INFEERENCE WITH IMPLICIT LIKELIHOODS FOR INFECTIOUS DISEASE MODELS

A Dissertation in Statistics
by
Roman A. Jandarov

© 2012 Roman A. Jandarov

Submitted in Partial Fulfillment of the Requirements for the Degree of Doctor of Philosophy

August 2012
The dissertation of Roman A. Jandarov was reviewed and approved* by the following:

Murali Haran  
Associate Professor of Statistics  
Dissertation Advisor, Chair of Committee

Francesca Chiaromonte  
Professor of Statistics and Public Health Sciences

David Hunter  
Professor of Statistics

Ottar Bjørnstad  
Professor of Entomology, Biology, and Statistics

David Hunter  
Professor of Statistics  
Head of the Department of Statistics

*Signatures are on file in the Graduate School.
Abstract

Probabilistic models for infectious diseases are important for understanding mechanisms underly- ing the spread of infection. I develop new models and computational approaches motivated by sev- eral research projects including a study of the dynamics of meningitis in Nigeria, measles infection in England and Wales, and gypsy moth infestations in Pennsylvania.

Likelihood functions for infectious disease models are often expensive to evaluate, making tradi- tional likelihood-based inference computationally infeasible. Furthermore, traditional inference may lead to poor parameter estimates and the fitted model may not capture important biological characteristics of the observed data. I explore efficient inferential methods based on so-called approximate Bayesian computation (ABC). I develop a new model for meningitis dynamics in Nigeria and show how a version of ABC is effective when it is relatively inexpensive to simulate from the model.

When the likelihood function is expensive to evaluate and simulations are time consum- ing ABC-based inference is infeasible. For such models I propose a novel approach that is inspired by recent work in emulation and calibration for complex computer models. My motivating example is the gravity time series susceptible-infectious-recovered (TSIR) model. My approach is based on obtaining a Gaussian process approximation to the model using key summary statistics calculated from model simulations. Unlike traditional likelihood-based inference, the new approach is computationally expedient, provides accurate parameter in- ference, and results in a good model fit. I apply my approach to the analysis of measles outbreaks in England and Wales. In general, my approach is applicable to many problems where traditional likelihood-based inference is computationally intractable or leads to poor inference. I demonstrate how this methodology can be used to learn about the parameters of a complex mixture random graph model.

The gypsy moth is the most important forest defoliating insect in the northeastern United States. Inferring gypsy moth population periodicities is challenging because population sizes
of the insect are not directly observable. I develop a new space-time Gaussian process model for inferring gypsy moth populations based on indirect information from defoliation data. I demonstrate via simulated examples that this approach provides accurate population periodicity estimates; I apply this methodology to a Pennsylvania defoliation data set.
# Table of Contents

List of Figures v

Preface xii

Acknowledgments xiii

Chapter 1

Introduction 1

1.1 Models for Infectious Diseases .......................... 1
1.2 Types of Mathematical Models .......................... 3
1.3 Statistical Inference for Infectious Disease Models ............. 11
1.4 Summary of Contributions ............................... 19
1.5 Dissertation Organization ............................... 20

Chapter 2

Approximate Bayesian Computation (ABC) and Inference for a Model
for Meningitis Dynamics 21

2.1 ABC Algorithms: Background ............................ 22
2.2 Outline of the Algorithms ............................... 23
2.3 Modeling and ABC-based Inference for Meningitis Dynamics ........ 25
   2.3.1 Background on Meningitis ............................. 26
   2.3.2 A Model for Meningitis Dynamics .................... 29
   2.3.3 Inference ........................................ 31
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.5.1</td>
<td>Computational Details of Gridded MCMC and Emulation-based Approaches</td>
<td>88</td>
</tr>
<tr>
<td>4.5.2</td>
<td>Application to Simulated Data</td>
<td>89</td>
</tr>
<tr>
<td>4.6</td>
<td>Results from Application to Measles Data</td>
<td>92</td>
</tr>
<tr>
<td>4.6.1</td>
<td>Description of Measles Data</td>
<td>92</td>
</tr>
<tr>
<td>4.6.2</td>
<td>Some Implications for Measles Dynamics</td>
<td>92</td>
</tr>
<tr>
<td>4.7</td>
<td>Discussion</td>
<td>93</td>
</tr>
</tbody>
</table>

Chapter 5
Emulator-Based Inference: Generalizations and Extensions 97
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.1</td>
<td>Emulator-Based Inference: General Overview</td>
<td>97</td>
</tr>
<tr>
<td>5.2</td>
<td>Simulated Examples</td>
<td>98</td>
</tr>
<tr>
<td>5.3</td>
<td>Indirect Inference</td>
<td>102</td>
</tr>
<tr>
<td>5.4</td>
<td>Other Approaches to Inference with Complicated Likelihoods</td>
<td>103</td>
</tr>
<tr>
<td>5.5</td>
<td>Inference for a Mixture Random Graphs Model</td>
<td>106</td>
</tr>
<tr>
<td>5.5.1</td>
<td>Background</td>
<td>106</td>
</tr>
<tr>
<td>5.5.2</td>
<td>Mixture Graphs Models</td>
<td>108</td>
</tr>
<tr>
<td>5.5.3</td>
<td>An Example: Affiliation Network</td>
<td>108</td>
</tr>
<tr>
<td>5.5.4</td>
<td>Emulator-Based Inference for Affiliation Networks</td>
<td>110</td>
</tr>
</tbody>
</table>

Chapter 6
Discussion and Future Work 115
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>6.1</td>
<td>Summary and Contributions</td>
<td>115</td>
</tr>
<tr>
<td>6.2</td>
<td>Caveats and Avenues for Future Research</td>
<td>116</td>
</tr>
<tr>
<td>6.3</td>
<td>Potential Avenues for Theoretical Research</td>
<td>118</td>
</tr>
</tbody>
</table>

Appendix A
Likelihood Function for the Gravity Model 121

Appendix B
Details of Constructing an Emulator 123

Bibliography 125
List of Figures

1.1 Movement of the host between the compartments ........................................ 4
1.2 Incident and prevalent cases of the 2001 foot-and-mouth epidemic in Cumbria (plot from Diggle (2006)). Here, incident cases within the time interval between successive frames are shown in black. Prevalent cases on the dates indicated are shown in grey. ......................................................... 8
1.3 Bi-weekly reported cases of measles for different cities (from the UK Registrar Generals data for 954 cities in England and Wales for years 1944-1966). .......................... 13

2.1 Inference for $\left(\beta_1, \beta_2, \gamma_1, \gamma_2\right)$: (a) Posterior density for $\beta_1$ when the true value of $\beta_1 = 0.4$; (b) Posterior density for $\beta_2$ when the true value of $\beta_2 = 0.4$; (c) Sample histogram for $\gamma_1$ when the true value of $\gamma_1 = 6$; (d) Sample histogram for $\gamma_2$ when the true value of $\gamma_2 = 13$. ......................................................... 36
2.2 Inference for $\left(\alpha_1, \alpha_2, \gamma_1, \gamma_2\right)$ when the incidence independent environmental effects can spark outbreaks: (a) Posterior density for $\alpha_1$ when the true value of $\alpha_1 = 0.48$; (b) Posterior density for $\alpha_2$ when the true value of $\alpha_2 = 0.45$; (c) Sample histogram for $\alpha_1$ when the true value of $\alpha_1 = 5$; (d) Sample histogram for $\alpha_2$ when the true value of $\alpha_2 = 15$. ......................................................... 36
2.3 Inference for $\left(\gamma_1, \gamma_2, \gamma_1, \gamma_2\right)$ when the transmission of the infection cannot spark outbreaks: (a) Posterior density for $\gamma_1$ when the true value of $\gamma_1 = 0.1$; (b) Posterior density for $\gamma_2$ when the true value of $\gamma_2 = 0.1$; (c) Sample histogram for $\gamma_1$ when the true value of $\gamma_1 = 2$; (d) Sample histogram for $\gamma_2$ when the true value of $\gamma_2 = 13$. ......................................................... 37
2.4 Inference for $\left(\gamma_1, \gamma_2, \gamma_1, \gamma_2\right)$ when the transmission of the infection can spark outbreaks: (a) Posterior density for $\gamma_1$ when the true value of $\gamma_1 = 1.5$; (b) Posterior density for $\gamma_2$ when the true value of $\gamma_2 = 1.5$; (c) Sample histogram for $\gamma_1$ when the true value of $\gamma_1 = 2$; (d) Sample histogram for $\gamma_2$ when the true value of $\gamma_2 = 13$. ......................................................... 38
2.5 Inference for \((\alpha_1, \alpha_2, \tau_1^\alpha, \tau_2^\alpha)\) when the climate effects cannot spark outbreaks: 
(a) Posterior density for \(\alpha_1\) when the true value of \(\alpha_1 = 1\); (b) Posterior density for \(\alpha_2\) when the true value of \(\alpha_2 = 1.5\); (c) Sample histogram for \(\tau_1^\alpha\) when the true value of \(\tau_1^\alpha = 5\); (d) Sample histogram for \(\tau_2^\alpha\) when the true value of \(\tau_2^\alpha = 15\). .................................................. 39

2.6 Simulated data example. Solid curves: simulated data; dotted curves: predictions; dashed curves: 95% prediction regions .................................................. 40

2.7 Inference for \((\beta_1, \beta_2, \tau_1^\beta, \tau_2^\beta)\) for the actual data: (a) Posterior density for \(\beta_1\); (b) Posterior density for \(\beta_2\); (c) Sample histogram for \(\tau_1^\beta\); (d) Sample histogram for \(\tau_2^\beta\). .................................................. 42

2.8 Inference for \((\alpha_1, \alpha_2, \tau_1^\alpha, \tau_2^\alpha)\) for the actual data: (a) Posterior density for \(\alpha_1\); (b) Posterior density for \(\alpha_2\); (c) Sample histogram for \(\tau_1^\alpha\); (d) Sample histogram for \(\tau_2^\alpha\). .................................................. 43

2.9 Inference for \(p_{\text{MSF}}\) and \(p_{\text{WHO}}\): (a) Posterior density for \(p_{\text{MSF}}\); (b) Posterior density for \(p_{\text{WHO}}\). .................................................. 46

3.1 Total number of forest defoliations aggregated over time in 2D ......... 59

3.2 Total number of forest defoliations aggregated over time in 3D ......... 60

3.3 Division of the data into separate blocks based on their closeness. ...... 62

3.4 Division of the data into separate blocks using different grids. ......... 63

3.5 Density Plots of Periodicities for Different Locations (LGP), along with estimates obtained with BD and PK: (a) All methods recover the truth; (b) LGP recovers the truth with some uncertainty; (c) LGP fails to recover the truth; (d) LGP outperforms BD and PK; (e) BD outperforms LGP; (f) PK outperforms LGP. .................................................. 65

3.6 Density plots of periodicities for different locations inferred from the defoliation data: (a) Density has one big and one small peaks; (b) Density is unimodal with almost all its mass at one point; (c) Density has two relatively steep peaks; (d) Density shows a lot of uncertainty; (e) Density has three modes; (f) Density is bimodal with some uncertainty. .................. 68

3.7 Histograms of periodicities by forest types obtained from posterior samples: (a) Oak-pine forests; (b) Oak-hickory forests; (c) Maple-beech-birch forests; (d) Non-forests (trees in parks close to populated areas). .................. 69
4.1 Inferred posterior 2D likelihood surface obtained for data with known parameters ($\theta' = 0.71$, $\tau_1 = 0.3$, $\tau_2 = 0.7$ and $\rho = 1$): (a) Marginal 2D likelihood surface for ($\theta'$, $\rho$); (b) Marginal 2D likelihood surface for ($\theta'$, $\rho$) assuming $\tau_1 = 0.3$ (true); (c) 2D likelihood surface for ($\theta'$, $\rho$) assuming $\tau_1 = 0.3$ (true) and $\tau_2 = 0.7$ (true); (d) 2D likelihood surface for ($\theta'$, $\rho$) assuming $\tau_1 = 0.5$ (any value) and $\tau_2 = 1$ (any value).

4.2 Inferred posterior 2D likelihood surface obtained for data with known parameters ($\theta' = 0.71$, $\tau_1 = 0.3$, $\tau_2 = 0.7$ and $\rho = 1$): (a) Marginal 2D likelihood surface for ($\tau_2$, $\rho$); (b) Marginal 2D likelihood surface for ($\tau_2$, $\rho$) assuming $\theta' = 0.71$ (true); (c) 2D likelihood surface for ($\tau_2$, $\rho$) assuming $\theta' = 0.71$ (true) and $\tau_1 = 0.3$ (true); (d) 2D likelihood surface for ($\tau_2$, $\rho$) assuming $\theta' = 1$ (any value) and $\tau_1 = 0.3$ (any value).

4.3 Inferred posterior 2D likelihood surface obtained for data with known parameters ($\theta' = 0.71$, $\tau_1 = 0.5$, $\tau_2 = 1$ and $\rho = 1$): Posterior 2D likelihood surface for ($\theta'$, $\rho$) assuming $\tau_1 = 0.5$ (true) and $\tau_2 = 1$ (true) has a shift and does not contain the true ($\theta'$, $\rho$) at its highest probability area.

4.4 Characteristics of simulated data at the parameters obtained via the traditional Bayes approach: (a) Simulated $\mathbf{M}$ from the data; (b) Simulated $\mathbf{P}$ from the data.

4.5 Approximation of a nonlinear function using a Gaussian process. Our “data” are denoted by black dots, the green curves are the independent error process without spatial dependence, red curves are linear Gaussian process model with exponential covariance, blue curves are with squared exponential covariance.

4.6 Approximation of nonlinear functions using a Gaussian process. The black dots represent ‘data’ from functions $f(x) = 2\sin(x)$ and $5\exp(-x/5)\sin(x)$. Both are fit with the same stochastic model: $f(x) = c + \epsilon(x)$ where $\epsilon(x)$ is a linear Gaussian process. Solid blue lines are predictions at new locations and the dotted blue lines reflect uncertainty.

4.7 95% C.I.’s for ($\theta'$, $\rho$) obtained via different methods (assuming that $\tau_1$ and $\tau_2$ are known): Solid line shows the 95% region obtained using the traditional Bayes method. Dashed line outlines the 95% region obtained via GP emulator: (a) Both regions contain the true parameter values; (b) Region obtained by the GP emulator contains the true values of the parameters, while the traditional Bayes region does not.

4.8 Characteristics of simulated data at the parameters chosen to minimize the discrepancy between the data and the simulation: (a) Simulated $\mathbf{M}$ from the data; (b) Simulated $\mathbf{P}$ from the data.
4.9 95% C.I. for \((\theta', \rho)\) obtained via fitting GP emulator to a part of the data: Solid line outlines the confidence region for parameters when data for years from 1944 -55 is used; Dashed line outlines the confidence region for parameters when data for years from 1956 -67 is used. ........................................ 94

4.10 95% C.I. for \((\theta', \rho)\) obtained via fitting GP emulator to a part of the data: Solid line outlines the confidence region for parameters when data only holidays periods is used; Dashed line outlines the confidence region for parameters when data for only non-holiday periods is used. ........................................ 95

5.1 True likelihood surface of the data. .................................................. 99

5.2 Inference based on Algorithm 1 with different tolerance levels \(\epsilon\): (a) \(\epsilon = 60\); (b) \(\epsilon = 70\); (c) \(\epsilon = 80\); (d) \(\epsilon = 90\). ........................................ 100

5.3 Inference based on modified Algorithm 2 with fixed tolerance levels \(\epsilon = 80\) and different values of \(R\): (a) \(R = 10\); (b) \(R = 50\); (c) \(R = 100\); (d) \(R = 500\). 100

5.4 Inference based on GP emulator: (a) (a) \(g = 5\) and \(r = 1\); (b) \(g = 10\) and \(r = 1\); (c) \(g = 5\) and \(r = 5\); (d) \(g = 10\) and \(r = 5\). ........................................ 101

5.5 A graph simulated from a mixture model ........................................ 109

5.6 Affiliation networks for different \(\xi\) ........................................ 110

5.7 A grid for \(\pi_1, \ldots, \pi_6\) ........................................ 112

5.8 Inference for identifiable parameters ........................................ 113

5.9 Flat posterior surface for \(\pi_5\) ........................................ 114
Preface

This dissertation includes the work of several multi-author papers for which input came from several people. In these papers, I am the first author. In the following, I will detail my contributions for each of these papers.

Chapter 2 of this dissertation is partially based on a paper that has been accepted for publication in the *Journal of Agricultural, Biological, and Environmental Statistics*: “A compartmental model for meningitis: separating transmission from climate effects on disease incidence”, by Jandarov, Haran, and Ferrari. My contributions in this work include designing details of an SIR model to describe the dynamics of meningitis, as well as implementation of an approximate Bayesian inference approach. I was also responsible for most of the writing of the manuscript.

Chapter 3 is partially based on our working paper titled “A latent Gaussian process model for inferring periodicities of gypsy moth populations” by Jandarov, Haran, and Bjørnstad. My contributions in this work include the development of the underlying latent model, implementation of the inference algorithm, and writing of the manuscript.

Chapter 4 is partially based on our working paper titled “Emulating a gravity model to infer the spatiotemporal dynamics of an infectious disease” by Jandarov, Haran, Bjørnstad, and Grenfell. My contributions here include the development of details and implementation of our inferential approach to study the dynamics of measles.

Finally, Chapter 5 is partially based on our working paper titled “Fast approximate inference for mixture random graph models” by Jandarov and Haran. Here, I develop a version of our approximate inferential approach in the context of a random graph model.
Acknowledgments

After all these years at Penn State, I have a list of people who contributed in some way to this dissertation, for which I would like to express my sincere gratitude. Without these people, this dissertation would not have been possible.

First, I want to thank my advisor, Dr. Murali Haran. It is extremely difficult to overstate my gratitude to him for his continuous support and guidance of my doctoral research, for his patience, motivation, enthusiasm, and immense knowledge of spatial and computational statistics. Murali has always been very understanding and has provided encouragement, sound advice, good teaching and a lot of good ideas. I can easily say that Murali was the greatest advisor I have ever had.

Besides my advisor, I would like to thank the rest of my dissertation committee and our helpful collaborators: Professors Ottar Bjørnstad, Francesca Chiaromonte, David Hunter and Matthew Ferrari. Professors Francesca Chiaromonte and David Hunter have always helped me with their useful and insightful comments, their encouragement and hard questions. Professors Ottar Bjørnstad and Matthew Ferrari have helped me in understanding the scientific aspects of the problem and allowed me to become a more interdisciplinary researcher. Professor Ottar Bjørnstad was also very helpful in preparation of my research presentations.

