DISSERTATION

PARAMETER INFERENCE AND MODEL SELECTION FOR DIFFERENTIAL EQUATION MODELS

Submitted by

Libo Sun

Department of Statistics

In partial fulfillment of the requirements

For the Degree of Doctor of Philosophy

Colorado State University

Fort Collins, Colorado

Summer 2015

Doctoral Committee:

Advisor: Jennifer A. Hoeting
Co-Advisor: Chihoon Lee

Wen Zhou
N. Thompson Hobbs

ABSTRACT

PARAMETER INFERENCE AND MODEL SELECTION FOR DIFFERENTIAL EQUATION MODELS

Firstly, we consider the problem of estimating parameters of stochastic differential equations with discrete-time observations that are either completely or partially observed. The transition density between two observations is generally unknown. We propose an importance sampling approach with an auxiliary parameter when the transition density is unknown. We embed the auxiliary importance sampler in a penalized maximum likelihood framework which produces more accurate and computationally efficient parameter estimates. Simulation studies in three different models illustrate promising improvements of the new penalized simulated maximum likelihood method. The new procedure is designed for the challenging case when some state variables are unobserved and moreover, observed states are sparse over time, which commonly arises in ecological studies. We apply this new approach to two epidemics of chronic wasting disease in mule deer.

Next, we consider the problem of selecting deterministic or stochastic models for a biological, ecological, or environmental dynamical process. In most cases, one prefers either deterministic or stochastic models as candidate models based on experience or subjective judgment. Due to the complex or intractable likelihood in most dynamical models, likelihood-based approaches for model selection are not suitable. We use approximate Bayesian computation for parameter estimation and model selection to gain further understanding of the dynamics of two epidemics of chronic wasting disease in mule deer. The main novel contribution of this work is that under a hierarchical model framework we compare three types of dynamical models: ordinary differential equation, continuous time Markov chain, and stochastic differential equation models. To our knowledge model selection between these types of models has not appeared previously. The practice of incorporating dynamical mod-

els into data models is becoming more common, the proposed approach may be useful in a variety of applications.

Lastly, we consider estimation of parameters in nonlinear ordinary differential equation models with measurement error where closed-form solutions are not available. We propose a new numerical algorithm, the data driven adaptive mesh method, which is a combination of the Euler and 4th order Runge-Kutta methods with different step sizes based on the observation time points. Our results show that both the accuracy in parameter estimation and computational cost of the new algorithm improve over the most widely used numerical algorithm, the 4th Runge-Kutta method. Moreover, the generalized profiling procedure proposed by Ramsay et al. (2007) doesn't have good performance for sparse data in time as compared to the new approach. We illustrate our approach with both simulation studies and ecological data on intestinal microbiota.

ACKNOWLEDGEMENTS

I would never have been able to finish my dissertation without the guidance of my committee members, help from friends, and support from my family and wife.

I would first like to thank my parents and sister for their unwavering support and love. I really appreciate their encouragement and help throughout my PhD study. I would also like to especially thank my wife, Jiwen Wu. She was always there cheering me up and stood by me through the good times and bad.

I would also like to express my deepest gratitude to my co-advisors, Jennifer Hoeting and Chihoon Lee, for their excellent support throughout this dissertation. My co-advisors have been great instructors, advisors, and research partners. I would like to thank them not only for their help with my dissertation but for their valuable mentoring throughout these past few years. Furthermore, I would like to thank Wen Zhou for his valuable research suggestions and guidance. I am grateful to N. Thompson Hobbs for his support and insights throughout this dissertation. I would also like to thank Michael W. Miller and the Colorado Division of Parks and Wildlife for sharing the data on chronic wasting disease that motivated much of the work in this dissertation. I would also like to express my gratitude to my fellow graduate students, the faculty, and the staff in the Statistics Department at Colorado State University for their help over the years.

The research in this dissertation was supported by the National Science Foundation (EF-0914489). This research also utilized the CSU ISTeC Cray HPS System which is supported by NSF Grant CSN-0923386.

DEDICATION

To my parents,

Jiyin Sun and Feng Liu

To my sister,

Xiao Liu

To my wife,

Jiwen Wu

And to my little boy,

Lucas W. Sun

for their unconditional love and support and patience

and for making it all worthwhile

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION AND BACKGROUND

1.1 Overview

Differential equations play a prominent role in many scientific areas including biology, ecology, economics, finance, bioinformatics, physics, and engineering. The reason why differential equations are so widely used is that they are mathematical equations that relate some function with their derivatives. In applications, the functions often represent physical quantities, the derivatives represent their rates of change, and the equation defines a relationship between the two.

In this dissertation we focus on two types of dynamical models, ordinary differential equation (ODE) and stochastic differential equation (SDE) models, for epidemics and other processes that evolve over time. We focus on ecological and biological applications. One of the most important differences between stochastic and deterministic models is that in a specified interval of time stochastic models define the probability of disease transmission between two individuals, whereas deterministic models state whether or not transmission will occur. A solution of a deterministic model is a function of time or space and is generally uniquely dependent on the initial data. A solution of a stochastic model is a probability distribution for each of the random variables. One sample path over time or space is one realization from this distribution. The process randomness or error described by stochastic models may arise from structural model misspecification or unpredictable random behavior of the underlying processes. Hence, stochastic epidemic models typically allow for more realistic description of the transmission of disease as compared to deterministic epidemic

models, especially when the number of members of the population subject to the epidemic is relatively small (Andersson and Britton, 2000; Daley and Gani, 2001).

Besides the process error, in real life applications there might be also uncertainty about the measurement and the numerical error due to numerical approximation in solving differential equations when the closed form solution is not available. Note that both ODE and SDE models could have both measurement error and numerical error.

In Chapter 2, we firstly consider stochastic differential equation models without measurement error. Then we extend the new algorithm we developed to the case that allows measurement error for stochastic differential equation models in Chapter 3. In Chapter 4, a new methodology for parameter estimation in ordinary differential equation models with measurement error is proposed. In Chapter 5, we consider the problem of parameter inference and model selection among both deterministic and stochastic models with or without measurement error. Chapter 6 concludes with a discussion.

1.2 Mathematical models for the outbreak and spread of disease

The outbreak and spread of disease have been studied for many years. The ability to make predictions about diseases could enable scientists to evaluate inoculation or isolation plans and may have a significant effect on the mortality rate of a particular epidemic. The modeling of infectious diseases is a tool which has been used to study the mechanisms by which diseases spread, to predict the future course of an outbreak and to evaluate strategies to control an epidemic (Daley and Gani, 2001).

1.2.1 SIR epidemic model

One of the most commonly used epidemic model is called the SIR for Susceptible – Infected – Recovered model. As a variant on this title SIR can also stand for Susceptible – Infected – Removed. Such types of models have been applied to many problems including

childhood diseases such as measles, mumps, and chickenpox (Anderson and May, 1992; Hethcote, 2000).

ODE SIR model Here we first consider a simple deterministic SIR model with no birth or natural death in a closed population. A set of ordinary differential equations (ODEs) describing the dynamics of an SIR epidemic are as follows:

$$\begin{aligned} \frac{dS}{dt} &= -\beta SI, \\ \frac{dI}{dt} &= \beta SI - \gamma I, \\ \frac{dR}{dt} &= \gamma I, \end{aligned} \tag{1}$$

where $\beta > 0$ is the transmission rate (unit = time^{-1}), $\gamma > 0$ is the disease recovery or death rate (unit = time^{-1}), initial conditions satisfy $S(0) > 0, I(0) > 0, R(0) \geq 0$, and $S(t) + I(t) + R(t) = N$, the total population size. The number of new infected individuals produced per susceptible individual per unit time is βI.

The dynamics of the epidemic model can be described by the basic reproduction number, which is the average number of secondary infections that occur when one infected individual is introduced into a completely susceptible population. For model (1), the basic reproduction number is defined as

$$\mathcal{R}_0 = \frac{\beta}{\gamma} N.$$

$\mathcal{R}_0$ also has an epidemiological interpretation. If $\mathcal{R}_0 \frac{S(0)}{N} > 1$, the population experiences an epidemic outbreak before the disease eventually disappears, and if $\mathcal{R}_0 \frac{S(0)}{N} \leq 1$, $I(t)$ decreases monotonically to zero and there is no epidemic (Allen, 2011). The basic reproduction number $\mathcal{R}_0$ is an important indicator because usually $I(0)$ is small compared to N and $R(0) = 0$, that is $S(0)/N \approx 1$.

SDE SIR model The stochastic differential equation (SDE) epidemic model is one of the commonly used stochastic epidemic models. It is based on a diffusion process, where both the time and state variables are continuous. Here, we briefly explain how to derive the corresponding SDE SIR model from the deterministic SIR model (1). See Allen (2011) for more details.

Let $\mathbf{X}(t) = (S(t), I(t))^T$, where $S(t)$ and $I(t)$ denote continuous random variables for the susceptible and infected individuals at time t, and let $\mathbf{X}_\delta = \mathbf{X}(t+\delta) - \mathbf{X}(t)$ be the increment during the time interval δ, where $\delta > 0$. If δ is sufficiently small, we can assume at most one animal is infected or died during the time interval δ. The probability of an event that more than one infection or death has occurred during time δ is of order δ^2, which can be neglected. Then we can approximate the mean of $\mathbf{X}_\delta$ for δ sufficiently small to order δ by

$$E[\mathbf{X}_\delta] \approx \begin{bmatrix} -\beta SI \\ \beta SI - \gamma I \end{bmatrix} \delta \equiv \boldsymbol{\mu}\delta.$$

Furthermore, we can also approximate the covariance of $\mathbf{X}_\delta$ for δ sufficiently small by

$$V[\mathbf{X}_\delta] = E[(\mathbf{X}_\delta)(\mathbf{X}_\delta)^T] - E(\mathbf{X}_\delta)E(\mathbf{X}_\delta)^T \approx E[(\mathbf{X}_\delta)(\mathbf{X}_\delta)^T] = \Sigma\delta,$$

where

$$\Sigma = \begin{bmatrix} \beta SI & -\beta SI \\ -\beta SI & \beta SI + \gamma I \end{bmatrix},$$

which is positive definite and hence has a positive definite square root $\boldsymbol{B} = \sqrt{\Sigma}$. By the Central Limit Theorem, we assume $\mathbf{X}_\delta$ follows a normal distribution with mean vector $\boldsymbol{\mu}\delta$

and covariance matrix $\boldsymbol{B}^2\delta = \Sigma\delta$. Thus,

$$\mathbf{X}(t+\delta) \approx \mathbf{X}(t) + \boldsymbol{\mu}\delta + \boldsymbol{B}\sqrt{\delta}\boldsymbol{\eta}, \tag{2}$$

where $\boldsymbol{\eta} \sim N(0, \boldsymbol{\mathcal{I}}_{2\times 2})$ and $\boldsymbol{\mathcal{I}}$ is the identity matrix. This is exactly one iteration of the Euler-Maruyama scheme for a SDE SIR model (Kloeden and Platen, 1992), which will be described further in Section 1.3, which is defined as follow:

$$\begin{aligned} dS &= -\beta SI dt + \boldsymbol{B}_{11} dW_1 + \boldsymbol{B}_{12} dW_2, \\ dI &= (\beta SI - \gamma I)\, dt + \boldsymbol{B}_{21} dW_1 + \boldsymbol{B}_{22} dW_2, \end{aligned} \tag{3}$$

where $S \in [0, N-I]$, $I \in [0, N-S]$, the matrix $\boldsymbol{B} = (\boldsymbol{B}_{ij})$, and W_1 and W_2 are two independent Wiener processes. The dynamical system (2) converges in the mean square sense to the system of SDEs (3) as $\delta \to 0$. Note that the number of recoveries or deaths at time t is given by $R(t) = N - S(t) - I(t)$ since the total population size N is a constant.

For the stochastic SIR model, the probability that there is no epidemic equals 1 if $\mathcal{R}_0 \leq 1$ and $(\frac{1}{\mathcal{R}_0})^{I(0)}$ if $\mathcal{R}_0 > 1$ when $S(0)/N \approx 1$. If $R_0 > 1$ and the population is large but the epidemic is initiated by only a few initial infective individuals, there is still a possibility that the outbreak will never happen.

Three sample paths of the SDE SIR epidemic model (3) with parameter $\beta = 0.4, \gamma = 0.2$, and the initial condition $S(0) = 98, I(0) = 2, R(0) = 0$ are shown in Figure 1.1. In this case, the basic reproduction number $\mathcal{R}_0 = 2$ and $S(0)/N \approx 1$, hence for the SDE SIR model the epidemic will occur with the probability 0.75; for the deterministic SIR model the epidemic is guaranteed to occur which may not be realistic for a given disease. This is one of the practical motivations of using a stochastic epidemic model.

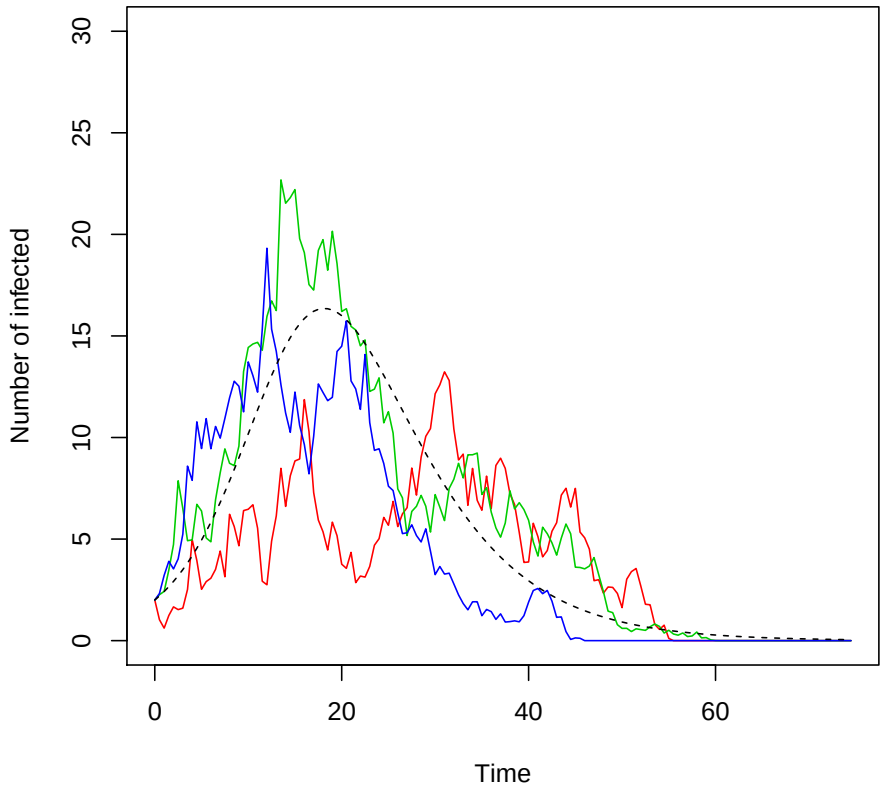

Figure 1.1: Three sample paths of the SDE SIR model with parameters $\beta = 0.4, \gamma = 0.2$, and the initial condition $(S(0) = 98, I(0) = 2, R(0) = 0)$. The deterministic trajectory is the dashed line.

1.2.2 Chronic wasting disease (CWD)

Deer populations and ecosystems can be severely disrupted by the contagious prion disease, chronic wasting disease (CWD) (Miller et al., 2006). CWD has been documented to have a widely spread throughout North America. Figure 1.2 shows the distribution of CWD in North America by April, 2014. Weight loss is the most obvious and consistent clinical sign of CWD and CWD always causes death. Since there are no effective vaccines or therapies, it is important to understand the transmission mechanisms of CWD in order to reduce the potential damage caused by CWD. Several deterministic epidemic models were proposed by Miller et al. (2006) in order to portray the transmission of CWD. In Chapter 2 and 5

we consider several of the ODE models proposed by Miller et al. (2006) as well as several extensions including SDE models and continuous time Markov chain models.

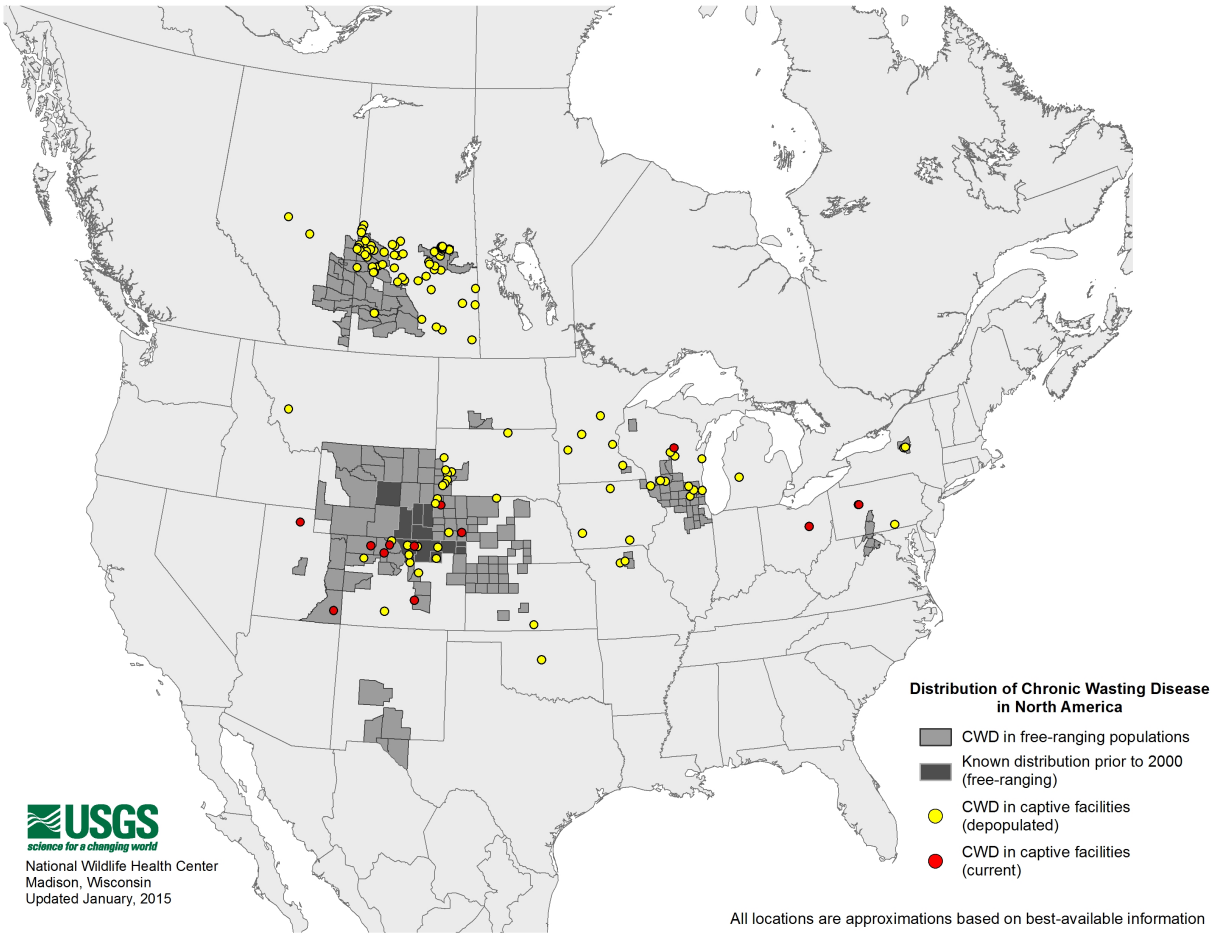

Figure 1.2: Distribution of CWD in North America by Jan, 2015. Source: USGS, National Wildlife Health Center.
http://www.nwhc.usgs.gov/disease_information/chronic_wasting_disease

In Chapters 2 and 5 we apply our methods to a dataset studied in Miller et al. (2006), which consisted of annual observations of cumulative mortality from two distinct CWD epidemics in captive mule deer held at the Colorado Division of Wildlife Foothills Wildlife Research Facility in Fort Collins, Colorado. The first epidemic occurred from 1974 to 1985. The second epidemic occurred in a new deer herd from 1992 to 2001. The dataset also includes the annual number of new deer added to the herd and the per capita losses due

to natural deaths and removals. Note that we only observed the number of deaths but the numbers of susceptible and infected are unknown.

1.3 Parameter inference and model selection for differential equation models

1.3.1 General SDE models without measurement error

Firstly, we consider a general SDE model without measurement error. Let $\mathbf{X}(t) = \{X_1(t), \ldots, X_k(t)\}^T$ denote a k-dimensional state variable vector at time t. Consider a general multivariate Itô SDE model,

$$d\mathbf{X}(t) = f(\mathbf{X}(t), \boldsymbol{\theta})dt + g(\mathbf{X}(t), \boldsymbol{\theta})d\mathbf{W}(t) \tag{4}$$

with known initial condition $\mathbf{X}(0) = \mathbf{x}_0$, and $\boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^p$ is an unknown p-dimensional parameter vector, $\mathbf{W}$ is a k-dimensional standard Wiener process, and both functions $f : \mathbb{R}^k \times \Theta \to \mathbb{R}^k$ and $g : \mathbb{R}^k \times \Theta \to \mathbb{R}^{k \times k}$ are known.

A solution of (4) is a solution of the integral equation,

$$\mathbf{X}(t) = \mathbf{X}(0) + \int_0^t f(\mathbf{X}(s), \boldsymbol{\theta})ds + \int_0^t g(\mathbf{X}(s), \boldsymbol{\theta})d\mathbf{W}(s),$$

where the first integral is a Riemann integral and the second integral is an Itô stochastic integral.

It can be shown that the SDE (4) has a unique Markov process solution if there exist constants $C, D > 0$ such that

$$|f(\mathbf{X}(t), \boldsymbol{\theta}) - f(\mathbf{Y}(t), \boldsymbol{\theta})| + |g(\mathbf{X}(t), \boldsymbol{\theta}) - g(\mathbf{Y}(t), \boldsymbol{\theta})| \leq C|\mathbf{X}(t) - \mathbf{Y}(t)|,$$
$$|f(\mathbf{X}(t), \boldsymbol{\theta})| + |g(\mathbf{X}(t), \boldsymbol{\theta})| \leq D(1 + |\mathbf{X}(t)|)$$

for $\mathbf{X}(t), \mathbf{Y}(t) \in \mathbb{R}^k$ and $t \in [0, T]$ (Kuo, 2006; Øksendal, 2010).

The Euler-Maruyama scheme (Kloeden and Platen, 1992) is a common approach to approximate the transition density between two discrete observations $\mathbf{X}(t + \delta)$ and $\mathbf{X}(t)$. The approximation is given by

$$\mathbf{X}(t + \delta) - \mathbf{X}(t) \approx f(\mathbf{X}(t), \boldsymbol{\theta})\delta + g(\mathbf{X}(t), \boldsymbol{\theta})(\mathbf{W}(t + \delta) - \mathbf{W}(t)),$$

where $\delta > 0$ is called the step size, $\mathbf{W}(t+\delta)-\mathbf{W}(t)$ follows a multivariate normal distribution with mean zero and variance matrix $\delta \boldsymbol{\mathcal{I}}_{k\times k}$, where $\boldsymbol{\mathcal{I}}$ is the identity matrix. The Euler-Maruyama scheme works well if the step size δ is small sufficiently.

We consider the problem of estimating parameters of SDEs with discrete-time observations that are completely or partially observed. The process defined by an SDE is in continuous time, but the data are always sampled in discrete time. The transition density between two observations is known in only a few univariate cases, and it has to be approximated in most cases. Different methodologies have been proposed in the literature to estimate the parameters of an SDE. These include importance sampling (Pedersen, 1995b; Santa-Clara, 1997; Durham and Gallant, 2002; Stramer and Yan, 2007a; Lindström, 2012), Hermite polynomials expansion (Aït-Sahalia, 2002), Bayesian and Markov Chain Monte Carlo (MCMC) approaches (Elerian et al., 2001; Eraker, 2001; Golightly and Wilkinson, 2005, 2006; Beskos et al., 2006), estimating functions (Bibby et al., 2004), and generalized method of moments (Clement, 2001). Jimenez et al. (2005) and Sørensen (2004) are good summaries of those methods for different situations.

In Chapter 2, we propose an importance sampling approach with an auxiliary parameter, penalized simulated maximum likelihood (PSML), which provides more accurate estimates of the parameters of an SDE when the transition density is unknown. We show via simulation studies that our approach improves the accuracy of parameter estimates and computational efficiency compared to several other methods. In Chapter 3, we extend our method to general SDE models with measurement error.

1.3.2 General ODE models with measurement error

In Chapter 4, we focus on the case when ODE models have measurement error. Suppose state variable $\boldsymbol{X}(t) = \{X_1(t), \ldots, X_k(t)\}^T$ is modeled by ordinary differential equations

$$\frac{d\boldsymbol{X}}{dt} = \boldsymbol{f}(\boldsymbol{X}(t), \boldsymbol{\theta}) \tag{5}$$

with known initial condition $\boldsymbol{X}(0) = \boldsymbol{x}_0$, where the map $\boldsymbol{f} : \mathbb{R}^p \times \Theta \to \mathbb{R}^p$ is assumed to be smooth. We observe the data with measurement error $\boldsymbol{\epsilon}(t)$,

$$\boldsymbol{Y}(t) = \boldsymbol{X}(t, \boldsymbol{\theta}) + \boldsymbol{\epsilon}(t)$$

where $\boldsymbol{\theta} \in \Theta \subset \mathbb{R}^d$; and $\boldsymbol{\epsilon}(t) \in L^2(T)$ (i.e. square integrable on the time domain) has zero mean and covariance function $\mathrm{Cov}(\boldsymbol{\epsilon}(t), \boldsymbol{\epsilon}(s)) = \sigma^2 \delta_{ts} \mathcal{I}$, where $\delta_{ts} = 1$ if $t = s$ and 0 elsewhere.

If a closed-form solution $\boldsymbol{X}(t)$ of ODE (5) is available, then the standard nonlinear least squares estimator can be used to estimate unknown parameters $\boldsymbol{\theta}$. However, a closed-form solution of (5) is not available for most of cases in practice. Numerical methods, such as Euler (Euler, 1913) and Runge-Kutta method (Runge, 1895; Kutta, 1901), are needed to

solve ODEs. The Euler method is given by the following recursive scheme

$$\boldsymbol{y}_{s+1} = \boldsymbol{y}_s + h\boldsymbol{f}(\boldsymbol{y}(t_s), \boldsymbol{\theta}),$$

where h is the step size and $\boldsymbol{y}_s$ is a numerical estimate of the exact solution $\boldsymbol{y}(t_s)$. The Runge-Kutta method with order four (RK4) is one of most widely used numerical methods, which is given by

$$\begin{aligned}
\boldsymbol{y}_{s+1} &= \boldsymbol{y}_s + \frac{h}{6}(\boldsymbol{k}_1 + 2\boldsymbol{k}_2 + 2\boldsymbol{k}_3 + \boldsymbol{k}_4), \\
\boldsymbol{k}_1 &= \boldsymbol{f}(t_s, \boldsymbol{y}_s) \\
\boldsymbol{k}_2 &= \boldsymbol{f}(t_s + h/2, \boldsymbol{y}_s + \boldsymbol{k}_1 h/2) \\
\boldsymbol{k}_3 &= \boldsymbol{f}(t_s + h/2, \boldsymbol{y}_s + \boldsymbol{k}_2 h/2) \\
\boldsymbol{k}_4 &= \boldsymbol{f}(t_s + h, \boldsymbol{y}_s + \boldsymbol{k}_3 h)
\end{aligned}$$

For discrete observation points, $t_1, \ldots, t_n$, and we are able to estimate $\boldsymbol{\theta}$ using nonlinear least squares, which is given by

$$\widehat{\boldsymbol{\theta}} = \text{argmin}_{\boldsymbol{\theta}} \sum_{i=1}^{n} ||\boldsymbol{Y}(t_i) - \widetilde{\boldsymbol{X}}(t_i, \boldsymbol{\theta})||^2,$$

where $\widetilde{\boldsymbol{X}}$ is an estimator of $\boldsymbol{X}$ by ODE numerical solver (e.g. Euler or RK4 method).

Although the RK4 has better accuracy, the computation cost of RK4 is much higher than Euler's method for the same step size. In Chapter 4, we propose a new method, data driven adaptive mesh (DDAM), that balances the accuracy and computational cost, especially when the data are sparsely observed on time. It implements the RK4 method with a smaller step size h around the data points and the Euler method with a larger step size H on elsewhere.

1.3.3 Model selection

Lastly, we consider the problem of selecting deterministic or stochastic models for a biological, ecological, or environmental dynamical process with or without measurement error. In most cases, one prefers either deterministic or stochastic models as candidate models based on experience or subjective judgement. In Chapter 5, we incorporate an Bayesian algorithm, approximate Bayesian computation (ABC), into a hierarchical model framework, and perform parameter estimation (with credible intervals) and model selection among a set of ordinary differential equation, continuous time Markov chain, and stochastic differential equation models that arise as models for the transmission of CWD. To our knowledge model selection between these types of models has not appeared previously.

CHAPTER 2

A PENALIZED SIMULATED MAXIMUM LIKELIHOOD APPROACH IN PARAMETER ESTIMATION FOR STOCHASTIC DIFFERENTIAL EQUATIONS

2.1 Introduction

It is very important for ecologists and wildlife managers to understand the dynamics of infectious diseases, such as chronic wasting disease (CWD) is a fatal disease in cervid populations (Miller et al., 2006). Several ordinary differential equation models have been proposed by Miller et al. (2006) to describe the transmission mechanism of CWD. Stochastic epidemic models allow more realistic description of the transmission of disease as compared to deterministic epidemic models (Becker, 1979; Andersson and Britton, 2000). However, parameter estimation is challenging for discretely observed data for stochastic models (Sørensen, 2004; Jimenez et al., 2005). Stochastic differential equation (SDE) models are a natural extension of ordinary differential equation models and they may be simpler to derive and apply than Markov chain models. For example, the transition matrix in Markov chain models can be very complicated when there are several interacting populations (Allen and Allen, 2003; Allen et al., 2005). Moreover, SDEs have broader application areas, which include not only ecology and biology but also economics, finance, bioinformatics, and engineering.

Various methods for inferential problems for SDEs have been developed. The Hermite polynomial expansion approach proposed by Aït-Sahalia (2002, 2008) may perform poorly if the data are sparsely sampled (Stramer and Yan, 2007b). Moreover, this approach has some restrictions which could limit its application, especially for multivariate models (Lindström, 2012). Särkkä and Sottinen (2008) proposed an approach which uses an alternative SDE as

an importance process and the Girsanov theorem to help evaluate the likelihood ratios of two SDEs. However, the diffusion coefficient of their model is state-independent, whereas general SDE models allow for a state-dependent diffusion coefficient. Recent developments have mainly been focused on Bayesian approaches (Eraker, 2001; Golightly and Wilkinson, 2005, 2006, 2011; Donnet et al., 2010), which can suffer a very slow rate of convergence as the dimension of the model increases and the data are sparsely sampled. We propose a penalized simulated maximum likelihood (PSML) approach which is computationally feasible.

For a SDE model the transition density between two observations is known in only a few univariate cases. Pedersen (1995b) firstly proposed a simulated maximum likelihood (SML) approach which integrates out the unobserved states using Monte Carlo integration with importance sampling. We refer to the basic sampler in this approach as the Pedersen sampler. Although the Pedersen sampler may provide estimates that are arbitrarily close to the true transition density, it is computationally expensive in practice. Durham and Gallant (2002) proposed several different importance samplers in a SML framework to improve the efficiency of the Pedersen sampler. They concluded their modified Brownian bridge (MBB) sampler has the best performance in terms of accuracy in root mean square error and efficiency in time. Richard and Zhang (2007) proposed an efficient importance sampling technique which converts the problem of minimizing the variance of an approximate likelihood to a recursive sequence of auxiliary least squares optimization problems. Pastorello and Rossi (2010) applied Richard and Zhang's approach to estimate the parameters of some univariate SDE models. However, the extension to multivariate SDEs with partially observed data is not trivial. Lindström (2012) introduced a regularized bridge sampler, which is a weighted combination of the Pedersen sampler and the MBB sampler, for sparsely sampled data.

