He et al., 2010) 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. Other *plug-and-play* methods proposed for partially observed Markov models include approximate Bayesian computations implemented via sequential Monte Carlo (Liu and West, 2001; Toni et al., 2008), an asymptotically exact Bayesian technique combining sequential Monte Carlo with Markov chain Monte Carlo (Andrieu et al., 2010), simulation-based forecasting (Kendall et al., 1999), and simulation-based spectral analysis (Reuman et al., 2006).

Stochastic expectation-maximization and Markov chain Monte Carlo methods for partially observed Markov process models have been extensively studied (reviewed by Cappé et al., 2005) but have limitations and complications when the unobserved dynamic system operates in continuous time (Roberts and Stramer, 2001; Beskos et al., 2006). These difficulties arise because standard stochastic expectation-maximization and Markov chain Monte Carlo approaches require the evaluation of transition probability densities for trajectories of the unobserved dynamic system at various values of the unknown parameters. These transition densities can be singular (for example, in many diffusion models) or close to singular (for example, in models that are well approximated by diffusion processes). The need for evaluating transition probability densities also means that stochastic expectation-maximization and Markov chain Monte Carlo methods typically do not enjoy the plug-and-play property. We see, therefore, that plug-and-play methods are not only computationally convenient but also reduce the technical difficulties of working with continuous time dynamic process models.

In Section 2.2 we develop a new theoretical framework 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. Furthermore, the analogous result to Theorem 2.2.1 below (Theorem 1 of Ionides et al., 2006) erroneously fails to include the two separate scales  $\tau$  and  $\sigma$ . 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 complete theory presented here gives additional insights into the capabilities, limitations and practical implementation of iterated filtering.

Section 2.2 presents our main results, drawing on the theories of stochastic optimization and sequential Monte Carlo. Our goal is not to employ the most recent

results available in each of these research areas, but rather to show that some fundamental and well-known results from both areas can be combined with a theorem extending Ionides et al. (2006) to synthesize a new theoretical understanding of iterated filtering. Proofs of the theorems stated in Section 2.2 are presented to Section 2.3. Section 2.4 discusses the relationship between this theory and the practice of iterated filtering. Since the asymptotic justification of iterated filtering in Section 2.2 is somewhat technical (in contrast to the simplicity of the algorithm itself), Section 2.4 carries the important burden of explaining how our results relate to practical data analysis considerations. As one aspect of this relationship, we argue that users of iterated filtering methodology can draw on the extensive literature on the practice and theory of stochastic optimization by simulated annealing (Kirkpatrick et al., 1983; Ingber, 1993; Spall, 2003).

## 2.2 Notation and main results

Let  $\{X(t), t \in T\}$  be a Markov process taking values in  $\mathfrak{X} \subset \mathbb{R}^{d_x}$  (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  $t_1 < t_2 < \dots < t_N$  at which  $X(t)$  is observed, together with some initial time  $t_0 < t_1$ . We write  $X_{0:N} = (X_0, \dots, X_N) = (X(t_0), \dots, X(t_N))$ . We denote the sequence of observations by  $y_{1:N} = (y_1, \dots, y_N)$ , with  $y_n$  taking a value in  $\mathbb{R}^{d_y}$ . We refer to the observation sequence  $y_{1:N}$  as the data, and this is considered as fixed throughout this chapter. The data are modeled as a realization of a sequence of random variables  $Y_{1:N} = (Y_1, \dots, Y_N)$ . 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  $\theta$  taking a value in  $\mathbb{R}^p$ . A POMP may then be specified at times  $t_0, t_1, \dots, t_N$  by an initial density  $f_{X_0}(x_0; \theta)$ , conditional transition densities  $f_{X_n|X_{n-1}}(x_n | x_{n-1}; \theta)$  for  $1 \leq n \leq N$ , and the conditional densities

of the observation process which have the form  $f_{Y_n|Y_{1:n-1}, X_{1:n}}(y_n | y_{1:n-1}, x_{1:n}; \theta) = f_{Y_n|X_n}(y_n | x_n; \theta)$ . The subscripts of  $f$  denote the variables to which the density corresponds. We use a semicolon to separate the values of random variables from the values of static parameters. We write  $f$  without subscripts to denote the full collection of densities and conditional densities, and we call  $f$  the *generic density* of a POMP.

Iterated filtering involves introducing a sequence of approximations to the model  $f$  in which a time-varying parameter process  $\{\Theta_n, 0 \leq n \leq N\}$  is introduced. Specifically, equations (2.1–2.3) define a model  $g$  for a Markov process  $\{(X_n, \Theta_n), 0 \leq n \leq N\}$  and observation process  $Y_{1:N}$ .

$$g_{X_n, \Theta_n | X_{n-1}, \Theta_{n-1}}(x_n, \theta_n | x_{n-1}, \theta_{n-1}; \theta, \sigma, \tau) = f_{X_n | X_{n-1}}(x_n | x_{n-1}; \theta_{n-1}) \sigma^{-p} \kappa\left(\frac{\theta_n - \theta_{n-1}}{\sigma}\right),$$

$$g_{Y_n | X_n, \Theta_n}(y_n | x_n, \theta_n; \theta, \sigma, \tau) = f_{Y_n | X_n}(y_n | x_n; \theta_n), \quad (2.2)$$

$$g_{X_0, \Theta_0}(x_0, \theta_0; \theta, \sigma, \tau) = f_{X_0}(x_0; \theta_0) \tau^{-p} \kappa\left(\frac{\theta_0 - \theta}{\tau}\right). \quad (2.3)$$

Here,  $\kappa$  is a probability density function on  $\mathbb{R}^p$  which specifies a random walk for  $\theta_n$ . From (2.1), the increments of the random walk are independent of the current state of the process  $x_n$ . We suppose that the distribution corresponding to  $\kappa$  has mean zero and covariance matrix  $\Sigma$ , so that

$$E_g[\Theta_n | \Theta_{n-1}] = \Theta_{n-1}, \quad \text{Var}_g(\Theta_n | \Theta_{n-1}) = \sigma^2 \Sigma, \quad (2.4)$$

$$E_g[\Theta_0] = \theta, \quad \text{Var}_g(\Theta_0) = \tau^2 \Sigma. \quad (2.5)$$

The subscripts for  $E_g$  and  $\text{Var}_g$  identify that expectation and variance are taken with respect to the model  $g$ . One natural choice of  $\kappa$  is a multivariate normal density, which must be truncated to meet condition (A3) of Theorem 2.2.1.

Assuming  $f$  is continuously parameterized as a function of  $\theta$ , which follows

from (A2) below, we see from (2.1–2.3) that  $g_{X_{0:N}, Y_{1:N}}(x_{0:N}, y_{1:N}; \theta, \sigma, \tau)$  approaches  $f_{X_{0:N}, Y_{1:N}}(x_{0:N}, y_{1:N}; \theta)$  as both  $\sigma \rightarrow 0$  and  $\tau \rightarrow 0$ . We call the model  $g$  a *perturbation* of the model  $f$ . We refer to  $\sigma, \tau, \kappa$  and  $\Sigma$  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  $\ell(\theta) = \log f_{Y_{1:N}}(y_{1:N}; \theta)$ . We write  $\nabla$  for a vector of partial derivatives with respect to each component of  $\theta$ , and  $\nabla^2$  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 2.2.1 relates this derivative to the filtering means and prediction variances for  $g$ , defined as

$$\begin{aligned} \theta_n^F &= \theta_n^F(\theta, \sigma, \tau) = E_g[\Theta_n | Y_{1:n} = y_{1:n}] = \int \theta_n g_{\Theta_n | Y_{1:n}}(\theta_n | y_{1:n}; \theta, \sigma, \tau) d\theta_n \\ V_n^P &= V_n^P(\theta, \sigma, \tau) = \text{Var}_g(\Theta_n | Y_{1:n-1} = y_{1:n-1}) \end{aligned} \quad (2.6)$$

for  $n = 1, \dots, N$ , with  $\theta_0^F = \theta$ . 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) (i) There is a constant  $C_1(\theta)$  such that  $\sup_{x_n} f_{Y_n | X_n}(y_n | x_n; \theta) \leq C_1(\theta)$  for all  $n$ . Additionally,  $C_1(\theta)$  is bounded on compact subsets of  $\mathbb{R}^p$ . (ii) For all  $\theta$ ,  $f_{Y_{1:N}}(y_{1:N}; \theta) > 0$ .

(A2) Defining  $\theta^{[k]}$  to be  $k$  concatenated copies of  $\theta$ ,  $g_{Y_{1:n} | \Theta_{0:n}}(y_{1:n} | \theta_{0:n})$  is twice continuously differentiable as a function of  $\theta_{0:n}$  at  $\theta_{0:n} = \theta^{[n+1]}$ .

(A3) There is a constant  $C_2$  with  $\kappa(\theta) = 0$  for  $|\theta| > C_2$  and  $\kappa(\theta) > 0$  for  $|\theta| \leq C_2$ .

(A4)  $\kappa(\theta)$  is twice differentiable on  $\{|\theta| < C_2\}$ , and  $\nabla^2 \kappa(\theta)$  is Lipschitz continuous.

Condition (A1) is not restrictive. Condition (A2) gives a way of specifying that the likelihood surface is smoothly parameterized. The conditional density

$$g_{Y_{1:n}|\Theta_{0:n}}(y_{1:n}|\theta_{0:n}) = \int f_{X_0}(x_0;\theta_0) \prod_{k=1}^n f_{Y_k|X_k}(y_k|x_k;\theta_k) f_{X_k|X_{k-1}}(x_k|x_{k-1};\theta_{k-1}) dx_{0:n} \quad (2.7)$$

does not depend on  $\theta$ ,  $\sigma$ ,  $\tau$  or the choice of the perturbation kernel  $\kappa$ . The relationship between smoothness of the likelihood surface, the transition density  $f_{X_k|X_{k-1}}(x_k|x_{k-1};\theta)$ , and the observation density  $f_{Y_k|X_k}(y_k|x_k;\theta)$  is simple to establish only under the restrictive condition that  $\mathfrak{X}$  is a compact set (Jensen and Petersen, 1999). Therefore, we note an alternative to (A2) which is more restrictive but more readily checkable:

(A2')  $\mathfrak{X}$  is a compact subset of  $\mathbb{R}^{d_x}$ . Both  $f_{X_k|X_{k-1}}(x_k|x_{k-1};\theta)$  and  $f_{Y_k|X_k}(y_k|x_k;\theta)$  are twice continuously differentiable with respect to  $\theta$ . These derivatives are also continuous with respect to  $x_{k-1}$  and  $x_k$ .

Conditions (A3) and (A4) can be satisfied by the choice of the algorithmic parameters. The assumption of a spherical support for  $\kappa$  in (A3) is mathematically convenient but we believe this requirement could be relaxed to some more general assumption of compact support.

**Theorem 2.2.1.** *Suppose conditions (A1–A4). Let  $\sigma$  be a function of  $\tau$  with  $\sigma\tau^{-3} \rightarrow 0$  as  $\tau \rightarrow 0$ . Using notation from (2.6),*

$$\lim_{\tau \rightarrow 0} \sum_{n=1}^N (V_n^P)^{-1} (\theta_n^F - \theta_{n-1}^F) = \nabla \ell(\theta). \quad (2.8)$$

A proof of Theorem 2.2.1 is given in Section 2.3.1, based on a Taylor series expansion around  $\theta_n = \theta_{n-1}^F$  of  $g_{Y_n|Y_{1:n-1},\Theta_n}(y_n|y_{1:n-1},\theta_n;\theta,\sigma,\tau)$ . Theorem 2.2.1 builds on a result of Ionides et al. (2006), however both the assumptions employed and the details of the proof differ substantially from this previous article. Our new

result is necessary for the following Theorems 2.2.2 and 2.2.3.

The quantities  $\theta_n^F$  and  $V_n^P$  in Theorem 2.2.1 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 2.2.1 may be useful. However, one might suspect that there is no “free lunch” and therefore the numerical calculation of the left hand side of (2.8) should become fragile as  $\sigma$  and  $\tau$  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 method presented as Algorithm 1. Sequential Monte Carlo 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 sequential Monte Carlo 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., 2010). We suspect that the qualitative conclusions obtained here would apply to variations on Algorithm 1.

To calculate Monte Carlo estimates of the quantities in (2.6), we apply Algorithm 1 to the model  $g$  with  $Z_n = (X_n, \Theta_n)$ ,  $\psi = (\theta, \sigma, \tau)$  and  $J$  particles. We write  $Z_{n,j}^F = (X_{n,j}^F, \Theta_{n,j}^F)$  and  $Z_{n,j}^P = (X_{n,j}^P, \Theta_{n,j}^P)$  for the Monte Carlo samples from the filtering and prediction calculations in Algorithm 1. Then, using  $x^T$  to denote the transpose of  $x$ , we define

$$\begin{aligned}\tilde{\theta}_n^F &= \tilde{\theta}_n^F(\theta, \sigma, \tau, J) = \frac{1}{J} \sum_{j=1}^J \Theta_{n,j}^F, \\ \tilde{V}_n^P &= \tilde{V}_n^P(\theta, \sigma, \tau, J) = \frac{1}{J-1} \sum_{j=1}^J (\Theta_{n,j}^P - \tilde{\theta}_{n-1}^F)(\Theta_{n,j}^P - \tilde{\theta}_{n-1}^F)^T.\end{aligned}\tag{2.9}$$

**Input:**

- POMP model described by a generic density  $h$  having parameter vector  $\psi$  and corresponding to a Markov process  $Z_{0:N}$ , observation process  $Y_{1:N}$ , and data  $y_{1:N}$
- number of particles,  $J$

**Procedure:**

- 1 initialize filter particles  $Z_{0,j}^F \sim h_{Z_0}(z_0; \psi)$  for  $j$  in  $1 : J$
- 2 for  $n$  in  $1 : N$
- 3   for  $j$  in  $1 : J$  draw prediction particles  $Z_{n,j}^P \sim h_{Z_n|Z_{n-1}}(z_n | Z_{n-1,j}^F; \psi)$
- 4   set  $w(n, j) = h_{Y_n|Z_n}(y_n | Z_{n,j}^P; \psi)$
- 5   draw  $k_1, \dots, k_J$  such that  $\mathbb{P}\{k_j=i\} = w(n, i) / \sum_{\ell} w(n, \ell)$
- 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. For the unperturbed model, set  $Z_n = X_n$ ,  $h = f$  and  $\psi = \theta$ . For the perturbed model, set  $Z_n = (X_n, \Theta_n)$ ,  $h = g$  and  $\psi = (\theta, \sigma, \tau)$ . 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.

We now present, as Theorem 2.2.2, an analogue to Theorem 2.2.1 in which the filtering means and prediction variances are replaced by their Monte Carlo counterparts. A proof of this result is given in Section 2.3.3. The stochasticity in Theorem 2.2.2 is due to Monte Carlo variability, conditional on the data  $y_{1:N}$ , and we write  $\tilde{E}$  and  $\widetilde{\text{Var}}$  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 2.2.2.** *Let  $\{\sigma_m\}$ ,  $\{\tau_m\}$  and  $\{J_m\}$  be positive sequences with  $\tau_m \rightarrow 0$ ,  $\sigma_m \tau_m^{-3} \rightarrow 0$  and  $\tau_m J_m \rightarrow \infty$ . Define  $\tilde{\theta}_{n,m}^F = \tilde{\theta}_n^F(\theta, \sigma_m, J_m)$  and  $\tilde{V}_{n,m}^P = \tilde{V}_{n,m}^P(\theta, \sigma_m, J_m)$  via (2.9). Supposing conditions (A1–A4),*

$$\lim_{m \rightarrow \infty} \tilde{E} \left[ \sum_{n=1}^N (\tilde{V}_{n,m}^P)^{-1} (\tilde{\theta}_{n,m}^F - \tilde{\theta}_{n-1,m}^F) \right] = \nabla \ell(\theta), \quad (2.10)$$

$$\lim_{m \rightarrow \infty} \tau_m^2 J_m \widetilde{\text{Var}} \left( \sum_{n=1}^N (\tilde{V}_{n,m}^P)^{-1} (\tilde{\theta}_{n,m}^F - \tilde{\theta}_{n-1,m}^F) \right) < \infty, \quad (2.11)$$

with convergence being uniform for  $\theta$  in compact sets.

Theorem 2.2.2 suggests that a Monte Carlo method which leans on Theorem 2.2.1 will require a sequence of Monte Carlo sample sizes,  $J_m$ , which increases faster than  $\tau_m^{-1}$ . Otherwise, the Monte Carlo bias in estimating  $\theta_n^F - \theta_{n-1}^F$ , which is of order  $\tau_m/J_m$ , will eventually dominate the information in  $\theta_n^F - \theta_{n-1}^F$  about  $\nabla \ell(\theta)$ , which is of order  $\tau_m^2$ . Even with  $\tau_m J_m \rightarrow \infty$ , we see from (2.11) that the estimated derivative in (2.10) may have increasing Monte Carlo variability as  $m \rightarrow \infty$ . 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; Andrieu et al., 2005).

Theorem 2.2.3 gives an example of a stochastic approximation procedure, defined by the recursive sequence  $\hat{\theta}_m$  in (2.12). Because each step of this recursion involves an application of the filtering procedure in Algorithm 1, we call (2.12) below 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  $\zeta(t)$  to be a solution to  $d\zeta/dt = \nabla \ell(\zeta(t))$ . Suppose that  $\hat{\theta}$  is an *asymptotically stable equilibrium point*, meaning that (i) for every  $\eta > 0$  there exists a  $\delta(\eta)$  such that  $|\zeta(t) - \hat{\theta}| \leq \eta$  for all  $t > 0$  whenever  $|\zeta(0) - \hat{\theta}| \leq \delta$ , and (ii) there exists a  $\delta_0$  such that  $\zeta(t) \rightarrow \hat{\theta}$  as  $t \rightarrow \infty$  whenever  $|\zeta(0) - \hat{\theta}| \leq \delta_0$ .

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

The conditions (B1–B2) are the basis of the classic results of Kushner and Clark (1978). Although research into stochastic approximation theory has continued (e.g., Kushner and Yin, 2003; Andrieu et al., 2005; Maryak and Chin, 2008), (B1–B2) remain a textbook approach (Spall, 2003). The relative simplicity and elegance of Kushner and Clark (1978) makes an appropriate foundation for investigating the links between iterated filtering, sequential Monte Carlo and stochastic approximation theory. There is, of course, scope for variations on our results based on the diversity of available stochastic approximation theorems. Although neither (B1–B2) nor alternative sufficient conditions 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 chapter. Practical implementation issues are discussed in Section 2.4 below.

**Theorem 2.2.3.** *Let  $\{a_m\}$ ,  $\{\sigma_m\}$ ,  $\{\tau_m\}$  and  $\{J_m\}$  be positive sequences with  $\tau_m \rightarrow 0$ ,  $\sigma_m \tau_m^{-3} \rightarrow 0$ ,  $J_m \tau_m \rightarrow \infty$ ,  $a_m \rightarrow 0$ ,  $\sum_m a_m = \infty$  and  $\sum_m a_m^2 J_m^{-1} \tau_m^{-2} < \infty$ . Specify a recursive sequence of parameter estimates  $\{\hat{\theta}_m\}$  by*

$$\hat{\theta}_{m+1} = \hat{\theta}_m + a_m \sum_{n=1}^N (\tilde{V}_{n,m}^P)^{-1} (\tilde{\theta}_{n,m}^F - \tilde{\theta}_{n-1,m}^F), \quad (2.12)$$

where  $\tilde{\theta}_{n,m}^F = \tilde{\theta}_n^F(\hat{\theta}_m, \sigma_m, J_m)$  and  $\tilde{V}_{n,m}^P = \tilde{V}_{n,m}^P(\hat{\theta}_m, \sigma_m, J_m)$  are defined in (2.9) via an application of Algorithm 1. Assuming conditions (A1–A4) and (B1–B2),  $\lim_{m \rightarrow \infty} \hat{\theta}_m = \hat{\theta}$  with probability one.

The proof of Theorem 2.2.3, given in Section 2.3.4, is based on applying Theo-

rem 2.2.2 in the context of existing results on stochastic approximation.

## 2.3 Proofs of the main results

We employ Landau notation for the limit  $\tau \rightarrow 0$ , namely, we write  $\alpha = O(\beta)$  to mean that  $\alpha(\tau)/\beta(\tau)$  is bounded, and  $\alpha = o(\beta)$  to mean that  $\lim_{\tau \rightarrow 0} \alpha(\tau)/\beta(\tau) = 0$ . In Sections 2.3.1 and 2.3.2, we write  $\nabla_\theta$  and  $\nabla_{\theta_n}$  for vectors of partial derivatives with respect to the components of  $\theta$  and  $\theta_n$  respectively.  $\nabla_{\theta_n} g_{Y_n|Y_{1:n-1}, \Theta_n}(y_{1:n} | \theta_n = \phi; \theta, \sigma, \tau)$  to denote partial derivatives evaluated at  $\theta_n = \phi$ .

