Statistical Inference for Nonlinear Dynamical Systems

by

Carles Bretó

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Statistics)
in The University of Michigan
2007

Doctoral Committee:

Assistant Professor Edward L. Ionides, Chair
Associate Professor Mercedes Pascual
Associate Professor Kerby A. Shedden
Assistant Professor Aaron A. King

To my family, for their love and support

ACKNOWLEDGEMENTS

This dissertation is the end result of my stay at the University of Michigan during which I have benefited from contact and varied relationships with many members of its community. I thank my advisor Professor Edward Ionides for his patient advice, generosity with his time and for the encouragement to join the statistics research community as a co-author of some of his work. In addition, Professor Ionides recruited me for the cholera project, which has served as a motivation for the statistical results in this dissertation as well as an important funding source throughout the program. I also thank the rest of the people involved in the cholera project, Mercedes Pascual in particular for her role in its organization. Access to Aaron King's cluster of computers has been crucial for reasonable computation times, which have made the whole process substantially more enjoyable. I also thank him for organizing and inviting me to the NCEAS group on inference for dynamical systems. I thank all three for very interesting and stimulating discussions.

This dissertation has benefited enormously from the resources that the university provides students with. The faculty and libraries have been a bottomless well of knowledge and the staff at the statistics department kind and helpful. I also thank my fellow graduate students, many of which offered their friendship, help and company in the long hours spent in the shared space.

Finally, I would like to add that without the love and support of my family, especially my wife Maria, this thesis would not have been possible.

Additional acknowledgements are required regarding collaborations. The second chapter has been published, with some minor changes, in the National Proceedings of the National Academy of Science of the USA (2007, Vol. 103, pp. 18438–18443) in collaboration with Edward L. Ionides and Aaron A. King. The third and fourth chapters are in the process of being submitted and are in collaboration with Edward L. Ionides and Aaron A. King.

TABLE OF CONTENTS

DEDICATION . . . . **ii**

ACKNOWLEDGEMENTS . . . . **iii**

LIST OF FIGURES . . . . **vii**

LIST OF TABLES . . . . **ix**

CHAPTER

I. Introduction 1

II. Maximum Likelihood Via Iterated Filtering 6

2.1 Introduction 6
2.2 Maximum likelihood via iterated filtering 9
2.3 Example: a compartment model for cholera 12
2.4 Results 14
2.4.1 Testing the method using simulated data 14
2.4.2 Applying the method to cholera mortality data 16
2.5 Discussion 20
2.6 Theoretical basis for MIF 22
2.7 Implementing MIF 25
2.7.1 A basic SMC algorithm 25
2.7.2 Numerical stability 26
2.7.3 Heuristics 27
2.7.4 Diagnostics 29
2.7.5 Initial values 30
2.7.6 Some recommendations for stochastic likelihood maximization . . . 32
2.7.7 Sufficient conditions for convergence of iterated filtering 34
2.7.8 Standard errors and confidence intervals 37
2.7.9 Comments on Procedure 1 39

III. Time Series Analysis Via Mechanistic Models 41

3.1 Introduction 41
3.2 A class of implicitly defined models 44
3.3 Plug-and-play inference methodology 50
3.4 A mechanistic model for competing strains of cholera 51
3.5 Discussion 62
3.6 Appendix 65
3.6.1 Proofs of Propositions III.1 and III.2 65
3.6.2 Equidispersion of Poisson Systems 71

IV. Over-dispersed Continuous Time Markov Counting Processes 72

4.1 Introduction 72
4.2 Dispersion of Continuous Time Markov Counting Processes 75
4.3 Equi-dispersed Continuous Time Markov Counting Processes 78
4.3.1 Poisson Process 79
4.3.2 Simple Linear Pure Death Process 80
4.3.3 Simple Linear Pure Birth Process 81
4.4 Over-dispersed Continuous Time Markov Counting Processes 82
4.4.1 The Poisson Gamma Process 84
4.4.2 The binomial gamma process 85
4.4.3 The binomial beta process 87

BIBLIOGRAPHY **92**

LIST OF FIGURES

Figure

2.1 Diagrammatic representation of a model for cholera population dynamics. Each individual is in S (susceptible), I (infected) or one of the classes R^j (recovered). Compartments B, C, and D allow for birth, cholera mortality, and natural death, respectively. The arrows show rates, interpreted as described in the text. 12

2.2 (A) One realization of the model in Sec. 2.3 using the parameter values in Table 2.1. (B) Historic monthly cholera mortality data for Dhaka, Bangladesh. (C) Southern oscillation index (SOI), smoothed with local quadratic regression (Cleveland et al., 1993) using a bandwidth parameter (span) of 0.12. 14

2.3 (A–C) Convergence plots for four MIFs, shown for three parameters. The dotted line shows θ^*. The parabolic lines give the sliced likelihood through $\hat{\theta}$, with the axis scale at the top right. (D–F) Corresponding closeups of the sliced likelihood. The dashed vertical line is at $\hat{\theta}$. 16

2.4 Profile likelihood for the environmental reservoir parameter. The larger of two MIF replications was plotted at each value of ω (circles), maximizing over the other parameters. Local quadratic regression (Cleveland et al., 1993; Ionides, 2005) with a bandwidth parameter (span) of 0.5 was used to estimate the profile likelihood (solid line). The dotted lines construct an approximate 99% confidence interval (Barndorff-Nielsen and Cox (1994) and 2.7.8) of $[75 \times 10^{-6}, 210 \times 10^{-6}]$. 17

2.5 Superimposed annual cycles of cholera mortality in Dhaka, 1891–1940. 17

3.1 Euler scheme corresponding to a numerical solution of the Markov chain specified by (3.2). 47

3.2 Implementation of likelihood maximization by iterated filtering. $N[\mu, \Sigma]$ corresponds to a normal random variable with mean vector μ and covariance matrix Σ. $X(t_n)$ takes values in $\mathcal{R}^{d_x}$, y_n takes values in $\mathcal{R}^{d_y}$, θ takes values in $\mathcal{R}^{d_\theta}$ and has components $\{\theta_i, i = 1, \ldots, d_\theta\}$. 52

3.3 Biweekly cholera cases for Matlab, Bangladesh, obtained from hospital records of the International Center for Diarrheal Disease Research, Bangladesh. Cases are separated by serotype into Inaba (dashed) and Ogawa (solid grey) 53

3.4 Flow diagram for cholera, including interactions between the two major serotypes. Each individual falls in one compartment: S, susceptible to both Inaba and Ogawa serotypes; I_1, infected with Inaba; I_2, infected with Ogawa; S_1, susceptible to Inaba (but immune to Ogawa); S_2, susceptible to Ogawa (but immune to Inaba); I_1^*, infected with Inaba (but immune to Ogawa); I_2^*, infected with Ogawa (but immune to Inaba); R, immune to both serotypes. Births enter S, and all individuals have a mortality rate m. 56

3.5 A simulated realization from regime A (top) and regime B (bottom), showing cases of Inaba (dashed) and Ogawa (solid grey). Compared to the data in Fig. 3.3, realizations from regime A typically have too many tall, narrow spikes in disease incidence. 60

3.6 Cross-immunity profile likelihood computed as described in chapter II yielding a 99% confidence interval for γ of (0.20, 0.61). Local quadratic regression with a bandwidth parameter (span) of 0.6 was used to estimate the profile likelihood (solid line). 61

LIST OF TABLES

Table

2.1 Parameters used for the simulation in Fig. 2.2A together with estimated parameters and their SEs where applicable. Also shown are log likelihoods, ℓ, evaluated with a Monte Carlo standard deviation of 0.1. 15

3.1 Interpretation of Figure 3.4 via the multinomial process with random rates in (3.2), with $X(t) = (S(t),\, I_1(t),\, I_2(t),\, S_1(t),\, S_2(t),\, I_1^*(t),\, I_2^*(t),\, R(t),\, B(t),\, D(t))$. Compartments B and D are introduced for demographic considerations: births are formally treated as transitions from B to S and deaths as transitions into D. All transitions not listed above have zero rate. $\xi_2(t)$ and $\xi_1(t)$ are independent gamma noise processes, both with infinitesimal variance parameter σ^2. Transition rates are noise-free unless specified otherwise. $\{s_i(t), i = 1, \dots, 6\}$ is a basis of periodic cubic B-splines, with $s_i(t)$ attaining its maximum at $t = (i-1)/6$. The population size $P(t)$ is assumed known, interpolated from census data. The birth process is treated as a covariate, i.e., the analysis is carried out conditional on the process $N_{BS}(t) = \lfloor P(t) - P(0) + \int_0^t mP(s)\,ds \rfloor$, where $\lfloor x \rfloor$ is the integer part of x. There is a small stochastic discrepancy between $S(t)+I_1(t)+I_2(t)+S_1(t)+S_2(t)+I_1^*(t)+I_2^*(t)+R(t)$ and $P(t)$. In principle, one could condition on the demographic data by including a census measurement model—we saw no compelling reason to add this extra complexity for the current purposes. Numerical solutions of sample paths were calculated using the algorithm in Figure 3.1, with $\delta = 2/365$. 57

3.2 Parameter estimates from both regimes. In both regimes, the mortality rate m is fixed at $1/38.8$ years^{-1}. The units of r, b_0, and ω are year^{-1}; σ has units year$^{1/2}$; and ρ, γ, ϕ, α, and $b_1, \dots, b_6$ are dimensionless. ℓ is the average of two log-likelihood evaluations using a particle filter with 120,000 particles. Optimization was carried out using the iterated filtering in Figure 3.2, with $M = 30$, $a = 0.95$ and $J = 15,000$. Optimization parameters were selected via diagnostic convergence plots (Ionides et al., 2006). Standard errors were derived via a Hessian approximation (Ionides et al., 2006). These standard errors are quickly obtained and give a reasonable idea of the scale of uncertainty, but profile likelihood based confidence intervals are more appropriate for careful inferences. 61

4.1 Mean and variance of the increment of Poisson, simple linear death and simple linear birth counting processes used in population modeling. The dispersion and aggregate dispersion indices show that all three processes, in spite of not being equi-dispersed on aggregate, are indeed infinitesimally equi-dispersed. This motivates the results in section 4.4 regarding over-dispersed continuous time Markov counting processes. 83

CHAPTER I

Introduction

This thesis addresses the analysis of nonlinear dynamical systems via statistical inference from time series data about parameters in stochastic models. Dynamical systems are studied in many diverse fields of engineering , social sciences and natural sciences. Examples include economics (Fernandez-Villaverde and Rubio-Ramirez, 2005; Shephard and Pitt, 1997), molecular biochemistry (Kou et al., 2005), ecology (Newman and Lindley, 2006; Thomas et al., 2005), cell biology (Ionides et al., 2004), signal processing (Anderson and Moore, 1979), meteorology (Houtekamer and Mitchell, 2001), neuroscience (Brown et al., 1998), and the study of infectious diseases (Kermack and McKendrick, 1927; Bartlett, 1960; Anderson and May, 1991; Finkenstädt and Grenfell, 2000; Ionides et al., 2006). The goal of the analysis is usually to increase understanding of the dynamic system and to make predictions. A better understanding of the system may assist in managing it and in decision making. In the study of infectious diseases, for example, it may help minimize the disease impact by informing decision makers about necessary numbers of vaccines or where and how much of the available resources should be allocated. Understanding the system may also help eradicate the disease whenever possible if considered when designing immunization or education programs.

In this thesis we consider model based analysis of dynamical systems. There are two fundamental aspects of this type of analysis: the model proposed and the statistical tools used. Regarding the model, there are a number of models that have been proposed in the literature which may be classified according to different criteria. A stochastic model is pre-requisite for time series analysis, since chance variability is required to explain the difference between the data and deterministic models. Models may treat time as discrete or continuous. Although observations will typically be at discrete times, many systems evolve in continuous time and there are dangers in using a discrete time model to analyze a system evolving in continuous time (see Glass et al., 2003). Other criteria include whether the variables in the model are real valued, as in stochastic differential equations (Karlin and Taylor, 1981; Basawa and Prakasa Rao, 1980; Øksendal, 1998), or integer valued, as in continuous time Markov chains (Brémaud, 1999) and point processes (Snyder and Miller, 1991). In addition, models may include stochasticity via variability in the underlying dynamics, or measurement error, or both.

State space models (Durbin and Koopman, 2001) are a very flexible class of Markov models that allow for different sources of stochasticity, both continuous time and discrete time modeling and both real and integer valued models. In addition, they are suited for partially observed processes. Although linear Gaussian models give an adequate representation of some processes, nonlinear behavior and non-Gaussianity are essential properties of many systems. Some previous likelihood-based methods have been proposed and, despite considerable work (Anderson and Moore, 1979; Doucet et al., 2001; Liu and West, 2001; Hürzeler and Künsch, 2001; Cappé et al., 2005; Clark and Bjornstad, 2004; Liu, 2001), statistical methodology which is readily applicable for a wide range of models, including nonlinear and/or non-Gaussian models, has remained elusive. This motivates the development of the tools presented in chapter II. While chapter II focuses on inference tools for the analysis, chapters III and IV consider the modeling aspect and introduce, from different perspectives, a novel class of models for dynamical systems composed of interacting populations of individuals. Applications of these inference tools and of the modeling to the dynamics of cholera infections are included in chapters II and III.

Chapter II introduces the method of likelihood maximization using iterated filtering for state space models, along with other tools for likelihood-based analysis. This new methodology makes maximum likelihood estimation feasible for complex nonlinear systems by exploiting the structure of state space models to avoid using a standard maximization algorithm. It is hence suited for the wide range of applications that arise in the many disciplines where dynamical systems are studied. The chapter presents both theoretical results and an application to historical cholera mortality using data collected between 1891 and 1940 in Dhaka, Bangladesh. The implementation suggested in the chapter relies on particle filtering (Doucet et al., 2001). The application to cholera focuses on both the role of a cholera reservoir in the environment and of the El Niño Southern Oscillation (ENSO) index in the disease dynamics.

Chapter III includes an application of this new inference methodology to cholera incidence data from a more recent period (1975-2005) collected in Matlab, Bangladesh. The analysis could have been based on standard continuous time Markov population models, common in ecology and epidemiology (Bartlett, 1960; Jacquez, 1996) and other disciplines. However, previous analysis of cholera data (Koelle and Pascual, 2004; Ionides et al., 2006) suggest that having stochastically varying individual rates is necessary to include sufficient variability to capture the dynamics of the disease.

In ecology and epidemiology the additional variability due to varying rates is usually referred to as *environmental stochasticity*, since it may be argued that the changes in the rates are due to changes in the environment (Renshaw, 1991). The variability produced by models with fixed individual rates is referred to as *demographic stochasticity*. Continuous time Markov processes with stochastic rates have received some attention in the literature (Snyder and Miller, 1991; Wolpert and Ickstadt, 1998), but these models might not retain the Markov property after the rates become stochastic. Continuous time Markov population models with stochastic rates seem to have been disregarded and chapters III and IV present continuous time population models with stochastic rates which do retain the Markov property.

Chapter III analyzes the more recent cholera data using a model based on the limit of coupled multinomial processes with random rates. The main focus of this application is in better understanding the strain structure of the disease and the role of strain cross-immunity in particular. A key element in the implementation of the method as presented in chapter II is simulation from the proposed model, taking advantage of recent advances in simulation-based nonlinear filtering. Analytical properties or calculations using the model, such as transition densities and their derivatives, are not required for the inference. This is a nice feature of the methodology since analytical results are not readily available in this case. Based on this, chapter III proposes an inference framework where the emphasis of the modeling is on simulation from the model, possibly using a numerical scheme. This allows analysis based on models where analytical properties are hard to derive, as is likely the case when models are based on scientifically proposed mechanisms and not chosen for statistical convenience.

While chapter III presents an instance of data analysis via over-dispersed continuous time Markov counting processes, chapter IV presents a more complete theory of these Markov counting processes. In particular, the relationship between continuous time Markov population processes with stochastic rates (such as the death processes of chapter III) and over-dispersed continuous time Markov counting processes is studied. The emphasis in this chapter is on the analytic properties and it includes the derivation of moments, infinitesimal moments and infinitesimal generators of some basic processes. In this more general framework, processes with unbounded states are dealt with and the limit of discrete time Markov processes is considered for constructing over-dispersed continuous time Markov counting processes.

CHAPTER II

Maximum Likelihood Via Iterated Filtering

2.1 Introduction

Nonlinear stochastic dynamical systems are widely used to model systems across the sciences and engineering. Such models are natural to formulate and can be analyzed mathematically and numerically. However, difficulties associated with inference from time-series data about unknown parameters in these models have been a constraint on their application. This chapter presents a new method that makes maximum likelihood estimation feasible for partially-observed nonlinear stochastic dynamical systems (also known as state-space models) where this was not previously the case. Sec. 2.2 describes the method, which is based on a sequence of filtering operations that are shown to converge to a maximum likelihood parameter estimate. We make use of recent advances in nonlinear filtering in the implementation of the algorithm. We apply the method to the study of cholera in Bangladesh in Sec. 2.4.2. We construct confidence intervals, perform residual analysis, and apply other diagnostics. The analysis, based upon a model capturing the intrinsic nonlinear dynamics of the system, reveals some effects overlooked by previous studies.

State space models have applications in many areas, including signal processing (Anderson and Moore, 1979), economics (Shephard and Pitt, 1997), cell biology (Ionides et al., 2004), meteorology (Houtekamer and Mitchell, 2001), ecology (Thomas et al., 2005), neuroscience (Brown et al., 1998), and various others (Shumway and Stoffer, 2000; Durbin and Koopman, 2001; Doucet et al., 2001). Formally, a state space model is a partially observed Markov process. Real world phenomena are often well modeled as Markov processes, constructed according to physical, chemical, or economic principles, about which one can make only noisy or incomplete observations.

It has been noted repeatedly that estimating parameters for state space models is simplest if the parameters are time-varying random variables that can be included in the state space (Anderson and Moore, 1979; Kitagawa, 1998). Estimation of parameters then becomes a matter of reconstructing unobserved random variables, and inference may proceed using standard techniques for filtering and smoothing. This approach is of limited value if the true parameters are thought not to vary with time, or to vary as a function of measured covariates rather than as random variables. A major motivation for this work has been the observation that the particle filter (Gordon et al., 1993; Kitagawa, 1998; Doucet et al., 2001; Liu, 2001; Arulampalam et al., 2002) is a conceptually simple, flexible, and effective filtering technique for which the only major drawback was the lack of a readily-applicable technique for likelihood maximization in the case of time-constant parameters. The contribution of this chapter is to show how time-varying-parameter algorithms may be harnessed for use in inference in the fixed-parameter case. The key result, Theorem II.1, shows that an appropriate limit of time-varying-parameter models can be used to locate a maximum of the fixed-parameter likelihood. This result is then used as the basis for a procedure for finding maximum likelihood estimates for previously intractable models.

We use the method to further our understanding of the mechanisms of cholera transmission. Cholera is a disease endemic to India and Bangladesh which has recently become reestablished in Africa, south Asia and South America (Sack et al., 2004). It is highly contagious, and the direct fecal-oral route of transmission is clearly important during epidemics. A slower transmission pathway, via an environmental reservoir of the pathogen, *Vibrio cholerae*, is also believed to be important, particularly in the initial phases of epidemics (Zo et al., 2002). The growth rate of *V. cholerae* depends strongly on water temperature and salinity, which can fluctuate markedly on both seasonal and interannual timescales (Huq et al., 1984; Pascual et al., 2002). Important climatic fluctuations, such as the El Niño-Southern Oscillation (ENSO), affect temperature and salinity, and operate on a timescale comparable to that associated with loss of immunity (Pascual et al., 2000; Rodó et al., 2002). It is therefore critical to disentangle the intrinsic dynamics associated with cholera transmission through the two main pathways and with loss of immunity, from the extrinsic forcing associated with climatic fluctuations (Koelle and Pascual, 2004).

We consider a model for cholera dynamics that is a continuous-time version of a discrete-time model considered by Koelle and Pascual (2004), who in turn followed a discrete-time model for measles (Finkenstädt and Grenfell, 2000). Discrete-time models have some features that are accidents of the discretization; working in continuous time avoids this, and also allows inclusion of covariates measured at disparate time intervals. Maximum likelihood inference has various convenient asymptotic properties: it is efficient, standard errors are available based on the Hessian matrix, and likelihood can be compared between different models. The transformation-invariance of maximum likelihood estimates allows modeling at a natural scale. Non-likelihood approaches typically require a variance-stabilizing transformation of the data, which may confuse scientific interpretation of results. Some previous likelihood-based methods have been proposed (Liu and West, 2001; Hürzeler and Künsch, 2001; Cappé et al., 2005; Clark and Bjornstad, 2004). However, the fact that non-likelihood-based statistical criteria such as least square prediction error (Turchin, 2003) or gradient matching (Ellner et al., 2002) are commonly applied to ecological models of the sort considered here is evidence that likelihood-based methods continue to be difficult to apply. Recent advances in nonlinear analysis have brought to the fore the need for improved statistical methods for dealing with continuous-time models with measurement error and covariates (Bjornstad and Grenfell, 2001).

2.2 Maximum likelihood via iterated filtering

A state space model consists of an unobserved Markov process, x_t, called the *state process*, and an *observation process*, y_t. Here, x_t takes values in the *state space* $\mathbb{R}^{d_x}$, and y_t in the *observation space* $\mathbb{R}^{d_y}$. The processes depend on an (unknown) vector of parameters, θ, in $\mathbb{R}^{d_\theta}$. Observations take place at discrete times, $t = 1, \dots, T$; we write the vector of concatenated observations as $y_{1:T} = (y_1, \dots, y_T)$; $y_{1:0}$ is defined to be the empty vector. A model is completely specified by the conditional transition density $f(x_t|x_{t-1}, \theta)$, the conditional distribution of the observation process $f(y_t|y_{1:t-1}, x_{1:t}, \theta) = f(y_t|x_t, \theta)$, and the initial density $f(x_0|\theta)$. Throughout, we adopt the convention that $f(\cdot \,|\, \cdot)$ is a generic density specified by its arguments, and we assume that all densities exist. The likelihood is given by the identity $f(y_{1:T}|\theta) = \prod_{t=1}^{T} f(y_t|y_{1:t-1}, \theta)$. The state process, x_t, may be defined in continuous or discrete time but only its distribution at the discrete times $t = 1, \dots, T$ directly affects the likelihood. The challenge is to find the maximum of the likelihood as a function of θ.

The basic idea of our method is to replace the original model with a closely related model, in which the time-constant parameter θ is replaced by a time-varying process θ_t. The densities $f(x_t|x_{t-1},\theta)$, $f(y_t|x_t,\theta)$ and $f(x_0|\theta)$ of the time-constant model are replaced by $f(x_t|x_{t-1},\theta_{t-1})$, $f(y_t|x_t,\theta_t)$ and $f(x_0|\theta_0)$. The process θ_t is taken to be a random walk in $\mathbb{R}^{d_\theta}$. Our main algorithm (Procedure 1 below) and its justification (Theorem II.1 in Sec. 2.6) depend only on the mean and variance of the random walk, which are defined to be

$$\begin{aligned} &E[\theta_t \mid \theta_{t-1}] = \theta_{t-1} && \operatorname{Var}(\theta_t \mid \theta_{t-1}) = \sigma^2\Sigma \\ &E[\theta_0] = \theta && \operatorname{Var}(\theta_0) = \sigma^2 c^2 \Sigma \end{aligned} \tag{2.1}$$

In practice, we use the normal distributions specified by Eq. 2.1. Here, σ and c are scalar quantities and the new model in Eq. 2.1 is identical to the fixed-parameter model when $\sigma = 0$. The objective is to obtain an estimate of θ by taking the limit as $\sigma \to 0$. Σ is typically a diagonal matrix giving the respective scales of each component of θ; more generally, it can be taken to be an arbitrary positive-definite symmetric matrix. Procedure 1 below is standard to implement, as the computationally challenging step 2(i) requires using only well-studied filtering techniques (Anderson and Moore, 1979; Arulampalam et al., 2002) to calculate

$$\begin{aligned} \hat{\theta}_t &= \hat{\theta}_t(\theta,\sigma) = E[\theta_t|y_{1:t}] \\ V_t &= V_t(\theta,\sigma) = \operatorname{Var}(\theta_t|y_{1:t-1}) \end{aligned} \tag{2.2}$$

for $t = 1,\ldots,T$. We call this procedure MIF for Maximum likelihood via Iterated Filtering.

Procedure 1. ***(MIF)***

1. *Select starting values* $\hat{\theta}^{(1)}$*, a discount factor* $0 < \alpha < 1$*, an initial variance multiplier* c^2*, and the number of iterations* N*.*

2. For n *in* $1, \dots, N$

(i) Set $\sigma_n = \alpha^{n-1}$. *For* $t = 1, \dots, T$, *evaluate* $\hat{\theta}_t^{(n)} = \hat{\theta}_t(\hat{\theta}^{(n)}, \sigma_n)$ *and* $V_{t,n} = V_t(\hat{\theta}^{(n)}, \sigma_n)$.

(ii) Set $\hat{\theta}^{(n+1)} = \hat{\theta}^{(n)} + V_{1,n} \sum_{t=1}^{T} V_{t,n}^{-1}(\hat{\theta}_t^{(n)} - \hat{\theta}_{t-1}^{(n)})$, *where* $\hat{\theta}_0^{(n)} = \hat{\theta}^{(n)}$.

3. Take $\hat{\theta}^{(N+1)}$ *to be a maximum likelihood estimate of the parameter* θ *for the fixed parameter model.*

The quantities $\hat{\theta}_t^{(n)}$ can be considered local estimates of θ, in the sense that they depend most heavily on the observations around time t. The updated estimate is a weighted average of the values $\hat{\theta}_t^{(n)}$, as explained in Sec. 2.6 and Sec. 2.7.9. A weighted average of local estimates is a heuristically reasonable estimate for the fixed "global" parameter θ. In addition, taking a weighted average and iterating to find a fixed point obviates the need for a separate optimization algorithm. Theorem II.1 asserts that (under suitable conditions) the weights in Procedure 1 result in a maximum likelihood estimate in the limit as $\sigma \to 0$. Taking a weighted average is not so desirable when the information about a parameter is concentrated in a few observations: this occurs for initial value parameters, and modifications to Procedure 1 are appropriate for these parameters (Sec. 2.7.5).