The methods of Pedersen (1995b) and Durham and Gallant (2002) have mainly been applied in the area of econometrics. Here we propose a methodology to improve the MBB sampler and the regularized sampler and extend them to the area of ecology. From an inferential viewpoint, practitioners must contend with two major challenges: (a) in the

multivariate state space, some state variables are completely unobserved; (b) observed data are quite sparse over time. These are common features of ecological data. Moreover, the time interval between two consecutive observations could be very long, usually weeks or even months. With such partially observed sparse data, the MBB approach no longer has the same promising results as in the univariate case. Although the regularized sampler in Lindström (2012) is designed for sparsely sampled data, the optimal choice of the weight parameter ρ (which is denoted as α in the cited paper) needs to be determined. We propose an importance sampling approach with an auxiliary parameter which provides more accurate estimates of the parameters of an SDE when the transition density is unknown.We embed the auxiliary importance sampler in a penalized maximum likelihood framework. The penalty term we add to the log likelihood is a constraint on selecting the importance sampler. We show via simulation studies that our approach improves the accuracy of parameter estimates and computational efficiency compared to the MBB sampler and the regularized sampler.

The remainder of the chapter is organized as follows. In Section 2, we present the general multivariate SDE model. Section 3 provides brief descriptions of the Pedersen, MBB and regularized samplers. Section 4 describes our methodology in detail. Section 5 presents simulation studies for different models. Section 6 illustrates our method on a CWD dataset as a real world example. Section 7 concludes with a discussion.

2.2 Background

We begin with the general multivariate SDE model where some state variables are unobserved. Let $\boldsymbol{X}(t) = \{X_1(t), \ldots, X_k(t)\}^T$ denote a k-dimensional state variable vector at

time $t \geq 0$. Consider a multivariate SDE model,

$$d\boldsymbol{X}(t) = f(\boldsymbol{X}(t), \boldsymbol{\theta})dt + g(\boldsymbol{X}(t), \boldsymbol{\theta})d\boldsymbol{W}(t) \tag{6}$$

with known initial condition $\boldsymbol{X}(t_0) = \boldsymbol{x}_0$, where $\boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^p$ is an unknown p-dimensional parameter vector, $\boldsymbol{W}$ is a k-dimensional standard Wiener process, and both functions $f : \mathbb{R}^k \times \Theta \to \mathbb{R}^k$ and $g : \mathbb{R}^k \times \Theta \to \mathbb{R}^{k \times k}$ are known. Note that the derivation below still holds for the case with unknown initial condition $\boldsymbol{X}(t_0)$, which can be treated as another unknown parameter. We assume that the SDE (6) has a unique weak solution. See Øksendal (2010, Chapter 5) for conditions that ensure this.

We assume that only a subset of the state process $\{\boldsymbol{X}_{\text{obs}}(t)\}_{t \geq 0}$ can be observed at discrete time points. It is natural to suppose only $\boldsymbol{X}_{\text{obs}}(t_i) = \{X_j(t_i), \ldots, X_k(t_i)\}$ is observed at t_i, for $1 < j \leq k$ and $i = 1, \ldots, n$, and all other state variables $\boldsymbol{X}_{-\text{obs}}(t_i) = \{X_1(t_i), \ldots, X_{j-1}(t_i)\}$, are unobserved. In the case of complete observation, that is when $j = 1$, a similar derivation as below can be obtained. Note that time intervals do not have to be equidistant.

The discrete-time likelihood of model (6) is given by

$$L(\boldsymbol{\theta}) = p(\boldsymbol{X}_{\text{obs}}(t_1)|\boldsymbol{X}(t_0), \boldsymbol{\theta}) \prod_{i=2}^{n} p(\boldsymbol{X}_{\text{obs}}(t_i)|\boldsymbol{X}(t_0), \boldsymbol{X}_{\text{obs}}(t_1 : t_{i-1}), \boldsymbol{\theta}) \tag{7}$$

where $\boldsymbol{X}_{\text{obs}}(t_1 : t_{i-1})$ denotes all observations of $\boldsymbol{X}_{\text{obs}}$ from time t_1 to t_{i-1}. We omit the parameter $\boldsymbol{\theta}$ for brevity from now on. Notice that the term $p(\boldsymbol{X}_{\text{obs}}(t_i)|\boldsymbol{X}(t_0), \boldsymbol{X}_{\text{obs}}(t_1 : t_{i-1}))$ is not available in closed form except for simple cases. However, factoring the likelihood as in (7) allows us to evaluate the likelihood given by

$$p(\boldsymbol{X}_{\text{obs}}(t_i)|\boldsymbol{X}(t_0), \boldsymbol{X}_{\text{obs}}(t_1 : t_{i-1})) = \int p(\boldsymbol{X}_{\text{obs}}(t_i)|\boldsymbol{X}(t_{i-1}))p(\boldsymbol{X}_{-\text{obs}}(t_{i-1})|\boldsymbol{X}(t_0), \boldsymbol{X}_{\text{obs}}(t_1 : t_{i-1}))d\boldsymbol{X}_{-\text{obs}}(t_{i-1}).$$

A feasible approach to evaluate this integral is via Monte Carlo integration. That requires a method to draw samples from the distribution of $\boldsymbol{X}_{-\text{obs}}(t_{i-1})|\boldsymbol{X}(t_0), \boldsymbol{X}_{\text{obs}}(t_1 : t_{i-1})$. It can be shown that (cf. Durham and Gallant, 2002)

$$p(\boldsymbol{X}_{-\text{obs}}(t_i)|\boldsymbol{X}(t_0), \boldsymbol{X}_{\text{obs}}(t_1 : t_i)) \propto \int p(\boldsymbol{X}(t_i)|\boldsymbol{X}(t_{i-1}))p(\boldsymbol{X}_{-\text{obs}}(t_{i-1})|\boldsymbol{X}(t_0), \boldsymbol{X}_{\text{obs}}(t_1 : t_{i-1}))d\boldsymbol{X}_{-\text{obs}}(t_{i-1}), \tag{8}$$

for $i = 1, ..., n$. Therefore, assuming $p(\boldsymbol{X}(t_i)|\boldsymbol{X}(t_{i-1}))$ is known (see below), iterative application of Monte Carlo integration (8) yields an approximation of $\boldsymbol{X}_{-\text{obs}}(t_\ell)|\boldsymbol{X}(t_0), \boldsymbol{X}_{\text{obs}}(t_1 : t_\ell)$ for $\ell \geq 1$. This is similar in spirit to a particle filter (Durham and Gallant, 2002; Pitt and Shephard, 1999), but our model does not include measurement errors. The algorithmic form of this simple sequential Monte Carlo algorithm is provided in Appendix.

It is left to approximate the transition probability density $p(\boldsymbol{X}(t_i)|\boldsymbol{X}(t_{i-1}))$ which has no closed form in most cases. The Euler-Maruyama scheme (Kloeden and Platen, 1992) is a common approach to approximate the solution of an SDE, which is given by

$$\boldsymbol{X}(t+\delta) - \boldsymbol{X}(t) \approx f(\boldsymbol{X}(t), \boldsymbol{\theta})\delta + g(\boldsymbol{X}(t), \boldsymbol{\theta})(\boldsymbol{W}(t+\delta) - \boldsymbol{W}(t)), \tag{9}$$

where δ is the step size and $\mathbf{W}(t+\delta) - \mathbf{W}(t)$ follows a multivariate normal distribution with variance matrix $\delta\boldsymbol{\mathcal{I}}_{k\times k}$, where $\boldsymbol{\mathcal{I}}$ is the identity matrix. This Euler-Maruyama scheme works well if the step size is small. Hence, if the time interval between two observations is small enough, we can approximate $p(\boldsymbol{X}(t_i)|\boldsymbol{X}(t_{i-1}))$ using a multivariate normal density.

If the time interval between observations is large, the above approximation will introduce bias. We can partition the interval t_{i-1} to t_i to M subintervals such that $\delta = (t_i - t_{i-1})/M$ is small enough for the Euler-Maruyama scheme. By the Markov property, Pedersen (1995b)

proved that $p(\boldsymbol{X}(t_i)|\boldsymbol{X}(t_{i-1}))$ can be approximated by

$$p^{(M)}\big(\boldsymbol{X}(t_i)|\boldsymbol{X}(t_{i-1})\big) \equiv \int \prod_{m=1}^{M} p^{(1)}\big(\boldsymbol{X}(t_{i-1}+m\delta)|\boldsymbol{X}(t_{i-1}+(m-1)\delta)\big) \\ d\boldsymbol{X}\big((t_{i-1}+\delta):(t_i-\delta)\big), \tag{10}$$

where $p^{(1)}(\cdot)$ is the multivariate normal density approximated by Euler-Maruyama scheme.

Then, our goal is to compute $p^{(M)}(\boldsymbol{X}(t_i)|\boldsymbol{X}(t_{i-1}))$. Using importance sampling, we draw i.i.d. J samples, $\{\boldsymbol{X}^{(j)}((t_{i-1}+\delta):(t_i-\delta)), j=1,\cdots,J\}$, from an importance sampler q, then (10) can be approximated by

$$\frac{1}{J}\sum_{j=1}^{J} h\Big(\boldsymbol{X}^{(j)}\big((t_{i-1}+\delta):(t_i-\delta)\big)\Big), \tag{11}$$

where

$$h\Big(\boldsymbol{X}^{(j)}\big((t_{i-1}+\delta):(t_i-\delta)\big)\Big) \equiv \frac{\prod_{m=1}^{M} p^{(1)}\big(\boldsymbol{X}^{(j)}(t_{i-1}+m\delta)|\boldsymbol{X}^{(j)}(t_{i-1}+(m-1)\delta)\big)}{q\big(\boldsymbol{X}^{(j)}((t_{i-1}+\delta):(t_i-\delta))\big)}. \tag{12}$$

The convergence of the importance sampling estimator (11) to (10) as $J\to\infty$ is shown by Geweke (1989). The estimator (11) is an unbiased estimator, regardless of the choice of the importance sampler q. The variance of (11) is given by

$$\mathrm{Var}\left(\frac{1}{J}\sum_{j=1}^{J} h\left(\boldsymbol{X}^{(j)}\big((t_{i-1}+\delta):(t_i-\delta)\big)\right)\right) = \frac{1}{J}\mathrm{Var}\left(h\Big(\boldsymbol{X}\big((t_{i-1}+\delta):(t_i-\delta)\big)\Big)\right) \\ = \frac{1}{J}\left[\int \frac{\prod_{m=1}^{M}\Big[p^{(1)}\big(\boldsymbol{X}(t_{i-1}+m\delta)|\boldsymbol{X}(t_{i-1}+(m-1)\delta)\big)\Big]^2}{q\big(\boldsymbol{X}((t_{i-1}+\delta):(t_i-\delta))\big)} \right. \\ \left. d\boldsymbol{X}\big((t_{i-1}+\delta):(t_i-\delta)\big) - \big[p^{(M)}(\boldsymbol{X}(t_i)|\boldsymbol{X}(t_{i-1}))\big]^2\right], \tag{13}$$

which attains its minimum of 0 when

$$q\big(\boldsymbol{X}((t_{i-1}+\delta):(t_i-\delta))\big) = \frac{\prod_{m=1}^{M} p^{(1)}(\boldsymbol{X}(t_{i-1}+m\delta)|\boldsymbol{X}(t_{i-1}+(m-1)\delta))}{p^{(M)}(\boldsymbol{X}(t_i)|\boldsymbol{X}(t_{i-1}))}. \tag{14}$$

Thus in theory a single sample is sufficient to approximate $p^{(M)}(\boldsymbol{X}(t_i)|\boldsymbol{X}(t_{i-1}))$. However, in practice this is infeasible because $p^{(M)}(\boldsymbol{X}(t_i)|\boldsymbol{X}(t_{i-1}))$ is unknown.

In order to decrease the variance (13) and reduce the sample size J, we want to choose an importance sampler $q\big(\boldsymbol{X}((t_{i-1}+\delta):(t_i-\delta))\big)$ that is as close as possible to $\prod_{m=1}^{M} p(\boldsymbol{X}(t_{i-1}+m\delta)|\boldsymbol{X}(t_{i-1}+(m-1)\delta))$, which is the principle of choosing the proposal density in importance sampling.

2.3 Importance samplers for simulated maximum likelihood

Here we review three importance samplers for approximating the transition probability density $p^{(M)}(\boldsymbol{X}(t_i)|\boldsymbol{X}(t_{i-1}))$ in (10). These approaches can be used to compute maximum likelihood estimates of the parameters of the SDE model (6) (i.e., simulated maximum likelihood estimation). In Section 2.4 we propose a new penalized simulated maximum likelihood approach which can be used to improve the performance of all three methods described below.

2.3.1 Pedersen sampler

The Pedersen sampler is the first importance sampler proposed to approximate a transition density (Pedersen, 1995b; Santa-Clara, 1997). The Pedersen sampler constructs the importance sampler q by simulating J paths on each subinterval just using the Euler-Maruyama scheme conditional on $\boldsymbol{X}(t_{i-1})$, so the first $M-1$ terms in (12) are canceled. Hence, (11)

reduces to

$$\frac{1}{J}\sum_{j=1}^{J} p^{(1)}(\boldsymbol{X}(t_i)|\boldsymbol{X}^{(j)}(t_i-\delta)). \tag{15}$$

One can simulate J trajectories of all k-dimensional state process $\boldsymbol{X}$ from time t_{i-1} to time $t_i - \delta$ by using the Euler-Maruyama scheme with the step size δ. Although the Pedersen sampler has a very simple form, it is well known that it is computationally intensive in practice (Durham and Gallant, 2002), especially for a multivariate SDE model. The Pedersen sampler can introduce excessive variance in the simulation of all possible transition probabilities even with a very large number of simulated trajectories.

2.3.2 Modified Brownian bridge sampler

A more efficient importance sampler is called the modified Brownian bridge (MBB) sampler, which is originally proposed by Durham and Gallant (2002) for the univariate case and modified by Golightly and Wilkinson (2006) for the multivariate case. Instead of simulating a path on each subinterval using the Euler approximation based on $\boldsymbol{X}(t_{i-1})$ as in Pedersen sampler, this method draws $\boldsymbol{X}((t_{i-1}+\delta) : (t_i - \delta))$ conditional on $\boldsymbol{X}(t_{i-1})$ and $\boldsymbol{X}_{\text{obs}}(t_i)$. Here, we outline the procedure. See Golightly and Wilkinson (2006) for more details.

Let $\boldsymbol{X}^m$ denote $\boldsymbol{X}(t_{i-1}+m\delta)$ and partition the drift and diffusion functions in (6) as

$$f(\boldsymbol{X}) = \begin{pmatrix} f_{-\text{obs}}(\boldsymbol{X}) \\ f_{\text{obs}}(\boldsymbol{X}) \end{pmatrix}$$

and

$$g^T(\boldsymbol{X})g(\boldsymbol{X}) = \begin{bmatrix} G_{-\text{obs},-\text{obs}}(\boldsymbol{X}) & G_{-\text{obs},\text{obs}}(\boldsymbol{X}) \\ G_{\text{obs},-\text{obs}}(\boldsymbol{X}) & G_{\text{obs},\text{obs}}(\boldsymbol{X}) \end{bmatrix}.$$

Then the MBB sampler draws $\boldsymbol{X}^{m+1}$ from the density

$$q(\boldsymbol{X}^{m+1}|\boldsymbol{X}^m, \boldsymbol{X}_{\text{obs}}(t_i)) = \phi(\boldsymbol{X}^{m+1}; \boldsymbol{X}^m + \boldsymbol{\eta}_m\delta, \Sigma_m\delta), \tag{16}$$

where $\phi(\boldsymbol{X}; \boldsymbol{\mu}, \Sigma)$ is a multivariate normal density with mean vector $\boldsymbol{\mu}$ and covariance matrix Σ. Here

$$\boldsymbol{\eta}_m = \begin{pmatrix} f_{-\text{obs}}(\boldsymbol{X}^m) + \frac{G_{-\text{obs},\text{obs}}(\boldsymbol{X}^m)}{\delta(M-m)G_{\text{obs},\text{obs}}(\boldsymbol{X}^m)}\Delta_{\text{obs}} \\ (\boldsymbol{X}_{\text{obs}}(t_i) - \boldsymbol{X}_{\text{obs}}(t_{i-1}+m\delta))/[\delta(M-m)] \end{pmatrix}, \tag{17}$$

and

$$\Sigma_m = \begin{bmatrix} G_{-\text{obs},-\text{obs}}(\boldsymbol{X}^m) - \frac{G_{-\text{obs},\text{obs}}(\boldsymbol{X}^m)G_{\text{obs},-\text{obs}}(\boldsymbol{X}^m)}{(M-m)G_{\text{obs},\text{obs}}(\boldsymbol{X}^m)} & \frac{M-m-1}{M-m}G_{-\text{obs},\text{obs}}(\boldsymbol{X}^m) \\ \frac{M-m-1}{M-m}G_{\text{obs},-\text{obs}}(\boldsymbol{X}^m) & \frac{M-m-1}{M-m}G_{\text{obs},\text{obs}}(\boldsymbol{X}^m) \end{bmatrix}, \tag{18}$$

where

$$\Delta_{\text{obs}} = \boldsymbol{X}_{\text{obs}}(t_i) - [\boldsymbol{X}_{\text{obs}}(t_{i-1}+m\delta) + f_{\text{obs}}(\boldsymbol{X}^m)(M-m-1)\delta]$$

for $m = 0, 1, \cdots, M-2$. For $m = M-1$, we draw $\boldsymbol{X}_{-\text{obs}}(t_i)$ conditional on $\boldsymbol{X}^{M-1} = \boldsymbol{X}(t_i-\delta)$ and $\boldsymbol{X}_{\text{obs}}(t_i)$, which is conditional multivariate normal by the Euler-Maruyama scheme. By recursively drawing from (16) one can obtain a Brownian bridge, $\boldsymbol{X}((t_{i-1}+\delta) : (t_i-\delta))$ conditioned on starting at $\boldsymbol{X}(t_{i-1})$ and finishing at $\boldsymbol{X}_{\text{obs}}(t_i)$.

2.3.3 Regularized sampler

The MBB sampler can produce a poor approximation because its linear interpolation between two observations ignores the dynamics of the model in constructing the sample paths, especially when the diffusion dynamics are dominated by the drift term for sparsely sampled data (Lindström, 2012). A regularized sampler which is a weighted combination of the Pedersen sampler and the MBB sampler is proposed by Lindström (2012) to overcome this limitation. Here we give the explicit form of this regularized sampler.

Let $\boldsymbol{\mu}_P$ and Σ_P be the mean and the variance of the Pedersen sampler and $\boldsymbol{\mu}_M$ and Σ_M be the mean and the variance of the MBB sampler. Then the regularized sampler draws $\boldsymbol{X}^{m+1}$ from the density

$$q_\rho(\boldsymbol{X}^{m+1}|\boldsymbol{X}^m, \boldsymbol{X}_{\text{obs}}(t_i)) = \phi(\boldsymbol{X}^{m+1}; (\mathcal{I} - \boldsymbol{V})\,\boldsymbol{\mu}_P + \boldsymbol{V}\boldsymbol{\mu}_M, (\mathcal{I} - \boldsymbol{V})\,\boldsymbol{\Sigma}_P + \boldsymbol{V}\boldsymbol{\Sigma}_M), \tag{19}$$

where $\mathcal{I}$ is the identity matrix and

$$\boldsymbol{V} = \frac{M - m}{M - m + \rho(M - m - 1)^2}\mathcal{I}, \tag{20}$$

where $\rho \in [0, 1]$. The regularized sampler is dominated by the Pedersen sampler initially and is dominated by the MBB sampler as $m \to (M - 1)$ in (39). The regularized sampler depends on the parameter ρ. A large ρ will make the regularized sampler similar to the Pedersen sampler and a smaller ρ will make it similar to the MBB sampler. Lindström (2012) used $\rho = 0.1$ throughout, however, did not propose an algorithm for choosing the optimal ρ. Hence, a practical algorithm for selecting the optimal ρ is needed for successful implementation of the regularized sampler. We propose one such approach in the next section.

2.4 Penalized simulated maximum likelihood and auxiliary importance sampling

To find an efficient importance sampler, we need to minimize (13), the variance of the approximation of the transition density. Here we propose a new approach to minimize the variance; (i) we augment the likelihood with an auxiliary parameter ρ which tunes the importance sampler to the model parameters and (ii) we maximize the log likelihood with a constraint on the coefficient of variation of the importance sampler.

2.4.1 Penalized simulated maximum likelihood

In our penalized simulated maximum likelihood (PSML) approach, we maximize the log likelihood subject to the sum of the coefficient of variation of the Monte Carlo approximation of the transition density being less than a prespecified level. Suppose a family of auxiliary importance samplers $\{q_\rho\}$ has been selected, where ρ is an auxiliary or nuisance parameter. Our goal is to find $\hat{\rho}$ that minimizes the sum of the coefficient of variation of the Monte Carlo approximation of the transition density.

Let h_ρ be the importance sampling weights to approximate $p(\boldsymbol{X}_{\text{obs}}(t_i)|\boldsymbol{X}(t_0), \boldsymbol{X}_{\text{obs}}(t_1 : t_{i-1}))$ in (6). Specifically,

$$h_\rho\Big(\boldsymbol{X}^{(j)}(t_{i-1} : (t_i - \delta))\Big) \equiv \frac{p^{(1)}(\boldsymbol{X}_{\text{obs}}(t_i)|\boldsymbol{X}^{(j)}(t_i - \delta)) \prod_{m=1}^{M-1} p^{(1)}(\boldsymbol{X}^{(j)}(t_{i-1} + m\delta)|\boldsymbol{X}^{(j)}(t_{i-1} + (m-1)\delta))}{q_\rho(\boldsymbol{X}^{(j)}(t_{i-1} : (t_i - \delta)))}, \quad (21)$$

where $\boldsymbol{X}^{(j)}(t_{i-1}) \equiv \{\boldsymbol{X}^{(j)}_{-\text{obs}}(t_{i-1}), \boldsymbol{X}_{\text{obs}}(t_{i-1})\}$ and q_ρ is the importance sampler density, e.g., (19) and (25) below. We adopt the notation h_ρ to indicate the expression in (21), suppressing the dependence on i and j for notational simplicity. The PSML estimator $(\hat{\boldsymbol{\theta}}, \hat{\rho})$ is defined

by

$$(\hat{\boldsymbol{\theta}}, \hat{\rho}) = \text{arg max} \sum_{i=1}^{n} \log \left(\frac{1}{J} \sum_{j=1}^{J} h_\rho \right) \text{ subject to } \sum_{i=1}^{n} \widehat{\text{cv}}\,(h_\rho) \leq s, \tag{22}$$

where $s \geq 0$ is a tuning parameter and $\widehat{\text{cv}}(h_\rho)$ is the sample coefficient of variation of h_ρ, which is the sample standard deviation of the J importance weights h_ρ divided by their sample mean. Notice that (22) is reminiscent of LASSO (Tibshirani, 1996), and is equivalent to maximizing a penalized log likelihood,

$$l^*(\boldsymbol{\theta}, \rho) = \sum_{i=1}^{n} \log \left(\frac{1}{J} \sum_{j=1}^{J} h_\rho \right) - \lambda \sum_{i=1}^{n} \widehat{\text{cv}}(h_\rho), \tag{23}$$

where λ in (23) has a one-to-one mapping to s in (22). The reason the coefficient of variation is chosen instead of the variance is that the former is a normalized measurement which is not affected by the magnitude of the data. This makes it easier to choose the tuning parameter λ in practice, as will be shown below. When the penalty term is omitted, the parameter estimates have a large variance because the importance sampler is not well tuned.

The constraint, $\sum_{i=1}^{n} \widehat{\text{cv}}(h_\rho) \leq s$ in (22), is equivalent to a constraint on the effective sample size (Givens and Hoeting, 2012, Chapter 6),

$$\widehat{N}(q_\rho, p) \equiv \frac{J}{1 + \frac{1}{n} \sum_{i=1}^{n} \hat{\text{cv}}^2(h_\rho)} \geq \frac{J}{1 + \frac{s^2}{n}}.$$

The effective sample size measures how much the auxiliary importance sampler density q_ρ differs from the target density p, and it can be interpreted as J weighted samples are worth $\widehat{N}(q_\rho, p)$ unweighted i.i.d. samples drawn exactly from target density p. Effective sample size can be used as a measure of computational efficiency.

The tuning parameter s controls how close the auxiliary importance sampler density q_ρ is to the product of transition probability densities, the numerator of (12). Let s^0 denote the sum of the coefficient of variation for the approximation of the transition density by the

Pedersen sampler. When $s < s^0$ the resulting auxiliary importance sampler will have smaller variance (13) than that from the Pedersen sampler (15). When $s = 0$, the constraint in (22) requires that the auxiliary importance sampler q_ρ attains its ideal case (14). However, as $s \to 0$, $\lambda \to \infty$ and therefore the log likelihood plays no role in estimating $\boldsymbol{\theta}$.

The tuning parameter λ in (23) can be estimated using various techniques. We choose the value that minimizes the estimated prediction error,

$$\epsilon_\lambda \equiv \frac{1}{nL} \sum_{\ell=1}^{L} \sum_{i=1}^{n} \| \widehat{\boldsymbol{X}}_{\text{obs}}^{(\ell)}(t_i) - \boldsymbol{X}_{\text{obs}}(t_i) \|, \tag{24}$$

where $\widehat{\boldsymbol{X}}_{\text{obs}}^{(\ell)}(t_i)$ is the ℓth simulated $\boldsymbol{X}_{\text{obs}}$ at observation time t_i by the Euler-Maruyama scheme (9) with $\boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}(\lambda)$ and $\|\boldsymbol{X}\|$ is a Euclidean norm of $\boldsymbol{X}$ in $\mathbb{R}^k$. The number of simulations L is chosen arbitrarily and is set to 1000 here. More details about selecting λ are given in Algorithm 1 in Section 2.4.3.

2.4.2 Auxiliary importance sampling

The first class of importance samplers with auxiliary parameter ρ in (21) is given by

$$q_\rho(\boldsymbol{X}^{m+1} | \boldsymbol{X}^m, \boldsymbol{X}_{\text{obs}}(t_i)) = \phi(\boldsymbol{X}^{m+1}; \boldsymbol{X}^m + \boldsymbol{\eta}_m \delta, \rho \Sigma_m \delta), \tag{25}$$

where $\boldsymbol{\eta}_m, \Sigma_m$ are defined in (17) and (18). Hence, $\rho \in [0, 1]$ is the shrinkage coefficient, which will be estimated as an auxiliary parameter in the penalized log likelihood (23). The penalty term in (22) allows estimation of the auxiliary parameter ρ, which is a feature of PSML, and leads to improved performance over the MBB and regularized sampler as will be illustrated in Section 5.5. Note that the MBB sampler is a special case of our auxiliary importance sampler (25) when $\lambda = 0$ and $\rho = 1$.

We consider the regularized sampler (19) as another class of auxiliary importance samplers with auxiliary parameter ρ. The optimal choice of ρ can be determined by maximizing

the penalized log likelihood (22). As will be shown below in Section 5.5, this leads to improved performance of the regularized sampler as compared to fixing $\rho = 0.1$ as in Lindström (2012).

One can also choose other families of auxiliary importance samplers, but the two classes considered above are a good starting point for illustration of the method. Other distributions, such as the Student's t distribution, might also be a suitable choice.

2.4.3 Algorithm for PSML

The Algorithm for penalized simulated maximum likelihood estimation is given in Algorithm 1. We consider two stopping criteria for this algorithm, ϵ_0 and δ_ϵ. First, ϵ_0 monitors the estimated prediction error ϵ_λ in (24). If the estimated prediction error is sufficiently small, $\epsilon_\lambda \leq \epsilon_0$, then there is no need to tune λ and the algorithm stops. The criterion δ_ϵ monitors the change in the estimated prediction error. If the improvement is small or there is no improvement at all, that is $\epsilon_\lambda - \epsilon_{\lambda_*} \leq \delta_\epsilon$, then the algorithm stops.

For Step 1, we find that the initial value $\lambda_0 \in (0.1, 0.5)$ works well for our models considered in Section 5.5. The values ϵ_0 and δ_ϵ are data dependent. The parameter δ_λ is the step size for exploring the space of λ values. We use $\delta_\lambda = 0.025$.

Note that, although this procedure looks computationally intensive, the algorithm converges quickly and is robust to the choice of λ_0 (as will be illustrated in the simulation studies in Section 5.5). Based on our simulation studies, we find that the first three steps, Steps 1 to 3, in the procedure are already sufficient to gain an improvement over the MBB sampler or the regularized sampler.

Note that Algorithm 1 can be extended to a parallel procedure by repeating Step 2 through a grid search with grid width δ_λ on the interval $\lambda \in (0, c)$, where c is a constant. In our experience $\lambda \in (0, 1)$ is reasonable. In this case, the parameter estimates $(\hat{\boldsymbol{\theta}}, \hat{\rho})$ that correspond to the smallest ϵ_λ would be the output.

Algorithm 1: Algorithm for penalized simulated maximum likelihood estimation.

Step 1. Pick $\lambda_0 > 0$, $\epsilon_0 > 0$, $\delta_\lambda > 0$, and $\delta_\epsilon > 0$. Let $\lambda = \lambda_0$.

Step 2. Find the maximizer $(\hat{\boldsymbol{\theta}}, \hat{\rho})$ of the penalized log likelihood in (23). Compute the estimated prediction error ϵ_λ in (24).

Step 3. If $\epsilon_\lambda < \epsilon_0$ then stop, otherwise go to Step 4.

Step 4. Let $\lambda_* = \lambda - \delta_\lambda$. Compute ϵ_{λ_*}.

Step 5. If $\epsilon_\lambda - \epsilon_{\lambda_*} > \delta_\epsilon$ then update $\lambda = \lambda_*$ and go back to Step 3, otherwise go to Step 6.

Step 6. If $\lambda < \lambda_0$ (i.e. λ was updated from the initial λ_0) then stop, otherwise go to Step 7.

Step 7. If $\epsilon_\lambda < \epsilon_0$ then stop, otherwise go to Step 8.

Step 8. Let $\lambda_* = \lambda + \delta_\lambda$. Compute ϵ_{λ_*} as in Step 2.

Step 9. If $\epsilon_\lambda - \epsilon_{\lambda_*} > \delta_\epsilon$ then update $\lambda = \lambda_*$ and go back to Step 7, otherwise stop.

We use the parametric bootstrap (Efron, 1982, Chapter 5) to obtain confidence intervals for the estimator, which proceeds as follows. First, based on the parameter estimates from the original dataset of interest, we can generate a large number of new *datasets* by using the Euler-Maruyama scheme for the SDE model (9). For each new simulated dataset, we obtain estimates of parameters using the PSML method described in Algorithm 1. Then we compute the confidence interval from those estimates using the corresponding quantiles.

2.5 Simulation studies

Here, we compare the performances of the MBB sampler, the regularized sampler with $\rho = 0.1$, and our PSML with the modified MBB class (25) and the regularized class (19) on simulated datasets for three different models. We refer to PSML with the modified MBB class (25) as PSML-MBB and refer to PSML with the regularized class (19) as PSML-Reg. For all the optimization algorithms in this chapter, we use an implementation of the Nelder-

Mead algorithm for derivative-free optimization (Varadhan and Borchers, 2011) in R (R Development Core Team, 2011) on an Intel Xeon W3565 3.2 GHz with CentOS 6 computer. The iterations for optimization of Step 2 of Algorithm 1 will stop when the absolute difference in function values between successive iterations is below 10^{-6}, which is the default value in the R dfoptim package. We also use the default value for the maximum number of objective function evaluations allowed, which is 1500 for all three models below. The initial values for the parameters are chosen arbitrarily (we tried different initial values and obtained similar results). No parallel algorithm is involved in all the reported computation times. The time to compute the confidence intervals is not included.

2.5.1 Ornstein-Uhlenbeck process

We first consider a univariate SDE, the Ornstein - Uhlenbeck process

$$dX = (\theta_1 - \theta_2 X)dt + \theta_3 dW, \tag{26}$$

with known initial condition $X(t_0)$, and the parameter $\boldsymbol{\theta} = (\theta_1, \theta_2, \theta_3) \in \mathbb{R} \times \mathbb{R}_+ \times \mathbb{R}_+$. The parameter θ_2 is the speed of reversion, θ_1/θ_2 is the long-run equilibrium value of the process, and θ_3 is interpreted as the volatility. We generate 100 datasets, each including 100 observations, with initial condition $X(t_0) = 1$, the time interval $t_i - t_{i-1} = 1$, and parameter $\boldsymbol{\theta}_0 = (0.0187, 0.2610, 0.0224)$ as reported in Aït-Sahalia (2002).