### 2.3.1 A proof of Theorem 2.2.1

Suppose inductively that  $|\theta_{n-1}^F - \theta| = O(\tau^2)$ , which holds for  $n = 1$  by construction. We make a Taylor series expansion of  $g_{Y_n|Y_{1:n-1}, \Theta_n}(y_n | y_{1:n-1}, \theta_n)$  about  $\theta_n = \theta_{n-1}^F$ , suppressing the dependence of  $g$  on  $\theta, \sigma$  and  $\tau$ , to give

$$\begin{aligned} g_{Y_n|Y_{1:n-1}, \Theta_n}(y_n | y_{1:n-1}, \theta_n) &= g_{Y_n|Y_{1:n-1}, \Theta_n}(y_n | y_{1:n-1}, \theta_{n-1}^F) \\ &\quad + (\theta_n - \theta_{n-1}^F)^\top \nabla_{\theta_n} g_{Y_n|Y_{1:n-1}, \Theta_n}(y_n | y_{1:n-1}, \theta_{n-1}^F) + R_1(\theta_n). \end{aligned} \tag{2.13}$$

Integrating (2.13), we calculate

$$\begin{aligned} g_{Y_n|Y_{1:n-1}}(y_n | y_{1:n-1}) &= \int g_{Y_n|Y_{1:n-1}, \Theta_n}(y_n | y_{1:n-1}, \theta_n) g_{\Theta_n|Y_{1:n-1}}(\theta_n | y_{1:n-1}) d\theta_n \\ &= g_{Y_n|Y_{1:n-1}, \Theta_n}(y_n | y_{1:n-1}, \theta_{n-1}^F) + R_2 \end{aligned} \tag{2.14}$$

where

$$R_2 = \int R_1(\theta_n) g_{\Theta_n|Y_{1:n-1}}(\theta_n | y_{1:n-1}) d\theta_n. \tag{2.15}$$

From Lemma 2.3.3 in Section 2.3.2, there is a constant  $C_3$  such that  $|R_1(\theta_n)|$  is bounded by  $C_3|\theta_n - \theta_{n-1}^F|^2/2$  on the set  $B(\tau) = \{\theta_n : |\theta_n - \theta| \leq C_2(\tau - n\sigma)\}$ , where

$C_2$  does not depend on  $\tau$ . From (A1),  $R_1(\theta_n)$  is also bounded by some constant  $C_4$ , uniformly in  $\tau$ , on the set  $\overline{B(\tau)} = \{\theta_n : |\theta_n - \theta| \leq C_2(\tau + n\sigma)\}$ . We therefore conclude that  $R_2 = O(\tau^2 + \sigma/\tau)$ , and so  $R_2 = O(\tau^2)$  by assumption. Dividing (2.13) by (2.14), and applying Bayes' formula, we obtain

$$\frac{g_{\Theta_n|Y_{1:N}}(\theta_n | y_{1:n})}{g_{\Theta_n|Y_{1:N-1}}(\theta_n | y_{1:n-1})} = 1 + (\theta_n - \theta_{n-1}^F)^\top \nabla_{\theta_n} \log g_{Y_n|Y_{1:n-1}, \Theta_n}(y_n | y_{1:n-1}, \theta_{n-1}^F) + R_3(\theta_n). \quad (2.16)$$

The bounds on  $R_1$  and  $R_2$ , together with the observation that  $g_{Y_n|Y_{1:n-1}, \Theta_n}(y_n | y_{1:n-1}, \theta_n)$  is uniformly bounded away from zero on  $B(\tau)$ , imply that there are constants  $C_5$  and  $C_6$  such that  $|R_3(\theta_n)| < C_5|\theta_n - \theta_{n-1}^F|^2$  on  $B(\tau)$  and  $|R_3(\theta_n)| < C_6$  on  $\overline{B(\tau)}$ . We now calculate

$$\begin{aligned} \theta_n^F - \theta_{n-1}^F &= E_g[\Theta_n - \theta_{n-1}^F | Y_{1:n} = y_{1:n}] \\ &= \int (\theta_n - \theta_{n-1}^F) g_{\Theta_n|Y_{1:n}}(\theta_n | y_{1:n}) d\theta_n \end{aligned} \quad (2.17)$$

$$= V_n^P \nabla_{\theta_n} \log g_{Y_n|Y_{1:n-1}, \Theta_n}(y_n | y_{1:n-1}, \theta_{n-1}^F) + R_4 \quad (2.18)$$

where

$$R_4 = \int (\theta_n - \theta_{n-1}^F) R_3(\theta_n) g_{\Theta_n|Y_{1:n-1}}(\theta_n | y_{1:n-1}) d\theta_n. \quad (2.19)$$

Equation (2.18) follows from (2.17) using (2.16), and we see that  $R_4 = O(\tau(\tau^2 + \sigma/\tau))$  so  $R_4 = o(\tau^2)$  by assumption. Applying Lemma 2.3.3 from Section 2.3.2 to  $\nabla_{\theta_n} \log g_{Y_n|Y_{1:n-1}, \Theta_n}(y_n | y_{1:n-1}, \theta_{n-1}^F)$ , we deduce from (2.18) that

$$\theta_n^F - \theta_{n-1}^F = V_n^P \nabla_{\theta} \log f_{Y_n|Y_{1:n-1}}(y_n | y_{1:n-1}; \theta) + o(\tau^2). \quad (2.20)$$

The inductive assumption that  $|\theta_n^F - \theta| = O(\tau^2)$  is justified by (2.20), since  $V_n^P = O(\tau^2)$  by construction. A similar argument for the prediction variance gives

$$\begin{aligned}
V_{n+1}^P &= \text{Var}_g(\Theta_{n+1} | Y_{1:n} = y_{1:n}) = \text{Var}_g(\Theta_n | Y_{1:n} = y_{1:n}) + \sigma^2 \Sigma \\
&= E_g[(\Theta_n - \theta_n^F)(\Theta_n - \theta_n^F)^\top | Y_{1:n} = y_{1:n}] + \sigma^2 \Sigma \\
&= E_g[(\Theta_n - \theta_{n-1}^F)(\Theta_n - \theta_{n-1}^F)^\top | Y_{1:n} = y_{1:n}] - (\theta_n^F - \theta_{n-1}^F)(\theta_n^F - \theta_{n-1}^F)^\top + \sigma^2 \Sigma
\end{aligned} \tag{2.21}$$

$$= V_n^P + \sigma^2 \Sigma + o(\tau^2), \tag{2.22}$$

where (2.22) follows from (2.21) via (2.16) and (2.20) together with the observation that  $|\theta_n^F - \theta_{n-1}^F| = o(\tau^2)$ . It follows from (2.22) that  $|V_n^P - V_1^P| = o(\tau^2)$ . Noting that  $V_1^P = (\tau^2 + \sigma^2)\Sigma$ , one can multiply (2.20) through by  $(V_n^P)^{-1}$  and sum over  $n$  to give (2.8). This completes the proof of Theorem 2.2.1.

### 2.3.2 Lemmas required for the proof of Theorem 2.2.1

The passage from (2.18) to (2.20) 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  $\sigma$  and  $\tau$  with  $\sigma\tau^{-3} \rightarrow 0$ . If the variability of  $g_{\Theta_{1:n}|\Theta_0}(\theta_{1:n} | \theta_0; \sigma)$  is small compared to the variability of  $g_{\Theta_0}(\theta_0; \theta, \sigma, \tau)$  then, heuristically, one expects  $g_{\Theta_{0:n-1}|Y_{1:n}, \Theta_n}(\theta_{0:n-1} | y_{1:n}, \theta_n; \theta, \sigma, \tau)$  to be concentrated around  $\theta_n$  in the limit as  $\tau \rightarrow 0$ . Lemma 2.3.3 takes advantage of a formalization of this limit. However, the issue may be of minor relevance in practice because one expects that  $g_{\Theta_{n-k:n-1}|Y_{1:n}, \Theta_n}(\theta_{n-k:n-1} | y_{1:n}, \theta_n)$  will indeed be concentrated around  $\theta_n$  when  $k \ll n$  even if  $\sigma$  is not small compared to  $\tau$ . 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 prop-

erties of filtering procedures (e.g., Crisan and Doucet, 2002). We conjecture that one could achieve a result similar to Lemma 2.3.3 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 chapter.

**Lemma 2.3.1.** *(A1–A4) implies*

$$\nabla_{\theta_n} g_{Y_{1:n}|\Theta_n}(y_{1:n} | \theta_n; \theta, \sigma, \tau) = \sum_{i=0}^n U_i - V \quad (2.23)$$

$$\nabla_{\theta_n}^2 g_{Y_{1:n}|\Theta_n}(y_{1:n} | \theta_n; \theta, \sigma, \tau) = \sum_{i=0}^n \sum_{j=0}^n W_{i,j} + X - \sum_{i=0}^n (Y_i + Y_i^T) \quad (2.24)$$

where

$$\begin{aligned} U_i &= \int [\nabla_{\theta_i} g_{Y_{1:n}|\Theta_{0:n}}(y_{1:n} | \theta_{0:n})] g_{\Theta_{0:n-1}|\Theta_n}(\theta_{0:n-1} | \theta_n; \theta, \sigma, \tau) d\theta_{0:n-1} \\ V &= \nabla_{\theta} g_{Y_{1:n}|\Theta_{0:n}}(y_{1:n} | \theta_n; \theta, \sigma, \tau) \\ W_{i,j} &= \int [\nabla_{\theta_i} \nabla_{\theta_j}^T g_{Y_{1:n}|\Theta_{0:n}}(y_{1:n} | \theta_{0:n})] g_{\Theta_{0:n-1}|\Theta_n}(\theta_{0:n-1} | \theta_n; \theta, \sigma, \tau) d\theta_{0:n-1} \\ X &= \nabla_{\theta}^2 g_{Y_{1:n}|\Theta_{0:n}}(y_{1:n} | \theta_n; \theta, \sigma, \tau) \\ Y_i &= \nabla_{\theta} \int [\nabla_{\theta_i}^T g_{Y_{1:n}|\Theta_{0:n}}(y_{1:n} | \theta_{0:n})] g_{\Theta_{0:n-1}|\Theta_n}(\theta_{0:n-1} | \theta_n; \theta, \sigma, \tau) d\theta_{0:n-1} \quad \left( = \nabla_{\theta} U_i^T \right) \end{aligned}$$

*Proof.* We start the derivation of (2.23) by integrating  $g_{Y_{1:n}, \Theta_{0:n-1}|\Theta_n}(y_{1:n}, \theta_{0:n-1} | \theta_n; \theta, \sigma, \tau)$  over  $\theta_{0:n-1}$ . We then employ (A2) to justify passing  $\nabla_{\theta_n}$  through the resulting integral. Noting that  $g_{Y_{1:n}|\Theta_{0:n}}(y_{1:n} | \theta_{0:n}; \theta, \sigma, \tau)$  does not depend on  $\theta$ ,  $\sigma$  or  $\tau$ , we calculate

$$\nabla_{\theta_n} g_{Y_{1:n}|\Theta_n}(y_{1:n} | \theta_n; \theta, \sigma, \tau) = U_n + \mathbb{A} + \mathbb{B}_n \quad (2.25)$$

for

$$\begin{aligned} \mathbb{A} &= g_{Y_{1:n}, \Theta_n}(y_{1:n}, \theta_n; \theta, \sigma, \tau) \nabla_{\theta_n} \left[ \frac{1}{g_{\Theta_n}(\theta_n; \theta, \sigma, \tau)} \right] \\ \mathbb{B}_i &= \int \frac{g_{Y_{1:n} | \Theta_{0:n}}(y_{1:n} | \theta_{0:n})}{g_{\Theta_n}(\theta_n; \theta, \sigma, \tau) \sigma^{np\tau p}} \left[ \nabla_{\theta_i} \kappa \left( \frac{\theta_i - \theta_{i-1}}{\sigma} \right) \right] \kappa \left( \frac{\theta_0 - \theta}{\tau} \right) \prod_{j \neq i} \kappa \left( \frac{\theta_j - \theta_{j-1}}{\sigma} \right) d\theta_{0:n-1} \quad \text{for } i \geq 1 \end{aligned} \quad (2.26)$$

$$\mathbb{B}_0 = \int \frac{g_{Y_{1:n} | \Theta_{0:n}}(y_{1:n} | \theta_{0:n})}{g_{\Theta_n}(\theta_n; \theta, \sigma, \tau) \sigma^{np\tau p}} \left[ \nabla_{\theta_0} \kappa \left( \frac{\theta_0 - \theta}{\tau} \right) \right] \prod_{j=1}^n \kappa \left( \frac{\theta_j - \theta_{j-1}}{\sigma} \right) d\theta_{0:n-1} \quad (2.27)$$

Noticing that  $\nabla_{\theta_i} \kappa([\theta_i - \theta_{i-1}]/\sigma) = -\nabla_{\theta_{i-1}} \kappa([\theta_i - \theta_{i-1}]/\sigma)$  and applying integration by parts to (2.26) one finds that

$$\mathbb{B}_i = U_{i-1} + \mathbb{B}_{i-1} \quad \text{for } 1 \leq i \leq n. \quad (2.28)$$

A further calculation gives

$$\begin{aligned} \mathbb{B}_0 &= - \int g_{Y_{1:n} | \Theta_{0:n}}(y_{1:n} | \theta_{0:n}) \frac{\nabla_{\theta} g_{\Theta_{0:n}}(\theta_{0:n}; \theta, \sigma, \tau)}{g_{\Theta_n}(\theta_n; \theta, \sigma, \tau)} d\theta_{0:n-1} \\ &= - \left( V - \nabla_{\theta} \left[ \frac{1}{g_{\Theta_n}(\theta_n; \theta, \sigma, \tau)} \right] g_{Y_{1:n}, \Theta_n}(y_{1:n}, \theta_n; \theta, \sigma, \tau) \right) \end{aligned} \quad (2.29)$$

$$= -V - \mathbb{A} \quad (2.30)$$

(2.30) follows from (2.29) because  $g_{\Theta_n}(\theta_n; \theta, \sigma, \tau)$  is a function of  $\theta_n - \theta$ . Combining (2.25), (2.28) and (2.30) gives (2.23). To show (2.24), we write

$$\nabla_{\theta_n}^2 g_{Y_{1:n}, \Theta_n}(y_{1:n} | \theta_n; \theta, \sigma, \tau) = \mathbb{C} + \mathbb{D} + \mathbb{D}^T + W_{n,n} + \mathbb{E}_{n,n} + \mathbb{E}_{n,n}^T + \mathbb{F}_n \quad (2.31)$$

where

$$\mathbb{C} = \nabla_{\theta_n}^2 \left[ \frac{1}{g_{\Theta_n}(\theta_n; \theta, \sigma, \tau)} \right] g_{Y_{1:n}, \Theta_n}(y_{1:n}, \theta_n; \theta, \sigma, \tau) \quad (2.32)$$

$$\mathbb{D} = \nabla_{\theta_n} \left[ \frac{1}{g_{\Theta_n}(\theta_n; \theta, \sigma, \tau)} \right] \nabla_{\theta_n}^T g_{Y_{1:n}, \Theta_n}(y_{1:n}, \theta_n; \theta, \sigma, \tau) \quad (2.33)$$

$$\mathbb{E}_{i,j} = \int \frac{\nabla_{\theta_i} g_{Y_{1:n} | \Theta_{0:n}}(y_{1:n} | \theta_{0:n})}{g_{\Theta_n}(\theta_n; \theta, \sigma, \tau) \sigma^{np} \tau^p} \left[ \nabla_{\theta_j}^T \kappa \left( \frac{\theta_j - \theta_{j-1}}{\sigma} \right) \right] \kappa \left( \frac{\theta_0 - \theta}{\tau} \right) \prod_{k \neq j} \kappa \left( \frac{\theta_k - \theta_{k-1}}{\sigma} \right) d\theta_{0:n-1} \quad \text{for } j \geq 1 \quad (2.34)$$

$$\mathbb{E}_{i,0} = \int \frac{\nabla_{\theta_i} g_{Y_{1:n} | \Theta_{0:n}}(y_{1:n} | \theta_{0:n})}{g_{\Theta_n}(\theta_n; \theta, \sigma, \tau) \sigma^{np} \tau^p} \left[ \nabla_{\theta_0}^T \kappa \left( \frac{\theta_0 - \theta}{\tau} \right) \right] \prod_{k=1}^n \kappa \left( \frac{\theta_k - \theta_{k-1}}{\sigma} \right) d\theta_{0:n-1} \quad (2.35)$$

$$\mathbb{F}_i = \int \frac{g_{Y_{1:n} | \Theta_{0:n}}(y_{1:n} | \theta_{0:n})}{g_{\Theta_n}(\theta_n; \theta, \sigma, \tau) \sigma^{np} \tau^p} \left[ \nabla_{\theta_i}^2 \kappa \left( \frac{\theta_i - \theta_{i-1}}{\sigma} \right) \right] \kappa \left( \frac{\theta_0 - \theta}{\tau} \right) \prod_{j \neq i} \kappa \left( \frac{\theta_j - \theta_{j-1}}{\sigma} \right) d\theta_{0:n-1} \quad \text{for } i \geq 1 \quad (2.36)$$

$$\mathbb{F}_0 = \int \frac{g_{Y_{1:n} | \Theta_{0:n}}(y_{1:n} | \theta_{0:n})}{g_{\Theta_n}(\theta_n; \theta, \sigma, \tau) \sigma^{np} \tau^p} \left[ \nabla_{\theta_0}^2 \kappa \left( \frac{\theta_0 - \theta}{\tau} \right) \right] \prod_{j=1}^n \kappa \left( \frac{\theta_j - \theta_{j-1}}{\sigma} \right) d\theta_{0:n-1} \quad (2.37)$$

Applying the identity  $\nabla_{\theta_j} \kappa([\theta_j - \theta_{j-1}]/\sigma) = -\nabla_{\theta_{j-1}} \kappa([\theta_j - \theta_{j-1}]/\sigma)$  to (2.34) and then integrating by parts gives

$$\mathbb{E}_{i,j} = W_{i,j-1} + \mathbb{E}_{i,j-1} \quad \text{for } 1 \leq j \leq n. \quad (2.38)$$

A similar calculation for  $j = 0$  gives

$$\begin{aligned} \mathbb{E}_{i,0}^T &= - \left( Y_i - \nabla_{\theta} \left[ \frac{1}{g_{\Theta_n}(\theta_n; \theta, \sigma, \tau)} \right] \int \left[ \nabla_{\theta_i}^T g_{Y_{1:n} | \Theta_{0:n}}(y_{1:n} | \theta_{0:n}) \right] g_{\Theta_{0:n}}(\theta_{0:n}; \theta, \sigma, \tau) d\theta_{0:n-1} \right) \\ &= -Y_i + \mathbb{G}_i \end{aligned} \quad (2.39)$$

for

$$\mathbb{G}_i = \nabla_{\theta} \left[ \frac{1}{g_{\Theta_n}(\theta_n; \theta, \sigma, \tau)} \right] g_{\Theta_n}(\theta_n; \theta, \sigma, \tau) U_i^T. \quad (2.40)$$

The same procedure applied to (2.36) gives

$$\mathbb{F}_i = W_{i-1,i-1} + \mathbb{E}_{i-1,i-1} + \mathbb{E}_{i-1,i-1}^T + \mathbb{F}_{i-1} \quad \text{for } 1 \leq i \leq n. \quad (2.41)$$

For  $i = 0$ , we calculate

$$\mathbb{F}_0 = X - \mathbb{H} - \mathbb{H}^T - \mathbb{C} \quad (2.42)$$

where

$$\mathbb{H} = \nabla_{\theta} \left[ \frac{1}{g_{\Theta_n}(\theta_n; \theta, \sigma, \tau)} \right] \int g_{Y_{1:n}|\Theta_{0:n}}(y_{1:n} | \theta_{0:n}) \nabla_{\theta}^T g_{\Theta_{0:n}}(\theta_{0:n}; \theta, \sigma, \tau) d\theta_{0:n-1} \quad (2.43)$$

$$= \nabla_{\theta} \left[ \frac{1}{g_{\Theta_n}(\theta_n; \theta, \sigma, \tau)} \right] \nabla_{\theta}^T g_{Y_{1:n}, \Theta_n}(y_{1:n}, \theta_n; \theta, \sigma, \tau) \quad (2.44)$$

A further calculation gives

$$\begin{aligned} \mathbb{D} &= -\nabla_{\theta} \left[ \frac{1}{g_{\Theta_n}(\theta_n; \theta, \sigma, \tau)} \right] \left\{ g_{\Theta_n}(\theta_n; \theta, \sigma, \tau) \nabla_{\theta_n}^T g_{Y_{1:n}|\Theta_n}(y_{1:n} | \theta_n; \theta, \sigma, \tau) \right. \\ &\quad \left. + g_{Y_{1:n}|\Theta_n}(y_{1:n} | \theta_n; \theta, \sigma, \tau) \nabla_{\theta_n}^T g_{\Theta_n}(\theta_n | \theta, \sigma, \tau) \right\} \\ &= \sum_{i=0}^n \mathbb{G}_i + \nabla_{\theta} \left[ \frac{1}{g_{\Theta_n}(\theta_n; \theta, \sigma, \tau)} \right] \left\{ g_{\Theta_n}(\theta_n; \theta, \sigma, \tau) \nabla_{\theta}^T g_{Y_{1:n}}(y_{1:n}; \theta, \sigma, \tau) \right. \\ &\quad \left. - g_{Y_{1:n}|\Theta_n}(y_{1:n} | \theta_n; \theta, \sigma, \tau) \nabla_{\theta_n}^T g_{\Theta_n}(\theta_n; \theta, \sigma, \tau) \right\} \\ &= \sum_{i=0}^n \mathbb{G}_i + \nabla_{\theta} \left[ \frac{1}{g_{\Theta_n}(\theta_n; \theta, \sigma, \tau)} \right] \left\{ g_{\Theta_n}(\theta_n; \theta, \sigma, \tau) \nabla_{\theta}^T g_{Y_{1:n}}(y_{1:n}; \theta, \sigma, \tau) \right. \\ &\quad \left. + g_{Y_{1:n}|\Theta_n}(y_{1:n} | \theta_n; \theta, \sigma, \tau) \nabla_{\theta}^T g_{\Theta_n}(\theta_n; \theta, \sigma, \tau) \right\} \\ &= \sum_{i=0}^n \mathbb{G}_i + \mathbb{H} \end{aligned} \quad (2.45)$$

Combining (2.31), (2.38), (2.39), (2.41), (2.42), (2.45) and the identity  $W_{i,j} = W_{j,i}^T$  gives (2.24).  $\square$

**Lemma 2.3.2.** *Let  $\Theta_{0:n}$  be the collection of random variables defined in Sec-*

tion 2.2, with joint density  $g_{\Theta_{0:n}}(\theta_{0:n} | \theta, \sigma, \tau) = \frac{1}{\sigma^{np} \tau^p} \kappa\left(\frac{\theta_0 - \theta}{\tau}\right) \prod_{i=1}^n \kappa\left(\frac{\theta_i - \theta_{i-1}}{\sigma}\right)$  with  $\kappa$  satisfying conditions (A3) and (A4). Let  $\psi(\theta_{0:n})$  be a continuous function of  $\theta_{0:n}$  taking values in  $\mathbb{R}^d$  for some  $d$ . Define  $\theta^{[k]} = (\theta, \dots, \theta)$  for  $k$  concatenated copies of  $\theta$ . Set  $B(\tau) = \{\theta_n : |\theta_n - \theta| < C_2(\tau - n\sigma)\}$  and  $\overline{B(\tau)} = \{\theta_n : |\theta_n - \theta| < C_2(\tau + n\sigma)\}$ . Then,

$$\lim_{\tau \rightarrow 0} \sup_{\theta_n \in \overline{B(\tau)}} \left| \psi(\theta^{[n+1]}) - \int \psi(\theta_{0:n}) g_{\Theta_{0:n-1} | \Theta_n}(\theta_{0:n-1} | \theta_n; \theta, \sigma, \tau) d\theta_{0:n-1} \right| = 0, \quad (2.46)$$

$$\lim_{\tau \rightarrow 0} \sup_{\theta_n \in B(\tau)} \left| \int \psi(\theta_{0:n}) \nabla_{\theta} g_{\Theta_{0:n-1} | \Theta_n}(\theta_{0:n-1} | \theta_n; \theta, \sigma, \tau) d\theta_{0:n-1} \right| = 0. \quad (2.47)$$

$$\lim_{\tau \rightarrow 0} \sup_{\theta_n \in B(\tau)} \left| \int \psi(\theta_{0:n}) \nabla_{\theta}^2 g_{\Theta_{0:n-1} | \Theta_n}(\theta_{0:n-1} | \theta_n; \theta, \sigma, \tau) d\theta_{0:n-1} \right| = 0. \quad (2.48)$$

*Proof.* To show (2.46) we note that, for any  $\theta_n \in \overline{B(\tau)}$ ,  $g_{\Theta_{0:n-1} | \Theta_n}(\theta_{0:n-1} | \theta_n; \theta, \sigma, \tau)$  defines a probability density for  $\Theta_{0:n-1}$  supported on

$$K(\tau) = \left\{ \theta_{0:n-1} : |\theta_j - \theta| \leq C_2(\tau + j\sigma) \forall j \in \{0, 1, \dots, n-1\} \right\}. \quad (2.49)$$