Procedure 1, with step 2(i) implemented using a sequential Monte Carlo method (see Arulampalam et al. (2002) and Sec. 2.7.1), permits flexible modeling in a wide variety of situations. The methodology requires only that Monte Carlo samples can be drawn from $f(x_t|x_{t-1})$, even if only at considerable computational expense, and that $f(y_t|x_t, \theta)$ can be numerically evaluated. We demonstrate this below with an analysis of cholera data, using a mechanistic continuous-time model. Sequential Monte Carlo is also known as "particle filtering" since each Monte Carlo realization can be viewed as a particle's trajectory through the state space. Each particle filter-

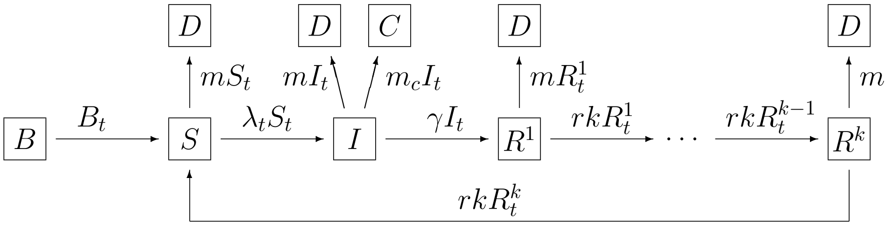

Figure 2.1: Diagrammatic representation of a model for cholera population dynamics. Each individual is in S (susceptible), I (infected) or one of the classes R^j (recovered). Compartments B, C, and D allow for birth, cholera mortality, and natural death, respectively. The arrows show rates, interpreted as described in the text.

ing step prunes particles in a way analogous to Darwinian selection. Particle filtering for fixed parameters, like natural selection without mutation, is rather ineffective. This explains heuristically why Procedure 1 is necessary to permit inference for fixed parameters via particle filtering. However, Procedure 1 and the theory of Sec. 2.6 apply more generally, and could be implemented using any suitable filter.

2.3 Example: a compartment model for cholera

In a standard epidemiological approach (Kermack and McKendrick, 1927; Bartlett, 1956), the population is divided into disease status classes. Here, we consider classes labeled susceptible (S), infected and infectious (I) and recovered ($R^1, \ldots, R^k$). The k recovery classes allow flexibility in the distribution of immune periods, a critical component of cholera modeling (Koelle and Pascual, 2004). Three additional classes B, C and D allow for birth, cholera mortality, and death from other causes respectively. S_t denotes the number of individuals in S at time t, with similar notation for other classes. We write N^{SI}_t for the integer-valued process (or its real-valued approximation) counting transitions from S to I, with corresponding definitions of N^{BS}_t, N^{SD}_t, etc. The model is shown diagrammatically in Fig. 2.1. To interpret the diagram in Fig. 2.1 as a set of coupled stochastic equations, we write

$$\begin{aligned}
dS_t &= dN_t^{BS} - dN_t^{SI} - dN_t^{SD} + dN_t^{R^kS} \\
dI_t &= dN_t^{SI} - dN_t^{IR^1} - dN_t^{IC} - dN_t^{ID} \\
dR_t^1 &= dN_t^{IR^1} - dN_t^{R^1R^2} - dN_t^{R^1D} \\
&\vdots \\
dR_t^k &= dN_t^{R^{k-1}R^k} - dN_t^{R^kS} - dN_t^{R^kD}
\end{aligned}$$

The population size P_t is presumed known, interpolated from census data. Transmission is stochastic, driven by Gaussian white noise:

$$\begin{aligned}
dN_t^{SI} &= \lambda_t S_t \, dt + \varepsilon (I_t/P_t) S_t \, dW_t \\
\lambda_t &= \beta_t I_t / P_t + \omega
\end{aligned} \tag{2.3}$$

In Eq. 2.3, we ignore stochastic effects at a demographic scale (infinitesimal variance proportional to S_t). We model the remaining transitions deterministically:

$$\begin{aligned}
dN_t^{IR^1} &= \gamma I_t \, dt; & dN_t^{R^{j-1}R^j} &= rkR_t^{j-1} \, dt; \\
dN_t^{R^kS} &= rkR_t^k \, dt; & dN_t^{SD} &= mS_t \, dt; \\
dN_t^{ID} &= mI_t \, dt; & dN_t^{R^jD} &= mR_t^j \, dt; \\
dN_t^{IC} &= m_c I_t \, dt; & dN_t^{BS} &= dP_t + mP_t \, dt.
\end{aligned} \tag{2.4}$$

Time is measured in months. Seasonality of transmission is modeled by $\log(\beta_t) = \sum_{k=0}^{5} b_k s_k(t)$, where $\{s_k(t)\}$ is a periodic cubic B-spline basis (Powell, 1981) defined so that $s_k(t)$ has a maximum at $t = 2k$ and normalized so that $\sum_{k=0}^{5} s_k(t) = 1$; ε is an *environmental stochasticity* parameter (resulting in infinitesimal variance proportional to S_t^2); ω corresponds to a non-human *reservoir* of disease; $\beta_t I_t/P_t$ is *human-to-human* transmission; $1/\gamma$ gives mean time to recovery; $1/r$ and $1/(kr^2)$ are respectively the mean and variance of the immune period; $1/m$ is the life expectancy

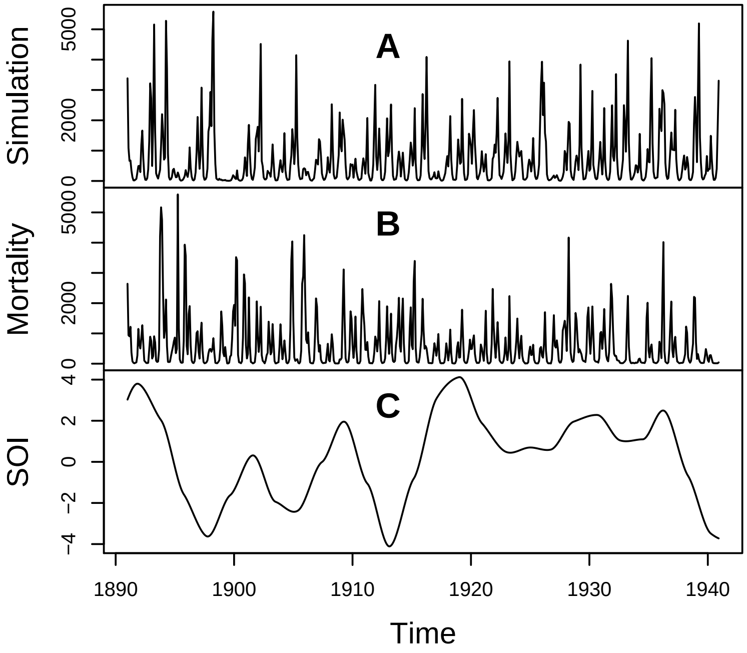

Figure 2.2: (A) One realization of the model in Sec. 2.3 using the parameter values in Table 2.1. (B) Historic monthly cholera mortality data for Dhaka, Bangladesh. (C) Southern oscillation index (SOI), smoothed with local quadratic regression (Cleveland et al., 1993) using a bandwidth parameter (span) of 0.12.

excluding cholera mortality, and m_c is the mortality rate for infected individuals. The equation for dN_t^{BS} in Eq. 2.4 is based on cholera mortality being a negligible proportion of total mortality. The stochastic system was solved numerically using the Euler-Maruyama method (Kloeden and Platen, 1999) with time increments of 1/20 month. The data on observed mortality were modeled as $y_t \sim \mathcal{N}[C_t - C_{t-1}, \tau^2(C_t - C_{t-1})^2]$, where $C_t = N_t^{IC}$. In the terminology of Sec. 2.2, the state process x_t is a vector representing counts in each compartment.

2.4 Results

2.4.1 Testing the method using simulated data

This section provides evidence that the MIF methodology successfully maximizes the likelihood. Likelihood maximization is a key tool not just for point estimation,

	θ^*	$\hat{\theta}$	SE($\hat{\theta}$)
b_0	-0.58	-0.50	0.13
b_1	4.73	4.66	0.15
b_2	-5.76	-5.58	0.42
b_3	2.37	2.30	0.14
b_4	1.69	1.77	0.08
b_5	2.56	2.47	0.09
$\omega \times 10^4$	1.76	1.81	0.26
τ	0.25	0.26	0.01
ε	0.80	0.78	0.06
$1/\gamma$	0.75		
m_c	0.046		
$1/m$	600		
$1/r$	120		
k	3		
ℓ	-3690.4	-3687.5	

Table 2.1: Parameters used for the simulation in Fig. 2.2A together with estimated parameters and their SEs where applicable. Also shown are log likelihoods, ℓ, evaluated with a Monte Carlo standard deviation of 0.1.

via the maximum likelihood estimate (MLE), but also for profile likelihood calculation, parametric bootstrap confidence intervals, and likelihood ratio hypothesis tests (Barndorff-Nielsen and Cox, 1994).

We present MIF on a simulated dataset (Fig. 2.2A), with parameter vector θ^* given in Table 2.1, based on data analysis and/or scientifically plausible values. Visually, the simulations are comparable to the data in Fig. 2.2B. Table 2.1 also contains the resulting estimated parameter vector $\hat{\theta}$ from averaging 4 MIFs, together with the maximized likelihood. A preliminary indicator that MIF has successfully maximized the likelihood is that $\ell(\hat{\theta}) > \ell(\theta^*)$. Further evidence that MIF is closely approximating the MLE comes from convergence plots and sliced likelihoods (described in Sec. 2.7.3), shown in Fig. 2.3. The SEs in Table 2.1 were calculated via the sliced likelihoods, as described in Sec. 2.7.3 and elaborated in Sec. 2.7.8. Since inference on initial values is not of primary relevance here, we do not present standard errors for their estimates. Were they required, we would recommend profile likelihood methods for uncertainty estimates of initial values. There is no asymptotic justification of the

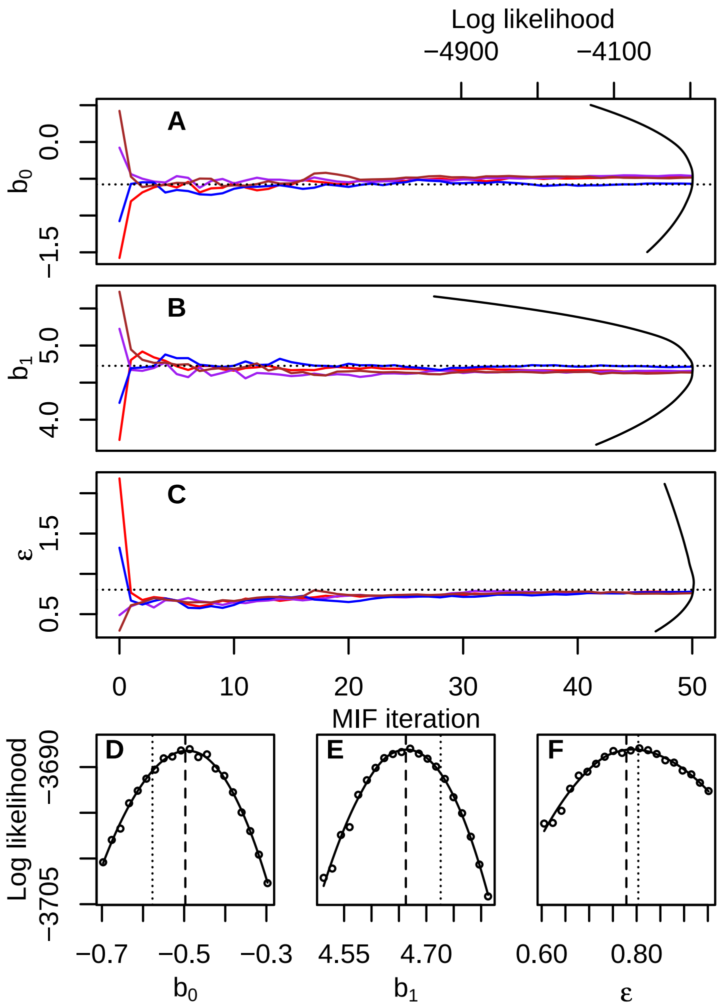

Figure 2.3: (A–C) Convergence plots for four MIFs, shown for three parameters. The dotted line shows θ^*. The parabolic lines give the sliced likelihood through $\hat{\theta}$, with the axis scale at the top right. (D–F) Corresponding closeups of the sliced likelihood. The dashed vertical line is at $\hat{\theta}$.

quadratic approximation for initial value parameters, since the information in the data about such parameters is typically concentrated in a few early time points.

2.4.2 Applying the method to cholera mortality data

We use the data in Fig. 2.2B and the model in Sec. 2.3 to address two questions: the strength of the environmental reservoir effect, and the influence of ENSO on cholera dynamics. The reader is referred to Rodó et al. (2002) and Koelle and Pascual (2004) for more extended analyses of these data. A full investigation of

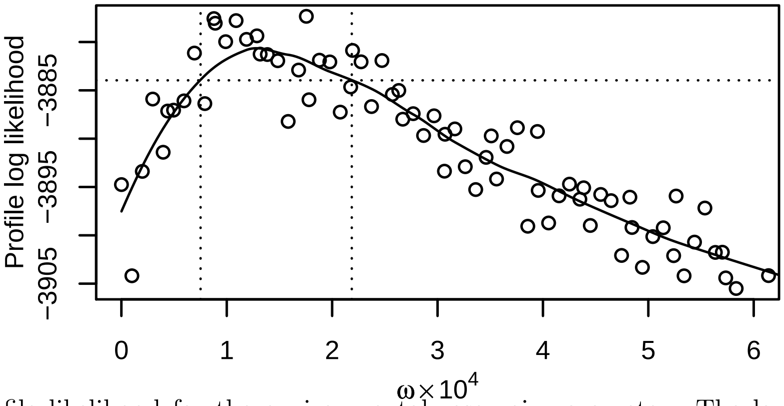

Figure 2.4: Profile likelihood for the environmental reservoir parameter. The larger of two MIF replications was plotted at each value of ω (circles), maximizing over the other parameters. Local quadratic regression (Cleveland et al., 1993; Ionides, 2005) with a bandwidth parameter (span) of 0.5 was used to estimate the profile likelihood (solid line). The dotted lines construct an approximate 99% confidence interval (Barndorff-Nielsen and Cox (1994) and 2.7.8) of $[75 \times 10^{-6}, 210 \times 10^{-6}]$.

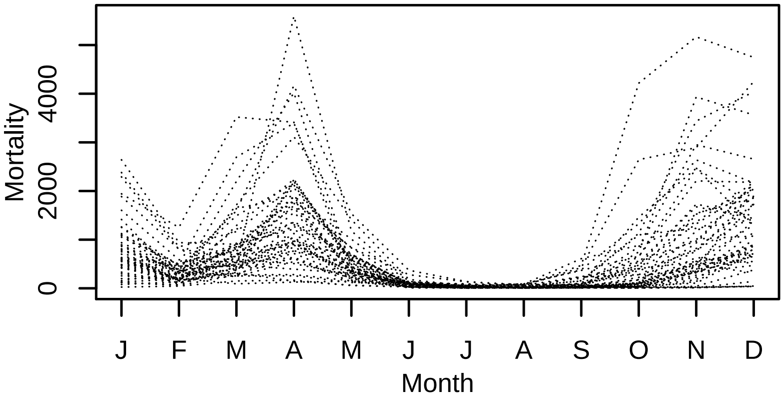

Figure 2.5: Superimposed annual cycles of cholera mortality in Dhaka, 1891–1940.

the likelihood function is challenging, due to multiple local maxima and poorly-identified combinations of parameters. Here, these problems are reduced by treating two parameters (m and r) as known. A value $k = 3$ was chosen based on preliminary analysis. The remaining 15 parameters (the first eleven parameters in Table 2.1 and the initial values S_0, I_0, R_0^1, R_0^2, R_0^3, constrained to sum to P_0) were estimated. There is scope for future work by relaxing these assumptions.

For cholera, the difference between human-to-human transmission and transmission via the environment is not clear-cut. In the model, the environmental reservoir contributes a component to the force of infection which is independent of the number of infected individuals. Previous data analysis for cholera using a mechanistic model (Koelle and Pascual, 2004) was unable to include an environmental reservoir since it would have disrupted the log-linearity required by the methodology. Fig. 2.4 shows the profile likelihood of ω and resulting confidence interval, calculated using MIF. This translates to between 29 and 83 infections per million inhabitants per month from the environmental reservoir, since the model implies a mean susceptible fraction of 38%. At least in the context of this model, there is clear evidence of an environmental reservoir effect (likelihood ratio test, $p < 0.001$). Although our assumption that environmental transmission has no seasonality is less than fully reasonable, this mode of transmission is only expected to play a major role when cholera incidence is low, typically during and after the summer monsoon season (see Fig. 2.5). Human-to-human transmission, by contrast, predominates during cholera epidemics.

Links between cholera incidence and ENSO have been identified (Pascual et al., 2000; Rodó et al., 2002). Such large-scale climatic phenomena may be the best hope for forecasting disease burden (Thomson et al., 2006). We looked for a relationship between ENSO and the prediction residuals (defined in Sec. 2.7.3). Prediction residuals are robust to the exact form of the model—they depend only on the data and the predicted values, and all reasonable models should usually make similar predictions. The low-frequency component of the southern oscillation index (SOI), graphed in Fig. 2.2C, is a measure of ENSO available during the period 1891–1940 (Rodó et al., 2002); low values of SOI correspond to El Niño events. Rodó et al. (2002) showed that low SOI correlates with increased cholera cases during the period 1980–2001 but found only weak evidence of a link with cholera deaths during the 1893–1940 period. Simple correlation analysis of standardized residuals or mortality with SOI reveals no clear relationship. Breaking down by month, we find that SOI is strongly correlated with the standardized residuals for August and September (in each case, $r = -0.36$, $p = 0.005$), at which time cholera mortality historically began its seasonal increase following the monsoon (see Fig. 2.5). This suggests a narrow window of opportunity within which ENSO can act. This is consistent with the mechanism conjectured by Rodó et al. (2002) whereby the warmer surface temperatures associated with an El Niño event lead to increased human contact with the environmental reservoir and greater pathogen growth rates in the reservoir. Mortality itself did not correlate with SOI in August ($r = -0.035$, $p = 0.41$). Some weak evidence of negative correlation between SOI and mortality appeared in September ($r = -0.22$, $p = 0.063$). Earlier work (Koelle and Pascual, 2004), based on a discrete-time model and with no allowance for an environmental reservoir, failed to resolve this connection between ENSO and cholera mortality in the historical period: to find clear evidence of the external climatic forcing of the system, it is essential to use a model capable of capturing the intrinsic dynamics of disease transmission.

2.5 Discussion

Procedure 1 is dependent on the viability of solving the filtering problem, i.e., calculating $\hat{\theta}_t$ and V_t in Eq. 2.2. This is a strength of the methodology, in that the filtering problem has been extensively studied. Filtering does not require stationarity of the stochastic dynamical system, enabling covariates (such as P_t in Sec. 2.3) to be included in a mechanistically plausible way. Missing observations and data collected at irregular time intervals also pose no obstacle for filtering methods. Filtering can be challenging, particularly in nonlinear systems with a high-dimensional state space (d_x large). One example is data assimilation for atmospheric and oceanographic science, where observations (satellites, weather stations, etc.) are used to inform large spatio-temporal simulation models: approximate filtering methods developed for such situations (Houtekamer and Mitchell, 2001) could be used to apply the methods of this chapter.

The goal of maximum likelihood estimation for partially observed data is reminiscent of the Expectation-Maximization (EM) algorithm (Dempster et al., 1977), and indeed Monte Carlo EM methods have been applied to nonlinear state space models (Cappé et al., 2005). The Monte Carlo EM algorithm, and other standard Monte Carlo Markov Chain methods, cannot be used for inference on the environmental noise parameter ε for the model of Sec. 2.3, since these methods rely upon different sample paths of the unobserved process x_t having densities with respect to a common measure (Roberts and Stramer, 2001). Diffusion processes, such as the solution to the system of stochastic differential equations in Sec. 2.3, are mutually singular for different values of the infinitesimal variance. Modeling using diffusion processes (as in Sec. 2.3) is by no means necessary for the application of Procedure 1, but continuous-time models for large discrete populations are well approximated by diffusion processes, so a method that can handle diffusion processes may be expected to be more reliable for large discrete populations.

Procedure 1 is well suited for maximizing numerically estimated likelihoods for complex models largely because it requires neither analytic derivatives, which may not be available, nor numerical derivatives, which may be unstable. The iterated filtering effectively produces estimates of the derivatives smoothed at each iteration over the scale at which the likelihood is currently being investigated. Although general stochastic optimization techniques do exist for maximizing functions measured with error (Spall, 2003), these methods are inefficient in terms of the number of function evaluations required (Wu, 1985). General stochastic optimization techniques have not to our knowledge been successfully applied to examples comparable to that presented here.

Each iteration of MIF requires similar computational effort to one evaluation of the likelihood function. The results in Fig. 2.3 demonstrate the ability of Procedure 1 to optimize a function of 13 variables using 50 function evaluations, with Monte Carlo measurement error and without knowledge of derivatives. This feat is only possible because Procedure 1 takes advantage of the state-space structure of the model; however, this structure is general enough to cover relevant dynamical models across a broad range of disciplines. The EM algorithm is similarly “only” an optimization trick, but in practice it has led to the consideration of models that would be otherwise intractable. The computational efficiency of Procedure 1 is essential for the model in Sec. 2.3, where Monte Carlo function evaluations each take approximately 15 min on a desktop computer.

Implementation of Procedure 1 using particle filtering conveniently requires little more than being able to simulate paths from the unobserved dynamical system. The new methodology is therefore readily adaptable to modifications of the model, allowing relatively rapid cycles of model development, model fitting, diagnostic analysis and model improvement.

2.6 Theoretical basis for MIF

Recall the notation of Sec. 2.2, and specifically the definitions in Eqs. 2.1 & 2.2.

Theorem II.1. *Assuming conditions (R1–R3) below,*

$$\lim_{\sigma\to 0}\sum_{t=1}^{T} V_t^{-1}(\hat{\theta}_t - \hat{\theta}_{t-1}) = \nabla \log f(y_{1:T}|\theta, \sigma{=}0) \tag{2.5}$$

where ∇g is defined by $[\nabla g]_i = \partial g/\partial\theta_i$ and $\hat{\theta}_0 = \theta$. Furthermore, for a sequence $\sigma_n \to 0$, define $\hat{\theta}^{(n)}$ recursively by

$$\hat{\theta}^{(n+1)} = \hat{\theta}^{(n)} + V_{1,n}\sum_{t=1}^{T} V_{t,n}^{-1}(\hat{\theta}_t^{(n)} - \hat{\theta}_{t-1}^{(n)}) \tag{2.6}$$

where $\hat{\theta}_t^{(n)} = \hat{\theta}_t(\hat{\theta}^{(n)}, \sigma_n)$ and $V_{t,n} = V_t(\hat{\theta}^{(n)}, \sigma_n)$. If there is a $\hat{\theta}$ with $|\hat{\theta}^{(n)} - \hat{\theta}|/\sigma_n^2 \to 0$ then $\nabla \log f(y_{1:T}|\theta = \hat{\theta}, \sigma{=}0) = 0$.

Theorem II.1 asserts that (for sufficiently small σ_n) Procedure 1 iteratively updates the parameter estimate in the direction of increasing likelihood, with a fixed point at a local maximum of the likelihood. Step 2(ii) of Procedure 1 can be rewritten as $\hat{\theta}^{(n+1)} = V_{1,n}\{\sum_{t=1}^{T-1}(V_{t,n}^{-1} - V_{t+1,n}^{-1})\hat{\theta}_t^{(n)} + (V_{T,n}^{-1})\hat{\theta}_T^{(n)}\}$. This makes $\hat{\theta}^{(n+1)}$ a weighted average, in the sense that $V_1\{\sum_{t=1}^{T-1}(V_t^{-1} - V_{t+1}^{-1}) + V_T^{-1}\} = I_{d_\theta}$ where I_{d_θ} is the $d_\theta \times d_\theta$ identity matrix. The weights are necessarily positive for sufficiently small σ_n (Sec. 2.7.9).

The exponentially decaying σ_n in step 2(i) of Procedure 1 is justified by empirical demonstration, provided by the simulation study in Sec. 2.4.1. Slower decay,

$\sigma_n^2 = n^{-\beta}$ with $0 < \beta < 1$, can give sufficient conditions for a Monte Carlo implementation of Procedure 1 to converge successfully (Sec. 2.7.7). In our experience, exponential decay yields equivalent results, considerably more rapidly. Analogously, simulated annealing provides an example of a widely used stochastic search algorithm where a geometric "cooling schedule" is often more effective than slower, theoretically motivated schedules (Press et al., 2002).

In the proof of Theorem II.1, we define $f_t(\psi) = f(y_t|y_{1:t-1}, \theta_t{=}\psi)$. The dependence on σ may be made explicit by writing $f_t(\psi) = f_t(\psi, \sigma)$. We assume that $y_{1:T}$, c and Σ are fixed: for example, the constant B in (R1) may depend on $y_{1:t}$. We use the Euclidean norm for vectors and the corresponding norm for matrices, i.e., $|M| = \sup_{|u|\leq 1} |u'Mu|$, where u' denotes the transpose of u. We assume the following regularity conditions.

(R1) The Hessian matrix is bounded, i.e., there are constants B and σ_0 such that, for all $\sigma < \sigma_0$ and all $\theta_t \in \mathbb{R}^{d_\theta}$, $|\nabla^2 f_t(\theta_t, \sigma)| < B$.

(R2) $E[|\theta_t - \hat{\theta}_{t-1}|^2 \mid y_{1:t-1}] = O(\sigma^2)$.

(R3) $E[|\theta_t - \hat{\theta}_{t-1}|^3 \mid y_{1:t-1}] = o(\sigma^2)$.

(R1) is a global bound over $\theta_t \in \mathbb{R}^{d_\theta}$, comparable to global bounds used to show the consistency and asymptotic normality of the MLE (Cramér, 1946; Jensen and Petersen, 1999). It can break down, for example, when the likelihood is unbounded. This problem can be avoided by reparameterizing to keep the model away from such singularities, as is common practice in mixture modeling (McLachlan and Peel, 2000). (R2–R3) require that a new observation cannot often have a large amount of new information about θ. For example, they are satisfied if $\theta_{0:t}$, $x_{1:t}$ and $y_{1:t}$ are jointly Gaussian. We conjecture that they are satisfied whenever the state space model is smoothly parametrized and the random walk θ_t does not have long tails.