I would like to also acknowledge some of the other professors in our department: Professors Bing Li, Donald Richards, Naomi Altman, John Fricks, Aleksandra Slavkovic, and Arkady Tempelman for being great mentors and lecturers, and for their regular feedback throughout my research at Penn State. They have always helped me with suggestions and assistance on both statistical and personal matters.

My sincere thanks also goes to Professors Bruce Lindsay and William Harkness for their support that allowed me to attend statistical conferences to present my work and learn about the ongoing research in the field of statistics and biostatistics. I would like to thank Jennifer Parkes, Laurie Roan, Marina Tempelman, Melissa McCloskey and other staff in our department for their help in many ways and their encouragement.

I thank my fellow students and friends: John Hughes, Sham Bhat, Ivan Simeonov, Isaac Dialsingh, Qing Wang, Esra Kurum, Duy Vu, Kuang-Yao Lee, Junyi Lin, Prabhani Ku-
ruppumullage Don, Lucia Tabacu, Marcus Nunes, Matthew Tibbits and Jialin Xu for their continuous support and advice.

Last but not least, I would like to thank my family: my wife, my parents and my siblings for theirs love and spiritual support throughout my life.

This work was partially supported by the Bill and Melinda Gates Foundation. Any opinions, findings, and conclusions expressed in this work are those of the authors alone, and do not necessarily reflect the views of the Bill and Melinda Gates Foundation.
Chapter 1

Introduction

My dissertation focuses on statistical modeling and inference for infectious disease and population dynamics. In this chapter, I start with a general overview of mathematical models that are used to describe infectious diseases and their importance in improving our understanding of mechanisms underlying the spread of infection. I then briefly describe statistical and computational challenges in statistical modeling and inference for infectious diseases. I provide illustrative examples and review existing statistical approaches for disease dynamics models. Finally, I summarize my contributions to the area of modeling and inference for infectious diseases.

1.1 Models for Infectious Diseases

Despite the progress in sanitation, medicine, and vaccination programs, infectious diseases are still a major cause of suffering and mortality worldwide. This is partly due to the fact that many infections adapt and evolve, and new infectious diseases emerge. At the same time, some existing diseases may re-emerge (cf. Hethcote, 2009; Levins et al., 1994). For example, some of the newly identified diseases include Lyme disease (1975), Legionnaires disease (1976), hepatitis C (1989), hepatitis E (1990), and hantavirus (1993). The human immunodeficiency virus (HIV) emerged in 1981. Since then, it has become a very important sexually-transmitted disease throughout the world. Drug and antibiotic resistance in many people have also become a serious problem for diseases such as tuberculosis, malaria, and gonorrhea. With climate change, infectious diseases like malaria, dengue, and yellow fever
have re-emerged and are spreading into new territories. Outbreaks of life-threatening diseases such as plague, cholera, and hemorrhagic fevers (Ebola, Marburg, Bolivian, etc.) continue to have outbreaks in many countries. In order to improve disease interventions and control strategies, a thorough analysis of outbreaks of infectious diseases with its connections to global climate change and population growth requires sound quantitative methods based on mathematical modeling. A general discussion of the importance of infectious disease modeling is available in Hethcote (2009).

Mathematical models are important tools in analyzing the spread and control of infectious diseases. Simple disease models were used in the eighteenth century. For example, a model for smallpox was first introduced and carefully studied by Daniel Bernoulli in 1760 in order to analyze mortality due to the disease and to evaluate the effectiveness of vaccination of healthy people with the smallpox virus (Bernoulli, 1760; Dietz and Heesterbeek, 2000). In the early 1900s, more sophisticated dynamical systems approaches to modeling infectious diseases started to become popular (cf. Becker, 1979; Castillo-Chávez et al., 1989; Dietz, 1967, 1988; Dietz et al., 1985; Hethcote, 1994; Hethcote and Levin, 1989; Hethcote et al., 1981; Wickwire, 1977). For a detailed history of the progress of the field of disease dynamics models, readers are referred to the books by Bailey et al. (1975), Anderson and May (1992), Grenfell and Dobson (1995), Daley et al. (2001), and Hethcote (2000).

Mathematical models may be thought of as a tool to conceptualize and explain how biological systems behave. Mathematical models use the language of mathematics to produce a refined and precise description of the system. For disease dynamics systems, mathematical models allow us to link the behavior of the system at various scales and/or to extrapolate from a known set of conditions to another. Mathematical models can be very different in their complexity. Some mathematical models are highly complex and require a range of experts to create and maintain them. Other models are very simple and can be easily understood and modified. Similarly, mathematical models can be deterministic and probabilistic. Whether deterministic or stochastic, the models are usually based on well-understood disease transmission principles and may have different levels of flexibility. The practical use of disease dynamics models relies heavily on the realism put into the models by incorporating the mechanisms in the simplest possible fashion so as to maintain major components that influence the spread of the infection. Regardless of their complexity, these models may be useful for answering important questions about the underlying mechanisms of the spread of infection and providing possible means of control of the disease or epidemic.
As discussed in Keeling and Rohani (2008a), mathematical models are often used for prediction and understanding of the underlying complex biological processes. These two important roles of the models are related to their accuracy and interpretability. Models used for predictions are usually required to be very accurate, while models that are used to increase our understanding of the process need to be easily interpretable. There are many interesting and important examples that show how different models can be a great tool for predictions guiding difficult policy decisions where two or more control strategies exist. For example, during the UK foot-and-mouth epidemic in 2001, models led to locally targeted culling that reduced the overall loss of livestock by reducing the number of cases (Keeling, 2005; Keeling and Rohani, 2008a). Predictions from relatively simple (Meltzer et al., 2001) to complex models (Halloran et al., 2002) are also used to choose between mass-vaccination and targeted measures as the best methods of control of smallpox. In another example, models for meningitis are used to estimate a threshold number of cases of the disease that signifies the start of an epidemic (Stollenwerk and Jansen, 2003).

When models are used to increase our understanding of the disease process, they provide a way to examine different complex factors separately to study the effect of these factors on how infectious disease spreads. With such models, for example, one can study the effect of variable numbers of partners on the spread of sexually transmitted diseases (Smith, 1991; Szmuness et al., 1975), the effects of increased transmission between children during school terms (cf. Cauchemez et al., 2008; Glass and Barnes, 2007; Mikolajczyk et al., 2008), or the effects of localized spread of the infection. Even though it might seem that these models are driven to answer only theoretical questions, the insights obtained from such models are often very generic and can be applied to a variety of practical problems. For instance, by understanding all the rich complexities of the real disease dynamics process better, we can develop more complex and more accurate predictive models to solve a wide array of challenging practical problems.

1.2 Types of Mathematical Models

The process of disease dynamics modeling is similar to ecological modeling in general, with the ultimate goal of understanding the spread of a species (Anderson et al., 1979; Bascompte and Rodríguez-Trelles, 1998; Earn et al., 1998; May et al., 1979). In ecology, modelers attempt to describe the precise abundance of a species. Since predicting the exact amount of,
for instance, virus particles in the population is infeasible, disease dynamics models focus on
categorizing the host population according to their infectious status. Therefore, these models
can be compared to the metapopulation models in ecology (cf. Hanski et al., 1997), where
each individual host is a patch for the infection, with transmission and recovery similar to
dispersal and extinction (Nee, 1994; Rohani et al., 2002). I now briefly describe a few classes
of models for infectious diseases, namely, Susceptible-Infectious-Recovered (SIR) models,
models based on point processes and network models.

SIR Models

One approach to disease dynamics modeling is to classify the host individuals within a pop-
ulation as Susceptible (if previously unexposed to the infection), Infectious (if currently
colonized by the infection), and Recovered (if they have successfully recovered from the dis-
ease). These categories or classes are often denoted by S, I, and R. Compartmental models
based on this assumption are called SIR models (Dietz, 1967). The compartmental models
were first studied in Kermack and McKendrick (1927, 1932, 1933). These three papers have
had a great influence on the development of mathematical models for infectious diseases and
are still relevant in many epidemic situations. The first of these papers lays out the SIR for-
alism. In this paper, as in many modern disease dynamics models, the population is taken
to be a constant, and no births or deaths other than from the disease are possible, which
is consistent with the course of an epidemic being short compared with the life time of an
individual. If a group of infectious individuals is introduced into a large population, a basic
problem is to describe the spread of the infection within the population as a function of time.
In the course of time the epidemic may come to an end. One of the most important questions
in epidemiology is then to understand whether this occurs only when all of the susceptible
individuals have contracted the disease or if some infectivity, recovery, and mortality factors
may result in extinction of the infection with many susceptibles still present in the unaffected
population.

![Figure 1.1: Movement of the host between the compartments](image)

In the simplest case, when population demography is ignored, the SIR model describes
the transitions from S to I and from I to R. This is illustrated by a diagram in Figure 1.1. The second transition of the host is simple and is based on the assumption that individuals can move to the recovered class only after they recover. For many popular infectious diseases, it is assumed that individuals recover after a certain time called the infectious period. This time is modeled via a constant recovery rate $\gamma$, which is the inverse of the infectious period. The movement of the individuals from S to I is more complex. This involves factors like the prevalence of the infectious individuals, the underlying population contact structure, and the probability of transmission when contact between individuals from different classes occurs. The connection between these factors and their effect on the transition of the individuals to the infectious class depends on many parameters. One of these parameters is the force of infection, $\lambda$, which is defined as the per capita rate at which susceptible individuals contract the infection. From this definition, the rate at which new infectious individuals are produced is equal to $\lambda X$, where $X$ is the number of individuals in the susceptibles class. The force of infection is often proportional to the amount of individuals in I. For directly transmitted infections, there are two possible ways to define the force of infection: $\lambda = \frac{\beta Y}{N}$ (for frequency dependent transmission) and $\lambda = \beta Y$ (for frequency independent transmission). Here, $Y$ is the number of individuals in the infectious class, $N$ is the population size, and $\beta$ is the product of the contact rates and transmission probability (or transmission rate). In Chapter 2 and Chapter 4, we discuss more complex models with more parameters that allow for flexible transition structures. Another important epidemiological quantity in the context of disease dynamics models is called the basic reproductive ratio (this ratio is denoted by $R_0$). $R_0$ is interpreted as the average number of secondary cases that arise from a single infectious individual in an entirely susceptible population. For the simplest SIR model with frequency-dependent transmission, $R_0$ is equal to $\frac{\beta}{\gamma}$. The importance of $R_0$ is partially related to the threshold of the proportion of the number of susceptibles that is needed for an infection to invade the entire population (cf. Hamer, 1897; Ransome, 1880, 1881).

In order to model the progress of an epidemic in a large population, the vast majority of the disease dynamics models in the literature are in the form of differential equations (cf. Anderson and May, 1992; Keeling and Rohani, 2008b). This is partly because of the assumption that the process of disease transmission is a continuous-time process and the variability in factors like the infectious period may be dynamically important. The simplest continuous-time SIR model is given in Equations (1.1 - 1.3). This model considers a “closed population” with no births, deaths, or migration and assumes a homogeneous mixing of the host individuals. Here, $S$, $I$, and $R$ represent proportions of people in Susceptible, Infectious,
and Recovered classes respectively. As was mentioned above, the parameter $\gamma$ is the removal rate of the infection and $\frac{1}{\gamma}$ determines the average infectious period. The parameter $\beta$ is interpreted as the transmission rate. For this model, it can be shown that $R_0 = \frac{\beta N}{\gamma}$.

$$\frac{dS}{dt} = -\beta SI$$  \hspace{1cm} (1.1)

$$\frac{dI}{dt} = \beta SI - \gamma I$$  \hspace{1cm} (1.2)

$$\frac{dR}{dt} = \gamma I$$  \hspace{1cm} (1.3)

Some infectious disease models are written in discrete time (cf. Bjørnstad et al., 2002a; Finkenstädt and Grenfell, 2000; Mollison and Ud Din, 1993; Xia et al., 2004). Such models were first developed by Reed and Frost in 1928 (Abbey, 1952) in the form of probabilistic equations with binomial transmission probabilities. These “chain binomial” models are carefully studied in Bailey et al. (1975) and Daley and Gani (1999). In a deterministic setting, discrete-time disease dynamics models are obtained by discretization of the underlying continuous-time process and are given in the form of difference equations. In these models, the time increment aims to represent the “generation” length of the infection through a host individual. This, however, may be difficult for some diseases especially when latent and infectious periods are substantially different. Another drawback of models with discrete time is their mathematical instability (Glass et al., 2003). The major advantage of using discrete-time models is that it is much easier to parameterize them using discrete-time data than the associated differential equation counterparts (Bailey et al., 1975; Finkenstädt and Grenfell, 2000; Mollison and Ud Din, 1993). In Chapter 2 and Chapter 4, we discuss more complex discrete-time SIR metapopulation models for measles and meningitis dynamics. For a detailed comparison of deterministic and stochastic models for infectious diseases, readers are referred to Allen and Burgin (2000) and Bartlett (1956a).

**Point Processes**

Another class of disease models is based on using spatio-temporal point processes (cf. Besag and Newell, 1991; Diggle, 2006; Diggle et al., 2005; Keeling et al., 2001; Ma et al., 2006; Meyer, 2010; Meyer et al., 2011; Mollié, 1996). The main idea of these models is to describe
the dynamics of infectious diseases via a spatio-temporal point process by specifying its conditional intensity at location \( x \) and time \( t \), given the history of the process up to time \( t \).

I briefly describe here a point process model used for the UK foot-and-mouth epidemic in 2001, following (Diggle, 2006) and (Keeling et al., 2001). Foot-and-mouth disease is an infectious disease of farm livestock. The virus that causes the disease spreads directly between animals over short distances via airborne droplets and/or indirectly over longer distances, for example, via the movement of different contaminated material between farms. Figure 1.2 shows the evolving pattern of incident and prevalent cases of foot-and-mouth disease as a discrete-time sequence in Cumbria county in the north-west of England in 2001. From this figure, one can easily see that the number of cases of the disease show the typical pattern of an infectious disease process, with possible spatial aggregation of cases resulting from short-range transmissions between neighboring farms, together with more spontaneous cases occurring at relatively long distances.

In order to describe the point process-based model for foot-and-mouth disease, we closely follow the notation introduced in Keeling et al. (2001) and Diggle (2006). Let \( \lambda_{ji}(t) \) denote the conditional rate of transmission of the virus from farm \( j \) to farm \( i \), given the history \( H_t \). Let \( n_1i \) and \( n_2i \) denote the numbers of cows and sheep held on farm \( i \). Let \( I_{ji}(t) \) denote an at-risk indicator for transmission of infection from farm \( j \) to farm \( i \) at time \( t \); it is assumed that \( I_{ji}(t) = 1 \) if farm \( i \) is not infected and not slaughtered and farm \( j \) is infected and not slaughtered by time \( t \). It is then assumed that the reporting date is the infection date plus a constant time \( \tau \), corresponding to the latent period of the disease plus a fixed reporting delay.

The model specifies that

\[
\lambda_{ji}(t) = \lambda_0(t) A_j B_i f(\|x_j - x_i\|) I_{ji}(t),
\]

where \( f(\cdot) \) is a transmission kernel defined as

\[
f(u) = \exp \left( - \left( \frac{u}{\phi} \right)^\kappa \right) + \rho.
\] (1.4)

The function \( \lambda_0(t) \) is an arbitrary baseline hazard,

\[
A_i = \alpha n_1i + n_2i,
\]
Figure 1.2: Incident and prevalent cases of the 2001 foot-and-mouth epidemic in Cumbria (plot from Diggle (2006)). Here, incident cases within the time interval between successive frames are shown in black. Prevalent cases on the dates indicated are shown in grey.
and

\[ B_i = \beta n_{1i} + n_{2i}. \]

The parameters \( \alpha \) and \( \beta \) represent the relative infectiousness and susceptibility, respectively, of cows to sheep.

The important feature of the model, the transmission kernel in Equation (1.4), allows for the separation of the effects of direct and indirect transmission of the virus on the disease dynamics. Here, the powered exponential term corresponds to direct transmission of the infection between farms located close to each other, whereas the parameter \( \rho \) allows for occasional and spontaneous cases occurring at farms located far from all currently infectious areas. For more details about this model and inferential approaches used to learn about its parameters, readers are referred to Keeling et al. (2001) and Diggle (2006). For a detailed review of modeling and inference for spatio-temporal point processes in general, readers are referred to Møller and Waagepetersen (2004).

Often, disease dynamics models based on point processes have analytically intractable likelihood functions. Therefore, in similar fashion to inference for SIR models, likelihood-based inference that relies on Monte Carlo methods require careful tuning to each application and may not always result in reliable estimates. For some of the point process disease dynamics models, a popular inferential approach that is used in the literature is based on partial or profile likelihood methods. We discuss these and other existing approaches for inference for infectious diseases in the following subsections.

**Network Models**

Since networks and epidemiology of directly transmitted infectious diseases are fundamentally linked, another type of models that has been used to describe the structure of relationships between hosts within populations in the study of disease dynamics are the network models (cf. Balcan et al., 2009; Ghani et al., 1997; Keeling and Eames, 2005; Klovdahl, 1985). These models are useful in practice since they account for the fact that each individual has a finite set of contacts to whom they can pass the infection. The goal of these models is to describe the potential pathways of disease transmission within the population using graphs. Knowledge of the structure of the network allows such models to compute the epidemic dynamics at the population scale from the individual-level behavior of infections. Therefore, for many infectious diseases, characteristics of mixing networks, and how these deviate from
the random-mixing norm, may enhance the understanding and prediction of outbreaks and intervention measures for such diseases.

In the analysis of disease dynamics models based on networks, popular random graph concepts that are used to study social networks are often employed in a similar manner. In such models, the vertices or nodes of the graph represent the individuals (or groups of individuals) in a population, and as in social network analysis, the edges between the nodes correspond to contacts among individuals (groups). The definition of a contact between individuals varies between different diseases. For example, for a sexually transmitted disease, an edge would represent a sexual relationship between the two individuals, whereas for a disease such as measles, an edge may be related physical proximity.

In the literature, various disease dynamics models based on networks have been used in order to study different infectious diseases in humans, livestock, and wildlife (cf. Balcan et al., 2009; Ball et al., 2010; Bansal et al., 2006; Eubank et al., 2004; Ghani et al., 1997; Hamede et al., 2009; Kao et al., 2006; Keeling and Eames, 2005; Kiss et al., 2006; Klovdahl, 1985; Meyers, 2007; Meyers et al., 2005; Perkins et al., 2009). Some of the early examples of network-type models used for infectious disease dynamics include Frisch and Hammersley (1963) and Dietz (1967). In Frisch and Hammersley (1963), authors use network-type models to study the spread of a blight through trees. Dietz (1967) overviews several different classes of models based on networks, both deterministic and stochastic, considering models with time-dependent transmission rates with spatial components. More recently, Meyers et al. (2005) applied different network-based models to study the spread of SARS (Severe Acute Respiratory Syndrome), Kao et al. (2006) and Kiss et al. (2006) employed random networks to analyze the risk of foot-and-mouth disease for livestock in Great Britain, and Groendyke (2011) used such models to investigate measles epidemic that spread through the small town of Hagelloch in Germany. For a detailed review of such models and examples, readers are referred to Keeling and Eames (2005) and Meyers (2007).