Since  $K(\tau)$  converges to the point  $\theta^{[n]}$  as  $\tau \rightarrow 0$ , (2.46) is guaranteed by the continuity of  $\psi(\theta_{0:n})$ . To show (2.47), we write the identity

$$\begin{aligned} \nabla_{\theta} \int \psi(\theta_{0:n}) g_{\Theta_{0:n-1} | \Theta_n}(\theta_{0:n-1} | \theta_n; \theta, \sigma, \tau) d\theta_{0:n-1} \\ = \int \psi(\theta_{0:n}) \mu(\theta_{0:n-1} | \theta_n; \sigma) \nabla_{\theta} \left[ \frac{g_{\Theta_0}(\theta_0; \theta, \tau)}{g_{\Theta_n}(\theta_n; \theta, \sigma, \tau)} \right] d\theta_{0:n-1} \end{aligned}$$

where  $\mu(\theta_{0:n-1} | \theta_n; \sigma) = \frac{1}{\sigma^{np}} \prod_{k=1}^n \kappa\left(\frac{\theta_k - \theta_{k-1}}{\sigma}\right)$ . Since it is clear that  $\mu(\theta_{0:n-1} | \theta_n; \sigma)$  converges to a point mass at  $\theta_n^{[n]}$  in the limit  $\sigma \rightarrow 0$ , it remains to show that

$$\lim_{\tau \rightarrow 0} \sup_{\theta_n \in B(\tau), |\theta_0 - \theta_n| \leq C_2 \sigma n} \left| \nabla_{\theta} \left[ \frac{g_{\Theta_0}(\theta_0; \theta, \tau)}{g_{\Theta_n}(\theta_n; \theta, \sigma, \tau)} \right] \right| = 0. \quad (2.50)$$

To show (2.50), we start by noting that

$$g_{\Theta_n}(\theta_n; \theta, \sigma, \tau) = \frac{1}{\tau}(\kappa * \nu)\left(\frac{\theta_n - \theta}{\tau}\right) \quad (2.51)$$

where  $*$  denotes the convolution  $(\kappa * \nu)(\psi) = \int \kappa(\phi)\nu(\psi - \phi) d\phi$ , and  $\nu(\psi) = \frac{1}{\sigma/\tau}\kappa^n\left(\frac{\psi}{\sigma/\tau}\right)$  with  $\kappa^n$  being the  $n$ -fold convolution of  $\kappa$  with itself. One way to derive (2.51) is to consider the rescaled random variable  $\tilde{\Theta}_n = (\Theta_n - \theta)/\tau$  and to observe that  $\tilde{\Theta}_n$  has density  $(\kappa * \nu)(\tilde{\theta}_n)$ . Using (2.51), we write

$$\begin{aligned} \nabla_{\theta} \left[ \frac{g_{\Theta_0}(\theta_0; \theta, \tau)}{g_{\Theta_n}(\theta_n; \theta, \sigma, \tau)} \right] &= \frac{\frac{1}{\tau^2}([\nabla\kappa] * \nu)\left(\frac{\theta_n - \theta}{\tau}\right) \times \frac{1}{\tau}\kappa\left(\frac{\theta_0 - \theta}{\tau}\right) - \frac{1}{\tau}(\kappa * \nu)\left(\frac{\theta_n - \theta}{\tau}\right) \times \frac{1}{\tau^2}[\nabla\kappa]\left(\frac{\theta_0 - \theta}{\tau}\right)}{\frac{1}{\tau^2}[(\kappa * \nu)\left(\frac{\theta_n - \theta}{\tau}\right)]^2} \\ &= \frac{\frac{1}{\tau} \left[ ([\nabla\kappa] * \nu)(\tilde{\theta}_n) \times \kappa(\tilde{\theta}_0) - (\kappa * \nu)(\tilde{\theta}_n) \times [\nabla\kappa](\tilde{\theta}_0) \right]}{[(\kappa * \nu)(\tilde{\theta}_n)]^2} \end{aligned} \quad (2.52)$$

where  $\tilde{\theta}_n = (\theta_n - \theta)/\tau$  and  $\tilde{\theta}_0 = (\theta_0 - \theta)/\tau$ . This change of variables maps  $B(\tau)$  to  $\tilde{B}(\tau) = \{\tilde{\theta}_n : |\tilde{\theta}_n| \leq C_2(1+n\sigma/\tau)\}$ , and  $\{|\theta_0 - \theta_n| \leq C_2\sigma n\}$  to  $\{|\tilde{\theta}_0 - \tilde{\theta}_n| \leq C_2n\sigma/\tau\}$ .

The denominator on the right hand side of (2.52) is uniformly bounded away from zero on  $\tilde{B}(\tau)$  since

$$\inf_{\tilde{\theta}_n \in \tilde{B}} (\kappa * \nu)(\tilde{\theta}_n) \geq \inf_{|\tilde{\theta}_n| \leq C_2} \kappa(\tilde{\theta}_n) > 0, \quad (2.53)$$

with the second inequality following from (A3) and being independent of  $\tau$ . Now we note that  $(\kappa * \nu)$  converges uniformly to  $\kappa$  at rate  $\sigma/\tau$ , in the sense that

$$\sup_{\tilde{\theta}_n \in \tilde{B}(\tau)} |(\kappa * \nu)(\tilde{\theta}_n) - \kappa(\tilde{\theta}_n)| \leq \sup_{\tilde{\theta}_n \in \tilde{B}(\tau), |\tilde{\phi} - \tilde{\theta}_n| \leq C_2n\sigma/\tau} \left| \kappa(\tilde{\phi}) - \kappa(\tilde{\theta}_n) \right| = O(\sigma/\tau). \quad (2.54)$$

Here, we use the Lipschitz continuity of  $\kappa$  guaranteed by (A4). The Lipschitz

continuity of  $\kappa$  also implies from (2.54) that

$$\sup_{\tilde{\theta}_n \in B(\tilde{\tau}), |\tilde{\theta}_0 - \tilde{\theta}_n| \leq C_2 n \sigma / \tau} \left| (\kappa * \nu)(\tilde{\theta}_n) - \kappa(\tilde{\theta}_0) \right| = O(\sigma / \tau). \quad (2.55)$$

Since  $\nabla \kappa * \nu$  converges uniformly to  $\nabla \kappa$  at rate  $\sigma$  on  $B(\tilde{\tau})$ , we obtain

$$\sup_{\tilde{\theta}_n \in B(\tilde{\tau}), |\tilde{\theta}_0 - \tilde{\theta}_n| \leq C_2 n \sigma} \left| ([\nabla \kappa] * \nu)(\tilde{\theta}_n) - \nabla \kappa(\tilde{\theta}_0) \right| = O(\sigma / \tau). \quad (2.56)$$

Combining (2.52), (2.53), (2.55) and (2.56) gives

$$\sup_{\theta_n \in B(\tau), |\theta_0 - \theta_n| \leq C_2 \sigma n} \left| \nabla_{\theta} \left[ \frac{g_{\Theta_0}(\theta_0; \theta, \tau)}{g_{\Theta_n}(\theta_n; \theta, \sigma, \tau)} \right] \right| = O(\sigma \tau^{-2}), \quad (2.57)$$

from which (2.50), and hence (2.46), follows by assumption. The demonstration of (2.47) follows along similar lines, by checking that

$$\lim_{\tau \rightarrow 0} \sup_{\theta_n \in B(\tau), |\theta_0 - \theta_n| \leq C_2 \sigma n} \left| \nabla_{\theta}^2 \frac{g_{\Theta_0}(\theta_0; \theta, \tau)}{g_{\Theta_n}(\theta_n; \theta, \sigma, \tau)} \right| = 0. \quad (2.58)$$

Following the reparameterization above, (2.58) is equivalent to

$$\begin{aligned} \lim_{\tau \rightarrow 0} \sup_{\substack{\tilde{\theta}_n \in B(\tilde{\tau}) \\ |\tilde{\theta}_0 - \tilde{\theta}_n| \leq C_2 n \sigma / \tau}} \frac{1}{\tau^2} & \left| \frac{[\kappa * \nu(\tilde{\theta}_n)] \nabla^2 \kappa(\tilde{\theta}_0) - \kappa(\tilde{\theta}_0) [\nabla^2 \kappa * \nu(\tilde{\theta}_n)]}{[\kappa * \nu(\tilde{\theta}_n)]^2} + \frac{2\kappa(\tilde{\theta}_0) [\nabla \kappa * \nu](\tilde{\theta}_n) [\nabla^T \kappa * \nu](\tilde{\theta}_n)}{[\kappa * \nu(\tilde{\theta}_n)]^3} \right. \\ & \left. - \frac{[\nabla \kappa * \nu](\tilde{\theta}_n) \nabla^T \kappa(\tilde{\theta}_0) + \nabla \kappa(\tilde{\theta}_0) [\nabla^T \kappa * \nu](\tilde{\theta}_n)}{[\kappa * \nu(\tilde{\theta}_n)]^2} \right| = 0. \end{aligned} \quad (2.59)$$

Combining (2.53), (2.55) and (2.56) with an analogous result for the second derivative,

$$\sup_{\tilde{\theta}_n \in B(\tilde{\tau}), |\tilde{\theta}_0 - \tilde{\theta}_n| \leq C_2 n \sigma / \tau} \left| \nabla^2 \kappa * \nu(\tilde{\theta}_n) - \nabla^2 \kappa(\tilde{\theta}_0) \right| = O(\sigma / \tau),$$

gives (2.59) via the assumption that  $\sigma = o(\tau^3)$ , completing the proof of Lemma 2.3.2.  $\square$

**Lemma 2.3.3.** *Suppose the conditions (A1–A4). For  $B(\tau) = \{\theta_n : |\theta_n - \theta| < C_2(\tau - n\sigma)\}$ ,*

$$\lim_{\tau \rightarrow 0} \sup_{\theta_n \in B(\tau)} \left| \nabla_{\theta_n} \log g_{Y_n | Y_{1:n-1}, \Theta_n}(y_n | y_{1:n-1}, \theta_n; \theta, \sigma, \tau) - \nabla_{\theta} \log f_{Y_n | Y_{1:n-1}}(y_n | y_{1:n-1}; \theta) \right| = 0 \quad (2.60)$$

$$\lim_{\tau \rightarrow 0} \sup_{\theta_n \in B(\tau)} \left| \nabla_{\theta_n}^2 \log g_{Y_n | Y_{1:n-1}, \Theta_n}(y_n | y_{1:n-1}, \theta_n; \theta, \sigma, \tau) - \nabla_{\theta}^2 \log f_{Y_n | Y_{1:n-1}}(y_n | y_{1:n-1}; \theta) \right| = 0 \quad (2.61)$$

*Proof.* Applying Lemma 2.3.2 with  $\psi(\theta_{0:n}) = \nabla_{\theta_i} g_{Y_n | Y_{1:n-1}, \Theta_{0:n}}(y_{1:n} | \theta_{0:n})$ , we find a limit for  $U_i$  from Lemma 2.3.1 of

$$\lim_{\tau \rightarrow 0} \sup_{\theta_n \in B(\tau)} \left| U_i - \nabla_{\theta_i} g_{Y_n | Y_{1:n-1}, \Theta_{0:n}}(y_{1:n} | \theta_{0:n} = \theta^{[n+1]}) \right| = 0. \quad (2.62)$$

Write  $V$  from Lemma 2.3.1 as

$$V = \int g_{Y_n | Y_{1:n-1}, \Theta_{0:n}}(y_{1:n} | \theta_{0:n}) \nabla_{\theta} g_{\Theta_{0:n-1} | \Theta_n}(\theta_{0:n-1} | \theta_n; \theta, \sigma, \tau) d\theta_{0:n-1}$$

Then Lemma 2.3.2 applied to  $g_{Y_{1:n} | \Theta_{0:n}}(y_{1:n} | \theta_{0:n})$  gives

$$\lim_{\tau \rightarrow 0} \sup_{\theta_n \in B(\tau)} |V| = 0. \quad (2.63)$$

Noticing that  $g_{Y_{1:n} | \Theta_{0:n}}(y_{1:n} | \theta_{0:n} = \theta^{[n+1]}) = f_{Y_{1:n}}(y_{1:n}; \theta)$ , it follows that

$$\nabla_{\theta} f_{Y_{1:n}}(y_{1:n}; \theta) = \sum_{i=0}^n \nabla_{\theta_i} g_{Y_{1:n} | \Theta_{0:n}}(y_{1:n} | \theta^{[n+1]}). \quad (2.64)$$

Combining (2.62), (2.63) and (2.64) with the identity in (2.23) gives

$$\lim_{\tau \rightarrow 0} \sup_{\theta_n \in B(\tau)} \left| \nabla_{\theta_n} g_{Y_n | Y_{1:n-1}, \Theta_n}(y_n | y_{1:n-1}, \theta_n; \theta, \sigma, \tau) - \nabla_{\theta} f_{Y_n | Y_{1:n-1}}(y_n | y_{1:n-1}; \theta) \right| = 0, \quad (2.65)$$

from which (2.60) follows by a few routine steps. The argument for (2.61) is similar, and we start by writing

$$\nabla_{\theta}^2 f_{Y_{1:n}}(y_{1:n}; \theta) = \sum_{i=0}^n \sum_{j=1}^n \nabla_{\theta_i} \nabla_{\theta_j}^T g_{Y_{1:n} | \Theta_{0:n}}(y_{1:n} | \theta_{0:n} = \theta^{[n+1]}). \quad (2.66)$$

Applying Lemma 2.3.2 to each term on the right hand side of (2.24) gives

$$\lim_{\tau \rightarrow 0} \sup_{\theta_n \in B(\tau)} \left| W_{ij} - \nabla_{\theta_i} \nabla_{\theta_j}^T g_{Y_{1:n} | \Theta_{0:n}}(y_{1:n} | \theta^{[n+1]}) \right| = 0, \quad (2.67)$$

$$\lim_{\tau \rightarrow 0} \sup_{\theta_n \in B(\tau)} |X| = 0, \quad \lim_{\tau \rightarrow 0} \sup_{\theta_n \in B(\tau)} |Y_i| = 0. \quad (2.68)$$

Combining (2.67), (2.68) and (2.66) with the identity in (2.24) gives

$$\lim_{\tau \rightarrow 0} \sup_{\theta_n \in B(\tau)} \left| \nabla_{\theta_n}^2 g_{Y_n | Y_{1:n-1}, \Theta_n}(y_n | y_{1:n-1}, \theta_n; \theta, \sigma, \tau) - \nabla_{\theta}^2 f_{Y_n | Y_{1:n-1}}(y_n | y_{1:n-1}; \theta) \right| = 0. \quad (2.69)$$

Convergence of the second derivative in (2.69) and the first derivative in (2.65) implies (2.61) via a few routine steps.  $\square$

### 2.3.3 A proof of Theorem 2.2.2

Our approach is based on two general theorems on sequential Monte Carlo by Crisan and Doucet (2002) and Del Moral and Jacod (2001), stated in our notation as Theorems 2.3.1 and 2.3.2 below. Both these theorems are stated for a POMP model with generic density  $h$ , parameter vector  $\psi$ , Markov process  $Z_{0:N}$  and observation process  $Y_{1:N}$  with observed sequence  $y_{1:N}$ . For application to the unperturbed model one sets  $h = f$ ,  $Z_n = X_n$  and  $\psi = \theta$ . For application to the

perturbed model one sets  $h = g$ ,  $Z_n = (X_n, \Theta_n)$  and  $\psi = (\theta, \sigma, \tau)$ .

**Theorem 2.3.1.** (CRISAN AND DOUCET, 2002) *Let  $h$  be a generic density for a POMP model having parameter vector  $\psi$ , unobserved Markov process  $Z_{0:N}$ , observation process  $Y_{1:N}$  and data  $y_{1:N}$ . Define  $Z_{n,j}^F$  via applying Algorithm 1 with  $J$  particles. Assume that  $h_{Y_n|Z_n}(y_n | z_n; \psi)$  is bounded as a function of  $z_n$ . For any  $\phi : \mathbb{R}^{d_z} \rightarrow \mathbb{R}$ , denote the filtered mean of  $\phi(z_n)$  and its Monte Carlo estimate by*

$$\phi_n^F = \int \phi(z_n) h_{Z_n|Y_{1:n}}(z_n | y_{1:n}; \psi) dz_n, \quad \tilde{\phi}_n^F = \frac{1}{J} \sum_{j=1}^J \phi(Z_{n,j}^F). \quad (2.70)$$

There is a  $C_7$  independent of  $J$  such that

$$\tilde{E} \left[ (\tilde{\phi}_n^F - \phi_n^F)^2 \right] \leq \frac{C_7 \sup_x |\phi(x)|^2}{J}. \quad (2.71)$$

Specifically,  $C_7$  can be written as a linear function of 1 and  $\eta_{n,1}, \dots, \eta_{n,n}$  defined as

$$\eta_{n,i} = \prod_{k=n-i+1}^n \left( \frac{\sup_{z_k} h_{Y_k|Z_k}(y_k | z_k; \psi)}{h_{Y_k|Y_{1:k-1}}(y_k | y_{1:k-1}; \psi)} \right)^2. \quad (2.72)$$

*Proof.* We focus on the assertion that the constant  $C_7$  in equation (2.71) can be written as a linear function of 1 and the quantities  $\eta_{n,1}, \dots, \eta_{n,n}$  in (2.72). This was not explicitly mentioned by Crisan and Doucet (2002) but is a direct consequence of their argument. Crisan and Doucet (2002, Section V) constructed the following recursion, for which  $c_{n|n}$  is the constant  $C_7$  in equation (2.71). For  $n = 1, \dots, N$  and  $c_{0|0} = 0$ , define

$$c_{n|n} = \left( \sqrt{C} + \sqrt{\tilde{c}_{n|n}} \right)^2 \quad (2.73)$$

$$\tilde{c}_{n|n} = 4c_{n|n-1} \left( \frac{\|h\|_n}{h_{Y_n|Y_{1:n-1}}(y_n | y_{1:n-1}; \psi)} \right)^2 \quad (2.74)$$

$$c_{n|n-1} = \left( 1 + \sqrt{c_{n-1|n-1}} \right)^2 \quad (2.75)$$

where  $\|h\|_n = \sup_{z_n} h_{Y_n|Z_n}(y_n | z_n; \psi)$ . Here,  $C$  is a constant that depends on the

resampling procedure but not on the number of particles  $J$ . Now, (2.73–2.75) can be reformulated by routine algebra as

$$c_{n|n} \leq K_1 + K_2 \tilde{c}_{n|n} \quad (2.76)$$

$$\tilde{c}_{n|n} \leq K_3 q_n c_{n|n-1} \quad (2.77)$$

$$c_{n|n-1} \leq K_4 + K_5 c_{n-1|n-1} \quad (2.78)$$

where  $q_n = \|h\|_n^2 [h_{Y_n|Y_{1:n-1}}(y_n|y_{1:n-1}; \psi)]^{-2}$  and  $K_1, \dots, K_5$  are constants which do not depend on  $h, \psi, y_{1:N}$  or  $J$ . Putting (2.77) and (2.78) into (2.76),

$$\begin{aligned} c_{n|n} &\leq K_1 + K_2 K_3 q_n c_{n|n-1} \\ &\leq K_1 + K_2 K_3 K_4 q_n + K_2 K_3 K_5 q_n c_{n-1|n-1}. \end{aligned} \quad (2.79)$$

Since  $\eta_{n,i} = q_n \eta_{n-1,i}$  for  $i < n$ , and  $\eta_{n,n} = q_n$ , the required assertion follows from (2.79).  $\square$

**Theorem 2.3.2.** (DEL MORAL AND JACOD, 2001) *Let  $h, \psi, Z_{0:N}, Y_{1:N}$  and  $y_{1:N}$  describe a POMP model, as in Theorem 2.3.1. Let  $\phi : \mathbb{R}^{d_z} \rightarrow \mathbb{R}$  be a bounded function, with  $\phi_n^F$  and  $\tilde{\phi}_n^F$  specified in (2.70). Define the un-normalized filtered mean  $\phi_n^U$  and its Monte Carlo estimate  $\tilde{\phi}_n^U$  by*

$$\phi_n^U = \phi_n^F \prod_{k=1}^n h_{Y_k|Y_{1:k-1}}(y_k | y_{1:k-1}; \psi), \quad \tilde{\phi}_n^U = \tilde{\phi}_n^F \prod_{k=1}^n \frac{1}{J} \sum_{j=1}^J w(k, j). \quad (2.80)$$

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

$$\tilde{E}[\tilde{\phi}_n^U] = \phi_n^U, \quad (2.81)$$

$$\tilde{E}\left[(\tilde{\phi}_n^U - \phi_n^U)^2\right] \leq \frac{(n+1) \sup_x |\phi(x)|^2}{J} \prod_{k=1}^n \sup_{z_k} h_{Y_k|Z_k}(y_k | z_k; \psi)^2. \quad (2.82)$$

To apply Theorems 2.3.1 and 2.3.2 in the context of Theorem 2.2.2, we define

$u_{n,m} = (\theta_{n,m}^F - \theta_{n-1,m}^F)/\tau_m$  and  $v_{n,m} = V_{n,m}^P/\tau_m^2$ . The corresponding Monte Carlo estimates of these quantities are  $\tilde{u}_{n,m} = (\tilde{\theta}_{n,m}^F - \tilde{\theta}_{n-1,m}^F)/\tau_m$  and  $\tilde{v}_{n,m} = \tilde{V}_{n,m}^P/\tau_m^2$ . We argue that there are constants  $C_8, \dots, C_{11}$  with

$$|\tilde{E}[\tilde{u}_{n,m} - u_{n,m}]| \leq C_8/J_m \quad |\tilde{E}[\tilde{v}_{n,m} - v_{n,m}]| \leq C_9/J_m \quad (2.83)$$

$$\tilde{E}[|\tilde{u}_{n,m} - u_{n,m}|^2] \leq C_{10}/J_m \quad \tilde{E}[|\tilde{v}_{n,m} - v_{n,m}|^2] \leq C_{11}/J_m \quad (2.84)$$

uniformly for  $\theta$  in any compact set. Previous bounds similar to (2.83,2.84) 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 (2.83,2.84) is that the model is varying with  $\sigma_m$  and  $\tau_m$ . However, the bounds  $|u_{n,m}| \leq 2C_2(1 + n\sigma/\tau)$  and  $|v_{n,m}| \leq 4pC_2^2(1 + \sigma/\tau)^2$ , together with the continuity of  $g(y_n | z_n; \theta, \sigma, \tau)$  as a function of  $\sigma$  and  $\tau$ , is enough to show via Theorem 2.3.1 that the uniform bound in (2.84) holds. To show that (2.83) follows from (2.84) we follow the approach of Del Moral and Jacod (2001, Equation 3.3.14). Noting that  $\phi_n^F = \phi_n^U/1_n^U$  and  $\tilde{\phi}_n^F = \tilde{\phi}_n^U/\tilde{1}_n^U$ , Theorem 2.3.2 implies the identity

$$\tilde{E}[\tilde{\phi}_n^F - \phi_n^F] = \tilde{E}\left[(\tilde{\phi}_n^F - \phi_n^F)\left(1 - \frac{\tilde{1}_n^U(1)}{1_n^U(1)}\right)\right]. \quad (2.85)$$

Applying the Cauchy-Schwarz inequality, together with (2.71) and (2.82), gives

$$|\tilde{E}[\tilde{\phi}_n^F - \phi_n^F]| \leq C_{12} \frac{\sup_x |\phi(x)|}{J}. \quad (2.86)$$

We now proceed to carry out a Taylor series expansion:

$$\begin{aligned} \tilde{v}_{n,m}^{-1} \tilde{u}_{n,m} &= v_{n,m}^{-1} u_{n,m} + v_{n,m}^{-1} (\tilde{u}_{n,m} - u_{n,m}) \\ &\quad - v_{n,m}^{-1} (\tilde{v}_{n,m} - v_{n,m}) v_{n,m}^{-1} \tilde{u}_{n,m} + R_5 \end{aligned} \quad (2.87)$$

where  $|R_5| < C_{13}(|\tilde{u}_{n,m} - u_{n,m}|^2 + |\tilde{v}_{n,m} - v_{n,m}|^2)$  for some constant  $C_{13}$ . The existence of such a  $C_{13}$  is guaranteed since the determinant of  $v_{n,m}$  is bounded away from zero. Taking expectations of both sides of (2.87) and applying (2.83,2.84) gives

$$|\tilde{E}[\tilde{v}_{n,m}^{-1}\tilde{u}_{n,m}] - v_{n,m}^{-1}u_{n,m}| \leq C_{14}/J_m. \quad (2.88)$$

Another Taylor series expansion,

$$\tilde{v}_{n,m}^{-1}\tilde{u}_{n,m} = v_{n,m}^{-1}u_{n,m} + R_6 \quad (2.89)$$

with  $|R_6| < C_{15}(|\tilde{u}_{n,m} - u_{n,m}| + |\tilde{v}_{n,m} - v_{n,m}|)$  implies

$$\widetilde{\text{Var}}(\tilde{v}_{n,m}^{-1}\tilde{u}_{n,m}) \leq C_{16}/J_m. \quad (2.90)$$

Putting together (2.88) and (2.90), we deduce that

$$\tau_m J_m \left\{ \tilde{E}[(\tilde{V}_{n,m}^P)^{-1}(\tilde{\theta}_{n,m}^F - \tilde{\theta}_{n-1,m}^F)] - (V_{n,m}^P)^{-1}(\theta_{n,m}^F - \theta_{n-1,m}^F) \right\}$$

and

$$\tau_m^2 J_m \left\{ \widetilde{\text{Var}}[(\tilde{V}_{n,m}^P)^{-1}(\tilde{\theta}_{n,m}^F - \tilde{\theta}_{n-1,m}^F)] \right\}$$

are bounded as  $m \rightarrow \infty$ . Theorem 2.2.2 then follows by applying Theorem 2.2.1, making use of the assumed continuity with respect to  $\theta$ .