Proof of Theorem II.1. Suppose inductively that $|V_t| = O(\sigma^2)$ and $|\hat{\theta}_{t-1} - \theta| = O(\sigma^2)$. This holds for $t = 1$ by construction. Bayes' formula gives

$$\frac{f(\theta_t|y_{1:t})}{f(\theta_t|y_{1:t-1})} = \frac{f_t(\theta_t)}{\int f_t(\theta_t) f(\theta_t|y_{1:t-1})\, d\theta_t} \tag{2.7}$$

$$= \frac{f_t(\hat{\theta}_{t-1}) + (\theta_t - \hat{\theta}_{t-1})' \, \nabla f_t(\hat{\theta}_{t-1}) + R_t}{f_t(\hat{\theta}_{t-1}) + O(\sigma^2)} \tag{2.8}$$

$$\begin{aligned} &= \{1 + (\theta_t - \hat{\theta}_{t-1})' \nabla \log f_t(\hat{\theta}_{t-1}) + R_t / f_t(\hat{\theta}_{t-1})\} \\ &\qquad \times (1 + O(\sigma^2)) \end{aligned} \tag{2.9}$$

The numerator in Eq. 2.8 comes from a Taylor series expansion of $f_t(\hat{\theta}_t)$, and (R1) implies $|R_t| \leq B|\theta_t - \hat{\theta}_{t-1}|^2/2$. The denominator then follows from applying this expansion to the integral in Eq. 2.7, invoking (R2), and observing that Eq. 2.1 implies $E[\theta_t|y_{1:t-1}] = \hat{\theta}_{t-1}$. We now calculate

$$\hat{\theta}_t - \hat{\theta}_{t-1} = E[\theta_t - \hat{\theta}_{t-1}|y_{1:t}]$$

$$= \int (\theta_t - \hat{\theta}_{t-1}) f(\theta_t|y_{1:t})\, d\theta_t \tag{2.10}$$

$$= V_t \nabla \log f_t(\hat{\theta}_{t-1}) + o(\sigma^2) \tag{2.11}$$

$$= V_t \nabla \log f_t(\theta, \sigma{=}0) + o(\sigma^2). \tag{2.12}$$

Eq. 2.11 follows from Eq. 2.10 using Eq. 2.9 and (R3). Eq. 2.12 follows from Eq. 2.11 using the induction assumptions on $\hat{\theta}_{t-1}$ and V_t; Eq. 2.12 then justifies this assumption for $\hat{\theta}_t$. A similar argument gives

$$V_{t+1} = \mathrm{Var}(\theta_{t+1}|y_{1:t}) = \mathrm{Var}(\theta_t|y_{1:t}) + \sigma^2\Sigma$$

$$= E[(\theta_t - \hat{\theta}_t)(\theta_t - \hat{\theta}_t)'|y_{1:t}] + \sigma^2\Sigma$$

$$\begin{aligned} &= E[(\theta_t - \hat{\theta}_{t-1})(\theta_t - \hat{\theta}_{t-1})'|y_{1:t}] \\ &\qquad - (\hat{\theta}_t - \hat{\theta}_{t-1})(\hat{\theta}_t - \hat{\theta}_{t-1})' + \sigma^2\Sigma \end{aligned} \tag{2.13}$$

$$= V_t + \sigma^2\Sigma + o(\sigma^2), \tag{2.14}$$

where Eq. 2.14 follows from Eq. 2.13 via Eqs. 2.9 and 2.12 and the induction hypothesis on V_t. Eq. 2.14 in turn justifies this hypothesis. Summing Eq. 2.12 over t produces

$$\sum_{t=1}^{T} V_t^{-1}(\hat{\theta}_t - \hat{\theta}_{t-1}) \quad = \quad \sum_{t=1}^{T} \nabla \log f_t(\theta, \sigma{=}0) + o(1)$$

which leads to Eq. 2.5. To see the second part of the theorem, note that Eq. 2.6 and the requirement that $|\hat{\theta}^{(n)} - \hat{\theta}|/\sigma_n^2 \to 0$ imply that

$$\sum_{t=1}^{T} V_t^{-1}(\hat{\theta}^{(n)}, \sigma_n) \left(\hat{\theta}_t(\hat{\theta}^{(n)}, \sigma_n) - \hat{\theta}_{t-1}(\hat{\theta}^{(n)}, \sigma_n)\right) = o(1).$$

Continuity then gives

$$\lim_n \sum_{t=1}^{T} V_t^{-1}(\hat{\theta}, \sigma_n) \left(\hat{\theta}_t(\hat{\theta}, \sigma_n) - \hat{\theta}_{t-1}(\hat{\theta}, \sigma_n)\right) = 0,$$

which, together with Eq. 2.5, yields the required result. □

2.7 Implementing MIF

2.7.1 A basic SMC algorithm

Sequential Monte Carlo (SMC), also known as the “particle filter”, is a numerical method for filtering and prediction. SMC has aroused considerable practical and theoretical interest since its development in the 1990s (Gordon et al., 1993; Kitagawa, 1998; Doucet et al., 2001; Liu, 2001; Arulampalam et al., 2002). Here we present a basic version, which is sufficient for the purposes of this chapter. A Monte Carlo filter draws a sample from $f(x_t|y_{1:t}, \theta)$, and similarly one-step prediction involves drawing from $f(x_{t+1}|y_{1:t}, \theta)$. SMC is based on the identities

$$\begin{aligned} f(x_t|y_{1:t}, \theta) &= \frac{f(x_t|y_{1:t-1}, \theta) f(y_t|x_t, \theta)}{\int f(x_t|y_{1:t-1}, \theta) f(y_t|x_t, \theta) dx_t} \\ f(x_{t+1}|y_{1:t}, \theta) &= \int f(x_{t+1}|x_t, \theta) f(x_t|y_{1:t}, \theta) dx_t \end{aligned}$$

which give rise to the following algorithm:

1. Suppose recursively that $X^F_{t,1}, \dots, X^F_{t,J}$ have (approximately) a marginal density of $f(x_t|y_{1:t}, \theta)$.

2. Make $X^P_{t+1,j}$ a draw from $f(x_{t+1}|x_t{=}X^F_{t,j}, \theta)$. Then $X^P_{t+1,j}$ has (approximately) a marginal density of $f(x_{t+1}|y_{1:t}, \theta)$.

3. Now draw $X^F_{t+1,j}$ from $\{X^P_{t+1,k}\}$ with probabilities proportional to the resampling weights $w_k = f(y_t|x_t{=}X^P_{t,k}, \theta)$. $X^F_{t+1,j}$ has (approximately) a marginal density of $f(x_{t+1}|y_{1:t+1}, \theta)$. Independent draws can be used, but we use a more efficient systematic scheme (Algorithm 2 of Arulampalam et al., 2002).

4. The conditional log likelihood at time t, defined as $\ell_t(\theta) = \log f(y_t|y_{1:t-1}, \theta)$, is estimated by $\log\left(J^{-1}\sum_{j=1}^{J} w_j\right)$.

The log likelihood is calculated via the identity $\ell(\theta) = \log f(y_{1:T}|\theta) = \sum_{t=1}^{T} \ell_t(\theta)$.

When applying Procedure 1, the time varying parameter θ_t is included in the state space, so x_t is replaced by (x_t, θ_t). $\hat{\theta}_t$ and V_t are calculated as the sample mean over the filter particles $X^F_{t,j}$ and the sample variance over the prediction particles $X^P_{t,j}$ respectively.

We used $J = 10^4$ for MIF in Table 2.1 and $J = 3 \times 10^4$ for MIF in Fig. 2.4.

2.7.2 Numerical stability

If the number J of particles is not sufficiently large, the conditional distribution $f(x_t|y_{1:t})$ may not be well sampled by $\{X^F_{t,j}, j = 1, \dots, J\}$. Put another way, there may be few (or zero) particles $X^P_{t,j}$ consistent with the observation y_t. The few consistent particles get relatively large resampling weights and dominate the evolution of the state process — an effect known as particle depletion (Arulampalam et al., 2002). In the context of MIF, the particle filter estimates of $\hat{\theta}_t$ and V_t (say, $\hat{\theta}^e_t$ and V^e_t) then become poor. Procedure 1 is more stable if $[V_t]_{ij}$ is approximated by 0 for $i \neq j$ and by $[V^e_t]_{ii}$ for $i = j$. This forces V_t away from singularity. Supposing Σ is diagonal, Eq. 2.14 reassures us that V_t/σ^2 is asymptotically diagonal as $\sigma \to 0$, so the approximation is justified by theory for small σ and by practical stability concerns for large σ. For successful maximum likelihood estimation, depletion should become a negligible issue as θ approaches $\hat{\theta}$, and that matches what we found for the example of Sec. 2.3. When tackling problems that stretch available computational capacity, particle depletion can still be common in the early iterations of MIF, where θ may still be far from the MLE.

Even more algorithmic stability can be achieved by using the updating rule

$$\hat{\theta}^{(n+1)} = \frac{1}{T}\sum_{t=1}^{T}\hat{\theta}^{(n)}_t. \tag{2.15}$$

Although Eq. 2.15 is attractively simple and robust to particle depletion, it does not have the theoretical property of producing a sequence of estimators converging to the MLE. We found empirically that employing Eq. 2.15 on the first 5 iterations of MIF added stability without adversely affecting the final estimator.

2.7.3 Heuristics

Heuristically, α can be thought of as a “cooling” parameter, analogous to that used in simulated annealing (Spall, 2003, Chapter 8). If α is too small, the convergence will be “quenched” and fail to locate a maximum. If α is too large, the algorithm will fail to converge in a reasonable time interval. A value of $\alpha = 0.95$ was used in Sec. 2.4.

Supposing that θ_i has a plausible range $[\theta^{\text{lo}}_i, \theta^{\text{hi}}_i]$ based on prior knowledge, then each particle is capable of exploring this range in early iterations of MIF (unconditional on the data) provided $\sqrt{\Sigma_{ii}T}$ is on the same scale as $\theta_i^{\mathrm{hi}} - \theta_i^{\mathrm{lo}}$. We use $\Sigma_{ii}^{1/2} = (\theta_i^{\mathrm{hi}} - \theta_i^{\mathrm{lo}})/2\sqrt{T}$ with $\Sigma_{ij} = 0$ for $i \neq j$.

Although the asymptotic arguments do not depend on the particular value of the dimensionless constant c, looking at convergence plots led us to take $c^2 = 20$ in Sec. 2.4. Large values $c^2 \approx 40$ resulted in increased algorithmic instability, as occasional large decreases in the prediction variance V_t resulted in large weights in Procedure 1 step 2(ii). Small values $c^2 \approx 10$ were diagnosed to result in appreciably slower convergence. We found it useful, in choosing c, to check that $[V_t]_{ii}$ plotted against t was fairly stable. In principle, a different value of c could be used for each dimension of θ; for our example, a single choice of c was found to be adequate.

If the dimension of θ is even moderately large (say, $d_\theta \approx 10$) it can be challenging to investigate the likelihood surface, to check that a good local maximum has been found, and to get an idea of the standard deviations and covariance of the estimators. A useful diagnostic, the "sliced likelihood" (Fig. 2.3B), plots $\ell(\hat{\theta} + h\delta_i)$ against $\hat{\theta}_i + h$, where δ_i is a vector of zeros with a one in the i^{th} position. If $\hat{\theta}$ is located at a local maximum of each sliced likelihood then $\hat{\theta}$ is a local maximum of $\ell(\theta)$, supposing $\ell(\theta)$ is continuously differentiable. Computing sliced likelihoods requires moderate computational effort, linear in the dimension of θ. A local quadratic fit is made to the sliced log likelihood (as suggested by Ionides, 2005), because $\ell(\hat{\theta} + h\delta_i)$ is calculated with a Monte Carlo error. Calculating the sliced likelihood involves evaluating $\log f(y_t|y_{1:t-1}, \hat{\theta} + h\delta_i)$ which can then be regressed against h to estimate $(\partial/\partial\theta_i) \log f(y_t|y_{1:t-1}, \hat{\theta})$. These partial derivatives may then be used to estimate the Fisher information (Barndorff-Nielsen and Cox, 1994, and Sec. 2.7.8) and corresponding standard errors (SEs). Profile likelihoods (Barndorff-Nielsen and Cox, 1994) can be calculated using MIF, but at considerably more computational expense than sliced likelihoods. SEs and profile likelihood confidence intervals, based on asymptotic properties of MLEs, are particularly useful when alternate ways to find standard errors, such as bootstrap simulation from the fitted model, are prohibitively expensive to compute. Our experience, consistent with previous advice (McCullagh and Nelder, 1989), is that SEs based on estimating Fisher information provide a computationally frugal method to get a reasonable idea of the scale of uncertainty, but profile likelihoods and associated likelihood based confidence intervals are more appropriate for drawing careful inferences.

2.7.4 Diagnostics

Our main MIF diagnostic is to plot parameter estimates as a function of MIF iteration; we call this a convergence plot. Convergence is indicated when the estimates reach a single stable limit from various starting points. Convergence plots were also used for simulations with a known true parameter, to validate the methodology. The investigation of quantitative convergence measures might lead to more refined implementations of Procedure 1.

As in regression, residual analysis is a key diagnostic tool for state space models. The standardized prediction residuals are $\{u_t(\hat{\theta})\}$ where $\hat{\theta}$ is the MLE and $u_t(\theta) = [\mathrm{Var}(y_t|y_{1:t-1},\theta)]^{-1/2}(y_t - E[y_t|y_{1:t-1},\theta])$. Other residuals may be defined for state space models (Durbin and Koopman, 2001), such as $E[\int_{t-1}^{t} dW_s|y_{1:T},\hat{\theta}]$ for the model of Sec. 2.3. Prediction residuals have the property that, if the model is correctly specified with true parameter vector θ^*, $\{u_t(\theta^*)\}$ is an uncorrelated sequence. This has two useful consequences: it gives a direct diagnostic check of the model, i.e., $\{u_t(\hat{\theta})\}$ should be approximately uncorrelated; it means that prediction residuals are an (approximately) pre-whitened version of the observation process, which makes them particularly suitable for using correlation techniques to look for relationships with other variables (Shumway and Stoffer, 2000), as demonstrated in Sec. 2.4.2.

The prediction residuals, $u_t(\hat{\theta}) = [\mathrm{Var}(y_t|y_{1:t-1},\hat{\theta})]^{-1/2}(y_t - E[y_t|y_{1:t-1},\hat{\theta}])$, can be calculated via

$$\begin{aligned} E[y_t|y_{1:t-1}] &\approx \frac{1}{J}\sum_{j=1}^{J} E[y_t|x_t = X^P_{t,j}] \\ \mathrm{Var}(y_t|y_{1:t-1}) &= E[\mathrm{Var}(y_t|x_t)|y_{1:t-1}] + \mathrm{Var}(E[y_t|x_t] \mid y_{1:t-1}) \\ &\approx \frac{1}{J}\sum_{j=1}^{J} \mathrm{Var}[y_t|x_t = X^P_{t,j}] + \frac{1}{J-1}\sum_{j=1}^{J} (\hat{y}_{t,j} - \hat{y}_{t,\bullet})(\hat{y}_{t,j} - \hat{y}_{t,\bullet})' \end{aligned}$$

where $\hat{y}_{t,j} = E[y_t|x_t = X^P_{t,j}]$ and $\hat{y}_{t,\bullet} = (1/J)\sum_{j=1}^{J} \hat{y}_{t,j}$.

2.7.5 Initial values

The property that Procedure 1 updates as a weighted average of local parameter estimates is less appropriate when the information about a parameter is not spread out across time. A good example of such a parameter is an initial value parameter (IVP). Other situations where information about a parameter is concentrated in time, such as modeling a structural break, can be treated in a similar way. We describe θ as an IVP if $f(x_0) = f(x_0|\theta)$, but $f(x_t|x_{t-1})$ and $f(y_t|x_t)$ do not depend on θ for $t > 0$. As a particular case, if x_0 is supposed to be fixed and unknown then one can take $\theta = x_0$. There may not be any IVP in a model; for example, if x_0 is drawn from the stationary distribution of a time homogeneous Markov transition density $f(x_t|x_{t-1},\theta)$.

For IVPs, we develop Procedure 2 based on Lemma II.2. To maximize the likelihood, we introduce a prior distribution $f(\theta)$ with prior variance $\mathrm{Var}(\theta) = \sigma^2\Sigma$.

Lemma II.2. *Let $\hat{\theta}_0$ be the prior mode, i.e., $\hat{\theta}_0 = \mathrm{argmax} f(\theta)$. Let $\hat{\theta}_T$ be the posterior mode, i.e., $\hat{\theta}_T = \mathrm{argmax} f(\theta|y_{1:T})$. Then*

$$f(y_{1:T}|\hat{\theta}_T) \geq f(y_{1:T}|\hat{\theta}_0).$$

Proof.

$$\frac{f(y_{1:T}|\theta{=}\hat{\theta}_T)}{f(y_{1:T}|\theta{=}\hat{\theta}_0)} = \frac{f(\theta{=}\hat{\theta}_T|y_{1:T})}{f(\theta{=}\hat{\theta}_0|y_{1:T})} \times \frac{f(\theta{=}\hat{\theta}_0)}{f(\theta{=}\hat{\theta}_T)} \geq 1$$

The inequality holds by the definition of $\hat{\theta}_0$ and $\hat{\theta}_T$, since both terms in the product are at least one. □

Procedure 2. ***(MIF for initial values)***

1. *Select starting values $\hat{\theta}^{(1)}$ and σ_1, a discount factor $0 < \alpha < 1$, a fixed lag T_0 and the number of iterations N.*

2. *For n in $1, \ldots, N$*

   (i) Evaluate $\hat{\theta}^{(n)}_{T_0}$ using $\hat{\theta}_0 = \hat{\theta}^{(n)}$ and $\sigma = \sigma_1 \alpha^{n-1}$.

   (ii) Set $\hat{\theta}^{(n+1)} = \hat{\theta}^{(n)}_{T_0}$.

3. *Take $\hat{\theta}^{(N+1)}$ to be an estimate of θ.*

Approximating $f(\theta|y_{1:T})$ by $f(\theta|y_{1:T_0})$ in step 2(i) of Procedure 2 is a standard method to facilitate nonlinear filtering, termed fixed lag smoothing (Anderson and Moore, 1979). It is certainly necessary for a particle filter implementation. The fixed lag smoothing approximation to $f(\theta|y_{1:T})$ is only reliable when the information in the data about θ is concentrated at small t values. Applying Procedure 2 to non-IVP parameters with $T_0 = T$ is a direct way to attempt inference for time-constant parameters. The difficulty of doing this in practice was exactly the motivation for developing Procedure 1. Procedure 2 is essentially an exhaustive search over a sequence of increasingly refined IVP values. An advantage of this procedure is that it fits in computationally with Procedure 1, allowing IVPs to be estimated simultaneously with other parameters.

2.7.6 Some recommendations for stochastic likelihood maximization

This section describes our approach to carrying out inference based on Procedure 1. When investigating a likelihood surface, there is a trade-off between effort spent on global searching and local searching. An effective way to investigate large-scale properties of the likelihood, and simultaneously to check that the maximization procedure is successful, is to initialize the maximization at a range of parameter values. This approach is formalized in Procedure 3, below:

Procedure 3. ***(Investigating the likelihood surface)***

1. *Pick K starting values (for example, by sampling each component of θ uniformly within an assigned plausible range) and apply Procedure 1 to get K pairs $\{(\hat{\theta}_k, \ell_k)\}$ of estimates and associated log likelihoods.*

2. *If there is a clear global maximum – i.e., there are many pairs (θ_k, ℓ_k) with $(\max_j \ell_j - \ell_k)$ small and $|\hat{\theta}_{\mathrm{argmax}_j \ell_j} - \hat{\theta}_k|$ small – then take the MLE to be the average of these global maximum estimates.*

3. *If there is not a clear global maximum – many pairs (θ_k, ℓ_k) have $(\max_j \ell_j - \ell_k)$ small but $|\hat{\theta}_{\mathrm{argmax}_j \ell_j} - \hat{\theta}_k|$ not small – then some combination of the parameters is poorly identifiable. Investigate this by plotting the components of $\{\hat{\theta}_k\}$ and calculating correlations. Perhaps make extra assumptions to improve identifiability and return to step 1.*

Procedure 3 requires manual oversight. This is appropriate for diagnostic checking of the maximization procedure and investigation of the global structure of the likelihood. Manual intervention is not necessary for each maximization of a profile likelihood or parametric bootstrap computation, since these require only local optimization in the neighborhood of the MLE (which is also the true parameter vector for bootstrap simulations). The only situation where local searches would be inappropriate for profile likelihood or bootstrap computations arise when the global likelihood has two (or more) separated modes of almost equal likelihood. These modes should be identified by Procedure 3 and require local maximization about each mode. Procedure 1 can be adapted for local maximization by decreasing α, c and Σ. This also demands a smaller value of N, the number of iterations, which is helpful for implementing these computationally intensive finite sample procedures.

One subtlety in Procedure 3 is the use of the average in step 2. In our applications, the Monte Carlo error in evaluating the likelihood is typically large compared to the actual difference in the likelihood between MIF estimates that have converged to the same mode. This occurs because MIF seeks the maximum by averaging Monte Carlo error over many iterations. Thus, we chose to average MIF estimates rather than to take the one with the highest evaluated likelihood.

To implement step 2 of Procedure 3 one must determine what is meant by “small”. As this procedure is intended to be used on a broad variety of models, we think automation is premature. A general observation is that “small” differences in the likelihood are of the order of one unit of log likelihood.

Some simple methods are available to check that the likelihood is being maximized effectively on simulated data, with a known parameter vector θ^*. Setting $\hat{\theta} = \arg\max \ell(\theta)$, an asymptotic result for regular parametric models is that $2(\ell(\hat{\theta}) - \ell(\theta^*))$ has approximately the distribution of $\chi^2(d_\theta)$, a chi-squared random variable on d_θ degrees of freedom (Barndorff-Nielsen and Cox, 1994). Thus, beyond the basic property that $\ell(\hat{\theta}) \geq \ell(\theta^*)$, one can expect $\ell(\hat{\theta}) - \ell(\theta^*) \approx d_\theta/2$. If estimates of the maximized log likelihood compared with the likelihood at θ^* are not unusual for $(1/2)\chi^2(d_\theta)$, we view this as some evidence for successful maximization. The sliced likelihood plots described in Sec. 2.7.3 give the formal demonstration of successful maximization, but require extra computation.

2.7.7 Sufficient conditions for convergence of iterated filtering

Theorem II.3 provides a complementary result to Theorem II.1, giving sufficient conditions on the sequence $\sigma_n \to 0$ for Procedure 1 to convergence successfully. Although stated as a global result, Theorem II.3 implies corresponding local behavior that is more relevant in practice.

Theorem II.3. *Suppose that $\ell(\theta)$ is twice continuously differentiable, with a uniform convexity property that there exist $0 > a > b$ such that*

$$a > u'\nabla^2\ell(\theta)u > b \quad \textit{for all } \theta \textit{ and all unit vectors, } |u| = 1. \tag{2.16}$$

Define the sequence $\{\hat{\theta}^{(n)}\}$ by a stochastic difference equation,

$$\hat{\theta}^{(n+1)} = \hat{\theta}^{(n)} + \sigma_n^2 M(\nabla\ell(\hat{\theta}^{(n)}) + \eta_n). \tag{2.17}$$

Take $M = (c^2+1)\Sigma$, so that M is a positive definite symmetric matrix and $\sigma_n^2 M = V_{1,n}$ in the notation of Theorem II.1. Suppose that $\lim_n \sigma_n^2 n^{1-\beta} > 0$ for some $\beta \in (0,1)$. Suppose also that the sequence $\{\eta_n\}$ has $E[\eta_n] = o(1)$, $\mathrm{Var}(\sigma_n^2\eta_n) = o(1)$ and $\mathrm{Cov}(\eta_m, \eta_n) = 0$ for $m \neq n$. If there is a $\hat{\theta}$ with $\nabla\ell(\hat{\theta}) = 0$ then $\hat{\theta}^{(n)}$ converges in probability to $\hat{\theta}$.

To see how Theorem II.3 applies to MIF, implemented using a Monte Carlo filter, we need some assumptions. We suppose that the Monte Carlo filter is unbiased: this is not quite true for sequential Monte Carlo with a finite sample size, but it become exactly true if we accept the goal of maximizing the expected Monte Carlo log likelihood rather than the true log likelihood. Theorem II.1 then gives $E[\eta_n] = o(1)$ as long as $\sigma_n \to 0$; we have to assume that this convergence is uniform over θ. A reasonable model for the variance of a derivative based on Monte Carlo likelihood evaluations in a neighborhood of size σ_n is $\mathrm{Var}(\eta_n) = O(\sigma_n^{-2})$, implying the condition $\mathrm{Var}(\sigma_n^2\eta_n) = o(1)$. Formally, to apply Theorem II.3, one must assume that this rate is also uniform over θ. If the Monte Carlo filter uses independent sequences of random numbers for each iteration, $\mathrm{Cov}(\eta_m, \eta_n) = 0$ for $m \neq n$.

Proof of Theorem II.3. The fundamental theorem of calculus gives

$$\nabla\ell(\theta) = \int_0^1 \nabla^2\ell(s\theta + (1-s)\hat{\theta})(\theta - \hat{\theta})\, ds.$$

This can also be written as $\nabla\ell(\theta) = H(\theta)(\theta - \hat{\theta})$ where $H(\theta) = \int_0^1 \nabla^2\ell(s\theta + (1-s)\hat{\theta})\, ds$. We re-write Eq. 2.17 as

$$\hat{\theta}^{(n+1)} = \hat{\theta}^{(n)} + \sigma_n^2 M(H_n(\hat{\theta}^{(n)} - \hat{\theta}) + \eta_n) \tag{2.18}$$

where $H_n = H(\hat{\theta}^{(n)})$. Eq. 2.18 can be written as

$$\begin{aligned}\hat{\theta}^{(n+1)} - \hat{\theta} &= \prod_{k=1}^{n}(I + \sigma_k^2 M H_k)(\hat{\theta}^{(1)} - \hat{\theta}) \\ &\quad + \sum_{m=1}^{n-1}\Big\{\prod_{k=m+1}^{n}(I + \sigma_k^2 M H_k)\Big\}\sigma_m^2 M\eta_m + \sigma_n^2 M\eta_n.\end{aligned} \tag{2.19}$$

$H(\theta)$ satisfies the same inequality in Eq. 2.16 as $\nabla^2\ell(\theta)$, which guarantees a uniform bound on the eigenvalues of $\sigma_k^2 M H_k n^{1-\beta}$. Lemma II.4, with A taken to be $\sigma_k^2 M H_k$, then secures the existence of a constant $c > 0$ such that, for sufficiently large k,

$$\log|I + \sigma_k^2 M H_k| < -ck^{\beta-1}.$$

A comparison of $\sum_{k=m}^{n} k^{\beta-1}$ with $\int_m^n x^{\beta-1}dx$ then gives

$$\log\prod_{k=m}^{n}|I + \sigma_k^2 M H_k| < c\beta^{-1}(m^\beta - n^\beta). \tag{2.20}$$

Lemma II.5 can be applied to Eq. 2.20 to demonstrate that

$$\sum_{m=1}^{n-1}|\sigma_m^2|\prod_{k=m+1}^{n}|I + \sigma_k^2 M H_k| = O(1).$$

Lemma II.6 can then be applied, with $w_{m,n-1} = |\sigma_m^2| \prod_{k=m+1}^n |I + \sigma_k^2 M H_k|$ and $b_n = E[\eta_n]$. This gives

$$E\Big[\sum_{m=1}^{n-1} \Big\{ \prod_{k=m+1}^{n} (I + \sigma_k^2 M H_k) \Big\} \sigma_m^2 \eta_m \Big] \to 0. \tag{2.21}$$

A very similar argument, replacing $w_{m,n-1}$ by $|\sigma_m^2| \prod_{k=m+1}^n |I + \sigma_k^2 M H_k|^2$ and b_n by $\mathrm{Var}(\sigma_n^2 \eta_n)$, allows the use of Lemma II.6 to give

$$\mathrm{Var}\Big[\sum_{m=1}^{n-1} \Big\{ \prod_{k=m+1}^{n} (I + \sigma_k^2 M H_k) \Big\} \sigma_m^2 \eta_m \Big] \to 0. \tag{2.22}$$

In addition, Eq. 2.20 implies that

$$\prod_{k=1}^{n} (I + \sigma_k^2 M H_k)(\hat{\theta}^{(1)} - \hat{\theta}) \to 0. \tag{2.23}$$

Eq. 2.21, Eq. 2.22 and Eq. 2.23 imply convergence in probability for Eq. 2.19, which completes the proof. □

Lemma II.4. *If A is a negative definite matrix with $|A| < 1$ and with largest eigenvalue $\pi < 0$ then* $\log |I + A| < \pi$.