The transition density between two observations is given by

$$X(t_{i+1})|X(t_i) \sim N\left(\frac{\theta_1}{\theta_2} + \left(X(t_i) - \frac{\theta_1}{\theta_2}\right)e^{-\theta_2\Delta}, \frac{\theta_3^2(1 - e^{-2\theta_2\Delta})}{2\theta_2}\right),$$

where $\Delta = t_{i+1} - t_i$ for $i = 1, \ldots, n-1$ (Iacus, 2009, Chapter 3). Hence, the exact likelihood is known for this case and we can obtain the exact maximum likelihood estimator of the parameters $\boldsymbol{\theta}$. We compute the bias and the root mean square error (RMSE) of the simulated maximum likelihood estimators $\widehat{\boldsymbol{\theta}}_r$ with respect to the exact maximum likelihood estimators

Table 2.1: The bias and RMSE of the simulated maximum likelihood estimates with respect to the exact maximum likelihood estimates for the Ornstein-Uhlenbeck process (26). All results are multiplied by 10^4. Both PSML-MBB and PSML-Reg have better performance than the MBB sampler and the regularized sampler in terms of reducing bias and RMSE, especially when the number of sample paths is small ($J = 8$).

		$J = 8$			$J = 16$		
	Method	θ_1	θ_2	θ_3	θ_1	θ_2	θ_3
Bias ($\times 10^{-4}$)	MBB	46	382	35	17	121	12
	Regularized	12	85	8	7	37	4
	PSML-MBB	-6	-69	-5	-8	-77	-5
	PSML-Reg	-2	2	3	-5	-39	5
RMSE ($\times 10^{-4}$)	MBB	115	982	93	74	585	57
	Regularized	67	505	49	52	406	43
	PSML-MBB	16	114	10	15	105	8
	PSML-Reg	21	142	15	22	135	13

$\widehat{\boldsymbol{\theta}}_{\text{MLE}}$, defined by $\frac{1}{100}\sum_{r=1}^{100}(\widehat{\boldsymbol{\theta}}_r - \widehat{\boldsymbol{\theta}}_{\text{MLE}})$ and $\sqrt{\frac{1}{100}\sum_{r=1}^{100}(\widehat{\boldsymbol{\theta}}_r - \widehat{\boldsymbol{\theta}}_{\text{MLE}})^2}$, respectively. For all the methods, we consider $M = 8$ subintervals but with different levels of the number of simulated sample paths J. We set the $\epsilon_0 = 0.04$, $\lambda_0 = 0.25$, $\delta_\lambda = 0.025$ and $\delta_\epsilon = 0.001$. The initial values for optimization for θ_1, θ_2, and θ_3 are $0.05, 0.5$ and 0.05, respectively.

Table 2.1 shows that both PSML-MBB and PSML-Reg have better performance than the MBB sampler and the regularized sampler in terms of reducing bias and RMSE, especially when the number of sample paths is small ($J = 8$). We find that more accurate estimates can be achieved by introducing the penalty term in the PSML and selecting the optimal ρ for the regularized class, which is in contrast to the fixed ρ case for the regularized sampler as studied in Lindström (2012).

Figure 2.1 shows that some estimates are far away from the exact maximum likelihood estimates for the MBB sampler and the regularized sampler. This typically happens when the estimates based on the log likelihood approximated by the MBB sampler or the regularized sampler get stuck at the local maxima for optimization. This may be an indication of the poor approximation of the likelihood, since as the number of sample paths J increases, fewer estimates have large bias. For a small J a poor choice of proposal distribution, e.g., the

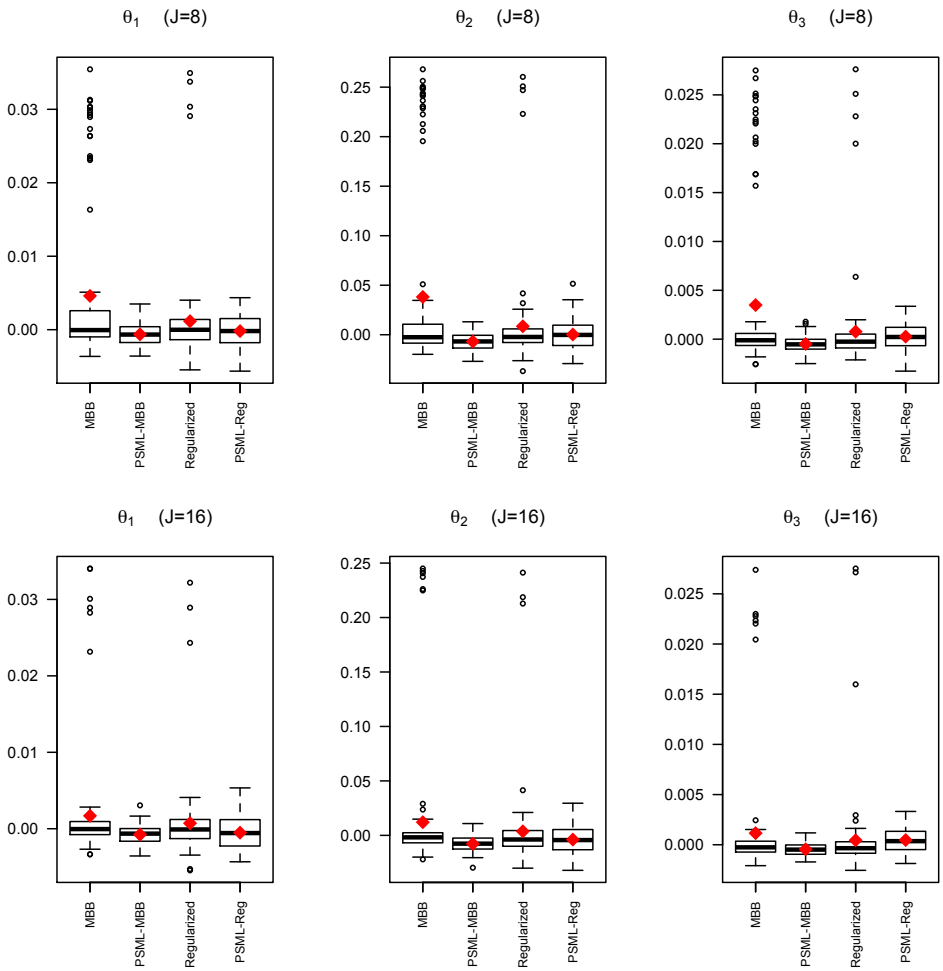

Figure 2.1: Boxplot of the bias of 100 estimates with $J = 8$ and $J = 16$ for the Ornstein-Uhlenbeck process (26). The red bold diamond points are the mean. Note that the PSML greatly reduces the Monte Carlo variability of the estimates.

MBB and the regularized sampler, could result in a poor approximation to the likelihood because the approximated likelihood surface is generally more wiggly when J is small. Thus when J is small it is more common for these methods to mistakenly select a local maximum. The proposed PSML-MBB and PSML-Reg have better performance in this regard.

Figure 2.2 indicates that the performance of PSML is robust to the choice of λ_0. The improvements of the PSML-MBB with various λ_0's over the MBB sampler are similar. This makes the algorithm easy to implement in practice. Clearly, when J is small the difference between the PSML-MBB and the MBB is very large. As the number of sample paths J increases, the difference between the MBB sampler with the PSML-MBB decreases. However, this is not always true for other SDE models. See the Lorenz model (27) in the next section for more details.

To obtain a similar level of accuracy as the PSML-MBB with $J = 8$, the MBB sampler requires at least $J = 96$. However, the PSML-MBB with $J = 8$ requires much less time (around 1/5) than the MBB with $J = 96$. For the computation time, the PSML with $J = 8$ takes 90 – 110 seconds and 180 – 200 seconds for $J = 16$ (for both PSML classes in Section 2.4.2). The MBB sampler or the regularized sampler takes 65 – 75 seconds to implement for $J = 8$, 120 – 140 seconds for $J = 16$, and 750 – 950 seconds for $J = 96$. The computational time grows approximating linearly in J for both algorithms.

For the PSML-MBB with $\lambda_0 = 0.25$, the mean of $\hat{\rho}$ equals 0.94 and the mean of $\hat{\lambda}$ equals 0.24. Note that though $\hat{\rho}$ is close to 1 for the PSML-MBB, the MBB is equivalent to the PSML-MBB only when $\rho = 1$ *and* $\lambda = 0$. For the PSML-Reg with $\lambda_0 = 0.25$, the mean of $\hat{\rho}$ equals 0.33 and the mean of $\hat{\lambda}$ equals 0.23. We note that the performance of PSML-Reg presented in Tables 2.1 – 2.3 is based on the estimated $\hat{\rho}$. (We fix $\rho = 0.1$ for the regularized sampler as in Lindström (2012).) This indicates the regularized sampler can be improved when ρ is estimated in (19). We have observed similar $\hat{\rho}$ values for the other models in the simulation studies considered in this section.

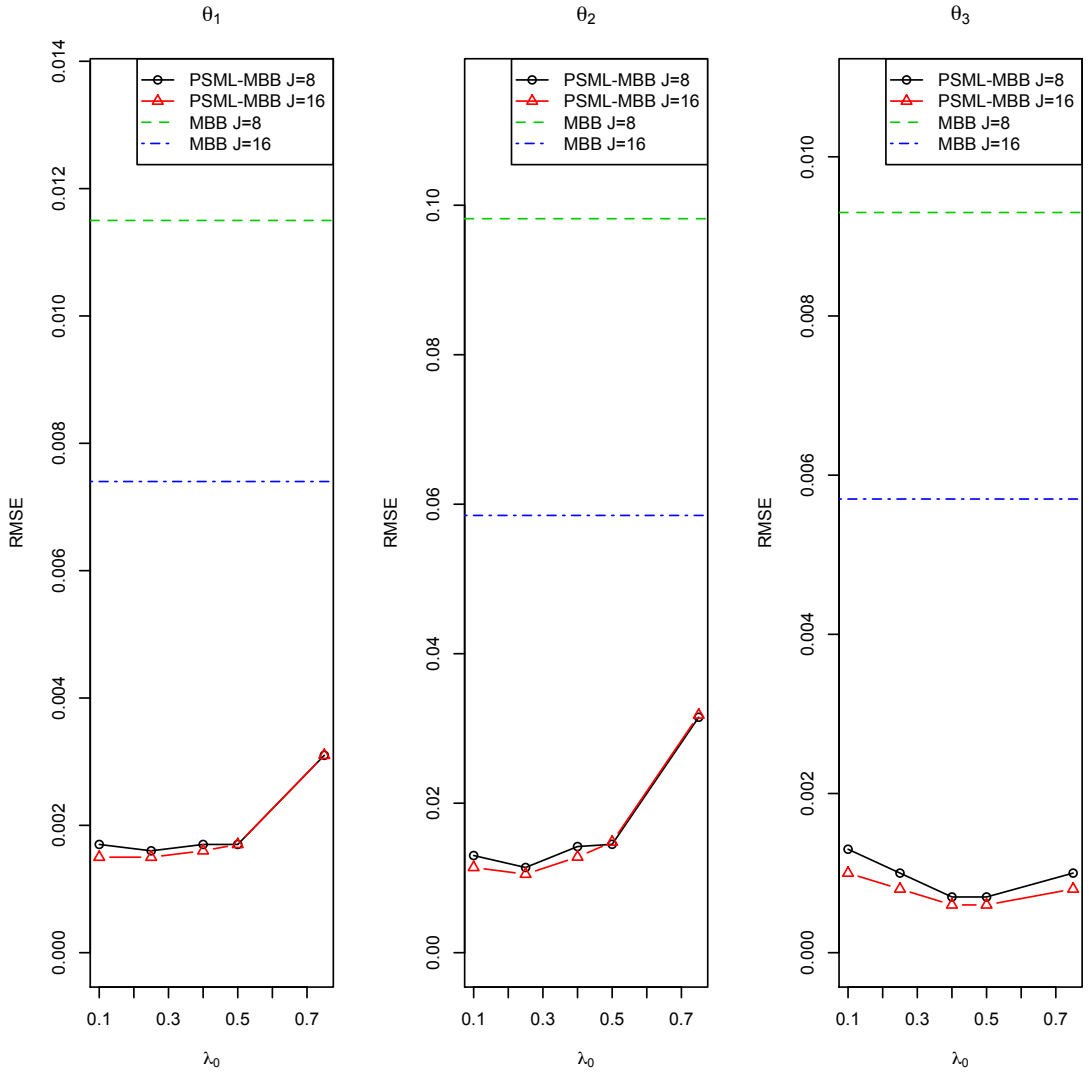

Figure 2.2: The RMSE of the MBB and PSML-MBB estimates with different λ_0 for the Ornstein-Uhlenbeck process (26).

2.5.2 Stochastic Lorenz 63 model

Next, we consider the stochastic version of the well-known chaotic Lorenz 63 model (Lorenz, 1963; Bengtsson et al., 2003), which is given by

$$d\begin{pmatrix} X_1 \\ X_2 \\ X_3 \end{pmatrix} = \begin{pmatrix} s(X_2 - X_1) \\ rX_1 - X_2 - X_1X_3 \\ X_1X_2 - bX_3 \end{pmatrix} dt + \sigma d \begin{pmatrix} W_1 \\ W_2 \\ W_3 \end{pmatrix}, \tag{27}$$

where W_1, W_2, and W_3 are three independent Wiener processes.

We again generate 100 datasets, each including 21 observations, with initial condition $(-10, -10, 30)$, time interval $t_i - t_{i-1} = 0.05$, and commonly used parameter values $\boldsymbol{\theta}_0 = (s_0 = 10, r_0 = 28, b_0 = 8/3, \sigma_0 = 2)$ (Bengtsson et al., 2003). We assume all state variables, $(X_1, X_2, X_3)^T$, are observed at t_i for $i = 0, \ldots, n$. In this case, the exact transition density is no longer available. We can only compute the bias and the RMSE of the simulated maximum likelihood estimators $\widehat{\boldsymbol{\theta}}_r$ with respect to the true parameters $\boldsymbol{\theta}_0$, defined by $\frac{1}{100}\sum_{r=1}^{100}(\widehat{\boldsymbol{\theta}}_r - \boldsymbol{\theta}_0)$ and $\sqrt{\frac{1}{100}\sum_{r=1}^{100}(\widehat{\boldsymbol{\theta}}_r - \boldsymbol{\theta}_0)^2}$, respectively. Estimates are obtained using different sample paths J and $M = 10$ subintervals. We set $\lambda_0 = 0.5$, $\epsilon_0 = 3.5$, $\delta_\lambda = 0.025$ and $\delta_\epsilon = 0.1$. The initial values for optimization are $(15, 30, 5, 1)$.

As shown in Table 2.2 the MBB sampler performs poorly for the stochastic Lorenz 63 model (27), especially for the parameters r, b, and σ as Lindström (2012) has observed. This is because the dynamics of the Lorenz model is dominated by the drift term. The MBB sampler ignores the dynamics of the model and generates paths far from the actual realization. Moreover, Figure 2.3 shows that there is no significant increasing trend in the accuracy of the regularized sampler as the number of sample paths J increases, especially for parameter s. A large J but a fixed ρ, which controls the weight between the Pedersen and the MBB sampler, still cannot assure the regularized sampler generates paths close to the actual trajectories. However, the improvement of the PSML-Reg over the MBB and the

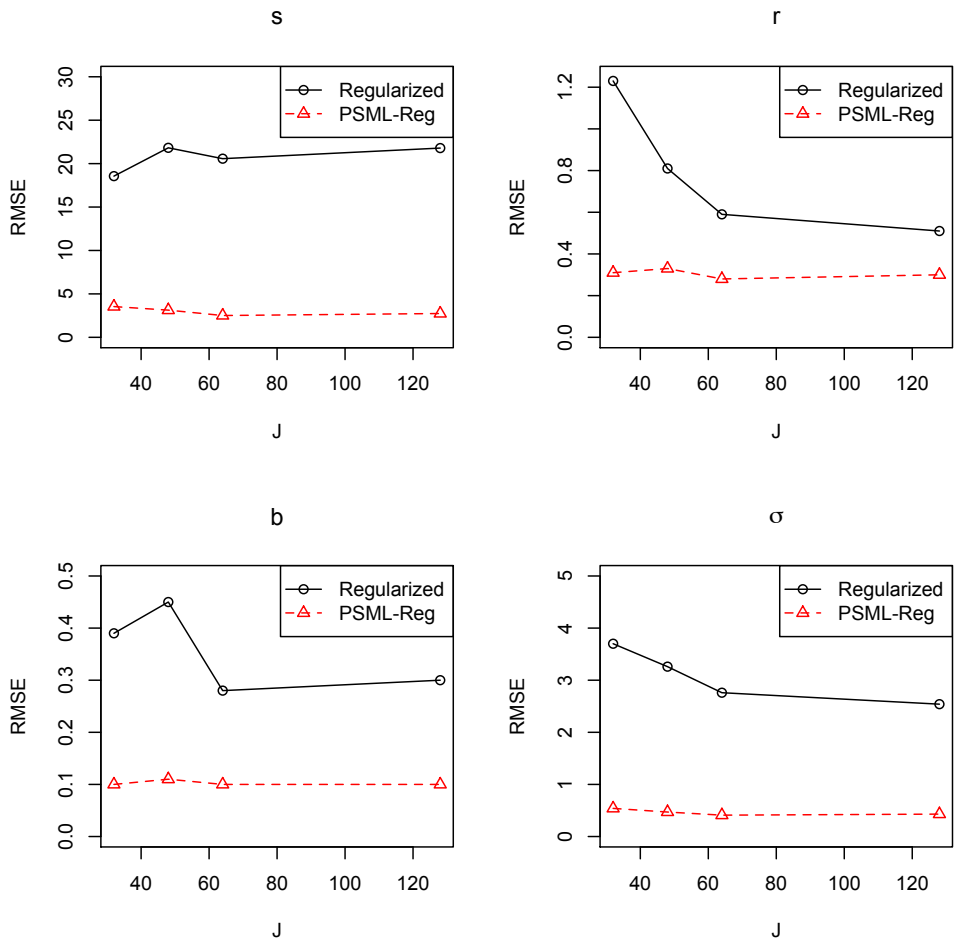

Figure 2.3: The RMSE of the regularized and the PSML-Reg estimates with different J for the stochastic Lorenz 63 model (27).

Table 2.2: The bias and RMSE of the simulated maximum likelihood estimates with respect to the true parameters for the stochastic Lorenz 63 model (27). The improvement of the PSML-Reg over the MBB and the regularized sampler with different J's is evident.

	Method	s	r	b	σ
Bias	MBB ($J = 128$)	1.86	−15.70	8.92	9.59
	Regularized ($J = 32$)	2.59	−0.25	−0.03	1.54
	Regularized ($J = 48$)	3.79	−0.07	−0.06	1.25
	Regularized ($J = 64$)	3.86	−0.07	0.00	1.17
	Regularized ($J = 128$)	3.89	−0.02	−0.03	0.99
	PSML-Reg ($J = 32$)	−1.75	0.00	0.01	0.38
RMSE	MBB ($J = 128$)	13.31	20.70	15.52	15.36
	Regularized ($J = 32$)	18.56	1.23	0.39	3.70
	Regularized ($J = 48$)	21.82	0.81	0.45	3.26
	Regularized ($J = 64$)	20.57	0.59	0.28	2.76
	Regularized ($J = 128$)	21.79	0.51	0.30	2.54
	PSML-Reg ($J = 32$)	3.54	0.31	0.10	0.54

regularized sampler is evident. An estimated $\hat{\rho}$ based on the data in the PSML-Reg plays an important role in generating efficient proposal trajectories.

In terms of computational time, both the MBB and the regularized samples need 1800 – 2100 seconds for $J = 32$, 2700 – 3000 seconds for $J = 48$, 3500 – 4000 seconds for $J = 64$, and 7000 – 8000 seconds for $J = 128$. The PSML-MBB with $J = 32$ requires 2000 – 2300 seconds.

2.5.3 CWD direct transmission model

The specific model and background are described in Section 6. Again, we generate 100 datasets, each including 21 annual observations from two distinct CWD epidemics similar to the real dataset in Section 6, by using the CWD direct transmission model (28) with parameter $(\beta_0 = 0.03, \mu_0 = 0.20)$. The initial condition $\boldsymbol{X}(t_0) = (S(t_0), I(t_0), C(t_0))^T$ is set to be the same as the real dataset. The step size of the Euler-Maruyama scheme is 1/12 of the time interval between each pair of observations, which is one month in this case. We set $\lambda_0 = 0.5$, $\epsilon_0 = 5$, $\delta_\lambda = 0.025$ and $\delta_\epsilon = 0.5$. The initial values for optimization for β and μ are 0.05 and 0.5, respectively. Parameter estimates are obtained using $J = 72$ sample paths

Table 2.3: The bias and RMSE of the simulated maximum likelihood estimates with respect to the true parameters for CWD direct transmission model (28). Both PSML-MBB and PSML-Reg have better performance than the MBB sampler and the regularized sampler.

	Method	β	μ
Bias	MBB ($J = 72$)	0.02	0.07
	Regularized ($J = 72$)	0.01	0.07
	PSML-MBB ($J = 48$)	0.01	0.02
	PSML-Reg ($J = 48$)	0.01	0.04
RMSE	MBB ($J = 72$)	0.07	0.11
	Regularized ($J = 72$)	0.04	0.12
	PSML-MBB ($J = 48$)	0.02	0.05
	PSML-Reg ($J = 48$)	0.02	0.06

for the MBB sampler and the regularized sampler, which require 2400 – 2700 seconds, and $J = 48$ for the PSML-MBB and the PSML-Reg, which require 2000 – 2300 seconds. The exact transition density is not available for this case. The bias and RMSE of the simulated maximum likelihood estimates with respect to the true parameters are shown in Table 2.3, which indicate similar improvements of the PSML-MBB and the PSML-Reg over the MBB sampler and the regularized sampler. For this simulation the states S and I are unobserved and the time between observations is long (yearly). The fact that the PSML does well in this context is promising for this and other applications in ecology. Since the best results are obtained by the PSML-MBB with $J = 48$ sample paths, we use the same setting in the real data example in Section 2.6.

2.6 Chronic wasting disease example

Deer populations and ecosystems can be severely disrupted by the contagious prion disease, known as chronic wasting disease (CWD) (Miller et al., 2006). In order to reduce the potential damages caused by CWD, it is important to understand the transmission mechanisms of CWD. Several deterministic epidemic models were proposed by Miller et al. (2006) in order to portray the transmission of CWD. Here, based on one of those deterministic models, we firstly derive a CWD SDE model using the technique described in Allen (2003,

Chapter 8). Then, we implement the proposed PSML method to the dataset studied in Miller et al. (2006). Their dataset consists of annual observations of cumulative mortality from two distinct CWD epidemics (Figure 2.4 upper display) in captive mule deer held at the Colorado Division of Wildlife Foothills Wildlife Research Facility in Fort Collins, Colorado. The first epidemic occurred from 1974 to 1985 and the second epidemic occurred in a new deer herd from 1992 to 2001. The dataset also includes the annual number of new deer added to the herd and the per capita losses due to natural deaths and removals. We note that the dataset contains no measurement or observation error since it was recorded in a captive laboratory facility. We assume the direct transmission coefficient β and the per capita CWD mortality rate μ do not change between two epidemics as such parameters are innate characteristics of the associated disease. Hence we can combine two epidemics as a single dataset for estimating the parameters.

2.6.1 CWD direct transmission model

CWD may be transmitted to susceptible animals directly from infected animals (Miller and Williams, 2003). We portray this direct transmission using an SDE model. Let $\boldsymbol{X}(t) = (S(t), I(t), C(t))^T$, where S is the number of susceptible animals, I is the number of infected animals, C is the total number of accumulate deaths from CWD over time. We assume the initial condition $\boldsymbol{X}(t_0) = (S(t_0), I(t_0), C(t_0))^T$ is known. Also, our basic assumption is that only C can be observed at t_i, for $i = 1, \ldots, n$, and the other two state variables, S and I, are unobserved. The unknown parameters to be estimated in the epidemic model are denoted by $\boldsymbol{\theta} = (\beta, \mu)$, where β is the direct transmission coefficient (unit = time^{-1}), μ is the per capita CWD mortality rate (unit = time^{-1}). Then the direct transmission SDE model is

given by

$$d\begin{pmatrix} S \\ I \\ C \end{pmatrix} = \begin{pmatrix} a - S(\beta I + m) \\ \beta SI - I(\mu + m) \\ \mu I \end{pmatrix} dt + \boldsymbol{B} d\boldsymbol{W} \tag{28}$$

where a is the known number of susceptible animals annually added to the population via births or importation, m is the known per capita natural mortality rate, $\boldsymbol{W} = (W_1, W_2, W_3)^T$ is a 3-dimensional standard Wiener process, and $\boldsymbol{B} = \sqrt{\Sigma}$ is the positive definite square root of the covariance matrix with

$$\Sigma = \begin{bmatrix} a + S(\beta I + m) & -\beta SI & 0 \\ -\beta SI & \beta SI + I(\mu + m) & -\mu I \\ 0 & -\mu I & \mu I \end{bmatrix}. \tag{29}$$

Although the SDE model (28) relaxes the assumption of discrete states and non-negative nature of S, I, and dC, similar SDE models have been used to approximate the transmissions of epidemics in several recent articles (Ionides et al., 2006; Bhadra et al., 2011; Golightly and Wilkinson, 2011). We also monitor the frequency of negative estimates in S, I, and dC; they were rare to the point of negligibility in our analysis.

Here, we briefly explain how the above SDE model is derived. See Allen (2003, Chapter 8) for more details. Let $\boldsymbol{X}_\delta = \boldsymbol{X}(t + \delta) - \boldsymbol{X}(t)$ be the increment during the time interval δ. If δ is sufficiently small, we can assume at most one animal is infected or died during the time interval δ. The probability of an event that more than one infection or death has occurred during time δ is of order δ^2, which can be neglected. Then we can approximate the mean of

$\boldsymbol{X}_\delta$ for δ sufficiently small to order δ by

$$E[\boldsymbol{X}_\delta] \approx f\delta = \begin{pmatrix} a - S(\beta I + m) \\ \beta SI - I(\mu + m) \\ \mu I \end{pmatrix} \delta. \tag{30}$$

Furthermore, we can also approximate the covariance of $\boldsymbol{X}_\delta$ for δ sufficiently small by

$$V[\boldsymbol{X}_\delta] = E[(\boldsymbol{X}_\delta)(\boldsymbol{X}_\delta)^T] - E(\boldsymbol{X}_\delta)E(\boldsymbol{X}_\delta)^T \approx E[(\boldsymbol{X}_\delta)(\boldsymbol{X}_\delta)^T] = \Sigma\delta. \tag{31}$$

The matrix Σ in (5.3.2) is positive definite and hence has a positive definite square root $\boldsymbol{B} = \sqrt{\Sigma}$. It can be shown that (30) and (31) are quantities of order δ. We also assume $\boldsymbol{X}_\delta$ follows normal distribution with mean vector $f\delta$ and covariance matrix $\boldsymbol{B}^2\delta = \Sigma\delta$. Thus,

$$\boldsymbol{X}(t+\delta) \approx \boldsymbol{X}(t) + f\delta + \boldsymbol{B}\sqrt{\delta}\boldsymbol{\eta}, \tag{32}$$

where $\boldsymbol{\eta} \sim N(0, \boldsymbol{\mathcal{I}}_{3\times 3})$ and $\boldsymbol{\mathcal{I}}$ is the identity matrix. This is exactly one iteration of the Euler-Maruyama scheme for a system of SDEs (28). As a result, the dynamical system (63) converges in the mean square sense to the system of SDEs (28) as $\delta \to 0$.

2.6.2 Results

The simulated maximum likelihood estimates based on the PSML-MBB with $J = 48$, and $M = 12$ are $\widehat{\boldsymbol{\theta}}_{\text{PSML-MBB}} = (\hat{\beta}, \hat{\mu}) = (0.03, 0.21)$ (unit = year^{-1}) with 95% confidence intervals $[0.027, 0.120]$ and $[0.143, 0.388]$, respectively, and $\hat{\rho} = 0.86$. Although Durham and Gallant (2002) did not provide confidence intervals based on the MBB approach or a method to compute them, we use the parametric bootstrap as described in Section 2.4.3 to obtain them. The estimates based on the MBB approach with $J = 48$ are $\widehat{\boldsymbol{\theta}}_{\text{MBB}} = (0.03, 0.27)$ with

95% confidence intervals $[0.027, 0.186]$ and $[0.148, 0.599]$, which are much wider than those from the PSML-MBB.

To measure the goodness of fit, 100 simulated trajectories of cumulative number of deaths for CWD using $\widehat{\boldsymbol{\theta}}_{\text{PSML-MBB}}$ are shown in Figure 2.4. For such a small sample size the estimated parameters from the PSML-MBB and the CWD direct transmission model capture the pattern of the CWD death data over time. The fit for the second epidemic is not as good as the first epidemic because we are estimating parameters (μ and β, which remain unchanged between epidemics) from a theoretical SDE model (28), not estimating a least squares fit to the observed data. The theoretical model does a remarkably good job at following the observed data. A non-parametric model would likely provide a close fit to the data in Figure 2.4 but would not provide the scientifically relevant interpretation sought by biologists. Miller et al. (2006) proposed a more complex deterministic model, which we could also extend to a corresponding stochastic model, however the model quickly becomes over-parameterized due to the limited sample size and complexity of the model. Therefore, we only consider the direct transmission model.

The basic reproductive number R_0, which is the average number of secondary cases generated by one infected individual over the course of its infectious period when the entire population is susceptible, is important in biology and epidemiology (Anderson and May, 1992). Usually people consider the situation in which the majority of a closed population is susceptible, that is $S(t_0)/N \approx 1$. For deterministic models, if $R_0 > 1$ then the infection will be spread in a population, and if $R_0 \leq 1$, the infection will die out monotonically. For stochastic models, the probability that there is no epidemic equals 1 if $R_0 \leq 1$ and $(\frac{1}{R_0})^{I(t_0)}$ if $R_0 > 1$ (Allen and Burgin, 2000), where $I(t_0)$ is the initial number of infected animals. The traditional interpretation of R_0 is not available here because the population is not closed; a in (28) is the known number of susceptible annually added to the population. However, we want to point out that our method can be used to estimate R_0 for cases when the population is closed and the other assumptions of R_0 hold. For example, assuming a natural mortality

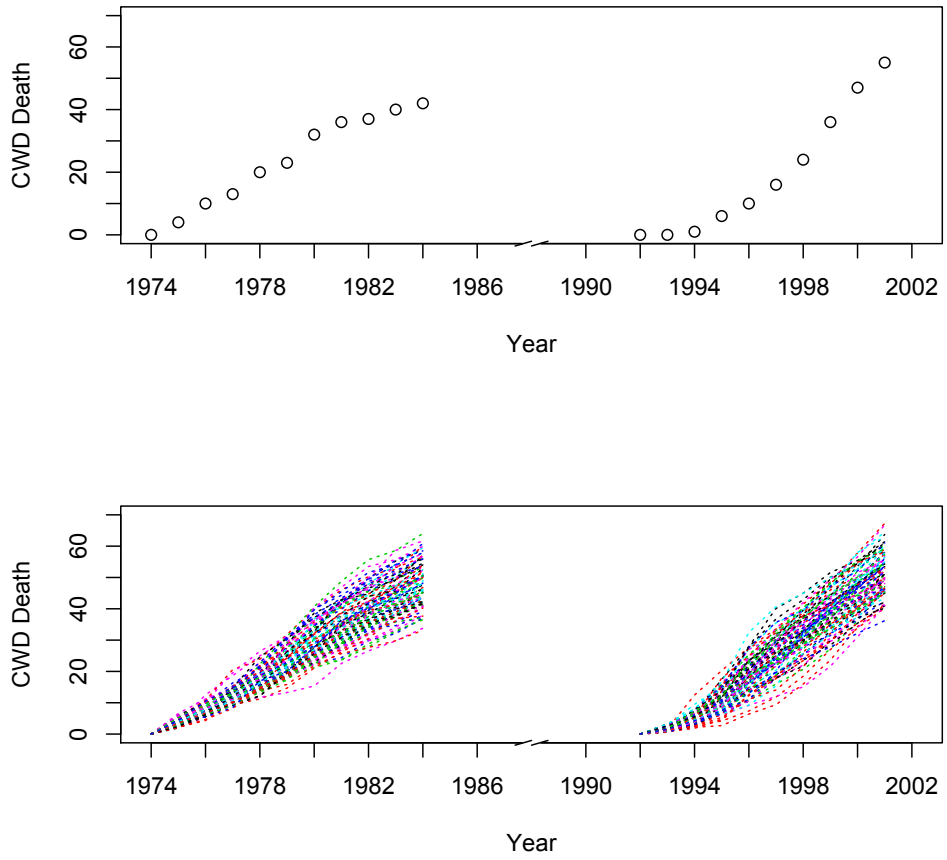

Figure 2.4: Upper display: observed cumulative number of deaths for CWD. Lower display: the 100 simulated trajectories of the cumulative number of deaths for CWD are obtained by using CWD direct transmission model (28) with estimated parameters from the PSML-MBB. The circled points are the observed CWD data.

rate of $m = 0.15$ (Miller et al., 2006), the corresponding estimate for the basic reproductive number R_0 equals $\widehat{\beta}N_0/(\widehat{\mu} + m) \approx 0.16N_0$ with 95% confidence interval $[0.06N_0, 0.42N_0]$, where N_0 is the initial population or susceptible size. Hence, we would expect that CWD will spread if a few infected animals, like one or two, are introduced to a closed susceptible population with size at least $1/0.06 \approx 17$ animals.