#### 2.3.4 A proof of Theorem 2.2.3

Theorem 2.2.3 follows directly from a general stochastic approximation result, presented as Theorem 2.3.3 below. In the context of Theorem 2.2.3, conditions (B4) and (B5) of Theorem 2.3.3 hold from Theorem 2.2.2 and the remaining assumptions of Theorem 2.3.3 hold by hypothesis.

**Theorem 2.3.3.** Let  $\ell(\theta)$  be a continuously differentiable function  $\mathbb{R}^p \rightarrow \mathbb{R}$  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 2.2 together with the following conditions:

$$(B3) \quad a_m > 0, \quad a_m \rightarrow 0, \quad \sum_m a_m = \infty.$$

$$(B4) \quad \sum_m a_m^2 \sup_{|\theta| < r} \widetilde{\text{Var}}(D_m(\theta)) < \infty \text{ for every } r > 0.$$

$$(B5) \quad \lim_{m \rightarrow \infty} \sup_{|\theta| < r} |\tilde{E}[D_m(\theta)] - \nabla\ell(\theta)| = 0 \text{ for every } r > 0.$$

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

Theorem 2.3.3 is a special case of Theorem 2.3.1 of Kushner and Clark (1978). The most laborious step in deducing Theorem 2.3.3 from Kushner and Clark (1978) is to check that (B1–B5) imply that, for all  $\epsilon > 0$ ,

$$\lim_{n \rightarrow \infty} \mathbb{P} \left[ \sup_{j \geq 1} \left| \sum_{m=n}^{n+j} a_m \{D_m(\hat{\theta}_m) - \tilde{E}[D_m(\hat{\theta}_m) | \hat{\theta}_m]\} \right| \geq \epsilon \right] = 0, \quad (2.91)$$

which in turn implies condition A2.2.4 of Kushner and Clark (1978). To show (2.91), we define  $\xi_m = D_m(\hat{\theta}_m) - \tilde{E}[D_m(\hat{\theta}_m) | \hat{\theta}_m]$  and

$$\xi_m^k = \begin{cases} \xi_m & \text{if } |\hat{\theta}_m| \leq k \\ 0 & \text{if } |\hat{\theta}_m| > k \end{cases}. \quad (2.92)$$

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),

$$\mathbb{P} \left[ \sup_j |M_j^{n,k}| \geq \epsilon \right] \leq \frac{1}{\epsilon^2} \sum_{m=n}^{\infty} a_m^2 \sup_{|\theta| < k} \widetilde{\text{Var}}(D_m(\theta)). \quad (2.93)$$

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 (B4) and (2.93) that  $\lim_{n \rightarrow \infty} \mathbb{P}\{F_{n,k}\} = 0$  for each  $k$ . In light of the non-divergence assumed in (B2), this implies  $\lim_{n \rightarrow \infty} \mathbb{P}\{F_n\} = 0$  which is exactly (2.91).

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  $\mathbb{P}(\Omega) = 1$ . Since the sequence of events  $\{\Omega_k\}$  is increasing up to  $\Omega$ , we have  $\lim_{k \rightarrow \infty} \mathbb{P}(\Omega_k) = \mathbb{P}(\Omega) = 1$ . 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  $\Omega_k$  when  $j \geq k$ . Then,

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

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

## 2.4 Discussion of the theory and practice of iterated filtering

The value of all asymptotic theory, such as presented in Section 2.2, 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 anneal-

ing are often too slow for practical implementation, yet variations on simulated annealing which endure less theoretical support have been found to be widely applicable (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. Indeed, simulated annealing can be studied within the framework of stochastic approximation theory (Spall, 2003, Chapter 8). To relate simulated annealing and iterated filtering, it is helpful to adopt from simulated annealing an analogy whereby  $\sigma_m$  and  $\tau_m$  are thought of as temperatures which approaching freezing as  $\sigma_m \rightarrow 0$  and  $\tau_m \rightarrow 0$ . 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 a 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.

The incorporation of iterated filtering into the framework of stochastic approximation, which underlies the proof of Theorem 2.2.3, 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.

An alternative heuristic approach to understanding iterated filtering is based on thinking of the quantity  $\tilde{\theta}_{n,m}^F$  in (2.12) as a time-localized estimate of  $\theta$ , in the sense that it depends most heavily on observations directly preceding and including  $y_n$ . Perturbing a state space model by applying a random walk in parameter space to reduce numerical instabilities arising in particle filtering was popularized by the influential work of Kitagawa (1998) and Liu and West (2001). Unlike these previous approaches, iterated filtering then reduces the intensity of the random walk to identify the maximum of the likelihood function for the original, unperturbed model. The updating step in iterated filtering is a weighted average of the time-localized estimates  $\{\tilde{\theta}_{n,m}^F, n = 0, \dots, N\}$ , in the sense that the coefficients on the right hand side of (2.12) add up to unity. These coefficients are not necessarily positive, though they become so asymptotically (Ionides et al., 2006). One can therefore think of the sequential Monte Carlo particles in each iteration of Algorithm 1 as exploring parameter space and their discoveries being gathered together by (2.12) to give the starting point for the next iteration. This heuristic explains how one iteration of iterated filtering (which has essentially the same computational effort as one evaluation of the likelihood function) can result in considerable progress toward finding appropriate values of  $\theta$  to match the data. Iterated filtering has been shown to effectively maximize the likelihood for a 13 dimensional parameter space based on 50 iterations (Ionides et al., 2006). By contrast, a direct attempt to construct one single noisy estimate of the derivative of the log likelihood would usually require  $13 + 1$  function evaluations in this context. When pushing model complexity to the computational limits permissible for likelihood-based inference, numerical efficiency becomes a relevant consideration.

If one departs from the plug-and-play paradigm, then one would expect a reduction in the required computational effort. For example, sequential Monte Carlo schemes that have access to derivatives of  $f_{X_n|X_{n-1}}(x_n | x_{n-1}; \theta)$  and  $f_{Y_n|X_n}(y_n | x_n; \theta)$  with respect to  $\theta$  can estimate the derivative of the log likelihood in a single smoothing operation (Poyiadjis et al., 2009). Access to evaluation of  $f_{X_n|X_{n-1}}(x_n | x_{n-1}; \theta)$  makes available other standard algorithms for the calculation and maximization of the likelihood function via sequential Monte Carlo (e.g., Pitt and Shepard, 1999; Pitt, 2002). However, we suspect that there will be a continuing demand for plug-and-play inference methodology for dynamic systems just as there is a continuing demand for derivative-free procedures to optimize deterministic functions.

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 sequential Monte Carlo 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  $J_m \rightarrow \infty$  in Theorem 2.2.3). As another example, methods based on investigating the likelihood function by fitting a spline approximation to sequential Monte Carlo estimates (Olsson and Rydén, 2008) become feasible on increasingly large problems once the maximum has been identified to within a reasonable amount of Monte Carlo uncertainty.

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 computationally efficient identification of statistically plausible models—the main strength of iterated filtering—is also the key step in model-based data analysis.

A limitation of the mathematical analysis in this chapter is the use of the basic sequential Monte Carlo scheme in Algorithm 1. Various strategies have been proposed to improve the numerical efficiency of sequential Monte Carlo (Cappé et al., 2007). However, the effectiveness of these algorithms is dependent on the details of specific models. Further, these schemes typically do not enjoy the plug-and-play property. In our experience, the usual cause of poor numerical performance (i.e., high Monte Carlo variability) is an attempt to fit a model that is inappropriate for the data. Heuristically, this is to be expected because numerical instability occurs when none or few of the sequential Monte Carlo particles are consistent with an observation. Since plug-and-play methods facilitate the development of new models and the investigation of variations on existing models, a practitioner using plug-and-play methodology can focus on developing a suitable model rather than becoming sidetracked in the pursuit of a customized inference algorithm to handle the numerical consequences of fitting an inappropriate model.

Beyond model mis-specification, another cause of poor numerical performance in sequential Monte Carlo schemes can be a model featuring highly accurate mea-

surements. A small measurement error leads to few of the particles being consistent with successive observations and consequent degeneracy of the resampling weights in Algorithm 1. In our view, this arises less often than might be expected because superficially accurate measurements typically have an uncertain relationship to the unobserved system. Practical dynamic models are idealizations of a system, and one can expect some error and uncertainty in the relationship between the idealized system variables and the measurable quantities. Such uncertainty naturally enters a statistical model as stochastic variability, even though it could be thought of as unknown (and perhaps unknowable) systematic error.

We do not wish to discourage the development of increasingly sophisticated sequential Monte Carlo schemes. Indeed, implementations of iterated filtering stand to benefit from the potential numerical efficiency of such techniques since the fundamental justification of iterated filtering (i.e., Theorem 2.2.1) simply calls for the existence of a numerically tractable filter. However, the arguments in the preceding two paragraphs help to explain why plug-and-play methods based on sequential Monte Carlo, such as the iterated filtering algorithm studied here, are more widely applicable than might have been anticipated.

#### **2.4.1 Case studies of iterated filtering**

Scientific applications of partially observed Markov process modeling typically require an entire paper to describe the scientific context, the model developed, the data, the inference procedures applied, the results and the conclusions. Substantial applications are, however, the ultimate demonstration of the potential of an inference approach. Such case studies exist for iterated filtering (King et al., 2008; Bretó et al., 2009; He et al., 2010; Ionides et al., 2006) and we direct the reader to these for fully worked examples. We limit ourselves here to discussion of points of general interest arising from these applications. These practical implementations

did not employ the increasing Monte Carlo sample size suggested by Theorem 2.2.3 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 sequential Monte Carlo can provide an unbiased estimate of the likelihood function (a consequence of Theorem 2.3.2 in Section 2.3.3) it is relatively straightforward to confirm whether the likelihood has indeed been successfully maximized.

A consideration for improving the performance of many Monte Carlo parameter estimation procedures, including both Markov chain Monte Carlo and iterated filtering, is to ensure that the scale of the random jumps in the parameter space is comparable to the scale of the uncertainty in each parameter. This is equivalent to reparameterizing the model so that the uncertainty in each parameter is approximately at a unit scale. Working with positive parameters on a logarithmic scale and  $(0, 1)$  valued parameters on a logistic scale has been an adequate resolution in our experience. Diagnostics and heuristics for convergence of iterated filtering were discussed by Ionides et al. (2006).

Maximization of the likelihood, which is the central topic of this chapter, is a basic building block for a complete data analysis. Successful likelihood maximization permits not just point estimates but also profile likelihood analysis (to construct confidence intervals) and likelihood-based model comparisons. Plug-and-play methodology facilitates the fitting of variations on the primary model, and the maximized likelihoods can then be compared by likelihood ratio tests or Akaike's information criterion. The likelihoods of simple alternatives, such as linear regression models or autoregressive moving average models, should be computed as benchmark comparisons to check whether more sophisticated models in fact provide a superior explanation of the data. Residual analyses can be carried out, adapting to dynamic models the techniques that have become standard for

regression analysis.

The existing demonstrations of the scientific value of iterated filtering methodology all take advantage of the plug-and-play property. Iterated filtering implementations can in principle take advantage of numerically efficient variations on sequential Monte Carlo which lack the plug-and-play property. However, the vista of new models which can be analyzed given the availability of effective plug-and-play methodology makes an attractive motivation for focusing on this property. Beyond case studies employing iterated filtering, other recently proposed plug-and-play methodology has also led to the development and analysis of new scientific models (Andrieu et al., 2010). As plug-and-play methodology becomes more widely employed, an increasing number of models will be developed which take advantage of the general applicability of such techniques.

## CHAPTER 3

# Malaria in Northwest India: Data analysis via partially observed stochastic differential equation models driven by Lévy noise

### 3.1 Introduction

Malaria is currently a widespread tropical and sub-tropical disease, with approximately 500 million cases per year (Snow et al., 2005) resulting in over one million deaths (Hay et al., 2005). Malaria is caused by infection with a protozoan parasite which is transmitted between humans by mosquitoes. The disease was eliminated from North America and Europe during the first half of the 20th century, primarily by sanitary and agricultural developments which reduced contact between humans and mosquitoes below the level required to sustain disease transmission (Packard, 2007). From 1955 to 1969 the World Health Organization ran an ambitious Global Malaria Eradication Program, based on mosquito control by extensive spraying with the insecticide DDT and treatment with the anti-malarial drug chloroquine (Packard, 2007). In India, malaria incidence declined dramatically during the Global Malaria Eradication Program. A crippling burden of approximately 75 million cases per year was reduced to a reported incidence of 49,151 in 1961 (Kumar et al., 2007). However, rather than continuing

this decline, malaria incidence crept up through the 1960's. The re-emergence in India has been attributed to the increasing cost and decreasing supply of DDT, resistance developed by mosquitoes to DDT, and increasing resistance of malaria parasites to chloroquine (Kumar et al., 2007; Sharma, 1996). After increasing to over 6 million reported cases annually in the 1970's, malaria incidence has since stabilized at around 2 million cases per year (Kumar et al., 2007). These official statistics are an indication of the trend of incidence but fail to include many cases which are treated outside the public health system. A more accurate estimate of recent incidence may be 11 million cases per year (World Health Organization, 2008).

Hopes for a global eradication of malaria have recently been raised once more. Eradication has been stated as an explicit goal of the Bill and Melinda Gates Foundation, with the endorsement of the World Health Organization and the Roll Back Malaria Partnership (Roberts and Enserink, 2007). The main technologies underpinning this aspiration are long lasting insecticide-treated bed nets and a new generation of artemisinin-derived anti-malarial drugs. Although global eradication is probably unrealistic with currently available tools (Greenwood, 2009), there is great potential to reduce the heavy global burden of malaria. One of the lessons learned from the previous eradication program is that effective control requires adaptation to local patterns of disease transmission (Greenwood, 2009). Improved quantitative understanding of transmission is therefore a necessary component of control and prevention efforts.

The early mathematical models of Ross (1911) and Macdonald (1957) have long been a foundation for developing malaria control strategies (McKenzie and Samba, 2004). Many extensions have been proposed to these mathematical models, allowing for biological aspects such as genetic diversity of the parasite (Gupta et al., 1994; McKenzie et al., 2008), the mosquito and parasite lifecycle (McKenzie and

Bossert, 2005), the development of drug resistance (Koella and Antia, 2003; Klein et al., 2008), and exposure-dependent partial immunity (Dietz et al., 1974; Aron and May, 1982; Filipe et al., 2007). Given the size of the public health issue and the extent of the research into malaria transmission, it may be surprising how few studies investigate the relationship between these dynamic models and available population-level time series data. Investigations relating disease models (which are typically partially observed nonlinear Markov processes) to time series data have a long tradition of inspiring developments in statistical analysis of stochastic dynamic systems (Bartlett, 1960; Ellner et al., 1998; Finkenstädt and Grenfell, 2000; Ionides et al., 2006; Cauchemez et al., 2008). Indeed, the most convenient disease systems to study, such as measles, are still considered a challenge for statistical inference (Cauchemez and Ferguson, 2008; He et al., 2010). Analysis of measles dynamics is simplified by clear clinical diagnosis, direct human-to-human transmission, lifelong immunity following infection, and the availability of extensive spatio-temporal incidence data. The study of malaria dynamics is hindered by nonspecific symptoms; one usually has to work under the assumption that malaria is the cause of sickness for patients who have a high fever and are found, by inspection of a blood slide under a microscope, to be infected with *Plasmodium* parasites. However, asymptomatic *Plasmodium* infections are not unusual, and there are many alternative potential causes of fever. Secondly, human immunity to malaria wanes with time and gives varying levels of protection to diverse disease strains. Clinical immunity (i.e., protection to symptomatic infection) can result from repeated infections, and leads to infections with a reduced transmissibility. Thirdly, malaria transmission is dependent on mosquito abundance. Malaria transmission is highly sensitive to the density, longevity and biting habits of the mosquito vector. These entomological quantities vary considerably in space and time, both within and between vector species (Packard, 2007). Time series of vector abundance and behavior directly

relevant to long-term population-level studies are therefore generally unavailable.

In Section 3.2, we develop a quantitative approach to relate malaria transmission to available time series data. We aim to construct statistical models of the population-level transmission dynamics which are at once sophisticated enough to capture the important features of the biological system and simple enough that they can be rigorously assessed using available data. Mathematically, our models are a set of coupled nonlinear system of stochastic differential equations driven by Lévy noise. Whereas certain specific models could be constructed using the more usual choice of Gaussian noise, a general framework which satisfies necessary non-negativity constraints can more readily be built using non-negative noise built from non-decreasing Lévy processes such as the Gamma process. Lévy process models have been proposed for a range of applications, ranging from option pricing in finance to quantum mechanics (Applebaum, 2004). However, statistically efficient inference from general classes of nonstationary partially observed systems driven by Lévy noise has not previously, to our knowledge, been demonstrated. Here, we use the term *statistically efficient* in an informal sense, to describe methodology leading to parameter estimates whose uncertainty approximates that of Bayesian or likelihood-based estimates. Statistical efficiency becomes an important consideration when building models whose complexity is at, or close to, the limit which the available data can support.

Recently, statistically efficient methodology for general partially observed Markov process (POMP) models has been proposed (Ionides et al., 2006; Andrieu et al., 2010). The generality of such methodology is based on possession of the so-called *plug-and-play* property (Bretó et al., 2009; He et al., 2010; Ionides et al., 2009); methodology for POMP models is said to have the plug-and-play property if the dynamic model enters into the inference procedure only through the availability of numerical solutions (i.e., simulated sample paths). The theory of numerical solu-

tion of SDEs driven by general Lévy noise closely follows the well-studied special case of Gaussian noise (Protter and Talay, 1997; Jacod, 2004). One might hope, therefore, that plug-and-play methodology would be applicable to such models. In our case study, we demonstrate that this is indeed the case, by carrying out inference as a routine application of a recently developed likelihood-based plug-and-play technique called iterated filtering (Ionides et al., 2006). By comparison, standard expectation-maximization and Markov chain Monte Carlo approaches (Cappé et al., 2005) require the evaluation of transition densities—this can cause difficulties, or even complete failure, on continuous time POMP models (Roberts and Stramer, 2001). We predict that the development of plug-and-play methodology will greatly extend the classes of dynamic models used for data analysis.

Section 3.3 presents a data analysis, through which we aim both to demonstrate our statistical approach and to draw conclusions about the respective roles of immunity and climate variability for epidemic malaria transmission. Epidemic or ‘unstable’ malaria (Molineaux, 1988; Kiszewski and Teklehaimanot, 2004) occurs when conditions are only occasionally favorable for disease transmission, for example due to cold or dry seasons which preclude mosquito activity. Waning of immunity during the absence of exposure to malaria can lead to high levels of severe infection in epidemics. By contrast, the repeated exposures in regions of endemic or ‘stable’ malaria result in acquisition of immunity that protects from severe forms of the disease. We focus on two questions. Firstly, what is the appropriate degree of model complexity which is necessary to understand population dynamics of epidemic malaria? This issue is basic to developing scientifically acceptable models for malaria which quantitatively match population-level incidence data. Secondly, what is the role of climate fluctuations, such as interannual changes in rainfall patterns, for determining the interannual variability of disease incidence? Despite agreement on the sensitivity of the mosquito vector to environ-

mental conditions, there has been considerable controversy on the respective roles of environmental forcing vs. epidemiological considerations, fueled by the lack of a quantitative statistical approach which can make a formal comparison of rival hypotheses. In particular, for malaria in East African highlands, some investigators have found that interannual variability in rainfall and temperature can explain a substantial share of the variability in regional malaria incidence time series (Pascual et al., 2008; Zhou et al., 2004), whereas others have proposed that oscillating levels of immunity in the population act as the major driver (Hay et al., 2000, 2002). We broaden this specific debate by analyzing data from another unstable malaria transmission environment, in an arid region of Northwest India, where the role of rainfall variability is less controversial but has not been addressed together with immunity in the context of the population dynamics of the disease. It is in desert and highland regions, at the edge of the distribution of the disease, that we expect climate variability and climate change to be potentially most relevant to disease dynamics due to the limiting roles of rainfall and temperature. The data analysis in this chapter focuses on a newly available malaria incidence time series for the Kutch district, an arid region in the state of Gujarat. The scientific argument is expanded on elsewhere (Laneri et al., 2010), and our primary goal here is to describe the statistical foundations for building and analyzing dynamic models of population-level malaria transmission that can be confronted to time series data.