Proof. Let u be an arbitrary vector with $|u| = 1$.

$$\begin{aligned} \log |I + A| &= \log(\sup_u |u'(I + A)u|) \\ &= \log(\sup_u |1 + u'Au|) \end{aligned}$$

By hypothesis $u'Au > -1$, and so $\sup_u |1 + u'Au| = 1 + \sup_u u'Au$. Therefore,

$$\log |I + A| = \log(1 + \sup_u u'Au) = \log(1 + \pi) < \pi,$$

where we use the inequality $\log(1 + \pi) < \pi$ for $\pi \in (-1, 0)$. □

Lemma II.5. *If $c > 0$ and $0 < \beta < 1$ then*

$$\sum_{m=1}^{n} \exp\{c(m^\beta - n^\beta)\} m^{\beta - 1} = O(1). \tag{2.24}$$

Proof. We write the sum in Eq. 2.24 as

$$n^{\beta}\frac{1}{n}\sum_{m=1}^{n}\exp\{-c(1-(m/n)^{\beta})n^{\beta}\}\times\left(\frac{m}{n}\right)^{\beta-1}. \tag{2.25}$$

As $n \to \infty$, Eq. 2.25 can be compared to the integral

$$n^{\beta}\int_0^1 \exp\{-c(1-x^{\beta})n^{\beta}\}x^{\beta-1}\,dx.$$

This can be analyzed in two parts. Firstly,

$$\begin{aligned} n^{\beta}\int_0^{1/2}\exp\{-c(1-x^{\beta})n^{\beta}\}x^{\beta-1}\,dx &< n^{\beta}\int_0^{1/2}\exp\{-(1-(1/2)^{\beta})cn^{\beta}\}x^{\beta-1}\,dx \\ &= n^{\beta}\exp\{-(1-(1/2)^{\beta})cn^{\beta}\}(1/2)^{\beta}/\beta \to 0. \end{aligned} \tag{2.26}$$

For the second part, change variable to $y=(1-x^{\beta})$ and proceed as follows:

$$\begin{aligned} n^{\beta}\int_{1/2}^{1}\exp\{-c(1-x^{\beta})n^{\beta}\}x^{\beta-1}\,dx &= n^{\beta}\int_0^{1-(1/2)^{\beta}}\exp\{-cyn^{\beta}\}\beta x^{2(\beta-1)}\,dy \\ &< n^{\beta}2^{2(1-\beta)}\int_0^{\infty}\exp\{-cyn^{\beta}\}\,dy = 2^{2(1-\beta)}/c. \end{aligned} \tag{2.27}$$

Eq. 2.26 and Eq. 2.27 together yield the required result. □

Lemma II.6. *Suppose* $b_n \to 0$ *and* $\sum_{m=1}^{n}|w_{m,n}| < C$ *with* $w_{m,n} \to 0$ *as* $n \to \infty$ *for each* m*. Then* $\sum_{m=1}^{n} b_n w_{m,n} \to 0$.

Proof. b_n is bounded, say $|b_n| < K$. For $\epsilon > 0$, $\exists n_0 : |b_n| < \epsilon \ \forall n > n_0$. Also, $\exists n_1 : |w_{m,n}| < \epsilon/n_0$ whenever $m \le n_0$ and $n > n_1$. Then, for $n > \max(n_0, n_1)$, $|\sum_{m=1}^{n_0} b_n w_{m,n}| < K\epsilon$ and $|\sum_{m=n_0+1}^{n} b_n w_{m,n}| < C\epsilon$. Thus, $|\sum_{m=1}^{n} b_n w_{m,n}| < (K + C)\epsilon$. □

2.7.8 Standard errors and confidence intervals

The Fisher information can be estimated by

$$\hat{\mathcal{I}}_{ij} = \sum_{t=1}^{T}\frac{\partial}{\partial\theta_i}\log f(y_t|y_{1:t-1},\hat{\theta})\frac{\partial}{\partial\theta_j}\log f(y_t|y_{1:t-1},\hat{\theta}) \tag{2.28}$$

leading to corresponding standard errors $\mathrm{SE}(\hat{\theta}_i) = [\hat{\mathcal{I}}^{-1/2}]_{ii}$. Procedure 4 details how this was implemented in this chapter.

Procedure 4. ***(Standard errors)***

1. *Let $\hat{\theta}$ be the sample mean of the (vector) estimates $\{\hat{\theta}_k, k = 1, \ldots, K\}$ from K replications of Procedure 1. Calculate $\ell_{t,ij} = \log f(y_t|y_{1:t-1}, \hat{\theta} + h_{ij}\delta_i)$ for $1 \leq i \leq m$ and $1 \leq j \leq q$, where δ_i is a vector of zeros with a one in the i^{th} position. $\{h_{ij}\}$ can be the offsets used for a sliced likelihood diagnostic plot. Alternatively, one can use $q = 2$ with $h_{i1} = 0$ and $h_{i2} = h\sqrt{\Phi_{ii}}$, where Φ is the sample covariance matrix of $\{\hat{\theta}_k\}$. The constant h is chosen by trial and error, and Φ gives the relative scale of the uncertainty in the components of θ.*

2. *Regress $\ell_{t,ij}$ on h_{ij} for each i, giving rise to regression coefficients $\dot{\ell}_{t,i}$ with variance estimates $\hat{\mathrm{Var}}(\dot{\ell}_{t,i})$.*

3. *Estimate the Fisher information by $\hat{I}_{ij} = \sum_t \dot{\ell}_{t,i}\dot{\ell}_{t,j}$ and estimate the derivative of the log likelihood at $\hat{\theta}$ by $\dot{\ell}_i = \sum_{t=1}^{T} \dot{\ell}_{t,i}$.*

Procedure 4 step 2 calculates numerical derivatives, averaging over a neighborhood given by $\{h_{ij}\}$. If $\{h_{ij}\}$ are too small, the Monte Carlo error in the likelihood evaluation will dominate the numerical derivative. Since $E[\dot{\ell}_{t,i}] \approx \partial\ell/\partial\theta_i$, $\sum_{t=1}^{T} E[\dot{\ell}_{t,i}^2] \approx \sum_{t=1}^{T} \left\{(\partial\ell/\partial\theta_i)^2 + \mathrm{Var}(\dot{\ell}_{t,i})\right\}$. Thus the bias of $\hat{I}_{ii}$ as an estimator of I_{ii} is approximately $\sum_{t=1}^{T} \hat{\mathrm{Var}}(\dot{\ell}_{t,i})$. We monitor this quantity and trust the estimate $\hat{I}_{ii}$ only if $\hat{I}_{ii} \gg \sum_{t=1}^{T} \hat{\mathrm{Var}}(\dot{\ell}_{t,i})$. Otherwise, either the neighborhood used to calculate the numerical derivative or the Monte Carlo sample size must be increased. There could be some advantage in calculating the numerical derivatives in the directions of the eigenvectors of Φ, with the eigenvalues giving the appropriate scaling in each direction. We prefer not to do this, since K is not necessarily large compared to m.

In particular, if $K \leq m$ then Φ is singular.

Note that one can use $\tilde{\theta} = \hat{\theta} + \hat{I}^{-1}\dot{\ell}$ as a possibly improved parameter estimate, based on a quadratic approximation to the local likelihood surface (Ionides, 2005). However, $\tilde{\theta}$ involves the potentially inaccurate Monte Carlo derivative estimates that MIF carefully avoids, and in our experience $\hat{\theta}$ is more reliable for the situation arising in this chapter.

Standard errors are usually interpreted in the context of a normal approximation for the MLE: one is invited to think of $\hat{\theta}_i \pm 2\,\mathrm{SE}(\hat{\theta}_i)$ as an approximate 95% confidence interval. A more accurate confidence interval comes from the profile log likelihood (Barndorff-Nielsen and Cox, 1994). Profile likelihoods can be calculated using MIF, but at considerably more computational expense than the SEs from Procedure 4. If θ is partitioned into two components ζ and η, of dimensions d_ζ and d_η respectively, then the profile log likelihood of η is defined by $\ell_{(p)}(\eta) = \sup_\zeta \ell(\zeta, \eta)$. An approximate 95% confidence interval for η is given by $\{\eta : 2[\ell_{(p)}(\hat{\eta}) - \ell_{(p)}(\eta)] < \chi^2_{0.95}(d_\eta)\}$ where $\chi^2_{0.95}(d_\eta)$ is the 0.95 quantile of a χ^2 random variable on d_η degrees of freedom, and $\hat{\eta} = \operatorname{argmax} \ell_{(p)}(\eta)$.

2.7.9 Comments on Procedure 1

Remark II.7. For a stationary time series, if $\sigma > 0$ is fixed and T grows, one expects (under suitable mixing conditions such as those of Jensen and Petersen, 1999) that $V_t(\sigma) \to V_\infty(\sigma)$. If $V_t \approx V_\infty$ for $t = 1, 2, \ldots$ then Procedure 1 gives $\hat{\theta}^{(n)} \approx \hat{\theta}^{(n-1)}_T$. On the other hand, fixing T, letting $\sigma \to 0$ and using Eq. 2.14, gives a rather different result of $V_t = (c^2 + t)\sigma^2\Sigma + o(\sigma^2)$. In this case,

$$\hat{\theta}^{(n)} \approx \sum_{t=1}^{T-1} \hat{\theta}^{(n-1)}_t \frac{c^2+1}{(c^2+t)(c^2+t+1)} + \hat{\theta}^{(n-1)}_T \frac{c^2+1}{c^2+T}. \tag{2.29}$$

A consequence of Eq. 2.29 is that, for sufficiently small σ, all the weights in the weighted average representation of Procedure 1 are positive. Eq. 2.29 also helps to explain why small values of c may lead to slow convergence, since small values of c lead to low weights for large t.

Remark II.8. If the assumption in Eq. 2.1 is relaxed to

$$E[\theta_t|\theta_{t-1}] = \theta_{t-1} + O(\sigma^2) \tag{2.30}$$

then Theorem 1 holds with $\hat{\theta}_{t-1}$ in Eq. 2.5 replaced by $E[\theta_t|y_{1:t-1}]$. The weaker assumption in Eq. 2.30 may be appropriate if θ lies in a bounded set, and θ_t is constrained to stay in this set. In this case, the weighted average interpretation of Procedure 1 is lost. Our solution to boundary issues for θ_t is to reparameterize to remove the difficulty, or just to ignore the difficulty if it disappears by itself for sufficiently small σ.

CHAPTER III

Time Series Analysis Via Mechanistic Models

3.1 Introduction

The purpose of time series analysis via mechanistic models is to reconcile the known or hypothesized structure of a dynamical system with observations collected over time. Motivated by examples in population biology, we develop in this chapter a framework for constructing models and carrying out inference. We build on recent advances in inference methodology for partially observed Markov models. As a case-study, we present a mechanistic analysis of cholera incidence data, involving interaction between two competing strains of the pathogen *Vibrio cholerae.* This leads us to develop inference for a new class of Markov chain models with stochastic transition rates.

A dynamical system is a process whose state varies with time. A mechanistic approach to understanding such a system is to write down equations, based on scientific understanding of the system, which describe how it evolves with time. Further equations describe the relationship of the state of the system to available observations on the system. Mechanistic time series analysis concerns drawing inferences from the available data about the hypothesized equations. Questions of general interest include the following. Are the data consistent with a particular model? If so, for what range of values of model parameters? Does one mechanistic model describe the data better than another?

The defining principle of mechanistic modeling is that the model structure should be chosen based on scientific considerations, rather than statistical convenience. Although linear Gaussian models (Durbin and Koopman, 2001) give an adequate representation of some processes, nonlinear behavior is an essential property of many systems. This leads to a need for statistical modeling and inference techniques applicable to rather general classes of processes. In the absence of alternative statistical methodology, a common approach to mechanistic investigations is to compare data, qualitatively or via some ad-hoc metric, with simulations from the model. A goal of this chapter is to increase the range of time series models for which formal statistical inferences, making efficient use of the data, can be made. Simulation of sample paths is still proposed as a basic tool for statistical analysis, but this does not preclude employing the framework of likelihood based inference. Inferential techniques that require only simulation from the model (i.e. for which the model could be replaced by a black box which inputs parameters and outputs sample paths) have been called "equation free" (Kevrekidis et al., 2004; Xiu et al., 2005). We will use the expression "plug and play," which we feel is more descriptive.

Here, we concern ourselves with partially observed, continuous-time, nonlinear, Markovian stochastic dynamical systems. The particular combination of properties listed above is chosen because it arises naturally when constructing a mechanistic model. Although observations will typically be at discrete times, mechanistic equations describing underlying continuous time systems are naturally described in continuous time. If all quantities important for the evolution of the system are explicitly modeled, then the future evolution of the system depends on the past only through the current state, i.e., the system is Markovian. A stochastic model is prerequisite for mechanistic time series analysis, since chance variability is required to explain the difference between the data and the solution to noise-free deterministic equations. Statistical analysis is simpler if stochasticity can be confined to the observation process (the statistical problem becomes nonlinear regression) or if the stochastic dynamical system is perfectly observed (Basawa and Prakasa Rao, 1980). Here we address the general case with both forms of stochasticity. Despite considerable work on such models (Anderson and Moore, 1979; Liu, 2001; Doucet et al., 2001), statistical methodology which is readily applicable for a wide range of models has remained elusive.

Several inference techniques have previously been proposed which are compatible with plug-and-play inference from partially observed Markov processes. Nonlinear forecasting (e.g., Kendall et al., 1999) is a method of simulated moments which approximates the likelihood. Iterated filtering (Ionides et al., 2006) provides a way to calculate a maximum likelihood estimate via Sequential Monte Carlo, a plug-and-play filtering technique. An approximate Bayesian Sequential Monte Carlo method (Liu and West, 2001) has also been proposed. This chapter develops inference methodology based on the iterated filtering technique, together with describing rather general classes of models for which the methodology is applicable.

In Section 3.2, we discuss a conceptual and notational framework for mechanistic modeling. Section 3.3 is concerned with inference methodology. Section 3.4 develops a concrete example. Section 3.5 discusses various extensions and alternatives to the statistical analyses developed in this chapter. The motivating example in this chapter has led to an emphasis on modeling infectious diseases. The issue of mechanistic modeling of time series data is too widespread to give a comprehensive review.

We instead list some examples: molecular biochemistry (Kou et al., 2005); wildlife ecology (Newman and Lindley, 2006); cell biology (Ionides et al., 2004); economics (Fernandez-Villaverde and Rubio-Ramirez, 2005); engineering (Arulampalam et al., 2002); data assimilation for numerical models (Houtekamer and Mitchell, 2001). The study of infectious disease, however, has a long history of motivating new modeling and data analysis methodology (Kermack and McKendrick, 1927; Bartlett, 1960; Anderson and May, 1991; Finkenstädt and Grenfell, 2000; Ionides et al., 2006). The freedom to carry out formal statistical analysis based on mechanistically motivated, non-linear, non-stationary, continuous time stochastic models is a new development which promises to be a useful tool for a variety of applications.

3.2 A class of implicitly defined models

We introduce a class of mechanistic models which is described implicitly, meaning that the model is written in such a way as to facilitate numerical solution without giving an explicit closed-form expression for transition probabilities or sample paths. The ability to analyze such models is a powerful property of plug-and-play methodology: one can carry out statistical inference on algorithms which compute sample paths, reducing the separation between algorithmic methods and model-based analyses (Breiman, 2001). The models introduced here are developed with the epidemiological application of Section 3.4 in mind, however the framework has broader relevance.

Many mechanistic models can be viewed in terms of flows between compartments (Jacquez, 1996; Matis and Kiffe, 2000). A general compartment model is a vector valued process $X(t) = (X_1(t), \dots, X_c(t))$ denoting the (integer or real-valued) counts in each of c compartments. The basic characteristic of a compartment model is that

$X(t)$ can be written in terms of the flows $N_{ij}(t)$ from i to j, via a "conservation of individuals" identity:

$$X_i(t) = X_i(0) + \sum_{j \neq i} N_{ji}(t) - \sum_{j \neq i} N_{ij}(t). \tag{3.1}$$

Each *flow* N_{ij} is associated with a *rate* function $\mu_{ij} = \mu_{ij}(t, X(t))$. There are many ways to develop concrete interpretations of such a compartment model. Here, we give a specification sufficient to cover the example of Section 3.4, while discussing alternatives and generalizations in Section 3.5. For the time being, we take $X_i(t)$ to be non-negative integer valued, so $X(t)$ models a population divided into c disjoint categories and μ_{ij} is the rate at which each individual in compartment i moves to j. The conservation equation (3.1) makes the compartment model closed in the sense that individuals cannot enter or leave the population. However, processes such as immigration, birth or death can be modeled via the introduction of additional source and sink compartments.

We wish to introduce white noise to model stochastic variation in the rates (discussion of this decision is postponed to Section 3.5). We refer to *white noise* as the derivative of an *integrated noise* process with stationary independent increments (Karlin and Taylor, 1981, Chapter 15). The integral of a white noise process over an interval is thus well defined, even when the sample paths of the integrated noise process are not formally differentiable. Specifically, we introduce a collection of gamma processes $\{\Gamma_{ij}(t), 1 \leq i \leq c, 1 \leq j \leq c\}$ (Sato, 1999). The collection of increments $\{\Gamma_{ij}(t_2) - \Gamma_{ij}(t_1), 1 \leq i \leq c, 1 \leq j \leq c\}$ is presumed to be independent of $\{\Gamma_{ij}(t_4) - \Gamma_{ij}(t_3), 1 \leq i \leq c, 1 \leq j \leq c\}$ for all $t_1 < t_2 < t_3 < t_4$. We have not assumed that different noise processes Γ_{ij} and Γ_{kl} are independent; their increments could be correlated, or even equal. Marginally, we suppose that $\Gamma_{ij}(t+\delta) - \Gamma_{ij}(t) \sim \mathit{Gamma}(\delta/\sigma_{ij}^2, \sigma_{ij}^2)$, the gamma distribution whose shape parameter is δ/σ_{ij}^2 and scale parameter is σ_{ij}^2, with corresponding mean δ and variance $\delta\sigma_{ij}^2$. We call σ_{ij}^2 an *infinitesimal variance* parameter (Karlin and Taylor, 1981). These gamma processes define a collection of gamma noise processes given by $\xi_{ij}(t) = \frac{d}{dt}\Gamma_{ij}(t)$. Since $\Gamma_{ij}(t)$ is increasing, $\xi_{ij}(t)$ is non-negative and $\mu_{ij}\xi_{ij}(t)$ can be interpreted as a rate with multiplicative white noise. The choice of gamma noise is made for convenience and to give a concrete example. A wide range of Lévy processes (Sato, 1999) could be equivalently employed.

We proceed to interpret a compartment model as a continuous time Markov chain via the limit of coupled multinomial processes with random rates. Let $\Delta N_{ij} = N_{ij}(t+\delta) - N_{ij}(t)$ and $\Delta\Gamma_{ij} = \Gamma_{ij}(t+\delta) - \Gamma_{ij}(t)$. We suppose that

$$P[\Delta N_{ij} = n_{ij}, \text{ for all } 1 \le i \le c, 1 \le j \le c, i \ne j \mid X(t) = (x_1, \ldots, x_c)]$$

$$= E\left[\prod_{i=1}^{c}\left\{\binom{x_i}{n_{i1} \;\ldots\; n_{ii-1}\; n_{ii+1} \;\ldots\; n_{ic}\;\; r_i}(1 - \textstyle\sum_{k\ne i} p_{ik})^{r_i}\displaystyle\prod_{j\ne i} p_{ij}^{n_{ij}}\right\}\right] + o(\delta) \tag{3.2}$$

where $r_i = x_i - \sum_{k\ne i} n_{ik}$, $\binom{n}{n_1 \ldots n_c}$ is a multinomial coefficient and

(3.3)

$$p_{ij} = p_{ij}(\{\mu_{ij}(t, X(t))\}, \{\Delta\Gamma_{ij}(t)\}) = (1 - \exp\{-\sum_k \mu_{ik}\Delta\Gamma_{ik}\})\mu_{ij}\Delta\Gamma_{ij}/\sum_k \mu_{ik}\Delta\Gamma_{ik}$$

with $\mu_{ij} = \mu_{ij}(t, X(t))$. If the limit in (3.2) is well defined, then it specifies infinitesimal probabilities which define a continuous time Markov chain (Brémaud, 1999). When the limit can be calculated exactly, then exact simulation methods are available (Gillespie, 1977), though in practice numerical schemes based on Euler approximations may be preferable (Gillespie, 2001; Tian and Burrage, 2004). The implicit representation in (3.2) suggests a numerical approximation where the $o(\delta)$ term in (3.2) is ignored. Discretizing time into units of δ and ignoring the term $o(\delta)$ in (3.2) corresponds to a multinomial death process Euler approximation to a population process with noise added to the parameters (Figure 3.1). The strength of

```
Divide the interval $[0,T]$ into $N$ intervals of width $\delta = T/N$
Set initial value $X(0)$
FOR $n = 0$ to $N-1$
    Generate noise increments $\{\Delta\Gamma_{ij} = \Gamma_{ij}(n\delta+\delta) - \Gamma_{ij}(n\delta)\}$
    Generate process increments $(\Delta N_{i1}, \dots, \Delta N_{i,i-1}, \Delta N_{i,i+1}, \Delta N_{ic}, R_i)$
        $\sim$ Multinomial$(X_i(n\delta), p_{i1}, \dots, p_{i,i-1}, p_{i,i+1}, \dots, p_{ic}, 1 - \sum_{k \neq i} p_{ik})$
    with $p_{ij} = p_{ij}(\{\mu_{ij}(n\delta, X(n\delta))\}, \{\Delta\Gamma_{ij}\})$ given in (3.3)
    Set $X_i(n\delta + \delta) = R_i + \sum_{j \neq i} \Delta N_{ji}$
END FOR
```

Figure 3.1: Euler scheme corresponding to a numerical solution of the Markov chain specified by (3.2).

the implicit representation, combined with plug-and-play methodology, is that it lets one proceed with modeling and data analysis even when the model is not analytically tractable.

Proposition III.1 demonstrates by construction some conditions under which (3.2) does indeed specify a Markov chain. Proposition III.2 shows that, with some extra assumptions, we can find a tractable form for the resulting infinitesimal probabilities. Proofs of these results are given in section 3.6.1.

Proposition III.1. *The following gives a construction of the process in (3.2), supposing that Γ_{ij} is independent of Γ_{ik} for $j \neq k$ and that $\mu_{ij}(t,x)$ is uniformly continuous as a function of t. Give the individuals in the population labels $1, \dots, \sum_i X_i(0)$. Let $C(\zeta, m)$ be the index of the compartment containing individual ζ after the individual's mth transition, for $1 \leq \zeta \leq \sum_i X_i(0)$, with $C(\zeta, 0)$ giving the location at time $t = 0$. Set $\tau_{\zeta,0} = 0$, and generate independent Exponential(1) random variables $M_{\zeta,0,j}$ for each ζ and $j \neq C(\zeta,0)$. Define $\tau_{\zeta,m,j}$ recursively for $m \geq 1$ and $j \neq C(\zeta, m-1)$ by*

$$\tau_{\zeta,m,j} = \inf\left\{t : \int_{\tau_{\zeta,m-1}}^{t} \mu_{C(\zeta,m-1),j}(s, X(s))\, d\Gamma_{C(\zeta,m-1),j}(s) > M_{\zeta,m-1,j}\right\}.$$

Individual ζ *makes its* m*th move at time* $\tau_{\zeta,m} = \min_j \tau_{\zeta,m,j}$ *into state* $C(\zeta, m) = \arg\min_j \tau_{\zeta,m,j}$*, at which time new independent transition clocks* $\{M_{\zeta,m,j}, j \neq C(\zeta, m)\}$ *are generated.*

Proposition III.2. *Suppose, in addition to the assumptions of Proposition III.1, that the integrated noise processes* $\{\Gamma_{ij}\}$ *are all independent. The transition probabilities given in (3.2) are*

$$P[\Delta N_{ij} = n_{ij},\ \textit{for all } i \neq j \mid X(t) = (x_1, \ldots, x_c)] = \prod_i \prod_{j \neq i} \pi(n_{ij}, x_i, \mu_{ij}, \sigma_{ij}) + o(\delta)$$

where

$$\pi(n, x, \mu, \sigma) = 1_{\{n=0\}} + \delta \binom{x}{n} \sum_{k=0}^{n} \binom{n}{k} (-1)^{n-k+1} \sigma^{-2} \ln\big(1 + \sigma^2 \mu(x - k)\big). \tag{3.4}$$

In the special case where $\sigma_{ij} = 0$, we interpret $\xi_{ij}(t) = 1$. If $\sigma_{ij} = 0$ for all i and j, then (3.2) becomes the Poisson system widely used to model demographic stochasticity in population models (Brémaud, 1999; Bartlett, 1960). Constructions similar to Proposition III.1 are standard for Poisson systems (Brémaud, 1999), but here care is required to deal with the novel inclusion of white noise in the rate process. Our formulation for adding noise to Poisson systems can be seen as a generalization of subordinated Lévy processes (Sato, 1999), though we are not aware of previous work on the more general Markov processes constructed here. It is only the recent development of appropriate inference methodology that has led to the need for flexible Markov chain models with random rates.

Following what might be called the "plug and play principle," one could suppose that simulation from an arbitrarily accurate numerical approximation is sufficient to answer the questions that the model has been constructed to address. In particular, any property of the model which is stable as the numerical approximation timestep,

δ, approaches 0 may be presumed to be a property of the limiting continuous time Markov process. This need not always be true, which is one reason why analytic properties, such as Propositions III.1 and III.2, are valuable.

Another reason for being content with a numerical approximation for sufficiently small δ is that there may be no scientific reason to prefer a true continuous time model over a fine discretization. For example, when modeling population dynamics, continuous time models of adequate simplicity for data analysis typically will not include diurnal effects. Thus, there is no particular reason to think the continuous time model more credible than a discrete time model with a step of one day. One can think of a set of equations defining a continuous time process, combined with a specified discretization, as a way of writing down a discrete time model, rather than treating the continuous time model as a gold standard against which all discretizations must be judged.