2.7 Conclusion and discussion

The dynamics of many ecological problems can be well described by a multivariate stochastic differential equation system. However, the transition densities of discrete-time observations are unknown for most interesting models. We propose the penalized simulated maximum likelihood approach, which provides a balanced approach to achieve accurate parameter estimates with efficient computation times for these complex stochastic models. The key idea is the introduction of a penalty term to select a better importance sampler in order to reduce the number of simulated sample paths. We compare the new method to the MBB sampler and the regularized sampler for three different models in simulation studies and also show an application for a real dataset. From those results, we conclude that the penalized simulated maximum likelihood approach is an improvement over the MBB sampler and the regularized sampler while keeping the computational cost low.

Note that it is possible to extend the penalized simulated maximum likelihood approach to allow for measurement errors in our observed data. The main challenge is still constructing effective and efficient importance samplers to approximate the transition probability density. The detailed statistical procedures are left as further work. Alternative approaches, such as methods that do not require evaluation of the likelihood function, have been proposed in both frequentist (Bretó et al., 2009) and Bayesian analysis (Andrieu et al., 2010; Sun et al., 2015a).

Markov jump processes offer an alternative approach to using SDE models (Toni et al., 2009; Drovandi and Pettitt, 2011). A Markov jump model particularly takes into account the discreteness of the data. However, a Markov jump model may be too simple. For example, the SDE models considered here allow modeling of the covariance between state variables. In contrast, a Markov jump model cannot capture such a dependence structure among the state variables.

Stramer and Yan (2007a) concluded the optimal choice for the number of Monte Carlo simulations J in (11) is of the order $O(M^2)$ for the MBB approach, where M is the number of subintervals between two observations. One can choose a number smaller than this as a starting point for the proposed penalized simulated maximum likelihood method in practice. More formal guidance is under investigation. Moreover, a formal study about the tuning parameter λ needs further development.

We find it is quite challenging to derive the theoretical properties of the maximum likelihood estimator based on either simulated likelihood (e.g., Pedersen and MBB) or penalized simulated likelihood (e.g., PSML). Pedersen (1995) and Geweke (1989) showed that the importance sampling estimator (11) converges to the transition density $p(\boldsymbol{X}(t_i)|\boldsymbol{X}(t_{i-1}))$. However, the properties of the MLE based on (11) (e.g., the estimator based on MBB or PSML) have not been established. The theoretical work, such as the convergence and the asymptotic distribution of the estimators, will be considered as future work.

Note that uncertainty in $\hat{\rho}$ is not accounted for in the bootstrap confidence intervals for the SDE parameters. This parameter is a nuisance parameter and is not used for simulated new datasets in the bootstrap algorithm. Methods to account for the effect of estimating ρ on bootstrap intervals for the process model parameters are a topic of future research.

CHAPTER 3

PSML EXTENSION AND THEORETICAL PROPERTIES

3.1 Extension with measurement error

We extend the penalized simulated maximum likelihood approach in Chapter 2 to the case that allows measurement or observation error and unknown initial conditions.

Consider a multivariate SDE model,

$$d\boldsymbol{X}(t) = f(\boldsymbol{X}(t), \boldsymbol{\theta})dt + g(\boldsymbol{X}(t), \boldsymbol{\theta})d\boldsymbol{W}(t) \tag{33}$$

with *unknown* initial condition $\boldsymbol{X}(t_0)$. Instead of directly observing a subset of the state process $\{\boldsymbol{X}_{obs}(t)\}_{t \geq t_0}$ at discrete time points, we assume $\boldsymbol{X}_{\text{obs}}(t)$ is subject to measurement error. That is

$$\boldsymbol{Y}(t) \sim r(\boldsymbol{X}_{obs}(t), \boldsymbol{\psi}) \tag{34}$$

is observed at t_i for $i = 0, 1, \ldots, n$, where r is a known density function and $\boldsymbol{\psi}$ is an unknown parameter vector. Note that all other assumptions are the same as the model in Section 2.2 of Chapter 2.

Then the discrete-time likelihood of model (33) is given by

$$L(\boldsymbol{\theta}, \boldsymbol{\psi}) = \prod_{i=1}^{n} p(\boldsymbol{Y}(t_i)|\boldsymbol{Y}(t_0 : t_{i-1}); \boldsymbol{\theta}, \boldsymbol{\psi}) \tag{35}$$

where $\boldsymbol{Y}(t_0 : t_{i-1})$ denotes all observations of $\boldsymbol{Y}$ from time t_0 to t_{i-1}, and

$$\begin{aligned}
&p(\boldsymbol{Y}(t_i)|\boldsymbol{Y}(t_0 : t_{i-1}); \boldsymbol{\theta}, \boldsymbol{\psi}) \\
&= \int r(\boldsymbol{Y}(t_i)|\boldsymbol{X}_{obs}(t_i); \boldsymbol{\psi}) p(\boldsymbol{X}(t_i)|\boldsymbol{Y}(t_0 : t_{i-1}); \boldsymbol{\theta}, \boldsymbol{\psi}) d\boldsymbol{X}(t_i) \\
&= \int r(\boldsymbol{Y}(t_i)|\boldsymbol{X}_{obs}(t_i); \boldsymbol{\psi}) p(\boldsymbol{X}(t_i)|\boldsymbol{X}(t_{i-1}); \boldsymbol{\theta}) p(\boldsymbol{X}(t_{i-1})|\boldsymbol{Y}(t_0 : t_{i-1}); \boldsymbol{\theta}, \boldsymbol{\psi}) d\boldsymbol{X}(t_{i-1} : t_i).
\end{aligned}$$

A feasible approach to evaluate this integral is via Monte Carlo integration. That requires a method to draw samples from the distribution of $\boldsymbol{X}(t_{i-1})|\boldsymbol{Y}(t_0 : t_{i-1})$, which can be obtained sequentially using the following equation,

$$\begin{aligned}
&p(\boldsymbol{X}(t_i)|\boldsymbol{Y}(t_0 : t_i); \boldsymbol{\theta}, \boldsymbol{\psi}) \propto \\
&\qquad r(\boldsymbol{Y}(t_i)|\boldsymbol{X}_{obs}(t_i); \boldsymbol{\psi}) p(\boldsymbol{X}(t_i)|\boldsymbol{X}(t_{i-1}); \boldsymbol{\theta}) p(\boldsymbol{X}(t_{i-1})|\boldsymbol{Y}(t_0 : t_{i-1}); \boldsymbol{\theta}, \boldsymbol{\psi})
\end{aligned} \tag{36}$$

for $i = 1, ..., n$. Therefore, assuming $p(\boldsymbol{X}(t_i)|\boldsymbol{X}(t_{i-1}); \boldsymbol{\theta})$ is known, iterative application of Monte Carlo integration (36) yields an approximation of $\boldsymbol{X}(t_\ell)|\boldsymbol{Y}(t_0 : t_{\ell-1})$ for $\ell \geq 1$. This is the idea of a particle filter (Durham and Gallant, 2002; Pitt and Shephard, 1999). The algorithmic form of this sequential Monte Carlo algorithm is provided in Appendix.

Again, it is left to approximate the transition probability density $p(\boldsymbol{X}(t_i)|\boldsymbol{X}(t_{i-1}); \boldsymbol{\theta})$. The procedure to approximate $p(\boldsymbol{X}(t_i)|\boldsymbol{X}(t_{i-1}); \boldsymbol{\theta})$ is still the same as introduced in Chapter 2. However, the importance samplers described in Section 2.3 of Chapter 2, such as the MBB and the regularized samplers, need to be adjusted to include measurement error. Note that the Pedersen sampler stays the same because it does not depend on the ending point.

3.1.1 Modified Brownian bridge sampler with measurement error

The MBB sampler draws $\boldsymbol{X}^{m+1} \equiv \boldsymbol{X}(t_{i-1}+(m+1)\delta)$ conditional on $\boldsymbol{X}^m$ and $\boldsymbol{X}_{\text{obs}}(t_i)$ as described in Section 2.3 of Chapter 2. Here, we wish to draw $\boldsymbol{X}^{m+1}$ conditional on $\boldsymbol{X}^m$ and the noisy observation $\boldsymbol{Y}(t_i)$. This is achieved by sampling from a Gaussian approximation to $p(\boldsymbol{X}^{m+1}|\boldsymbol{X}^m, \boldsymbol{Y}(t_i))$. See Golightly and Wilkinson (2008) for more details.

Suppose the distribution of the measurement error can be approximated by a Gaussian distribution, that is $\boldsymbol{Y}(t) \sim N(\boldsymbol{X}_{obs}(t), \boldsymbol{\Sigma}_{obs})$, then the modified Brownian bridge with measurement error (MBBE) sampler draws $\boldsymbol{X}^{m+1}$ from the density

$$q(\boldsymbol{X}^{m+1}|\boldsymbol{X}^m, \boldsymbol{Y}(t_i)) = \phi(\boldsymbol{X}^{m+1}; \boldsymbol{X}^m + \boldsymbol{\eta}_m\delta, \boldsymbol{\Sigma}_m\delta), \tag{37}$$

where $\phi(\boldsymbol{X}; \boldsymbol{\mu}, \Sigma)$ is a multivariate normal density with mean vector $\boldsymbol{\mu}$ and covariance matrix Σ. Here

$$\boldsymbol{\eta}_m = f_{obs}(\boldsymbol{X}^m) + \boldsymbol{C}_m[\boldsymbol{G}_{obs,obs}(\boldsymbol{X}^m)(M-m)\delta + \boldsymbol{\Sigma}_{obs}]^{-1}[\boldsymbol{Y}(t_i) - (\boldsymbol{X}^m + f_{obs}(\boldsymbol{X}^m)(M-m)\delta)]$$

and

$$\boldsymbol{\Sigma}_m = \boldsymbol{G}(\boldsymbol{X}^m) - \boldsymbol{C}_m[\boldsymbol{G}_{obs,obs}(\boldsymbol{X}^m)(M-m)\delta + \boldsymbol{\Sigma}_{obs}]^{-1}\boldsymbol{C}'_m\delta,$$

where

$$\boldsymbol{C}'_m = [\boldsymbol{G}_{\text{obs,-obs}}(\boldsymbol{X}^m), \boldsymbol{G}_{\text{obs,obs}}(\boldsymbol{X}^m)]$$

for $m = 0, 1, \cdots, M-1$. Note that when $\boldsymbol{\Sigma}_{obs} = 0$, MBBE degenerates to MBB.

3.1.2 Regularized sampler with measurement error

Similarly, we can adjust the regularized sampler (Section 2.3 of Chapter 2) to incorporate measurement error (RegE hereafter). Let $\boldsymbol{\mu}_P$ and $\boldsymbol{\Sigma}_P$ be the mean and the variance of the Pedersen sampler and $\boldsymbol{\mu}_{ME}$ and $\boldsymbol{\Sigma}_{ME}$ be the mean and the variance of the MBBE sampler.

Then the regularized sampler draws $\boldsymbol{X}^{m+1}$ from the density

$$q_\rho(\boldsymbol{X}^{m+1}|\boldsymbol{X}^m, \boldsymbol{Y}(t_i)) = \phi(\boldsymbol{X}^{m+1}; (\mathcal{I} - \boldsymbol{V})\,\boldsymbol{\mu}_P + \boldsymbol{V}\boldsymbol{\mu}_{ME}, (\mathcal{I} - \boldsymbol{V})\,\boldsymbol{\Sigma}_P + \boldsymbol{V}\boldsymbol{\Sigma}_{ME}), \tag{38}$$

where $\mathcal{I}$ is the identity matrix and

$$\boldsymbol{V} = \frac{M-m}{M-m+\rho(M-m-1)^2}\mathcal{I}, \tag{39}$$

where $\rho \in [0, 1]$.

3.1.3 Penalized simulated maximum likelihood with measurement error

Using importance sampling, we draw i.i.d. J samples, $\{\boldsymbol{X}^{(j)}(t_{i-1} : t_i), j = 1, \cdots, J\}$, from an importance sampler q_ρ. Let h_ρ be the importance sampling weights to approximate $p(\boldsymbol{Y}(t_i)|\boldsymbol{Y}(t_1 : t_{i-1}))$ in (35). Specifically,

$$\begin{aligned} &h_\rho\Big(\boldsymbol{X}^{(j)}(t_{i-1} : t_i)\Big) \\ &\qquad \equiv \frac{r(\boldsymbol{Y}(t_i)|\boldsymbol{X}_{\text{obs}}(t_i))\prod_{m=1}^{M} p^{(1)}(\boldsymbol{X}^{(j)}(t_{i-1}+m\delta)|\boldsymbol{X}^{(j)}(t_{i-1}+(m-1)\delta))}{q_\rho(\boldsymbol{X}^{(j)}(t_{i-1} : (t_i - \delta)))}, \end{aligned} \tag{40}$$

where q_ρ is the importance sampler density, e.g., (38) and (41) below. The definition and algorithm of the PSML estimator is the same as described in Chapter 2.

We again consider two classes of importance samplers with auxiliary parameter ρ in (40) for the case with measurement errors. The first one is given by

$$q_\rho(\boldsymbol{X}^{m+1}|\boldsymbol{X}^m, \boldsymbol{X}_{\text{obs}}(t_i)) = \phi(\boldsymbol{X}^{m+1}; \boldsymbol{X}^m + \boldsymbol{\eta}_m\delta, \rho\Sigma_m\delta), \tag{41}$$

where $\boldsymbol{\eta}_m, \Sigma_m$ are defined in (37). The second class is the regularized sampler with measurement error (38).

3.1.4 Simulation studies

Here, we compare the performance of the MBBE sampler, the RegE sampler with $\rho = 0.1$, and our PSML with the modified MBBE class (41) and the RegE class (38) on simulated datasets for two different models. We refer to PSML with the modified MBB class (41) as PSML-MBBE and refer to PSML with the RegE class (38) as PSML-RegE.

We first consider the Ornstein - Uhlenbeck process with measurement error ϵ,

$$\begin{aligned} dX(t) &= (\theta_1 - \theta_2 X(t))dt + \theta_3 dW(t), \\ Y(t) &= X(t) + \epsilon(t), \end{aligned} \tag{42}$$

and the parameter $\boldsymbol{\theta} = (\theta_1, \theta_2, \theta_3, \sigma) \in \mathbb{R} \times \mathbb{R}_+ \times \mathbb{R}_+ \times \mathbb{R}_+$, where $\epsilon(t) \sim N(0, \sigma^2)$ for $t \geq t_0$. We assume noisy observations, Y, are observed at t_i for $i = 0, \ldots, n$. We consider two scenarios with different time intervals $t_i - t_{i-1} = 1$ or 2. For each scenario, we generate 100 datasets with parameter $\boldsymbol{\theta}_0 = (0.0187, 0.2610, 0.0224, 0.5)$. For each dataset, the length of overall process time $t_n - t_0 = 100$, that means the sample size is $n = 100$ if the time interval $t_i - t_{i-1} = 1$ and $n = 50$ if $t_i - t_{i-1} = 2$.

With measurement error, the exact likelihood is not available for this case. We compute the bias and the root mean square error (RMSE) of the simulated maximum likelihood estimators $\widehat{\boldsymbol{\theta}}_r$ with respect to the true parameters $\boldsymbol{\theta}_0$, defined by $\frac{1}{100}\sum_{r=1}^{100}(\widehat{\boldsymbol{\theta}}_r - \boldsymbol{\theta}_0)$ and $\sqrt{\frac{1}{100}\sum_{r=1}^{100}(\widehat{\boldsymbol{\theta}}_r - \boldsymbol{\theta}_0)^2}$, respectively. For all the methods, we consider $M = 8$ subintervals the simulated sample paths $J = 8$. We set the $\epsilon_0 = 0.03$, $\lambda_0 = 0.25$, $\delta_\lambda = 0.05$ and $\delta_\epsilon = 0.001$. The initial values for optimization for θ_1, θ_2, θ_3 and σ_ϵ are $0.05, 0.5$, 0.05, and 0.05, respectively.

Table 3.1 shows that both PSML-MBB and PSML-Reg have better performance than the MBBE sampler and the RegE sampler in terms of reducing bias and RMSE, especially when the time interval is large ($t_i - t_{i-1} = 2$).

Table 3.1: The bias and RMSE of the simulated maximum likelihood estimates with respect to the true parameters for the Ornstein-Uhlenbeck process with measurement error (42). Both PSML-MBBE and PSML-RegE have better performance than the MBBE sampler and the RegE sampler in terms of reducing bias and RMSE.

		$t_i - t_{i-1} = 1$				$t_i - t_{i-1} = 2$			
	Method	θ_1	θ_2	θ_3	σ	θ_1	θ_2	θ_3	σ
Bias	MBBE	0.15	2.83	0.45	0.22	0.20	2.17	0.54	0.20
	RegE	0.28	3.98	0.43	0.26	0.19	3.55	0.40	0.26
	PSML-MBBE	−0.03	1.55	0.28	0.20	−0.11	0.66	0.19	0.19
	PSML-RegE	0.12	2.72	0.31	0.23	0.06	2.39	0.10	0.27
RMSE	MBBE	0.39	3.05	0.56	0.31	0.47	2.36	0.62	0.34
	RegE	0.47	4.17	0.55	0.33	0.46	3.76	0.51	0.34
	PSML-MBBE	0.24	1.64	0.40	0.29	0.25	0.79	0.32	0.28
	PSML-RegE	0.30	2.84	0.43	0.31	0.29	2.68	0.27	0.34

Next, we consider the stochastic version of the well-known chaotic Lorenz 63 model (Lorenz, 1963; Bengtsson et al., 2003) with measurement error, which is given by

$$
\begin{aligned}
d\begin{pmatrix} X_1(t) \\ X_2(t) \\ X_3(t) \end{pmatrix} &= \begin{pmatrix} s(X_2(t) - X_1(t)) \\ rX_1(t) - X_2 - X_1(t)X_3(t) \\ X_1(t)X_2(t) - bX_3(t) \end{pmatrix} dt + \sigma d \begin{pmatrix} W_1(t) \\ W_2(t) \\ W_3(t) \end{pmatrix}, \\
\begin{pmatrix} Y_1(t) \\ Y_2(t) \\ Y_3(t) \end{pmatrix} &= \begin{pmatrix} X_1(t) \\ X_2(t) \\ X_3(t) \end{pmatrix} + \begin{pmatrix} \epsilon_1(t) \\ \epsilon_2(t) \\ \epsilon_3(t) \end{pmatrix},
\end{aligned}
\tag{43}
$$

where W_1, W_2, and W_3 are three independent Wiener processes and $\epsilon_i \overset{iid}{\sim} N(0, \sigma_\epsilon^2)$.

We assume the noisy observations, $(Y_1, Y_2, Y_3)^T$, are observed at t_i for $i = 0, \ldots, n$. We consider three scenarios with different time interval $t_i - t_{i-1} = 0.025$, or 0.05, or 0.1. For each scenario, we generate 100 datasets using commonly used parameter values $\boldsymbol{\theta}_0 = (s_0 = 10, r_0 = 28, b_0 = 8/3, \sigma_0 = 2, \sigma_{\epsilon 0} = 0.6)$ (Bengtsson et al., 2003). For each dataset, the length of overall process time $T \equiv t_n - t_0 = 2$, that means the sample size $n = 80$ if the time interval $t_i - t_{i-1} = 0.025$, $n = 40$ if $t_i - t_{i-1} = 0.05$, $n = 20$ if $t_i - t_{i-1} = 0.1$. The

exact transition density is not available. We compute the RMSE of the simulated maximum likelihood estimators $\widehat{\boldsymbol{\theta}}_r$ with respect to the true parameters $\boldsymbol{\theta}_0$. Estimates are obtained using sample path $J = 48$ and $M = 10$ subintervals. We set $\lambda_0 = 0.5$, $\epsilon_0 = 5.5$, $\delta_\lambda = 0.05$ and $\delta_\epsilon = 0.5$. The initial values for optimization are $(15, 30, 5, 1)$.

Table 3.2: The RMSE of the simulated maximum likelihood estimates with respect to the true parameters for the stochastic Lorenz 63 model (43). The improvement of the PSML-Reg over the MBB and the regularized sampler with different J's is evident.

	Method	s	r	b	σ	σ_ϵ
$t_i - t_{i-1} = 0.1$	MBBE	29.59	1.47	0.97	8.60	0.65
	RegE	49.52	1.75	1.03	8.49	0.53
	PSML-MBBE	5.05	0.33	0.14	1.86	0.26
	PSML-RegE	3.55	0.34	0.13	1.70	0.34
$t_i - t_{i-1} = 0.05$	MBBE	5.01	0.33	0.23	2.02	0.30
	RegE	4.87	0.31	0.23	2.02	0.30
	PSML-MBBE	4.68	0.30	0.10	2.50	0.39
	PSML-RegE	2.22	0.29	0.10	1.10	0.16
$t_i - t_{i-1} = 0.025$	MBBE	8.24	0.42	0.43	3.09	0.34
	RegE	13.19	0.39	0.41	3.16	0.35
	PSML-MBBE	4.21	0.32	0.10	2.41	0.26
	PSML-RegE	4.17	0.32	0.14	2.26	0.21

We notice that the RegE sampler is not better than the MBBE sampler for the case that allows measurement error as shown in Table 3.2. The improvement of PSML algorithm over the MBBE or RegE sampler is evident, especially when the time interval $t_i - t_{i-1}$ is large. An estimated ρ based on the data in the PSML-RegE plays an important role in generating efficient proposal trajectories.

3.2 Consistency and asymptotic distribution

We show the consistency and asymptotic distribution of the PSML estimator under the setup without measurement error as described in Chapter 2. The proof for the case that allows measurement error, as described in Chapter 3, can be obtained by using similar arguments below.

Let $\ell_n(\theta)$ denote the log likelihood,

$$\ell_n(\theta) = \sum_{i=1}^{n} \log p(\boldsymbol{X}_{\text{obs}}(t_i)|\boldsymbol{X}(t_0), \boldsymbol{X}_{\text{obs}}(t_1 : t_{i-1})),$$

$\Psi_{n,M}(\theta)$ denote the approximate log likelihood,

$$\Psi_{n,M}(\theta) = \sum_{i=1}^{n} \log p^{(M)}(\boldsymbol{X}_{\text{obs}}(t_i)|\boldsymbol{X}(t_0), \boldsymbol{X}_{\text{obs}}(t_1 : t_{i-1})),$$

and $\Psi_{n,M,J}(\theta)$ be the penalized log likelihood,

$$\Psi_{n,M,J}(\theta) = \sum_{i=1}^{n} \log \left(\frac{1}{J} \sum_{j=1}^{J} h_\rho \right) - \lambda \sum_{i=1}^{n} \widehat{\text{cv}} \left(\frac{1}{J} \sum_{j=1}^{J} h_\rho \right),$$

where h_ρ is in (21).

Theorem 1. *Let $\Theta \subseteq \mathbb{R}^q$ be a compact subset, $\theta_0 \in \Theta$ denote the true parameter value, and P_θ denote a unique probability measure on the space $C([0,\infty), \mathbb{R}^k)$ of continuous trajectories from $[0,\infty)$ into $\mathbb{R}^k$ endowed with its Borel σ-field for each $\theta \in \Theta$. Furthermore, assume that the following two conditions are satisfied P_{θ_0} almost surely for some fixed $n, M \in \mathbb{N}$:*

(i) $\Psi_{n,M}(\theta)$ is continuous and has a unique maximum point $\hat{\theta}_{n,M} \in \Theta$.

(ii) $\Psi_{n,M,J}(\theta)$ is continuous, at least when J is larger than some fixed $J_0 \in \mathbb{N}$.

(iii) Finally, assume that there exists, with P_{θ_0} probability that tends to 1 as $n \to \infty$, a sequence $\{\hat{\theta}_n\}_{n=1}^{\infty} \subseteq \Theta$ of local maximum points for the log likelihood function $\ell_n(\theta)$ such that

(a) $\hat{\theta}_n \to \theta_0$ in probability under P_{θ_0} as $n \to \infty$;

(b) there exists a sequence $\{A_n(\theta_0)\}_{n=1}^{\infty}$ *of a non-random and nonsingular* $q \times q$ *matrices sequence such that*

$$A_n(\theta_0)(\hat{\theta}_n - \theta_0) \Rightarrow N_q(0, V(\theta_0))$$

in distribution under P_{θ_0} *as* $n \to \infty$*, where* $V(\theta_0)$ *is some non-random positive definite* $q \times q$ *matrix.*

Then

(i) there exists (P_{θ_0} *almost surely) sequence* $\{\hat{\theta}_{n,M,J}\}_{J=1}^{\infty} \subseteq \Theta$ *of maximum points for the functions* $\{\Psi_{n,M,J}(\theta)\}_{J=1}^{\infty}$ *for some fixed* $n, M \in \mathbb{N}$*;*

(ii) for any such sequence in (i), there exists subsequences $M(n) \to \infty$ *and* $J(n) \to \infty$*, with* P_{θ_0} *that tends to 1 as* $n \to \infty$*, a sequence* $\{\hat{\theta}_{n,M(n),J(n)}\}_{n=1}^{\infty}$ *such that*

$$\hat{\theta}_{n,M(n),J(n)} \to \theta_0$$

in probability under P_{θ_0} *as* $n \to \infty$*, and such that*

$$A_n(\theta_0)(\hat{\theta}_{n,M(n),J(n)} - \theta_0) \Rightarrow N_q(0, V(\theta_0))$$

in distribution under P_{θ_0} *as* $n \to \infty$*.*

Proof. The strong law of large numbers implies that

$$\frac{1}{J}\sum_{j=1}^{J} h_\rho \to p^{(M)}(\boldsymbol{X}_{\text{obs}}(t_i)|\boldsymbol{X}(t_0), \boldsymbol{X}_{\text{obs}}(t_1 : t_{i-1}))$$

almost surely as $J \to \infty$, then

$$\sum_{i=1}^{n} \log\left(\frac{1}{J}\sum_{j=1}^{J} h_\rho\right) \to \sum_{i=1}^{n} \log p^{(M)}(\boldsymbol{X}_{\text{obs}}(t_i)|\boldsymbol{X}(t_0), \boldsymbol{X}_{\text{obs}}(t_1 : t_{i-1}))$$

almost surely as $J \to \infty$, and

$$\begin{aligned}
&\sup_{\theta\in\Theta} |\Psi_{M,J}(\theta) - \Psi_M(\theta)| \\
&\leq \sup_{\theta\in\Theta} \left[\left| \sum_{i=1}^{n} \log\left(\frac{1}{J}\sum_{j=1}^{J} h_\rho\right) - \sum_{i=1}^{n} \log p^{(M)}(\boldsymbol{X}_{\text{obs}}(t_i)|\boldsymbol{X}(t_0), \boldsymbol{X}_{\text{obs}}(t_1 : t_{i-1})) \right| \right. \\
&\qquad \left. + \frac{\lambda}{\sqrt{J}} \sum_{i=1}^{n} \widehat{\text{cv}}\,(h_\rho) \right] \to 0
\end{aligned}$$

almost surely as $J \to \infty$.

Then Theorem 1 in Pedersen (1995a) implies that

$$\hat{\theta}_{n,M,J} \to \hat{\theta}_{n,M}$$

in probability under P_{θ_0} as $J \to \infty$ for some fixed $n, M \in \mathbb{N}$.

Based on Lemma A in the Appendix of Pedersen (1995a), there exists subsequences $M(n)$ and $J(n)$ such that

$$\hat{\theta}_{n,M(n),J(n)} \to \theta_0$$

in probability under P_{θ_0} as $n \to \infty$, and $\hat{\theta}_{n,M(n),J(n)}$ also is asymptotically normally distributed, that is

$$A_n(\theta_0)(\hat{\theta}_{n,M(n),J(n)} - \theta_0) \Rightarrow N_q(0, V(\theta_0))$$

in distribution under P_{θ_0} as $n \to \infty$.

■

Note that under Assumptions 1 – 3 for Theorem 2 in Pedersen (1995a), the conditions (c) and (d) in Theorem 1 are met. Billingsley (1961) showed that if the diffusion process corresponding to the stochastic differential equation is time-homogeneous and ergodic, the observation time intervals are equidistant, and the transition densities (exist and) satisfy some weak regularity conditions, then conditions (c) and (d) in Theorem 1 are met with $A_n(\theta_0) = -\ell_n''(\theta_0)$, the negative second derivative of the log likelihood, and $V(\theta_0) = \mathcal{I}$, the identity matrix. Then Theorem 1 implies that

$$-\ell_n''(\theta_0)^{1/2}(\hat{\theta}_{n,M(n),J(n)} - \theta_0) \Rightarrow N_q(0, \mathcal{I}),$$

in distribution under P_{θ_0} as $n \to \infty$.

CHAPTER 4

DATA DRIVEN ADAPTIVE MESH ESTIMATION IN NONLINEAR ORDINARY DIFFERENTIAL EQUATION MODELS WITH BOTH NUMERICAL AND MEASUREMENT ERRORS

4.1 Introduction

Ordinary differential equations (ODEs) describe how systems evolve in time and are essential tools in many scientific disciplines. For example, Gyllenberg and Webb (1988) propose a two-compartment model of the tumor cells in a typical avascular multicellular tumor spheroid. The model assumes that the tumor cells transition to and from a quiescent state at the rate $\gamma_0(N)$ and $\gamma_i(N)$, respectively, which can be described by a two dimensional ODE:

$$\begin{aligned} \frac{dP}{dt} &= (\beta - \mu_p - \gamma_0(N))P + \gamma_i(N)Q, \\ \frac{dQ}{dt} &= \gamma_0(N)P - (\gamma_i(N) + \mu_q)Q, \end{aligned} \tag{44}$$

where $N(t) = P(t) + Q(t)$, $P(t)$ is the density of proliferative cells at time t, $Q(t)$ is the density of quiescent cells at time t, β is a constant proliferation rate for proliferative cells, μ_p and μ_q are death rates for proliferative and quiescent cells, respectively. One observes the densities of proliferative cells and quiescent cells $\boldsymbol{C}(t) = \{P(t), Q(t)\}$ at discrete time points $t_i, i = 1, \ldots, t_n$, with measurement errors $\epsilon(t)$, $\boldsymbol{y}(t_i) = \boldsymbol{C}(t_i) + \boldsymbol{\epsilon}(t_i)$.

Similarly, Stein et al. (2013) propose a generalized Lotka-Volterra (LV) system to predict the temporal dynamics of intestinal microbiota and further understand microbial ecosystems. They describe the ecological dynamics using generalized LV equations with the addition of

external perturbations,

$$\frac{dx_i}{dt} = \mu_i x_i + x_i \sum_{j=1}^{L} M_{ij} x_j + x_i \sum_{l=1}^{P} \varepsilon_{il} \mu_l, \tag{45}$$

where $x_i(t)$ is the concentration of a focal species i, $i = 1, \ldots, L$, at time t, μ_i is its growth rate, M_{ij} is the effect of interaction of species j on species i, and ε_{il} is the susceptibility to the time dependent perturbation $\mu_l(t)$. The observed data are the state variables $\boldsymbol{x(t)} = \{x_1(t), \ldots, x_L(t)\}$ which are measured at discrete time points $t_i, i = 1, \ldots, t_n$, with measurement errors $\epsilon(t)$, so $\boldsymbol{y}(t_i) = \boldsymbol{x}(t_i) + \boldsymbol{\epsilon}(t_i)$.