In network-type models for disease dynamics, the network structure is often assumed to have some specified parametric form. Given the available epidemic data, the parameters of the network models need to be inferred using statistical approaches. However, when the number of nodes of the network is large, likelihood functions for these models is computationally intractable; this makes inference very challenging. We discuss these challenges and different existing and new approach to solve this problem in Chapter 5.
1.3 Statistical Inference for Infectious Disease Models

Generic models that do not have parameters that can be modified or adapted to data can be very useful by providing us with a deep understanding of the transmission and control of infectious diseases. For example, very simple and general models that distinguished between “core” and “non-core” groups (having high and low contact rates, respectively) led to the successful adoption of contact tracing for the control of gonorrhea in the US (Hethcote and Yorke, 1984). Recently Wallinga et al. (2010) demonstrated that in a wide range of situations, vaccinating individuals in the group experiencing the highest force of infection would be optimal for reducing the transmission of a novel infection.

Motivation for Parameter Inference

When one wishes to examine the spread mechanisms of a specific infection, the models need to be parameterized accordingly. From a modeling perspective, for example, only the parametrization makes models for smallpox and measles different (cf. Keeling and Rohani, 2008a). Therefore, while generic disease dynamics models provide a tool to obtain general and intuitive explanations about the disease dynamics, models tailored to a specific situation with parameters estimated as accurately as possible can be extremely helpful in providing detailed public health guidance. Fitting models to data with accurate parameter estimates is also important for epidemiology because it allows the modelers to compare different alternative models for discriminating between different hypotheses about underlying biological mechanisms (cf. King et al., 2008; Koelle and Pascual, 2004; Miller et al., 2006; New et al., 2009; Wearing and Rohani, 2006; Webb et al., 2006) and/or possible environmental drivers of the disease processes (cf. Ferrari et al., 2008; King et al., 2008; Koelle et al., 2005; Pascual et al., 2000; Shaman et al., 2010). Furthermore, in many cases, the relative gains from different epidemiological response options depend on quantitative values of parameters or on case-specific features of the disease process such as multiple transmission routes and their relative importance (cf. Blower et al., 2000; Ferguson et al., 2006; Keeling, 2005; Longini Jr and Halloran, 2005; Medlock and Galvani, 2009; Miller et al., 2006; Truscott et al., 2010).

In the simplest case, the general SIR model that accounts for population demography, for example, has four parameters: the birth rate, the natural death rate, the infectious period (or the inverse of the recovery rate $\gamma$), and the transmission rate $\beta$. We discussed some of these parameters in Section 1.2. More complex models with more parameters are introduced
in Chapter 2 and Chapter 4. A model for meningitis dynamics discussed in Chapter 2 has parameters that represent (i) the incidence-dependent effects of movement of individuals; and (ii) the incidence-independent climate effects on the transmission of the infection and number of cases of the disease. This model allows for the fact that these effects may seasonally vary by using additional, discrete, timing parameters. A gravity model for measles dynamics discussed in Chapter 4 accounts for seasonally changing transmission rates and focuses on the parameters describing the topology of the movement of the infection. These so-called gravity parameters separate the effects of population sizes of cities and distances between them. Both the models for meningitis and measles also account for the fact that there might be a substantial underreporting in the data. In these epidemiological models, whether simple or complex, when data are available for the disease counts, the unknown parameters of the models need to be inferred from the data. With accurate parameter estimates, scientists can answer important biological question about the spread of the infection, for example, to optimize epidemic response strategies, to identify major transmission pathways, and to predict emerging outbreaks. However, parameter inference for disease dynamics models is often a very difficult statistical problem that requires substantial innovations to develop statistical methods that perform well and are computationally tractable. Estimating the parameters of disease dynamics models is especially challenging because transmission of the infection is usually unobserved. We may know at best when and how a particular host individual contracted the infection, but apart from sexually transmitted diseases in humans, it is rare to know with any degree of certainty which other individuals may have passed the infection to them. Hence, a fundamental goal of parameter inference for disease dynamics models from epidemic data are to learn about the latent transmission process, and obtain accurate fitted models of transmission to answer important epidemiological questions: at what rate will hosts become infectious, and what factors of the host population and the environment impacts that rate (cf. Fine and Clarkson, 1982; Finkenstädt and Grenfell, 2000; King et al., 2008; Metcalf et al., 2009; Xia et al., 2005)? Statistical inference for disease dynamics models is the main topic of this dissertation. I introduce a new inference approach for complex disease dynamics models that is useful for a number of interesting and important probabilistic models for infectious diseases. This approach is inspired by work in the field of emulation and calibration for complex computer models (cf. Bayarri et al., 2007a; Craig et al., 2001; Kennedy and O’Hagan, 2001; Sacks et al., 1989a). Based on different examples and various epidemiological data, I will demonstrate that this approach results in reliable parameter estimates and a good model fit, and is also computationally efficient.
My motivating example for this approach is a so-called gravity model for measles dynamics discussed in Chapter 4. Measles is an extremely contagious, acute and immunizing infectious disease. Dynamics of measles possess all the properties of a complex epidemiological system. The gravity model for measles is a metapopulation model for regional dynamics of the disease that uses a gravity coupling model and a time series susceptible-infectious-recovered (TSIR) model for local dynamics of the infection. In Chapter 4, using our new approach, I analyze data for measles outbreaks for 952 cities in England and Wales between 1944 and 1965. As an illustration, these data are plotted in Figure 1.3 for selected cities of varying population sizes.

![Figure 1.3: Bi-weekly reported cases of measles for different cities (from the UK Registrar Generals data for 954 cities in England and Wales for years 1944-1966).](image)

A model like the gravity model for measles and large epidemiological data present a number of challenges:

1. Epidemiological data are usually scarce and incomplete. For example, in the context of the gravity model for measles, we have data on only three processes (population sizes, birth rates, and the number of new cases of measles) in a system with many state variables and other disease processes. This and the complexity of the models make the likelihood functions very complex and computationally expensive or even intractable.
2. There is substantial measurement error in epidemiological data and disease dynamics systems are almost invariably driven by endogenous dynamic processes with demographic and environmental noise. For instance, in the measles data we consider in Chapter 4, while the behavior of the system for big cities is almost deterministic, time series for small cities are extremely variable and affected by measurement error. This may pose difficulty to traditional likelihood-based statistics since the noise may hugely affect the joint probability density of the observable data and sometimes make it less useful as a basis for obtaining measures of statistical fit. This means that even when we resolve challenges related to the computational expense of the likelihood, traditional inference may lead to poor parameter estimates and the fitted model may not capture important biological characteristics of the observed data. We return to this in detail in Chapter 4.

3. There are possible identifiability issues between the parameters. We discuss this in more detail in Chapter 4.

Existing Approaches

There are at least two major classes of inferential approaches for disease dynamics models: (i) methods that involve explicit evaluations of the likelihood, often using Bayesian approaches and Markov chain Monte Carlo (MCMC) (cf. Cauchemez and Ferguson, 2008; Jewell et al., 2009a,b; Morton and Finkenstädt, 2005; New et al., 2009, and references therein); and (ii) methods that replace likelihood evaluations with simulations from the disease dynamics model (cf. He et al., 2010; Ionides et al., 2006; Kendall et al., 1999; King et al., 2008; Liu and Chen, 1998; Padhukasahasram et al., 2006; Pritchard et al., 1999; Sisson et al., 2007).

When evaluating the likelihood function at different parameter settings, if the major difficulty comes from the unobservable latent variables that often need to be inferred from the available data, a popular and particularly effective solution to learn about the parameters of interest is to use a Bayesian approach via numerical techniques like Markov chain Monte Carlo algorithms (cf. Cauchemez and Ferguson, 2008; Jewell et al., 2009a,b; Morton and Finkenstädt, 2005; New et al., 2009). The methodology incorporates a mechanism to infer unobserved or missing data, since it treats all unknown parameters and data alike as random variables, for which full posterior distribution can be estimated. However, for many interesting disease dynamics models, even MCMC cannot be applied directly since it assumes that
we can at least calculate the unnormalized likelihood function for each parameter/latent variables setting. This usually happens when the population gets larger and/or the process gets too intricate and the likelihood becomes mathematically intractable. Additionally, methods based on MCMC are very challenging for disease dynamics models because it is often very difficult to construct MCMC proposal distributions that can result in algorithms with acceptance and “mixing” rates high enough for the posterior distribution to be sampled well within a reasonable amount of time (cf. New et al., 2009). Our method is based on using MCMC as well. The difference of our approach from the traditional MCMC approaches that has been used in the literature is that we do not explicitly consider the likelihood function (normalized or unnormalized). Instead, we first obtain an approximate likelihood function for the model by focusing on the scientifically important aspects of the data. This new likelihood function is generally easier to evaluate at each parameter setting and allows for an efficient MCMC approach to Bayesian inference for the parameters. Therefore, we believe that our approach is widely applicable to problems where MCMC based on the likelihood functions directly is infeasible.

Inferential approaches that require simulations from the disease dynamics model without explicit evaluation of the likelihood function are sometimes called “plug-and-play” methods. These methods assume that while it is impractical to calculate the likelihood, it is easy and fast to simulate from the probability model corresponding to the likelihood. Plug-and-play methods may be Bayesian or Non-Bayesian. Bayesian plug-and-play methods are based on using approximate Bayesian computation (cf. Bortot et al., 2007; Grelaud et al., 2009; McKinley et al., 2009; Padhukasahasram et al., 2006; Pritchard et al., 1999; Ratmann et al., 2009; Tanaka et al., 2006; Toni et al., 2009; Wilson et al., 2009) via popular sampling techniques like MCMC (cf. Marjoram et al., 2003; McKinley et al., 2009) or sequential Monte Carlo (cf. Liu and Chen, 1998; Sisson et al., 2007). The main idea of ABC approaches is to combine an estimate of the likelihood with a prior to produce an approximate posterior, which is often referred to as the ABC posterior. I will discuss these methods in detail in Chapter 2.

Non-Bayesian plug-and-play methods are based on likelihood inference via an iterated filtering (cf. He et al., 2010; Ionides et al., 2006). The idea of these methods is to obtain approximate maximum likelihood estimates of the parameters by solving a sequence of problems involving calculations of means of conditional distributions of unobservable state vectors (filtering problem). Under certain conditions, the solutions of these filtering problems can be
shown to maximize the likelihood function over unknown model parameters (cf. Bretó et al., 2009; Ionides et al., 2006). At each step of the algorithm, filtering problems are solved by simulations from the dynamical model and sequential Monte Carlo.

In the following, in order to describe the basics of non-Bayesian plug-and-play methods, we closely follow Ionides et al. (2006). Let us suppose that we have a state space model that consists of an unobserved Markov process, \( x_t \), called the state process and an observation process \( y_t \). The processes depend on an unknown vector of parameters, \( \theta \). Observations take place at discrete times, \( t = 1, \ldots, T \); we write the vector of concatenated observations as \( y_{1:T} = (y_1, \ldots, y_T) \); \( y_{1:0} \) is defined to be the empty vector. We assume that the 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 \( f(x_0|\theta) \), where \( f(\cdot|\cdot) \) is a generic density specified by its arguments. The likelihood for the observation process is then given by the identity \( f(y_{1:T}|\theta) = \prod_{t=1}^{T} f(y_t|y_{1:t-1}, \theta) \). We assume the state process, \( x_t \), may be defined in continuous or discrete time, but only its distribution at the discrete times \( t = 1, \ldots, T \) directly affects the likelihood. The goal is then to find the maximum of the likelihood as a function of \( \theta \).

The basic idea of a plug-and-play method based on iterated filtering is to replace the original model with a similar model, in which the time-constant parameter \( \theta \) is replaced by a time-varying process \( \theta_t \). In this new model, the densities \( f(x_t|x_{t-1}, \theta) \), \( f(y_t|x_t, \theta) \), and \( f(x_0|\theta) \) of the model with constant \( \theta \) 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 \( \theta_t \) is assumed to be a random walk defined as

\[
\begin{align*}
E[\theta_t|\theta_{t-1}] &= \theta_{t-1} \\
\text{Var}(\theta_t|\theta_{t-1}) &= \sigma^2 \Sigma \\
E[\theta_0] &= \theta \\
\text{Var}(\theta_0) &= \sigma^2 c^2 \Sigma
\end{align*}
\]

Here, \( \sigma \) and \( c \) are scalars, and it can be seen that the model with time-varying \( \theta \) is identical to the original model when \( \sigma = 0 \). The matrix \( \Sigma \) is usually a diagonal matrix giving the respective scales of each component of \( \theta \). The objective is then to estimate \( \theta \) by taking the limit as \( \sigma \to 0 \). This estimate is obtained by the following algorithm:

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

2. For \( n \) in \( 1, \ldots, N \)
(i) Set $\sigma_n = \alpha^{n-1}$. For $t = 1, \ldots, T$, evaluate $\hat{\theta}^{(n)}_t = \hat{\theta}_t(\hat{\theta}^{(n)}, \sigma_n)$ and $V_{1,n} = V_t(\hat{\theta}^{(n)}, \sigma_n)$, where $\hat{\theta}_t(\theta, \sigma) := E[\theta_t | y_{1:t}]$ and $V_t(\theta, \sigma) := \text{Var}(\theta_t | y_{1:(t-1)})$.

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

3. Take $\hat{\theta}^{(N+1)}$ to be a maximum likelihood estimate of $\theta$ in the original model.

In this algorithm, the computationally challenging step 2 requires using only well studied filtering techniques (Anderson and Moore, 1979; Arulampalam et al., 2002) to calculate the conditional mean and variance of the random walk $\theta_t$. For more details and justification of this algorithm, readers are referred to Ionides et al. (2006) and Ionides et al. (2011). Another useful implementation of iterated filtering-based plug-and-play method can be found in the software package POMP (King et al., 2009) written for the R statistical computing environment (R Development Core Team 2006). This implementation follows the algorithm described by King et al. (2008, supplementary information therein).

Our inferential approach for disease dynamics models is based on simulating from the model at different parameter values and therefore can be classified as a plug-and-play method. However, even though ABC methods are slightly easier to implement, our method has important advantages over ABC: (i) it does not need to specify a tolerance level that can affect the inference; and (ii) it can be applied to problems where simulations from the model are expensive. In contrast to non-Bayesian plug-and-play methods, our method can be easily generalized to various problems involving inference for space-time data, while methods based on iterated filtering may require careful case-specific adjustments (mostly in the context of inference for time series data) to avoid “particle depletion”, the analog of poor mixing in MCMC (see section 3.2 of Newman et al. (2009)).

There are also methods based on different optimization criterion using least-squares trajectory matching (cf. Arora and Biegler, 2004; Banks et al., 1981; Biegler et al., 1986; Bock, 1982; Ciupé et al., 2006) and/or gradient matching (cf. Brunel, 2008; Ellner et al., 2002; Swartz and Bremermann, 1975; Varah, 1982). These so-called generalized profiling approaches are based on the likelihood approach to quantifying uncertainty. The numerical construction of these likelihood uncertainty regions requires solution of a series of constrained nonlinear optimization problems (cf. Cao and Ramsay, 2009; Hooker et al., 2011; Qi and Zhao, 2010; Ramsay et al., 2007), which makes the methodology computationally challenging for many practical problems.
For some disease dynamics models, including the models based on point processes, methods of inference via profile or partial likelihood functions are used (cf. Bjørnstad et al., 2002a; Diggle et al., 2005; Eisenberg et al., 2002; King et al., 2008). In general, however, methods based on profile likelihoods may lead to poor parameter inference (cf. Berger et al., 1999).

The gravity model for measles dynamics discussed in more detail in Chapter 4 was first proposed by Xia et al. (2004). The main goal of the gravity model is to understand the topology of the movement of the infection between communities to assess the effects of various factors on the transmission of measles. These factors are the population sizes of communities, distances between them and abundance of the infectious individuals at each community. In the gravity model, this topology is described by a set of parameters called the gravity parameters. The main statistical problem for this model is to infer these unknown gravity parameters from data. Since each likelihood evaluation for the gravity model is computationally very expensive, the only previous approach to learn about the gravity parameters is based on minimizing ad hoc objective functions. This approach is used to obtain point estimates of the parameters without a proper uncertainty quantification (Xia et al., 2004).

In Chapter 4, we first demonstrate that traditional likelihood-based approaches (Bayesian or maximum likelihood) to inference for the gravity model are problematic and may lead to poor parameter estimates. We also show that forward simulations of the model at the obtained parameter settings do not reproduce epidemiological features of the data deemed key in Xia et al. (2004). We then describe how our approach can be used to infer the parameters of the gravity model and illustrate that it resolves the above issues with the likelihood-based inference. We demonstrate in a number of examples that our method recovers the true parameters and the resultant fitted model captures biologically relevant features of the data. We note that existing plug-and-play approaches to inference cannot be applied to the gravity model as it is computationally expensive to simulate from the model.

Generally, the methodology we describe in my dissertation may be useful for models where the likelihood is expensive to evaluate or in situations where the likelihood is unable to capture characteristics of the model that are of scientific interest. It is also applicable to problems where current traditional likelihood-based methods may lead to poor inference. In the context of another challenging problem, in Chapter 5, I demonstrate how this methodology can be used to learn about the parameters of complex mixture random graph models. These models are an important tool used in describing protein-protein interactions, gene regulatory networks, Web-pages, and social networks.
In Chapter 2, we describe our new model for meningitis dynamics to investigate different hypotheses about the transmission of the disease that may help in designing efficient public health responses. Easily interpretable parameters of our model allow us to study and compare differences in the attack rates, rates of transmission and the possible underlying environmental effects on meningitis dynamics during different seasons. For reasons similar to ones encountered in inference for the gravity model, standard maximum likelihood or Bayesian inference for the meningitis model is problematic. However, forward simulations for this model are relatively cheap and therefore we use it to demonstrate the usefulness of an ABC approach to infer the unknown parameters of the model. Using simulated data examples, we show that it is possible to learn about some of the important parameters of the meningitis model using our methodology. I note again, however, that ABC is generally not applicable to problems where simulations from the model are expensive. For these models, our emulator-based approach is a viable alternative.