The full independence of $\{\Gamma_{ij}\}$ assumed in Proposition III.2 gives a form for the limiting probabilities where multiple individuals can move simultaneously between some pair of compartments i and j, but no simultaneous transitions occur between different compartments. In more generality, the limiting probabilities do not have this simple structure. In the setup for Proposition III.1, where Γ_{ij} is independent of Γ_{ik} for $j \neq k$, no simultaneous transitions occur out of some compartment i into different compartments $j \neq k$, but simultaneous transitions from i to j and from i' to j' cannot be ruled out for $i \neq i'$. The assumption in Proposition III.1 that Γ_{ij} is independent of Γ_{ik} for $j \neq k$ is not necessary for the construction of a process via (3.2), but simplifies the subsequent analysis. Without this assumption, a construction similar to Proposition III.1 would have to specify a rule for what happens when an individual who has two simultaneous event times, i.e., when $\min_j \tau_{\zeta,m,j}$ is not uniquely attained. Although independence assumptions are useful for analytical results, a major purpose of the formulation in (3.2) is to allow the practical use of models that surpass currently available mathematical analysis. In particular, it may be natural for different transition processes to share the same noise process, if they correspond to transitions between similar pairs of states.

3.3 Plug-and-play inference methodology

Inference can be carried out for the framework of Section 3.2 using the iterated filtering methodology proposed by Ionides et al. (2006), implemented as described in Figure 3.2. This technique maximizes the likelihood for a partially observed Markov model, permitting calculation of maximum likelihood point estimates, confidence intervals (via profile likelihood, bootstrap or Fisher information), and likelihood ratio hypothesis tests. For non-linear non-Gaussian partially observed Markov models, the likelihood function can typically be evaluated only inexactly and at considerable computational expense. The iterated filtering procedure takes advantage of the partially observed Markov structure to enable computationally efficient maximization. A useful property of partially observed Markov models is that, if the parameter θ is replaced by a random walk θ_n with $\theta_0 = \theta$, the calculation of $\hat{\theta}_n = E[\theta_n|y_{1:n}]$ and $V_n = Var(\theta_n|y_{1:n-1})$ is a well-studied and computationally convenient filtering problem. Ionides et al. (2006) showed that a procedure which iteratively updates an estimate of θ by a weighted average of $\hat{\theta}_{1:N}$, with weights depending on $V_{1:N}$, while progressively decreasing the variance of the random walk, converges to the maximum of the likelihood function (under appropriate conditions). If the filtering technique is plug-and-play then maximization by iterated filtering also has this property. Basic sequential Monte Carlo filtering techniques, although usually written in terms of transition densities (Arulampalam et al., 2002; Doucet et al., 2001), do have the plug-and-play property. In Figure 3.2 we emphasize this by specifying a Markov process at a sequence of times $t_0 < t_1 < \cdots < t_N$ via a recursive transition rule,

$$X(t_n) = f(X(t_{n-1}), t_{n-1}, t_n, \theta, W),$$

where it is understood that W is some random variable which is drawn independently each time $f(.)$ is evaluated. In the context of the plug-and-play philosophy, $f(.)$ is the algorithm to generate a simulated sample path of $X(t)$ at the discrete times $t_1, \ldots, t_N$ given an initial value $X(t_0)$.

Other plug-and-play inference methodologies applicable to the models of Section 3.2 have been developed. Nonlinear forecasting (Kendall et al., 1999) has neither the statistical efficiency of a likelihood-based method nor the computational efficiency of a filtering-based method. The Bayesian sequential Monte Carlo approximation of Liu and West (2001) combines likelihood-based inference with a filtering algorithm, but is not supported by theoretical guarantees comparable to those presented by Ionides et al. (2006) for iterated filtering.

3.4 A mechanistic model for competing strains of cholera

We consider a compartment model for cholera dynamics subject to competing strains of pathogen, the bacterium *Vibrio cholerae.* All infectious diseases have a variety of strains, and a good understanding of the strain structure is key to understanding the epidemiology of the disease, developing effective vaccines and vaccination strategies, and understanding evolution of resistance to medication (Grenfell et al., 2004). If strain structure is excluded from a disease model, any features due to strain variability will be attributed to other elements of the model, which is likely to result in ineffective early warning systems and/or inefficient vaccination strategies.

MODEL INPUT: $f(\cdot)$, $g(\cdot|\cdot)$, $y_1, \dots, y_N$, $t_0, \dots, t_N$

ALGORITHMIC PARAMETERS: integers J, L, M; scalars $0 < a < 1$, $b > 0$; vectors $X_I^{(1)}$, $\theta^{(1)}$; positive definite symmetric matrices Σ_I, Σ_θ.

```
FOR $m = 1$ to $M$
    $X_I(t_0, j) \sim N[X_I^{(m)}, a^{m-1}\Sigma_I], \quad j = 1, \dots, J$
    $X_F(t_0, j) = X_I(t_0, j)$
    $\theta(t_0, j) \sim N[\theta^{(m)}, ba^{m-1}\Sigma_\theta]$
    $\bar{\theta}(t_0) = \theta^{(m)}$
    FOR $n = 1$ to $N$
        $X_P(t_n, j) = f(X_F(t_{n-1}, j), t_{n-1}, t_n, \theta(t_{n-1}, j), W)$
        $w(n, j) = g(y_n | X_P(t_n, j), t_n, \theta(t_{n-1}, j))$
        draw $k_1, \dots, k_J$ such that $\mathrm{Prob}(k_j = i) = w(n, i) / \sum_\ell w(n, \ell)$
        $X_F(t_n, j) = X_P(t_n, k_j)$
        $X_I(t_n, j) = X_I(t_{n-1}, k_j)$
        $\theta(t_n, j) \sim N[\theta(t_{n-1}, k_j), a^{m-1}(t_n - t_{n-1})\Sigma_\theta]$
        Set $\bar{\theta}_i(t_n)$ to be the sample mean of $\{\theta_i(t_{n-1}, k_j), j = 1, \dots, J\}$
        Set $V_i(t_n)$ to be the sample variance of $\{\theta_i(t_n, j), j = 1, \dots, J\}$
    END FOR
    $\theta_i^{(m+1)} = \theta_i^{(m)} + V_i(t_1) \sum_{n=1}^{N} V_i^{-1}(t_n)(\bar{\theta}_i(t_n) - \bar{\theta}_i(t_{n-1}))$
    Set $X_I^{(m+1)}$ to be the sample mean of $\{X_I(t_L, j), j = 1, \dots, J\}$
END FOR
```

RETURN
maximum likelihood estimate for parameters, $\hat{\theta} = \theta^{(M+1)}$
maximum likelihood estimate for initial values, $\hat{X}(t_0) = X_I^{(M+1)}$
maximized log likelihood estimate, $\lambda(\hat{\theta}) = \sum_n \log(\sum_j w(n, j)/J)$

Figure 3.2: Implementation of likelihood maximization by iterated filtering. $N[\mu, \Sigma]$ corresponds to a normal random variable with mean vector μ and covariance matrix Σ. $X(t_n)$ takes values in $\mathcal{R}^{d_x}$, y_n takes values in $\mathcal{R}^{d_y}$, θ takes values in $\mathcal{R}^{d_\theta}$ and has components $\{\theta_i, i = 1, \dots, d_\theta\}$.

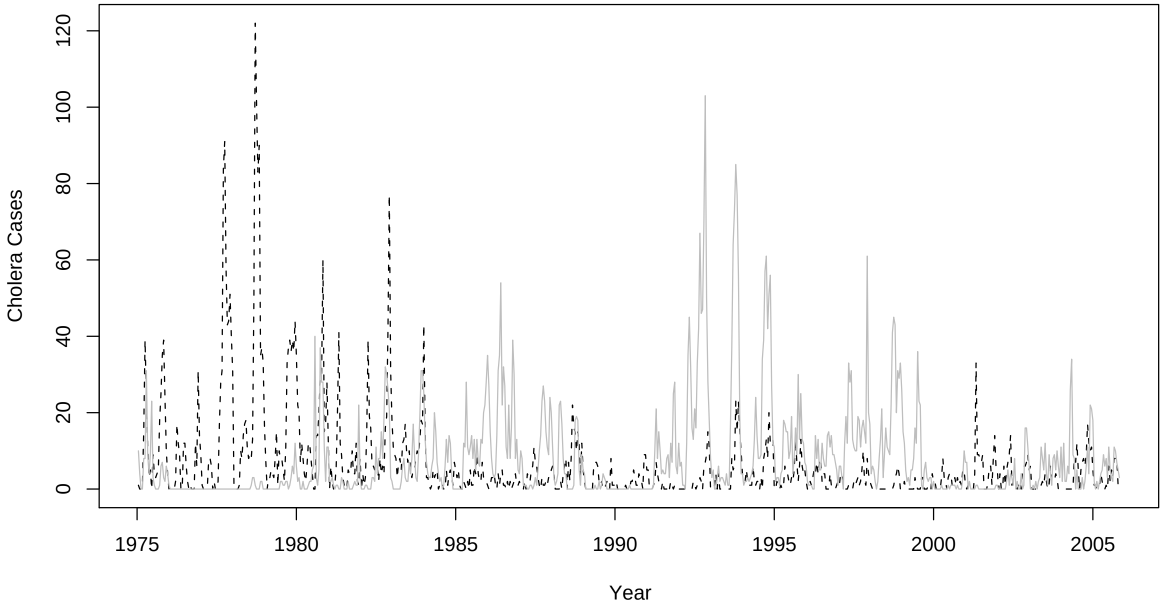

Figure 3.3: Biweekly cholera cases for Matlab, Bangladesh, obtained from hospital records of the International Center for Diarrheal Disease Research, Bangladesh. Cases are separated by serotype into Inaba (dashed) and Ogawa (solid grey)

Previous analyses relating mathematical consequences of strain structure to disease data include studies of malaria (Gupta et al., 1994), dengue (Ferguson et al., 1999), influenza (Ferguson et al., 2003; Koelle et al., 2006a) and cholera (Koelle et al., 2006b). These previous analyses of strain-structured time series data have drawn statistical inferences based on ad-hoc comparisons of simulations from the model with observed data. The goal of this current example is to demonstrate that the mechanistic modeling framework developed here permits likelihood based inference for mechanistically motivated stochastic models of strain-structured disease systems. We analyze a time series recording 30 years of biweekly cholera incidence in Matlab, Bangladesh (Figure 3.3), previously studied by Koelle et al. (2006b). Each cholera case was classified into one of two serotypes, Inaba and Ogawa. Exposure to one serotype results in strong immunity to that serotype, and weaker cross-immunity to the other.

Figure 3.4 describes the compartments for our two-strain model, with arrows showing possible transitions and their labels showing the corresponding rate of flow. This model combines the multistrain modeling approach of Kamo and Sasaki (2002) with previous compartmental models for single-strain cholera (Koelle and Pascual, 2004). Table 3.1 gives a formal interpretation of our model as a Markov chain with stochastic rates, in the framework of Section 3.2. Here, λ_1 is the force of infection for the Inaba serotype, i.e. the mean rate at which susceptible individuals become infected; ξ_1 is the stochastic noise on this rate; λ_2 and ξ_2 are the corresponding force of infection and noise for Ogawa; $\beta(t)$ is the rate of transmission between individuals, parameterized with a trend and a smooth seasonal component; ω gives the rate of infection from an environmental reservoir, independent of the current number of contagious individuals; the exponent α allows for inhomogeneous mixing of the population; r is the recovery rate from infection; γ measures the strength of cross-immunity between serotypes. In this model, following Koelle et al. (2006b), acquired immunity to a given serotype is life-long subsequent to infection with that serotype. The argument for giving both strains common variability is that they are believed to be biologically similar except in regard to immune response. The argument for giving the separate strains independent noise components is that noise represents chance events, such as a contaminated feast or a single community water source which is transiently in a favorable condition for contamination, and such events spread whichever strain is in the required place at the required time.

In addition, a measurement model is required. Biweekly aggregated cases for Inaba and Ogawa strains are denoted by $C_{i,t} = N_{SI_i}(t) - N_{SI_i}(t-1) + N_{S_iI_i^*}(t) - N_{S_iI_i^*}(t-1)$ for $i = 1, 2$ respectively. Reporting rates $\rho_{1,t}$ and $\rho_{2,t}$ are taken to be independent $Gamma(1/\phi, \rho\phi)$ random variables. Conditional on $\rho_{1,t}$ and $\rho_{2,t}$, the observations are modeled as independent Poisson counts,

$$Y_{i,t}|\rho_{i,t}, C_{i,t} \sim Poisson(\rho_{i,t}C_{i,t}), \quad i = 1, 2.$$

Thus, $Y_{i,t}$ given $C_{i,t}$ has a negative binomial distribution with $E[Y_{i,t}|C_{i,t}] = \rho C_{i,t}$ and $Var(Y_{i,t}|C_{i,t}) = \rho C_{i,t} + \phi\rho^2 C_{i,t}^2$. This model allows for the possibility of both under-reporting and over-reporting (mis-diagnosis), as well as both demographic stochasticity (i.e., Poisson variability) and environmental stochasticity (i.e., Gamma variability on the rates).

Note that this model is of the SIR (susceptible-infectious-removed) type for each strain, in contrast with the SIRS model proposed for the analysis of the cholera mortality data in chapter II. SIRS models allow for loss of immunity, i.e. individuals may cycle from susceptible to infectious to immune and back to susceptible several times in the course of their lives. This agrees with the observed pattern that an individual recently infected with cholera is unlikely to become infected again soon but could very well become infected in the future. The multistrain SIR on the other hand only allows each individual to become infectious with cholera twice, once with each strain.

Although these might appear to be two fundamentally different models for cholera immunity, the following observation reconciles the use of these two models. Given the right parameters, the two serotype model predicts that serotypes will alternate as the dominant serotype. An analysis that ignored the serotype structure, as is the case in chapter II, would then conclude that cholera confers temporary immunity. The duration of immunity would be confounded with the duration of dominance of the serotype of the first infection. Individuals that became infectious with the replacing serotype would seem to have lost the immunity conferred by the previous infection.

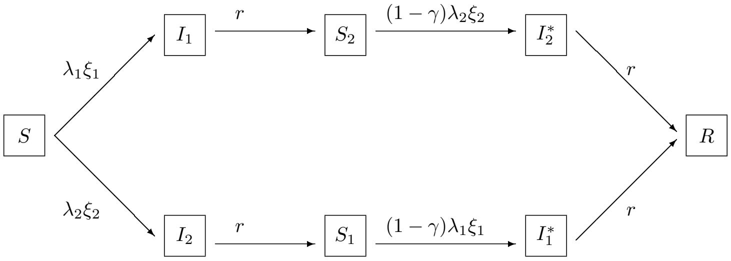

Figure 3.4: Flow diagram for cholera, including interactions between the two major serotypes. Each individual falls in one compartment: S, susceptible to both Inaba and Ogawa serotypes; I_1, infected with Inaba; I_2, infected with Ogawa; S_1, susceptible to Inaba (but immune to Ogawa); S_2, susceptible to Ogawa (but immune to Inaba); I_1^*, infected with Inaba (but immune to Ogawa); I_2^*, infected with Ogawa (but immune to Inaba); R, immune to both serotypes. Births enter S, and all individuals have a mortality rate m.

A multistrain SIRS model would allow for loss of serotype specific immunity. In this model more infections over the period of time for which the data is observed would be possible for any given value of γ. However, the overall picture in terms of number of infections could be similar if the possibility for additional infections were compensated by a lower infection force, which would be the case for smaller values of b_i, ω, α or any combination of those. The values in the new profile likelihood would most likely be higher and the curve flatter since the multistrain SIR model is the limit of the SIRS letting the rate of loss of immunity go to infinity. Exploring this more complex model is straightforward but it represents an additional considerable computational expense.

Some results from fitting the model in Figure 3.4 via the method in Figure 3.2 are shown in Table 3.2. The two sets of parameter values $\hat{\theta}_A$ and $\hat{\theta}_B$ in Table 3.2 are maximum likelihood estimates, with $\hat{\theta}_A$ having the additional constraints $\rho = 0.06$ and $r = 38.4$. The additional constraints results in a qualitatively different fitted model, and we refer to the neighborhoods of these two parameter sets as regimes A and B. These constraints were used by Koelle et al. (2006b) so, to the extent that the model there and in this chapter are comparable, A corresponds to a regime comparable to Koelle et al. (2006b). Regime B can be distinguished by a much higher

λ_1	$=$	$\beta(t)(I_1(t)+I_1^*(t))^\alpha/P(t)+\omega$	λ_2	$=$	$\beta(t)(I_2(t)+I_2^*(t))^\alpha/P(t)+\omega$
$\log\beta(t)$	$=$	$b_0(t-1990)+\sum_{i=1}^{12} b_i s_i(t)$			
μ_{SI_1}	$=$	λ_1	μ_{SI_2}	$=$	λ_2
$\mu_{S_1 I_1^*}$	$=$	$(1-\gamma)\lambda_1$	$\mu_{S_2 I_2^*}$	$=$	$(1-\gamma)\lambda_2$
$\mu_{I_1 S_2}$	$=$	$\mu_{I_2 S_1}=r$	$\mu_{I_2^* R}$	$=$	$\mu_{I_1^* R}=r$
$\mu_{X_j D}$	$=$	m for $X_j\in\{S,I_1,I_2,S_1,S_2,I_1^*,I_2^*,R\}$			
ξ_{SI_2}	$=$	$\xi_{S_2 I_2^*}=\xi_2(t)$	ξ_{SI_1}	$=$	$\xi_{S_1 I_1^*}=\xi_1(t)$

Table 3.1: Interpretation of Figure 3.4 via the multinomial process with random rates in (3.2), with $X(t)=(S(t),\ I_1(t),\ I_2(t),\ S_1(t),\ S_2(t),\ I_1^*(t),\ I_2^*(t),\ R(t),\ B(t),\ D(t))$. Compartments B and D are introduced for demographic considerations: births are formally treated as transitions from B to S and deaths as transitions into D. All transitions not listed above have zero rate. $\xi_2(t)$ and $\xi_1(t)$ are independent gamma noise processes, both with infinitesimal variance parameter σ^2. Transition rates are noise-free unless specified otherwise. $\{s_i(t), i=1,\dots,6\}$ is a basis of periodic cubic B-splines, with $s_i(t)$ attaining its maximum at $t=(i-1)/6$. The population size $P(t)$ is assumed known, interpolated from census data. The birth process is treated as a covariate, i.e., the analysis is carried out conditional on the process $N_{BS}(t)=\lfloor P(t)-P(0)+\int_0^t mP(s)\,ds\rfloor$, where $\lfloor x\rfloor$ is the integer part of x. There is a small stochastic discrepancy between $S(t)+I_1(t)+I_2(t)+S_1(t)+S_2(t)+I_1^*(t)+I_2^*(t)+R(t)$ and $P(t)$. In principle, one could condition on the demographic data by including a census measurement model—we saw no compelling reason to add this extra complexity for the current purposes. Numerical solutions of sample paths were calculated using the algorithm in Figure 3.1, with $\delta=2/365$.

reporting rate ($\rho=0.65$). Fig. 3.6 shows a profile likelihood for cross-immunity in regime A, which can be considered a formal statistical analysis of the aspects of the disease dynamics analyzed by Koelle et al. (2006b).

These two regimes demonstrate two distinct uses of a statistical model—firstly, to investigate the consequences of a set of assumptions and, secondly, to challenge those assumptions. If we decide to limit the investigation to regimes restricted by estimates obtained from previous studies, then the resulting parameter estimates $\hat{\theta}_A$ are broadly consistent with previous understanding of cholera dynamics, except that cross-immunity appears to be lower than previously thought. However, the results in Table 3.2 raise an alternative possibility, that the data are better explained by scenario B, for which the epidemiologically relevant cases are only the severe cases that are likely to result in hospitalization. Unlike in regime A, asymptomatic cholera cases play almost no role in regime B and cross-immunity is high, corresponding to lower numbers of infections and rate of disease transmission than in A. This is consistent with the calculations in Koelle et al. (2006b) that conclude that cross-immunity should be high (but not with $\rho = 0.06$, which is assumed in some sections of Koelle et al., 2006b).

The contrast between these regimes highlights a conceptual limitation of compartment models: disease severity and level of infectiousness are continuous, not discrete or binary as in basic compartment models. Differences in the level of morbidity required to be classified as "infected" result in re-interpretation of the parameters of the model, and consequently of other fundamental model characteristics such as the basic reproductive ratio of an infectious disease (Anderson and May, 1991). Despite this limitation, it remains the case that compartment models are a fundamental tool for understanding and describing disease dynamics, so identification and comparison of different interpretations is a worthwhile exercise.

Demonstrating the existence of a regime such as B shows that there is room for improvement in the model by departing from the assumptions in A, but care is required to interpret the finding scientifically. Extending the model to include differing levels of severity might permit a combination of the scientific interpretation of A with the data-matching properties of B. This explanation would be sensible if low reporting in the past was due to inaccurate reporting rather than infectious individuals which did not have symptoms. Re-interpreting cholera epidemiology is beyond the scope of this chapter, but we refer the reader to King et al. (2007) for an example of how the discovery of an unexpected combination of parameters which explain the data well can be used as the basis of a scientific argument. King et al. (2007) investigated a model for cholera, without multiple strain structure, and without an assumption of lifelong immunity following infection. Permitting loss of immunity over time, the best fitting model has immunity waning more quickly than previously supposed. Corroborative evidence was then found to support the epidemiological importance of short-term immunity acquired by exposure to low doses of the pathogen.

Likelihood based modeling of ecological and epidemiological systems has been criticized for leading to models for which simulated realizations do not look qualitatively like the data, and which have correspondingly poor medium and long term forecasting capabilities (good short term forecasting is a necessary consequence of the likelihood criterion for fitting models, resulting from the factorization $f(y_{1:T}|\theta) = \prod_{t=1}^{T} f(y_t|y_{1:t-1}, \theta)$). If simulations from the fitted model do not resemble the data, that provides another diagnostic that the class of models is inadequate. Here, simulations from B are generally qualitatively similar to the data, whereas those from A are too explosive (Figure 3.5). Building the desirable features of B into strain-structured models would help to realize the potential forecasting improvement arising from the availability of strain information.

Previous analysis of multistrain models have emphasized the possibility of strain cycling. Further analysis of strain cycling might reveal and allow for prediction of serotype switching. Wavelet time series analysis of simulations from the stochastic model using $\hat{\theta}_A$ and $\hat{\theta}_B$ would help understand the implications of regimes A and B in terms of dominant frequencies and highlight strain cycling if present. Alternatively, spectral analysis could be used. Spectral analysis is most useful when the model considered is stationary. If the time variable in the multistrain SIRS model is replaced by, say, the initial time point, spectral analysis can be used to study simulations from this modified version of the model which is stationary. The same could be done at other times. Since the model includes a time-varying population size, this would

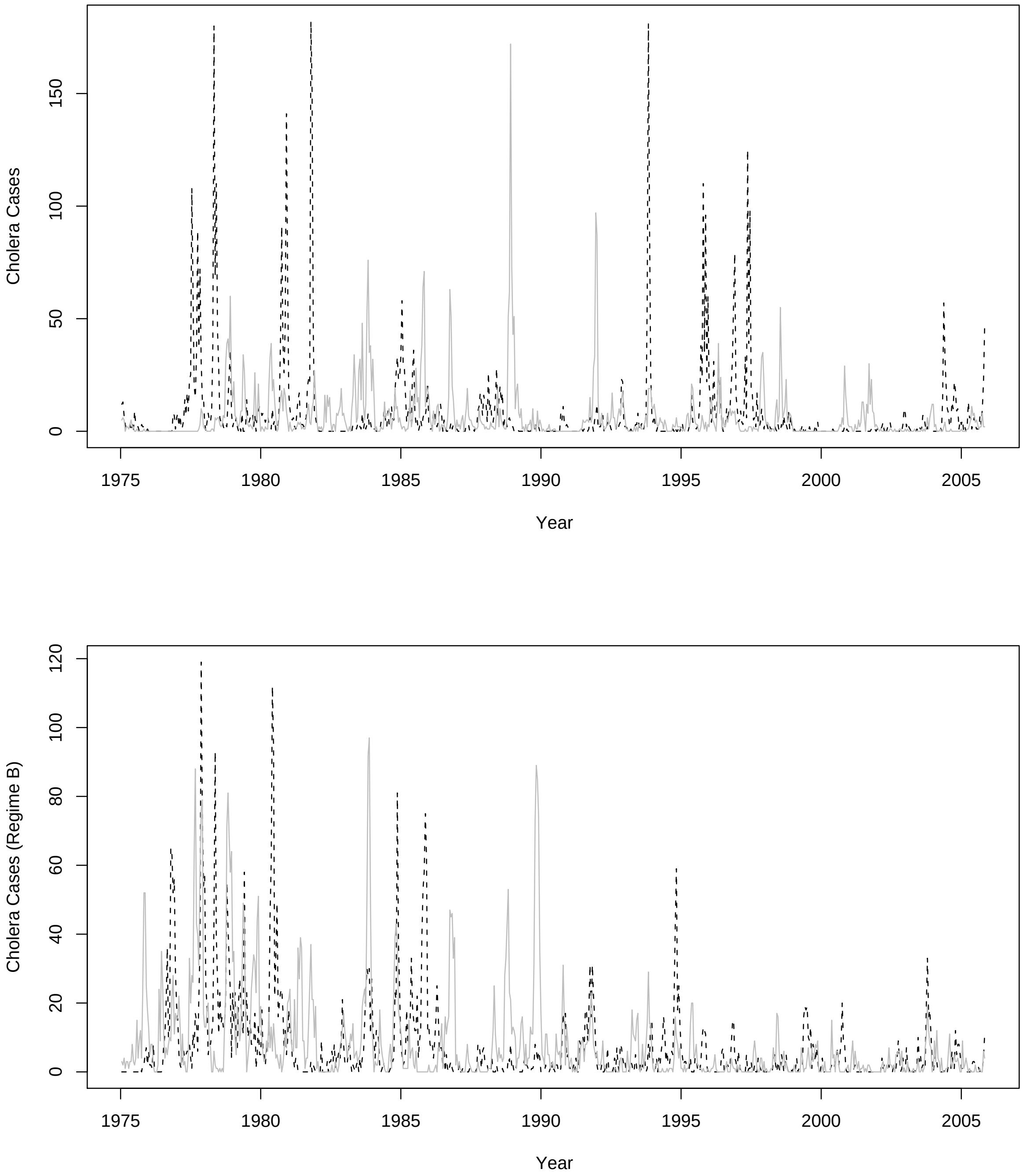

Figure 3.5: A simulated realization from regime A (top) and regime B (bottom), showing cases of Inaba (dashed) and Ogawa (solid grey). Compared to the data in Fig. 3.3, realizations from regime A typically have too many tall, narrow spikes in disease incidence.