The objective of this work is to estimate the parameters in ODEs, $\boldsymbol{\theta} = \{\beta, \mu_p, \mu_q, \gamma_0(N), \gamma_i(N)\}$ in (44) or $\boldsymbol{\theta} = \{\mu_i, M_{ij}, \varepsilon_{il}\}$ in (45), from the noisy observations $\boldsymbol{y}(t_i)$, $i = 1, \ldots, n$. Most ODEs cannot be solved analytically due to their complex nature. Numerical methods, such as the Euler and Runge-Kutta methods, are needed to solve the ODEs. Based on the approximated ODE solution, the parameter of interest $\boldsymbol{\theta}$ can be estimated by least squares. More accurate numerical methods, such as the Runge-Kutta method, usually have higher computational costs than simpler methods for the same step size, such as the Euler method. Here, we develop a new approach called the data driven adaptive mesh (DDAM) that balances accuracy and computational time. The basic idea of DDAM is that we can implement a more accurate method around the data points and a less accurate method elsewhere.

As an alternative to the numerical ODE solver approachs, Ramsay et al. (2007) proposed the generalized profiling (GP) procedure which uses non-parametric basis functions to smooth the data. GP is a two-stage optimization procedure. Firstly, the coefficients of the basis functions are estimated by fitting the observations with an ODE related penalty using initial choice of the parameters of interest in the inner optimization. Then the parameter of interest $\boldsymbol{\theta}$ is estimated by minimizing this penalized data-dependent fitting criterion in the outer optimization.

We show the following by the simulation study in Section 4.4. For high frequency data (the time interval $t_i - t_{i-1}$ is small), the GP procedure and the DDAM have a similar level of accuracy for parameter estimation. However, the computational time of the latter is shorter than the former. When the data are sparse, the accuracy of the GP procedure estimator is poor compared with the DDAM estimator.

The remainder of the paper is organized as follows. In Section 4.2, we present the details of the proposed method. Section 4.3 provides some theoretical properties of the new method. Section 4.4 presents simulation studies for different models. Section 4.5 illustrates the new method on a real dataset. Section 4.6 concludes with a discussion.

4.2 Methodology

4.2.1 Model setup

Consider $\boldsymbol{Y}_{p\times 1}(t)$ modeling time series for p quantities of interest, which could be modeled by

$$\boldsymbol{Y}(t) = \boldsymbol{u}(t, \boldsymbol{\theta}) + \boldsymbol{\epsilon}(t) \tag{46}$$

where $\boldsymbol{\theta} \in \Theta \subset \mathbb{R}^d$; and $\boldsymbol{\epsilon}(t) \in L^2(T)$ (i.e. square integrable on the time domain) has zero mean and covariance function $\mathrm{Cov}(\boldsymbol{\epsilon}(t), \boldsymbol{\epsilon}(s)) = \sigma^2 \delta_{ts} \mathcal{I}$, where $\delta_{ts} = 1$ if $t = s$ and 0 elsewhere. State variable $\boldsymbol{\mu}$ is modeled by differential equations

$$\frac{d\boldsymbol{\mu}}{dt} = \boldsymbol{F}(\boldsymbol{\mu}(t), \boldsymbol{\theta}) \tag{47}$$

where the map $\boldsymbol{F} : \mathbb{R}^p \times \Theta \to \mathbb{R}^p$ is assumed to be smooth (when $\boldsymbol{F}$ is an operator, we assume it is at least twice Fréchét-Differentiable).

If a closed-form solution $\boldsymbol{\mu}_{\boldsymbol{\theta}}(t)$ of ODE (47) is available, then the standard nonlinear least squares estimator can be used to estimate unknown parameters $\boldsymbol{\theta}$. However, a closed-form solution is not available for most of cases in practice. Let $\widetilde{\boldsymbol{\mu}}$ be an estimator of $\boldsymbol{\mu}$ such that

$||\boldsymbol{\mu} - \widetilde{\boldsymbol{\mu}}|| = O(h^p)$ for some h to be defined later and $p > 1$, then (46) can be approximated by the nonlinear regression model

$$\boldsymbol{Y}(t) = \widetilde{\boldsymbol{\mu}}(t, \boldsymbol{\theta}) + \boldsymbol{\epsilon}(t). \tag{48}$$

For discrete observation points, consider $\{t_1, \ldots, t_n\} \subset T$ (the time domain). Then the parameter vector $\boldsymbol{\theta}$ can be estimated using nonlinear least squares, which is given by

$$\widehat{\boldsymbol{\theta}} = \text{argmin}_{\boldsymbol{\theta}} \sum_{i=1}^{n} ||\boldsymbol{Y}(t_i) - \widetilde{\boldsymbol{u}}(t_i, \boldsymbol{\theta})||^2, \tag{49}$$

or both $\boldsymbol{\theta}$ and σ can be estimated via maximum likelihood, which is given by

$$(\widehat{\boldsymbol{\theta}}, \widehat{\sigma}) = \text{argmax } L\left(\boldsymbol{Y}_1, \cdots, \boldsymbol{Y}_n | \widetilde{\boldsymbol{\mu}}(t_1, \boldsymbol{\theta}), \cdots, \widetilde{\boldsymbol{\mu}}(t_n, \boldsymbol{\theta}), \sigma^2\right).$$

4.2.2 Data driven adaptive mesh method (DDAM)

The estimator $\widehat{\boldsymbol{\mu}}$ can be constructed using a numerical ODE solver, which requires a partition of the time interval $[0, T]$ by m grid points. Let $t_0 = s_0 < s_1 < \cdots < s_{m-1} = T$ be the grid points, and $h = s_{j+1} - s_j$ be the step size, and $\boldsymbol{\mu}_j^h$ and $\boldsymbol{\mu}_{j+1}^h$ be the numerical approximation to the true solution $\boldsymbol{\mu}(s_j)$ and $\boldsymbol{\mu}(s_{j+1})$, respectively. Suppose we consider a one-step numerical method, given by

$$\boldsymbol{\mu}_{j+1}^h = \boldsymbol{\mu}_j^h + h\Psi(s_j, \boldsymbol{\mu}_j^h, \boldsymbol{\mu}_{j+1}^h, h), \tag{50}$$

where the operator Ψ depends on the specific numerical method. The Euler method and 4th order Runge-Kutta (RK4) method are the most commonly used numerical methods to approximate the solution to a set of ODEs. We use $\widetilde{\boldsymbol{\mu}}(t, \boldsymbol{\theta})$ to denote the interpolated numerical solution of $\boldsymbol{\mu}(t, \boldsymbol{\theta})$ obtained from the one-step numerical method (50) for given $\boldsymbol{\theta}$. If the observations points $(t_i, i = 1, 2, \ldots, n)$ are not coincident with the grid points

$(s_j, j = 1, 2, \ldots, m-1)$, then cubic Hermite interpolation (De Boor, 1978) is often used to approximate the solution at the grid points.

For both methods the global truncation error $||\boldsymbol{\mu}(s_j) - \boldsymbol{\mu}_j^h||$ is a function of the step size h. For the Euler and RK4 methods, the error is of $O(h)$ and $O(h^4)$, respectively. Although the RK4 method has a smaller global truncation error, the computation cost of the RK4 is much higher than the Euler method. Hence, in order to balance the accuracy and computational cost, we implement the RK4 method with a smaller step size h around the data points and the Euler method with a larger step size H elsewhere. For each observation t_i, $i = 1, \ldots, n$, we implement the RK4 method from $t_i - H$ to $t_i + H$, and implement the Euler method from $t_i + H$ to $t_{i+1} - H$. Note that we require $H \leq \Delta/2$, where $\Delta = \min_{1 \leq i \leq n-1} t_{i+1} - t_i$. The detailed steps from t_i to t_{i+1} for the DDAM method are shown in Table 4.1 and Figure 4.1.

Table 4.1: The numerical steps from t_i to t_{i+1} for the data driven adaptive mesh method.

From	To	Method	Step size	Number of steps
t_i	$t_i + H/2$	RK4	h	$H/(2h)$
$t_i + H/2$	$t_{i+1} - H/2$	Euler	H	$(t_{i+1} - t_i - H)/H$
$t_{i+1} - H/2$	t_{i+1}	RK4	h	$H/(2h)$

The DDAM method has a lower computation cost than the RK4 method because of the Euler steps between $t_i + H/2$ to $t_{i+1} - H/2$. In the meantime, the accuracy of the estimator in terms of bias and root mean square error (RMSE) is not sacrificed, especially when the data are sparsely sampled (see Section 4.4 for details).

4.3 Selection of λ

The ratio $\lambda \equiv H/h$ can be chosen using cross-validation as is commonly used in selecting the tuning parameter practice. We seek the optimal ratio λ that minimizes the prediction error $E[Y(t') - \widetilde{\mu}(t', \widehat{\boldsymbol{\theta}})]$ for a fixed $t' \in [t_0, T]$, where E is the expectation with respect to

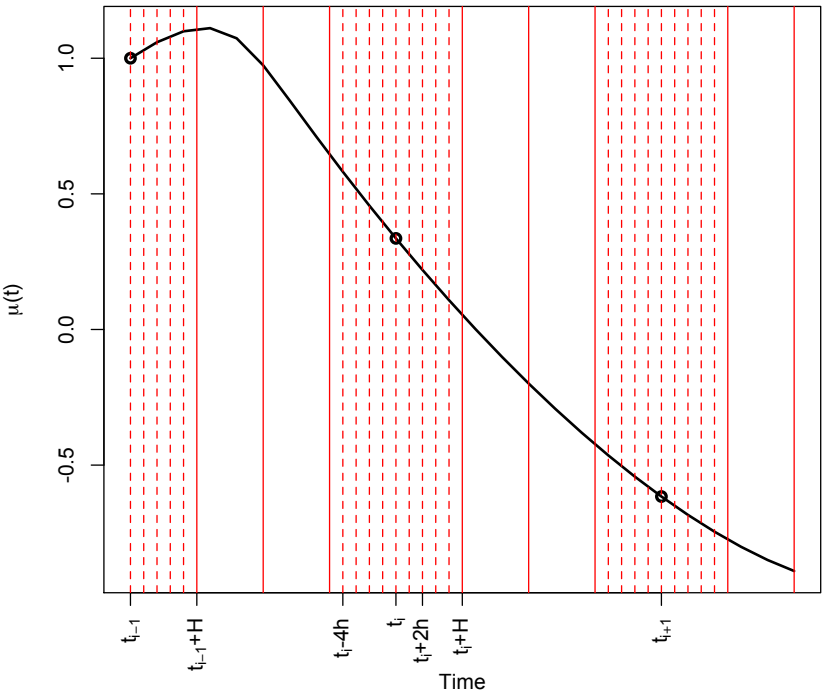

Figure 4.1: The numerical steps for the data driven adaptive mesh method when $H = 5h$. The RK4 method is used between dashed lines, and the Euler method is used between solid lines.

$P_{\boldsymbol{\theta}_0}$, the joint probability distribution of $(t, Y(t))$ at true value $\boldsymbol{\theta}_0$. Note that

$$\begin{aligned} E||Y(t') - \widetilde{\boldsymbol{\mu}}(t', \widehat{\boldsymbol{\theta}})||^2 &= E||Y(t') - \boldsymbol{\mu}(t, \boldsymbol{\theta}_0) + \boldsymbol{\mu}(t', \boldsymbol{\theta}_0) - \widetilde{\boldsymbol{\mu}}(t', \widehat{\boldsymbol{\theta}})||^2 \\ &= E||Y(t') - \boldsymbol{\mu}(t', \boldsymbol{\theta}_0)||^2 + E||\boldsymbol{\mu}(t', \boldsymbol{\theta}_0) - \widetilde{\boldsymbol{\mu}}(t', \widehat{\boldsymbol{\theta}})||^2 \\ &\quad + 2E[\boldsymbol{\epsilon}^T(t')]E[(\boldsymbol{\mu}(t', \boldsymbol{\theta}_0) - \widetilde{\boldsymbol{\mu}}(t', \widehat{\boldsymbol{\theta}}))] \\ &= p\sigma^2 + E||\boldsymbol{\mu}(t', \theta_0) - \widetilde{\boldsymbol{\mu}}(t', \widehat{\boldsymbol{\theta}})||^2 \end{aligned}$$

where $||\boldsymbol{x}||$ is the Euclidean norm of $\boldsymbol{x}$ and the second term $E[\boldsymbol{\mu}(t', \theta_0) - \widetilde{\boldsymbol{\mu}}(t', \widehat{\boldsymbol{\theta}})]^2$ is related to the accuracy of the DDAM method. Hence, we study the global truncation error of the DDAM method next.

First we make the following assumptions:

Assumption 1. $\boldsymbol{\theta} \in \Theta$, *where* Θ *is a compact subset of* $\mathbb{R}^d$ *with a finite diameter* $R_{\boldsymbol{\theta}}$.

Assumption 2. $\Omega = \{\boldsymbol{\mu}(t, \boldsymbol{\theta}) : t \in [t_0, T], \boldsymbol{\theta} \in \Theta\}$ *is a closed and bounded convex subset of* $\mathcal{R}^p$.

Assumption 3. *There exist two constants* $-\infty < c_1 < c_2 < \infty$ *such that* $c_1 \leq \boldsymbol{Y}(t) \leq c_2$ *for all* $t \in [t_0, T]$.

Assumption 4. *All partial derivatives of* $\boldsymbol{F}(\boldsymbol{\mu}(t), \boldsymbol{\theta})$ *up to order 4 with respect to* t *and* μ *exist and are continuous.*

Assumption 5. *For any* $\boldsymbol{\theta} \in \Theta$, $E_t[\boldsymbol{\mu}(t, \boldsymbol{\theta}) - \boldsymbol{\mu}(t, \boldsymbol{\theta}_0)]^2 = 0$ *if and only if* $\boldsymbol{\theta} = \boldsymbol{\theta}_0$, *the true value.*

Assumption 6. *The first and second partial derivatives,* $\frac{d\boldsymbol{\mu}(t,\boldsymbol{\theta})}{d\boldsymbol{\theta}}$ *and* $\frac{d^2\boldsymbol{\mu}(t,\boldsymbol{\theta})}{d\boldsymbol{\theta}d\boldsymbol{\theta}^T}$, *exist and are continuous and uniformly bounded for all* $t \in [t_0, T]$ *and* $\boldsymbol{\theta} \in \Theta$.

Assumption 7. *For the ODE numerical solution* $\widetilde{\boldsymbol{\mu}}(t, \boldsymbol{\theta})$, *the first and second partial derivatives,* $\frac{d\widetilde{\boldsymbol{\mu}}(t,\boldsymbol{\theta})}{d\boldsymbol{\theta}}$ *and* $\frac{d^2\widetilde{\boldsymbol{\mu}}(t,\boldsymbol{\theta})}{d\boldsymbol{\theta}d\boldsymbol{\theta}^T}$, *exist and are continuous and uniformly bounded for all* $t \in [t_0, T]$ *and* $\boldsymbol{\theta} \in \Theta$.

Assumption 8. *Let* $0 < c_3 < c_4 < \infty$ *be two constants. For random design points,* $t_1, \ldots, t_n$*, are i.i.d. The joint density function* $\phi(t, \boldsymbol{y})$ *of* $(t, \boldsymbol{Y})$ *satisfies* $c_3 \leq \phi(t, \boldsymbol{y}) \leq c_4$ *for all* $(t, \boldsymbol{y}) \in [t_0, T] \times [c_1, c_2]$*. Moreover,* $\max_i(t_i - t_{i-1}) = O(T/(n-1))$*.*

Assumption 9. *The true parameter* $\boldsymbol{\theta}_0$ *is an interior point of* Θ*.*

Assumption 10. $V_1 = \sigma^2\{E_t\left(\frac{d\boldsymbol{\mu}}{d\boldsymbol{\theta}}(\boldsymbol{\theta}_0)\frac{d^T\boldsymbol{\mu}}{d\boldsymbol{\theta}}(\boldsymbol{\theta}_0)\right)\}^{-1}$ *is positive definite, where* E_t *is expectation with respect to* t*.*

Assumption 11. *There exists a* $\alpha \geq 1$ *such that* $H \asymp n^{-\alpha}$*.*

Note that Assumptions 1-4 are general requirements for existence of numerical solutions of ODE models. For the consistency and asymptotic normality of the parameter estimator (49) based on the DDAM method, Theorems 2 and 3 are proved by Xue et al. (2010).

Theorem 2. *Under Assumptions 1-11,* $\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0 \to 0$*, almost surely under* $\mathcal{P}_{\boldsymbol{\theta}_0}$

Theorem 3. *Under Assumptions 1-11,* $\sqrt{n}\left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right) \to N(0, V_1)$*.*

Now, we focus on the global truncation error of the DDAM method.

Theorem 4. *Assume that* $\max_i(t_i - t_{i-1}) = O(T/(n-1))$ *then under Assumptions 1-4 and 11, the global truncation error of the data driven adaptive mesh method at step* i *for given* $\boldsymbol{\theta}$ *is given by*

$$||\boldsymbol{\varepsilon}_i||_\infty \equiv ||\boldsymbol{\mu}(s_i, \boldsymbol{\theta}) - \widetilde{\boldsymbol{\mu}}(s_i, \boldsymbol{\theta})||_\infty = O(h^{2\wedge 1+1/\alpha}).$$

Proof. For one-step numerical methods (50), such as the Euler and RK4 methods, the global truncation error

$$||\boldsymbol{\varepsilon}_i||_\infty \leq \exp(LT_i)\left[(1 + Lh')|\boldsymbol{\varepsilon}_0| + cT_i h'^p\right],$$

where $T_i = s_i - t_0$, L is the Lipschitz constant for Ψ function in (50), c is a constant, h' is the step size, and $p = 1$ or 4 for the Euler or RK4 method, respectively.

Without loss of generality we can assume $t_0 = s_0 = 0$ and $t_n = s_m = T$. For the DDAM method, if $s_i \in [0, H]$, since $\boldsymbol{\varepsilon}_0 = 0$

$$||\boldsymbol{\varepsilon}_i||_\infty \leq \exp(L_R(s_i - t_0))c_R(s_i - t_0)h^4 \leq \exp(L_R H)c_R H h^4 = O(Hh^4),$$

where $c_R > 0$ is a constant and L_R is the Lipschitz constant for the RK4 method.

If $s_i \in [H, t_1 - H]$,

$$\begin{aligned}
||\boldsymbol{\varepsilon}_i||_\infty &\leq e^{L_E(s_i - H)}\left[(1 + L_E H)||\boldsymbol{\varepsilon}_H||_\infty + c_E(s_i - H)H\right] \\
&\leq e^{L_E(s_i - H)}\left[(1 + L_E H)e^{L_R H}c_R H h^4 + c_E(s_i - H)H\right] \\
&\leq e^{L_E(s_i - H)}\left[(1 + L_E H)e^{L_R H}c_R H h^4 + c_E\left(c'\frac{T}{(n-1)} - 2H\right)H\right] \\
&= O(Hh^4) + O(H^{1+1/\alpha}) + O(H^2),
\end{aligned}$$

where $c_E > 0$ and $c' > 0$ are constants and L_E is the Lipschitz constant for the Euler method.

If $s_i \in [t_1 - H, t_1 + H]$,

$$\begin{aligned}
||\boldsymbol{\varepsilon}_i||_\infty &\leq e^{L_R(s_i - kH)}\left[(1 + L_R h)||\boldsymbol{\varepsilon}_{t_1 - H}||_\infty + c_R(s_i - kH)h^4\right] \\
&= e^{L_R(s_i - kH)}\left[(1 + L_R h)\left(O(Hh^4) + O(H^{1+1/\alpha}) + O(H^2)\right) + c_R 2Hh^4\right] \\
&= O(Hh^4) + O(H^{1+1/\alpha}) + O(H^2),
\end{aligned}$$

where $k = (t_1 - H)/H$.

Similarly, if $s_i \in [t_1 + H, t_n]$, $||\boldsymbol{\varepsilon}_i||_\infty = O(Hh^4) + O(H^{1+1/\alpha}) + O(H^2)$.

Since $\lambda = H/h$, then $||\boldsymbol{\varepsilon}_i||_\infty = O(h^{2\wedge(1+1/\alpha)})$ for $0 \leq i \leq m - 1$.

■

Lemma 1. *Assume that* $\max_i(t_i - t_{i-1}) = O(T/(n-1))$ *then under Assumptions 1-4 and 11, then* $\sup_{t\in[t_0,T]} ||\widetilde{\boldsymbol{\mu}}(t, \boldsymbol{\theta}) - \boldsymbol{\mu}(t, \boldsymbol{\theta})||_\infty = O(h^{2\wedge(1+1/\alpha)})$ *for any given* $\theta \in \Theta$ *in* (47).

Proof. By Theorem 4, the global truncation error of the DDAM method is

$$\max_{0\le i\le m-1} ||\boldsymbol{\mu}(s_i,\boldsymbol{\theta}) - \widetilde{\boldsymbol{\mu}}(s_i,\boldsymbol{\theta})||_\infty = O(h^{2\wedge(1+1/\alpha)}) \text{ for given } \theta \in \Theta.$$

The cubic Hermite interpolation (De Boor, 1978) can be used when t is not coincident with the grid points of the DDAM method. In that case,

$$\max_{t\in[t_0,T]\setminus\{s_i:0\le i\le m-1\}} ||\boldsymbol{\mu}(s_i,\boldsymbol{\theta}) - \widetilde{\boldsymbol{\mu}}(s_i,\boldsymbol{\theta})||_\infty = O(h^4).$$

Therefore,

$$\begin{aligned}\max_{t\in[t_0,T]} ||\boldsymbol{\mu}(s_i,\boldsymbol{\theta}) - \widetilde{\boldsymbol{\mu}}(s_i,\boldsymbol{\theta})||_\infty \le & \max_{t\in\{s_i:0\le i\le m-1\}} ||\boldsymbol{\mu}(s_i,\boldsymbol{\theta}) - \widetilde{\boldsymbol{\mu}}(s_i,\boldsymbol{\theta})||_\infty \\ & + \max_{t\in[t_0,T]\setminus\{s_i:0\le i\le m-1\}} ||\boldsymbol{\mu}(s_i,\boldsymbol{\theta}) - \widetilde{\boldsymbol{\mu}}(s_i,\boldsymbol{\theta})||_\infty \\ & = O(h^{2\wedge(1+1/\alpha)}) + O(h^4) = O(h^{2\wedge(1+1/\alpha)}).\end{aligned}$$

■

Theorem 5. *Under Assumptions 1-11, the prediction error* $E[Y(t') - \widetilde{\boldsymbol{\mu}}(t',\widehat{\boldsymbol{\theta}})]$ *for a fixed* $t' \in [t_0,T]$ *can be minimized by an optimal* $\lambda \equiv H/h$ *on* $[1, T/(n-1)]$ *for fixed* n *and* h.

Proof. For notation and presentation simplicity, we outline the proof for the univariate case below. The proof for the multivariate case are the same.

$$\begin{aligned}E[\mu(t',\boldsymbol{\theta_0}) - \widetilde{\mu}(t',\widehat{\boldsymbol{\theta}})]^2 &= E[\widetilde{\mu}(t',\widehat{\boldsymbol{\theta}}) - E(\widetilde{\mu}(t',\widehat{\boldsymbol{\theta}})) + E(\widetilde{\mu}(t',\widehat{\boldsymbol{\theta}})) - \mu(t',\boldsymbol{\theta_0})]^2 \\ &= E[\widetilde{\mu}(t',\widehat{\boldsymbol{\theta}}) - E(\widetilde{\mu}(t',\widehat{\boldsymbol{\theta}}))]^2 + [E(\widetilde{\mu}(t',\widehat{\boldsymbol{\theta}})) - \mu(t',\boldsymbol{\theta_0})]^2 \\ &= \mathrm{Var}(\widetilde{\mu}(t',\widehat{\boldsymbol{\theta}})) + [E(\widetilde{\mu}(t',\widehat{\boldsymbol{\theta}})) - \mu(t',\boldsymbol{\theta_0})]^2,\end{aligned}$$

where the second term

$$\begin{aligned}[E(\widetilde{\mu}(t',\widehat{\boldsymbol{\theta}}))-\mu(t',\boldsymbol{\theta_0})]^2 &= [E(\widetilde{\mu}(t',\widehat{\boldsymbol{\theta}}))-\widetilde{\mu}(t',\boldsymbol{\theta}_0)+\widetilde{\mu}(t',\boldsymbol{\theta}_0)-\mu(t',\boldsymbol{\theta_0})]^2\\ &= [E(\widetilde{\mu}(t',\widehat{\boldsymbol{\theta}}))-\widetilde{\mu}(t',\boldsymbol{\theta}_0)]^2+[\widetilde{\mu}(t',\boldsymbol{\theta}_0)-\mu(t',\boldsymbol{\theta_0})]^2\\ &\quad +[E(\widetilde{\mu}(t',\widehat{\boldsymbol{\theta}}))-\widetilde{\mu}(t,\boldsymbol{\theta}_0)][\widetilde{\mu}(t',\boldsymbol{\theta}_0)-\mu(t',\boldsymbol{\theta_0})]\\ &= I_1+I_2+I_3\end{aligned}$$

where

$$\begin{aligned}I_1 &= [E(\widetilde{\mu}(t',\widehat{\boldsymbol{\theta}}))-\widetilde{\mu}(t',\boldsymbol{\theta}_0)]^2\\ &= \left[E\left(\frac{d\widetilde{\mu}}{d\theta}^T(\boldsymbol{\theta}_0)(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)+\frac{1}{2}(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)^T\frac{d^2\widetilde{\mu}}{d\theta d\theta^T}(\boldsymbol{\theta}_*)(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)\right)\right]^2\\ &= \left[E\left(\frac{d\mu}{d\theta}^T(\boldsymbol{\theta}_0)(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)+\left(\frac{d\widetilde{\mu}}{d\theta}^T(\boldsymbol{\theta}_0)-\frac{d\mu}{d\theta}^T(\boldsymbol{\theta}_0)\right)(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)\right.\right.\\ &\quad \left.\left.+\frac{1}{2}(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)^T\frac{d^2\widetilde{\mu}}{d\theta d\theta^T}(\boldsymbol{\theta}_*)(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)\right)\right]^2\\ &= \left[\frac{d\mu}{d\theta}^T(\boldsymbol{\theta}_0)O(1/\sqrt{n})+O(H^{1\wedge(\frac{1}{2}+\frac{1}{2\alpha})}/\sqrt{n})+O(1/n)\right]^2\\ &= O(1/n),\end{aligned}$$

$$I_2 = [\widetilde{\mu}(t',\boldsymbol{\theta}_0)-\mu(t',\boldsymbol{\theta}_0)]^2 = O(H^{4\wedge(2+2/\alpha)}),$$

and

$$I_3 = [E(\widetilde{\mu}(t',\widehat{\boldsymbol{\theta}}))-\widetilde{\mu}(t',\boldsymbol{\theta}_0)][\widetilde{\mu}(t',\boldsymbol{\theta}_0)-\mu(t',\boldsymbol{\theta_0})] = O(H^{2\wedge(1+1/\alpha)}/\sqrt{n})$$

The first term

$$\mathrm{Var}(\widetilde{\mu}(t', \widehat{\boldsymbol{\theta}})) = E_t[\mathrm{Var}_{Y|t}(\widetilde{\mu}(t', \widehat{\boldsymbol{\theta}})|t)] + \mathrm{Var}_t[E_{Y|t}(\widetilde{\mu}(t', \widehat{\boldsymbol{\theta}})|t)],$$

where

$$\begin{aligned}
&E_t[\mathrm{Var}_{Y|t}(\widetilde{\mu}(t, \widehat{\boldsymbol{\theta}})|t)] \\
&\quad = E_t\left[\frac{1}{n}\left(\frac{d\widetilde{\mu}}{d\theta}^T(\boldsymbol{\theta}_0)V_1\frac{d\widetilde{\mu}}{d\theta}(\boldsymbol{\theta}_0)\right)\right] \\
&\quad \leq \frac{1}{n}\left[\frac{d\mu}{d\theta}^T(\boldsymbol{\theta}_0)V_1\frac{d\mu}{d\theta}(\boldsymbol{\theta}_0) + 2H^{1\wedge(\frac{1}{2}+\frac{1}{2\alpha})}\mathbf{1}^TV_1\frac{d\mu}{d\theta}(\boldsymbol{\theta}_0) \quad + H^{2\wedge(1+1/\alpha)}\mathbf{1}^TV_1\mathbf{1}\right] \\
&\quad = O(1/n) + O(H^{1\wedge(\frac{1}{2}+\frac{1}{2\alpha})}/n) + O(H^{2\wedge(1+1/\alpha)}/n),
\end{aligned}$$

and

$$\mathrm{Var}_t[E_{Y|t}(\widetilde{\mu}(t', \widehat{\boldsymbol{\theta}})|t)] = \mathrm{Var}_t\left[\widetilde{\mu}(t', \boldsymbol{\theta}_0)\right] = \mathrm{Var}_t\left[\mu(t', \boldsymbol{\theta}_0) + O(H^{2\wedge(1+1/\alpha)})\right] = 0.$$

Therefore,

$$\begin{aligned}
E[\mu(t', \boldsymbol{\theta_0}) - \widetilde{\mu}(t', \widehat{\boldsymbol{\theta}})]^2 &= O(1/n) + O(H^{1\wedge(1/2+\frac{1}{2\alpha})}/n) + O(H^{2\wedge(1+1/\alpha)}/n) \\
&\quad + O(H^{4\wedge(2+2/\alpha)}) + O(H^{2\wedge(1+1/\alpha)}/\sqrt{n}) \\
&= O(1/n) + O(H^{1\wedge(1/2+\frac{1}{2\alpha})}/n) + O(H^{2\wedge(1+1/\alpha)}/\sqrt{n}) \\
&= O(1/n) + \lambda^* O(h^{1\wedge(1/2+\frac{1}{2\alpha})}/n) + \lambda^{*2} O(h^{2\wedge(1+1/\alpha)}/\sqrt{n}),
\end{aligned}$$

where $\lambda^* = \lambda^{1\wedge(1/2+\frac{1}{2\alpha})}$. Hence, $E[\mu(t', \boldsymbol{\theta_0}) - \widetilde{\mu}(t', \widehat{\boldsymbol{\theta}})]^2$ is a quadratic function in λ^*. It can be minimized on $\lambda \in \left[1, \frac{T}{2(n-1)h}\right]$ for fixed n and h.

■

4.4 Simulation Studies

Here, we compare the performances of the GP, RK4, and DDAM method on simulated datasets for two different models. The GP procedure is implemented in the `CollocInfer` R package (Hooker et al., 2014). The inner and outer optimization method used in the GP procedure are PORT routines (Gay, 1990) and a Gauss-Newton algorithm (Bates and Watts, 1988), respectively. The optimization method used in the RK4 method and the adaptive method is also Gauss-Newton algorithm (Bates and Watts, 1988). A third-order B-spline with knots at each data point and the tuning parameter $\lambda^* = 10^4$ are used for the GP procedure. For both RK4 and DDAM methods, we set $h = 0.01$ and consider $H = 10h$, $H = 5h$, and $H = h$.

4.4.1 FitzHugh-Nagumo equations

We firstly consider the FitzHugh-Nagumo equations (FitzHugh, 1961; Nagumo et al., 1962) to compare the performance of the DDAM, RK4, and GP methods. The FitzHugh-Nagumo ODEs describe the behavior of spike potentials in the giant axon of squid neurons, which can be written as the following:

$$\begin{aligned} \frac{dV}{dt} &= c(V - \frac{V^3}{3} + R), \\ \frac{dR}{dt} &= -\frac{(V - a + bR)}{c}, \end{aligned} \tag{51}$$

where V is the membrane potential and R is a recovery variable.