In Chapter 3, we present a study of a population dynamics of an insect called the gypsy moth. This insect is the most important cause of forest defoliation in the northeastern United States. It is known that the gypsy moth outbreaks occur with periodicities that differ from region to region. Learning about the gypsy moth population periodicities is a challenging problem because population sizes are not directly observable: there is only indirect information in the form of binary space-time defoliation data. In order to infer the unknown population periodicities of the gypsy moth, we describe an approach based on a space-time Gaussian process model for gypsy moth populations. We apply the approach to Pennsylvania defoliation data. My results from a simulation study demonstrate well the flexibility offered by a latent space-time Gaussian process model.

### 1.4 Summary of Contributions

From my perspective, this dissertation makes the following main contributions:

- We develop a new inferential approach that is particularly useful when the likelihood function is computationally intractable and in cases where simulations from the probability model might be too expensive to allow the use of other popular inferential methods. This approach provides a way to directly incorporate important scientifically-relevant characteristics of the data while accounting for imperfections in how well the model compares to reality (data discrepancy).
• We provide a rigorous approach for answering some important questions about measles dynamics.

• We develop a new model for meningitis dynamics to answer important scientific questions that may be useful for designing epidemic response strategies. This model allows us to separate the effect of transmission from environmental effects on disease incidence. Such models are particularly relevant to study the impact of climate on disease dynamics.

• We develop a model and computational approach to infer unknown gypsy moth population periodicities from indirect information by exploiting dependencies across space and time.

1.5 Dissertation Organization

The rest of this dissertation is organized as follows. In Chapter 2, I describe a new model for meningitis dynamics. In the context of this model I explain how one can infer the unknown parameters of the model using approximate Bayesian computation methods. These methods are useful for inference problems when the likelihood functions are expensive, but simulations from the model are relatively cheap. In Chapter 3, I discuss the basics of Gaussian processes and covariance functions for spatially dependent data. I demonstrate how Gaussian processes can be used to model latent data in the problem of inference of population periodicities of the Gypsy moth in Northeastern United States. In Chapter 4, I describe my new, Gaussian process-based inferential approach in the context of a gravity model for measles. In Chapter 5, I apply this method to infer the unknown parameters of a mixed membership graph model. Finally, in Chapter 6, I discuss the results and contributions of this dissertation and overview avenues for future work.
Chapter 2

Approximate Bayesian Computation (ABC) and Inference for a Model for Meningitis Dynamics

Approaches for sampling from posterior distributions typically assume that it is possible to repeatedly evaluate the likelihood function quickly. However, for many complex probability models such likelihoods are either too computationally expensive or mathematically intractable. For some of these complex models, forward simulation are computationally cheap. Approximate Bayesian computation is a class of methods of inference for such models. These methods replace calculations of the likelihood function by a step which involves simulating data for different parameter values, and comparing summary statistics of the simulated data to summary statistics of the observed data. In this chapter, I will first give a brief overview of these approaches for models with intractable likelihood functions. I will then provide an example of the application of an ABC method to a new model for meningitis dynamics from our paper titled “A compartmental model for meningitis: separating transmission from climate effects on disease incidence” accepted for publication in the Journal of Agricultural, Biological, and Environmental Statistics. In the context of meningitis dynamics, the new model and ABC-based inferential approach allows us to answer important epidemiological questions that may help with building optimal vaccination strategies.
2.1 ABC Algorithms: Background

Approximate Bayesian computation (ABC) methods assume that while it is impractical to calculate the likelihood, it is easy and fast to simulate from the probability model corresponding to the likelihood. ABC approaches combine an estimate of the likelihood with a prior to produce an approximate posterior, which is often referred to as the ABC posterior. The use of ABC first became popular within population genetics, where simulation from a range of population genetics models can be done quickly, but where calculating the likelihood is very expensive for any realistic sized datasets. Pritchard et al. (1999) was first to use ABC in the context of inference about human demographic history. Others have applied the method to study inference for recombination rates (Padhukasahasram et al., 2006), evolution of pathogens (Wilson et al., 2009) and protein networks (Ratmann et al., 2009). In more recent work, the idea has been applied within epidemiology (McKinley et al., 2009; Tanaka et al., 2006), inference for extremes (Bortot et al., 2007), dynamical systems (Toni et al., 2009), and Gibbs random fields (Grelaud et al., 2009).

Let us suppose that we have a probabilistic model, a collection of probability distributions \( \{\pi(y|\theta)\} \) that depends on some unknown \( d \)-dimensional parameter \( \theta = (\theta_1, \cdots, \theta_d) \), and let \( D \) denote the data. Given the prior distribution \( \pi(\theta) \), the goal is to approximate the posterior \( \pi(\theta|D) \propto L(D|\theta)\pi(\theta) \), where \( L(D|\theta) = \pi(D|\theta) \) is the likelihood of \( \theta \) given the data \( D \) and \( \pi(\cdot) \) is a prior of the parameters.

If \( U = U(\theta) \) is simulated data from \( \pi(y|\theta) \), we can define the ABC posterior in terms of (i) a function \( S(\cdot) \) which calculates the summary statistics from \( U \), (ii) a density kernel \( K(\cdot) \), which integrates to 1; and (iii) a bandwidth \( \epsilon > 0 \). Let \( s_{\text{obs}} = S(D) \). We can now define an approximation to the likelihood as

\[
\pi(\theta|s_{\text{obs}}) = \int L(y|\theta)K((S(y) - s_{\text{obs}})/\epsilon)dy.
\]

Then, the ABC posterior can be defined as

\[
\pi_{\text{ABC}}(\theta|s_{\text{obs}}) \propto \pi(\theta)\pi(\theta|s_{\text{obs}}).
\]

The main idea of ABC is that the ABC posterior is an approximation to the true posterior for \( \theta \) given the appropriate choice of \( S(\cdot) \), \( K(\cdot) \) and \( \epsilon \). If \( S(\cdot) \) is the identity function, for example, the resultant \( \pi(\theta|s_{\text{obs}}) \) will just be a kernel density approximation to the likelihood
function. When $\epsilon \to 0$, the ABC posterior converges to the true posterior. For choices of $S(\cdot)$ that are different from the identity function, the kernel function measures the closeness between summary statistics calculated from the model output $Y$ and the data $D$. The advantage of using this construction is that it makes it easy to construct Monte Carlo algorithms that approximate the ABC posterior. These algorithms only require simulations from $\pi(y|\theta)$.

In the following sections, I describe ABC algorithms assuming that $K(x) = 1$ for simplicity. We note that this assumption imposes no restriction, as if $K_0(x)$ is a density kernel, then so is $K(x) = h_0^{-d}K_0(x/h_0)$ for any $h_0 > 0$. Thus we can choose $h_0$ so that $\max K(x) = 1$ and note that a bandwidth $\epsilon$ for kernel $K_0(x)$ is equivalent to a bandwidth $h = \epsilon h_0$ for $K(x)$, so the value of $h_0$ just redefines the units of the bandwidth $h$.

### 2.2 Outline of the Algorithms

When distances between full datasets are used, the ABC methods have the following generic form (cf. Toni et al., 2009):

- Sample a candidate parameter vector $\theta^*$ from some proposal distribution $\pi(\theta)$
- Simulate a dataset $D^*$ from $\pi(y|\theta^*)$.
- Compare a dataset, $D^*$, with the data $D$, using a distance function, $d(\cdot, \cdot)$, and a tolerance level $\epsilon$; if $d(D, D^*) \leq \epsilon$, accept $\theta^*$. The tolerance level $\epsilon$ is the desired level of agreement between $D$ and $D^*$.

The output of an ABC algorithm is a sample of parameters from a distribution $\pi(\theta|d(D, D^*) \leq \epsilon)$. If $\epsilon$ is sufficiently small, then the distribution $\pi(\theta|d(D, D^*) \leq \epsilon)$ will be a good approximation for the posterior distribution $\pi(\theta|D)$. When it is not easy to define a suitable distance function $d(\cdot, \cdot)$ between two full datasets, as was mentioned in the previous subsection, one may instead use a distance defined on some summary statistics, $S(D)$ and $S(D^*)$, of the datasets.

The simplest ABC algorithm is based on the rejection sampler (Pritchard et al., 1999). We call this Algorithm 1 and it is given below.

The disadvantage of Algorithm 1 is that it has a low acceptance rate when the prior distribution is very different from the posterior. To avoid this issue, Marjoram et al. (2003)
Algorithm 1 ABC algorithm based on rejection sampler

1. Sample $\theta$ from $\pi(\theta)$.
2. Simulate a dataset $D^*$ from $\pi(y|\theta^*)$.
3. If $d(D, D^*) \leq \epsilon$, accept $\theta^*$, otherwise reject.
4. Return to 1.

Algorithm 2 ABC algorithm based on MCMC

1. Initialize $\theta_i, i = 0$.
2. Propose $\theta^*$ according to a proposal distribution $q(\theta|\theta_i)$.
3. Simulate a dataset $D^*$ from $\pi(y|\theta^*)$.
4. If $d(D, D^*) \leq \epsilon$, go to 5, otherwise set $\theta_{i+1} = \theta_i$ and go to 6.
5. Set $\theta_{i+1} = \theta^*$ with probability

\[
\alpha = \min \left( 1, \frac{\pi(\theta^*)q(\theta_i|\theta^*)}{\pi(\theta_i)q(\theta^*|\theta_i)} \right)
\]

and $\theta_{i+1}$ with probability $1 - \alpha$.
6. Set $i = i + 1$, go to 2.

The output of this algorithm is a Markov chain with the stationary distribution $\pi(\theta|d(D, D^*) \leq \epsilon)$ (Marjoram et al., 2003). Another version of an ABC algorithm was developed by Sisson et al. (2007) and is based on sequential Monte Carlo (SMC). This algorithm is a generalization of Algorithm 1 and is called the ABC SMC.

McKinley et al. (2009) presents a modification of Algorithm 2, where for each proposed candidate value $\theta^*$, the algorithms simulates $R$ datasets, $D^*_1, \ldots, D^*_R$ from $\pi(y|\theta^*)$, and
calculates
\[ \hat{L}(D|\theta^*) = \frac{1}{R} \sum_{r=1}^{R} I(d(D', D) < \epsilon). \]

This algorithm accepts the proposed \( \theta^* \) with the probability
\[ \alpha = \min(1, \frac{\hat{L}(D|\theta^*)\pi(\theta^*)q(\theta_i|\theta^*)}{\hat{L}(D|\theta_i)\pi(\theta_i)q(\theta^*|\theta_i)}). \]

McKinley et al. (2009) also argues that in some cases, it is better to use a composite approximate likelihood instead of using the usual approximate likelihood. We avoid the discussion of their approach here since it is lengthy and problem-specific.

The ABC algorithms are limited to probability models that are easy to simulate from, and therefore are not useful for complex infectious disease models like the one described in Chapter 4, where simulations from the model are expensive as well. However, even though there still seem to be a number of methodological issues to resolve, ABC-based methods can be applied in a fairly large set of problems. In the next section, I provide an illustration of how an ABC method works in the context of a real world inference problem for a new model for meningitis dynamics.

\[ \text{2.3 Modeling and ABC-based Inference for Meningitis Dynamics} \]

The timing and size of many infectious disease outbreaks depend on climatic influences. Meningitis is an example of such a disease. Every year countries in the so-called African meningitis belt are afflicted with meningococcal meningitis disease outbreaks. The timing of these outbreaks coincide with the dry season that starts in February and ends in late May. There are two main hypotheses about this strong seasonal effect. The first hypothesis assumes that during the dry season there is an increase in the risk that an individual will transition from being an asymptomatic carrier to having invasive disease. The second hypothesis states that the incidence of meningitis increases due to higher transmission of the infection during the dry season. These two biological hypotheses suggest dynamics that would necessitate different public health responses: the first would result in broadly correlated outbreak dynamics, and thus a regional vaccination response; the second would result
in locally correlated outbreaks, spreading from location to location, for which a localized response may be effective in containing regional spread. In this section of my dissertation, we describe a new statistical model to investigate these hypotheses. Easily interpretable parameters of the model allow us to study and compare differences in the attack rates, rates of transmission and the possible underlying environmental effects during the dry and non-dry seasons. Standard maximum likelihood or Bayesian inference for this model is infeasible as there are potentially tens of thousands of latent variables in the model and each evaluation of the likelihood is expensive. We therefore propose an approximate Bayesian computation (ABC) approach to infer the unknown parameters. Using simulated data examples, we demonstrate that it is possible to learn about some of the important parameters of our model using our methodology. We apply our modeling and inferential approach to data on cases of meningitis for 34 communities in Nigeria from Medicins Sans Frontières (MSF) and World Health Organization (WHO) for 2009. For this particular dataset we are able to find weak evidence in favor of the first hypothesis, suggesting a regional vaccination response.

2.3.1 Background on Meningitis

Meningococcal meningitis is a disease caused by a commensal bacterium called *Neisseria meningitidis* (cf. Ryan et al., 2004). Asymptomatic carriage of *N. meningitidis* is common in most populations, and invasive disease is a rare outcome of infection (Trotter et al., 2005). The population prevalence of meningococcal carriage varies from 3% to 30% according to different studies (cf. Trotter and Greenwood, 2007).

The epidemiology of meningitis varies by geographical area and time. Outbreaks are most severe in the sub-Saharan African meningitis belt, where they have recurrent since the early 20th century (Lapeyssonnie and Organization, 1963). This belt was first described in Lapeyssonnie and Organization (1963) and now includes 15 countries (Greenwood, 1999, 2006; Molesworth et al., 2002). Climatic conditions may influence the risk of meningococcal disease: epidemics start during the dry season, i.e. from February until May (Greenwood et al., 1979; Sultan et al., 2005), during which it is typically hot and arid, and ends with the beginning of the rainy season (Harrison et al., 2009). Despite the strong correlation of outbreaks with the dry season, it has been impossible to predict epidemics during a given year in a particular area. Previous studies also show that crowding may play a role in meningococcal disease risk (Brundage and Zollinger, 1987). Behavioral risk factors, such as smoking, going to bars, kissing, and university dormitory attendance have all been associated
with meningococcal disease and/or meningococcal carriage (Cookson et al., 1998; Fischer et al., 1997; Harrison et al., 2009; Imrey et al., 1996; Tappero et al., 1996).

From a public health policy perspective, there are two important hypotheses about this strong seasonal effect (Trotter and Greenwood, 2007). The first hypothesis assumes that there is an increase in the risk of transitioning from asymptomatic carriage to invasive disease during the dry season; environmental conditions during the dry season may damage the mucosal defenses making it more likely that bacteria will move from the nasopharynx (where they are often commensal) to the bloodstream and then the central nervous system, leading to invasive disease. In favor of this hypothesis, Molesworth et al. (2003) finds that highly seasonal variables like absolute humidity, dust and rainfall profiles, land-cover type are all strongly associated with the location of epidemics. It is known that these variables vary between years because of the climate change as well. Using a stepwise regression, Thomson et al. (2006) shows that changes in rainfall and dust are the most consistent predictors of changes in meningitis incidence.

The second hypothesis states that the incidence of meningitis increases due to higher transmission of the bacterium. If the second hypothesis is true, then it would be expected that both prevalence of carriage and invasive disease would increase during the dry season. Evidence for this hypothesis is equivocal; longitudinal carriage studies conducted in the meningitis belt have concluded that there is no association between carriage prevalence and season (Trotter and Greenwood, 2007) and therefore do not support the second hypothesis, while others have found higher prevalence of carriage in the dry season (Kristiansen et al., 2011). We briefly note that the lack of longitudinal surveillance on meningitis carriage is a significant impediment to the understanding of the dynamics of meningitis transmission. Thus a key challenge is to differentiate among the predictions of these competing models with respect to incidence of invasive disease, which is more commonly monitored than bacterial carriage. We now summarize the hypotheses as follows:

\( H_1 \): increase in the number of incidences in the dry season is entirely due to seasonally forced changes in disease status from asymptomatic to infectious.

\( H_2 \): increase in the number of incidences in the dry season is entirely due to seasonally forced changes in the transmission rate during the dry season.

These two biological hypotheses suggest dynamics that would necessitate different public health responses: \( H_1 \) would result in broadly correlated outbreak dynamics, and thus a regional vaccination response, \( H_2 \) would result in locally correlated outbreaks, spreading from
location to location, for which a localized response may be effective in containing regional spread.

In this section of the dissertation, we propose a parsimonious statistical model to investigate these competing hypothesis. Intuitively meaningful parameters of our model allow us to conveniently study the relative contribution of the transition from carriage to invasive disease, which is expected to be driven by environmental conditions and independent of the incidence of infection, and transmission, which is expected to scale with the incidence of infection. Since evaluation of the likelihood for the model is expensive, we propose a fast approximate Bayesian computation (ABC) method to infer the unknown parameters (cf. Beaumont et al., 2002; Marjoram et al., 2003; Pritchard et al., 1999).

We present two important simulated examples. True parameters of the first example are chosen so that the effect of the movement of infection from neighboring communities is relatively small while climatic effects are large; therefore an outbreak is sparked mainly because of the change in environmental conditions. In the second simulated example, we choose the incidence dependent movement parameters to be relatively large and make the incidence independent environmental parameters very small; making the migration of the infection from neighboring areas the predominant factor in starting an outbreak of the disease. The scenario when both the movement and environmental effects are small or large at the same time resulted in unrealistic simulated data. Our goal in considering these examples is to investigate scenarios that represent the extremes of the prevailing hypothesis about meningitis dynamics and to demonstrate that our computational approach results in reliable parameter inference. Once we have established that our approach is reliable and computationally expedient, we apply our methods to the problem that motivated this research: data on meningitis from Medecins Sans Frontieres (MSF) and World Health Organization (WHO) for 34 communities in Nigeria in 2009.