## **3.2 Malaria transmission: A statistical model**

We start by describing some relevant biology; for a more complete introduction we recommend Warrell and Gilles (2002) or the article on malaria in Wikipedia (2010). The unicellular protozoan parasites of the genus *Plasmodium* which cause malaria are transmitted between humans by the female of certain species of *Anophe-*

les mosquito. The *Plasmodium* lifecycle consists of multiple stages in both human and mosquito hosts. When a mosquito takes a blood meal from an infected human, male and female *Plasmodium* gametocytes may be ingested. Sexual reproduction of the parasite takes place within a vector mosquito's stomach, resulting in the formation of sporozoites which migrate to the mosquito's salivary glands. Upon a subsequent blood meal, the sporozoites can infect another human—entering the bloodstream, becoming sequestered in the liver, reemerging into the blood, reproducing asexually in erythrocyte stages, and eventually producing gametocytes to complete the cycle. During the stages in a human host, the *Plasmodium* must do battle with the complex human immune system which attacks sporozoite, erythrocyte and gametocyte stages (Artavanis-Tsakonas et al., 2003). The effectiveness of the immune response depends, amongst other things, on system memory from previous exposure to related parasites. Transmission of malaria relies upon the availability of infected humans, susceptible humans, and mosquitoes having sufficient longevity. The mosquito longevity is critical for the viability of the *Plasmodium* lifecycle since the time taken for the *Plasmodium* to undergo ingestion, reproduction, development and retransmission to a human host is comparable to the mean lifespan of the mosquito.

The majority of severe and fatal human malaria cases are caused by infection with *P. falciparum*. The other widespread species is *P. vivax*, which is characterized by less severe symptoms with the possibility of relapse many months after infection. To develop a quantitative representation, we will write down a model for falciparum malaria (i.e., disease resulting from infection with *P. falciparum*) which captures some key aspects of the human, parasite and vector dynamics. This model could be extended to vivax malaria by the inclusion of relapse. Our goal is to present a *statistical model* in the sense that it is sufficiently parsimonious that the parameters can be estimated directly from available data, as carried out in Section 3.3.

We divide humans into five distinct classes:  $S_1$ , fully susceptible to infection;  $S_2$ , protected from severe infection, but susceptible to mild reinfection;  $E$ , exposed (i.e., carrying *Plasmodium* parasites which have not yet matured into gametocytes);  $I_1$ , infected and gametocytemic;  $I_2$ , possessing a mild, asymptomatic infection with reduced gametocyte levels (Klein et al., 2008; Filipe et al., 2007). An innovative feature of this framework, compared to other epidemiological models previously fitted to population-level time series data, is the inclusion of an explicit representation of the vector dynamics: A mosquito stage  $\kappa$  represents the latent force of infection, capturing the likelihood of successful transmission from human to human; a mosquito stage  $\lambda$  represents the current force of infection, which consists of the latent infection lagged by a distributed delay corresponding to the time for development of the *Plasmodium* parasite among surviving mosquitoes. By representing mosquito dynamics through a model for the force of infection of humans, we avoid explicit consideration of mosquito abundance, survival and behavior. In other words, we limit our inclusion of vector dynamics to the aspect that is most directly relevant to the human disease.

Figure 3.1 represents diagrammatically the modeled flows between these classes, formally defined by equations (3.1–3.9) below. We write  $\mu_{XY}$  for the rate of transition from class  $X$  to class  $Y$ , for  $X$  and  $Y$  in  $\{S_1, S_2, E, I_1, I_2\}$ . In addition, we introduce a per-capita birth rate,  $\mu_{BS_1}$ , into the completely susceptible class. Deaths occur at a constant rate  $\mu_{XD} = \delta$  from each class  $X \in \{S_1, S_2, E, I_1, I_2\}$ . As mortality from acute malarial infection has become small in India, we do not include disease-induced mortality in our model. The total population size  $P(t)$  is supposed known by interpolation from the decennial census. The *force of infection*,  $\lambda(t)$ , is simply an epidemiological term for  $\mu_{S_1E}$  and so we have the identity  $\mu_{S_1E}(t) = \lambda(t)$ . Transition from  $S_2$  to  $I_2$  can be interpreted as reinfection with clinical immunity, i.e., reduced symptoms which do not lead the patient to seek medical

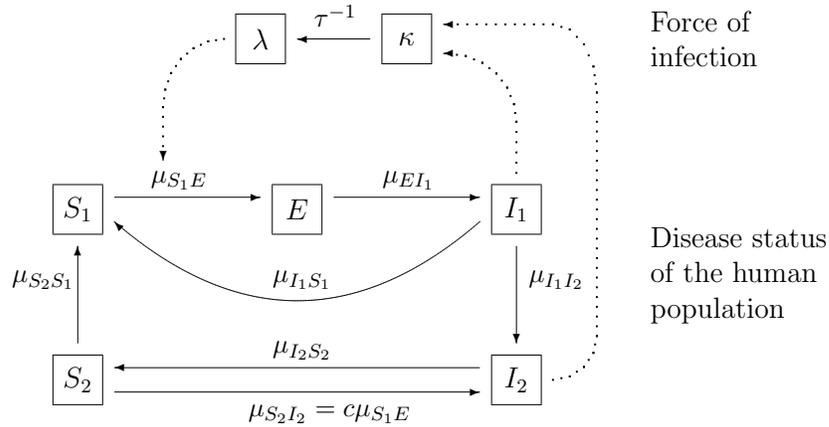


Figure 3.1: Flow diagram for a compartment model of malaria transmission. Human classes are  $S_1$  (susceptible),  $S_2$  (partially protected),  $E$  (exposed, carrying a latent infection),  $I_1$  (infected and infectious) and  $I_2$  (asymptomatic, with reduced infectivity). The possibility of transition between class  $X$  and  $Y$  is denoted by a solid arrow, with the corresponding rate written as  $\mu_{XY}$ . The dotted arrows represent interactions between the human and mosquito stages of the parasite. Mosquito dynamics are modeled via the two stages  $\kappa$  (the latent force of infection) and  $\lambda$  (the current force of infection), with  $\tau$  being the mean latency time. The model, which we call VS<sup>2</sup>EI<sup>2</sup> with ‘V’ for ‘vector’ followed by a list of the human classes with their multiplicities as superscripts, is formalized by equations (3.1–3.8). We also consider the subcase with  $\mu_{I_2S_2} = \infty$  and  $\mu_{S_2I_2} = \mu_{I_1S_1} = 0$ . The class  $I_2$  can then be eliminated, and so transition directly from  $I_1$  to  $S_2$  becomes possible. Also, individuals in  $S_2$  are fully protected in this case. The remaining classes  $\{S_1, E, I_1, S_2\}$  can then be mapped onto the classes  $\{S, E, I, R\}$  in a standard epidemiological susceptible-exposed-infected-recovered model (Anderson and May, 1991; Keeling and Ross, 2008) with added vector dynamics and waning immunity; we therefore call this special case VSEIR.

attention (Gupta et al., 1999; Artavanis-Tsakonas et al., 2003). We suppose that  $\mu_{S_2 I_2} = c\mu_{S_1 E}$  with some constant of proportionality  $0 \leq c \leq 1$ . Our model also includes the possibility of failing to acquire any protective immunity following infection, by transitioning directly from  $I_1$  back to  $S_1$  without passing through  $I_2$  and  $S_2$ . This can arise through prompt treatment with antimalarial drugs, in which case the body does not have time to build an immune response (Klein et al., 2008). Alternatively, it can be a consequence of the necessity for multiple infections before the body learns to mount an effective general-purpose defense against clinical symptoms in the face of the genetic diversity of the *Plasmodium* (McKenzie et al., 2008). The resulting system of coupled nonlinear stochastic differential equations is as follows:

$$dS_1/dt = \mu_{BS_1}P - \mu_{S_1E}S_1 + \mu_{I_1S_1}I_1 + \mu_{S_2S_1}S_2 - \mu_{S_1D}S_1 \quad (3.1)$$

$$dS_2/dt = \mu_{I_2S_2}I_2 - \mu_{S_2S_1}S_2 - \mu_{S_2I_2}S_2 - \mu_{S_2D}S_2 \quad (3.2)$$

$$dE/dt = \mu_{S_1E}S_1 - \mu_{EI_1}E - \mu_{ED}E \quad (3.3)$$

$$dI_1/dt = \mu_{EI_1}E - \mu_{I_1S_1}I_1 - \mu_{I_1I_2}I_1 - \mu_{I_1D}I_1 \quad (3.4)$$

$$dI_2/dt = \mu_{I_1I_2}I_1 + \mu_{S_2I_2}S_2 - \mu_{I_2S_2}I_2 - \mu_{I_2D}I_2 \quad (3.5)$$

$$d\kappa/dt = d\lambda_0/dt = (f(t) - \kappa) n_\lambda \tau^{-1} \quad (3.6)$$

$$d\lambda_i/dt = (\lambda_{i-1} - \lambda_i) n_\lambda \tau^{-1} \quad \text{for } i = 1, \dots, n_\lambda - 1 \quad (3.7)$$

$$d\lambda/dt = d\lambda_{n_\lambda}/dt = (\lambda_{[n_\lambda-1]} - \lambda) n_\lambda \tau^{-1} \quad (3.8)$$

The malarial status of the human population is represented by the differential equations in (3.1–3.5), which correspond to a large population limit of homogeneous individual-level interactions where each individual has exponentially distributed transition times. Sometimes, consideration of non-exponential transition times can be worthwhile (Wearing et al., 2005), though we consider this only in the mosquito stages. Specifically, (3.6–3.8) correspond to Gamma-distributed transitions for

the latent period of the force of infection. The extra flexibility in the shape of the Gamma distribution over the exponential distribution may be appropriate since the development time of the *Plasmodium* in the mosquito gives rise to a (temperature-dependent) lower bound. We suppose that the main stochasticity in this system arises from variations in vector abundance and behavior, which is modeled in the specification of  $f(t)$  by including multiplicative Gamma noise:

$$f(t) = \frac{I_1(t) + qI_2(t)}{N(t)} \bar{\beta} \exp \left\{ \sum_{i=1}^{n_s} \beta_i s_i(t) + Z_t \beta \right\} \frac{d\Gamma}{dt}. \quad (3.9)$$

Here,  $q$  represents the transmissibility, relative to full-blown infections, from asymptomatic infections in partially immune individuals; the seasonality of disease transmission is modeled by the coefficients  $\{\beta_i\}$  corresponding to a periodic cubic B-spline basis  $\{s_i(t), i = 1, \dots, n_s\}$  constructed using  $n_s$  evenly spaced knots; time-varying covariates enter via the row vector  $Z_t$  with coefficients in a column vector  $\beta$ ; the dimensional constant  $\bar{\beta}$  is required to give  $f(t)$  units of  $t^{-1}$ , and we set  $\bar{\beta} = 1\text{yr}^{-1}$ .  $\Gamma(t)$  is a Gamma process representing integrated noise with intensity  $\sigma^2$ . This is defined as a process with stationary independent increments such that  $\Gamma(t) - \Gamma(s) \sim \text{Gamma}((t-s)/\sigma^2, \sigma^2)$  where  $\text{Gamma}(a, b)$  is the Gamma distribution with mean  $ab$  and variance  $ab^2$ . Although  $\Gamma(t)$  is a jump process, and therefore its sample paths are not differentiable, one can interpret the noise process  $d\Gamma/dt$  in (3.9) as multiplicative Gamma noise (Bretó et al., 2009). The reason to choose Gamma noise over the more familiar Gaussian noise is to enforce the positivity of  $f(t)$  and hence all the state variables in (3.1–3.8). Gamma noise is perhaps the simplest and most-studied non-negative Lévy noise process (Applebaum, 2004; Bretó et al., 2009). We solve (3.1–3.8) numerically via the Euler method (Protter and Talay, 1997; Jacod, 2004) with a time-step of one day. Whereas all state variables in the unavailable exact solutions to (3.1–3.8) are non-negative, it is possible for

the Euler method to generate numerical approximations violating this constraint. We monitored the frequency of these occurrences; they were rare to the point of negligibility in our analysis.

At first inspection, (3.1–3.9) may appear to be a dauntingly complex model specification based on many assumptions that one cannot hope to validate. However, this work builds on a long history of developing and using similar models (Keeling and Ross, 2008; Anderson and May, 1991). All the parameters in (3.1–3.9) have interpretable scientific meaning and can therefore be discussed in the context of the literature on malaria transmission. Indeed, our model can also be criticized as an over-simplification, since we do not incorporate many of the biological aspects developed in previous models (such as Chitnis et al., 2006; Gupta et al., 1994; McKenzie and Bossert, 2005; Koella and Antia, 2003). In addition, our model does not make allowances for spatial, socio-economic, age-related and genetic inhomogeneities among the population. Such structure could play an important role. Nevertheless, models based on homogeneous populations are often sufficient to describe the major features of disease transmission dynamics (Earn et al., 2000; Keeling and Ross, 2008). In the face of biological complexity, a major part of the value of constructing and analyzing dynamic models is to develop an understanding of the key components driving the behavior of the biological system. In our modeling framework, alternative model specifications can readily be analyzed and compared, building on the results reported here.

A measurement model provides a formal connection between the dynamic process model and available data. Here we give an abstract representation, deferring concrete discussion of data to Section 3.3. We write  $\{t_n, n = 1, \dots, N\}$  for the times of the  $N$  observations, and we suppose that the model is initialized at some time  $t_0 < t_1$ . We define the number of new cases in the  $n$ th interval to be  $C_n = \int_{t_{n-1}}^{t_n} \mu_{EI_1} E(s) ds$ . The reported number of confirmed cases,  $y_n$ , is then

modeled conditional on  $C_n$  as

$$y_n | C_n \sim \text{Negbin}(\rho C_n, \psi^2), \quad (3.10)$$

where  $\text{Negbin}(\alpha, \beta)$  is the negative binomial distribution with mean  $\alpha$  and variance  $\alpha + \alpha^2\beta$ . This distribution allows for the possibility of over-reporting or under-reporting, and can be viewed as an over-dispersed Poisson distribution with dispersion parameter  $\psi$ . We refer to  $\rho$  as the reporting rate. It is known that only a small fraction of malaria cases are treated in the public clinics which contribute to district statistics (Kumar et al., 2007), so we expect  $\rho \ll 1$ . The exact interpretation of  $\rho$  is necessarily sensitive to the severity of disease that is required to be classified as a case.

Although environmental covariates affect many biological systems, quantifying their dynamic role can be a formidable task, both from a scientific and a statistical perspective (Bjørnstad and Grenfell, 2001). The flexibility of plug-and-play statistical methodology permits scientific considerations to determine ways in which covariates might appropriately be included in the analysis. Here, we take  $Z_t$  to be a scalar covariate measuring the thresholded rainfall integrated over a time interval  $[t - u, t]$ . Specifically, from the accumulated rainfall data  $\{r_n, n = 1, \dots, N\}$  at times  $t_1, \dots, t_N$  we interpolated a continuous-time cubic spline  $r(t)$  and then set

$$\tilde{Z}_t = \max \left\{ \int_{t-u}^t r(s) ds - v, 0 \right\}. \quad (3.11)$$

The specification in (3.11) is designed to represent parsimoniously the threshold and lag effects which are to be expected in biological systems (Stenseth et al., 2004). The covariate was standardized by setting  $Z_t = (\tilde{Z}_t - \bar{Z})/\sigma_Z$ , where  $\bar{Z} = (t_N - t_0)^{-1} \int_{t_0}^{t_N} \tilde{Z}_s ds$  and  $\sigma_Z^2 = (t_N - t_0)^{-1} \int_{t_0}^{t_N} (Z_s - \bar{Z})^2 ds$ . This standardization makes the coefficient  $\beta$  a dimensionless quantity which is expected to vary on a

unit scale.

### 3.3 Data analysis

Figure 3.2 shows a plot of the monthly confirmed cases of *P. falciparum* and monthly rainfall in the district of Kutch in the state of Gujarat in Northwest India between January, 1987 and December, 2006. The record of the malaria cases was obtained from the National Institute of Malaria Research in India, and was originally compiled by the office of the District Malaria Officer. The rainfall time series was obtained from a local district weather station run by the Indian Meteorology Department. Rainfall in Kutch is concentrated within the monsoon season, and Kutch experiences the seasonal epidemic malaria typical of arid regions of India (Swaroop, 1949; Bouma et al., 1996; Kiszewski and Teklehaimanot, 2004). Visually, a lag relationship, with rainfall leading malaria, may seem evident from this figure. Since rainfall typically peaks during the summer monsoon and malaria typically peaks a few months later, in late fall, one might see the appearance of a lag relationship in the absence of a direct link. The correlation between total monsoon rainfall (aggregated over June-August) and total fall cases (aggregated over October-December) is 0.84 over these twenty years, which is suggestive of a causal relationship. However, the intensity of monsoon rainfall has cycles of 2-4 years which matches cycles that are predicted in malaria due to the building up of population immunity in epidemics followed by subsequent waning of immunity and birth of newly susceptible children (Pascual et al., 2008). This confounding of intrinsic cycles (e.g., immunity) with the effect of extrinsic cycles (e.g., climate variability) adds difficulty to the interpretation of such correlations. Modeling both intrinsic and extrinsic effects simultaneously provides a way to strengthen scientific conclusions. This is analogous to using multiple regression to control for potential confounding variables, but here we must take into account the nonlinear stochastic

feedbacks and lagged relationships in the dynamic system. In this investigation, we fixed the rainfall covariate in (3.11) by setting  $u = 5\text{mo}$  and  $v = 200\text{mm}$ . Additional analysis of the relationship with rainfall will be published elsewhere (Laneri et al., 2010), but it should be clear that the approach we develop here has the flexibility to address alternative hypotheses concerning this as well as many other questions about malaria dynamics.

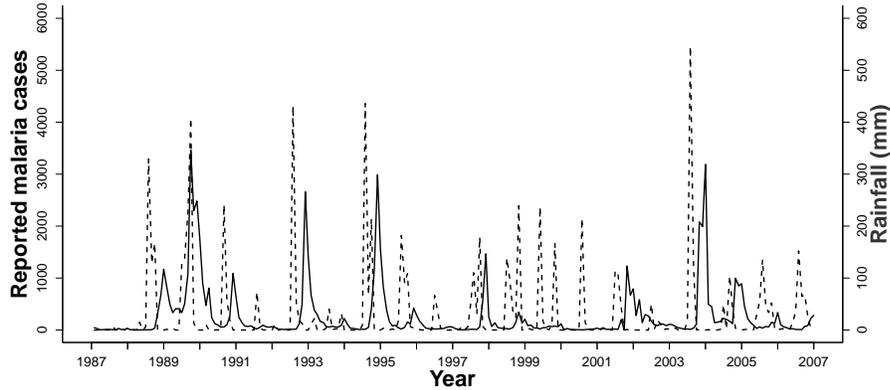


Figure 3.2: Monthly reported *P. falciparum* malaria cases (solid line) and monthly rainfall from a local weather station (broken line) for Kutch.

We carried out likelihood-based inference via iterated filtering, a plug-and-play sequential Monte Carlo procedure for calculating maximum likelihood estimates which was introduced by Ionides et al. (2006). Iterated filtering was implemented using the `pomp` software package (King et al., 2009) which encodes the algorithm presented by King et al. (2008, supplementary text). Computer code to generate an Euler solution to the dynamic model described by equations (3.1–3.9) and to evaluate the density of the measurement model in (3.10) is all that the user need supply to embark on statistical analysis via general-purpose software implementing such a plug-and-play procedure. We refer the reader to the online supplement, and to the relevant literature (King et al., 2008; Bretó et al., 2009; He et al., 2010; Ionides et al., 2009), for further discussion of iterated filtering methodology.

There are tuning parameters which affect the numerical efficiency of the maximization algorithm; the values we used are reported in the online supplement. These algorithmic parameters are inconsequential for the inferential conclusions once numerical convergence has been confirmed by checking consistency over a range of starting values for the likelihood maximization. If all the model parameters share a unit scale of variability, selection of the algorithmic parameters is simplified. With this in mind, we worked with the logarithmic transform of non-negative parameters and the logit transform of parameters taking values in the interval  $(0, 1)$ . On this common scale, standard values of the algorithmic parameters gave acceptable optimization performance. All reported results are transformed back to the original scale.

A simple, but valuable, diagnostic for the specification of a mechanistic model is to compare the goodness of fit with standard non-mechanistic statistical models. One can argue that part of the point of fitting a mechanistic model to data is to discover which aspects of the data are not captured by a model describing current scientific knowledge about the system under investigation. Somewhat equivalently, one might understand that requiring a model to have scientific interpretability may lead to a cost in terms of the ability to match data statistically. In this sense, it may not be a scientific goal to achieve a level of fit comparable to flexible statistical models which do not seek scientific interpretability. On the other hand, to carry out formal hypothesis tests, or to interpret parameter estimates and their uncertainty, it is helpful if the model can be shown to give an adequate statistical fit to the data. In Table 3.1, we include as a benchmark comparison a model in which  $\{\log(y_n + 1), n = 1, \dots, N\}$  is supposed to follow a Gaussian SARIMA specification. The large number of additional parameters in the mechanistic models appears to be justified relative to this log-SARIMA model, according to the AIC criterion. Log-SARIMA models are theoretically appealing as simple models for

disease transmission, since (in common with many other biological populations) the *Plasmodium* demonstrates annual cycles of abundance which consist approximately of a period of exponential growth followed by a period of exponential decay. As another benchmark, we included the rainfall covariate  $Z_t$  into the log-SARIMA model (via the ARMAX framework; Shumway and Stoffer, 2006), also reported in Table 3.1. The improvement in model fit from including the covariate is comparable, in terms of units of log likelihood, to the improvement seen in the VSEIR and VS<sup>2</sup>EI<sup>2</sup> models.

It is a substantial computational challenge to investigate a non-convex and potentially multimodal likelihood function, with around 20 parameters, based on Monte Carlo estimates of the likelihood which involves integrating over all the unobserved state variables at  $(t_N - t_0)/\Delta = 20 \times 365$  time points. However, verifying that this function is indeed adequately maximized, once this has been achieved, is relatively straightforward. One check is to confirm that the maximization is robust to different starting values and Monte Carlo replications (i.e., choices of the random number generation seed). In addition, we construct profile likelihood plots (one example is given in Figure 3.4) and check that each profile consistently attains the maximized likelihood. We have found profile likelihood calculations particularly useful for ensuring that the dynamic system is investigated across a range of parameter values, facilitating the discovery of new modes of the likelihood function.

From Table 3.1, we see that all the four mechanistic models analyzed beat the benchmark non-mechanistic log-SARIMA model by a large margin of AIC. Having established that these models are adequate statistical explanations of the data, we compare these models amongst each other. Likelihoods for both the VS<sup>2</sup>EI<sup>2</sup> model and the simpler VSEIR submodel (described in the caption to Figure 3.1) improve significantly when the rainfall covariate is used ( $p < 0.001$  for the likelihood ratio

test, using a chi-square approximation on one degree of freedom). After concluding that inclusion of rainfall does indeed help to describe malaria dynamics, we proceed to compare the VSEIR and the  $VS^2EI^2$  models, both including rainfall. These two models have different numbers of parameters and we can compare their Akaike Information Criterion (AIC) values, which favors the  $VS^2EI^2$  model with rainfall. Since these two models are nested, one can also carry out a likelihood ratio test of the null hypothesis that the data follow the VSEIR model ( $p < 0.001$ , chi-square test on 5 degrees of freedom). The nesting is nonstandard (e.g., when  $\mu_{I_2S_2} \rightarrow \infty$  the initial value  $[I_2]_0$  becomes undefined), however, the chi-square test is expected to be conservative in such situations (Self and Liang, 1987; Anisimova et al., 2001). We consider this comparison to be evidence for the value of incorporating characteristic aspects of the human immune response to malaria into models used for time series analysis. However, models based on simpler SEIR descriptions of human immunity will continue to be central to the study of disease dynamics, and our results also support a position that the VSEIR model is not entirely discredited. It produces parameter estimates which are qualitatively similar to the  $VS^2EI^2$  model, and its log likelihood is much closer to that of the  $VS^2EI^2$  than to the log-SARIMA benchmark. To understand the relative strengths and weaknesses of different models, one pertinent question to consider is which parts of the time series are better explained by each model. In Figure 3.3 we plot the difference of the conditional log likelihoods of the  $VS^2EI^2$  model with rainfall and the VSEIR model with rainfall, at each point in time. We note that during many of the epidemics, most notably in the fall of 1989, 1990, 1992, 1994 and 1997, the simpler VSEIR model fits the data better as the epidemic approaches its peak. Predicting the peak of an epidemic is of particular public health interest, as it determines the maximum case burden experienced by the health care system. The more complex  $VS^2EI^2$  model, which fits the data better overall, may have little or no advantage

for this specific purpose.