	$\hat{\theta}_A$	$\hat{\sigma}_A$	$\hat{\theta}_B$	$\hat{\sigma}_B$
r	38.42	–	36.91	3.88
ρ	0.067	–	0.653	0.069
γ	0.400	0.087	0.9996	0.41
σ	0.1057	0.0076	0.0592	0.0075
ϕ	0.014	0.030	0.0004	0.024
$\omega \times 10^3$	0.099	0.022	0.0762	0.0072
α	0.860	0.015	0.864	0.017
b_0	-0.0275	0.0017	-0.0209	0.0015
b_1	4.608	0.098	3.507	0.083
b_2	5.342	0.074	3.733	0.091
b_3	5.723	0.075	4.448	0.055
b_4	5.022	0.076	3.534	0.065
b_5	5.508	0.064	4.339	0.053
b_6	5.804	0.059	4.339	0.039
ℓ	-3560.23		-3539.11	

Table 3.2: Parameter estimates from both regimes. In both regimes, the mortality rate m is fixed at $1/38.8$ years^{-1}. The units of r, b_0, and ω are year^{-1}; σ has units year$^{1/2}$; and ρ, γ, ϕ, α, and $b_1, \ldots, b_6$ are dimensionless. ℓ is the average of two log-likelihood evaluations using a particle filter with 120,000 particles. Optimization was carried out using the iterated filtering in Figure 3.2, with $M = 30$, $a = 0.95$ and $J = 15,000$. Optimization parameters were selected via diagnostic convergence plots (Ionides et al., 2006). Standard errors were derived via a Hessian approximation (Ionides et al., 2006). These standard errors are quickly obtained and give a reasonable idea of the scale of uncertainty, but profile likelihood based confidence intervals are more appropriate for careful inferences.

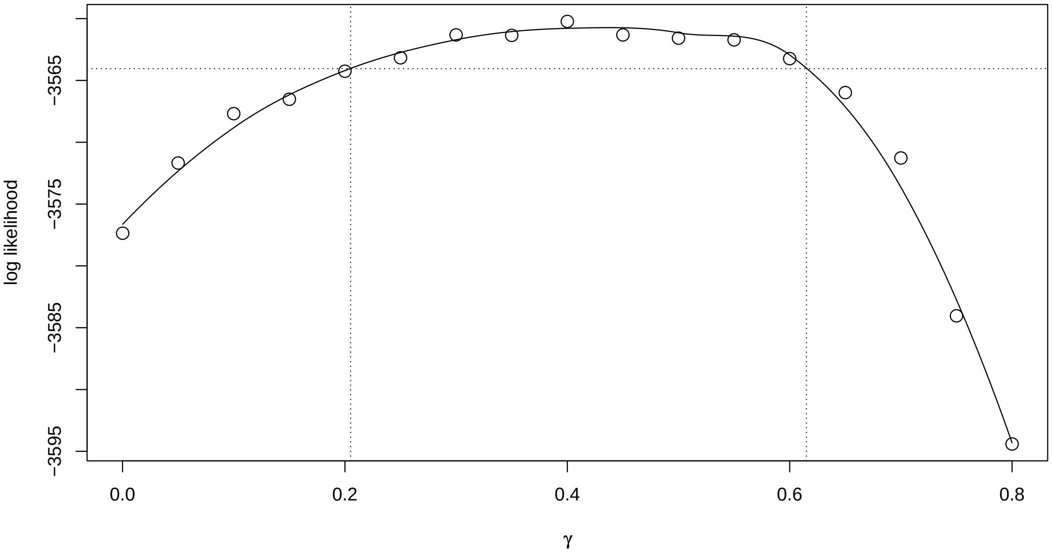

Figure 3.6: Cross-immunity profile likelihood computed as described in chapter II yielding a 99% confidence interval for γ of (0.20, 0.61). Local quadratic regression with a bandwidth parameter (span) of 0.6 was used to estimate the profile likelihood (solid line).

have to be fixed at the chosen time point. A more detailed study would involve comparing the cycling properties of the deterministic skeleton for $\hat{\theta}_A$ and $\hat{\theta}_B$ with the wavelet of spectral analysis of the stochastic counterpart.

In addition to strain specific prediction, overall case prediction ability is likely to be increased by incorporating information from environmental drivers, which have been shown to play a role in cholera dynamics in recent periods, like ENSO and rainfall (Pascual et al., 2000; Koelle et al., 2005). A residual analysis similar to that of chapter II may corroborate these relationships, as may a correlation analysis of the covariates and the structural residuals. This information could be incorporated in the force of infection as a linear component or some other form if suggested by the results.

3.5 Discussion

This discussion limits itself to the framework of compartment models, however these provide a rather broad perspective on the general topic of mechanistic models. Given rates μ_{ij}, one interpretation of a compartment model is to write the flows as coupled ordinary differential equations (ODEs),

$$\frac{d}{dt}N_{ij} = \mu_{ij}X_i(t). \tag{3.5}$$

Data analysis via ODE models has challenges in its own right (Ramsay et al., 2007). Previous work (Swishchuk and Wu, 2003) has included stochasticity by adding a slowly varying function to the derivative in (3.5).

Alternatively, one can add Gaussian white noise to give a set of coupled stochastic differential equations (SDEs) (e.g. Øksendal, 1998). For example, if $\{W_{ijk}(t)\}$ is a collection of independent standard Brownian motion processes, and $\sigma_{ijk} = \sigma_{ijk}(t, X(t))$, an SDE interpretation of a compartment model is given by

$$dN_{ij} = \mu_{ij} X_i(t)\, dt + \sum_k \sigma_{ijk} dW_{ijk}, \tag{3.6}$$

following the custom of writing the SDE as an infinitesimal equation rather than dividing through by dt. SDEs have some favorable properties for mechanistic modeling, such as the ease with which stochastic models can be written down and interpreted in terms of infinitesimal mean and variance (Ionides et al., 2006). However, there are several reasons to prefer integer-valued stochastic processes over SDEs for modeling population processes. Populations consist of discrete individuals, and, when a population becomes small, that discreteness can become important. For infectious diseases, there may be temporary extinctions, or “fade-outs,” of the disease in a population or sub-population. Even if the SDE is an acceptable approximation to the disease dynamics, there are technical reasons to prefer a discrete model. Standard methods allow exact simulation for continuous time Markov chains (Brémaud, 1999; Gillespie, 1977), whereas for an SDE this is at best difficult (Beskos et al., 2006). In addition, if an approximate Euler solution for a compartment model is required, non-negativity constraints can more readily be accommodated for Markov chain models, particularly when the model is specified by a limit of multinomial approximations, as in (3.2). The most basic discrete population compartment model is the Poisson system (Brémaud, 1999), given by

$$P[\Delta N_{ij} = n_{ij} | X(t) = (x_1, \ldots, x_c)] = \prod_i \prod_{j \neq i} (\mu_{ij} x_i \delta)^{n_{ij}} (1 - \mu_{ij} x_i \delta) + o(\delta). \tag{3.7}$$

The Poisson system is a Markov chain whose transitions consist of single individuals moving between compartments, i.e., the infinitesimal probability is negligible of either simultaneous transitions between different pairs of compartments or multiple transitions between a given pair of compartments. As a consequence of this the Poisson system is “equidispersed,” meaning that the infinitesimal mean of the increments equals the infinitesimal variance (section 3.6.2). Overdispersion is a ubiquitous feature of data (McCullagh and Nelder, 1989), and this leads us to consider models such as (3.2) for which the infinitesimal variance can exceed the infinitesimal mean. As a consequence, instantaneous transitions of more than one individual are possible. This may be scientifically plausible: a cholera-infected meal or water-jug may lead to several essentially simulaneous cases; many people could be simultaneously exposed to an influenza patient on a crowded bus. Even dis-regarding scientifically plausibility of multiple simultaneous transitions, if one wishes to write down an over-dispersed Markov model the inclusion of such possibilities is unavoidable. Simultaneity in the limiting continuous time model can alternatively be justified by arguing that the model only claims to capture macroscopic behavior over sufficiently long time intervals.

Note that the multinomial gamma limit used in (3.2) could be replaced by alternatives, such as Poisson gamma or negative binomial gamma. The latter are more natural for unbounded processes, such as birth processes. For equidispersed processes, the Euler approximation through taking these three different processes produces the same limiting process. For overdispersed processes, these limits differ. In particular, the Poisson gamma and negative binomial gamma Euler limits have unbounded jump distributions and so are less readily applicable to finite populations.

The approach in (3.2) of adding white noise to the transition rates differs from previous approaches of making the rates a slowly varying random function of time, i.e., adding low frequency “red noise” to the rates. There are several motivations for introducing this new class of models. Most simply, adding white noise is a more parsimonious parameterization, since the intensity but not the spectral shape of the noise needs to be considered. For smoothly varying rates, the infinitesimal mean and variance are still equal (section 3.6.1). At least for the cholera example, high-frequency variability in the rate of infection appears necessary to describe the data. We do not wish to imply that white noise should always be used to model variability in rates, but we do think that a demonstration of the possibility is of general interest.

Time series analysis is, by tradition, data oriented, and so the quantity and quality of available data may limit the questions that the data can reasonably answer. This forces a limit on the number of parameters that can be estimated for a model. Thus, a time series model termed mechanistic might be a simplification of a more complex model which more fully describes reductionist scientific understanding of the dynamical system. As one example, one could certainly argue for including age structure or other population inhomogeneities into Figure 3.4. Indeed, determining which additional model components lead to important improvement in the statistical description of the observed process is a key data analysis issue.

3.6 Appendix

3.6.1 Proofs of Propositions III.1 and III.2

We proceed to construct increasingly complex over-dispersed Markov chains. A fundamental building block of processes such as (3.2) is the over-dispersed binomial death process. Conditional on a Gamma process $\Gamma(t)$, with infinitesimal variance σ and corresponding noise process $\xi(t) = \frac{d}{dt}\Gamma(t)$, individuals from a population of initial size $X(0)$ each "die" from compartment X to Y at rate $\mu\xi(t)$. Here, $N(t) = N_{XY}(t) = X(0) - X(t)$ counts the total number of deaths occurring by time t. Also, we define $\Delta N = N(t+\delta) - N(t)$ and $\Delta\Gamma = \Gamma(t+\delta) - \Gamma(t) \sim \mathit{Gamma}(\delta/\sigma^2, \sigma^2)$. To give a construction of a Markov chain, suppose independent $\mathit{Exponential}(1)$ random variables $\{M_\zeta, \zeta = 1, \ldots, X(0)\}$ are generated at $t = 0$ and assigned to each member of the initial population. Individual ζ dies at time $\tau_\zeta = \inf\{t : \mu\Gamma(t) > M_\zeta\}$. $X(t)$ constructed in this way is a Markov chain, due to the memoryless property of $\{M_\zeta\}$ and the independent increments of $\Gamma(t)$. In this case,

$$P[\Delta N = n|X(t) = x] = E\Big[\binom{x}{n}(1-p)^{n-x}p^n\Big] \tag{3.8}$$

where $p = 1 - \exp\{-\mu\Delta\Gamma\}$. The limiting behavior of (3.8) as $\delta \to 0$ is analytically tractable, and is given in Lemma III.3.

Lemma III.3. *The limiting probabilities for (3.8) are given by* $P(\Delta N = n|X(t) = x) = \pi(n, x, \mu, \sigma) + o(\delta)$ *where* $\pi(n, x, \mu, \sigma)$ *is given in (3.4). The infinitesimal moments are*

$$\begin{aligned} E[\Delta N|X(t) = x] &= \delta x\sigma^{-2}\ln(1+\mu\sigma^2) + o(\delta) \\ Var[\Delta N|X(t) = x] &= \delta x\sigma^{-2}\Bigg\{x\ln\bigg(\frac{(1+\mu\sigma^2)^2}{1+2\mu\sigma^2}\bigg) + \ln\bigg(\frac{1+2\mu\sigma^2}{1+\mu\sigma^2}\bigg)\Bigg\} + o(\delta) \end{aligned}$$

Proof. For convenience, we define $a = \sigma^{-2}$ and $b = (\mu\sigma^2)^{-1}$. Thus, $\mu\Delta\Gamma \sim Gamma(a\delta, b^{-1})$

and, using G to denote the gamma function,

$$\begin{aligned}
P(\Delta N = n | X(t) = x) &= \int\limits_0^\infty \binom{x}{n} \left[1 - e^{-\lambda}\right]^n \left[e^{-\lambda}\right]^{x-n} \frac{\lambda^{\delta a - 1} e^{-\lambda b} b^{\delta a}}{G(\delta a)} d\lambda \\
&= \binom{x}{n} \int\limits_0^\infty \left[\sum_{k=0}^{n} \binom{n}{k} (-e^{-\lambda})^{n-k}\right] e^{-\lambda(x-n)} \frac{\lambda^{\delta a - 1} e^{-\lambda b} b^{\delta a}}{G(\delta a)} d\lambda \\
&= \binom{x}{n} \int\limits_0^\infty \sum_{k=0}^{n} \binom{n}{k} (-1)^{n-k} e^{-\lambda(x-k)} \frac{\lambda^{\delta a - 1} e^{-\lambda b} b^{\delta a}}{G(\delta a)} d\lambda \\
&= \binom{x}{n} \sum_{k=0}^{n} \binom{n}{k} (-1)^{n-k} \frac{b^{\delta a}}{(b + x - k)^{\delta a}} \int\limits_0^\infty \frac{\lambda^{\delta a - 1} e^{-\lambda(b + x - k)} (b + x - k)^{\delta a}}{G(\delta a)} d\lambda \\
&= \binom{x}{n} \sum_{k=0}^{n} \binom{n}{k} (-1)^{n-k} \left(1 + \frac{x-k}{b}\right)^{-\delta a} \\
&= \binom{x}{n} \sum_{k=0}^{n} \binom{n}{k} (-1)^{n-k} \left(1 - \delta a \ln\left(1 + \frac{x-k}{b}\right) + o(\delta)\right) \\
&= 1_{\{n=0\}} + \delta \binom{x}{n} \sum_{k=0}^{n} \binom{n}{k} (-1)^{n-k+1} a \ln\left(1 + \frac{x-k}{b}\right) + o(\delta)
\end{aligned}$$

The expression for $\pi(n, x, \mu, \sigma)$ in (3.4) then follows by substituting in the appropriate definitions. To calculate the infinitesimal moments, we note that the moment generating function of a gamma random variable (e.g., Casella and Berger, 1990) gives $E[e^{-\mu\Delta\Gamma}] = (1 + \mu\tau)^{-\delta/\tau}$, where $\tau = \sigma^2$. It follows that

$$Var[e^{-\mu\Delta\Gamma}] \;=\; E[e^{-2\mu\Delta\Gamma}] - \left(E[e^{-\mu\Delta\Gamma}]\right)^2 = (1 + 2\mu\tau)^{-\delta/\tau} - (1 + \mu\tau)^{-2\delta/\tau}$$

A Taylor series expansion around $\delta = 0$ gives

$$\begin{aligned}
E[e^{-\mu\Delta\Gamma}] &= 1 - \delta\tau^{-1} \ln(1 + \mu\tau) + o(\delta) \\
Var[e^{-\mu\Delta\Gamma}] &= \delta\tau^{-1}\left\{2\ln(1 + \mu\tau) - \ln(1 + 2\mu\tau)\right\} + o(\delta)
\end{aligned}$$

Conditional on $X(t)$ and $\Delta\Gamma$, $\Delta N \sim \mathit{Binomial}(X(t), 1 - e^{-\mu\Delta\Gamma})$. Standard identities for the conditional mean and variance (e.g., Casella and Berger, 1990, Chapter 4) then complete the proof. □

The infinitesimal moments calculated in Lemma III.3 are not needed for Propositions III.1 and III.2, but provide a contrast with the equidispersed property of Poisson systems discussed in section 3.6.2.

Now we return to the overdispersed Poisson system constructed in Proposition III.1. Recall that individual ζ is initially in compartment $C(\zeta,0)$ and independent $Exponential(1)$ random variables $M_{\zeta,0,j}$ are generated for each ζ and each $j \neq C(\zeta,0)$. Define $\tau_{\zeta,0}=0$ and, for $m \geq 1$ and $j \neq C(\zeta, m-1)$, recursively define "event times"

$$\tau_{\zeta,m,j} = \inf\Big\{t : \int_{\tau_{\zeta,m-1}}^{t} \mu_{C(\zeta,m-1),j}(s,X(s))d\Gamma_{C(\zeta,m-1),j} > M_{\zeta,m-1,j}\Big\}.$$

Individual ζ makes its mth move at the "transition time" $\tau_{\zeta,m} = \min_j \tau_{\zeta,m,j}$ from compartment $C(\zeta,m-1)$ into $C(\zeta,m) = \arg\min_j \tau_{\zeta,m,j}$, at which time a new independent transition clock $M_{\zeta,m,j}$ is generated. $X(t)$ constructed in this way is a Markov chain, due to the memoryless property of $\{M_{\zeta,j,m}\}$ and the independent increments of $\{\Gamma_{ij}(t)\}$.

To prove Propositions III.1 and III.2, we construct a sequence of related processes each of which are shown to give rise to transition probabilities over the time interval $[0,\delta]$ which differ by $o(\delta)$. Descriptively, $\{N'_{ij}(t)\}$ will fix the transition rates for each individual to their values at $t-0$; $\{N''_{ij}(t)\}$ will replace the white noise in the transition rates by its average over the interval $[0,\delta]$, and will consider only the first jump for each individual; $\{N'''_{ij}(t)\}$ will modify $\{N''_{ij}(t)\}$ by counting all the event times, rather than just the transition times.

Formally, $\{N'_{ij}(t)\}$ is constructed with event times given by

$$\tau'_{\zeta,j,m} = \inf\Big\{t : \int_{\tau'_{\zeta,m-1}}^{t} \mu_{C(\zeta,m-1),j}(0,X(0))d\Gamma_{C(\zeta,m-1),j} > M_{\zeta,m-1,j}\Big\},$$

where $C'(\zeta,0) = C(\zeta,0)$, $\tau'_{\zeta,m} = \min_j \tau'_{\zeta,j,m}$ and $C'(\zeta,m) = \arg\min_j \tau'_{\zeta,j,m}$. Now define

$$\begin{aligned} S &= \bigcap_{\zeta,j}\{\tau_{\zeta,2,j} > \delta\}\bigcap_{k\neq j}\{\tau_{\zeta,1,j} < \delta, \tau_{\zeta,1,k} < \delta\}^c \\ S' &= \bigcap_{\zeta}\bigcap_{j}\{\tau'_{\zeta,2,j} > \delta\}\bigcap_{k\neq j}\{\tau'_{\zeta,1,j} < \delta, \tau'_{\zeta,1,k} < \delta\}^c, \end{aligned}$$

the sets on which no individual has more than one event time $\tau_{\zeta,m,j}$ (or $\tau'_{\zeta,m,j}$) in $(0,\delta]$. Due to the assumption of independence between $\Gamma_{ij}(t)$ and $\Gamma_{ik}(t)$ for $j \neq k$, $P[\{\tau_{\zeta,1,j} < \delta, \tau_{\zeta,1,k} < \delta\}] = o(\delta)$. Similarly, $P[\{\tau'_{\zeta,1,j} < \delta, \tau'_{\zeta,1,k} < \delta\}] = o(\delta)$. It follows that

$$(3.9) \qquad P[S] = 1 - o(\delta), \quad P[S'] = 1 - o(\delta)$$

(though the chance of multiple individuals making transitions in $[0,\delta]$ may still be $O(\delta)$.) The uniform continuity of $\mu_{ij}(t, X(t))$ as a function of t, together with the observation that

$$P[X(t) \text{ makes} > 1 \text{ transition in } [0,\delta]] = o(\delta),$$

means that

$$(3.10) \qquad P[\tau_{\zeta,j,m} < \delta, \tau'_{\zeta,j,m} > \delta] = o(\delta), \qquad P[\tau_{\zeta,j,m} > \delta, \tau'_{\zeta,j,m} < \delta] = o(\delta).$$

From (3.9) and (3.10), transition probabilities for $\{N'_{ij}(\delta)\}$ can differ from those for $\{N_{ij}(\delta)\}$ by at most $o(\delta)$. Now construct $\{N''_{ij}(t)\}$ via event times given by

$$\tau''_{\zeta,j,1} = \inf\Big\{t : (t/\delta)\int_0^\delta \mu_{C(\zeta,m-1),j}(0,X(0))d\Gamma_{C''(\zeta,0),j} > M_{\zeta,0,j}\Big\},$$

with $C''(\zeta,0) = C(\zeta,0)$, $\tau''_{\zeta,1} = \min_j \tau''_{\zeta,j,1}$ and $C''(\zeta,1) = \arg\min_j \tau''_{\zeta,j,1}$, and

$$N''_{ij}(t) = \sum_{\zeta} I\{C''(\zeta,0) = i, C''(\zeta,1) = j, \tau''_{\zeta,1} \leq t\},$$

where I is an indicator function. Conditional on $\{\Gamma_{ij}\}$ and $X(0)$, $\{N''_{ij}(t)\}$ is constructed as a family of independent multinomial death processes on the time interval

$[0, \delta]$ for each $i = 1, \ldots, c$, and so the conditional distribution of $\{N''_{ij}(\delta)\}$ is a product of multinomial distributions. Therefore,

$$P[N''_{ij}(\delta) = n_{ij}, \text{ for all } 1 \le i \le c, 1 \le j \le c, i \ne j \mid X(0) = (x_1, \ldots, x_c), \{\Gamma_{ij}(\delta)\}]$$

(3.11)

$$= \prod_{i=1}^{c} \left\{ \begin{pmatrix} & & & x_i & & & \\ n_{i1} & \ldots & n_{ii-1} & n_{ii+1} & \ldots & n_{ic} & r_i \end{pmatrix} (1 - \textstyle\sum_{k \ne i} p_{ik})^{r_i} \displaystyle\prod_{j \ne i} p_{ij}^{n_{ij}} \right\}$$

with $r_i = x_i - \sum_{k \ne i} n_{ik}$ and $p_{ij} = p_{ij}(\{\mu_{ij}(0, X(0))\}, \{\Gamma_{ij}(\delta)\})$ given in (3.3). Notice that, for outcomes restricted to S' in (3.9), $N'_{ij}(t) = N''_{ij}(t)$. Therefore, transition probabilities for the process $N'_{ij}(\delta)$ can differ from those for $N''_{ij}(\delta)$ by at most $o(\delta)$. Taking expectations of both sides of (3.11), conditional on $X(t)$, matches (3.2) up to a term $o(\delta)$. It follows that the limiting probabilities specified by (3.2) agree with the construction in Proposition III.1 up to a term $o(\delta)$. Infinitesimal transition probabilities for which terms $o(\delta)$ are uniform in t characterize a finite state Markov chain. Therefore, the specification in (3.2) is well defined and results in the same Markov chain as the construction of Proposition III.1.

Now we define $N'''_{ij}(t) = \sum_{\zeta} \sum_{j \ne C(\zeta,0)} I\{C''(\zeta, 0) = i, \tau''_{\zeta,j,1} < t\}$. Under the hypothesis of Proposition III.2 the event times $\{\tau''_{\zeta,j,1}\}$ in this sum are independent, and so an application of Lemma III.3 gives

$$P[N'''_{ij}(\delta) = n_{ij}, \text{ for all } 1 \le i \le c, 1 \le j \le c, i \ne j \mid X(0) = (x_1, \ldots, x_c)]$$

$$= \prod_{i} \prod_{j \ne i} E\Big[\binom{x_i}{n_{ij}} (1 - \tilde{p}_{ij})^{n_{ij} - x_i} \tilde{p}_{ij}^{n_{ij}}\Big] \tag{3.12}$$

$$= \prod_{i} \prod_{j \ne i} \pi(x_i, n_{ij}, \mu_{ij}, \sigma_{ij}) + o(\delta) \tag{3.13}$$

where $\tilde{p}_{ij} = 1 - \exp(-\mu_{ij}(0, X(0))\Gamma_{ij}(\delta))$. On the event S in (3.9), $N'''_{ij} = N''_{ij}$, and so the transition probabilities for N'''_{ij} and N''_{ij} differ by $o(\delta)$. The calculation (3.13) thus proves Proposition III.2.

3.6.2 Equidispersion of Poisson Systems

For the Poisson system in (3.7),

$$P[\Delta N_{ij} = 0 | X(t) = (x_1, \ldots, x_c)] \;=\; 1 - \mu_{ij} x_i \delta + o(\delta) \tag{3.14}$$

$$P[\Delta N_{ij} = 1 | X(t) = (x_1, \ldots, x_c)] \;=\; \mu_{ij} x_i \delta + o(\delta) \tag{3.15}$$

where $\mu_{ij} = \mu_{ij}(t, x)$. Since the state space of $X(t)$ is finite, it is not a major restriction to suppose that there is some uniform bound $\mu_{ij}(t,x)x_i \leq \nu$, and that the terms $o(\delta)$ in (3.14,3.15) are uniform in x and t. Then, $P[\Delta N_{ij} > k | X(t)] \leq \bar{F}(k, \delta\nu)$ where $\bar{F}(k, \lambda) = \sum_{j=k+1}^{\infty} \lambda^j e^{-\lambda}/j!$. It follows that $\sum_{k=1}^{\infty} P[\Delta N_{ij} > k | X(t)] = o(\delta)$, and so

$$E[\Delta N_{ij} | X(t) = x] = \sum_{k=0}^{\infty} P[\Delta N_{ij} > k | X(t) = x] = \mu_{ij} x_i \delta + o(\delta) \tag{3.16}$$

Similarly,

$$E[(\Delta N_{ij})^2 | X(t) = x] = \sum_{k=0}^{\infty} (2k+1) P[\Delta N_{ij} > k | X(t) = x] = \mu_{ij} x_i \delta + o(\delta) \tag{3.17}$$

and so $Var(\Delta N_{ij} | X(t)) = \mu_{ij} X_i(t) \delta + o(\delta)$. If the rate functions $\mu_{ij}(X(t), t)$ are themselves stochastic, with $X(t)$ being a conditional Markov chain given $\{\mu_{ij}, 1 \leq i \leq c, 1 \leq j \leq c\}$, a similar calculation applies so long as a uniform bound ν still exists. In this case,

$$E[\Delta N_{ij} | X(t)] \;=\; \delta E[\mu_{ij} X_i(t) | X(t)] + o(\delta) \tag{3.18}$$

$$Var(\Delta N_{ij} | X(t)) \;=\; \delta E[\mu_{ij} X_i(t) | X(t)] + o(\delta) \tag{3.19}$$

The necessity of the uniform bound ν is demonstrated by the inconsistency between (3.18, 3.19) and the result in Proposition III.2 for the addition of white noise to the rates.

CHAPTER IV

Over-dispersed Continuous Time Markov Counting Processes

4.1 Introduction

This chapter presents a more complete theory of over-dispersed continuous time Markov counting processes. While in chapter III the stress was on a specific class of processes required for the data analysis, in this chapter we consider other types of population processes and focus on their analytic properties.