The section is organized as follows. Subsection 2.3.2 describes our statistical model in detail. Subsection 2.3.3 describes our inferential approach and computational challenges posed by the model. In Subsection 3.2.5, we describe the data and our results. Finally, in Subsection 3.2.6, we summarize our results and discuss our scientific conclusions.
2.3.2 A Model for Meningitis Dynamics

Using a common theoretical framework used to describe the dynamics of infectious diseases, we divide the human host population into groups containing susceptible, infectious and recovered individuals (cf. Keeling and Rohani, 2008a). Our goal is then to study the dynamics of the number of people within each group. Let $I^i_t$ and $S^i_t$ denote the number of infectious and susceptible individuals respectively in disease generation $t$ in community $i$ and let $m^i_t$ be the number of infectious people moving to community $i$ at time $t$. We assume that the movement of individuals is transient in a way that does not reflect changes in the absolute size of communities. Let $V^i_t$ be the proportion of vaccinated individuals in community $i$ at time $t$. We assume that this variable is equal to zero until the first week of the vaccination campaign. Additionally, denote the size of community $i$ at time $t$ by $N^i_t$. Our model can then be described as follows. First, we model the number of cases of meningitis by

\[ I^i_{t+1} \sim \text{Binomial} \left( S^i_t, 1 - \exp \left( -\alpha_t - \beta_t \left( \frac{I^i_t + m^i_t}{N^i_t} \right) \right) \right) \tag{2.1} \]

where $\alpha_t = \alpha_1 + \alpha_2 \mathcal{I}_{(\tau_1^\alpha, \tau_2^\alpha)}(t)$ is an infection rate that is independent of current incidence (including the rate at which asymptomatic carriers transition to meningitis disease), and $\beta_t = \beta_1 + \beta_2 \mathcal{I}_{(\tau_1^\beta, \tau_2^\beta)}(t)$ controls the component of the infection rate that is dependent on current incidence (i.e. direct transmission), $t = 1, \cdots, T$ and $i = 1, \cdots, K$. Note that here and in the following, the function $\mathcal{I}_{(t_1, t_2)}(t)$ is an indicator function for the interval $[t_1, t_2]$, that is, $\mathcal{I}_{(t_1, t_2)}(t) = 1$ if $t \in [t_1, t_2]$ and $\mathcal{I}_{(t_1, t_2)}(t) = 0$ otherwise. This indicator function identifies periods of time where either the incidence independent ($\alpha_t$) or incidence independent ($\beta_t$) is elevated due to seasonal (assumed to be environmental) factors. Parameters $(\alpha_1, \alpha_2, \beta_1, \beta_2)$ are all non-negative real numbers, while $(\tau_1^\alpha, \tau_2^\alpha)$ and $(\tau_1^\beta, \tau_2^\beta)$ take on integer values from 0 to $T$. Note here that since the ranges for the parameters $\alpha_1$ and $\alpha_2$ are very small, we use re-parametrizations $\alpha'_1 = \log_{10}(-\alpha_1)$ and $\alpha'_2 = \log_{10}(-\alpha_2)$ throughout the section.

The susceptibles are modeled as follows

\[ S^i_{t+1} = (S^i_t - I^i_t)(1 - V^i_t), \tag{2.2} \]

where $t = 1, \cdots, T$ and $i = 1, \cdots, K$, reflecting how susceptibles are depleted by infection and vaccination ($V^i_t$). Here we assume that the vaccination is a known covariate.
We describe the number of moving infectious individuals in community $i$ at time $t$ by

$$m_t^i = \gamma_t \sum_{j \neq i} \frac{I_t^j}{d_{ij}},$$

(2.3)

where $d_{ij}$ is the distance between communities $i$ and $j$ and $\gamma_t = \gamma_1 + \gamma_2 \mathcal{I}(\tau_1^\gamma,\tau_2^\gamma)(t)$ represents the migration rate. Note that while migration between $i$ and $j$ is assumed to decline with distance apart, the overall migration rate may be seasonal; thus, $\gamma_1 \geq 0$, $\gamma_2 \geq 0$ and $(\tau_1^\gamma,\tau_2^\gamma)$ are integers from the interval $[0,T]$.

Finally, if we denote the number of reported cases by MSF and WHO by $R_t^{\text{MSF},i}$ and $R_t^{\text{WHO},i}$ for community $i$ at time $t$ respectively, we can model these variables by

$$R_t^{\text{MSF},i} \sim \text{Binomial}(I_t^i, p_{\text{MSF}})$$
$$R_t^{\text{WHO},i} \sim \text{Binomial}(I_t^i, p_{\text{WHO}})$$

where $p_{\text{MSF}}$ and $p_{\text{WHO}}$ are the reporting rates for the MSF and WHO surveillance, respectively, and both lie in the interval $(0,1)$.

We note that the model we have developed above is closely related to the compartment models widely used in modeling disease dynamics (cf. Keeling and Rohani, 2008a, Chapter 6). However, this particular formulation in the context of meningitis is new.

As can be seen from Equations (4.1)-(4.3), each of the unknown parameters of the model has a natural interpretation. The parameter $\alpha_2$ is the seasonal increase in density independent infection rate, which would reflect a broad-scale regional increase in incidence of meningitis disease and may recommend a regionally-based control strategy. The parameter $\beta_2$ represents a possible seasonal increase in the incidence-dependent infection rate and $\gamma$s are the parameters related to the movement between neighboring areas. Together these parameters may suggest greater role for local (within and between community) transmission in epidemic progression and would recommend a more localized control strategy; i.e. if $\gamma_1$ is 0, then epidemics cannot be initiated by neighboring areas and may only start due to the spatially uniform regional increase in infection risk.

Using these interpretations of the parameters, we can now write our two important hypotheses about meningitis dynamics as $H_1 = \{\gamma_2 = 0 \text{ and } \alpha_2 > 0\}$ and $H_2 = \{\gamma_2 > 0 \text{ and } \alpha_2 = 0\}$. Our main goal is then to distinguish between the effects of (i) the broadscale
environmental effect (α’s) which is expected to be uniform across communities and therefore induce a spatially random pattern of epidemic emergence, and (ii) transmission, which is expected to scale with the incidence of infection within each community (β’s) and hence generate spatial correlation (spatial aggregation of cases for close locations) in the introduction of epidemics between neighboring communities (γ’s).

2.3.3 Inference

In order for readers to easily follow the model description along with implementation details, we begin with a quick overview of this section. We have described the likelihood function in detail immediately below, and the probability model on which it is based was described previously in Equations (4.1), (4.2), and (4.3). We provide a detailed description of how to construct the ABC-MCMC algorithm in Section 3.1, prior specification details in 3.2, and implementation details in 3.3.

Let us first assume that \( p_{\text{MSF}} \) and \( p_{\text{WHO}} \) are known and \( \{I\} \) can be observed. If we denote all the remaining unknown parameters by \( \Theta \), we can write down the likelihood function for the model,

\[
L(I_{-0}, S_{-0}|I_0, S_0, \Theta) =
\begin{cases}
  \prod_{t=1}^{T-1} \prod_{i=1}^K f_{\text{binom}}(S_t^i, 1 - \exp(-\alpha_t - \beta_t(I_t^i + m_t^i)/N_t^i)) \left[ I_t^i \right], & \text{if } S_{t+1}^i = (S_t^i - I_t^i)(1 - V_t^i) \forall i, t \\
  0, & \text{otherwise},
\end{cases}
\]

where \( f_{\text{binom}}(n, p)[k] = \frac{n!}{k!(n-k)!} p^k (1 - p)^{n-k} \) is the probability mass function of the binomial distribution and \( I_{-0} = \{I_t^i \}_{t=1, \ldots, T} \), \( S_{-0} = \{S_t^i \}_{t=1, \ldots, T} \), \( I_0 = \{I_0^i \}_{i=1, \ldots, K} \) and \( S_0 = \{S_0^i \}_{i=1, \ldots, K} \).

Evaluation of this likelihood function for each parameter setting is extremely expensive since it involves integration over the unobserved \( \{S\} \), which has dimension \( K \times T \). For our data example in Section 4.6.1, \( K \times T = 34 \times 37 = 1,258 \). If we remove the assumption that \( p_{\text{MSF}} \) and \( p_{\text{WHO}} \) are known and assume that \( \{I\} \) are only observed indirectly via \( R_{\text{MSF}} \) and \( R_{\text{WHO}} \), we will double the dimensionality of the unobserved variables making the likelihood evaluations even more expensive. Therefore inference based on traditional maximum likelihood or Bayesian methods for this model is computationally infeasible. Hence, we propose an alternative approach to inference using approximate Bayesian computation.
2.3.3.1 Approximate Bayesian Computation

In a Bayesian framework, we are interested in posterior distribution of $\Theta$ given the observed data $D$. In other words, we are interested in studying

$$f(\Theta|D) \propto L(D|\Theta)\pi(\Theta),$$

where $L(D|\Theta)$ is the likelihood function of the model. If the model is a set of probability distributions $\{\pi(y|\Theta)\}$, our data $D$ is a realization from $\pi(y|\Theta)$ for some unknown $\Theta$ and $L(D|\Theta) = \pi(D|\Theta)$. For any given $\Theta'$, we can simulate data $D'$ from $\pi(y|\Theta')$. If $D = D'$, $\Theta'$ is a draw from the posterior likelihood $L(\Theta|D) = f(\Theta|D)$. Since an exact match of the simulated and the observed data will not occur in most realistic settings, Pritchard et al. (1999) suggests using a suitably selected distance function $d(D', D)$ between $D'$ and $D$. Using $d(D', D)$, it is easy to generate draws from the approximate posterior $f(\Theta|d(D', D) < \epsilon)$, where $\epsilon$ can be thought of as a tolerance level. The smaller the $\epsilon$ used the more accurate the approximation to true posterior distribution sampling, with the tradeoff being that far fewer samples are returned when the tolerance is set to be very small.

To infer the parameters of our model based on samples from $f(\Theta|d(D', D) < \epsilon)$, we use an ABC algorithm introduced by Marjoram et al. (2003) and Toni et al. (2009). A more efficient variant of this algorithm is carefully studied in McKinley et al. (2009). It is similar to the standard Metropolis-Hasting’s algorithm and can result in high acceptance rates given a good distance function and a tolerance level. Determining an appropriate distance and tolerance level for ABC has been a subject of recent research (cf. Fearnhead and Prangle, 2010).

At each iteration of the algorithm, we do the following. Suppose that the most recent value sampled is $\Theta^{(i)}$. We draw a new candidate parameter $\Theta'$ from the proposal distribution $q(\cdot)$ and calculate

$$\rho = \min \left(1, \frac{\hat{f}(D|\Theta')\pi(\Theta')q(\Theta^{(i)}|\Theta')}{\hat{f}(D|\Theta^{(i)})\pi(\Theta^{(i)})q(\Theta'|\Theta^{(i)})} \right),$$

where

$$\hat{f}(D|\Theta') = \frac{1}{R} \sum_{r=1}^{R} I(d(D'_r, D) < \epsilon)$$

and $\pi(\cdot)$ is the prior distribution of the parameters. We set $\Theta^{(i+1)} = \Theta'$ with probability $\rho$, and $\Theta^{(i+1)} = \Theta^{(i)}$ with probability $1 - \rho$. 
2.3.3.2 Prior Specification

Our choice of these priors is pragmatic: Based on our study of multiple simulated examples with various parameter settings we found parameter settings outside a certain range of values resulted in data that were unrealistic. Based on these ranges of viable parameter values, we assume uniform priors on the interval $[0, 2]$ for the continuous parameters $\beta_1, \beta_2, \alpha_1$ and $\alpha_2$. For the parameters $\gamma_1$ and $\gamma_2$, we use a mixture of the degenerate distribution at 0 and the uniform distribution on the interval $[0, 2]$ with equal mixing probabilities. As a prior distribution for the reporting rates, $p_{\text{MSF}}$ and $p_{\text{WHO}}$, we use Uniform$[0, 1]$. We also assume that each pair of the discrete parameters $(\tau_{\beta_1}, \tau_{\beta_2}), (\tau_{\alpha_1}, \tau_{\alpha_2})$ and $(\tau_{\gamma_1}, \tau_{\gamma_2})$ is distributed uniformly on the upper left triangle of the square $[0, T] \times [0, T]$ to insure that the lower bound is no greater than the upper bound.

2.3.3.3 Computational Details

We now describe some of the details for our ABC-MCMC algorithm. For the continuous parameters, we use a mixture of the Uniform$[\nu^{\text{(current)}} - 0.2, \nu^{\text{(current)}} + 0.2]$ and the prior distribution with mixing probabilities $p_1 = 0.7$ and $p_2 = 0.3$ as a proposal distribution, where $\nu = \beta_1, \beta_2, \alpha_1, \alpha_2, \gamma_1, \gamma_2$ and ‘current’ means the current value of the parameter. These proposal distributions were chosen by trial and error to produce an efficient Markov chain Monte Carlo algorithm.

For each pair of the discrete parameters, the proposal distribution is a mixture of two distributions with probabilities $p_1 = 0.7$ and $p_2 = 0.3$ as well. This time, however, the first distribution is distributed uniformly on the upper left triangle of the smaller square $[(\tau_{\nu}^{\text{(current)}})^{\text{(current)}} - 3, (\tau_{\nu}^{\text{(current)}})^{\text{(current)}} + 3] \times [(\tau_{\nu}^{\text{(current)}})^{\text{(current)}} - 3, (\tau_{\nu}^{\text{(current)}})^{\text{(current)}} + 3]$ and the second distribution is equal to the prior for $(\tau_{\nu_1}, \tau_{\nu_2})$, where $\nu = \beta, \alpha, \gamma$.

We use Euclidean distance as our distance function in the construction of the approximate likelihood $\hat{f}(\cdot)$. When we assume that reporting rates $p_{\text{MSF}}$ and $p_{\text{WHO}}$ are known and we have data for $\{I\}$, we calculate the $\hat{f}(\cdot)$ using Equation (4.4). When $p_{\text{MSF}}$ and $p_{\text{WHO}}$ are not known and we observe $\{I\}$ via sources of information from MSF ($R_{\text{MSF}}$) and WHO ($R_{\text{WHO}}$), we construct our approximate likelihood function as

\[
\hat{f}(D|\Theta') = \frac{1}{R} \sum_{r=1}^{R} \mathcal{I}(d(D'_r, D_{\text{MSF}})<\epsilon) \mathcal{I}(d(D'_r, D_{\text{WHO}})<\epsilon),
\] (2.5)
where \( D = (D_{\text{MSF}}, D_{\text{WHO}}) \), \( D_{\text{MSF}} \) and \( D_{\text{WHO}} \) are the data from MSF and WHO respectively. This construction is similar to using a composite approximate likelihood function described in Ratmann et al. (2007) and Pritchard et al. (1999). Based on our simulated data examples in Section 2.3.3.4, this approach results in better inference with confidence regions for the parameters that are much narrower than confidence regions obtained by using Equation (4.4). Both in our simulated examples and the analysis of the original data, we used Markov chains of length 200,000 for the parameters, which was adequate for producing posterior estimates with small Monte Carlo standard errors (Flegal et al., 2008; Jones et al., 2006). Based on a study of different tolerance levels \( \epsilon \) and their effect on our inference, we chose to use \( \epsilon = 350 \) in simulated data examples. For the original data, this value was set to 900. In each case, \( \epsilon \) was obtained heuristically, striking a balance between having a reasonable MCMC acceptance rate and keeping the value of \( \epsilon \) small enough so that the original and simulated datasets (part of the Metropolis-Hastings accept-reject step) are reasonably close.

Our MCMC standard error calculation assumes that the Markov chain converges to the target distribution. As such it is a way of controlling the simulation standard error which, in theory, can be driven down to 0. It is not, however, a way of controlling the approximation error introduced by the tolerance level(\( \epsilon \)) used in the ABC algorithm. The latter is difficult to problem and no current approach exists for estimating it. Furthermore, running our algorithm for much longer has no impact on this error. Our use of MCMC standard error is therefore a “diagnostic” in that it is an approach for determining a reasonable length for the Markov chain.

2.3.3.4 Application to Simulated Data Examples

In this section, we consider two important simulated examples. In the first example, we make the movement parameters \( \gamma_1 \) and \( \gamma_2 \) very small so that outbreaks are unlikely to spread from neighboring areas. At the same time, we assume that the environmental parameters \( \alpha_1 \) and \( \alpha_2 \) are large; suggesting a broad-scale regional increase in the transition from carriage to disease.

In the second example, we assume that the environmental effects are small and make the movement parameters larger so that our model generates dynamics where outbreak are initiated due to the importation of the infection from neighboring communities. We note here that when both the movement and environmental effects are important factors in sparking an outbreak (i.e. both \( \gamma \)'s and \( \alpha \)'s are relatively large) the model generates dynamics that
are very different from the observed data; the percentages of incidences of meningitis in these simulations are higher than 8% of the population for many communities. The actual percentage of cases of meningitis is less than 1% even for the worst epidemics in these regions. (Harrison et al., 2009).

To clarify our approach for simulation: We generate the simulated datasets under the two different hypotheses defined in Section 4.1, except with minor modifications. Rather than setting $\gamma_2$ identically to 0 (as in $H_1$), we set it equal to a small non-zero positive value and rather than setting $\alpha_2$ identically to 0 (as in $H_2$), we set it equal to a small non-zero positive value. These small non-zero values allow for more realistic parameter settings than those obtained when setting them identically to 0. Of course it is possible in principle to simulate data using other combinations of parameter settings. However, as mentioned above, both $(\alpha_2, \gamma_2)$ small and both large are not reasonable combinations as they result in very unrealistic data – not enough outbreaks and too many outbreaks respectively. Both $(\alpha_2, \gamma_2)$ moderate is possible but is not of interest from a public health/epidemiological perspective; it is therefore a case we did not consider.

In our first example, we simulated data using the model described in Section 2.3.2 with parameters $(\beta_1, \beta_2) = (0.4, 0.4), (\tau_{\beta_1}^\beta, \tau_{\beta_2}^\beta) = (6, 13), (\gamma_1, \gamma_2) = (0.1, 0.1), (\tau_{\gamma_1}^\gamma, \tau_{\gamma_2}^\gamma) = (2, 13), (\alpha_1, \alpha_2) = (0.48, 0.45) and (\tau_{\alpha_1}^\alpha, \tau_{\alpha_2}^\alpha) = (5, 15)$. These parameter values produced biologically realistic data that resembled the real observations. Based on our preliminary analysis of simulated data, values of $\gamma_1$ and $\gamma_2$ here were small enough not to spark an outbreak in the absence of the incidence independent climate effects. The dimensionality of the simulated data was also chosen to be equal to the dimensions of the original data matrix, with $K = 34$ and $T = 37$. Our goal here is to see how well our approach can recover the true parameter values used to simulate the data.

In Figure 2.1, we plot the posterior density function for $(\beta_1, \beta_2)$ along with the histogram for discrete $(\tau_{\beta_1}^\beta, \tau_{\beta_2}^\beta)$ (calculated from samples obtained from the corresponding posterior distributions). In Figures 2.1(a)-(d), we can see that the true values of the parameters are located very close to the global maxima of the density or to the peak of the histogram. This means that data contain information that allows us to learn about these four parameters. Note, however, that the 95% credible regions, showed by two vertical dashed lines, are quite wide. These regions are obtained using the highest posterior density approach (Chen et al., 2000).

In Figure 2.2, we study the posterior densities and histograms for $(\alpha_1, \alpha_2, \tau_{\alpha_1}^\alpha, \tau_{\alpha_2}^\alpha)$. We
Figure 2.1: Inference for \((\beta_1, \beta_2, \tau_1^\beta, \tau_2^\beta)\): (a) Posterior density for \(\beta_1\) when the true value of \(\beta_1 = 0.4\); (b) Posterior density for \(\beta_2\) when the true value of \(\beta_2 = 0.4\); (c) Sample histogram for \(\tau_1^\beta\) when the true value of \(\tau_1^\beta = 6\); (d) Sample histogram for \(\tau_1^\beta\) when the true value of \(\tau_2^\beta = 13\).