We simulate the state variables V and R at n discrete time points from 0 to 20, $\{t_0 = 0, \ldots, t_n = 20\}$, with observation error Normal$(0, \sigma_\epsilon^2 \mathcal{I}_2)$ based on parameter $\boldsymbol{\theta}_0 = \{a = 0.2, b = 0.2, c = 3\}$ and initial conditions $\{V(t_0), R(t_0)\} = \{-1, 1\}$. We consider $\sigma_\epsilon = 1$, 0.5, and 0.1 and $n = 11$, 21, and 51 for the GP, RK4, and DDAM methods with $h = 0.01$ and $\lambda = 1$, 5, and 10.

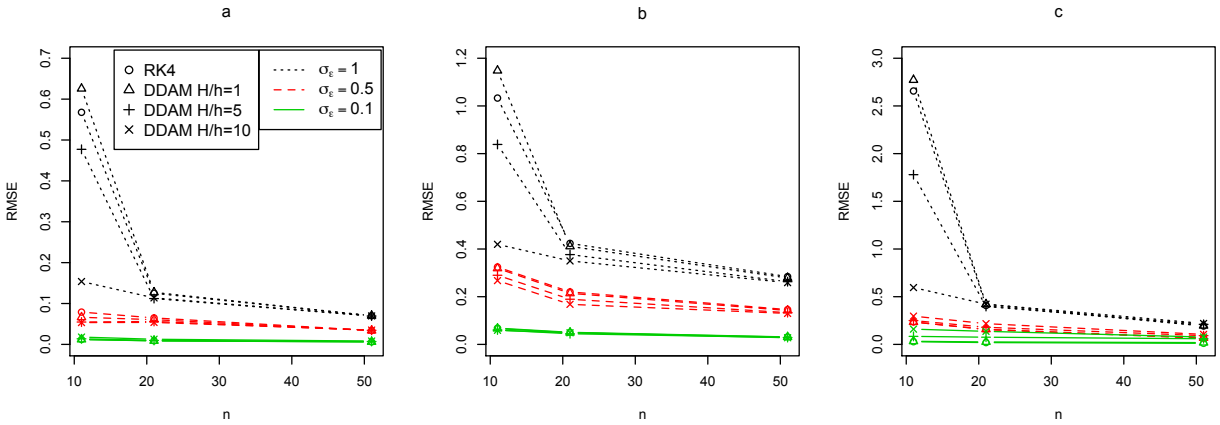

Figure 4.2: The RMSE of parameter estimates for FitzHugh-Nagumo equation (51). Note that the RMSE based on the GP method are not shown for a fair comparison between the rest methods.

Table 4.2 and Figure 4.2 show the bias and RMSE of parameter estimates under those various scenarios. Note that the RMSE based on the GP method are not shown on Figure 4.2 to allow for a fair comparison between the other methods. Both bias and RMSE for all the methods improve as n increases and σ_ϵ decreases. Both RK4 and DDAM methods have much better performance than the GP method. The DDAM method with $\lambda = 5$ or 10 has smaller RMSE over the RK4 method for small n and large σ_ϵ.

4.4.2 Gyllenberg-Webb model

Based on the two-compartment GW model (44), Alzahrani et al. (2014) proposed a modified three-compartment GW model. The key assumption of this three-compartment GW model is that the dead cells are removed from the tumor at a constant rate d. The model can be written as the following:

$$\begin{aligned} \frac{dP}{dt} &= (b - \gamma_0(N))P + \gamma_i(N)Q, \\ \frac{dQ}{dt} &= \gamma_0(N)P - (\gamma_i(N) + \mu)Q, \\ \frac{dD}{dt} &= \mu Q - dD, \end{aligned} \tag{52}$$

Table 4.2: The bias and RMSE of parameter estimates for the FitzHugh-Nagumo equations (51).

			Bias			RMSE			Time
			a	b	c	a	b	c	
$\sigma_\epsilon = 1$	$n = 11$	GP	$> 10^4$	$< -10^4$	$< -10^4$	$> 10^4$	$> 10^4$	$> 10^4$	59.98
		RK4	0.0136	−0.0942	−0.8694	0.5677	1.0330	2.6566	39.31
		DDAM $\lambda = 1$	−0.0158	−0.0280	−0.8334	0.6256	1.1482	2.7722	14.23
		DDAM $\lambda = 5$	−0.0525	−0.0321	−0.6644	0.4772	0.8385	1.7794	4.33
		DDAM $\lambda = 10$	−0.0129	−0.0861	−0.4922	0.1538	0.4196	0.5959	6.78
	$n = 21$	GP	0.0113	0.0247	0.6450	0.4201	1.9266	2.2701	71.50
		RK4	−0.0262	−0.1183	−0.2533	0.1251	0.4236	0.4080	29.27
		DDAM $\lambda = 1$	−0.0180	−0.1207	−0.2737	0.1273	0.4111	0.4215	21.61
		DDAM $\lambda = 5$	−0.0263	−0.1091	−0.2823	0.1121	0.3766	0.3971	9.54
		DDAM $\lambda = 10$	−0.0245	−0.1013	−0.3281	0.1138	0.3500	0.4216	13.18
	$n = 51$	GP	0.0368	0.0906	0.1043	0.1695	0.4870	0.7640	80.68
		RK4	−0.0084	−0.0146	−0.0924	0.0700	0.2843	0.1927	26.56
		DDAM $\lambda = 1$	−0.0093	−0.0121	−0.1043	0.0701	0.2784	0.1964	17.44
		DDAM $\lambda = 5$	−0.0129	0.0000	−0.1383	0.0694	0.2630	0.2104	14.78
		DDAM $\lambda = 10$	−0.0149	0.0053	−0.1546	0.0696	0.2590	0.2193	28.55
$\sigma_\epsilon = 0.5$	$n = 11$	GP	−0.0298	2.4047	3.7986	0.4544	20.3164	32.9224	13.16
		RK4	0.0023	−0.0763	−0.1382	0.0790	0.3244	0.2354	17.59
		DDAM $\lambda = 1$	0.0014	−0.0694	−0.1495	0.0663	0.3190	0.2341	5.73
		DDAM $\lambda = 5$	−0.0034	−0.0708	−0.1944	0.0547	0.2899	0.2496	2.00
		DDAM $\lambda = 10$	−0.0099	−0.0694	−0.2605	0.0534	0.2675	0.2962	2.04
	$n = 21$	GP	0.0088	0.2153	0.3062	0.1969	0.5892	0.7862	17.12
		RK4	0.0065	−0.0658	−0.0790	0.0648	0.2201	0.1548	15.93
		DDAM $\lambda = 1$	0.0065	−0.0600	−0.0931	0.0605	0.2142	0.1584	12.34
		DDAM $\lambda = 5$	0.0004	−0.0548	−0.1389	0.0561	0.1894	0.1797	3.68
		DDAM $\lambda = 10$	−0.0052	−0.0588	−0.1919	0.0538	0.1680	0.2174	4.82
	$n = 51$	GP	0.0518	0.1307	−0.1615	0.0957	0.2244	0.3574	28.46
		RK4	−0.0066	0.0144	−0.0197	0.0339	0.1459	0.0645	26.17
		DDAM $\lambda = 1$	−0.0073	0.0136	−0.0332	0.0339	0.1422	0.0680	5.20
		DDAM $\lambda = 5$	−0.0099	0.0134	−0.0733	0.0345	0.1331	0.0915	5.11
		DDAM $\lambda = 10$	−0.0116	0.0145	−0.0904	0.0350	0.1299	0.1049	7.35
$\sigma_\epsilon = 0.1$	$n = 11$	GP	−0.0134	0.5795	−0.6592	0.0949	0.7258	0.8727	43.78
		RK4	−0.0018	0.0003	−0.0033	0.0120	0.0682	0.0281	10.67
		DDAM $\lambda = 1$	−0.0034	0.0012	−0.0187	0.0120	0.0663	0.0326	3.69
		DDAM $\lambda = 5$	−0.0088	−0.0044	−0.0810	0.0137	0.0605	0.0841	1.24
		DDAM $\lambda = 10$	−0.0148	−0.0211	−0.1587	0.0179	0.0590	0.1600	5.35
	$n = 21$	GP	−0.0365	0.3794	0.3344	0.0479	0.3954	0.3582	83.13
		RK4	0.0008	−0.0022	−0.0041	0.0089	0.0507	0.0190	8.08
		DDAM $\lambda = 1$	−0.0007	−0.0011	−0.0187	0.0086	0.0491	0.0256	3.58
		DDAM $\lambda = 5$	−0.0057	−0.0109	−0.0741	0.0096	0.0446	0.0754	2.03
		DDAM $\lambda = 10$	−0.0101	−0.0309	−0.1363	0.0126	0.0487	0.1368	3.60
	$n = 51$	GP	0.0656	0.1238	−0.2345	0.0676	0.1342	0.2411	99.06
		RK4	0.0001	−0.0013	−0.0007	0.0061	0.0313	0.0114	7.23
		DDAM $\lambda = 1$	−0.0007	−0.0025	−0.0154	0.0061	0.0306	0.0189	4.21
		DDAM $\lambda = 5$	−0.0037	−0.0036	−0.0585	0.0071	0.0287	0.0594	4.51
		DDAM $\lambda = 10$	−0.0055	−0.0026	−0.0765	0.0082	0.0279	0.0771	6.50

Table 4.3: The bias and RMSE of parameter estimates for the Gyllenberg-Webb model (52).

			Bias					RMSE					Time
	n	Method	b	k	a	μ	d	b	k	a	μ	d	
		GP	-0.06	-60.85	-94.04	0.03	0.02	0.51	686.57	994.92	0.37	0.33	N/A
		RK4	0.32	4.71	1.37	0.36	0.38	0.79	19.64	4.69	1.03	1.07	72.72
	11	DDAM $\lambda = 1$	0.32	4.73	1.37	0.36	0.38	0.80	19.96	4.72	1.03	1.07	25.41
		DDAM $\lambda = 5$	0.33	4.65	1.42	0.37	0.39	0.80	20.21	4.74	1.03	1.07	11.24
		DDAM $\lambda = 10$	0.34	4.90	1.45	0.37	0.39	0.80	20.41	4.77	1.03	1.06	13.61
		GP	-0.07	0.67	0.25	0.01	0.01	0.24	10.73	4.10	0.14	0.16	N/A
$a_\epsilon = .3$		RK4	0.11	0.10	0.11	0.09	0.09	0.21	2.45	0.89	0.21	0.21	49.46
$b_\epsilon = 3$	21	DDAM $\lambda = 1$	0.12	0.23	0.17	0.10	0.11	0.19	2.13	0.74	0.18	0.18	20.17
$d = .5$		DDAM $\lambda = 5$	0.14	0.49	0.25	0.12	0.13	0.20	1.16	0.55	0.15	0.16	12.70
		DDAM $\lambda = 10$	0.14	0.40	0.25	0.12	0.12	0.20	1.39	0.56	0.15	0.16	16.80
		GP	-0.07	-0.45	-0.17	0.01	0.01	0.11	0.57	0.27	0.05	0.06	N/A
		RK4	0.12	0.30	0.18	0.11	0.12	0.14	0.57	0.31	0.12	0.13	46.40
	51	DDAM $\lambda = 1$	0.12	0.32	0.19	0.11	0.12	0.14	0.58	0.32	0.13	0.13	25.25
		DDAM $\lambda = 5$	0.12	0.28	0.19	0.11	0.12	0.15	1.07	0.44	0.13	0.14	24.50
		DDAM $\lambda = 10$	0.13	0.39	0.23	0.12	0.12	0.15	0.63	0.34	0.13	0.14	38.23
		GP	0.55	-323.98	-85.96	0.29	149.92	1.33	$> 10^3$	282.18	0.74	$> 10^3$	N/A
		RK4	0.05	-0.67	0.04	0.04	-0.03	0.16	3.42	0.78	0.12	0.82	72.22
	11	DDAM $\lambda = 1$	0.03	-0.52	-0.07	0.02	-0.16	0.16	3.37	1.06	0.14	0.87	28.37
		DDAM $\lambda = 5$	0.04	-0.27	0.01	0.03	-0.10	0.13	2.55	0.69	0.10	0.65	12.96
		DDAM $\lambda = 10$	0.07	-0.29	0.12	0.04	0.03	0.14	3.46	0.43	0.09	0.59	15.02
		GP	$> 10^4$	$> 10^4$	$> 10^4$	$> 10^4$	$< -10^4$	$> 10^4$	$> 10^4$	$> 10^4$	$> 10^4$	$> 10^4$	N/A
$a_\epsilon = .1$		RK4	0.03	0.12	0.07	0.02	-0.15	0.05	0.33	0.18	0.04	0.37	81.21
$b_\epsilon = 5$	21	DDAM $\lambda = 1$	0.03	0.13	0.08	0.02	-0.14	0.06	0.33	0.18	0.04	0.36	35.16
$d = 4$		DDAM $\lambda = 5$	0.04	0.17	0.10	0.03	-0.11	0.06	0.35	0.19	0.04	0.36	21.54
		DDAM $\lambda = 10$	0.04	0.21	0.12	0.03	-0.07	0.06	0.38	0.21	0.04	0.35	29.77
		GP	$> 10^4$	$> 10^4$	$> 10^4$	$> 10^4$	$< -10^4$	$> 10^4$	$> 10^4$	$> 10^4$	$> 10^4$	$> 10^4$	N/A
		RK4	0.03	0.10	0.06	0.02	-0.19	0.05	0.26	0.14	0.03	0.33	76.97
	51	DDAM $\lambda = 1$	0.03	0.11	0.07	0.02	-0.19	0.05	0.26	0.14	0.03	0.33	45.67
		DDAM $\lambda = 5$	0.04	0.14	0.08	0.03	-0.16	0.05	0.28	0.15	0.04	0.32	44.53
		DDAM $\lambda = 10$	0.04	0.16	0.09	0.03	-0.14	0.06	0.29	0.16	0.04	0.31	65.82

where

$$\gamma_i(N) = \gamma/(N+m) \text{ and } \gamma_0(N) = kN/(aN+1),$$

d is the dead cells removal rate from the tumor, $b = \beta - \mu_p$, and $\mu = \mu_q$.

We simulate the state variables P, Q, and D at n discrete time points from 0 to 20, $\{t_0 = 0, \ldots, t_n = 20\}$, with observation error Gamma(a_ϵ, b_ϵ) based on parameter $\boldsymbol{\theta}_0 = \{b = 1, k = 2, a = 1, m = 2, \mu = 0.5, \gamma = 1, d = 4\}$ and initial conditions $\{P(t_0), Q(t_0), D(t_0)\} = \{0.1, 0, 0\}$.

Table 4.3 lists the bias and RMSE of all the methods. Note that the GP method failed to reach convergence for most scenarios, so the computational time is not available. Again, for all values of λ we considered, the DDAM method has similar accuracy to the RK4 method, but has shorter computational time, especially for $\lambda = 5$.

4.5 Ecology of intestinal microbiota

The intestinal microbial community is very important to human health. It is important to understand the dynamic process of interaction between the species in the microbiota. Stein et al. (2013) extended the generalized LV equations (45) to describe microbiota ecology under external perturbations. They convert a ODE inverse problem to a linear regression problem by discretization of the ODE. Since the ODE is a continuous process, the discretization can introduce unnecessary approximation error. Here, we estimate the parameters of interest directly using our DDAM numerical solver to avoid the possible discretization error.

The data are from recent mouse experiments on antibiotic-mediated *Clostridium difficile* infection (Buffie et al., 2012). The experiment consisted of three distinct populations of mice and three mouse colonies for each population. We consider the second and third population due to small sample sizes in the first population. Both populations received the antibiotic clindamycin, but the third population was exposed to spores of the pathogen while the second population was not. We want to study whether the pathogen could make an impact on the microbial interactions between four genera of bacteria, *Barnesiella*, *und. Lachnospiraceae*, *uncl. Lachnospiraceae*, and *und. Enterobacteriaceae*.

For each population we consider a generalized LV system for the microbial interactions, which is given by

$$\frac{dx_i}{dt} = \mu_i x_i + x_i \sum_{j=1}^{4} M_{ij} x_j, \tag{53}$$

for species i, $i = 1, \ldots, 4$, where x_i is the concentration of species, μ_i is the growth rate, and M_{ij} is the effect of the interaction of species j on species i.

The estimated growth rates obtained by the DDAM method with $\lambda = 5$ for population 2 and population 3 are $(-0.36, -0.27, -0.19, -0.28)$ and $(-18.06, -15.42, -16.02, 41.56)$, respectively. The estimated interaction matrix M for two populations are shown in Figure 4.3. In population 2 the genus *uncl. Lachnospiraceae* has the strongest interactions with other species. Interactions were generally smaller in population 3 as compared with popu-

Table 4.4: Parameter estimates of the real data for the two populations. The confidence intervals are based on bootstrap estimates.

	Pop 2		Pop 3	
Parameter	Estimate	95% CI	Estimate	95% CI
μ_1	-0.36	$(-0.60, -0.24)$	-18.06	$(-23.24, 1.77)$
μ_2	-0.27	$(-0.53, -0.15)$	-15.42	$(-18.90, 8.18)$
μ_3	-0.19	$(-0.32, -0.12)$	-16.02	$(-21.70, 10.88)$
μ_4	-0.28	$(-0.53, -0.14)$	41.56	$(5.75, 45.52)$
$M_{11} \times 10^{11}$	-443.58	$(-980.82, -197.16)$	-5.49	$(-136.93, 6.74)$
$M_{21} \times 10^{11}$	-679.01	$(-1597.81, -426.41)$	26.21	$(-70.07, 44.16)$
$M_{31} \times 10^{11}$	-431.83	$(-1426.15, -25.81)$	50.37	$(-52.45, 73.31)$
$M_{41} \times 10^{11}$	697.83	(526.95, 1077.69)	-161.91	$(-468.52, 57.12)$
$M_{12} \times 10^{11}$	-429.82	$(-537.29, -162.30)$	146.73	$(-84.56, 595.25)$
$M_{22} \times 10^{11}$	-1240.94	$(-4173.15, -336.52)$	123.32	$(-48.51, 626.89)$
$M_{32} \times 10^{11}$	-1374.69	$(-4470.03, -404.13)$	120.43	$(-26.81, 703.03)$
$M_{42} \times 10^{11}$	1029.21	(45.78, 1824.22)	548.31	$(-89.65, 2240.34)$
$M_{13} \times 10^{11}$	4412.39	(1860.60, 5870.86)	-221.22	$(-2175.95, 139.26)$
$M_{23} \times 10^{11}$	3864.68	(2204.52, 7877.01)	-200.64	$(-2673.75, 11.97)$
$M_{33} \times 10^{11}$	1749.24	(848.21, 5061.24)	-209.17	$(-3423.22, -77.43)$
$M_{43} \times 10^{11}$	-2082.07	$(-4759.56, 14.51)$	-557.21	$(-6070.10, 1246.19)$
$M_{14} \times 10^{11}$	-0.19	$(-0.39, 0.22)$	15.70	$(-0.53, 27.29)$
$M_{24} \times 10^{11}$	0.59	(0.07, 1.53)	13.37	$(-6.27, 22.82)$
$M_{34} \times 10^{11}$	0.86	(0.07, 2.15)	13.90	$(-9.27, 25.50)$
$M_{44} \times 10^{11}$	-0.35	$(-0.73, -0.05)$	-35.79	$(-58.09, -4.41)$

lation 2. To obtain confidence intervals for those estimates, we use the weighted bootstrap method (Ma and Kosorok, 2005; Xue et al., 2010). Figure 4.4 shows the weighted bootstrap distributions of $\mu_1, \ldots, \mu_4$ for the two populations. The parameter estimates and confidence intervals for the two populations are shown in Table 4.4.

We investigate the performance of the model and DDAM method by predicting microbiota trajectories. In order to do so, we omit 1/4 of the observed points for each population and compare the observed hold-out data with the predicted values based on the remaining 3/4 of the observed data. Figure 4.5 shows the comparison. The high correlations between the observed and predicted data (0.94 and 0.90, respectively for populations #2 and #3) indicate a good performance of the LV model (53) and DDAM method.

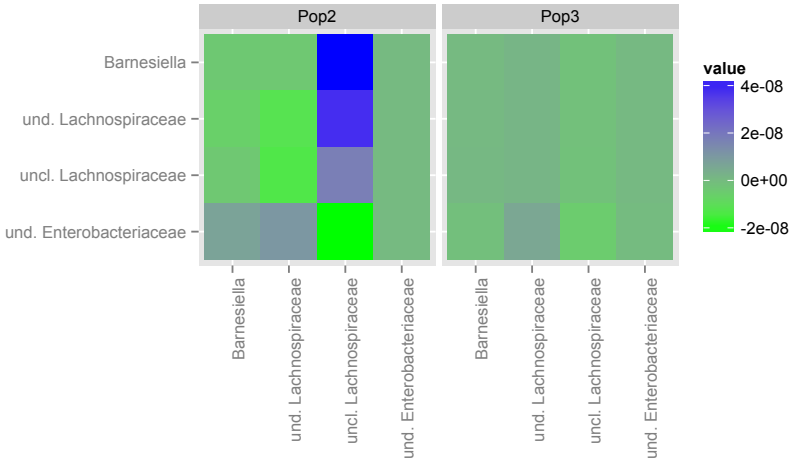

Figure 4.3: The estimated interaction matrix M (see (53)) for two populations, where M_{ij} represents the effect of genus j on i.

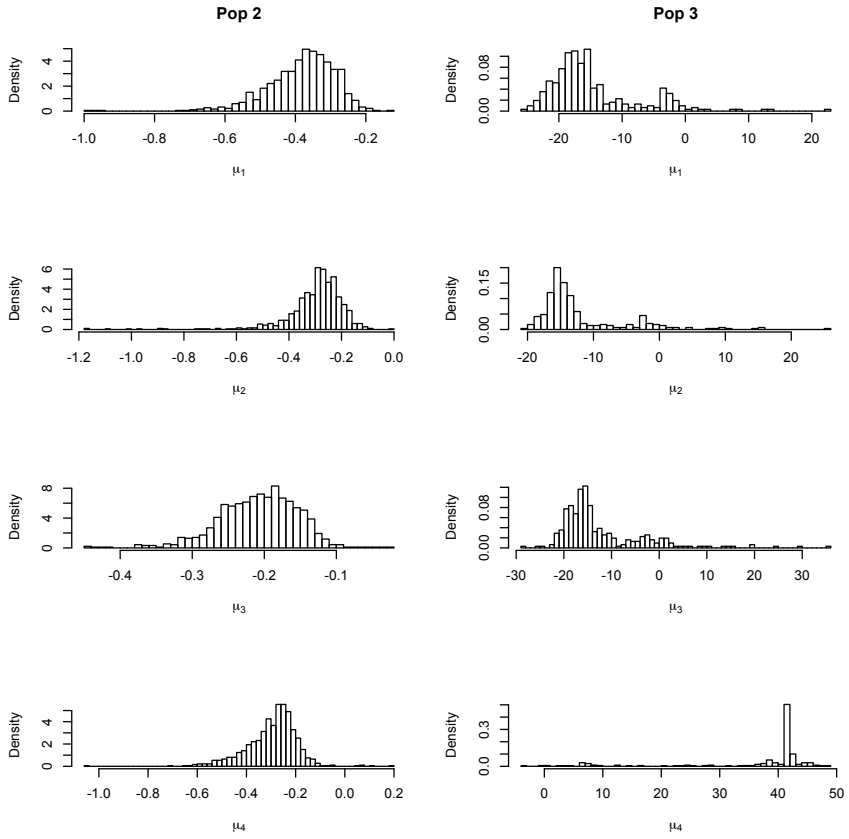

Figure 4.4: The weighted bootstrap distributions of $\mu_1, \ldots, \mu_4$ for the two populations.

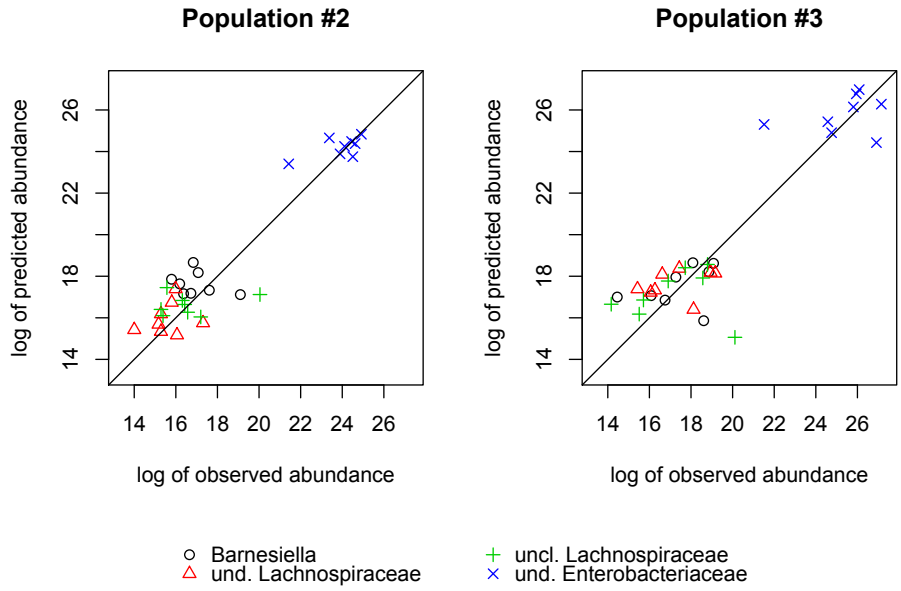

Figure 4.5: The comparison between the observed and predicted abundance. The correlations for population #2 and #3 are 0.94 and 0.90, respectively.

4.6 Conclusion and discussion

Although differential equations have a long and illustrious history in mathematical modeling for many scientific areas, very little statistical development has been done in parameter estimation and model selection for differential equation models. Our new approach, data driven adaptive mesh method, is a mixing numerical method between the Euler and 4th Runge-Kutta methods for parameter estimation in ordinary differential equations with measurement errors.

We compare the new method to the 4th Runge-Kutta method for two different models in simulation studies and also show an application for a real dataset. The new method has a shorter computation time and maintains a good accuracy at the same time. We also found that the generalized profiling procedure proposed by Ramsay et al. (2007) doesn't have good performance for the low frequency data (the time interval $t_i - t_{i-1}$ is large) compared to the new approach we developed.

The exact value of the optimal H/h ratio for the data drive adaptive mesh method for a general ordinary differential equation model still needs further investigation.

CHAPTER 5

PARAMETER INFERENCE AND MODEL SELECTION IN DETERMINISTIC AND STOCHASTIC DYNAMICAL MODELS VIA APPROXIMATE BAYESIAN COMPUTATION: MODELING A WILDLIFE EPIDEMIC

5.1 Introduction

In the study of a biological, ecological, or environmental dynamical process, the choice of underlying dynamical model (also known as the *process* model) is usually based upon expert knowledge or non-generalizable, ad hoc preference. Moreover, it is often the case that parameters of the model are not estimated using statistical functions of observed data. The objectives of this chapter are (a) *to investigate a systematic statistical approach to select a process model that is consistent with the observed data* and (b) *to produce parameter estimates and quantify associated uncertainties based on the observed data*. We undertake these goals under a hierarchical model framework and demonstrate our approach using ecological models for the transmission of chronic wasting disease (CWD) in mule deer.

In general, the hierarchical model (Berliner, 1996; Wikle, 2003) consists of three levels: a data model, a process model, and a parameter model. The data model represents measurement error in the observed data, which is very common in epidemiology, ecology and environmental science. For example, the number of deaths due to CWD in a wild population is subject to CWD test accuracy and the expense of data collection. The process model is the scientific model based on theories and simplifications of reality. Deterministic or stochastic models may be adopted as the process model. The parameter model acknowledges parameter uncertainty.

With regards to the *process* model, there could be several candidate models. For instance, in understanding the dynamics of infectious diseases in biology, ecology and environmental science, scientists can adopt a set of ordinary differential equations (ODEs) or a set of stochastic differential equations (SDEs), or a continuous time Markov chain (CTMC). A notable example is the Susceptible-Infected-Removed (SIR) model, which is a commonly used dynamical model (Anderson and May, 1992; Hethcote, 2000) in the study of disease transmission (see also Allen (2003)). Miller et al. (2006) proposed several ODE models to describe the transmission mechanism of CWD, a fatal contagious disease in cervid populations. Subsequently, an SDE model was proposed by Sun et al. (2015b) to provide more realistic description of the transmission process of CWD. There are pros and cons of those models; for example, stochastic process models allow process error but deterministic models do not. Due to their simplicity, deterministic dynamical models are typically preferred when studying a large community. Stochastic models define the probability of disease transmission between two individuals, while deterministic models describe the spread under the assumption of mass action. However for a specific dataset, the choice between deterministic or stochastic dynamical models is often subjective. Therefore, a *data-driven* approach to select between the deterministic and stochastic models based on the observed data is needed.

In many contexts model selection is typically performed via a likelihood ratio test, the Akaike information criterion or the Bayesian information criterion. However, such approaches are not suitable for the dynamical models that are often used in biology and ecology because the likelihood is intractable. Approximate Bayesian computation (ABC) is a methodology to estimate the model parameters when the likelihood is difficult to compute. A simulation-based procedure and a distance function between simulated data and the observed data are used instead of the likelihood in ABC. Various ABC algorithms have been proposed, such as rejection based ABC (Pritchard et al., 1999), regression based ABC (Beaumont et al., 2002), and ABC Markov chain Monte Carlo (MCMC) (Marjoram et al., 2003). Toni et al. (2009) developed an ABC method based on sequential Monte Carlo (SMC)

(Del Moral et al., 2006) for parameter estimation and model selection for dynamical models. This ABC SMC algorithm addresses a potential drawback of previous ABC algorithms, such as slow convergence rate, by sampling from a sequence of intermediate distributions. Beaumont (2010) provides a detailed review of ABC methods.

In this work, we incorporate the ABC SMC algorithm into a hierarchical model framework, and perform parameter estimation (with credible intervals) and dynamical model selection among a set of ODEs, SDEs, and CTMC that arise as models for the transmission of CWD. To our knowledge model selection between these types of models has not appeared previously. Since the practice of incorporating dynamical models into data models (i.e., a hierarchical framework) is becoming more common, the proposed approach may be useful in a variety of applications.

The remainder of the chapter is organized as follows. We provide a brief introduction to CWD in Section 5.2 and present the related hierarchical model framework used to investigate the transmission of CWD in Section 5.3. Section 5.4 briefly describes the ABC SMC algorithm in Toni et al. (2009). Section 5.5 presents the performance of the ABC SMC algorithm on simulated datasets. Section 5.6 shows the results based on data from two CWD epidemics. Section 5.7 concludes with a discussion.

5.2 Chronic wasting disease

Deer populations and ecosystems can be severely disrupted by the contagious prion disease, known as CWD (Miller et al., 2006). Deer populations in many U.S. states are intensely monitored due to hunting. Because of the impact of CWD on the number of deer, it is important to understand the transmission mechanisms of CWD. Several deterministic epidemic models were proposed by Miller et al. (2006) in order to portray the transmission of CWD. Here, based on those deterministic models, we derive CTMC and SDE models for CWD using the techniques described in Allen (2003, Chapter 8). Then, we implement the ABC SMC approach to the dataset studied in Miller et al. (2006). Their dataset consists of annual

observations of cumulative mortality from two distinct CWD epidemics (Figure 5.3 upper display) in captive mule deer held at the Colorado Division of Wildlife Foothills Wildlife Research Facility in Fort Collins, Colorado. The first epidemic occurred from 1974 to 1985 and the second epidemic occurred in a new deer herd from 1992 to 2001. The dataset also includes the annual number of new deer added to the herd and the per capita losses due to natural deaths and removals. We assume key model parameters, such as the direct transmission coefficient β, the per capita CWD mortality rate μ, the indirect transmission coefficient γ, the per capita rate of excretion of infectious material by infected animals ϵ, and the mass-specific rate of loss of infectious material from the environment τ, are innate characteristics of the population and the associated disease and do not change between these two epidemics. Biologists with considerable expertise in CWD have previously made the same assumption (Miller et al., 2006). Moreover, it is not possible to get accurate parameter estimates if you consider the two epidemics separately for such a small sample size.

5.3 Hierarchical model framework

A hierarchical model is a natural choice for many problems in ecology because there are typically multiple sources of uncertainty (Berliner, 1996; Wikle, 2003). There are three stages in the hierarchical model framework:

Data Model: Specify the distribution of the data given the process model.

Process Model: Describe the process conditional on process parameters.

Parameter Model: Account for uncertainty in the process parameters.

Below we develop several hierarchical models for the CWD data.