Table 3.1: A likelihood-based comparison of the fitted models. Corresponding point estimates are presented in Table 3.3. The column labeled  $p$  corresponds to the number of estimated parameters, including unknown initial conditions. Parameters which were not estimated are documented in Table 3.2. AIC is computed as  $AIC = -2\ell + 2p$ .

| Model  | log likelihood ( $\ell$ ) | $p$ | AIC    |
|--|---------------------------|-----|--------|
| VSEIR without rainfall   | -1275.0                   | 19  | 2588.0 |
| VSEIR with rainfall  | -1265.0                   | 20  | 2570.0 |
| VS <sup>2</sup> EI <sup>2</sup> without rainfall                       | -1261.1                   | 24  | 2570.2 |
| VS <sup>2</sup> EI <sup>2</sup> with rainfall                          | -1251.0                   | 25  | 2552.0 |
| Log-SARIMA (1, 0, 1) $\times$ (1, 0, 1) <sub>12</sub> without rainfall | -1329.0                   | 6   | 2670.0 |
| Log-SARIMA (1, 0, 1) $\times$ (1, 0, 1) <sub>12</sub> with rainfall    | -1322.6                   | 7   | 2659.2 |

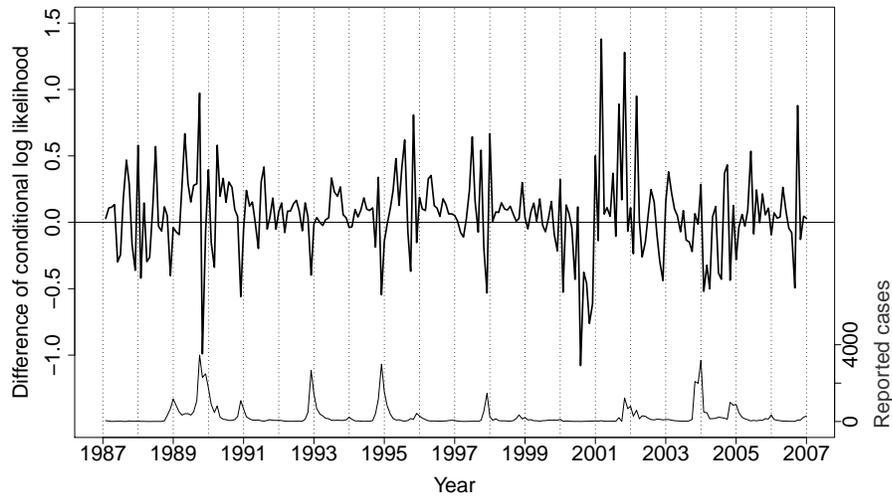


Figure 3.3: Difference between the conditional log likelihood of  $y_n$  given  $y_1, \dots, y_{n-1}$  for the VS<sup>2</sup>EI<sup>2</sup> model with rainfall and the VSEIR model with rainfall, plotted against time (bold line). For comparison, reported malaria cases in Kutch are also shown (thin line).

Table 3.2: List of symbols used in the chapter along with a brief description and units. Some parameters were not estimated as part of the analysis presented in this chapter, and the last column gives their fixed values. Alternative values of these fixed parameters were investigated, but did not affect the conclusions of the analysis. The values of the remaining parameters are give in Table 3.3.

| symbol                | brief description  | unit              | fixed value |
|-----------------------|--|-------------------|-------------|
| $\mu_{XY}$            | per-capita transition rate from X to Y; $X, Y \in \{S_1, S_2, E, I_1, I_2\}$ | yr <sup>-1</sup>  | -           |
| $[X]_0$               | initial fraction in compartment X; $X \in \{S_1, S_2, E, I_1, I_2\}$         | -                 | -           |
| $\kappa_0, \lambda_0$ | Initial values for the latent and current force of infection                 | -                 | -           |
| $\tau$                | mean development delay for mosquitoes  | yr                | -           |
| $\sigma$              | standard deviation of the process noise                                      | yr <sup>1/2</sup> | -           |
| $\rho$                | reporting fraction   | -                 | -           |
| $q$                   | relative infectivity of partially immune individuals                         | -                 | -           |
| $c$                   | coefficient of reinfection with clinical immunity                            | -                 | -           |
| $n_\lambda$           | shape parameter for the delay development kernel for mosquitoes              | -                 | 1           |
| $\psi$                | dispersion parameter of the observation noise                                | -                 | -           |
| $n_s$                 | number of splines describing seasonality                                     | -                 | 6           |
| $\beta_i$             | spline coefficients, for $i = 1, \dots, n_s$                                 | -                 | -           |
| $\bar{\beta}$         | dimensionality constant  | yr <sup>-1</sup>  | 1           |
| $\beta$               | coefficient of climate (rainfall) covariate                                  | -                 | -           |
| $u$                   | window for rainfall to affect transmission                                   | mo                | 5           |
| $v$                   | threshold for integrated rainfall  | mm                | 200         |
| $1/\delta$            | average life expectancy  | yr                | 50          |
| $\Delta$              | time step for stochastic Euler integration                                   | day               | 1           |

Table 3.3: Estimated model parameters. The columns marked ‘without rain’ correspond to maximum likelihood point estimates under the constraint  $\beta = 0$ . The last two columns give the lower and upper bounds for approximate 95% confidence intervals for the VS<sup>2</sup>EI<sup>2</sup> model with rainfall, derived from profile likelihood computations as shown in Figures 3.4 and 3.5; values of 0 and  $\infty$  correspond to confidence intervals extending to the boundary of the parameter space. Note: these models coincide with a subset of the models estimated by Laneri et al. (2010).

|                       | VSEIR<br>without rain | VS <sup>2</sup> EI <sup>2</sup><br>without rain | VSEIR<br>with rain | VS <sup>2</sup> EI <sup>2</sup><br>with rain | confidence interval  |
|-----------------------|-----------------------|---|--------------------|--|----------------------|
| $\mu_{I_1 S_2}$       | 13.587                | –   | 39.021             | –  | ( – , – )            |
| $\mu_{S_2 S_1}$       | 0.116                 | 0.230   | 5.657              | 0.334  | ( 0.067 , 3.270 )    |
| $\mu_{E I_1}$         | 7.301                 | 7.408   | 10.480             | 8.902  | ( 8.885 , 17.277 )   |
| $\mu_{I_1 I_2}$       | –                     | 11.544  | –                  | 5.511  | ( 3.218 , $\infty$ ) |
| $\mu_{I_2 S_2}$       | –                     | 0.004   | –                  | 0.035  | ( 0 , 0.073 )        |
| $\mu_{I_1 S_1}$       | –                     | 2.320   | –                  | 6.563  | ( 0 , $\infty$ )     |
| $\beta_1$             | -0.076                | -2.469  | 1.242              | 1.201  | ( -4.819 , 4.109 )   |
| $\beta_2$             | 1.287                 | 2.001   | 3.590              | 2.088  | ( -0.153 , 6.616 )   |
| $\beta_3$             | 4.446                 | 4.227   | 3.906              | 3.866  | ( 1.874 , 6.939 )    |
| $\beta_4$             | 2.868                 | 2.786   | 3.747              | 2.808  | ( 1.092 , 6.042 )    |
| $\beta_5$             | 6.709                 | 6.534   | 5.742              | 5.996  | ( 4.695 , 9.749 )    |
| $\beta_6$             | 6.319                 | 7.080   | 4.803              | 5.333  | ( 3.912 , 8.287 )    |
| $\tau$                | 0.025                 | 0.022   | 0.033              | 0.030  | ( 0.015 , 0.084 )    |
| $\sigma$              | 0.347                 | 0.309   | 0.225              | 0.243  | ( 0.162 , 0.259 )    |
| $\rho$                | 0.022                 | 0.030   | 0.005              | 0.015  | ( 0.007 , 0.025 )    |
| $q \times 10^4$       | –                     | 4.763   | –                  | 9.424  | ( 0.100 , 48.102 )   |
| $\psi$                | 0.384                 | 0.390   | 0.390              | 0.395  | ( 0.365 , 0.445 )    |
| $\beta$               | –                     | –   | 0.489              | 0.512  | ( 0.270 , 0.765 )    |
| $[S_1]_0$             | 0.494                 | 0.164   | 0.956              | 0.138  | ( 0.001 , 0.900 )    |
| $[S_2]_0$             | 0.505                 | 0.765   | 0.038              | 0.775  | ( 0.276 , 0.900 )    |
| $[E]_0$               | 0.003                 | 0.002   | 0.014              | 0.004  | ( 0.003 , 0.009 )    |
| $[I_1]_0$             | 0.011                 | 0.002   | 0.002              | 0.002  | ( 0 , 0.087 )        |
| $[I_2]_0$             | –                     | 0.067   | –                  | 0.080  | ( 0 , 0.754 )        |
| $\kappa_0 \times 10$  | 0.079                 | 0.133   | 0.189              | 0.171  | ( 0 , $\infty$ )     |
| $\lambda_0 \times 10$ | 0.050                 | 0.045   | 0.058              | 0.061  | ( 0 , $\infty$ )     |
| $c$                   | –                     | 0.004   | –                  | 0.010  | ( 0.001 , 0.067 )    |

When building mechanistic dynamic models for biological systems, there is a temptation to include as much biological detail as the available data will support. A price for this is that certain combinations of parameters may be weakly identified by the data. However, we can focus on conclusions which are robust to identifiability issues. For example, the model comparison via log likelihoods in Table 3.1 is valid despite any potential lack of identifiability. The profile likelihood for the reporting rate in Figure 3.4 shows that, without making any specific assumptions on the values of the 25 parameters estimated, there is evidence that the effective reporting rate is less than 2.5%. There is general agreement that malaria is substantially under-reported in South-East Asia (Snow et al., 2005) and a study in the city of Ahmedabad, Gujarat, found a reporting rate of 10% (Yadav et al., 2003). Much of the remaining discrepancy could be explained by a recent suggestion, based on a sensitive polymerase chain reaction diagnostic analysis in an epidemic malaria region of the East African highlands, that microscopy techniques may fail to detect two thirds of asymptomatic *Plasmodium* infections (Baliraine et al., 2009). There is potential for asymptomatic infections to play important dynamic roles, which can be hard to identify (King et al., 2008); for example, there could be an epidemiological role for boosted immunity due to mild infections that occur at blood parasite levels too low to be detected by standard field investigations. One cannot at this point rule out the possibility that the low estimated reporting rate could be an artifact due to unmodeled population inhomogeneity, or some other shortcoming of the model. Resolving such questions is beyond the scope of this chapter. The statistical interpretation, however, is more clearcut: Any attempt to learn about malaria via fitting epidemiological models of the type constructed here must take into account the discovery that unconventionally low reporting rates may give superior explanation of the data.

Many parameter estimates have large statistical uncertainty (Table 3.3, last

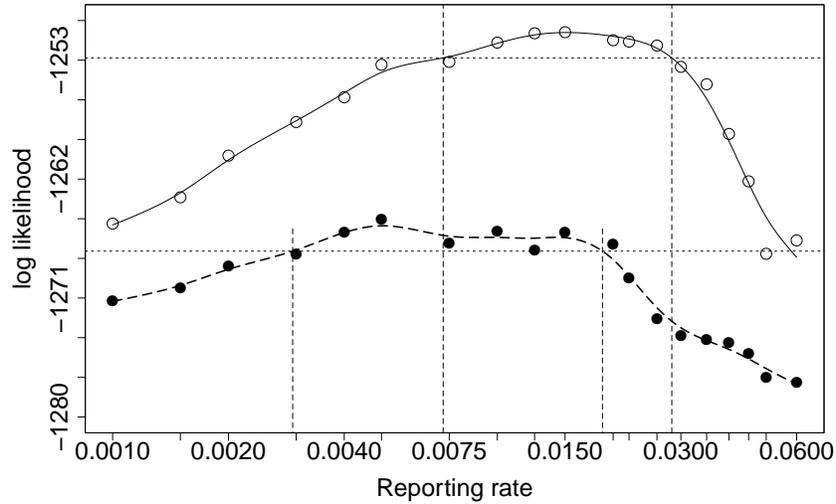


Figure 3.4: Profile likelihood plot for the reporting rate ( $\rho$ ) for the VS<sup>2</sup>EI<sup>2</sup> model with rainfall (solid line) and the VSEIR model with rainfall (broken line). The profile is estimated via fitting a smooth curve through Monte Carlo evaluations shown as open circles (VS<sup>2</sup>EI<sup>2</sup>) and filled circles (VSEIR). The dashed vertical lines construct approximate 95% confidence intervals (Barndorff-Nielsen and Cox, 1994).

two columns). One could investigate whether fixing some parameters at previously published scientific values helps to identify some other parameters. Conclusions from such an analysis should be made cautiously, since the variability and complexity of biological systems means that it is typically difficult to know to what extent previous investigations are indeed quantitatively relevant for the current model and data. This consideration would similarly complicate the development of a scientifically informed prior distribution, if one were to investigate a Bayesian approach.

Adding additional parameters to a model does not necessarily result in more weakly identified parameter estimates, particularly when the extended model provides substantial improvement in fit. Figure 3.5 provides one such example, where the *Plasmodium* development delay  $\tau$  is more precisely estimable in the larger VS<sup>2</sup>EI<sup>2</sup> model. Further, in the VS<sup>2</sup>EI<sup>2</sup> model the parameter values which are con-

sistent with the data are closer to the biological interpretation as a development delay—directly measured mean development times are around two weeks in this context. It could be nothing but a happy accident that, in this case, a model estimated on population data happens to match an individual-level biological interpretation. Given all the simplifications necessarily involved in the modeling process, it is hard to be sure that this parameter describes the biological interpretation in the strong sense that manipulation of the development time would affect the system only through the estimated value of  $\tau$ . Since development time is a well-studied function of temperature, in principle one could investigate this by seeing whether building this dependence into the model improves its explanation of the data. However, even without insisting on such a strong interpretation, when the data and the model and the desired biological interpretation are all mutually consistent then the model becomes validated as a conceptual tool for understanding the biological system.

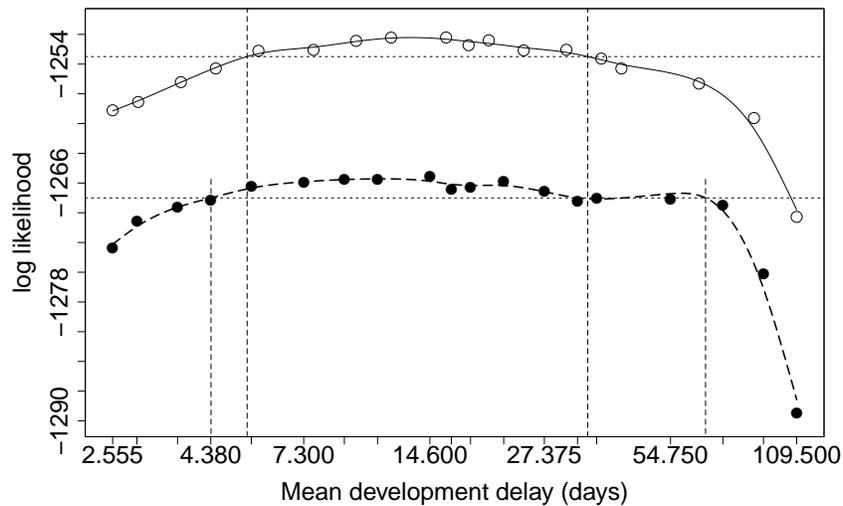


Figure 3.5: Profile likelihood plot for the mean development delay time of mosquitoes ( $\tau$ ) for the VS<sup>2</sup>EI<sup>2</sup> model with rainfall (solid line) and the VSEIR model with rainfall (broken line). The dashed vertical lines construct approximate 95% confidence intervals.

### 3.4 Discussion

One of the clearest scientific conclusions from our data analysis is that rainfall variability does indeed have a detectable effect on malaria dynamics in Kutch, even once one controls for seasonality and nonlinear dynamic effects of the force of infection and immunity. This is a contribution to the debate on the role of climate variability in malaria transmission, which has previously been lacking such an analysis (Hay et al., 2000, 2002; Zhou et al., 2004; Pascual et al., 2006, 2008; Briët et al., 2008; Clements et al., 2009). We have studied just one district here, in order to focus on the statistical principles behind our analysis. Investigation of another district in Northwest India leads to similar conclusions (Laneri et al., 2010). The statistical approach presented will facilitate similar investigations of other regions with endemic and epidemic malaria. Given geographical differences in mosquito species, social and agricultural practices, and many relevant ecological variables, one should however be cautious about extrapolating our quantitative results.

At the African summit on Roll Back Malaria in April 2000, forty four leaders of affected countries signed the Abuja declaration. One of the requirements of this declaration was that malaria epidemics should be detected, and effective control measures implemented, within two weeks. In practice, this timeline necessitates the use of malaria forecasts. Two major components of such a forecast should be measures of environmental suitability for transmission and the extent of residual immunity from previous epidemics. Seasonal rainfall forecasts (Bouma and van der Kaay, 1994, 2009) and satellite observations (Thomson et al., 2006) may have a role to play, though our results suggest that local rainfall acts at a sufficient lag to be a simple and useful predictor. Indeed, local rainfall was used as the primary component of epidemic malaria forecasts published for the semi-arid Punjab in the early part of the 20th century (Swaroop, 1949).

This chapter has adopted a likelihood-based, non-Bayesian inferential approach. Much recent work in the area of inference for POMP models has followed the Bayesian paradigm (e.g., Andrieu et al., 2010; Toni et al., 2009; Cauchemez and Ferguson, 2008; Boys et al., 2008). Our maximum likelihood methodology is a computationally viable alternative to these Bayesian approaches, in addition to being readily applicable due to the plug-and-play property. The analysis presented in this chapter is consistent with a recent study by Liu et al. (2009) which reported computational advantages for adopting a maximum likelihood approach over Bayesian methods for inference on complex phylogenetic models. Regardless of one's opinion on the epistemological value of asserting a prior distribution on unknown parameters, there may be computational advantages to exploring the likelihood surface rather than a posterior distribution.

Other vector-borne diseases, such as dengue and leishmaniasis, lead to statistical considerations and challenges similar to those for malaria. In a wider context, disease systems exemplify the issues at stake in developing an understanding of ecological processes from available time-series data (Bjørnstad and Grenfell, 2001). Quantitative understanding of ecosystems has growing importance as mankind is increasingly responsible for managing the biological resources of the planet. The broad scope of these responsibilities will continue to drive further developments in statistical methodology and data analysis.

## CHAPTER 4

# An adaptive particle allocation scheme for off-line sequential importance sampling algorithms

### 4.1 Introduction

In many engineering and scientific applications, state and parameter estimation in a hidden Markov model is of great importance. Particle filters, introduced by Gordon et al. (1993), have revolutionized the estimation problem. Prior to this, one had to resort to some variant of the Kalman filter, often requiring questionable linearization and assumptions of normality. Using particle filters, it is possible to carry out state and parameter estimation in a general class of state space models without these restricting assumptions. For a recent survey on the results on particle filters, see Cappé et al. (2007).

In this chapter, we focus on the problem of likelihood evaluation using particle filters. A serious challenge to particle filters in this context is the presence of outliers which results in high Monte Carlo variability of the estimated likelihood at the corresponding time points. Any implementation of a particle filter is limited by the computing power of the machine, which translates to a given total number of particles. Given this constraint, it is often useful to find a way of distributing

the particles across the time points in order to minimize a target criterion, such as the variance of the importance sampling estimate of the likelihood. Some previous non-rigorous result on the particle allocation problem for the online estimation case can be found in engineering literature, e.g. Fox (2003). In this chapter, we aim to give a more rigorous treatment to this problem for the off-line case. For many Sequential Monte Carlo (SMC) based algorithms, such as iterated filtering as described in chapter 2, the focus is on an off-line estimation of states, parameters or likelihood. An effective particle allocation technique should result in improved performance in these cases.

## 4.2 Problem statement and background

We specifically focus on the off-line estimation problem of the likelihood. With no prior information about the likelihood surface, the naive strategy is to use an equal number of particles for each time point. However, this naive initial run generates some informative diagnostics. One of them, the effective sample size (Liu, 2001) has proved to be a powerful, though somewhat informal indicator of the variance of the conditional likelihood at each time point. A low effective sample size is an indication of high variance in the estimated likelihood. In an off-line setting, this information can be used in the later iterations to determine which problems were problematic in the likelihood evaluation and allocate number of particles accordingly.

In sections 4.3 and 4.4 we deal with the particle allocation problem when the observations are independent and dependent respectively. The independent case provides an exact solution but the expression for SMC likelihood for dependent observations is complicated. We, therefore, model the conditional log-likelihood with an autoregressive structure in this case and solve the allocation problem for that model, rather than tackling the likelihood expression directly. This can be

heuristically argued by an exponential mixing of the likelihood (Crisan and Doucet, 2002). We now state a useful theorem and a lemma.

**Theorem 4.2.1.** (Oehlert, 1992) Suppose  $X_1, \dots, X_n$  are iid with  $k$  finite moments where  $k \geq 3$ . Let  $\bar{X}_n$  denote the sample mean and the mean and variance of  $X_1$  be denoted by  $\mu$  and  $\sigma^2$ . Suppose there is a function  $g$  that has  $k$  bounded derivatives. Then

$$\mathbb{E}[g(\bar{X}_n)] = g(\mu) + g^{(2)}(\mu)\sigma^2/(2n) + O(n^{-2})$$

**Lemma 4.2.1.** Consider the variables  $m_1, \dots, m_n$  so that  $m_1 + \dots + m_n = N$ . Let  $k_1, \dots, k_n$  be known constants. Then the values of  $m_i$  that minimize  $\sum_{i=1}^n k_i/m_i$  are given by,

$$m_i = \frac{N\sqrt{k_i}}{\sum_{i=1}^n \sqrt{k_i}} \quad i = 1, \dots, n$$

*Proof.* The Lagrangian function is given by,

$$f = \sum_{i=1}^n k_i/m_i + \lambda \left( \sum_{i=1}^n m_i - N \right)$$

Computing  $\partial f/\partial m_i$  and equating to 0 we get,

$$\begin{aligned} \frac{k_i}{m_i^2} - \lambda &= 0 \\ \text{i.e., } m_i &= \sqrt{\frac{k_i}{\lambda}} \end{aligned} \tag{4.1}$$

Using the constraint  $m_1 + \dots + m_n = N$ , gives the value of  $\lambda$  as,

$$\lambda = \frac{(\sum_{i=1}^n \sqrt{k_i})^2}{N^2}$$

Using this in equation (4.1) completes the proof.  $\square$

### 4.3 Applications to likelihood estimation (independent case)

The problem of optimal particle allocation in order to minimize overall variance for likelihood estimation takes a simple form for time independent observations. In this case since the total variance is the sum of point-wise variances, we can just minimize the variance at a given time point, without having to worry about the dependence structure in the data. The result is given in the following theorem.