Figure 2.2: Inference for \((\alpha_1, \alpha_2, \tau_1^\alpha, \tau_2^\alpha)\) when the incidence independent environmental effects can spark outbreaks: (a) Posterior density for \(\alpha_1\) when the true value of \(\alpha_1 = 0.48\); (b) Posterior density for \(\alpha_2\) when the true value of \(\alpha_2 = 0.45\); (c) Sample histogram for \(\tau_1^\alpha\) when the true value of \(\tau_1^\alpha = 5\); (d) Sample histogram for \(\tau_1^\alpha\) when the true value of \(\tau_2^\alpha = 15\).
can see again the data are informative regarding these four parameters (bold vertical lines coincide with the maxima) and conclude that it is possible to learn about \((\alpha_1, \alpha_2, \tau_1^\alpha, \tau_2^\alpha)\).

Figure 2.3: Inference for \((\gamma_1, \gamma_2, \tau_1^\gamma, \tau_2^\gamma)\) when the transmission of the infection cannot spark outbreaks: (a) Posterior density for \(\gamma_1\) when the true value of \(\gamma_1 = 0.1\); (b) Posterior density for \(\gamma_2\) when the true value of \(\gamma_2 = 0.1\); (c) Sample histogram for \(\tau_1^\gamma\) when the true value of \(\tau_1^\gamma = 2\); (d) Sample histogram for \(\tau_1^\gamma\) when the true value of \(\tau_2^\gamma = 13\).

The posterior densities and posterior histograms for \((\gamma_1, \gamma_2, \tau_1^\gamma, \tau_2^\gamma)\) are in Figure 2.3. In Figure 2.3, we see that the posterior density functions for \((\gamma_1, \gamma_2)\) are flat and the posterior maxima does not reflect the true parameter values. This indicates that the data may not have sufficient information about these parameters. Similarly, we can see that the histograms for \((\tau_1^\gamma, \tau_2^\gamma)\) in Figure 2.3 (c)-(d) seem to have more than one peak and therefore fail to reveal the true values of the parameters.

In the second example, we simulated data with parameters \((\beta_1, \beta_2) = (0.4, 0.4), (\tau_1^\beta, \tau_2^\beta) = (6, 13), (\gamma_1, \gamma_2) = (1.5, 1.5), (\tau_1^\gamma, \tau_2^\gamma) = (2, 13), (\alpha_1, \alpha_2) = (1, 1.5)\) and \((\tau_1^\alpha, \tau_2^\alpha) = (5, 15)\). In these simulations, outbreaks of meningitis can only be started by the movement from neighboring areas and the climate parameters \(\alpha_1\) and \(\alpha_2\) are too small to have a noticeable effect on the dynamics of the disease.
Since we used the same values for $\beta$'s, inference for $(\beta_1, \beta_2)$ and $(\tau_1^\beta, \tau_2^\beta)$ for this example was similar to the one obtained in the previous simulated example in Figure 2.1.

The posterior densities and histograms for $(\gamma_1, \gamma_2, \tau_1^\gamma, \tau_2^\gamma)$ are in Figure 2.4. In contrast to previous example, for this simulated data, we see that we can learn about these parameters and the posterior peaks of the corresponding densities or histograms are located very close to the true values of the parameters. As was noted above, the credible regions are still very wide.

![Figure 2.4: Inference for $(\gamma_1, \gamma_2, \tau_1^\gamma, \tau_2^\gamma)$ when the transmission of the infection can spark outbreaks: (a) Posterior density for $\gamma_1$ when the true value of $\gamma_1 = 1.5$; (b) Posterior density for $\gamma_2$ when the true value of $\gamma_2 = 1.5$; (c) Sample histogram for $\tau_1^\gamma$ when the true value of $\tau_1^\gamma = 2$; (d) Sample histogram for $\tau_1^\gamma$ when the true value of $\tau_2^\gamma = 13$.](image)

Figure 2.5 contains the posterior densities and histograms for $(\alpha_1, \alpha_2, \tau_1^\alpha, \tau_2^\alpha)$. As can be seen in Figures 2.5(a) and (b), the posterior density functions for $\alpha_1$ and $\alpha_2$ are now relatively flat. Therefore, it is clear that for this example, the data are not informative regarding $\alpha_1$ and $\alpha_2$ and we cannot infer these two environmental parameters. In Figures 2.5 (c)-(d), we see that the histograms for $\tau_1^\alpha, \tau_2^\alpha$ are also multimodal and do not reveal the true values of the parameters.
Figure 2.5: Inference for $(\alpha_1, \alpha_2, \tau_1^{\alpha}, \tau_2^{\alpha})$ when the climate effects cannot spark outbreaks: (a) Posterior density for $\alpha_1$ when the true value of $\alpha_1 = 1$; (b) Posterior density for $\alpha_2$ when the true value of $\alpha_2 = 1.5$; (c) Sample histogram for $\tau_1^{\alpha}$ when the true value of $\tau_1^{\alpha} = 5$; (d) Sample histogram for $\tau_2^{\alpha}$ when the true value of $\tau_2^{\alpha} = 15$.

Since our simulated examples resemble the real datasets, our study suggests that in similar settings we may be able to infer those parameters in our model that have a significant effect on starting an outbreak of meningitis. In both examples, our inferential approach also showed that the data do not have sufficient information regarding the parameters that do not have a noticeable effect on the dynamics of meningitis. An important caveat is that our conclusions are, of course, based on an empirical study involving two simulated data examples, and therefore may not extend to a wide variety of scenarios.

Finally, as an illustration of the model fit, we have provided in Figure 2.6 a few representative examples of the predictions overlaid on the real data for a few selected locations for the first simulated date example. As can be seen, the fitted model performs well as is apparent from simulations from the predictive distribution.
2.3.4 Application to Meningitis Data

In this section, we describe our data from MSF and WHO and apply our model and the inferential approach to investigate the key features of meningitis dynamics in Nigeria. In particular, we are interested in separating the effects of incidence-dependent transmission from environmental effects (incidence independent) on meningitis infection rates.

2.3.4.1 Meningitis Data

In 2009, a large meningitis epidemic affected a broad region of northern Nigeria and southern Niger, resulting in greater than 75,000 cases and 4000 deaths. *N. meningitidis* serogroup A was identified as the main causative agent. Here we analyze two independent surveillance datasets from this outbreak in the 34 local government areas (LGAs) of Katsina State, Nigeria. The first comprises weekly counts of cerebrospinal meningitis (CSM) cases that were collected through the Ministry of Health (MoH) surveillance system and reported to the World Health Organization (WHO). The second comprises weekly counts of CSM cases that
were collected by Medecins Sans Frontieres-Operational Center Paris (MSF-OCP) directly from local health centers. In both datasets the case definition of a suspected meningitis was: any person with sudden onset of fever (> 38.5°C rectal or 38.0°C axillary) and one of the following signs: neck stiffness, altered consciousness or other meningeal signs. The MSF surveillance system reported cases per LGA from week 50 of 2008 to week 22 of 2009, a total of 20,617 cases. The WHO data included weekly cases per LGA from week 50 of 2008 to week 37 of 2009, though no cases were reported after week 22, a total of 9,331 cases. LGA and ward population sizes were obtained from Katsina State authorities and are based on the extrapolation of the general population census of the year 2006.

In collaboration with the Federal and the State Ministry of Health, the Primary Health Care Development Agency and the national program of immunization, MSF-OCP conducted vaccination campaigns in 18 of 34 LGAs in Katsina state targeting all individuals between 2 and 30 years of age beginning in week 10 of the year. The administrative vaccine coverage achieved within each ward of each LGA during the 2009 vaccination campaigns were obtained from MSF-OCP. For LGAs where only a fraction of wards were vaccinated, the LGA-level vaccine coverage was calculated as follows: vaccination coverage (%) in the vaccinated wards multiplied by the percentage of wards vaccinated in the LGA.

It is worth clarifying that the WHO and MSF datasets are independent measures by different organizations of the same epidemic process. The WHO reporting system is a system of self reporting from clinics to the national level to the WHO that is designed to help detect outbreaks and involves weekly aggregate counts passed up that hierarchy. The MSF surveillance was done independently and involved field epidemiologists traveling to the clinics themselves and gathering the individual admission sheets for each reported case after the outbreak had occurred. One reason that the reporting rate is higher for these data is that admissions that occurred in a particular week but were recorded after the weekly summary was sent to the national level were often missed in the WHO system. The MSF system is thus the far more complete record of the outbreak. Alternatively, in overburdened clinics, aggregate counts may have been made, but not recorded on individual admission sheets, leading to disparities in the MSF surveillance.

While these two datasets may not be strictly statistically independent (as they are two measures on the same process) we submit that there are enough differences in the collection and reporting of these data to treat them roughly as such.
2.3.4.2 Results

We now fit the model to the data described in Section 4.6.1. In Figure 2.7, we plot the posterior densities and histograms for $\beta_1, \beta_2, \tau_1^\beta$ and $\tau_2^\beta$. Using the mode of the density function, we estimate $\hat{\beta}_1 = 0.46, \hat{\beta}_2 = 0.45, \tau_1^\beta = 3$ and $\tau_2^\beta = 13$. The 95% credible regions for $\beta_1$ and $\beta_2$ are indicated by two vertical dashed lines.

![Figure 2.7: Inference for ($\beta_1, \beta_2, \tau_1^\beta, \tau_2^\beta$) for the actual data: (a) Posterior density for $\beta_1$; (b) Posterior density for $\beta_2$; (c) Sample histogram for $\tau_1^\beta$; (d) Sample histogram for $\tau_2^\beta$.](image)

In Figure 2.8, we infer $\alpha_1, \alpha_2, \tau_1^\alpha$ and $\tau_2^\alpha$. Our estimates for these parameters are: $\hat{\alpha}_1 = 0.49, \hat{\alpha}_2 = 0.4, \tau_1^\alpha = 7$ and $\tau_2^\alpha = 15$. Note here that the histogram $\tau_2^\alpha$ has two peaks located in the interval $[13, 37]$ meaning that this parameter is only very weakly informed by the data.

Inference for the unknown movement parameters ($\gamma_1, \gamma_2, \tau_1^\gamma, \tau_2^\gamma$) from the data was similar to inference from the data in our first example in Section 2.3.3.4 in Figure 2.3. Posterior densities and histograms were relatively flat or multimodal indicating that the data did not have information regarding these parameters. This implied that our data followed a pattern of the first example and suggested that the parameters controlling the movement of the infection between neighboring communities were small.
Finally, we infer $p_{\text{MSF}}$ and $p_{\text{WHO}}$ in Figure 2.9. Our estimates based on the modes of the posterior density functions are: $\hat{p}_{\text{MSF}} = 0.72$ and $\hat{p}_{\text{WHO}} = 0.26$ with 95% credible regions (0.19, 0.993) and (0.002, 0.841) respectively. This difference is not surprising as the WHO surveillance system was implemented in real time during the epidemic and the MSF surveillance was conducted retrospectively. As such, any lags in the recording of cases would have been resolved in time to be recorded by the MSF surveillance system.

These results confirm a strong seasonal component to the risk of meningitis disease, consistent with the established correlation between meningitis incidence and the dry season. The direct transmission parameters, $\beta_1$ and $\beta_2$, were significantly greater than 0 and there was strong support for an increase in transmission over weeks 3 to 13. This suggests that the overall attack rate scales with the incidence of infection and the number of susceptibles, consistent with direct transmission within communities. The effect of transmission of the infection from neighboring communities in these data was similar to the one observed in the first example in Section 2.3.3.4 and was not noticeable at any point during the epidemic. This indicates little or no role for local transmission in generating spatial pattern at the
Table 2.1: Interpretation of the Bayes factor (adapted from Kass and Raftery (1995))

<table>
<thead>
<tr>
<th>Bayes factor BF</th>
<th>evidence against $M_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 − 3</td>
<td>very weak</td>
</tr>
<tr>
<td>3 − 20</td>
<td>positive</td>
</tr>
<tr>
<td>20 − 150</td>
<td>strong</td>
</tr>
<tr>
<td>&gt; 150</td>
<td>very strong</td>
</tr>
</tbody>
</table>

regional scale as would be expected by an epidemic spreading in a wave across the region. The parameters $\alpha_1$ and $\alpha_2$ were both significantly greater than 0, as was $\tau_1^\alpha$, indicating a general increase in the rate of meningitis disease during the dry season that was independent of the local incidence. This is consistent with a hypothesis of a generalized regional increase in the rate of transition from carriage to invasive meningitis. We speculate that the fact that $\tau_2^\alpha$ was not well identified may indicate that the end of epidemics are driven more by the reduction in susceptible individuals and a decrease in direct transmission rather than by a decline in the transition from carriage to invasive disease.

Taken together, these patterns suggest that the dry season is coincident with both an increase in direct transmission and increased transition to invasive disease at the community scale. The spatial distribution of epidemics is more consistent with a broad scale regional increase in incidence rate rather than wave-like spread from community to community. This distinction is significant in that it suggests that the occurrence of an outbreak locally is somewhat informative of the likelihood of outbreaks elsewhere in the region, but not particularly informative of elevated outbreak risk in neighboring areas. Vaccination for meningitis with polysaccharide vaccine has classically been done reactively after local incidence crosses a threshold value. These results suggest that prioritization of LGAs for vaccination campaigns should be assessed at a regional scale (e.g. the whole of Katsina state) as the proximity of outbreaks at the local LGA scale is not strongly predictive of outbreak risk among neighbors.

In general, the plausibility of two models $M_1$ and $M_2$ given the observed data $D$ can be assessed by the Bayes factor $BF$ that is given by

$$BF = \frac{P(D|M_1)}{P(D|M_2)}.$$  

A value of $BF > 1$ means that $M_1$ is more strongly supported by the data than $M_2$. However, when $M_2$ is a null hypothesis, $BF$ can be interpreted as the amount of evidence against $M_2$ in favor of an alternative hypothesis $M_1$ (see Table 2.1 for its interpretation).
We consider $M_1 = \{\text{a model under } H_1\}$ and $M_2 = \{\text{a model under } H_2\}$. In the context of our inferential approach, we include a model indicator as a new parameter $\mu$ in the sampling algorithm with a prior $P(\mu = M_1) = P(\mu = M_2) = 0.5$. The Bayes factor can then be estimated using the sample for $\mu$ (following Didelot et al. (2010)):

$$\hat{BF} = \frac{\sum_{\mu = M_1} 1}{\sum_{\mu = M_2} 1}.$$ 

For the meningitis dataset, our estimate of the Bayes factor $\hat{BF}$ was around 1.7. Even though this indicates that $M_1$ explains the data better, since $\hat{BF} < 3$, according to Table 1 (Kass and Raftery, 1995), this evidence is weak.

We note here that approximate Bayes factors calculated based on ABC may lead to incorrect conclusions in some situations. Such problems arise when inappropriate summary statistics are used in the ABC algorithm. In our approach, however, we consider the whole dataset rather than summary statistics. Hence, the Bayes factors we report should be reasonable. Additionally, we calculated the approximate deviance information criteria as described in Francois and Laval (2011) to distinguish $H_1$ from $H_2$. The obtained conclusions were not different from the ones we found using the Bayes factors.

2.3.5 Discussion

In the absence of detailed longitudinal surveillance of $N. meningitidis$ carriage, the relative role of direct transmission and environmental conditions in driving patterns of local and regional meningitis incidence have been a challenge to identify. Here, we have developed a statistical model to investigate the effects of incidence-dependent local transmission and incidence independent regional climate on the incidence of meningitis in Nigeria. Using the data described in Section 4.6.1, we infer the unknown parameters of our model using an ABC approach based on the approximate posterior distributions. We have shown that the local dynamics of meningitis epidemics tend to scale with incidence, which is characteristic of a role of direct transmission. However, the broad-scale regional pattern is more consistent with both an elevated transmission rate and an increased risk of severe disease, suggesting the role of regional environmental conditions rather than direct transmission among neighboring communities, in correlating local outbreak risk. This finding has implications for both
the general understanding of the role of transmission in meningitis epidemiology and the prioritization of communities for reactive vaccination campaigns.

In a more complex model, it is theoretically possible to allow for changes in the prevalence of the asymptomatic carriers that can affect the transmission process of meningitis. This hidden transmission is, however, unobservable. It is also possible to investigate a hypothesis about seasonally forced changes in the proportion of carriers and the new cases of the disease. However, since there is no information for the amounts of carriers, it would be infeasible to distinguish this hypothesis from $H_1$ and $H_2$ based on the data for the reported cases of meningitis.

This challenge of only partial observability is not unique to meningitis. Public health policy decisions frequently have to be made in such challenging situations. Hence, as mentioned in Section 4.1, our goal in this work is more to deal with what management has to do rather than trying to actually uncover all the underlying dynamics/biology of the system since the latter is perhaps not feasible.

We note that there are possible issues with parameter identifiability. One possible way to
reduce these issues would be to assume that the timing of the dry season is known. A number of papers (cf. Molesworth et al., 2003; Mueller and Gessner, 2010; Sultan et al., 2005; Thomson et al., 2006) suggest different approaches to learn about the timing of the underlying environmental drivers in various contexts. These approaches have identified several important climate/environmental correlates of meningitis outbreaks. However, they have failed to identify a definitive predictive covariate to explain the timing of meningitis outbreaks at small spatial scales. Therefore, we believe, that using these methods to obtain values for the periods of seasonal forcing would be ad-hoc and would not quantify correctly our uncertainty about this aspect of the model.

Our model does not account for stochasticity in the population sizes but this is not an issue here. In general, when population sizes of the communities are small, demographic stochasticity can affect the dynamics of the disease as the infection may become locally extinct. In our data, communities have quite large populations (mean = 183,051 and standard deviation = 48,315) and we therefore do not think demographic stochasticity has a role to play.

As an alternative to our Binomial model, we have also fit the model using Poisson and Negative Binomial observation models. This did not result in any difference in the final inference about the parameters of interest.

Our ABC-based approach is easy to implement and computationally expedient. While there may not be enough information about certain parameters in the data (the $\alpha$, $\gamma$ parameters, depending on the setting), our ABC approach appears to provide reliable inference for the well-informed parameters in the model. Based on our simulated examples it does not appear to be possible to make quantitative statements about the magnitude of the less important effect (where effect relates to either regional correlation or local transmission). These examples show that we can, however, still learn about the magnitude of the predominant effect. Our method in principle may hence be useful for disentangling the qualitative differences between the two hypotheses presented in Section 4.1. However, we find that for the real data there is only weak evidence in favor of one of the hypotheses ($H_1$) versus the other ($H_2$).