5.3.1 Data model

To allow for measurement and observation error in the observed counts, we consider two possible data models for the transmission of CWD. At time t let $S(t)$ denote the number of

susceptible animals, $I(t)$ denote the number of infected animals, $C(t)$ denote the true unobserved number of accumulated deaths from CWD, and $\tilde{C}(t)$ denote the *observed* accumulated CWD deaths. We assume that only $\tilde{C}(t)$ is observed at discrete time $t = t_0, t_1, \ldots, t_n$, and is modeled by

$$\tilde{C}(t) \sim \text{Binomial}\left(N(t); \frac{C(t)}{N(t)}\right), \tag{54}$$

where $N(t) = S(t) + I(t) + C(t)$ is the total number of animals (including deaths) at time t. As an alternative data model we also consider

$$\tilde{C}(t) \sim \text{Poisson}\left(C(t)\right). \tag{55}$$

Note that this model allows for the case where the observed number of animals at time t $\tilde{C}(t)$ to be larger than the total number of animals $N(t)$. When such an assumption is not reasonable, it is necessary to constrain $\tilde{C}(t) \leq N(t)$. Without loss of generality, we assume $C(t_0) = \tilde{C}(t_0) = 0$.

5.3.2 Process model

We consider five process models which describe the transmission mechanism of CWD. Note that combining the two different data models in Section 5.3.1 with the five process models described below, we consider a total of ten different models for CWD. The five process models, which are based on deterministic or stochastic models, are introduced below.

CWD direct transmission model CWD may be transmitted to susceptible animals directly from infected animals (Miller and Williams, 2003). We portray this direct transmission using ODE, CTMC and SDE models.

ODE model Miller et al. (2006) propose an ODE model for the direct (animal to animal) transmission of CWD given by

$$d\begin{pmatrix} S \\ I \\ C \end{pmatrix} = \begin{pmatrix} a - S(\beta I + m) \\ \beta SI - I(\mu + m) \\ \mu I \end{pmatrix} dt, \tag{56}$$

where a is the known number of susceptible animals annually added to the population via births or importation, m is the known per capita natural mortality rate, β is the unknown direct transmission coefficient (unit = time^{-1}) and μ is the unknown per capita CWD mortality rate (unit = time^{-1}). The unknown quantities to be estimated are $(\beta, \mu, S(t_0), I(t_0))$, where $S(t_0)$ and $I(t_0)$ are the unknown initial conditions.

CTMC model A continuous time Markov chain model can also be used to study a stochastic epidemic process. In a CTMC model time is continuous, but the random variables of interest are discrete. Based on the direct transmission ODE model (56), the probability equations for the CTMC model for the direct transmission of CWD are given by

$$P\left(\begin{array}{l|l} S(t+\delta) = i+1 & S(t) = i \\ I(t+\delta) = j & I(t) = j \\ C(t+\delta) = k & C(t) = k \end{array}\right) = a\delta + o(\delta), \tag{57a}$$

$$P\left(\begin{array}{l|l} S(t+\delta) = i-1 & S(t) = i \\ I(t+\delta) = j & I(t) = j \\ C(t+\delta) = k & C(t) = k \end{array}\right) = im\delta + o(\delta), \tag{57b}$$

$$P\left(\begin{array}{l|l} S(t+\delta) = i-1 & S(t) = i \\ I(t+\delta) = j+1 & I(t) = j \\ C(t+\delta) = k & C(t) = k \end{array}\right) = \beta ij\delta + o(\delta), \tag{57c}$$

$$P\left(\begin{array}{l|l} S(t+\delta)=i & S(t)=i \\ I(t+\delta)=j-1 & I(t)=j \\ C(t+\delta)=k & C(t)=k \end{array}\right) = jm\delta + o(\delta), \tag{57d}$$

$$P\left(\begin{array}{l|l} S(t+\delta)=i & S(t)=i \\ I(t+\delta)=j-1 & I(t)=j \\ C(t+\delta)=k+1 & C(t)=k \end{array}\right) = j\mu\delta + o(\delta), \tag{57e}$$

where $o(\delta) \to 0$ as the time interval $\delta \to 0$. Each probability statement in the CTMC model corresponds to a component of the ODE model (56). For example, (57a) is the probability that an additional susceptible deer is added due to birth or importation, (57b) accounts for the loss of a susceptible deer due to natural mortality, and (57d) is the corresponding probability for a loss of an infected deer due to natural mortality. More details about the derivation of a CTMC model based on an ODE model are given by Allen (2008).

SDE model SDE models are a natural extension of ODE models and they may be simpler to derive and apply than Markov chain models. For example, the transition matrix in a continuous time Markov chain model can be very complicated when there are several interacting populations (Allen and Allen, 2003; Allen et al., 2005). We consider the SDE

model for the direct transmission of CWD given by

$$d\begin{pmatrix} S \\ I \\ C \end{pmatrix} = \begin{pmatrix} a - S(\beta I + m) \\ \beta SI - I(\mu + m) \\ \mu I \end{pmatrix} dt + \boldsymbol{B} d\boldsymbol{W}, \tag{58}$$

where $\boldsymbol{W} = (W_1, W_2, W_3)^T$ is a 3-dimensional standard Wiener process and $\boldsymbol{B} = \sqrt{\boldsymbol{\Sigma}}$ is the positive definite square root of the covariance matrix with

$$\boldsymbol{\Sigma} = \begin{bmatrix} a + S(\beta I + m) & -\beta SI & 0 \\ -\beta SI & \beta SI + I(\mu + m) & -\mu I \\ 0 & -\mu I & \mu I \end{bmatrix}.$$

The derivation of the direct transmission SDE model (58) is given in Sun et al. (2015b); in the next section, we briefly illustrate the derivation of a more complex SDE model for CWD.

CWD indirect transmission model CWD may also be transmitted to susceptible animals from excreta left in the environment by infected animals. We describe this indirect transmission using both an ODE and an SDE model. The CTMC model is not suitable here, because excreta left in the environment is not a discrete variable. Let E denote the mass of infectious material in the environment.

ODE model An ODE model for the indirect transmission of CWD (Miller et al., 2006) is

$$d\begin{pmatrix} S \\ I \\ E \\ C \end{pmatrix} = \begin{pmatrix} a - S(\gamma E + m) \\ \gamma SE - I(\mu + m) \\ \epsilon I - \tau E \\ \mu I \end{pmatrix} dt, \tag{59}$$

where γ is the indirect transmission coefficient (unit = mass^{-1}time^{-1}), ϵ is the per capita rate of excretion of infectious material by infected animals (unit = time^{-1}), and τ is the mass-specific rate of loss of infectious material from the environment (unit = time^{-1}). The unknown quantities to be estimated are $(\gamma, \mu, \epsilon, \tau, S(t_0), I(t_0), E(t_0))$.

SDE model The corresponding SDE model for the indirect transmission of CWD can be derived as follows. Let $\boldsymbol{X}(t)$ denote $(S(t), I(t), E(t), C(t))^T$ and $\boldsymbol{X}_\delta = \boldsymbol{X}(t+\delta) - \boldsymbol{X}(t)$ be the increment during the time interval of length of length δ. If δ is sufficiently small, we can assume at most one animal is added, infected, or died during the time interval of length δ. The probability that more than one addition, infection, or death has occurred during that interval is of order δ^2, which can be neglected. Then we can approximate the mean of $\boldsymbol{X}_\delta$

for sufficiently small δ by

$$E[\boldsymbol{X}_\delta] \approx \begin{pmatrix} a - S(\gamma E + m) \\ \gamma SE - I(\mu + m) \\ \epsilon I - \tau E \\ \mu I \end{pmatrix} \delta \equiv \boldsymbol{f}\delta. \tag{60}$$

Furthermore, we can also approximate the covariance of $\boldsymbol{X}_\delta$ for sufficiently small δ by

$$V[\boldsymbol{X}_\delta] = E[(\boldsymbol{X}_\delta)(\boldsymbol{X}_\delta)^T] - E(\boldsymbol{X}_\delta)E(\boldsymbol{X}_\delta)^T \approx E[(\boldsymbol{X}_\delta)(\boldsymbol{X}_\delta)^T] = \boldsymbol{\Sigma}\delta, \tag{61}$$

where $\boldsymbol{\Sigma}$ is the covariance matrix given by

$$\boldsymbol{\Sigma} = \begin{bmatrix} a + S(\gamma E + m) & -\gamma SE & 0 & 0 \\ -\gamma SE & \gamma SE + I(\mu + m) & 0 & -\mu I \\ 0 & 0 & \epsilon I + \tau E & 0 \\ 0 & -\mu I & 0 & \mu I \end{bmatrix}. \tag{62}$$

The matrix $\boldsymbol{\Sigma}$ in (62) is positive definite and hence has a positive definite square root $\boldsymbol{B} = \sqrt{\boldsymbol{\Sigma}}$. It can be shown that (60) and (61) are quantities of order δ. We also assume $\boldsymbol{X}_\delta$

follows a normal distribution with mean vector $\boldsymbol{f}\delta$ and covariance matrix $\boldsymbol{B}^2\delta = \boldsymbol{\Sigma}\delta$. Thus,

$$\boldsymbol{X}(t+\delta) \approx \boldsymbol{X}(t) + \boldsymbol{f}\delta + \boldsymbol{B}\sqrt{\delta}\boldsymbol{\eta}, \tag{63}$$

where $\boldsymbol{\eta} \sim N(\boldsymbol{0}, \boldsymbol{\mathcal{I}}_{4\times 4})$ and $\boldsymbol{\mathcal{I}}$ is the identity matrix. This is exactly one iteration of the Euler-Maruyama scheme (Kloeden and Platen, 1992) for a system of SDE, which is given by

$$d\begin{pmatrix} S \\ I \\ E \\ C \end{pmatrix} = \begin{pmatrix} a - S(\gamma E + m) \\ \gamma SE - I(\mu + m) \\ \epsilon I - \tau E \\ \mu I \end{pmatrix} dt + \boldsymbol{B}d\boldsymbol{W}, \tag{64}$$

where $\boldsymbol{W} = (W_1, W_2, W_3, W_4)^T$ is a 4-dimensional standard Wiener process. The dynamical system (63) converges in the mean square sense to the system of SDEs (64) as $\delta \to 0$ (Kloeden and Platen, 1992).

5.3.3 Parameter model

We consider three sets of prior distributions. The two sets of informative prior distributions which were chosen based on expert knowledge. We selected distributions and elicited distribution moments with the assistance of N. Thompson Hobbs, an expert on CWD. The parameters β, μ, and ϵ are most likely be between 0 and 1; thus we used a Beta or uniform distribution as the informative priors for these parameters. Little is known about γ and τ and thus we used less informative prior distributions for these parameters. To investigate sensitivity to these priors, we also consider a set of noninformative prior distributions. The three sets of prior distributions for parameters $\boldsymbol{\theta}$ and initial conditions $(S(t_0), I(t_0), E(t_0))$ are shown in Table 5.1. In a non-Bayesian context, the parameter model can be omitted.

Table 5.1: The prior distributions for parameters and initial conditions. Recall β is the direct transmission coefficient, μ is the per capita CWD mortality rate, γ is the indirect transmission coefficient, ϵ is the per capita rate of excretion of infectious material by infected animals, and τ is the mass-specific rate of loss of infectious material from the environment.

	Informative I	Informative II	Noninformative	Initial	Prior
β	Beta(2,10)	U(0,1)	Gamma(0.1,0.1)	$S(t_0)$	Discrete U(10,50)
μ	Beta(2,5)	U(0,1)	Gamma(0.1,0.1)	$I(t_0)$	Discrete U(0,20)
γ	Gamma(0.01,0.01)	U(0,20)	Gamma(0.1,0.1)	$E(t_0)$	U(0,6)
ϵ	Beta(2,2)	U(0,1)	Gamma(0.1,0.1)		
τ	Gamma(0.01,0.01)	U(0,20)	Gamma(0.1,0.1)		

5.4 Approximate Bayesian computation

For all the process models described in Sections 5.3.2 and 5.3.2, we assume the data model is given in (54) or (55). That is, only $\tilde{C}(t)$ is observed at discrete time $t = t_0, t_1, \ldots, t_n$. To estimate the parameters in the process models via maximum likelihood, one needs to compute the likelihood,

$$\int \cdots \int \prod_{i=0}^{n} \left[p\left(\tilde{C}(t_i)|\boldsymbol{X}(t_i), \boldsymbol{\theta}\right) p\Big(\boldsymbol{X}(t_{i+1})|\boldsymbol{X}(t_i), \boldsymbol{\theta}\Big)\right] d\boldsymbol{X}(t_0) \cdots d\boldsymbol{X}(t_n), \tag{65}$$

where $p(\tilde{C}(t_i)|\boldsymbol{X}(t_i), \boldsymbol{\theta})$ is given by (54) or (55) and $\boldsymbol{X}(t) \equiv (S(t), I(t), C(t))^T$ or $(S(t), I(t), E(t), C(t))^T$, depending on the process model that is assumed. The likelihood (65) thus requires a multivariate integration over all unobserved state variables $\boldsymbol{X}(t)$, which can be computationally intensive or even infeasible.

To carry out Bayesian inference using a Markov chain Monte Carlo algorithm, one can treat all unobserved state variables $\boldsymbol{X}(t)$ as augmented data to avoid this complex integration (Golightly and Wilkinson, 2005, 2006, 2008). However, MCMC approaches are typically slow to converge for nonlinear multivariate dynamical models, particularly when the time interval between consecutive observations is large (Golightly and Wilkinson, 2008; Donnet and Samson, 2011), which is often the situation for ecological or environmental data. For example, in the CWD epidemic the number of deaths were recorded annually. In contrast

to the slow convergence in MCMC approaches, simulating data from the process models is relatively straightforward. For example, there are many numerical methods for solving ODEs, such as Euler's method and the Runge-Kutta method (Butcher, 2008). Based on the Markov property, simulating sample paths of a CTMC is straightforward (Allen, 2003, Chapter 5). Simple numerical solutions for SDEs include the Euler-Maruyama and the Milstein methods (Kloeden and Platen, 1992). Embedding these simulation methods in the approximate Bayesian computation with sequential Monte Carlo algorithm makes it a suitable choice for parameter inference and model selection for hierarchical models that are built upon dynamical processes.

The basic idea of ABC is that sample parameters are proposed from their corresponding prior distributions and data are simulated from the model based on the proposed parameters. The proposed parameters are accepted if the difference between the summary statistics $\eta(\cdot)$ of the simulated data D^* and the observed data D is small. The simplest ABC approach is the ABC rejection algorithm proposed by Tavare et al. (1997) and Pritchard et al. (1999). In the ABC SMC algorithm (Toni et al., 2009), N samples of parameters $\boldsymbol{\theta}$ are proposed through a sequence of intermediate distributions, $f(\boldsymbol{\theta}|\rho(\eta(D^*), \eta(D)) \leq \xi_t)$, with decreasing distance tolerances, $\xi_1 > \cdots > \xi_T > 0$, between prior distribution and target distribution, $f(\boldsymbol{\theta}|\rho(D^*, D) \leq \xi_T)$. Here, ρ is a distance function between the summary statistics $\eta(\cdot)$ of the simulated data D^* and the observed data D. For each distance tolerance ξ_t, $1 \leq t \leq T$, a new candidate sample parameter $\boldsymbol{\theta}^{**}$ is drawn from a proposal distribution $q_t(\boldsymbol{\theta}|\boldsymbol{\theta}^*)$, where $\boldsymbol{\theta}^*$ is a sample from the previous population of all proposals that have a distance tolerance ξ_{t-1}. The advantage of generating samples via a sequence of distributions is that it often avoids the problem of having low acceptance rates which is common in ABC rejection and ABC MCMC algorithms (Toni et al., 2009). The ABC SMC algorithm is given in Algorithm 2 (Toni et al. (2009) provide a similar algorithm).

Step 3 of Algorithm 2 requires selection of a proposal distribution from which to sample a set of candidate parameters. We chose the proposal distribution $q_t(\boldsymbol{\theta}|\boldsymbol{\theta}^*)$ to be a normal

Algorithm 2: The ABC SMC algorithm.

Step 1. Set the tolerance sequence $\xi_1 > \cdots > \xi_T > 0$, and $t = 1$.

Step 2. Set the sample index $i = 1$.

Step 3. Sample model index $\mathcal{M}^*$ from the model prior $\pi(\mathcal{M})$. If $t = 1$, sample $\boldsymbol{\theta}^{**}$ from the prior distribution $\pi_{\mathcal{M}^*}(\boldsymbol{\theta})$. Else, sample $\boldsymbol{\theta}^*$ from the previous population $\{\boldsymbol{\theta}^{(i)}_{t-1,\mathcal{M}^*}\}$ with weights $\omega_{t-1,\mathcal{M}^*}$ and sample $\boldsymbol{\theta}^{**}$ from the proposal distribution $q_t(\boldsymbol{\theta}|\boldsymbol{\theta}^*)$.

Step 4. If $\pi_{\mathcal{M}^*}(\boldsymbol{\theta}^{**}) = 0$, return to Step 3.

Step 5. Simulate B_t candidate datasets, $D_1, D_2, \ldots, D_{B_t}$, based on candidate parameter $\boldsymbol{\theta}^{**}$ and model $\mathcal{M}^*$. Calculate $b_t(\boldsymbol{\theta}^{**}) = \sum_{b=1}^{B_t} I(\rho(D_b, D) \leq \xi_t)$, where $I(x)$ is the indicator function.

Step 6. If $b_t(\boldsymbol{\theta}^{**}) = 0$, return to Step 3.

Step 7. Update $\mathcal{M}_t^{(i)} = \mathcal{M}^*$ and $\boldsymbol{\theta}_t^{(i)}(\mathcal{M}^*) = \boldsymbol{\theta}^{**}$. Update its weight,

$$\omega^{(i)}_{t,\mathcal{M}^*} = \begin{cases} b_t(\boldsymbol{\theta}^{**}), & \text{if } t = 1, \\ \frac{\pi_{\mathcal{M}^*}(\boldsymbol{\theta}^{**}) b_t(\boldsymbol{\theta}^{**})}{\sum_{j=1}^{N_{\mathcal{M}^*}} \omega^{(j)}_{t-1,\mathcal{M}^*} q_t\left(\boldsymbol{\theta}^{**}|\boldsymbol{\theta}^{(j)}_{t-1,\mathcal{M}^*}\right)}, & \text{if } t > 1, \end{cases}$$

where $N_{\mathcal{M}^*}$ is the number of samples for the model $\mathcal{M}^*$.

Step 8. If $i < N$, update $i = i + 1$ and go to Step 3.

Step 9. Normalize the weights for each model $\mathcal{M}$. If $t < T$, update $t = t + 1$ and go to Step 2.

or uniform random walk (that is, $\theta = \theta^* + \zeta$, where ζ is sampled from a normal or uniform distribution). We discuss this further for the specific examples below and in Table 5.3.

ABC also requires selection of a number of parameters and functions including selection of a set of summary statistics η, a distance function ρ, and two tuning parameters, $\boldsymbol{\xi} = \{\xi_1, \ldots, \xi_T\}$ and B_t. The determination of summary statistics requires some care. Marin et al. (2014) showed that model selection via ABC is only consistent when the summary statistics are either the full dataset or a set of sufficient statistics that are sufficient under all models under consideration (see Section 2.1 of Marin et al. for additional discussion of these

requirements). For our problem involving discrete-time observations of a dynamical process, no summary statistics are required because we can compare the simulated and observed datasets directly, so $\eta(D) = D$. In general one reasonable choice of the distance function $\rho(D^*, D)$ is $\frac{1}{n}\sum_i \|\mathbf{x}_i^* - \mathbf{x}_i\|$, where $\mathbf{x}_i^*$ and $\mathbf{x}_i$ are the corresponding ith observation in the simulated dataset D^* and observed dataset D, respectively, and $\|\cdot\|$ is a Euclidean norm for this case. A similar distance function is used in Toni et al. (2009). For the one dimensional CWD cumulative death data, this reduces to $\rho(D^*, D) = \frac{1}{n}\sum_i |\tilde{C}^*(t_i) - \tilde{C}(t_i)|$; hence it is equivalent to use the L^1 or infinity norm. The vector $\boldsymbol{\xi}$ such that $\xi_1 > \cdots > \xi_T > 0$ denotes the tolerance level for the cut-off for the distance function, $\rho(D^*, D) \leq \xi_i$ for $i = 1, \ldots, T$. Note that the tolerance level $\boldsymbol{\xi}$ does not have a strong influence on ABC output, but computational costs are significantly increased as $\boldsymbol{\xi}$ decreases (Marin et al., 2012). In practice one can select $\boldsymbol{\xi}$ as a small percentile of the simulated distance $\rho(D^*, D)$ (Beaumont et al., 2002). B_t is the number of simulated datasets for a given parameter $\boldsymbol{\theta}$ for stochastic models. For the deterministic model, one uses $B_t = 1$. A larger B_t may decrease the computational time of the ABC algorithm because it allows the algorithm more opportunities to generate a dataset that is sufficiently close to the observed dataset. For our model set-up we have found that using $B_t = 5$ or 10 is generally sufficient.

The outputs of the ABC SMC algorithm are the approximations of the marginal posterior distribution of the model parameter $P(\mathcal{M}|D)$, which is the proportion of times that model $\mathcal{M}$ is selected in N samples, and the marginal posterior distributions of parameters $P(\boldsymbol{\theta}|D, \mathcal{M})$ for models $\mathcal{M} = \mathcal{M}_1, \mathcal{M}_2, \ldots, \mathcal{M}_M$. We consider the ABC SMC algorithm in a model selection context where we simultaneously estimate parameters and perform model selection.

Consider the problem where one wishes to compare the posterior distributions of two models, $P(\mathcal{M}_1|D)$ and $P(\mathcal{M}_2|D)$. The ABC SMC output can be used to perform model

Table 5.2: Interpretation of the Bayes factor, where "strength of evidence" indicates evidence in favor of model 1 against model 2.

The Bayes factor B_{12}	Strength of evidence
1 to 3	Weak
3 to 20	Positive
20 to 150	Strong
>150	Very Strong

selection based on the Bayes factor,

$$B_{\mathcal{M}_1\mathcal{M}_2} = \frac{P(\mathcal{M}_1|D)/P(\mathcal{M}_2|D)}{\pi(\mathcal{M}_1)/\pi(\mathcal{M}_2)}, \tag{66}$$

where $\pi(\mathcal{M})$ is the model prior. A commonly used interpretation of the Bayes factor values, which is given by Kass and Raftery (1995), is shown in Table 5.2. In this work we adopt the model prior $\pi(\mathcal{M})$ as the discrete uniform distribution from 1 to M for models $\mathcal{M}_1$ to $\mathcal{M}_M$.

5.5 Simulation studies

We illustrate the performance of the ABC SMC algorithm on 100 simulated datasets. Each dataset includes 21 annual CWD death observations from two distinct CWD epidemics similar to the observed epidemic data in Section 5.6. We generate 100 datasets under two different scenarios: (a) the indirect transmission SDE process model (64) with the Binomial data model (54), parameters $(\gamma_0, \mu_0, \epsilon_0, \tau_0) = (0.15, 0.20, 0.50, 1.70)$, and a set of initial conditions for each epidemic given by $(S(t_0), I(t_0), E(t_0)) = (24, 5, 4.04)$ and $(22, 2, 0.87)$; (b) the direct transmission CTMC process model (57) with the Binomial data model (54), parameters $(\beta_0, \mu_0) = (0.04, 0.30)$, and initial conditions $(S(t_0), I(t_0)) = (12, 14)$ and $(30, 5)$. The parameters and initial conditions were selected so that the simulated trajectories are similar to the observed data (Section 5.6). We apply the ABC SMC algorithm on each dataset for parameter estimation and model selection among the ten models (five process models and

two data models) described in Section 5.3. The set-up for the ABC SMC algorithm is the same as the set-up we used for the observed real data and is described in Section 5.6.

To investigate model selection performance of the ABC SMC algorithm, we record the number of times that the true model (the indirect transmission SDE process model (64) with the Binomial data model (54) or the direct transmission CTMC process model (57) with the Binomial data model (54)) has the highest posterior model probability $P(\mathcal{M}|D)$ among the ten models for the 100 simulated datasets. We compute the Bayes factor between the true model and the model that has the highest probability for 100 simulated datasets for two scenarios.

For the first scenario, in 71 out of the 100 simulated datasets the true model (the indirect transmission SDE process model (64) with the Binomial data model (54)) has the highest posterior model probability among the ten models. Figure 5.1 left shows the histogram of the Bayes factor in favor of the model with highest posterior model probability against the true model over the 100 simulated datasets. Note that if the true model has the highest posterior model probability then the Bayes factor is 1. In 91 out of the 100 simulated datasets, the Bayes factor is less than 1.4. In fact, there is no dataset for which the Bayes factor is larger than 2.2. Although the ABC SMC algorithm does not always select the true model as the highest probability model, it is apparent that the strength of evidence in favor of the other models is very weak. For the second scenario, similar results are obtained. The true model (the direct transmission CTMC process model (57) with the Binomial data model (54)) was selected as the best model for 64 out of 100 simulated datasets (Figure 5.1 right) and as the second best model for 28 simulated datasets. The closest model to the true model, the direct transmission CTMC process model (57) with the Poisson data model (55), is selected as the best model in 25 simulated datasets and as second best model in 50 simulated datasets.

Note that in the ABC algorithm the proposed parameter is accepted if the simulated data based on it are close enough to the observed data. If the observed data were generated from a bad model, then the simulated data from the candidate model probably will be far

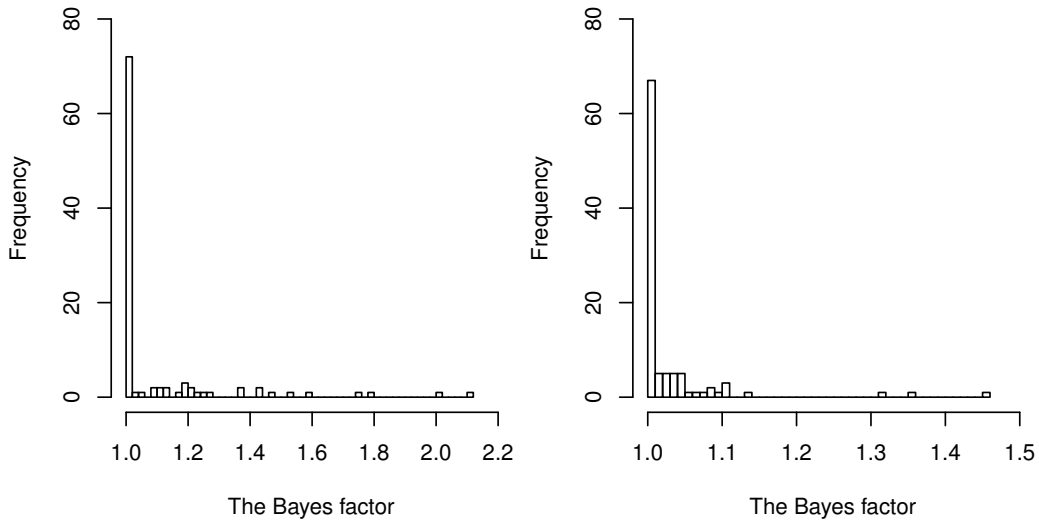

Figure 5.1: The histogram of the Bayes factor in favor of the model with the highest posterior model probability against the true model for 100 simulated datasets (scenario (a) left: the indirect transmission SDE model (64) with the Binomial data model (54); scenario (b) right: the direct transmission CTMC process model (57) with the Binomial data model (54)). Note that if the true model has the highest posterior model probability then the Bayes factor is 1.

away from the observed data. Hence, no proposed parameter will be accepted and the ABC algorithm will be unlikely to converge.

5.6 CWD application results

We apply the ABC SMC algorithm to the CWD epidemic data, which includes 21 annual CWD death observations from two distinct CWD epidemics as described in Section 5.2. To carry out model selection we compute the posterior model probability $P(\mathcal{M}|D)$ for each model and the Bayes factors to compare pairs of models. We compare the ten models in Section 5.3 and assume all models are equally likely by adopting a discrete uniform distribution as the prior distribution of the model parameter $\mathcal{M}$. We consider three sets of prior distributions for the other model parameters (Table 5.1).

For the ABC SMC algorithm the tolerance sequence is set to be $\xi = \{7, 6, 5, 4, 3.5, 3\}$, so $T = 6$, and $N = 2500$ samples of parameters are generated. The proposal distributions q_t for

Table 5.3: The proposal distributions for model parameters. We used a random walk proposal for each parameter. Below, the superscripts (t) and $(t-1)$ refer to iteration number in the ABC SMC algorithm. The initial conditions $S(t_0)$, $I(t_0)$ and $E(t_0)$ are the unknown values of the number of susceptible and infected animals, the unknown mass of infectious material in the environment at time t_0, respectively.

Parameter & proposal dist. q_t	Initial condition parameters & proposal dist. q_t
$\beta^{(t)} \sim N(\beta^{(t-1)}, 0.02^2)$	$S(t_0)^{(t)} = S(t_0)^{(t-1)} + \zeta$ where $\zeta \sim$ Discrete U$(-8, 8)$
$\mu^{(t)} \sim N(\mu^{(t-1)}, 0.2^2)$	$I(t_0)^{(t)} = I(t_0)^{(t-1)} + \zeta$ where $\zeta \sim$ Discrete U$(-3, 3)$
$\gamma^{(t)} \sim N(\gamma^{(t-1)}, 0.2^2)$	$E(t_0)^{(t)} = E(t_0)^{(t-1)} + \zeta$ where $\zeta \sim$ U$(-1, 1)$
$\epsilon^{(t)} \sim N(\epsilon^{(t-1)}, 0.2^2)$	
$\tau^{(t)} \sim N(\tau^{(t-1)}, 4)$	

parameters $\boldsymbol{\theta}$ and initial conditions $(S(t_0), I(t_0), E(t_0))$ are based on a random walk described in Table 5.3. For example, $\beta^{(t)} \sim N(\beta^{(t-1)}, 0.02^2)$. We chose a small variance for the proposal distribution for the parameters β, μ, γ and ϵ because these parameters are generally small. The parameter τ takes on larger values so we use a larger variance. The simulated data from the ODE models, (56) and (59), are generated using the `ode` function with default settings in the `deSolve` R package (Soetaert et al., 2010). The simulation method described in Allen (2003, Chapter 5) is used for simulating the CTMC process model (57). The sample paths of the SDE models, (58) and (64), are approximated using the Euler-Maruyama scheme (Kloeden and Platen, 1992) with time step $\delta = 1/12$ which is one month for the CWD epidemic data.

Posterior model probabilities $P(\mathcal{M}|D)$ and the Bayes factor in favor of the model constructed with the indirect SDE process model (64) with the Binomial data model (54) against the other models are shown in Table 5.4. The Bayes factor results indicate that the Binomial data model (54) is generally preferred over the Poisson data model (55). The marginal posterior model probability of the Binomial data model (54) practically remains unchanged under different prior sets, 0.72, 0.73, and 0.72. There is uncertainty about the form of the process model. There is a weak evidence in favor of the indirect transmission SDE process model (64) compared with the other process models considered in Section 5.3.2. It is of particular interest to biologists about whether the indirect CWD transmission model is supported by

Table 5.4: Posterior model probabilities for each model $P(\mathcal{M}|D)$ and the Bayes factor (BF) in favor of the indirect SDE process model (64) with the Binomial data model (54) against the other models for the CWD epidemic data based on 2500 samples of parameters of ABC SMC. The results are given in order of the posterior model probabilities $P(\mathcal{M}|D)$ from the informative prior set I. The three prior sets are listed in Table 5.1.