In model-based time series data analysis, time dependence is usually modeled via difference or differential relationships. Some data analysis techniques are based on entirely deterministic models, in which case a system of ordinary differential equations (ODE)

$$\dot{\boldsymbol{x}}(t) = \mu_{\boldsymbol{x}}\left(t, \boldsymbol{x}\left(t\right)\right) \tag{4.1}$$

is commonly used. In (4.1), $\boldsymbol{x}(t)$ is a vector valued function of time and $\dot{\boldsymbol{x}}(t)$ is the vector of first derivatives with respect to time. However, here we will consider analysis based on stochastic models. Stochastic differential equations (SDE) of the form

$$\boldsymbol{dX}(t) = \mu_{\boldsymbol{X}}\left(t, \boldsymbol{X}\left(t\right)\right) dt + \sigma_{\boldsymbol{X}}\left(t, \boldsymbol{X}\left(t\right)\right) \boldsymbol{dW}(t) \tag{4.2}$$

have been extensively studied (Øksendal, 1998). SDEs are intimately connected to deterministic ODEs. One may think of the solution to (4.2) as a process with the mean behavior driven by the solution to the ODE system specified by $\mu_{\boldsymbol{X}}$ (the infinitesimal mean function), with $\sigma_{\boldsymbol{X}}$ (the infinitesimal standard deviation function) determining the variability around it. Then, for "small enough" $\sigma_{\boldsymbol{X}}$, the solution to (4.2) may behave very much like the solution to (4.1), though even quite small amounts of process noise can have qualitative consequences (Coulson et al., 2004). Continuous time Markov chains (CTMC) form another family of commonly used models that has also been well studied (Brémaud, 1999). Examples of CTMCs used in different disciplines are: the Poisson process, linear pure birth process and linear pure death process. The linear pure birth and death processes in particular are CTMCs that may be seen as doubly stochastic Poisson processes (the intensity or rate of the stochastic process is random) with the rate depending on the process itself. Hence, we refer to $N(t)$ as a *self-exciting Poisson process* (Snyder and Miller, 1991) if it has transition probabilities given by

$$\begin{aligned}
P(\Delta N(t) = 0 | N(t) = n) &= 1 - \mu_N(t,n)\delta + o(\delta) \\
P(\Delta N(t) = 1 | N(t) = n) &= \mu_N(t,n)\delta + o(\delta) \\
P(\Delta N(t) > 1 | N(t) = n) &= o(\delta),
\end{aligned} \tag{4.3}$$

for $n \in \{0, 1, \dots\}$, and where the operator Δ is defined by $\Delta N(t) = N(t+\delta) - N(t)$. The dependence of $\Delta N(t)$ on δ will be suppressed. These univariate models, although quite simple, have been used in their own right and as building blocks for more complex models, giving rise to multivariate self-exciting Poisson processes, like queues and compartmental models (Brémaud, 1999; Jacquez, 1996). Although these multivariate extensions of self-exciting Poisson processes are more interesting from the point of view of actual applications and data analysis, we start by deriving analytical results for the simpler univariate case before addressing the more complex models. These models, like SDEs, are also tied to deterministic ODEs in the sense that the infinitesimal mean μ_N determines the mean behavior of the process. For large enough values of the integer valued process, the stochasticity becomes negligible and $N(t)$, as defined by (4.3), behaves very much like the solution of the corresponding ODE, $\dot{x}(t) = \mu_N(t, x(t))$.

An important difference between SDEs and self-exciting Poisson processes is the absence in the latter of some sort of counterpart of $\sigma_{\boldsymbol{X}}$ in (4.2), which would allow for a more flexible modeling of the variability. This is because, unlike in SDEs, there is a relationship between the mean and variance of $N(t)$ specified by (4.3). For this class of models, it is usual to choose a form for μ_N with a specific model in mind for the mean, which in turn implies a model for the variability. The exact mean-variance constraint is different for different processes. These constraints are derived explicitly for the homogeneous Poisson, linear pure birth and linear pure death processes in section 4.3. In particular, these processes are shown to have the same infinitesimal mean and variance, i.e. they are equi-dispersed, as discussed in section 4.2. This affects the properties of more complex models that use them as building blocks.

The techniques presented in chapter II provide general methods for fitting such models to time series data, as discussed in chapter III. From a data analysis point of view, this constraint on the moments is undesirable since it may result in a reduction in the goodness of the fit. Prior work has studied over-dispersion in count modeling (Gillespie, 1984; Takahata, 1987; Brown et al., 1998) but the tendency has been towards renewal processes (Snyder and Miller, 1991; Cutler, 2000; Wilson and Costello, 2005), moving away from the Markovian framework. The main contribution of this chapter is to present continuous time Markov counting processes for which it is possible to specify the first two infinitesimal moments independently, in the spirit of the SDE framework, making them capable of over-dispersion. These processes are constructed by modeling the individual event intensities or rates of the standard population models as independent gamma processes (although other Lévy processes could be used), making the rates state variables in a continuous time state space model. The resulting doubly stochastic self-exciting Poisson processes are in fact over-dispersed continuous time Markov counting processes. Adding noise to the event rates may be justified as a model for a stochastically changing environment or random media where the self-exciting Poisson process evolves.

Section 4.2 introduces the concept of dispersion of continuous time Markov counting processes which leads to the examples of equi-dispersed self-exciting Poisson processes of section 4.3. Then section 4.4 presents over-dispersed versions of homogeneous Poisson and linear pure death processes.

4.2 Dispersion of Continuous Time Markov Counting Processes

Previously considered measures of dispersion of continuous time Markov processes include $Var[N(t)]/E[N(t)]$ (Gillespie, 1984) and $Var(N(t))-E(N(t))$ (Brown et al., 1998). Given the conditional independence property of Markov processes and the treatment of time as continuous, we consider the ratio of infinitesimal moments a more appropriate measure of the dispersion of univariate Markov counting processes. We will refer to

$$D_N(t,n) \quad = \quad \frac{\lim_{\delta\downarrow 0}\delta^{-1}V[N(t+\delta)-N(t)|N(t)=n]}{\lim_{\delta\downarrow 0}\delta^{-1}E[N(t+\delta)-N(t)|N(t)=n]} \tag{4.4}$$

as the dispersion index of the Markov counting process $N(t)$. This index could be referred to as the infinitesimal dispersion index but, for ease of notation, we will usually omit "infinitesimal". The "non-infinitesimal" counterpart

$$AD_N(t, n_0) \quad = \quad \frac{V[N(t) - N(0)|N(0) = n_0]}{E[N(t) - N(0)|N(0) = n_0]}$$

is still of interest and we refer to it as the aggregated dispersion index of $N(t)$. A Markov counting process is then said to be equi-dispersed (on aggregate) if the (aggregate) dispersion index is 1, over-dispersed (on aggregate) if it is greater than 1 and under-dispersed (on aggregate) if it is smaller than 1.

Since the increments of a Markov counting processes are integer valued, the contribution of terms corresponding to increments of size zero and one is the same for all moments, i.e., for all $r \in \mathbb{N}$,

$$\begin{aligned} 0^r P(\Delta N(t){=}0|N(t){=}n) &= 0 \\ 1^r P(\Delta N(t){=}1|N(t){=}n) &= P(\Delta N(t){=}1|N(t){=}n), \end{aligned}$$

so that the difference between any two moments comes from terms corresponding to increments of size more than one. Usually, these terms are assumed to vanish as the time interval considered shrinks, i.e. $P(\Delta N(t){=}k|N(t){=}n) = o(\delta)$ for $k \geq 2$. Heuristically, this assumption implies that all infinitesimal moments should be equal, since these terms make the moments different and they vanish. This idea is formalized in lemma IV.2 at the end of this section. In addition, infinitesimally there is no difference between the second moment and the variance of the increments of a univariate Markov counting process under quite general conditions. It would follow from this that self-exciting Poisson processes are equi-dispersed since, as defined in (4.3), terms corresponding to increments of size more than one are $o(\delta)$. This additional insight is formalized in proposition IV.1 below, which provides sufficient conditions for equi-dispersion.

Note that the conditions in lemma IV.2 and proposition IV.1 are trivially satisfied if $\Delta N(t)$ is bounded (like in the cases of death processes), since it becomes a finite sum of $o(\delta)$ terms. Even though for unbounded process it is not so trivial to check whether these conditions hold, the insight provided by lemma IV.2 and proposition IV.1 is still useful. The homogeneous Poisson and simple linear birth processes are examples of processes where these probabilities disappear fast enough to cause equi-dispersion, which is checked by direct computation of the moments in section 4.3.

Proposition IV.1. *Provided that*

$$\sum_{k=2}^{\infty} k^2 P(\Delta N(t) = k | N(t) = n) \quad = \quad o(\delta),$$

the self-exciting Poisson process $N(t)$ is equi-dispersed, i.e.

$$\mu_N = \lim_{\delta \downarrow 0} \frac{E[\Delta N(t)|N(t) = n]}{\delta} = \sigma_N^2 = \lim_{\delta \downarrow 0} \frac{V[\Delta N(t)|N(t) = n]}{\delta}.$$

Proof. Since $\sum_{k=2}^{\infty} k^2 P(\Delta N(t) = k|N(t) = n) = o(\delta)$, it follows using lemma IV.2 that

$$E[\Delta N(t)|N(t) = n] - E[(\Delta N(t))^2|N(t) = n] \quad = \quad o(\delta).$$

(4.5)

Since

$$\begin{aligned} (E[\Delta N(t)|N(t) = n])^2 &= o(\delta) \\ V[\Delta N(t)|N(t) = n] &= E[(\Delta N(t))^2|N(t) = n] - (E[\Delta N(t)|N(t) = n])^2, \end{aligned}$$

the result follows. □

Lemma IV.2. *Provided that*

$$\sum_{k=2}^{\infty} k^r P(\Delta N(t) = k | N(t) = n) \quad = \quad o(\delta),$$

for $r \in \mathbb{N}$, any two infinitesimal moments of order smaller or equal to r are equal, i.e. for $s \leq r$

$$\lim_{\delta \downarrow 0} \frac{E[\Delta N^s(t)|N(t)=n]}{\delta} = \lim_{\delta \downarrow 0} \frac{E[\Delta N^r(t)|N(t)=n]}{\delta}.$$

Proof.

$$\begin{aligned} E[\Delta N(t)^s|N(t)=n] &= \sum_{k=0}^{\infty} k^s P(\Delta N(t)=k|N(t)=n) \\ &= P(\Delta N(t)=1|N(t)=n) + \sum_{k=2}^{\infty} k^s P(\Delta N(t)=k|N(t)=n), \end{aligned}$$

so that

$$\begin{aligned} E[\Delta N(t)^r|N(t)=n] - E[\Delta N(t)^s|N(t)=n] &= \sum_{k=2}^{\infty} (k^r - k^s) P(\Delta N(t)=k|N(t)=n) \\ &= o(\delta), \end{aligned}$$

since the condition that $\sum_{k=2}^{\infty} k^r P(\Delta N(t)=k|N(t)=n) = o(\delta)$, implies that

$$0 \leq \sum_{k=2}^{\infty} (k^r - k^s) P(\Delta N(t)=k|N(t)=n) \leq \sum_{k=2}^{\infty} k^r P(\Delta N(t)=k|N(t)=n) = o(\delta)$$

and the result follows. □

4.3 Equi-dispersed Continuous Time Markov Counting Processes

In this section we show that the counting processes associated with homogeneous Poisson, simple linear pure birth and simple linear death processes are all equi-dispersed and that, in spite of this, they are equi, over and under-dispersed on aggregate respectively. We do this by using the known exact distribution of these processes, which has been well known for a long time (Bailey, 1964; Bharucha-Reid, 1960), to explicitly compute the first two infinitesimal moments about the mean. These results are summarized in Table 4.1. This section provides the background for the over-dispersed processes introduced in section 4.4.

4.3.1 Poisson Process

According to (4.3), for $\mu_N(t,n) = \lambda$, $N(t)$ is a homogeneous Poisson process with rate λ. Under appropriate regularity conditions, the infinitesimal generator

$$\begin{aligned} q_{i,i} &= -\lim_{\delta \downarrow 0} \frac{1 - P(\Delta N(t) = 0 | N(t) = i)}{\delta} \\ q_{i,j} &= \lim_{\delta \downarrow 0} \frac{P(\Delta N(t) = j - i | N(t) = i)}{\delta} \text{ for } i < j, \\ q_{i,j} &= 0 \text{ for } j < i, \end{aligned}$$

for $i, j \in \{0, 1, \dots\}$ defines a continuous time Markov counting process (Brémaud, 1999). In the case of the homogeneous Poisson process, the infinitesimal generator is

$$\begin{aligned} q_{i,i} &= -\lambda \\ q_{i,i+1} &= \lambda \\ q_{i,j} &= 0, \text{ for } j \notin i, i+1. \end{aligned}$$

The increment process of a homogenous Poisson process follows a Poisson distribution with mean $\lambda\delta$, i.e.

$$P(\Delta N(t) = k | N(t) = n) = \frac{e^{-\lambda\delta}(\lambda\delta)^k}{k!},$$

for $k \in \{0, 1, \dots\}$. The mean and variance of the increment process are

$$\begin{aligned} E[\Delta N(t) | N(t) = n] &= \lambda\delta \\ V[\Delta N(t) | N(t) = n] &= \lambda\delta, \end{aligned}$$

from which it follows that the process is both equi-dispersed and equi-dispersed on aggregate, i.e.

$$D_N(t,n) = \frac{\lim_{\delta \downarrow 0} \delta^{-1} \lambda\delta}{\lim_{\delta \downarrow 0} \delta^{-1} \lambda\delta} = 1,$$

and

$$AD_N(t, n_0) \quad = \quad \frac{\lambda t - n_0}{\lambda t - n_0} = 1.$$

4.3.2 Simple Linear Pure Death Process

Setting $N(t)$ to satisfy (4.3) with $\mu_N(t, n) = \mu n 1_{\{n<n_0^*\}}$, where n_0^* is a positive integer representing the individuals in a population that are alive at time 0, $N(t)$ is the counting process associated with a simple linear death process with individual death rate μ. $N(t)$ then has infinitesimal generator

$$\begin{aligned} q_{i,i} &= -\mu i 1_{\{i<n_0^*\}} \\ q_{i,i+1} &= \mu i 1_{\{i<n_0^*\}} \\ q_{i,j} &= 0, \text{ for } j \notin i, i+1. \end{aligned}$$

We distinguish the standard death process, which we call $N^*(t)$, from its associated counting process defined by the infinitesimal generator above, $N(t)$, by having a star on the former. Starting with a population of size n_0^* at time 0, $N(t)$ counts the individuals that have died by time $t \geq 0$ and $N^*(t)$ is the number of individuals still alive by time t. It follows that $N(t) = n_0^* - N^*(t)$. The standard death process $N^*(t)$ is decreasing so it is not a counting process. To do a comparison with the other counting processes in this chapter, it is more relevant to consider $N(t)$. By definition, if $N(t) = n$, $N^*(t) = n_0^* - n$ and the distribution of the increment process of $N(t)$ conditional on $N(t) = n$ is binomial with parameters size $n_0^* - n$ and event probability $\pi(t) = 1 - e^{-\mu\delta}$, i.e.

$$P(\Delta N(t) = k | N(t) = n) \quad = \quad \binom{n_0^* - n}{k} \pi^*(t)^k (1 - \pi^*(t))^{(n_0^* - n) - k},$$

for $k \in \{0, 1, \ldots, n_0^* - n\}$. This implies, via the Taylor series expansion $0 \leq e^{-\mu\delta} = 1 - \mu\delta + o(\delta) \leq 1$,

$$
\begin{aligned}
E[\Delta N(t)|N(t) = n] &= (n_0^* - n)(1 - e^{-\mu\delta}) \\
&= (n_0^* - n)\mu\delta + o(\delta) \\
V[\Delta N(t)|N(t) = n] &= (n_0^* - n)(1 - e^{-\mu\delta})e^{-\mu\delta} \\
&= (n_0^* - n)(e^{-\mu\delta} - e^{-2\mu\delta}) \\
&= (n_0^* - n)\mu\delta + o(\delta) \\
D_N(t, n) &= \frac{\lim_{\delta\downarrow 0} \delta^{-1}((n_0^* - n)\mu\delta + o(\delta))}{\lim_{\delta\downarrow 0} \delta^{-1}((n_0^* - n)\mu\delta + o(\delta))} = 1.
\end{aligned}
$$

It follows that the process is equi-dispersed. To see that the process is under-dispersed on aggregate, note that $N(0) = n_0^* - N^*(0) = 0$ by definition, and

$$
AD_N(t, 0) = \frac{n_0^*(1 - e^{-\mu t})e^{-\mu t}}{n_0^*(1 - e^{-\mu t})} = e^{-\mu t} < 1,
$$

for $\mu > 0$.

4.3.3 Simple Linear Pure Birth Process

According to (4.3), for $\mu_N(t, n) = \beta n$, $N(t)$ is a simple linear birth process with individual birth rate β and has infinitesimal generator

$$
\begin{aligned}
q_{i,i} &= -\beta i \\
q_{i,i+1} &= \beta i \\
q_{i,j} &= 0, \text{ for } j \notin i, i + 1.
\end{aligned}
$$

Conditional on $N(t) = n$, the simple linear birth process $N(s)$ counts the number of individuals in the population at time $s \geq t$, which consists of the initial n individuals plus their descendants. The distribution of $N(s)$ given $N(t) = n$ is a negative binomial distribution with parameters number of successes n and success probability $\pi(t) = e^{-\beta\delta}$, i.e.

$$P(\Delta N(t) = k|N(t) = n) \quad = \quad \binom{n+k-1}{k}(e^{-\beta\delta})^n \left(1 - e^{-\beta\delta}\right)^k .$$

for $k \in \{0, 1, \dots\}$. This implies that, since $1 \leq e^{\beta\delta} = 1 + \beta\delta + o(\delta)$,

$$\begin{aligned} E[\Delta N(t)|N(t) = n] &= n(e^{\beta\delta} - 1) \\ &= n\beta\delta + o(\delta) \\ V[\Delta N(t)|N(t) = n] &= ne^{\beta\delta}(e^{\beta\delta} - 1) \\ &= n(e^{2\beta\delta} - e^{\beta\delta}) \\ &= n\beta\delta + o(\delta) \\ D_N(t, n) &= \frac{\lim_{\delta\downarrow 0} \delta^{-1}(n\beta\delta + o(\delta))}{\lim_{\delta\downarrow 0} \delta^{-1}(n\beta\delta + o(\delta))} = 1, \end{aligned}$$

from which equi-dispersion follows. However, the process is over-dispersed on aggregate,

$$AD_N(t, n_0) \quad = \quad \frac{n_0 e^{\beta t}(e^{\beta t} - 1)}{n_0(e^{\beta t} - 1)} = e^{\beta t} > 1,$$

for $\beta > 0$. Note that the expected value of $\Delta N(t)$ given $N(t)$ is the initial number of individuals in the population times

$$e^{\beta\delta} - 1 = \frac{1 - e^{-\beta\delta}}{e^{-\beta\delta}} = \frac{P(\text{Any individual gives birth to a new individual in } \delta)}{P(\text{Any individual does not give birth in } \delta)} ,$$

which is the odds of a birth (as opposed to the simple linear death process, where the expected value is the number of individuals alive times the *probability* of a death).

4.4 Over-dispersed Continuous Time Markov Counting Processes

This section presents over-dispersed counterparts of the homogeneous Poisson and linear pure death processes of section 4.3. For these processes we provide three

	Poisson	Birth	Death
$E[\Delta N(t) \vert N(t)]$	$\lambda\delta$	$N(t)(e^{\beta\delta}-1)$	$[n_0 - N(t)]\,(1-e^{-\mu\delta})$
$V[\Delta N(t) \vert N(t)]$	$\lambda\delta$	$N(t)e^{\beta\delta}(e^{\beta\delta}-1)$	$[n_0 - N(t)]\,(1-e^{-\mu\delta})e^{-\mu\delta}$
$AD(t, n_0)$	1	$e^{\beta t}$	$e^{-\mu t}$
$D(t, n)$	1	1	1

Table 4.1: Mean and variance of the increment of Poisson, simple linear death and simple linear birth counting processes used in population modeling. The dispersion and aggregate dispersion indices show that all three processes, in spite of not being equi-dispersed on aggregate, are indeed infinitesimally equi-dispersed. This motivates the results in section 4.4 regarding over-dispersed continuous time Markov counting processes.

results: their first two infinitesimal moments about the mean, which shows that they are indeed over-dispersed; the distribution of the counting processes, which allows for exact simulation of the counting processes; and a closed form for the infinitesimal generator, which may be used for exact simulation of the event times (point process).

To obtain these processes we add white noise to the rates of the equi-dispersed processes. The added noise should have positive increments to preserve the positiveness of the rates. Also, to retain the Markov property of the equi-dispersed processes, the increments should be independent. The class of Lévy processes provides a rich class of processes to choose from. The construction of these processes is similar to subordination of Lévy processes. In fact, the Poisson gamma process of section 4.4.1 is a subordinated Lévy process that has been analyzed before (Sato, 1999; Wolpert and Ickstadt, 1998), but not in the framework of over-dispersed Markov processes. The other processes we consider are not subordinated Lévy processes. Following the customary naming of subordinated Lévy processes, the name of the original process is placed first and followed by the name of the process proposed for the rate.

4.4.1 The Poisson Gamma Process

Consider the doubly stochastic homogeneous Poisson process $N(t)$ which, conditional on the gamma process $\Lambda(t)$, is the homogeneous Poisson process of section 4.3.1 with integrated rate $\Lambda(t)$ (see chapter III for a more detailed discussion), where $\Delta\Lambda(t) \sim \lambda\Gamma\left(\mu = \delta, \sigma^2 = \tau\delta\right)$, i.e.

$$P(\Delta N(t) = k | N(t), \Delta\Lambda) \quad = \quad \frac{e^{-\Delta\Lambda(t)}(\Delta\Lambda(t))^k}{k!}.$$

It is a standard result that the distribution of the increment process of $N(t)$ is negative binomial with probability mass function

$$P(\Delta N(t) = k | N(t) = n) = \frac{\Gamma\left(\tau^{-1}\delta + k\right)}{k!\Gamma(\tau^{-1}\delta)} p^{\tau^{-1}\delta} \left(1-p\right)^k,$$

where $p = \left(1+\omega\right)^{-1}$, $\omega = \tau\lambda$, $k \in \mathbb{N} \cup \{0\}$. It follows that the first two moments of $N(t)$ are $E[\Delta N(t)|N(t) = n] = \lambda\delta$ and $V[\Delta N(t)|N(t) = n] = (1+\omega)\lambda\delta$, giving a dispersion index of $D(t, n) = (1+\omega)$.

To obtain the limiting probabilities note that,

$$P(\Delta N(t) = 0 | N(t) = n) = p^{\tau^{-1}\delta} = 1 + \tau^{-1}\delta p^{\tau^{-1}\delta} \log p + o(\delta),$$

and, for $k \geq 1$,

$$\begin{aligned} P(\Delta N(t) = k | N(t) = n) \quad &= \quad \frac{\tau^{-1}\delta\,(k-1)!\Gamma(\tau^{-1}\delta) + o(\delta)}{k!\Gamma(\tau^{-1}\delta)} p^{\tau^{-1}\delta} \left(1-p\right)^k \\ &= \quad \tau^{-1} \frac{p^{\tau^{-1}\delta}\left(1-p\right)^k}{k} \delta + o(\delta), \end{aligned}$$

since, letting $\alpha = \tau^{-1}\delta$,

$$\Gamma(\alpha + k) \quad = \quad (\alpha + (k-1)) \times \ldots \times (\alpha + 1) \times (\alpha) \times \Gamma(\alpha),$$

$$(\alpha + (k-1)) \times (\alpha + (k-2)) \times \ldots \times (\alpha + 3) \times (\alpha + 2) \times \alpha \quad = \quad o(\delta) + (k-1)!\alpha,$$

and

$$\begin{aligned}\Gamma(\alpha+k) &= (o(\delta)+(k-1)!\alpha)\times(\alpha+1)\times\Gamma(\alpha)\\ &= (k-1)!\alpha\Gamma(\alpha)+o(\delta).\end{aligned}$$

The infinitesimal generator is then

$$\begin{aligned}q_{i,i} &= -\lim_{\delta\downarrow 0}\frac{1-p^{\tau^{-1}\delta}}{\delta}=\tau^{-1}\log p\\ q_{i,i+k} &= \tau^{-1}\frac{p^{\tau^{-1}\delta}\,(1-p)^k}{k},\text{ for } k\geq 1.\end{aligned}$$

4.4.2 The binomial gamma process

Consider the doubly stochastic process $N(t)$ which, conditional on the gamma process $\Lambda(t)$, is the counting process associated with a simple linear death process (as in section 4.3.2), with individual integrated death rate $\Lambda(t)$(again, see chapter III for further discussion), where $\Delta\Lambda(t)\sim\lambda\Gamma\left(\mu=\delta,\sigma^2=\tau\delta\right)$. Recall from section 4.3.2 that $N(t)$ (here conditional on $\Lambda(t)$) is the increasing process counting the dead individuals rather than the individuals still alive by time t. We distinguish the counting process $N(t)$ from the standard death process $N^*(t)$ by having a star on the latter. The derivation of the infinitesimal moments and the infinitesimal generator are essentially the same as in chapter III and are included here as well in order to make each chapter self-contained. Using the moments derived in section 4.3, the moments of the conditional increments of $N(t)$ can be derived as follows:

$$\begin{aligned}E[\Delta N(t)|N(t)=n] &= (n_0^*-n)(1-E[e^{-\Delta\Lambda(t)}|N(t)=n])\\ V[\Delta N(t)|N(t)=n] &= V[(n_0^*-n)(1-e^{-\Delta\Lambda(t)})|N(t)=n]+\\ &\quad +E[(n_0^*-n)(t)(e^{-\Delta\Lambda(t)}-e^{-2\Delta\Lambda(t)})|N(t)=n]\\ &= (n_0^*-n)^2V[e^{-\Delta\Lambda(t)}|N(t)=n]+\\ &\quad +(n_0^*-n)\big(E[e^{-\Delta\Lambda(t)}|N(t)=n]-E[e^{-2\Delta\Lambda(t)}|N(t)=n]\big)\end{aligned}$$

where n_0^* is the positive integer representing the individuals alive at time 0. Since $\Delta\Lambda(t) \sim \Gamma(\alpha = \tau^{-1}\delta, \beta = (\lambda\tau)^{-1})$, with moment generating function $E[e^{z\Delta\Lambda}] = (\frac{1}{1-z\lambda\tau})^{\tau^{-1}\delta}$ for $z\lambda\tau < 1$ and $\delta, \lambda, \tau > 0$. Then,

$$\begin{aligned}
E[e^{-\Delta\Lambda}] &= \Big(\frac{1}{1+\lambda\tau}\Big)^{\tau^{-1}\delta} \\
V[e^{-\Delta\Lambda}] &= E[e^{-2\Delta\Lambda}] - E[e^{-\Delta\Lambda}]^2 \\
E[e^{-2\Delta\Lambda}] &= \Big(\frac{1}{1+2\lambda\tau}\Big)^{\tau^{-1}\delta} \\
E[e^{-\Delta\Lambda}]^2 &= \Big(\frac{1}{1+\lambda\tau}\Big)^{2\tau^{-1}\delta}.
\end{aligned}$$