Finally, it is important to note that our conclusions are based on ABC-MCMC, which is an approximate approach. For instance, different tolerance levels (different $\epsilon$ values) can result in different inferential results (regardless of how long the ABC-MCMC algorithm is run) even though the underlying model remains unchanged. Based on our own experimentation
with varying tolerance levels, however, we are fairly confident that our inferential results are reasonable and would be fairly close to results based on a standard MCMC algorithm.
Chapter 3

Gaussian Processes and a Model for Gypsy Moth Populations

Our inferential approach described in the following chapters uses Gaussian processes extensively. In this chapter, therefore, I will be giving a brief overview of Gaussian processes, different covariance matrices, and separability assumptions based on Kronecker products. I will also describe how Gaussian processes are used in the area of computer model emulation on simple examples. Finally, in the context of a challenging real-world inferential problem, I will show how Gaussian processes can be used to model spatio-temporal dynamics of an insect called the gypsy moth to study its population periodicities. This analysis is the basis of a paper titled “A latent Gaussian process model for inferring periodicities of gypsy moth populations”.

3.1 Overview of Gaussian Processes

In this section, I will provide a brief overview of Gaussian processes as used for modeling space-time data and complex computer models. A more detailed discussion of Gaussian processes in these contexts are available in Cressie (1993), Banerjee et al. (2004a) and Stein (1999a).

A stochastic process with multidimensional index space, a “random field”, \( \{X(s), s \in E \subset \mathbb{R}^d\}, d \geq 1 \) is defined to be strictly stationary if for any \( h \in E \), \( (X(s_1), \cdots, X(s_k)) \) and \( (X(s_1 + h), \cdots, X(s_k + h)) \) have the same distribution for all finite \( k > 0 \). In other words, a
A stochastic process is strictly stationary if it is invariant to a constant shift in location. The process is defined to be *weakly stationary* if \( E[X(s)] = \mu \) and \( \text{Cov}[X(s), X(s+h)] = C(h) \) for some constant \( \mu \) and a function \( C(\cdot) \). Here, \( C(h) \) is called the covariance function of the process. It is easy to see that the condition of weak stationarity is weaker than the strict stationarity. The weak stationarity is also sometimes called *second order stationarity*. Weakly stationary processes are very popular because of their simplicity, interpretability, and, in particular, because a number of relatively simple parametric forms that are available as candidates for the covariance functions.

A weakly stationary stochastic process with multidimensional index space \( \{X(s), s \in E \subset \mathbb{R}^d\} \) is called a Gaussian process if for any finite \( k > 0 \) and numbers \( s_1, \ldots, s_k \in E \), \( (X(s_1), \ldots, X(s_k)) \) is a \( k \)-dimensional multivariate normal random variable with a finite mean vector and covariance matrix. In other words, a Gaussian process is an infinite dimensional stochastic process for which any finite dimensional subset has a multivariate normal distribution.

Gaussian processes are very flexible models for both dependent processes, especially space-time processes (cf. Cressie, 1993), and for approximating highly non-linear functions. For Gaussian processes, all marginal and conditional distributions are easily obtained. They have been widely used in the literature on emulation and calibration of complex computer models (cf. Bayarri et al., 2007a; Bhat et al., 2010a; Sacks et al., 1989a), mostly in the realm of deterministic equations, and in machine learning (cf. Rasmussen and Williams, 2005).

### 3.1.1 Gaussian Processes for Spatiotemporal Data

Spatial data are data that contain information about both the response variable and the locations of the measurements. These data are common in meteorology, ecology, and epidemiology. A spatial process can be used to model spatial data. A spatial process is defined as \( \{Z(s) : s \in E \subset \mathbb{R}^d\} \), where \( Z \) is the observation, and \( s \) is the location of the observation (Cressie, 1993). Spatial data are called geostatistical data if the locations where data are observed are pointwise and the domain \( E \) is continuous and fixed.

An interpolation at locations where data are unavailable is one of the main problems of geostatistics. Since spatial measurements are often spatially dependent, Gaussian processes offer a great tool to model geostatistical data. Generally, when the distance between locations increases, the dependence between the measurement from those locations tend to decrease.
Therefore, the covariance functions used in Gaussian processes provide an easy approach to account for this dependence. The concepts of distance here can obviously be extended from physical locations to more abstract notions of distance, such as distance between calibration parameter settings or distance in space and time.

Let the spatial process at location $s \in E$ be defined as

$$Z(s) = X(s)\beta + w(s), \text{ for } s \in E, \quad (3.1)$$

where $X(s)$ is a set of $p$ covariates associated with each location $s$, and $\beta$ is a $p$-dimensional vector of coefficients. Spatial dependence is incorporated by modeling $\{w(s) : s \in E\}$ as a zero mean stationary Gaussian process. If $w = (w(s_1), \ldots, w(s_n))^T$ and $\xi$ are the parameters of the model, then

$$w \mid \xi \sim N(0, \Sigma(\xi)), \quad (3.2)$$

where $\Sigma(\xi)$ is the symmetric and positive definite covariance matrix of $w$. One commonly used covariance function is the ‘exponential’ covariance with parameters $\xi = (\psi, \kappa, \phi)$, (with $\psi, \kappa, \phi > 0$), which has the form $\Sigma(\xi)_{ij} = \psi I(i = j) + \kappa \exp(-\|s_i - s_j\|/\phi)$, where $I$ is the indicator function and $\|s_i - s_j\|$ is the Euclidean distance between locations $s_i, s_j \in E$. This model is interpreted in the following manner: $\psi$ (often called the nugget) is the non-spatial error associated with each location, and $\kappa$ (called the sill) and $\phi$ (called the range) are parameters that define the scale and range of the spatial dependence respectively. The exponential covariance is a special case of the Matérn family (Handcock and Stein, 1993). We discuss this and other popular covariance functions in Section 3.1.2.

Expanding the concept of the linear Gaussian processes to temporal dependence or dependence of other forms is not difficult if there is no interaction between the different types of dependence (such as space and time). When this is the case, we can assume a separable covariance function using Kronecker products:

$$\text{Cov}(X(s, t), X(s + h, t + k)) = C_s(h)C_t(k).$$

We discuss details and properties of Kronecker products in Section 3.1.3.

Gneiting (2002) has determined a parametric form for a nonseparable space-time covariance functions whose separability depends on a single parameter being non-zero. In a Bayesian framework, one only needs to determine whether the credible region for this pa-
rameter includes zero to determine whether the covariance is separable. Generally, however, unless there is evidence of nonseparability, we assume a separable covariance function.

### 3.1.2 Covariance Functions

In this section, we describe different popular covariance functions that can be used to model dependent data.

**Spherical Covariance Function**

The spherical covariance function has the following form:

\[
C(h) = \begin{cases} 
0, & \text{if } h \geq \frac{1}{\phi} \\
\sigma^2 \left[1 - \frac{3}{2} \phi h + \frac{1}{2} (\phi h)^3\right], & \text{if } 0 < h \leq \frac{1}{\phi} \\
\tau^2 + \sigma^2, & \text{otherwise}
\end{cases}
\]

This function is very popular and widely used largely due to the fact that it offers clear illustrations of the nugget (\(\tau^2\)), sill (\(\tau^2 + \sigma^2\)), and range (1/\(\phi\)), three characteristics traditionally associated with covariance functions.

**Exponential Covariance Function**

The exponential covariance function has the following form:

\[
C(h) = \begin{cases} 
\sigma^2 \exp(-\phi h), & \text{if } h \geq 0 \\
\tau^2 + \sigma^2, & \text{otherwise}
\end{cases}
\]

Since exponential covariance function is always positive when \(\tau \neq 0\), the actual range for this function is infinite. However, for practical reasons, one can use a notion of an effective range, i.e., the distance at which there is essentially no lingering spatial correlation. This can be defined as the distance at which this correlation is dropped to only 0.05. With exponential covariance function, the nugget \(\tau^2\) is often viewed as a “nonspatial effect variance” while the partial sill \(\sigma^2\) is interpreted as a “spatial effect variance”.
Gaussian Covariance Function

The Gaussian covariance function has the form:

\[ C(h) = \begin{cases} 
\sigma^2 \exp(-\phi^2 h^2), & \text{if } h \geq 0 \\
\tau^2 + \sigma^2, & \text{otherwise} 
\end{cases} \]

The Gaussian covariance function is a very popular choice for smooth data. This function results in infinitely differentiable stochastic process. In many cases, however, it is very difficult to estimate smoothness from the data and therefore justify appropriateness of the Gaussian covariance function, especially when there is little information about dependence at locations which are very close to each other.

The Matérn Covariance Function

The Matérn covariance function is defined as follows:

\[ \text{Cov}(x; \psi, \kappa, \phi, \nu) = \begin{cases} 
\frac{\kappa^2 \nu}{2^{\nu-1} \Gamma(\nu)} (2\nu^{1/2} x/\phi)^\nu K_\nu(2\nu^{1/2} x/\phi) & \text{if } x > 0 \\
\psi + \kappa & \text{if } x = 0 
\end{cases} \]

where parameters \( \psi, \kappa, \phi, \) and \( \nu \) are all positive. Here, \( K_\nu(x) \) is a modified Bessel function of order \( \nu \) (Abramowitz and Stegun, 1964), where \( \nu \) determines the smoothness of the process. With Matérn covariance function, the smoothness of the process increases as \( \nu \) increases. Two popular covariance functions described above are part of the Matérn family: the exponential covariance function \((\nu = 0.5)\), and the Gaussian covariance function \((\nu \to \infty)\).

3.1.3 Kronecker Products

When we need to explore the posterior likelihood surface of the parameters of interest, we can use Markov chain Monte Carlo (MCMC) methods to simulate from the posterior distributions. The Metropolis-Hastings algorithm provides a general recipe for constructing a Markov chain with a target distribution \( \pi \). Under fairly general conditions, the ergodic theorem allows for the states of this Markov chain to be treated as samples from \( \pi \) in the steady-state (Hastings, 1970; Metropolis et al., 1953). In this work, for Gaussian process models, we typically use Metropolis random walk updates for the conditional distribution of
each of the parameters. At times, however, when we have a multimodal target distribution, using a combination of Gibbs and Metropolis-Hastings may not result in an algorithm that samples from all the modes of the posterior distribution in an efficient or accurate manner. In such circumstances, we use a slice sampling algorithm (Neal, 2003b), which allows for sampling from the target distribution throughout its support. We note, however, that slice sampling approach may be computationally expensive.

When one uses a Gaussian process for spatial modeling, the main bottleneck is often related to obtaining inversions of the covariance matrix, which requires $O(n^3)$ computations. One of the approaches that may provide significant computational gains is based on Kronecker products. Using Kronecker products helps to reduce the size of matrix manipulations and therefore decreases the required computational time drastically. Kronecker products are especially useful with separable covariance functions (Banerjee et al., 2004b). The definition of a Kronecker product (denoted by $\otimes$) is given below. Let $A = (a_{ij})$ be a $m \times n$ matrix, $B$ be a $p \times q$ matrix, then the Kronecker product of $A$ and $B$ is a matrix of dimension $mp \times nq,$

$$A \otimes B = \begin{bmatrix} a_{11}B & \cdots & a_{1m}B \\ \vdots & \ddots & \vdots \\ a_{p1}B & \cdots & a_{pm}B \end{bmatrix}.$$ 

Some of the useful properties of Kronecker products are below (see Anderson, 2003). Let $A$, $B$, $C$, and $D$ be matrices, then the following properties hold:

$$\begin{align*}
(A \otimes B) \otimes C &= A \otimes (B \otimes C), \\
(A \otimes B)(C \otimes D) &= (AC) \otimes (BD), \\
(A \otimes B)^{-1} &= (A)^{-1} \otimes (B)^{-1}.
\end{align*}$$

The third property is particularly useful since it reduces the dimensions of the matrices to be inverted, and therefore decreases the computation time substantially.

### 3.1.4 Prediction with Gaussian Processes

If $Z$ is the observed data modeled as a Gaussian process, predictions of the process, $Z^\ast = (Z(s^\ast_1), \ldots, Z(s^\ast_m))^T$, where $s^\ast_1, \ldots, s^\ast_m$ are new locations in $E$, are obtained using the posterior
predictive distribution,

$$\pi(Z^* | Z) = \int \pi(Z^* | Z, \xi, \beta) \pi(\xi, \beta | Z) d\Theta d\beta.$$  

Since $Z$ is a Gaussian process, the joint distribution of $Z$ and $Z^*$ given $\xi, \beta$, the covariance and mean parameters respectively, is described in Equation 3.4 below:

$$\begin{bmatrix} Z \\ Z^* \end{bmatrix} | \xi, \beta \sim N \left( \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix} \right),$$  

(3.4)

where $\mu_1$ and $\mu_2$ are the means of $Z$ and $Z^*$ respectively and $\Sigma_{11}, \Sigma_{12}, \Sigma_{21}, \Sigma_{22}$ are block partitions of the covariance matrix $\Sigma(\xi)$, given $\xi, \beta$. $\mu_1$ and $\mu_2$ are functions of $\beta$ and $\Sigma_{11}, \Sigma_{12}, \Sigma_{21}$, and $\Sigma_{22}$ are functions of $\xi$. Consequently, using multivariate normal theory (Anderson, 2003), $Z^* | Z, \beta, \xi$ is normally distributed with the mean and covariance given in Equation 3.5.

$$E(Z^* | Z, \beta, \xi) = \mu_2 + \Sigma_{21} \Sigma_{11}^{-1} (Z - \mu_1), \text{ and } \text{Var}(Z^* | Z, \beta, \xi) = \Sigma_{22} - \Sigma_{21} \Sigma_{11}^{-1} \Sigma_{12}. \quad (3.5)$$

Therefore, to obtain predictions of the process, we first obtain the estimates $\hat{\xi}, \hat{\beta} \sim \pi(\xi, \beta | Z)$, and then simulate from a multivariate normal density with the mean and covariance matrix in Equation (3.5).

### 3.2 A Latent Gaussian Process Model for Gypsy Moth Populations

The gypsy moth is the most important forest defoliating insect in the northeastern United States. Millions of dollars have been spent to control and eliminate gypsy moth populations to protect hardwood trees. Gypsy moth populations are known to vary from very low densities when they have almost undetectable effect on trees to levels when outbreaks cause large regions of the forest to defoliate. It is also known that these outbreaks occur with periodicities that differ from region to region. In this section, we describe a new statistical method to infer the periodicities of the gypsy moth populations from binary space-time defoliation data. Our method is based on reconstructing unobserved gypsy moth populations, modeling them via a latent Gaussian process. For each location and time, our method allows us to
borrow information from neighboring observations both in space and across time. Using a Markov chain Monte Carlo approach we show that it is possible to obtain estimates of the posterior densities of the population periodicities of the insect. We demonstrate the performance of our method on simulated examples under different scenarios. Remarkably, by incorporating space-time dependence, our approach produces more accurate inference for periodicities than current approaches even when the current approaches are provided with the actual gypsy moth populations. Using our approach, we analyze the defoliation data for forests in Pennsylvania between 1975 and 2005.

3.2.1 Background

The gypsy moth, *Lymantria dispar*, is a moth of Eurasian origin. It was introduced to North America in the late 1860s and has been expanding its range ever since (Liebhold and Elkinton, 1989; Williams and Liebhold, 1995). Today the area infested with the gypsy moth spans across the northeastern United States (Bjørnstad et al., 2010; Johnson et al., 2006). The gypsy moth is currently one of the most damaging tree defoliators in the US. It is known that gypsy moth population densities remain low in most years; when this is the case, their effect on forests is virtually undetectable. Occasional outbreaks of the gypsy moth when their population sizes are high, on the other hand, cause severe and widespread forest defoliation. It has been shown that these outbreaks are commonly periodic and synchronous when analyzed over large spatial scales over the northeaster United States (Montgomery and Wallner, 1988; Peltonen et al., 2002; Williams and Liebhold, 1995). At smaller spatial scales, when each forest location is less than a 2km² quadrant, the gypsy moth outbreaks seem to be spatially heterogeneous with possibly differing phases (Johnson et al., 2006).

The gypsy moth feeds on over 200 species of trees in North America (cf. Liebhold et al., 1995b). Gypsy moth caterpillars prefer to eat oak, apple, alder, aspen, basswood, birch, poplar, willow, hawthorn, hemlock, tamarack, pine, spruce, and witch hazel (Bess et al., 1947; Gottschalk, 1993; Houston and Valentine, 1977). Plants that gypsy moth avoids include ash, butternut, black walnut, locust, sycamore, yellow poplar, ferns, mountain laurel, redbud and rhododendron (cf. Gottschalk, 1993). Consequently, research studies show that forest susceptibility is related to the proportion of the trees that are preferred by gypsy moth larvae within each forest location (Bess et al., 1947; Gottschalk, 1993; Houston and Valentine, 1977). Therefore, as with most of the forest defoliating insects, the change in forest composition due to geographical variations may have an impact on periodicities and probabilities of the gypsy
moth outbreaks.

Recent studies investigate models for the dynamics of the gypsy moth populations that allow for periodicities to vary geographically and due to forest type (cf. Bjørnstad et al., 2010; Johnson et al., 2006). For instance, Johnson et al. (2006) show that the distribution of periodicities in the gypsy moth populations is not unimodal (see Bjørnstad et al., 1996, 1995, and examples therein), but instead is bimodal with two dominating periodicities of approximately 5 and 10 years. We note the approaches used for inferring these periodicities from small scale binary forest defoliation data has typically been wavelet analysis, with such an analysis carried out on time series for data aggregated across space and/or across forest type.

When data are aggregated across regions in which the gypsy moth outbreaks have different periodicities with different phases, the resulting periodicity estimates for the combined regions are likely to be highly unreliable. For example, if two regions have periodicities equal to 10 years in anti-phase, the analysis of the aggregated data would suggest an estimate of the periodicity equal to 5 years which would be very different from the truth. Bjørnstad et al. (2010) estimates the gypsy moth periodicities from the binary data using a spatial-smoothing method that attempts to resolve the problems that could arise from aggregation. This approach confirms the existence of differing periodicities among different forest types. The method is based on analyzing the defoliation times series for each forest location at their finest available resolution. Since this requires taking into account the possible spatial errors and inherent local stochasticities in estimating the unknown periodicities in the gypsy moth populations, Bjørnstad et al. (2010) suggests using a generalized time-lag autocorrelation function between locations that can reveal spatial patterns across broader spatial domains. Estimates based on these functions, however, may be biased due to two main reasons: (1) ignoring the spatiotemporal dependence across different locations, and (2) the defoliation data are treated as gypsy moth populations data rather than as a surrogate for the unobservable population size.