**Theorem 4.3.1.** *Suppose  $X_1, \dots, X_K$  are independent random variables with densities given by  $f_k$ , i.e.,  $X_k \sim f_k(\cdot; \theta)$  where  $\theta$  is a vector of parameters. Let the observations be denoted as  $y_1, \dots, y_K$  and conditional on  $X_k = x_k$ ,  $Y_k \sim h_k(\cdot | x_k; \theta)$  be the observation density. Also assume we can draw samples  $Z_{k,m}$  from  $Z_{k,m} \sim f_k(\cdot; \theta)$  for  $m = 1, \dots, M_k$  and  $\sum_{k=1}^K M_k = M$ . The likelihood functions  $l_k(\theta)$  is given by  $\int h_k(y_k | x; \theta) f_k(x; \theta) dx$  and let  $\hat{l}_k$  be the importance sampling estimate of  $l_k(\theta)$ , given by  $\hat{l}_k(\theta) = \frac{1}{M_k} \sum_{m=1}^{M_k} h_k(y_k | Z_{k,m}; \theta)$ . Assume that  $C_1 < h_k(\cdot | \cdot; \theta) < \infty$  for all  $k$  and some  $C_1 > 0$ . Then, there exists a constant  $M_0$  such that if  $M_k > M_0$  for all  $k$ , the value of  $M_k$  minimizing  $\text{Var}(\sum_{k=1}^K (\log \hat{l}_k))$  is given by,*

$$M_k^{\min} = \frac{M \sqrt{\phi_k}}{\sum_{i=1}^K \sqrt{\phi_i}}$$

where,  $\phi_k = \text{Var}_{f_k}(h_k(y_k | X_k; \theta)) / l_k^2$  is bounded and  $M$  is chosen large enough so that the condition  $M_k^{\min} > M_0$  is satisfied for all  $k$ .

*Proof.* We have,

$$\begin{aligned} l_k(\theta) &= \int h_k(y_k | x; \theta) f_k(x; \theta) dx \\ &= \text{E}_{f_k} [h_k(y_k | X; \theta)] \end{aligned}$$

where,  $X \sim f_k(\cdot; \theta)$ . Now, the importance sampling estimate of the likelihood is,

$$\hat{l}_k(\theta) = \frac{1}{M_k} \sum_{m=1}^{M_k} h_k(y_k | Z_{k,m}; \theta)$$

where  $Z_{k,m} \sim f_k(\cdot; \theta)$ ,  $E_{f_k}[h_k(y_k | Z_{k,m}; \theta)] = l_k$  and  $\text{Var}_{f_k}[h_k(y_k | Z_{k,m}; \theta)] = \phi_k l_k^2$  (from the definition of  $\phi_k$ ). We now apply Theorem 4.2.1 to the functions  $\log(\hat{l}_k)$  and  $(\log(\hat{l}_k))^2$  respectively. It is easy to see the conditions for Theorem 4.2.1 are satisfied since  $h_k(\cdot, \cdot, \theta)$  is bounded away from 0. This also means  $l_k$  and  $\phi_k$  are bounded away from 0 for all  $k$  and the boundedness condition for the derivatives are satisfied. Thus, we have

$$\begin{aligned} E_{f_k}[\log \hat{l}_k] &= \log l_k - \left(\frac{1}{l_k^2}\right) \frac{\phi_k l_k^2}{2M_k} + O(M_k^{-2}) \\ &= \log l_k - \frac{\phi_k}{2M_k} + O(M_k^{-2}) \end{aligned} \quad (4.2)$$

and

$$\begin{aligned} E_{f_k}[(\log \hat{l}_k)^2] &= (\log l_k)^2 + \left(\frac{2(1 - \log l_k)}{l_k^2}\right) \frac{\phi_k l_k^2}{2M_k} + O(M_k^{-2}) \\ &= (\log l_k)^2 + \frac{\phi_k}{M_k} - \frac{\phi_k \log l_k}{M_k} + O(M_k^{-2}) \end{aligned} \quad (4.3)$$

Then combining equations (4.2) and (4.3) we have

$$\begin{aligned} \text{Var}_{f_k}[\log \hat{l}_k] &= E_{f_k}[(\log \hat{l}_k)^2] - E_{f_k}^2[\log \hat{l}_k] \\ &= \frac{\phi_k}{M_k} + O(M_k^{-2}) \end{aligned}$$

Thus, for each  $k$ , there exists an  $M_{k_0}$  and a constant  $C'_k$  such that

$$\frac{\phi_k}{M_k} - \frac{C'_k}{M_k^2} \leq \text{Var}_{f_k}[\log \hat{l}_k] \leq \frac{\phi_k}{M_k} + \frac{C'_k}{M_k^2} \quad \text{for all } M_k > M_{k_0}, 0 < C'_k < \infty$$

$$\text{i.e.,} \quad \frac{\phi_k}{M_k}(1 - \epsilon_k) \leq \text{Var}_{f_k}[\log \hat{l}_k] \leq \frac{\phi_k}{M_k}(1 + \epsilon_k) \quad \text{for } \epsilon_k = C'_k/M_k\phi_k$$

Thus, we have,  $(1 - \bar{\epsilon}) \sum_{k=1}^K \phi_k/M_k \leq \sum_{k=1}^K \text{Var}_{f_k}(\log \hat{l}_k) \leq (1 + \bar{\epsilon}) \sum_{k=1}^K \phi_k/M_k$ , where  $\bar{\epsilon} > \max_k \epsilon_k = \max_k C'_k/M_k\phi_k$ . Such an  $\bar{\epsilon}$ , independent of  $k$ , will exist and can be made arbitrarily small (i.e., smaller than any given  $\delta$ ) if all  $M_k$  are chosen greater than a sufficiently large  $M_0$  which is also greater than each  $M_{k_0}$ . Now the problem is to minimize  $\text{Var}(\sum_{k=1}^K (\log \hat{l}_k)) = \sum_{k=1}^K \text{Var}_{f_k}(\log \hat{l}_k)$  (using independence), subject to  $\sum_{k=1}^K M_k = M$ . Since  $\bar{\epsilon}$  is arbitrarily small, we can shrink the interval containing  $\sum_{k=1}^K \text{Var}_{f_k}(\log \hat{l}_k)$  until it collapses to  $\sum_{k=1}^K \phi_k/M_k$  from both sides and using Lemma 4.2.1 we have,

$$M_k^{\min} = \frac{M\sqrt{\phi_k}}{\sum_{i=1}^K \sqrt{\phi_i}}$$

as the optimal value of  $M_k$  when  $M$  is chosen large enough so that the condition  $M_k^{\min} > M_0$  is satisfied for all  $k$ . Such an  $M$  can be chosen since  $\phi_k$  is bounded for all  $k$ .  $\square$

#### 4.4 Applications to likelihood estimation (dependent case)

For the dependent case, consider the time series of observations  $y_{1:K} = (y_1, \dots, y_K)$ . The data are modeled as a realization of some sequence of real valued random variables  $Y_{1:K}$ , taking value in  $\mathbb{R}^{d_y \times K}$ . The conditional density of  $Y_k$  given  $Y_{1:k-1}$  is assumed to exist for all  $k = 1, \dots, K$ . Now, let,  $g_k = \log f_{Y_k|Y_{1:k-1}}(y_k|y_{1:k-1}; \theta)$  and  $\hat{g}_k = \log \hat{f}_{Y_k|Y_{1:k-1}}(y_k|y_{1:k-1}; \theta)$ , where the subscripts denote the random variables to which the densities correspond and  $\hat{f}_{Y_k|Y_{1:k-1}}(y_k|y_{1:k-1}; \theta)$  denotes the sequential

Monte Carlo estimate of  $f_{Y_k|Y_{1:k-1}}(y_k|y_{1:k-1}; \theta)$ . It is known  $\hat{f}_{Y_k|Y_{1:k-1}}(y_k|y_{1:k-1}; \theta)$  is an unbiased estimator of  $f_{Y_k|Y_{1:k-1}}(y_k|y_{1:k-1}; \theta)$  (Del Moral and Jacod, 2001). We call  $g_k$  the conditional log likelihood at time  $k$  and note that  $\sum_{k=1}^K g_k = \log f_{Y_{1:K}}(y_{1:K}; \theta)$ , gives the total log likelihood. We use the SMC evaluations  $\hat{g}_k$  to fit the following AR(1) model

$$x_k = \mu_k + q(x_{k-1} - \mu_{k-1}) + \epsilon_k \quad (4.4)$$

where  $\mu_k = \mathbb{E}[x_k]$ ,  $\epsilon_k \sim N(0, \phi_k/M_k)$ ,  $\phi_k$  is positive and  $M_k$  is the number of particles used at time  $k$ . An AR(1) model for the conditional log likelihoods in SMC is suggested by exponential mixing of the SMC likelihood. We then proceed to solve the particle allocation problem for this simple AR(1) model. We have

$$\begin{aligned} \text{Var}(x_k) &= q^2 \text{Var}(x_{k-1}) + \frac{\phi_k}{M_k} \\ &= q^2 \text{Var}(qx_{k-2} + \epsilon_{k-1}) + \frac{\phi_k}{M_k} \\ &= q^4 \text{Var}(x_{k-2}) + q^2 \frac{\phi_{k-1}}{M_{k-1}} + \frac{\phi_k}{M_k} \\ &= \dots \\ &= \sum_{m=1}^k \frac{\phi_m q^{2(k-m)}}{M_m} \end{aligned}$$

and for  $i > k$

$$\begin{aligned}\text{Cov}(x_i, x_k) &= \text{Cov}(\mu_i + q(x_{i-1} - \mu_{i-1}) + \epsilon_i, x_k) \\ &= q\text{Cov}(x_{i-1}, x_k) \\ &= \dots \\ &= q^{i-k}\text{Cov}(x_k, x_k) \\ &= q^{i-k}\text{Var}(x_k) \\ &= q^{i-k} \sum_{m=1}^k \frac{\phi_m q^{2(k-m)}}{M_m}\end{aligned}$$

Thus,

$$\begin{aligned}
\text{Var}\left(\sum_{k=1}^K x_k\right) &= \sum_{k=1}^K \text{Var}(x_k) + 2 \sum_{i=1}^K \sum_{j=1, j < i}^K \text{Cov}(x_i, x_j) \\
&= \sum_{k=1}^K \text{Var}(x_k) + 2 \sum_{i=2}^K \sum_{j=1}^{i-1} \text{Cov}(x_i, x_j) \\
&= \sum_{k=1}^K \sum_{m=1}^k \frac{\phi_m q^{2(k-m)}}{M_m} + 2 \sum_{i=2}^K \sum_{j=1}^{i-1} q^{i-j} \sum_{m=1}^j \frac{\phi_m q^{2(j-m)}}{M_m} \\
&= \sum_{m=1}^K \sum_{k=m}^K \frac{\phi_m q^{2(k-m)}}{M_m} + 2 \sum_{i=2}^K \sum_{j=1}^{i-1} \sum_{m=1}^j \frac{\phi_m q^{(i+j-2m)}}{M_m} \\
&= \sum_{m=1}^K \frac{\phi_m q^{-2m}}{M_m} \sum_{k=m}^K q^{2k} + 2 \sum_{m=1}^K \frac{\phi_m q^{-2m}}{M_m} \sum_{j=1}^{K-1} q^j \sum_{i=j+1}^K q^i \\
&= \sum_{m=1}^K \frac{\phi_m q^{-2m}}{M_m} \cdot q^{2m} \cdot \frac{1 - q^{2(K-m+1)}}{1 - q^2} + 2 \sum_{m=1}^K \frac{\phi_m q^{-2m}}{M_m} \sum_{j=1}^{K-1} q^j \cdot q^{j+1} \frac{1 - q^{(K-j)}}{1 - q} \\
&= \frac{1}{1 - q^2} \sum_{m=1}^K \frac{\phi_m (1 - q^{2(K-m+1)})}{M_m} + \frac{2q}{1 - q} \sum_{m=1}^K \frac{\phi_m q^{-2m}}{M_m} \sum_{j=1}^{K-1} (q^{2j} - q^{(K+j)}) \\
&= \frac{1}{1 - q^2} \sum_{m=1}^K \frac{\phi_m (1 - q^{2(K-m+1)})}{M_m} \\
&\quad + \frac{2q}{1 - q} \sum_{m=1}^K \frac{\phi_m q^{-2m}}{M_m} \left( \frac{1 - q^{2(K-1)}}{1 - q^2} - q^K \cdot \frac{1 - q^{(K-1)}}{1 - q} \right) \\
&= C_1 \sum_{m=1}^K \frac{A_m}{M_m} + C_2 \sum_{m=1}^K \frac{B_m}{M_m} \\
&= \sum_{m=1}^K \frac{Z_m}{M_m}
\end{aligned}$$

where

$$C_1 = \frac{1}{1 - q^2} \quad (4.5)$$

$$C_2 = \frac{2q}{1 - q} \left( \frac{1 - q^{2(K-1)}}{1 - q^2} - q^K \cdot \frac{1 - q^{(K-1)}}{1 - q} \right) \quad (4.6)$$

$$A_m = \phi_m (1 - q^{2(K-m+1)}) \quad (4.7)$$

$$B_m = \phi_m q^{-2m} \quad (4.8)$$

$$Z_m := C_1 A_m + C_2 B_m \quad (4.9)$$

Thus, the variance minimization problem is now equivalent to minimizing  $\sum_{k=1}^K Z_k / M_k$  with the constraint  $\sum_{k=1}^K M_k = M$ . Using Lemma 4.2.1 we have,

$$M_k = \frac{M \sqrt{Z_k}}{\sum_{i=1}^K \sqrt{Z_i}}$$

as the optimal value of  $M_k$ . As a consistency check, note how the results presented in this section reduce to the results presented for the independent case (section 4.3) when  $q = 0$ .

#### 4.4.1 Joint estimation of $q$ and $\phi_k$

In order to implement the particle allocation scheme described in section 4.4, one would need to fit the model described in equation (4.4). As described before, without any prior knowledge of the likelihood surface, initially one just uses an equal allocation of particles for all the time points, i.e. an ordinary particle filter, say  $P$  times. This gives rise to estimates of  $\text{Var}(x_k)$ . Call these estimates  $\widetilde{\text{Var}}(x_k)$ . It is given by

$$\widetilde{\text{Var}}(x_k) = \frac{1}{P-1} \sum_{p=1}^P (\hat{g}_{k,p} - \bar{\hat{g}}_k)^2$$

where  $\hat{g}_{k,p}$  is the estimate of the conditional log likelihood at time point  $k$  for the  $p$ 'th filtering operation and  $\overline{\hat{g}_k} = \frac{1}{P} \sum_{p=1}^P \hat{g}_{k,p}$ . Now, note from equation (4.4)

$$\text{Var}(x_k) = q^2 \text{Var}(x_{k-1}) + \frac{\phi_k}{M_k}$$

Thus, in equation (4.7) and (4.8) we use the estimate  $\hat{\phi}_k$  of  $\phi_k$  defined as

$$\hat{\phi}_k = M_k \left( \widetilde{\text{Var}}(x_k) - q^2 \widetilde{\text{Var}}(x_{k-1}) \right)$$

Then  $Z_m$  in equation (4.9) becomes a function of just  $q$  whose estimate  $\hat{q}$  is then calculated as a number between -1 and 1 that minimizes the sum  $\sum_{m=1}^K Z_m / M_m$ .

## 4.5 Simulation Study

A simulation study was performed to compare the performance of the adaptive and non-adaptive filtering procedures. We work with a two dimensional AR(1) process observed with noise defined by the following state and observation equations,

$$X_k = \alpha X_{k-1} + \sigma \xi_k \quad (\text{state equation})$$

$$Y_k = \beta X_k + \tau \epsilon_k \quad (\text{observation equation})$$

Here,  $X_k = (X_k^1, X_k^2)$  and  $Y_k = (Y_k^1, Y_k^2)$  are in  $\mathbb{R}^2$  for all  $k = 1, \dots, K$ ;  $\alpha$  and  $\beta$  are  $2 \times 2$  constant matrices,  $\xi_k$  and  $\epsilon_k$  are independent bivariate standard normal random variables,  $\sigma$  is a lower-triangular  $2 \times 2$  matrix and  $\tau$  is a  $1 \times 2$  matrix with

both entries equal. For our example, we use

$$\begin{aligned}\alpha &= \begin{pmatrix} 0.9 & 0.0 \\ 0.0 & 0.99 \end{pmatrix} \\ \beta &= \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \\ \sigma &= \begin{pmatrix} 1.00 & 0 \\ 0 & 2.00 \end{pmatrix} \\ \tau &= \begin{pmatrix} 1 & 1 \end{pmatrix}\end{aligned}$$

These parameter values were chosen to be the same as in the vignette for King et al. (2009). One simulation of the state and observation models for  $k = 1, \dots, 100$  time points is shown in Figure 4.1(left panel). We then use 1000 simulations from the process model to compute the standard deviation at each time point. Outliers are then introduced at 4 randomly chosen points in  $Y^1$  as

$$Y_{k'}^{1*} = Y_{k'}^1 + 3 * Z * \text{sd}(Y_{k'}^1)$$

where the random variable  $Z \in \{-1, 1\}$  with probability 0.5 each and  $k' \in C$ , where  $C$  is a set of 4 time points randomly chosen between 1 and 100. Figure 4.1(right panel) shows  $Y^1$  before (top) and after (bottom) introducing the outliers. Thus, after introducing these so called outliers the hope is that an adaptive particle filter will perform better than the naive particle filter, given same amount of computer resource is available to both the adaptive and non-adaptive schemes.

Since this is clearly a dependent setting, we choose the particle allocation scheme detailed in section 4.4. Table 4.1 shows the result of applying the ordinary and adaptive filters on the data set in order to estimate the overall likelihood.

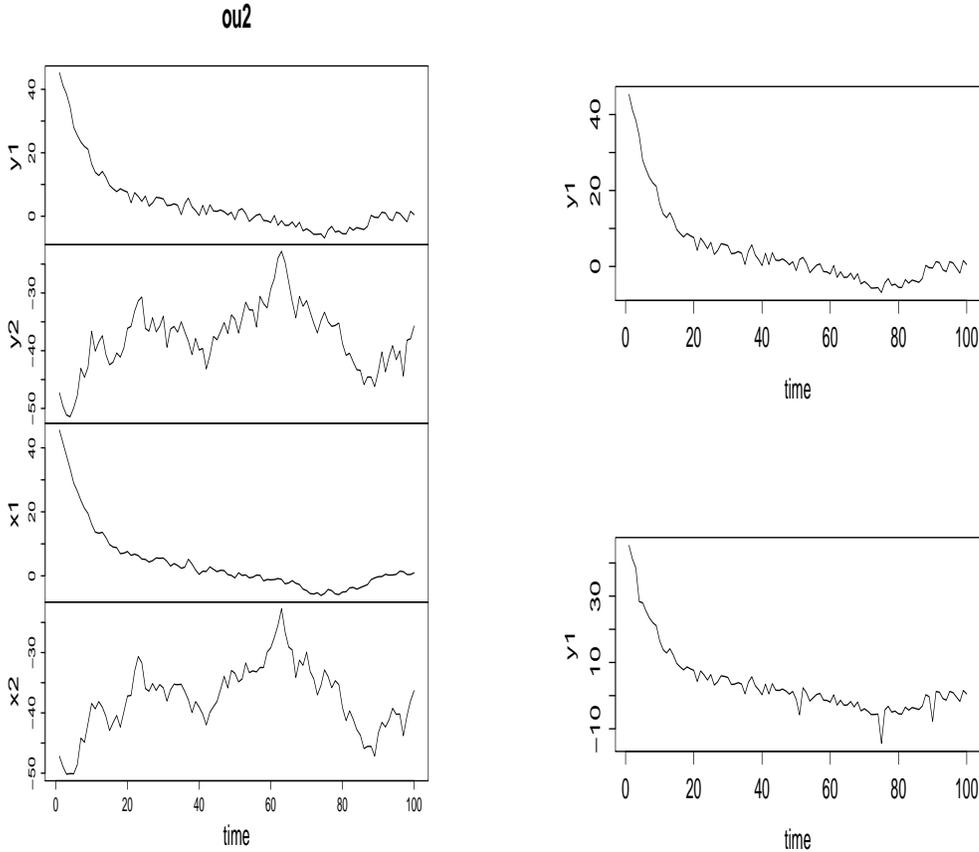


Figure 4.1: Plot of the two dimensional AR(1) process. Left panel shows the state  $(X^1, X^2)$  and observation  $(Y^1, Y^2)$  vectors. Outliers are introduced only in  $Y^1$  at time points 4, 51, 75 and 90. The vector  $Y^1$  is shown in the right panel, upper part is before introducing outliers (same as that of left panel) and lower part is after introducing outliers.

Here,  $N$  denotes the number of times a set of 4 outliers are introduced at randomly chosen points between 1 and 100 to generate a test set.  $M$  denotes the number of times the filters are applied on each such test set. The total time taken  $T$  is then the time for a total  $N \times M$  filtering operations. For each  $n = 1, \dots, N$  we then

compute the estimate of the variances of the overall log-likelihood as

$$\begin{aligned}\hat{V}_n^a &= \frac{1}{M-1} \sum_{m=1}^M (\hat{G}_{n,m}^a - \overline{\hat{G}_n^a})^2 \\ \hat{V}_n^o &= \frac{1}{M-1} \sum_{m=1}^M (\hat{G}_{n,m}^o - \overline{\hat{G}_n^o})^2\end{aligned}$$

where

$$\hat{G}_{n,m} = \sum_{k=1}^K \hat{g}_{k,n,m}$$

Here  $\hat{g}_{k,n,m}$  is the estimate of the conditional log-likelihood (as defined in section 4.4) at time  $k$  evaluated with the particle filter for the  $n$  th series and  $m$  th filtering iteration, the superscripts  $a$  and  $o$  denote the adaptive and ordinary (i.e. non-adaptive) particle filters respectively.  $\hat{G}_{n,m}$  is thus an estimate of the total log-likelihood and  $\overline{\hat{G}_n} = \frac{1}{M} \sum_{m=1}^M \hat{G}_{n,m}$ . The estimates of the average variance ( $\hat{V}$ ) of the independently generated test sets as shown in table 4.1 are then computed as  $\hat{V} = \frac{1}{N} \sum_{n=1}^N \hat{V}_n$ .