Let $f(\delta) = E[e^{-\Delta\Lambda}]$, then $f'(\delta) = -\tau^{-1}(1+\lambda\tau)^{-\tau^{-1}\delta}\ln(1+\lambda\tau)$ for $1+\lambda\tau > 0$. Using Taylor series expansion for $\delta_0 = 0$, $f(\delta) = f(0) + f'(0)\delta + o(\delta)$ and

$$\begin{aligned}
E[e^{-\Delta\Lambda}] &= 1 - \tau^{-1}\ln(1+\lambda\tau)\delta + o(\delta) \\
E[e^{-2\Delta\Lambda}] &= 1 - \tau^{-1}\ln(1+2\lambda\tau)\delta + o(\delta) \\
E[e^{-\Delta\Lambda}]^2 &= 1 - 2\tau^{-1}\ln(1+\lambda\tau)\delta + o(\delta) \\
V[e^{-\Delta\Lambda}] &= \tau^{-1}\delta(\ln((1+\lambda\tau)^2) - \ln(1+2\lambda\tau)) + o(\delta),
\end{aligned}$$

It follows that the moments of the increment process are

$$\begin{aligned}
E[\Delta N(t)|N(t)=n] &= (n_0^* - n)\tau^{-1}\delta\ln(1+\lambda\tau) + o(\delta) \\
V[\Delta N(t)|N(t)=n] &= V[(n_0^* - n)(1 - e^{-\Delta\Lambda(t)})|N(t)=n] + \\
&\quad + E[(n_0^* - n)(e^{-\Delta\Lambda(t)} - e^{-2\Delta\Lambda(t)})|N(t)=n] \\
&= (n_0^* - n)^2\tau^{-1}\delta\ln\bigg(\frac{(1+\lambda\tau)^2}{1+2\lambda\tau}\bigg) + \\
&\quad + (n_0^* - n)\tau^{-1}\delta\ln\bigg(\frac{1+2\lambda\tau}{1+\lambda\tau}\bigg) + o(\delta) \\
&= (n_0^* - n)\tau^{-1}\delta\ln(1+\lambda\tau) + \\
&\quad + (n_0^* - n)\tau^{-1}\delta\big((n_0^* - n) - 1\big)\ln\bigg(\frac{(1+\lambda\tau)^2}{1+2\lambda\tau}\bigg) + o(\delta)
\end{aligned}$$

Since $\frac{(1+\lambda\tau)^2}{1+2\lambda\tau} \geq 1$, it follows that the process is over-dispersed for $(n_0^* - n) > 1$ and equi-dispersed for $(n_0^* - n) = 1$. Even though the distribution of the increment process is not needed for computing the infinitesimal moments, it gives the infinitesimal generator, so we derive it below. For $k \in \{0, \dots, n_0^* - n\}$

$$\begin{aligned}
P(\Delta N = k|N(t) = n) &= \int_0^\infty \binom{n_0^* - n}{k} \left[1 - e^{-\lambda}\right]^k \left[e^{-\lambda}\right]^{(n_0^* - n) - k} \frac{\lambda^{\delta\alpha - 1} e^{-\lambda\beta} \beta^{\delta\alpha}}{\Gamma(\delta\alpha)} d\lambda \\
&= \binom{n_0^* - n}{k} \int_0^\infty \left[\sum_{j=0}^{k} \binom{k}{j} (-e^{-\lambda})^{k-j}\right] e^{-\lambda((n_0^* - n) - k)} \frac{\lambda^{\delta\alpha - 1} e^{-\lambda\beta} \beta^{\delta\alpha}}{\Gamma(\delta\alpha)} d\lambda \\
&= \binom{n_0^* - n}{k} \int_0^\infty \sum_{j=0}^{k} \binom{k}{j} (-1)^{k-j} e^{-\lambda((n_0^* - n) - j)} \frac{\lambda^{\delta\alpha - 1} e^{-\lambda\beta} \beta^{\delta\alpha}}{\Gamma(\delta\alpha)} d\lambda \\
&= \binom{n_0^* - n}{k} \sum_{j=0}^{k} \binom{k}{j} (-1)^{k-j} \frac{\beta^{\delta\alpha}}{(\beta + (n_0^* - n) - j)^{\delta\alpha}} \\
&\qquad \times \int_0^\infty \frac{\lambda^{\delta\alpha - 1} e^{-\lambda(\beta + (n_0^* - n) - j)} (\beta + (n_0^* - n) - j)^{\delta\alpha}}{\Gamma(\delta\alpha)} d\lambda \\
&= \binom{n_0^* - n}{k} \sum_{j=0}^{k} \binom{k}{j} (-1)^{k-j} \left(1 + \frac{(n_0^* - n) - j}{\beta}\right)^{-\delta\alpha} \\
&= \binom{n_0^* - n}{k} \sum_{j=0}^{k} \binom{k}{j} (-1)^{k-j} \left(1 - \delta\alpha \ln\left(1 + \frac{(n_0^* - n) - j}{\beta}\right) + o(\delta)\right) \\
&= 1_{\{k=0\}} + \delta \binom{n_0^* - n}{k} \sum_{j=0}^{k} \binom{k}{j} (-1)^{k-j+1} \alpha \ln\left(1 + \frac{(n_0^* - n) - j}{\beta}\right) + o(\delta)
\end{aligned}$$

4.4.3 The binomial beta process

Here we consider a slightly different approach for finding an over-dispersed simple linear death process. Instead of using a continuous time doubly stochastic process, we consider the limit of discrete time models.

Building on the counting process associated with the simple linear death process of section 4.3.2, $N(t)$, we now make the probability of death in the interval $[t, t + \delta]$, for an individual alive at time t, a random variable $\Pi(t)$. Since $\Delta N(t)|\Pi(t), N(t)$

follows a binomial distribution given by

$$P(\Delta N(t){=}k|N(t){=}n, \Pi(t){=}\pi(t)) = \binom{n_0^* - n}{k}(\pi(t))^k(1-\pi(t))^{(n_0^*-n)-k},$$

for $k \in \{0, \ldots, n_0^* - n\}$, and letting

$$\begin{aligned} \Pi(t)|N(t) = n \quad &\sim \quad Beta(\alpha = c(1-e^{-\mu\delta}), \beta = c(e^{-\mu\delta})) \\ c \quad &= \quad \begin{cases} \frac{(n_0^*-n)-1}{\omega} - 1 & \text{if } (n_0^* - n) > 1 \\ 1 & \text{if } (n_0^* - n) = 1 \end{cases} \end{aligned} \tag{4.6}$$

with $0 < \omega < (n_0^* - n) - 1$, it follows that $\Delta N(t)$ conditional on $N(t) = n$ has a beta binomial distribution with the corresponding parameters. This constraint in ω is necessary because both α and β of the beta distribution need to be positive. The value of c for $(n_0^* - n) = 1$ could be any other strictly positive real number since it does not affect the infinitesimal moments.

Note that $N(t)$ does not define a continuous time process anymore, since $\Pi(t)$ is not infinitely divisible. Nevertheless, it is still a useful process to consider since it is possible to derive the moments of $\Delta N(t)$ given $N(t) = n$ and an infinitesimal generator corresponding to (4.7) below, which does define a continuous time Markov counting process.

We now derive a continuous time Markov counting process defined by the limit of the transition probabilities of $\Delta N(t)$ conditional on $N(t) = n$. The beta binomial probability mass function of $\Delta N(t)$ given $N(t) = n$ is

$$P(\Delta N(t) = k|N(t) = n) =$$

$$= \binom{n_0^* - n}{k}\frac{\Gamma(\alpha+\beta)\Gamma(k+\alpha)\Gamma((n_0^*-n)-k+\beta)}{\Gamma(\alpha)\Gamma(\beta)\Gamma(\alpha+\beta+(n_0^*-n))} \tag{4.7}$$

$$= \binom{n_0^* - n}{k}\frac{\Gamma(\alpha+\beta)\Gamma(\alpha)\Gamma(\beta)\Gamma(k)\alpha\{\frac{\Gamma(c+(n_0^*-n)-k)}{\Gamma(c)} + O(\delta)\}}{\Gamma(\alpha+\beta)\Gamma(\alpha)\Gamma(\beta)\frac{\Gamma(c+(n_0^*-n))}{\Gamma(c)}} \tag{4.8}$$

$$= \binom{n_0^* - n}{k}\frac{\Gamma(k)\Gamma(c+(n_0^*-n)-k)}{\Gamma(c+(n_0^*-n))}c\mu\delta + o(\delta), \tag{4.9}$$

for $k \in \{0, \ldots, n_0^* - n\}$. Equation (4.8) follows from (4.7) via an application of lemma IV.3 at the end of this section. Specifically, using lemma IV.3 with $i = (n_0^* - n) - k$, it follows that

$$\Gamma((n_0^* - n) - k + \beta) \;=\; \{\frac{\Gamma(c + (n_0^* - n) - k)}{\Gamma(c)} + O(\delta)\}\Gamma(\beta), \tag{4.10}$$

and, since $\alpha + \beta = c$,

$$\begin{aligned} \Gamma(\alpha + \beta + (n_0^* - n)) &= \Gamma(c + (n_0^* - n)) \\ &= \frac{\Gamma(c + (n_0^* - n))}{\Gamma(c)}\Gamma(c) \\ &= \frac{\Gamma(c + (n_0^* - n))}{\Gamma(c)}\Gamma(\alpha + \beta). \end{aligned} \tag{4.11}$$

Plugging (4.10) and (4.11) into (4.7) gives (4.8). Then, using $\alpha = c\mu\delta + o(\delta)$ and canceling terms gives (4.9), which corresponds to the infinitesimal generator

$$\begin{aligned} q_{i,i+k} &= \binom{n_0^* - i}{k}\frac{\Gamma(k)\Gamma(c + (n_0^* - i) - k)}{\Gamma(c + (n_0^* - i))}c\mu \\ q_{i,j} &= 0 \text{ for } j > n_0^* - i, \end{aligned}$$

for $k \in \{1, \ldots, n_0^* - i\}$. We call the process defined by this infinitesimal generator, say $\tilde{N}(t)$, the binomial beta process. Its infinitesimal moments can be derived based on the moments of the discrete time process $N(t)$, defined at the beginning of this section. The moments of a beta binomial distribution are a standard result which gives

$$\begin{aligned} E[\Delta N(t)|N(t) = n] &= (n_0^* - n)\frac{\alpha}{\alpha + \beta} \\ &= (n_0^* - n)\mu\delta + o(\delta) \\ V[\Delta N(t)|N(t) = n] &= (n_0^* - n)\frac{\alpha\beta((n_0^* - n) + \alpha + \beta)}{(\alpha + \beta)^2(1 + \alpha + \beta)} \\ &= (n_0^* - n)(1 + \omega)\mu\delta + o(\delta). \end{aligned}$$

Since the binomial beta process has a finite number of states, it follows that the increment moments of the binomial beta process $\tilde{N}(t)$ are

$$\begin{aligned}
E[\Delta\tilde{N}(t)|\tilde{N}(t) = \tilde{n}] &= \sum_{k=0}^{n_0^*-\tilde{n}} kP(\Delta\tilde{N}(t)|\tilde{N}(t) = \tilde{n}) \\
&= \sum_{k=0}^{n_0^*-\tilde{n}} kP(\Delta N(t)|N(t) = \tilde{n}) + ko(\delta) \\
&= (n_0^* - \tilde{n})\mu\delta + o(\delta) \\
V[\Delta\tilde{N}(t)|\tilde{N}(t) = \tilde{n}] &= \sum_{k=0}^{n_0^*-\tilde{n}} k^2P(\Delta\tilde{N}(t)|\tilde{N}(t) = \tilde{n}) - ((n_0^* - \tilde{n})\mu\delta + o(\delta))^2 \\
&= \sum_{k=0}^{n_0^*-\tilde{n}} k^2P(\Delta N(t)|N(t) = \tilde{n}) + k^2o(\delta) + o(\delta) \\
&= (n_0^* - \tilde{n})(1+\omega)\mu\delta + o(\delta),
\end{aligned}$$

and it follows that the binomial beta process $\tilde{N}(t)$ is over-dispersed for $\omega > 0$ and $(n_0^* - \tilde{n}) > 1$. If $(n_0^* - \tilde{n}) = 1$, then using

$$\begin{aligned}
V[\Delta N(t)|N(t) = n] &= (n_0^* - n)\frac{\alpha\beta}{(\alpha+\beta)^2} \\
&= (n_0^* - n)\mu\delta + o(\delta)
\end{aligned}$$

as above, it follows that the process $\tilde{N}(t)$ is equi-dispersed, just like in the binomial gamma process. ω can be used to obtain a specific infinitesimal variance while with the binomial gamma it would be necessary to solve a nonlinear system of equations. In practice, these equations are not easy to solve, so one does not parameterize a binomial gamma process by the infinitesimal moments. Also note that the constraint $0 < \omega < (n_0^* - n) - 1$ gives a clear bound on the over-dispersion that is possible for a given population of size $(n_0^* - n)$. In the binomial gamma process it is not even obvious that such a constrain exists nor the exact bound of the over-dispersion. These are some advantages of the binomial beta process over the binomial gamma process.

Lemma IV.3. *For* $\alpha = c(1 - e^{-\mu\delta})$, $\beta = ce^{-\mu\delta}$, *as defined in (4.6),* $c > 0$ *and* $i \in \{1, 2, \dots\}$,

$$\Gamma(\beta + i) \quad = \quad \left\{ \frac{\Gamma(c+i)}{\Gamma(c)} + O(\delta) \right\} \Gamma(\beta).$$

Proof. Since $\beta = c - \alpha$, and by the definition of the gamma function, for $i \geq 1$,

$$\begin{aligned}
\Gamma(\beta + i) &= (c - \alpha + (i-1)) \times (c - \alpha + (i-2)) \times \cdots \times (c - \alpha) \times \Gamma(\beta) \\
&= \{(c + (i-1)) \times (c + (i-2)) \times \cdots \times (c) + O(\delta)\} \Gamma(\beta) \\
&= \left\{ \prod_{j=0}^{i-1} (c + j) + O(\delta) \right\} \Gamma(\beta) \\
&= \left\{ \frac{\Gamma(c+i)}{\Gamma(c)} + O(\delta) \right\} \Gamma(\beta).
\end{aligned}$$

□

Bibliography

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

Anderson, R. M. and May, R. M. (1991). *Infectious Diseases of Humans.* Oxford University Press, Oxford.

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

Bailey, N. T. (1964). *The elements of Stochastic Processes with applications to the natural sciences.* John Wiley & sons, Inc.

Barndorff-Nielsen, O. E. and Cox, D. R. (1994). *Inference and Asymptotics.* Chapman and Hall, London.

Bartlett, M. S. (1956). Deterministic and stochastic models for recurrent epidemics. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, Volume 4*, pages 81–109.

Bartlett, M. S. (1960). *Stochastic Population Models in Ecology and Epidemiology.* Wiley, New York.

Basawa, I. V. and Prakasa Rao, B. L. S. (1980). *Statistical Inference for Stochastic Processes*. Academic Press, New York.

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

Bharucha-Reid, A. T. (1960). *Elements of the Theory of Markov Processes and their Applications*. McGraw-Hill.

Bjornstad, O. N. and Grenfell, B. T. (2001). Noisy clockwork: Time series analysis of population fluctuations in animals. *Science*, 293:638–643.

Breiman, L. (2001). Statistical modeling: The two cultures. *Statistical Science*, 16(3):199–215.

Brémaud, P. (1999). *Markov Chains: Gibbs Fields, Monte Carlo Simulation, and Queues*. Springer, New York.

Brown, T. C., Hamza, K., and Xia, A. (1998). On the variance to mean ratio for random variables from markov chains and point processes. *Journal of Applied Probability*, 35(2):303–312.

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

Casella, G. and Berger, R. L. (1990). *Statistical Inference*. Wadsworth, Pacific Grove.

Clark, J. S. and Bjornstad, O. N. (2004). Population time series: Process variability, observation errors, missing values, lags, and hidden states. *Ecology*, 85:3140–3150.

Cleveland, W. S., Grosse, E., and Shyu, W. M. (1993). Local regression models. In Chambers, J. M. and Hastie, T. J., editors, *Statistical models in S*, pages 309–376. Chapman and Hall, London.

Coulson, T., Rohani, P., and Pascual, M. (2004). Skeletons, noise and population growth: the end of an old debate? *Trends in Ecology and Evolution*, 19:359–364.

Cramér, H. (1946). *Mathematical Methods of Statistics*. Princeton University Press, Princeton.

Cutler, D. (2000). Understanding the overdispersed molecular clock. *Genetics*, 154:1403–1417.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Ser. B*, 39:1–22.

Doucet, A., Freitas, N., and Gordon, N. (2001). An introduction to sequential monte carlo methods. In Doucet, A., de Freitas, N., and Gordon, N. J., editors, *Sequential Monte Carlo Methods in Practice*, pages 3–14. Springer, New York.

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

Ellner, S. P., Seifu, Y., and Smith, R. H. (2002). Fitting population dynamic models to time-series data by gradient matching. *Ecology*, 83:2256–2270.

Ferguson, N., Anderson, R., and Gupta, S. (1999). The effect of antibody-dependent enhancement on the transmission dynamics and persitence of multiple-strain pathogens. *Proceedings of the National Academy of Sciences of the USA*, 96:187–205.

Ferguson, N. M., Galvani, A. P., and Bush, R. M. (2003). Ecological and immunological determinants of influenza evolution. *Nature*, 422:428–433.

Fernandez-Villaverde, J. and Rubio-Ramirez, J. F. (2005). Estimating dynamic equilibrium economies: Linear versus nonlinear likelihood. *Journal of Applied Econometrics*, 20:891–910.

Finkenstädt, B. F. and Grenfell, B. T. (2000). Time series modelling of childhood diseases: A dynamical systems approach. *Applied Statistics*, 49:187–205.

Gillespie, D. T. (1977). Exact stochastic simulation of coupled chemical reactions. *Journal of Physical Chemistry*, 81:2340–2361.

Gillespie, D. T. (2001). Approximate accelerated stochastic simulation of chemically reacting systems. *Journal of Chemical Physics*, 115:1716–1733.

Gillespie, J. (1984). The molecular clock may be an episodic clock. *Proc. Natl. Acad. Sci. USA*, 81 No.24:8009–8013.

Glass, K., Xia, Y., and Grenfell, B. (2003). Interpreting time-series analyses for continuous-time biological models–measles as a case study. *Journal of Theoretical Biology*, 223:19–25.

Gordon, N., Salmon, D. J., and Smith, A. F. M. (1993). Novel approach to nolinear/non-gaussian bayesian state estimation. *IEE Proceedings-F*, 140:107–113.

Grenfell, B. T., Pybus, O. G., Gog, J. R., Wood, J. L. N., Daly, J. M., Mumford, J. A., and Holmes, E. C. (2004). Unifying the epidemiological and evolutionary dynamics of pathogens. *Science*, 303:327–332.

Gupta, S., Trenholme, K., Anderson, R., and Day, K. (1994). Antigenic diversity and the transmission dynamics of plasmodium falciparum. *Science*, 263:961–963.

Houtekamer, P. L. and Mitchell, H. L. (2001). Data assimilation using an ensemble Kalman filter technique. *Monthly Weather Review*, 129:123–137.

Huq, A., West, P. A., Small, E. B., Huq, M. I., and Colwell, R. R. (1984). Influence of water temperature, salinity, and pH on survival and growth of toxigenic *vibrio cholerae* serovar o1 associated with live copepods in laboratory microcosms. *Appl. Environ. Microbiol.*, 48:420–424.

Hürzeler, M. and Künsch, H. R. (2001). Approximating and maximising the likelihood for a general state-space model. In Doucet, A., de Freitas, N., and Gordon, N. J., editors, *Sequential Monte Carlo Methods in Practice*, pages 159–175. Springer, New York.

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

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

Ionides, E. L., Fang, K. S., Isseroff, R. R., and Oster, G. F. (2004). Stochastic models for cell motion and taxis. *Journal of Mathematical Biology*, 48:23–37.

Jacquez, J. A. (1996). *Comparmental Ananlysis in Biology and Medicine*. 3rd edition, BioMedware, Ann Arbor, MI.

Jensen, J. L. and Petersen, N. V. (1999). Asymptotic normality of the maximum likelihood estimator in state space models. *Annals of Statistics*, 27:514–535.

Kamo, M. and Sasaki, A. (2002). The effect of cross-immunity and seasonal forcing in a multi-strain epidemic model. *Physica D*, 165:228–241.

Karlin, S. and Taylor, H. M. (1981). *A Second Course in Stochastic Processes*, volume 1. Academic Press.

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

Kermack, W. O. and McKendrick, A. G. (1927). A contribution to the mathematical theory of epidemics. *Proceedings of the Royal Society of London, Ser. A*, 115:700–721.

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

King, A. A., Ionides, E. L., Pascual, M., and Bouma, M. J. (2007). Rapid loss of immunity is necessary to explain historical cholera epidemics. *submitted*.

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

Kloeden, P. E. and Platen, E. (1999). *Numerical Soluion of Stochastic Differential Equations*. Springer, New York, 3rd edition.

Koelle, K., Cobey, S., Grenfell, B., and Pascual, M. (2006a). Epochal evolution shapes the philodynamics of interpandemic influenza a (h5n2) in humans. *Science*, 314:1898–1903.

Koelle, K. and Pascual, M. (2004). Disentangling extrinsic from intrinsic factors in disease dynamics: a nonlinear time series approach with an application to cholera. *The American Naturalist*, 163:901–913.

Koelle, K., Pascual, M., and Yunus, M. (2006b). Serotype cycles in cholera dynamics. *Proceedings of the Royal Society of London B*, 273:2876–2889.

Koelle, K., Rodo, X., Pascual, M., Yunus, M., and Mostafa, G. (2005). Refractory periods and climate forcing in cholera dynamics. *Nature*, 436:696–700.

Kou, S. C., Xie, S., and Liu, J. S. (2005). Bayesian analysis of single-molecule experimental data. *Applied Statistics*, 54:469–506.

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

Liu, J. S. (2001). *Monte Carlo Strategies in Scientific Computing.* Springer, New York.

Matis, J. H. and Kiffe, T. R. (2000). *Stochastic Population Models. A Compartmental Perpective.* Springer.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models.* Chapman and Hall, London, 2nd edition.

McLachlan, G. and Peel, D. (2000). *Finite Mixture Models.* Wiley, New York.

Newman, K. B. and Lindley, S. T. (2006). Accounting for demographic and environmental stochasticity, observation error and parameter uncertainty in fish population dynamic models. *North American Journal of Fisheries Management*, 26:685–701.

Øksendal, B. (1998). *Stochastic Differential Equations.* Springer, New York, 5th edition.

Pascual, M., Bouma, M. J., and Dobson, A. P. (2002). Cholera and climate: revisiting the quantitative evidence. *Microbes Infect.*, 4:237–245.

Pascual, M., Rodó, X., Ellner, S. P., Colwell, R., and Bouma, M. J. (2000). Cholera dynamics and el niño–southern oscillation. *Science*, 289:1766–1769.

Powell, M. J. D. (1981). *Approximation Theory and Methods.* Cambridge University Press, Cambridge.

Press, W., Flannery, B., Teukolsky, S., and Vetterling, W. (2002). *Numerical Recipes in C++.* Cambridge University Press, Cambridge, 2nd edition.

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, Ser. B*, to appear.

Renshaw, E. (1991). *Modelling Biological Populations in Space and Time.* Cambridge University Press.

Roberts, G. O. and Stramer, O. (2001). On inference for partially observed nonlinear diffusion models using the Metropolis-Hastings algorithm. *Biometrika*, 88:603–621.

Rodó, X., Pascual, M., Fuchs, G., and Faruque, A. S. G. (2002). ENSO and cholera: A nonstationary link related to climate change? *Proceedings of the National Academy of Sciences of the USA*, 99:12901–12906.

Sack, D. A., Sack, R. B., Nair, G. B., and Siddique, A. K. (2004). Cholera. *The Lancet*, 363:223–233.

Sato, K. (1999). *Levy Processes and Inifinitely Divisible Distributions.* Cambridge University Press.

Shephard, N. and Pitt, M. K. (1997). Likelihood analysis of non-Gaussian measurement time series. *Biometrika*, 84:653–667.

Shumway, R. H. and Stoffer, D. S. (2000). *Time Series Analysis and Its Applications.* Springer, New York.

Snyder, D. and Miller, M. (1991). *Randon Point Processes in Time and Space.* Springer-Verlag.

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

Swishchuk, A. and Wu, J. (2003). *Evolution of Biological Systems in Random Media: Limit Theorems and Stability.* Kluwer Acamedic Publishers.

Takahata, N. (1987). On the overdispersed molecular clock. *Genetics*, 116:169–179.

Thomas, L., Buckland, S. T., Newman, K. B., and Harwood, J. (2005). A unified framework for modelling wildlife population dynamics. *Australian & New Zealand Journal of Statistics*, 47:19–34.

Thomson, M. C., Doblas-Reyes, F. J., Mason, S. J., Hagedorn, S. J., Phindela, T., Morse, A. P., and Palmer, T. N. (2006). Malaria early warnings based on seasonal climate forecasts from multi-model ensembles. *Nature*, 439:576–579.

Tian, T. and Burrage, K. (2004). Binomial leap methods for simulating stochastic chemical kinetics. *Journal of Chemical Physics*, 121:10356–10364.

Turchin, P. (2003). *Complex Population Dynamics. A Theoretical/Empirical Synthesis.* Princeton University Press.

Wilson, S. and Costello, M. (2005). Predicting future discoveries of european marine species by using a non-homogeneous renewal process. *Journal of the Royal Statistical Society Series C*, 54, Part.5:897–918.

Wolpert, R. and Ickstadt, K. (1998). Poisson/gamma random field models for spatial statistics. *Biometrika*, 85,2:251–267.

Wu, C. F. J. (1985). Efficient sequential designs with binary data. *Journal of the American Statistical Association*, 80:974–984.

Xiu, D., Kevrekidis, I. G., and Ghanem, R. (2005). An equation-free, multiscale approach to uncertainty quantification. *Computing in Science and Eng.*, 7(3):16–23.

Zo, Y.-G., Rivera, I. N. G., Russek-Cohen, E., Islam, M. S., Siddique, A. K., Yunus, M., Sack, R. B., Huq, A., and Colwell, R. R. (2002). Genomic profiles of clinical and environmental isolates of vibrio cholerae o1 in cholera-endemic areas of bangladesh. *Proceedings of the National Academy of Sciences of the USA*, 99(19):12409–12414.