Data	Process	Informative I		Informative II		Noninformative	
Model	Model	$P(\mathcal{M}\|D)$	BF	$P(\mathcal{M}\|D)$	BF	$P(\mathcal{M}\|D)$	BF
Binom (1)	Indirect SDE (64)	0.21	1.00	0.26	1.00	0.20	1.00
Binom (1)	Direct SDE (58)	0.18	1.15	0.18	1.41	0.17	1.17
Binom (1)	Direct ODE (56)	0.13	1.55	0.06	3.99	0.13	1.52
Binom (1)	Direct CTMC (57)	0.11	1.87	0.08	3.20	0.11	1.89
Binom (1)	Indirect ODE (59)	0.09	2.43	0.15	1.71	0.11	1.83
Pois (2)	Indirect SDE (64)	0.09	2.27	0.08	3.30	0.06	3.48
Pois (2)	Direct ODE (56)	0.06	3.48	0.03	9.24	0.06	3.15
Pois (2)	Direct SDE (58)	0.05	3.87	0.06	4.20	0.08	2.64
Pois (2)	Indirect ODE (59)	0.04	4.63	0.07	3.92	0.04	4.60
Pois (2)	Direct CTMC (57)	0.03	6.17	0.03	9.66	0.04	5.06

the data because indirect transmission makes CWD control efforts very challenging (Miller et al., 2006). For the informative prior set I there are no significant differences among the other four process models in terms of the Bayes factor. Since the evidence in favor of the indirect transmission SDE model (64) is not very strong, one could consider Bayesian model averaging (Hoeting et al., 1999). Model averaging can provide more accurate forecasts if the goal is to predict the development of the disease in the future. The results based on different prior sets are similar. It appears that the ABC SMC model selection is not sensitive to the priors we used for this study. The main difference is in the ordering the direct versus the indirect ODE process model under the binomial data model.

The marginal posterior distributions for the parameters for the indirect process model (64) and Binomial data model (54) are given in Figure 5.2 and Table 5.5 for the two informative prior sets. The results from the noninformative prior set is very similar (not shown). The results under the two sets of informative prior distributions are similar except the parameter γ which models the indirect transmission rate. This parameter is particularly challenging to estimate as we are estimating the effects due to some unknown mass of infectious material in the environment (see Section 5.3.2). The influence of the prior on the estimates of γ and

Table 5.5: The marginal posterior modes and 95% highest posterior density (HPD) intervals of the parameters of the indirect transmission SDE process model (64) with the Binomial data model (54) based on the CWD epidemic data.

	Informative prior set I		Informative prior set II	
Parameter	Mode	95% HPD	Mode	95% HPD
γ (Indirect transmission rate (mass^{-1}yr^{-1}))	0.05	(0.01, 0.36)	0.16	(0.02,0.63)
μ (CWD mortality rate (yr^{-1}))	0.20	(0.10, 0.59)	0.12	(0.07,0.41)
ϵ (Per capita rate of excretion of infectious agent (yr^{-1}))	0.47	(0.15, 0.91)	0.26	(0.02,0.89)
τ (Rate of loss of infectious agent (yr^{-1}))	0.88	(0.01, 4.52)	1.71	(0.01,5.07)
$S(0)$ of the first epidemic	18	(10,26)	20	(11,37)
$I(0)$ of the first epidemic	10	(5,18)	16	(0,18)
$E(0)$ of the first epidemic	1.73	(0.97,5.84)	4.93	(0.48,5.94)
$S(0)$ of the second epidemic	48	(24,50)	28	(20,48)
$I(0)$ of the second epidemic	2	(0,5)	1	(0,5)
$E(0)$ of the second epidemic	3.47	(0.24,4.85)	1.11	(0.02,4.59)

the wide highest posterior density (HPD) intervals are probably due to the small sample size. There is also a considerable uncertainty about ϵ, the per capita rate of excretion of the infectious agent, for both prior sets. This is not surprising as this transmission mechanism is difficult to quantify. While it has been demonstrated that CWD can be transmitted via the environment, the scientific community is still trying to understand the exact mechanisms of its transmission. Although the modes of the estimated density of the parameters are different based on the different prior sets, the HPD intervals for μ, ϵ, τ, and initial conditions are similar (Table 5.5).

To assess goodness of fit, we generated 100 simulated trajectories of the cumulative number of deaths for CWD. To construct the trajectories we used the CWD indirect transmission SDE process model (64) with the Binomial data model (54) and the modes of the estimated density of the parameters from the informative prior set I listed in Table 5.5. The simulated trajectories and the observed CWD data are overlaid in Figure 5.3. The simulated trajectories based on the mode estimates from the noninformative prior set are very similar (results not shown). If the dataset had more observations we would predict a hold-out set, but this is not reasonable for these data. The simulated trajectories in Figure 5.3 are close to the observed data given that they were based on a theoretical model for the process and not from a purely empirical model based only on the observed counts.

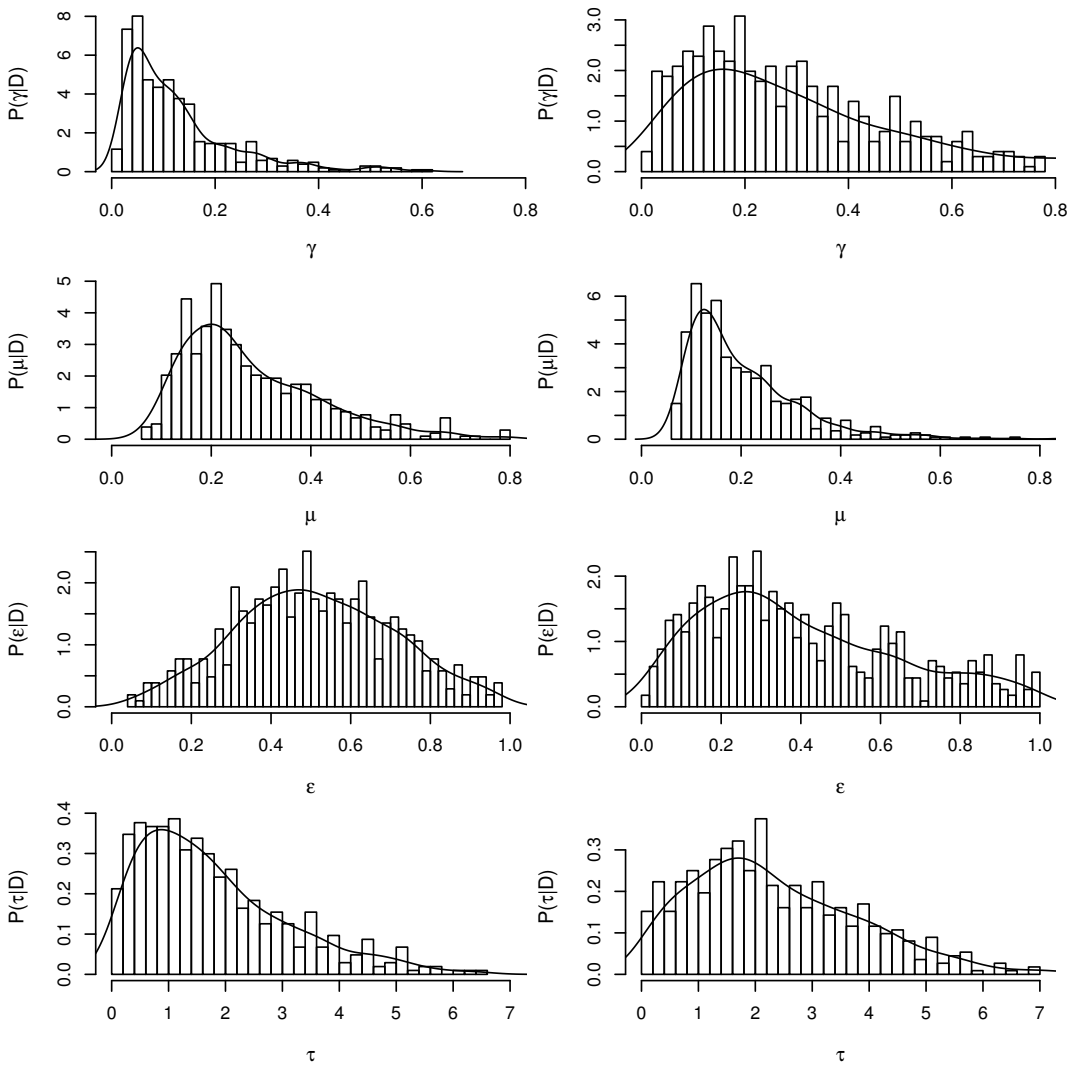

Figure 5.2: The marginal posterior distribution for the parameters of the indirect transmission SDE process model (64) with the Binomial data model (54) based on the CWD epidemic data. The left column is based on the informative prior set I and the right column is based on the informative prior set II listed in Table 5.1. A smoothed density has been super-imposed.

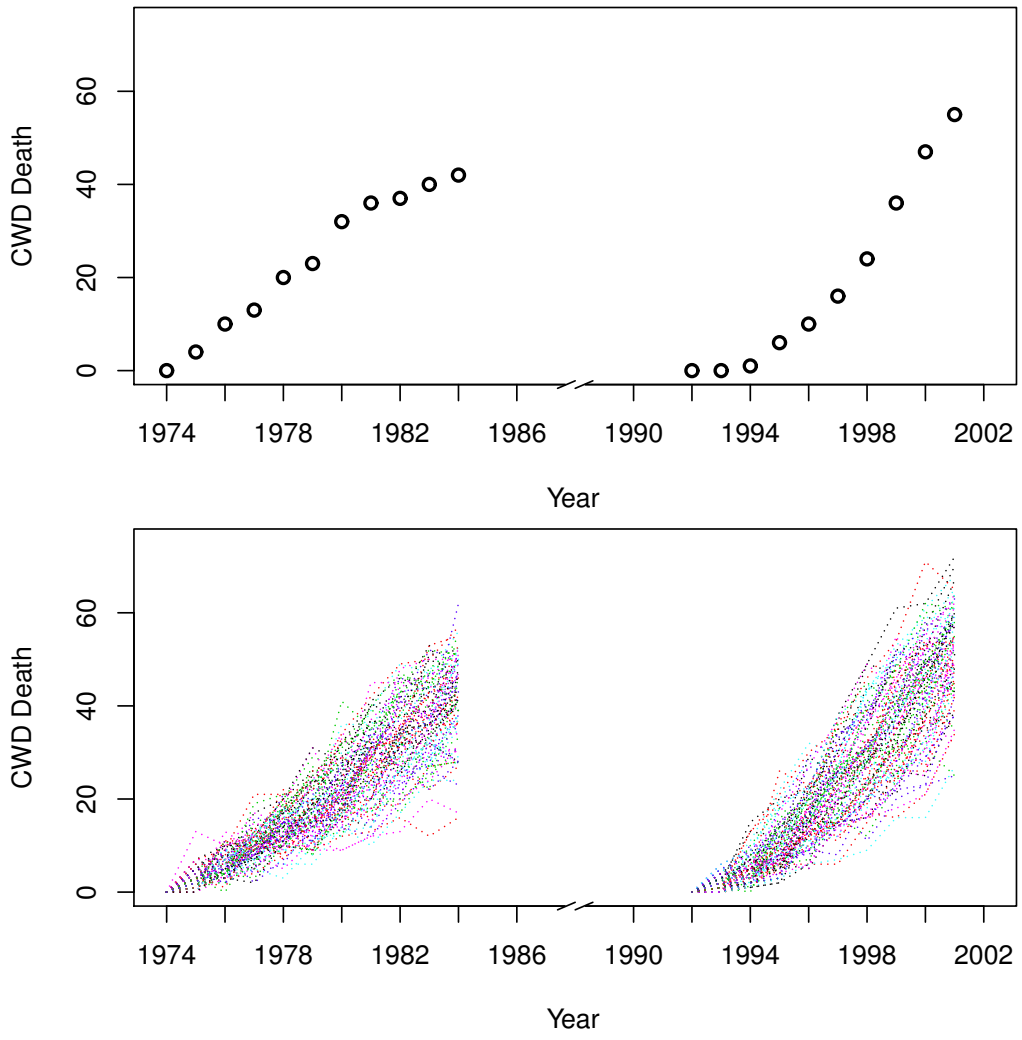

Figure 5.3: The 100 simulated trajectories of the cumulative number of deaths for CWD are obtained by using the CWD indirect transmission SDE process model (64) with the Binomial data model (54) and posterior estimates of both the parameters and the initial conditions from ABC SMC.

5.7 Conclusion and discussion

In the pursuit of gaining further understanding of ecological or environmental processes, it is important for statisticians to continue to develop tools for parameter inference and model selection for complex models. The parameters and models for the description of the transmission of CWD play a vital role in its ecological interpretation. A choice between deterministic or stochastic dynamic models is typically based on a scientific theory or personal (ad hoc) preference. We offer a systematic approach to select among these models based on empirical evidence. Although there has been considerable research focused on selecting ecological or environmental models among deterministic models, we are not aware of any previous work where deterministic and stochastic models are directly compared and selected. We illustrate a real world example which considers both deterministic and stochastic models based on the observed data via the ABC SMC algorithm. Simulation studies show the effectiveness of this approach.

We used Bayes factors for model selection because they are easy to calculate using ABC SMC. As described in Section 5.4, some care must be taken to ensure that the model selection results based on ABC are consistent. This has been an area of recent interest (Robert et al., 2011; Marin et al., 2014). There are many other options for model selection in addition to Bayes factors such as the deviance information criteria (DIC) (Spiegelhalter et al., 2002). All commonly used model selection methods have some desirable theoretical properties but there is no single method that can be used for all situations. For example, Bayes factors can be hard to estimate for some models and DIC can give incorrect results when the posterior distribution is not well summarized by the mean (Gelman et al., 2014). Most methods can give misleading results if the statistical model is inappropriate (e.g., Hoeting et al., 2006; Tenan et al., 2014). The debate about the properties of different model selection methods will continue and new model selection methods will continue to be proposed for the foreseeable future (e.g., Bové and Held, 2013; Watanabe, 2010).

The choice of distance function or summary statistics used in the ABC SMC algorithm is still an open research topic because sufficient statistics are not available for many applications. Marin et al. (2014, Section 2) give guidelines for deciding when a set of statistics is appropriate for ABC. Fearnhead and Prangle (2012) proposed a semi-automatic approach that can construct appropriate summary statistics for ABC. For the CWD epidemic models that we considered here, we found that this approach increases the complexity and decreases the efficiency of the ABC SMC algorithm.

CHAPTER 6

CONCLUSION AND FUTURE WORK

Although differential equations have a long and illustrious history in mathematical modeling for many scientific areas, very little statistical development has been done in parameter estimation and model selection for differential equation models. This dissertation focuses on estimating parameters and differential equation models with observational data.

Firstly, we propose the penalized simulated maximum likelihood approach, which provides a balanced approach to achieve accurate parameter estimates with efficient computation times for these complex stochastic models. The key idea is the introduction of a penalty term to select a better importance sampler in order to reduce the number of simulated sample paths. Then we extend the penalized simulated maximum likelihood approach to allow for measurement errors in the observed data. Note that a formal guidance of selecting the number of Monte Carlo simulations J in equation (11) is under investigation. Moreover, a formal study about the tuning parameter λ in equation (23) needs further development. Methods to account for the effect of estimating ρ in equation (21) on bootstrap intervals for the process model parameters are a topic of future research.

Secondly, a new numerical method, data driven adaptive mesh method, is developed to provide a balanced approach in accuracy and computational cost for parameter estimation in ordinary differential equations with measurement errors. The exact value of the optimal H/h ratio in minimizing the prediction error is under investigation.

Lastly, a systematic approach to select among deterministic and stochastic dynamic models based on empirical evidence is offered. Although there has been considerable research focused on selecting ecological or environmental models among deterministic models, we are not aware of any previous work where deterministic and stochastic models are directly

compared and selected. The choice of distance function or summary statistics used in the approximate Bayesian computation sequential Monte Carlo algorithm is still an open research topic because sufficient statistics are not available for many applications.

REFERENCES

Aït-Sahalia, Y. (2002). Transition densities for interest rate and other nonlinear diffusions. *The Journal of Finance*, 54(4):1361–1395.

Aït-Sahalia, Y. (2008). Closed-form likelihood expansions for multivariate diffusions. *The Annals of Statistics*, 2(36):906–937.

Allen, E. J., Allen, L. J., and Schurz, H. (2005). A comparison of persistence-time estimation for discrete and continuous stochastic population models that include demographic and environmental variability. *Mathematical biosciences*, 196(1):14–38.

Allen, L. J. (2003). *An Introduction to Stochastic Processes with Applications to Biology*. Pearson/Prentice Hall Upper Saddle River (New Jersey).

Allen, L. J. (2008). An introduction to stochastic epidemic models. In *Mathematical epidemiology*, pages 81–130. Springer.

Allen, L. J. (2011). *An introduction to stochastic processes with applications to biology*. Pearson Education Upper Saddle River, New Jersey, second edition.

Allen, L. J. and Allen, E. J. (2003). A comparison of three different stochastic population models with regard to persistence time. *Theoretical Population Biology*, 64(4):439–449.

Allen, L. J. and Burgin, A. M. (2000). Comparison of deterministic and stochastic SIS and SIR models in discrete time. *Mathematical Biosciences*, 163(1):1–33.

Alzahrani, E., Asiri, A., El-Dessoky, M., and Kuang, Y. (2014). Quiescence as an explanation of gompertzian tumor growth revisited. *Mathematical Biosciences*, 254:76–82.

Anderson, R. and May, R. (1992). *Infectious Diseases of Humans: Dynamics and Control*. Wiley Online Library.

Andersson, H. and Britton, T. (2000). *Stochastic epidemic models and their statistical analysis*. Springer Verlag.

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

Bates, D. M. and Watts, D. G. (1988). *Nonlinear regression analysis and its applications. 1988*. John Wiles & Sons, Inc.

Beaumont, M. A. (2010). Approximate Bayesian computation in evolution and ecology. *Annual Review of Ecology, Evolution, and Systematics*, 41:379–406.

Beaumont, M. A., Zhang, W., and Balding, D. J. (2002). Approximate Bayesian computation in population genetics. *Genetics*, 162(4):2025–2035.

Becker, N. (1979). The uses of epidemic models. *Biometrics*, 35(1):295–305.

Bengtsson, T., Snyder, C., and Nychka, D. (2003). Toward a nonlinear ensemble filter for high-dimensional systems. *Journal of Geophysical Research: Atmospheres (1984–2012)*, 108(D24).

Berliner, L. M. (1996). Hierarchical bayesian time series models. In *Maximum entropy and Bayesian methods*, pages 15–22. Springer.

Beskos, A., Papaspiliopoulos, O., Roberts, G. O., and Fearnhead, P. (2006). Exact and computationally efficient likelihood-based estimation for discretely observed diffusion processes (with discussion). *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(3):333–382.

Bhadra, A., Ionides, E. L., Laneri, K., Pascual, M., Bouma, M., and Dhiman, R. C. (2011). Malaria in northwest india: Data analysis via partially observed stochastic differential equation models driven by lévy noise. *Journal of the American Statistical Association*, 106(494):440–451.

Bibby, B. M., Jacobsen, M., Sørensen, M., and Universitet, K. (2004). *Estimating functions for discretely sampled diffusion-type models*. Department of Applied Mathematics and Statistics, University of Copenhagen.

Billingsley, P. (1961). *Statistical inference for Markov processes*, volume 2. University of Chicago Press Chicago.

Bové, D. S. and Held, L. (2013). Approximate Bayesian model selection with the deviance statistic. *arXiv preprint arXiv:1308.6780*.

Bretó, C., He, D., Ionides, E. L., and King, A. A. (2009). Time series analysis via mechanistic models. *The Annals of Applied Statistics*, 3(1):319–348.

Buffie, C. G., Jarchum, I., Equinda, M., Lipuma, L., Gobourne, A., Viale, A., Ubeda, C., Xavier, J., and Pamer, E. G. (2012). Profound alterations of intestinal microbiota following a single dose of clindamycin results in sustained susceptibility to clostridium difficile-induced colitis. *Infection and Immunity*, 80(1):62–73.

Butcher, J. C. (2008). *Numerical methods for ordinary differential equations*. John Wiley & Sons.

Clement, E. (2001). Estimation of diffusion processes by simulated moment methods. *Scandinavian Journal of Statistics*, 24(3):353–369.

Daley, D. J. and Gani, J. (2001). *Epidemic modelling: an introduction*. Cambridge University Press.

De Boor, C. (1978). *A practical guide to splines*. Springer, New York.

Del Moral, P., Doucet, A., and Jasra, A. (2006). Sequential Monte Carlo samplers. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(3):411–436.

Donnet, S., Foulley, J.-L., and Samson, A. (2010). Bayesian analysis of growth curves using mixed models defined by stochastic differential equations. *Biometrics*, 66(3):733–741.

Donnet, S. and Samson, A. (2011). EM algorithm coupled with particle filter for maximum likelihood parameter estimation of stochastic differential mixed-effects models. Technical Report hal-00519576 v2, Universite Paris Descartes MAP5.

Drovandi, C. C. and Pettitt, A. N. (2011). Estimation of parameters for macroparasite population evolution using approximate Bayesian computation. *Biometrics*, 67(1):225–233.

Durham, G. and Gallant, A. (2002). Numerical techniques for maximum likelihood estimation of continuous-time diffusion processes. *Journal of Business & Economic Statistics*, 20(3):297–338.

Efron, B. (1982). *The Jackknife, the Bootstrap and Other Resampling Plans*. Society for Industrial and Applied Mathematics.

Elerian, O., Chib, S., and Shephard, N. (2001). Likelihood inference for discretely observed nonlinear diffusions. *Econometrica*, 69(4):959–993.

Eraker, B. (2001). MCMC analysis of diffusion models with application to finance. *Journal of Business and Economic Statistics*, 19(2):177–191.

Euler, L. (1913). De integratione aequationum differentialium per approximationem. *Opera Omnia*, 11:424.

Fearnhead, P. and Prangle, D. (2012). Constructing summary statistics for approximate Bayesian computation: semi-automatic approximate Bayesian computation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 74(3):419–474.

FitzHugh, R. (1961). Impulses and physiological states in theoretical models of nerve membrane. *Biophysical Journal*, 1(6):445.

Gay, D. M. (1990). Usage summary for selected optimization routines. *Computing Science Technical Report*, 153:1–21.

Gelman, A., Hwang, J., and Vehtari, A. (2014). Understanding predictive information criteria for Bayesian models. *Statistics and Computing*, 24(6):997–1016.

Geweke, J. (1989). Bayesian inference in econometric models using Monte Carlo integration. *Econometrica*, 57(6):1317–1339.

Givens, G. and Hoeting, J. (2012). *Computational Statistics*. Wiley Series in Probability and Statistics, second edition.

Golightly, A. and Wilkinson, D. J. (2005). Bayesian inference for stochastic kinetic models using a diffusion approximation. *Biometrics*, 61(3):781–788.

Golightly, A. and Wilkinson, D. J. (2006). Bayesian sequential inference for nonlinear multivariate diffusions. *Statistics and Computing*, 16(4):323–338.

Golightly, A. and Wilkinson, D. J. (2008). Bayesian inference for nonlinear multivariate diffusion models observed with error. *Computational Statistics & Data Analysis*, 52(3):1674–1693.

Golightly, A. and Wilkinson, D. J. (2011). Bayesian parameter inference for stochastic biochemical network models using particle Markov chain Monte Carlo. *Interface Focus*, 1(6):807–820.

Gyllenberg, M. and Webb, G. F. (1988). Quiescence as an explanation of gompertzian tumor growth. *Growth, Development, and Aging: GDA*, 53(1-2):25–33.

Hethcote, H. W. (2000). The mathematics of infectious diseases. *SIAM review*, 42(4):599–653.

Hoeting, J. A., Davis, R. A., Merton, A. A., and Thompson, S. E. (2006). Model selection for geostatistical models. *Ecological Applications*, 16(1):87–98.

Hoeting, J. A., Madigan, D., Raftery, A. E., and Volinsky, C. T. (1999). Bayesian model averaging: a tutorial. *Statistical Science*, 14(4):382–401.

Hooker, G., Xiao, L., and Ramsay, J. (2014). *CollocInfer: Collocation Inference for Dynamic Systems*. R package version 1.0.1.

Iacus, S. M. (2009). *Simulation and inference for stochastic differential equations: with R examples*. Springer.

Ionides, E., Bretó, C., and King, A. (2006). Inference for nonlinear dynamical systems. *Proceedings of the National Academy of Sciences*, 103(49):18438–18443.

Jimenez, J., Biscay, R., and Ozaki, T. (2005). Inference methods for discretely observed continuous-time stochastic volatility models: a commented overview. *Asia-Pacific Financial Markets*, 12(2):109–141.

Kass, R. E. and Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, 90(430):773–795.

Kloeden, P. and Platen, E. (1992). *Numerical Solution of Stochastic Differential Equations*. Springer-Verlag.

Kuo, H.-H. (2006). *Introduction to stochastic integration*. Springer.

Kutta, W. (1901). Beitrag zur näherungweisen integration totaler differentialgleichungen. *Z. Math. Phys.*, 46:435–453.

Lindström, E. (2012). A regularized bridge sampler for sparsely sampled diffusions. *Statistics and Computing*, 22(2):615–623.

Lorenz, E. N. (1963). Deterministic nonperiodic flow. *Journal of the Atmospheric Sciences*, 20(2):130–141.

Ma, S. and Kosorok, M. R. (2005). Robust semiparametric M-estimation and the weighted bootstrap. *Journal of Multivariate Analysis*, 96(1):190–217.

Marin, J.-M., Pillai, N. S., Robert, C. P., and Rousseau, J. (2014). Relevant statistics for Bayesian model choice. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(5):833–859.

Marin, J.-M., Pudlo, P., Robert, C. P., and Ryder, R. J. (2012). Approximate Bayesian computational methods. *Statistics and Computing*, 22(6):1167–1180.

Marjoram, P., Molitor, J., Plagnol, V., and Tavaré, S. (2003). Markov chain Monte Carlo without likelihoods. *Proceedings of the National Academy of Sciences*, 100(26):15324–15328.

Miller, M., Hobbs, N., and Tavener, S. (2006). Dynamics of prion disease transmission in mule deer. *Ecological Applications*, 16(6):2208–2214.

Miller, M. W. and Williams, E. S. (2003). Prion disease: horizontal prion transmission in mule deer. *Nature*, 425(6953):35–36.

Nagumo, J., Arimoto, S., and Yoshizawa, S. (1962). An active pulse transmission line simulating nerve axon. *Proceedings of the IRE*, 50(10):2061–2070.

Øksendal, B. (2010). *Stochastic Differential Equations: An Introduction with Applications*. Springer.

Pastorello, S. and Rossi, E. (2010). Efficient importance sampling maximum likelihood estimation of stochastic differential equations. *Computational Statistics & Data Analysis*, 54(11):2753–2762.

Pedersen, A. (1995a). Consistency and asymptotic normality of an approximate maximum likelihood estimator for discretely observed diffusion processes. *Bernoulli*, 1(3):257–279.

Pedersen, A. (1995b). A new approach to maximum likelihood estimation for stochastic differential equations based on discrete observations. *Scandinavian Journal of Statistics*, 22(1):55–71.

Pitt, M. K. and Shephard, N. (1999). Filtering via simulation: Auxiliary particle filters. *Journal of the American statistical association*, 94(446):590–599.

Pritchard, J. K., Seielstad, M. T., Perez-Lezaun, A., and Feldman, M. W. (1999). Population growth of human Y chromosomes: a study of Y chromosome microsatellites. *Molecular Biology and Evolution*, 16(12):1791–1798.

R Development Core Team (2011). *R: A Language and Environment for Statistical Computing*. Vienna, Austria. ISBN 3-900051-07-0.

Ramsay, J. O., Hooker, G., Campbell, D., and Cao, J. (2007). Parameter estimation for differential equations: a generalized smoothing approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(5):741–796.

Richard, J. and Zhang, W. (2007). Efficient high-dimensional importance sampling. *Journal of Econometrics*, 141(2):1385–1411.

Robert, C. P., Cornuet, J.-M., Marin, J.-M., and Pillai, N. S. (2011). Lack of confidence in approximate Bayesian computation model choice. *Proceedings of the National Academy of Sciences*, 108(37):15112–15117.

Runge, C. (1895). Über die numerische auflösung von differentialgleichungen. *Mathematische Annalen*, 46(2):167–178.

Santa-Clara, P. (1997). Simulated likelihood estimation of diffusions with an application to the short term interest rate. *University of California at Los Angeles, Anderson Graduate School of Management*.

Särkkä, S. and Sottinen, T. (2008). Application of girsanov theorem to particle filtering of discretely observed continuous-time non-linear systems. *Bayesian Analysis*, 3(3):555–584.

Soetaert, K., Petzoldt, T., and Setzer, R. W. (2010). Solving differential equations in R: Package desolve. *Journal of Statistical Software*, 33(9):1–25.

Sørensen, H. (2004). Parametric inference for diffusion processes observed at discrete points in time: a survey. *International Statistical Review*, 72(3):337–354.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(4):583–639.

Stein, R. R., Bucci, V., Toussaint, N. C., Buffie, C. G., Rätsch, G., Pamer, E. G., Sander, C., and Xavier, J. B. (2013). Ecological modeling from time-series inference: insight into dynamics and stability of intestinal microbiota. *PLoS Computational Biology*, 9(12):e1003388.

Stramer, O. and Yan, J. (2007a). Asymptotics of an efficient Monte Carlo estimation for the transition density of diffusion processes. *Methodology and Computing in Applied Probability*, 9(4):483–496.

Stramer, O. and Yan, J. (2007b). On simulated likelihood of discretely observed diffusion processes and comparison to closed-form approximation. *Journal of Computational and Graphical Statistics*, 16(3):672–691.

Sun, L., Lee, C., and Hoeting, J. A. (2015a). Parameter inference and model selection in deterministic and stochastic dynamical models via approximate bayesian computation: modeling a wildlife epidemic. *arXiv preprint arXiv:1409.7715*.

Sun, L., Lee, C., and Hoeting, J. A. (2015b). A penalized simulated maximum likelihood approach in parameter estimation for stochastic differential equations. *Computational Statistics & Data Analysis*, 84:54–67.

Tavare, S., Balding, D. J., Griffiths, R., and Donnelly, P. (1997). Inferring coalescence times from DNA sequence data. *Genetics*, 145(2):505–518.

Tenan, S., O'Hara, R. B., Hendriks, I., and Tavecchia, G. (2014). Bayesian model selection: The steepest mountain to climb. *Ecological Modelling*, 283:62–69.

Tibshirani, R. (1996). Regression shrinkage and selection via the LASSO. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288.

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(31):187–202.

Varadhan, R. and Borchers, H. W. (2011). *dfoptim: Derivative-free Optimization*. R package version 2011.8-1.

Watanabe, S. (2010). Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *The Journal of Machine Learning Research*, 11:3571–3594.

Wikle, C. K. (2003). Hierarchical models in environmental science. *International Statistical Review*, 71(2):181–199.

Xue, H., Miao, H., and Wu, H. (2010). Sieve estimation of constant and time-varying coefficients in nonlinear ordinary differential equation models by considering both numerical error and measurement error. *Annals of Statistics*, 38(4):2351.

APPENDIX

A sequential Monte Carlo or particle filter scheme can be described as follows. Repeat the following steps for $i = 1, \ldots, n$,

(a). Sample $\boldsymbol{X}^{*(j)}_{-\text{obs}}(t_i) \sim q(\boldsymbol{X}_{-\text{obs}}(t_i)|\boldsymbol{X}^{(j)}(t_{i-1}), \boldsymbol{X}_{\text{obs}}(t_i))$ for $j = 1, \ldots, J$, where $\boldsymbol{X}^{(j)}(t_0) \equiv \boldsymbol{X}(t_0)$ and q is an importance sampler.

(b). Compute the weights

$$\omega_i^{(j)} = \frac{p^{(M)}(\boldsymbol{X}^{*(j)}(t_i)|\boldsymbol{X}^{(j)}(t_{i-1}))}{q(\boldsymbol{X}^{*(j)}_{-\text{obs}}(t_i)|\boldsymbol{X}^{(j)}(t_{i-1}), \boldsymbol{X}_{\text{obs}}(t_i))},$$

and $W_i^{(j)} \propto \omega_i^{(j)}$, where $\boldsymbol{X}^{*(j)}(t_i) \equiv \{\boldsymbol{X}^{*(j)}_{-\text{obs}}(t_i), \boldsymbol{X}_{\text{obs}}(t_i)\}$ and $p^{(M)}$ in the numerator is defined in (10).

(c). Resample J times with replacement from $\{\boldsymbol{X}^{*(1)}_{-\text{obs}}(t_i), \ldots, \boldsymbol{X}^{*(J)}_{-\text{obs}}(t_i)\}$ with probabilities given by $\{W_i^{(1)}, \ldots, W_i^{(J)}\}$ to obtain J equally-weighted particles $\{\boldsymbol{X}^{(1)}_{-\text{obs}}(t_i), \ldots, \boldsymbol{X}^{(J)}_{-\text{obs}}(t_i)\}$.