The gain or improvement  $\alpha$  in using the adaptive over ordinary particle filter is calculated as

$$\alpha = 1 - \frac{1}{N} \sum_{n=1}^N \frac{\hat{V}_n^a}{\hat{V}_n^o} \quad (4.10)$$

With approximately the same amount of time ( $T$ ) as shown in table 4.1,  $G$  was found to be 45.862%. A 1 sided t-test performed to test the hypothesis  $H_0 : \mu_{V^a/V^o} = 1$  and  $H_1 : \mu_{V^a/V^o} < 1$  had a p-value of  $2.379 \times 10^{-5}$ , indicating a statistically significant improvement achieved with the adaptive scheme.

|  | Adaptive filter (a) | Ordinary filter (o) |
|--|---------------------|---------------------|
| Number of tests ( $N$ )                                | 10                  | 10                  |
| Number of filtering operations per test ( $M$ )        | 100                 | 100                 |
| Number of total particles used per filtering operation | 15000               | 15000               |
| Total time taken(sec) ( $T$ )                          | 2222                | 2149                |
| Average variance ( $\hat{V}$ )                         | 0.995               | 1.854               |

Table 4.1: Comparison of adaptive and non-adaptive particle filters. The gain  $\alpha$  as defined in equation (4.10) was found to be 45.862%.

## BIBLIOGRAPHY

- Anderson, B. D. and Moore, J. B. (1979). *Optimal Filtering*. Prentice-Hall, New Jersey.
- Anderson, J. L. and Collins, N. (2007). Scalable implementations of ensemble filter algorithms for data assimilation. *Journal of Atmospheric and Oceanic Technology*, 24:1452–1463.
- Anderson, R. M. and May, R. M. (1991). *Infectious Diseases of Humans*. Oxford University Press, Oxford.
- Andrieu, C., Doucet, A., and Holenstein, R. (2010). Particle Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society, Series B (Statistical Methodology)*, 72:269–342.
- Andrieu, C., Éric Moulines, and Priouret, P. (2005). Stability of stochastic approximation under verifiable conditions. *SIAM Journal on Control and Optimization*, 44(1):283–312.
- Anisimova, M., Bielawski, J. P., and Yang, Z. (2001). Accuracy and power of the likelihood ratio test in detecting adaptive molecular evolution. *Molecular Biology and Evolution*, 18:1585–1592.
- Applebaum, D. (2004). Lévy processes: From probability to finance and quantum groups. *Notices of the American Mathematical Society*, 51:1336–1347.
- Aron, J. L. and May, R. M. (1982). The population dynamics of malaria. In Anderson, R. M., editor, *The Population Dynamics of Infectious Diseases*, pages 139–179. Chapman and Hall, London, UK.
- Artavanis-Tsakonas, K., Tongren, J. E., and Riley, E. M. (2003). The war between the malaria parasite and the immune system: Immunity, immunoregulation and immunopathology. *Clinical and Experimental Immunology*, 133:145–152.
- Arulampalam, M. S., Maskell, S., Gordon, N., and Clapp, T. (2002). A tutorial on particle filters for online nonlinear, non-Gaussian Bayesian tracking. *IEEE Transactions on Signal Processing*, 50:174 – 188.

- Bailey, N. T. J. (1955). Some problems in the statistical analysis of epidemic data. *Journal of the Royal Statistical Society, Series B*, 17:35–68.
- Baliraine, F., Afrane, Y., Ameyna, D., Bonizzoni, M., Menge, D., Zhou, G., Zhong, D., Vardo-Zalik, A., Githeko, A., and Yan, G. (2009). High prevalence of asymptomatic *plasmodium falciparum* infections in a highland area of Western Kenya: A cohort study. *The Journal of Infectious Diseases*, 200:66–74.
- Barndorff-Nielsen, O. E. and Cox, D. R. (1994). *Inference and Asymptotics*. Chapman and Hall, London.
- Barndorff-Nielsen, O. E. and Shephard, N. (2001). Non-gaussian ornstein-uhlenbeck-based models and some of their uses in financial economics. *Journal of the Royal Statistical Society, Series B (Statistical Methodology)*, 63(2):167–241.
- Bartlett, M. S. (1960). *Stochastic Population Models in Ecology and Epidemiology*. Wiley, New York.
- Basawa, I. V. and Prakasa Rao, B. L. S. (1980). *Statistical inference for stochastic processes*. Academic Press, London.
- Beskos, A., Papaspiliopoulos, O., Roberts, G., and Fearnhead, P. (2006). Exact and computationally efficient likelihood-based inference for discretely observed diffusion processes. *Journal of the Royal Statistical Society, Series B (Statistical Methodology)*, 68:333–382.
- Bhadra, A. (2010). Discussion of Particle Markov chain Monte Carlo methods by Andrieu et al. *Journal of the Royal Statistical Society, Series B (Statistical Methodology)*, 72:314–315.
- Bjørnstad, O. N. and Grenfell, B. T. (2001). Noisy clockwork: Time series analysis of population fluctuations in animals. *Science*, 293:638–643.
- Bouma, M. J., Dye, C., and van der Kaay, H. J. (1996). Falciparum malaria and climate change in the Northwest Frontier Province of Pakistan. *American Journal of Tropical Medicine and Hygiene*, 55(2):131–137.
- Bouma, M. J. and van der Kaay, H. J. (1994). Epidemic malaria in India and the El Niño Southern Oscillation. *The Lancet*, 344(8937):1638 – 1639.
- Bouma, M. J. and van der Kaay, H. J. (2009). The El Niño Southern Oscillation and the historic malaria epidemics on the Indian subcontinent and Sri Lanka: An early warning system for future epidemics? *Tropical Medicine and International Health*, 1:86–96.
- Boys, R. J., Wilkinson, D. J., and Kirkwood, T. B. L. (2008). Bayesian inference for a discretely observed stochastic kinetic model. *Statistica Sinica*, 18:125–135.

- 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.
- Briët, O., Vounatsou, P., Gunawardena, D., Galappaththy, G., and Amerasinghe, P. (2008). Temporal correlation between malaria and rainfall in Sri Lanka. *Malaria Journal*, 7(1):77.
- Cappé, O., Godsill, S., and Moulines, E. (2007). An overview of existing methods and recent advances in sequential Monte Carlo. *Proceedings of the IEEE*, 95(5):899–924.
- Cappé, O., Moulines, E., and Rydén, T. (2005). *Inference in Hidden Markov Models*. Springer, New York.
- Cauchemez, S. and Ferguson, N. M. (2008). Likelihood-based estimation of continuous-time epidemic models from time-series data: Application to measles transmission in London. *Journal of the Royal Society Interface*, 5:885–897.
- Cauchemez, S., Valleron, A., Boëlle, P., Flahault, A., and Ferguson, N. M. (2008). Estimating the impact of school closure on influenza transmission from sentinel data. *Nature*, 452:750–754.
- Chitnis, N., Cushing, J. M., and Hyman, J. M. (2006). Bifurcation analysis of a mathematical model for malaria transmission. *SIAM Journal on Applied Mathematics*, 67(1):24–45.
- Clements, A., Barnett, A., Cheng, Z., Snow, R., and Zhou, H. (2009). Space-time variation of malaria incidence in Yunnan province, China. *Malaria Journal*, 8(1):180.
- Crisan, D. and Doucet, A. (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*. Springer, New York.
- Del Moral, P. and Jacod, J. (2001). Interacting particle filtering with discrete observations. In Doucet, A., de Freitas, N., and Gordon, N. J., editors, *Sequential Monte Carlo Methods in Practice*, pages 43–75. Springer, New York.
- Dietz, K., Molineaux, L., and Thomas, A. (1974). A malaria model tested in the African savannah. *Bulletin of the World Health Organization*, 50:347–357.
- Doucet, A., de Freitas, N., and Gordon, N. J., editors (2001). *Sequential Monte Carlo Methods in Practice*. Springer, New York.
- Durbin, J. and Koopman, S. J. (2001). *Time Series Analysis by State Space Methods*. Oxford University Press, Oxford.

- Earn, D. J. D., Rohani, P., Bolker, B. M., and Grenfell, B. T. (2000). A simple model for complex dynamical transitions in epidemics. *Science*, 287:667–670.
- Ellner, S. P., Bailey, B. A., Bobashev, G. V., Gallant, A. R., Grenfell, B. T., and Nychka, D. W. (1998). Noise and nonlinearity in measles epidemics: Combining mechanistic and statistical approaches to population modeling. *American Naturalist*, 151:425–440.
- Ellner, S. P., Seifu, Y., and Smith, R. H. (2002). Fitting population dynamic models to time-series data by gradient matching. *Ecology*, 83(8):2256–2270.
- Ergun, A., Barbieri, R., Eden, U., Wilson, M., and Brown, E. (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 Rubio-Ramírez, J. F. (2007). Estimating macroeconomic models: A likelihood approach. *Review of Economic Studies*, 74:1059–1087.
- Ferrari, M. J., Grais, R. F., Bharti, N., Conlan, A. J. K., Bjornstad, O. N., Wolfson, L. J., Guerin, P. J., Djibo, A., and Grenfell, B. T. (2008). The dynamics of measles in sub-Saharan Africa. *Nature*, 451(7179):679–684.
- Filipe, J. A. N., Riley, E. M., Drakeley, C. J., Sutherland, C. J., and Ghani, A. C. (2007). Determination of the processes driving the acquisition of immunity to malaria using a mathematical transmission model. *PLoS Computational Biology*, 3(12):e255.
- Finkenstädt, B. F. and Grenfell, B. T. (2000). Time series modelling of childhood diseases: A dynamical systems approach. *Applied Statistics*, 49:187–205.
- Fox, D. (2003). Adapting the sample size in particle filters through KLD-sampling. *International Journal of Robotics Research*, 22:985–1003.
- Gilks, W. R. and Berzuini, C. (2001). Following a moving target-monte carlo inference for dynamic bayesian models. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 63(1):127–146.
- Godsill, S., Vermaak, J., Ng, W., and Li, J. (2007). Models and algorithms for tracking of maneuvering objects using variable rate particle filters. *Proceedings of the IEEE*, 95(5):925–952.
- Gordon, N. J., Salmond, D. J., and Smith, A. F. M. (1993). Novel approach to nonlinear/non-gaussian bayesian state estimation. *Radar and Signal Processing, IEE Proceedings F*, 140(2):107–113.
- Greenwood, B. (2009). Can malaria be eliminated? *Transactions of the Royal Society of Tropical Medicine and Hygiene*, 103(1, Supplement 1):S2–S5.

- Grenfell, B. T., Bjornstad, O. N., and Finkenstädt, B. F. (2002). Dynamics of measles epidemics: Scaling noise, determinism, and predictability with the TSIR model. *Ecological Monographs*, 72(2):185–202.
- Grimmett, G. R. and Stirzaker, D. R. (1992). *Probability and Random Processes*. Oxford University Press, Oxford.
- Gupta, S., Snow, R. W., Donnelly, C. A., Marsh, K., and Newbold, C. (1999). Immunity to non-cerebral severe malaria is acquired after one or two infections. *Nature Medicine*, 5(3):340–343.
- Gupta, S., Trenholme, K., Anderson, R., and Day, K. (1994). Antigenic diversity and the transmission dynamics of *plasmodium falciparum*. *Science*, 263:961–963.
- Hay, S. I., Cox, J., Rogers, D. J., Randolph, S. E., Stern, D. I., Shanks, G. D., Myers, M. F., and Snow, R. W. (2002). Climate change and the resurgence of malaria in the East African highlands. *Nature*, 415:906–909.
- Hay, S. I., Guerra, C. A., Tatem, A. J., Atkinson, P. M., and Snow, R. W. (2005). Tropical infectious diseases: Urbanization, malaria transmission and disease burden in Africa. *Nature Reviews Microbiology*, 3:81–90.
- Hay, S. I., Myers, M. F., Burke, D. S., Vaughn, D. W., Endy, T., Ananda, N., Shanks, G. D., Snow, R. W., and Rogers, D. J. (2000). Etiology of interepidemic periods of mosquito-borne disease. *Proceedings of the National Academy of Sciences of the USA*, 97(16):9335–9339.
- He, D., Ionides, E. L., and King, A. A. (2010). Plug-and-play inference for disease dynamics: Measles in large and small towns as a case study. *Journal of the Royal Society Interface*, 7:271–283.
- 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., Bhadra, A., and King, A. A. (2009). Iterated filtering. Submitted for publication, and available online at <http://arxiv.org/abs/0902.0347>.
- 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.
- Jacod, J. (2004). The Euler scheme for Lévy driven stochastic differential equations: Limit theorems. *Annals of Probability*, 32:1830–1872.
- Jensen, J. L. and Petersen, N. V. (1999). Asymptotic normality of the maximum likelihood estimator in state space models. *Annals of Statistics*, 27:514–535.

- Johannes, M., Polson, N., and Stroud, J. (2009). Optimal filtering of jump diffusions: Extracting latent states from asset prices. *Review of Financial Studies*, 22:2759–2799.
- Keeling, M. and Ross, J. (2008). On methods for studying stochastic disease dynamics. *Journal of the Royal Society Interface*, 5(19):171–181.
- Kendall, B. E., Briggs, C. J., Murdoch, W. W., Turchin, P., Ellner, S. P., McCauley, E., Nisbet, R. M., and Wood, S. N. (1999). Why do populations cycle? A synthesis of statistical and mechanistic modeling approaches. *Ecology*, 80:1789–1805.
- Kermack, W. O. and McKendrick, A. G. (1927). A contribution to the mathematical theory of epidemics. *Proceedings of the Royal Society of London, Series A*, 115:700–721.
- 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.
- Kiefer, J. and Wolfowitz, J. (1952). Stochastic estimation of the maximum of a regression function. *Annals of Mathematical Statistics*, 23:462–466.
- King, A. A., Ionides, E. L., Bretó, C. M., Ellner, S., and Kendall, B. (2009). *pomp: Statistical inference for partially observed Markov processes*. Available at [www.r-project.org](http://www.r-project.org), R package.
- King, A. A., Ionides, E. L., Pascual, M., and Bouma, M. J. (2008). Inapparent infections and cholera dynamics. *Nature*, 454:877–880.
- Kirkpatrick, S., Gelatt, C. D., J., and Vecchi, M. P. (1983). Optimization by simulated annealing. *Science*, 220(4598):671–680.
- Kiszewski, A. E. and Teklehaimanot, A. (2004). A review of the clinical and epidemiologic burdens of epidemic malaria. *American Journal of Tropical Medicine and Hygiene*, 71(2):128–135.
- Kitagawa, G. (1998). A self-organising state-space model. *Journal of the American Statistical Association*, 93:1203–1215.
- Klein, E., Smith, D., Boni, M., and Laxminarayan, R. (2008). Clinically immune hosts as a refuge for drug-sensitive malaria parasites. *Malaria Journal*, 7(1):67.
- Koella, J. and Antia, R. (2003). Epidemiological models for the spread of anti-malarial resistance. *Malaria Journal*, 2(1):3.
- Kumar, A., Valecha, N., Jain, T., and Dash, A. P. (2007). Burden of malaria in India: Retrospective and prospective view. *American Journal of Tropical Medicine and Hygiene*, 77(6):69–78.

- Kushner, H. J. and Clark, D. S. (1978). *Stochastic Approximation Methods for Constrained and Unconstrained Systems*. Springer-Verlag, New York.
- Kushner, J. and Yin, G. G. (2003). *Stochastic Approximation and Recursive Algorithms and Applications*. Springer, New York.
- Laneri, K., Bhadra, A., Ionides, E. L., Bouma, M., Yadav, R., Dhiman, R., and Pascual, M. (2010). Forcing versus feedback: Epidemic malaria and monsoon rains in NW India. *To appear in PLoS Computational Biology*.
- Le Cam, L. and Yang, G. L. (2000). *Asymptotics in Statistics*. Springer, New York, 2nd edition edition.
- Liu, J. and West, M. (2001). Combining parameter and state estimation in simulation-based filtering. In Doucet, A., de Freitas, N., and Gordon, N. J., editors, *Sequential Monte Carlo Methods in Practice*, pages 197–224. Springer, New York.
- Liu, J. S. (2001). *Monte Carlo Strategies in Scientific Computing*. Springer, New York.
- Liu, K., Raghavan, S., Nelesen, S., Linder, C. R., and Warnow, T. (2009). Rapid and accurate large-scale coestimation of sequence alignments and phylogenetic trees. *Science*, 324(5934):1561–1564.
- Macdonald, G. (1957). *The epidemiology and control of malaria*. Oxford University Press, London.
- Marjoram, P., Molitor, J., Plagnol, V., and Tavaré, S. (2003). Markov chain monte carlo without likelihoods. *Proceedings of the National Academy of Sciences of the United States of America*, 100(26):15324–15328.
- Maryak, J. and Chin, D. (2008). Global random optimization by simultaneous perturbation stochastic approximation. *IEEE Transactions on Automatic Control*, 53(3):780–783.
- McKenzie, F. E. and Bossert, W. H. (2005). An integrated model of Plasmodium falciparum dynamics. *Journal of Theoretical Biology*, 232(3):411–426.
- McKenzie, F. E. and Samba, E. M. (2004). The role of mathematical modeling in evidence-based malaria control. *American Journal of Tropical Medicine and Hygiene*, 71:94–96.
- McKenzie, F. E., Smith, D. L., O’Meara, W. P., and Riley, E. M. (2008). Strain theory of malaria: The first 50 years. *Advances in Parasitology*, 66:1–46.
- Molineaux, L. (1988). The epidemiology of human malaria as an explanation of its distribution, including some implications for its control. In Wernsdorfer, W. H. and McGregor, I., editors, *Malaria: Principles and Practice of Malariology, Volume 2*, pages 913–998. Churchill Livingstone, New York.

- Morton, A. and Finkenstadt, B. F. (2005). Discrete time modelling of disease incidence time series by using Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 54(3):575–594.
- Newman, K. B., Fernandez, C., Thomas, L., and Buckland, S. T. (2008). Monte Carlo inference for state-space models of wild animal populations. *Biometrics*, 65:572 – 583.
- Oehlert, G. (1992). A Note on the Delta Method. *The American Statistician*, 46(1):27–29.
- Olsson, J. and Rydén, T. (2008). Asymptotic properties of particle filter-based maximum likelihood estimators for state space models. *Stochastic Processes and their Applications*, 118(4):649 – 680.
- Packard, R. M. (2007). *The Making of a Tropical Disease: A Short History of Malaria*. Johns Hopkins University Press, Baltimore, Md.
- Pascual, M., Ahumada, J., Chaves, L. F., Rodo, X., and Bouma, M. (2006). Malaria resurgence in East African Highlands: temperature trends revisited. *Proceedings of the National Academy of Sciences of the USA*, 103:5829–5835.
- Pascual, M., Cazelles, B., Bouma, M., Chaves, L., and Koelle, K. (2008). Shifting patterns: malaria dynamics and rainfall variability in an african highland. *Proceedings of the Royal Society B: Biological Sciences*, 275(1631):123–132.
- Pitt, M. K. (2002). Smooth particle filters for likelihood evaluation and maximization. Technical report, University of Warwick.
- Pitt, M. K. and Shephard, N. (1999). Filtering via simulation: Auxillary particle filters. *Journal of the American Statistical Association*, 94:590–599.
- Pitt, M. K. and Shephard, N. (1999). Filtering via simulation: Auxiliary particle filters. *Journal of the American Statistical Association*, 94(446):590–599.
- Polson, N. G., Stroud, J. R., and Muller, P. (2008). Practical filtering with sequential parameter learning. *Journal of the Royal Statistical Society, Series B (Statistical Methodology)*, 70(2):413–428.
- Poyiadjis, G., Doucet, A., and Singh, S. (2009). Sequential monte carlo computation of the score and observed information matrix in state-space models with application to parameter estimation. Technical report, Cambridge University.
- Protter, P. and Talay, D. (1997). The Euler scheme for Lévy driven stochastic differential equations. *The Annals of Probability*, 25(1):393–423.
- Reuman, D. C., Desharnais, R. A., Costantino, R. F., Ahmad, O. S., and Cohen, J. E. (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.

- Roberts, G. O. and Stramer, O. (2001). On inference for partially observed nonlinear diffusion models using the Metropolis-Hastings algorithm. *Biometrika*, 88:603–621.
- Roberts, L. and Enserink, M. (2007). Malaria: Did they really say ... eradication? *Science*, 318(5856):1544–1545.
- Rogers, L. C. G. and Williams, D. (1994). *Diffusions, Markov Processes, and Martingales. Volume 1: Foundations*. Wiley, New York, 2nd edition.
- Ross, R. (1911). *The prevention of malaria*. John Murray, London.
- Self, S. G. and Liang, K.-Y. (1987). Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *Journal of the American Statistical Association*, 82:605–610.
- Sharma, V. P. (1996). Re-emergence of malaria in India. *Indian Journal of Medical Research*, 103:26–45.
- Shumway, R. H. and Stoffer, D. S. (2006). *Time Series Analysis and Its Applications*. Springer, New York, 2nd edition.
- Snow, R. W., Guerra, C. A., Noor, A. M., Myint, H. Y., and Hay, S. I. (2005). The global distribution of clinical episodes of *Plasmodium falciparum* malaria. *Nature*, 434(7030):214–217.
- Spall, J. C. (2003). *Introduction to Stochastic Search and Optimization*. Wiley, Hoboken.
- Stenseth, N. C., Chan, K.-S., Tavecchia, G., Coulson, T., Mysterud, A., Clutton-Brock, T., and Grenfell, B. (2004). Modelling nonadditive and nonlinear signals from climatic noise in ecological time series: Soay sheep as an example. *Proceedings of the Royal Society, Series B*, 271(1552):1985–1993.
- 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.
- Swaroop, S. (1949). Forecasting of epidemic malaria in the Punjab, India. *American Journal of Tropical Medicine and Hygiene*, S1-29(1):1–17.
- Thomson, M. C., Doblus-Reyes, F. J., Mason, S. J., Hagedorn, S. J., Phindela, T., Morse, A. P., and Palmer, T. N. (2006). Malaria early warnings based on seasonal climate forecasts from multi-model ensembles. *Nature*, 439:576–579.
- Toni, T., Welch, D., Strelkowa, N., Ipsen, A., and Stumpf, M. P. (2008). Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems. *Journal of the Royal Society Interface*, 6:187–202.

- 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.
- Warrell, D. A. and Gilles, H. A., editors (2002). *Essential Malariology*. Arnold, London.
- Wearing, H. J., Rohani, P., and Keeling, M. J. (2005). Appropriate models for the management of infectious diseases. *PLoS Medicine*, 2(7):e174.
- West, M. and Harrison, J. (1997). *Bayesian forecasting and dynamic models*. Springer-Verlag, New York.
- Wikipedia (2010). Malaria. in *Wikipedia, The Free Encyclopedia*. [Online; accessed 9-March-2010].
- World Health Organization (2008). *World Malaria Report*. WHO Press, Geneva, Switzerland.
- Xiu, D., Kevrekidis, I. G., and Ghanem, R. (2005). An equation-free, multiscale approach to uncertainty quantification. *Computing in Science and Engineering*, 7(3):16–23.
- Yadav, R. S., Bhatt, R. M., Kohli, V. K., and Sharma, V. P. (2003). The burden of malaria in Ahmedabad city, India: A retrospective analysis of reported cases and deaths. *Annals of Tropical Medicine and Parasitology*, 97:793–802.
- Zhou, G., Minakawa, N., Githeko, A. K., and Yan, G. (2004). Association between climate variability and malaria epidemics in the East African highlands. *Proceedings of the National Academy of Sciences of the USA*, 101(8):2375–2380.