In this study, we develop a statistical inferential approach to estimate the unknown periodicities of the gypsy moth populations from binary space-time defoliation data. Our approach is based on reconstructing unobserved gypsy moth populations by modeling them as a latent Gaussian process. By using such an approach, we more realistically account for the fact that the binary data on forest defoliation are related to underlying populations of the gypsy moth. For the populations of the gypsy moth at any given location and time, this
method allows us to borrow information from neighboring observations both in space and in time, thereby accounting also for spatiotemporal dependence. We specify a full Bayesian model and infer the latent gypsy moth populations via Markov chain Monte Carlo. Because of our sample-based approach, we account for variability in the gypsy moth populations at each location in space and time and obtain population periodicity estimates along with their associated uncertainties. We demonstrate that our approach works well in different situations – it is particularly useful when the data are observed with errors, as in forest defoliation that can only be detected at levels that are above 30% (Webb et al., 1961). Based on simulated examples, we show that when outbreaks of gypsy moth are synchronized within neighboring locations, our new method is especially effective: it can even outperform standard approaches for estimating periodicities with known sizes of gypsy moth populations.

The rest of the section is organized as follows. Section 4.1 describes the forest defoliation data. Section 3.2.3 describes our inferential approach. Since naive inferential approaches for the gypsy moth periodicities are computationally infeasible for the large space-time defoliation dataset, computational concerns play an important role in our method. These concerns and possible ways to resolve them are also discussed in Section 3.2.3. In Section 3.2.4, we use simulated examples to compare the performance of our approach to other methods. In Section 3.2.5, we apply our method to the defoliation data for forests in Pennsylvania between 1975 and 2005. Finally, in Section 3.2.6, we summarize our results and discuss our statistical approach.

### 3.2.2 Forest Defoliation Data

There are no direct observations for actual sizes of the gypsy populations for large regions. Therefore, analysis of the gypsy moth population periodicities can only be conducted using forest defoliation data as a proxy for the gypsy moth abundance. There are studies showing the validity of using defoliation data in connection with the gypsy moth population sizes measured as the number of egg masses per hectare for small regions (cf. Gansner et al., 1985; Gribko et al., 1995; Liebhold et al., 1995a, 1998, 1993; Williams et al., 1991). Some of the possible issues that may arise from using the defoliation data to study the dynamics of the gypsy moth are discussed in Bjørnstad et al. (2002b, 2010). These are related to the facts that the aerial surveys may have some level of spatial error (Ciesla, 2000) and the underling gypsy moth dynamics could be spatially dependent (Bjørnstad et al., 2010).
USDA forest service have been collecting data on forest defoliation caused by the gypsy moth outbreaks since the early 1960s (cf. Bjørnstad et al., 2010). These data were obtained via aerial surveys conducted in each state in the northeastern United States. In our study, we use $2 \times 2$ km grids of annual, binary forest defoliation data for years from 1975 to 2005 for the part of Pennsylvania that was invaded by gypsy moth prior to 1975. Our data are derived from geographical information system (GIS) layers. For each grid, the data indicates the presence or absence of forest defoliation and contain information regarding the forest type for each forest location. Figures 3.1 and 3.2 illustrate the variation in the total numbers of the gypsy moth outbreaks for different locations, based on aggregated forest defoliation data. From these plots, we can see that the data have more outbreaks in the northeastern forests which corresponds to the upper corner of the window in Figure 3.1. It is also seen here that there are very few data on the middle of the window.

![Figure 3.1: Total number of forest defoliations aggregated over time in 2D](image)

### 3.2.3 Inference for Gypsy Moth Periodicities

The basic idea of our approach in inferring the unknown periodicities in the gypsy moth populations is to first recover the unobservable population sizes of the insect. Our model for the population sizes of the moth is based on a Gaussian process, with dependence both in space and time. In our approach, by using the estimated population sizes obtained from a latent model, we infer the periodicities of the gypsy moth abundance using a Fourier
Figure 3.2: Total number of forest defoliations aggregated over time in 3D

transformation at each step of the algorithm. This allows us to obtain samples from the posterior densities of the population periodicities of the gypsy moth.

We first begin with some notation. Denote the defoliation data as $Z = \{Z_{ijt}\}$, where indices $i$ and $j$ are the coordinates of the grid and $t$ represents a time step. We assume that the unobservable population sizes of the gypsy moth is a latent process $Y = \{Y_{ijt}\}$. Finally, we assume that the defoliation data are related to the latent $Y$ by the following equation

$$Z_{ijt} = \mathcal{I}(\exp(Y_{ijt}) > 0),$$

(3.6)

where as in the previous chapter, $\mathcal{I}()$ is an indicator function.

We model $Y$ as a Gaussian process,

$$Y \mid \mu, \xi \sim N(\mu, \Sigma_\xi),$$

where $\mu$ is a constant mean of the process and $\xi$ is a parameter that specifies the covariance function. Details about these two parameters and the form the covariance function we use are given below.

After specifying the model, we obtain samples from the posterior distribution of $Y$ given the data using an MCMC algorithm. For each fixed $i$ and $j$, each sample $\{Y_{ijt}\}$ is a time series. Using an estimate of the spectral density via a periodogram for each of these time series, we can obtain samples of the periodicities for each forest location. This is done by
finding frequencies that maximize the spectral density via a periodogram for each time series. Using samples of the gypsy moth population periodicities, we then obtain estimates of the posterior densities of the periodicities for all forest locations.

The defoliation data contain information on 897 forest locations (grids) for the period of 31 years. This translates to $897 \times 31 = 27,807$ unobservable variables that need to be updated at each iteration of the MCMC algorithm. In order to make our MCMC approach computationally feasible, we first split the data into 9 regions (blocks) based on the closeness of the grids. We analyze each smaller region with approximately 100 forest locations separately which allows us to run the MCMC algorithm in parallel. These independent regions are illustrated in Figure 3.3. We note that this simplification may affect the accuracy of our estimates at the locations that correspond the borders of the blocks. In order to verify that this effect does not change our overall conclusions, we also run our inferential algorithm on different, shifted regions that are chosen so that different blocks do not have common borders. These two sets of blocks are shown in Figure 3.4 by blue and red lines. Finally, by comparing our estimates at the border locations from different sets of blocks, we confirm that the accuracy we may loose at these points is not noticeable.

For each block of the data, in order to avoid high dimensional matrix inversions, we further assume that the covariance is separable in space and time:

$$\Sigma_\xi = \Sigma_{1\xi_1} \otimes \Sigma_{2\xi_2}$$  \hspace{1cm} (3.7)

where $\Sigma_{1\xi_1}$ and $\Sigma_{2\xi_2}$ are space and time components of the covariance matrix $\Sigma_\xi$. Based on our exploratory data analysis, since there is no strong evidence of nonseparability, we believe that this will not have significant adverse effects on our estimates. We note that, in general, there is parametric form for a nonseparable space-time covariance functions whose separability depends on a single parameter being non-zero (cf. Gneiting, 2002). In a Bayesian framework, one only needs to determine whether the credible region for this parameter includes zero to determine whether the covariance is separable. This, however, requires very rich data.

Finally, we assume that

$$(\Sigma_{1\xi_1})_{s_i, s_j} = C_1(d(s_i, s_j))$$

and

$$(\Sigma_{2\xi_2})_{t_i, t_j} = C_2(d(t_i, t_j))$$

where as in the previous chapters, $d(\cdot, \cdot)$ is a bivariate function that returns the Euclidean
distance between its arguments. Functions $C_1()$ and $C_2()$ are the spherical covariance functions:

\[
C_1(d; \kappa_1, \tau_1) = \begin{cases} 
0 & \text{if } d > 4000, \\
\kappa_1 \left( 1 - \frac{3}{2} \frac{d}{4000} + \left( \frac{1}{2} \frac{d}{4000} \right)^3 \right) & \text{if } 4000 \geq d > 0, \\
\kappa_1 + \tau_1 & \text{otherwise.}
\end{cases}
\]

and

\[
C_2(d; \kappa_2) = \begin{cases} 
0 & \text{if } d > 5, \\
\kappa_2 \left( 1 - \frac{3}{2} \frac{d}{5} + \left( \frac{1}{2} \frac{d}{5} \right)^3 \right) & \text{if } 5 \geq d > 0, \\
\kappa_2 + \tau_2 & \text{otherwise.}
\end{cases}
\]

We note here that assuming a spherical covariance structure increases the efficiency of our MCMC algorithm by allowing us to be able to update each component of $Y$ one at a time while letting us to employ benefits of using Kronecker products in Equation (3.7). Fixed range parameters of the spherical covariance functions make it possible to use the simplification

\[
P(Y_{(ij)t}|Y_{-((ij)t)}) = P(Y_{(ij)t}|N(Y_{(ij)t})),
\]

Figure 3.3: Division of the data into separate blocks based on their closeness.
where $N(Y_{(ij)t})$ is a sub-matrix of elements of $Y$ which are close to $Y_{(ij)t}$ both in space and time, and $Y_{-((ij)t)}$ is a set of all the elements of $Y$ without $Y_{(ij)t}$. We fix the range parameters at these specific $\phi_1 = \frac{1}{4000}$ and $\phi_2 = \frac{1}{5}$ based on scientific assumptions about the ranges of spatial synchrony in interspecific population dynamics of the gypsy moth (cf. Johnson et al., 2005). In other words, we fix the range parameter of the spherical covariance function based on a scientific belief that there is no correlation between the observations when data are far apart. An additional benefit of fixing these parameter values is computational – these parameters are difficult to infer and by fixing them we also impose a certain degree of sparsity that allows for faster computation. When the space-time observations are binary, learning about parameters of a latent Gaussian process is difficult. See, for example, the discussion in Oliveira (2000) regarding identifiability issues.

According to the above notation, $\xi = (\xi_1, \xi_2)$, where $\xi_1 = (\kappa_1, \tau_1)$ and $\xi_2 = (\kappa_2, \tau_2)$. For the mean $\mu$, we use a normal prior centered at zero and with a variance equal to 100. In our MCMC algorithm, we update $\mu$ using a simple random walk. For the covariance parameters ($\xi_1, \xi_2$) and ($\xi_1, \xi_2$), we use improper positive uniform priors. These parameters
are updated using a slice sampler (cf. Neal, 2003a). Finally, full conditional distributions of the components of \( Y \) are truncated normal distributions (cf. McCulloch and Rossi, 1994), and are updated using a Gibbs sampler. For each accepted \( Y \), we obtain estimates of the periodicities of the gypsy moth fluctuations using a standard optimizer and periodogram functions in R (R Development Core Team, 2011).

### 3.2.4 Analysis of Simulated Examples

In this section, we demonstrate the performance of our inferential approach on simulated examples. In these examples, we choose 100 \( 2 \times 2 \) km forest grids with coordinates similar to the ones of the actual defoliation locations. For each selected location, we simulate time series using

\[
y = \sin(2x\pi/P(s)) + \text{error},
\]

where \( P(s) \) denotes the true periodicity that could be different for different locations \( s \). The defoliation data has 31 time steps. Therefore, in order to make our simulated data similar to the actual data, we use \( x = \{1, 2, ..., 30, 31\} \). Forest defoliation can be detected at levels that are above 30\% (Webb et al., 1961). Furthermore, because this information is obtained via satellite imagery, there is typically error associated with such data. In Equation (3.8), we assume normally distributed errors with standard deviations equal to 0.3, 0.75 and 1.1.

At the final step of the simulation algorithm, for each forest location, using the time series generated by Equation (3.8), we create simulated data by converting these time series into binary data via Equation (3.6).

In choosing the true periodicities, we consider two different scenarios. In the first scenario, we assume that there is no spatial synchrony between locations and choose periodicities for different forest grids randomly from a set of numbers from 4 to 11. In the second scenario, we divide the locations into subregions and randomly choose the true periodicities within each subregion making sure that periodicities for forest locations within same subregions are close to each other.

We now assume that the true periodicities for the simulated data sets are unknown and try to infer them using our approach. For simplicity, let us refer to our method as a latent Gaussian process-based method (LGP). We compare the performance of LGP to two default methods of inference. The first of these two methods is based on using binary time series for each forest location directly (method based on binary data (BD)). The second
methods is based on the assumption that population sizes of the gypsy moth, \( Y \), are known (methods based on known populations (PK)). This method therefore infers the population periodicities of the insect from the data for \( Y \). We refer to these methods as BD and PK respectively. We note that, in contrast to LGP, both of these methods do not exploit the possible spatiotemporal dependencies between locations. It is also clear that PK is unrealistic for the real defoliation data as information on the latent population sizes of the gypsy moth are only available from simulations.

Figure 3.5: Density Plots of Periodicities for Different Locations (LGP), along with estimates obtained with BD and PK: (a) All methods recover the truth; (b) LGP recovers the truth with some uncertainty; (c) LGP fails to recover the truth; (d) LGP outperforms BD and PK; (e) BD outperforms LGP; (f) PK outperforms LGP.

Figure 3.5 contains plots of some of the inferred posterior densities for different forest locations. The vertical lines in these plots represent the true periodicities, and estimates obtained from BD and PK. In Figure 3.5 (a), we can see that our method, LGP, recovers the true periodicity exactly. In this plot, both BD and PK capture the truth as well. In Figure 3.5 (b), LGP reveals the truth with some uncertainty, since the posterior density is bimodal. In Figure 3.5 (c), the peak of the density function is far from the true value of the
### Table 3.1: Summary of inference without spatiotemporal dependence

<table>
<thead>
<tr>
<th>Performance</th>
<th>Error SD = 0.3</th>
<th>Error SD = 0.75</th>
<th>Error SD = 1.1</th>
</tr>
</thead>
<tbody>
<tr>
<td>LGP recovers the truth exactly</td>
<td>76%</td>
<td>70%</td>
<td>56%</td>
</tr>
<tr>
<td>LGP is close to the truth</td>
<td>24%</td>
<td>26%</td>
<td>22%</td>
</tr>
<tr>
<td>LGP fails to recover the truth</td>
<td>0%</td>
<td>4%</td>
<td>22%</td>
</tr>
<tr>
<td>LGP is better than BD</td>
<td>0%</td>
<td>0%</td>
<td>6%</td>
</tr>
<tr>
<td>LGP is better than PK</td>
<td>0%</td>
<td>0%</td>
<td>10%</td>
</tr>
<tr>
<td>BD is better than LGP</td>
<td>0%</td>
<td>2%</td>
<td>2%</td>
</tr>
<tr>
<td>PK is better than LGP</td>
<td>4%</td>
<td>12%</td>
<td>16%</td>
</tr>
</tbody>
</table>

Table 3.2: Summary of inference with spatiotemporal dependence

<table>
<thead>
<tr>
<th>Performance</th>
<th>Error SD = 0.3</th>
<th>Error SD = 0.75</th>
<th>Error SD = 1.1</th>
</tr>
</thead>
<tbody>
<tr>
<td>LGP recovers the truth exactly</td>
<td>70%</td>
<td>68%</td>
<td>52%</td>
</tr>
<tr>
<td>LGP is close to the truth</td>
<td>30%</td>
<td>28%</td>
<td>30%</td>
</tr>
<tr>
<td>LGP fails to recover the truth</td>
<td>0%</td>
<td>4%</td>
<td>8%</td>
</tr>
<tr>
<td>LGP is better than BD</td>
<td>4%</td>
<td>24%</td>
<td>28%</td>
</tr>
<tr>
<td>LGP is better than PK</td>
<td>0%</td>
<td>4%</td>
<td>20%</td>
</tr>
<tr>
<td>BD is better than LGP</td>
<td>0%</td>
<td>0%</td>
<td>4%</td>
</tr>
<tr>
<td>PK is better than LGP</td>
<td>0%</td>
<td>0%</td>
<td>6%</td>
</tr>
</tbody>
</table>

Periodicity. This means that LGP is missing the truth for this location. In Figure 3.5 (d), inference based on LGP is clearly better than estimates obtained via other approaches. In Figure 3.5 (e), BD (method based on binary data directly) results in more accurate inference. Finally, plot in Figure 3.5 (f) shows the case when inference based on PK (inference based on the true population sizes) is better than LGP.

Tables 3.1 and 3.2 summarize our conclusions on inferring the unknown periodicities using LGP, BD and PK. In Table 3.1, when data do not have spatiotemporal dependencies, we see that all three methods perform relatively similar. Based on this simulated example, when data have a lot of noisy, the method of inference assuming that population sizes are known (PK) resulted in more accurate estimates in 16% of the locations, while our method outperformed PK in 10% of the locations. In Table 3.2, when data are spatiotemporally correlated, we see that the percentage of locations when our method performs better that other approaches increases with the magnitude of the added noisy. Here, for example, inference based on LGP was more accurate than inference based on PK in 20% of the locations, while PK outperformed our method only in 6% of the times.
3.2.5 Results

In this section, we apply our inferential approach to the forest defoliation data described in Section 3.2.2. Figure 3.6 shows some of the inferred posterior density functions of unknown gypsy moth population periodicities. We see here that the shapes of the density functions can be very different for various forest locations. While for some locations, the modes of the density functions are very peaked resulting in estimates of the periodicities with very small variability (as in Figure 3.6 (a) or (b)), there are forest locations where corresponding posterior densities are flat indicating a lot of uncertainty in inferring the unknown periodicities (as in Figure 3.6 (d) or (f)). For some locations, inferred density functions are multimodal (as in Figure 3.6 (c) or (e)). We note here that in our simulated examples in Section 3.2.4, all true periodicities were equal to integer numbers and our inferential approach resulted in very accurate estimates. However, when the true periodicities of the process are not equal to integers, using estimates based on periodograms may be slightly inaccurate. This is due to the fact that these estimates will be chosen based on canonical frequencies of the corresponding periodograms (calculated from time series with integer time steps) that are closest to the true non-integer periodicities.

In order to see the possible differences in periodicities in populations of the gypsy moth between forests with different tree compositions, we combine all samples by forest types. As mentioned above, our inferential approach can only result in discretized estimates of the true periodicities. Therefore, instead of plotting the obtained posterior density functions (like we did in Figure 3.6), we chose to display histograms of the samples from corresponding posterior distributions. We also note that estimates of periodicities equal to 2 are artifacts of the estimation procedure resulted from the largest canonical frequencies and are not supported by data. We remove these in post-processing of our samples. Resulting histograms of the unknown periodicities calculated from the pooled samples are in Figure 3.7. Based on these plots, we see that the gypsy moth cycles in oak-pine forest have a periodicity that varies between 4 to 8 years, with two dominant periodicity equal to 4 and 8 years. For oak-hickory forests, the dominant periodicity is equal to 8 years. For forest with high proportion of maple, beech, and birch, the dominant periodicity of the gypsy moth abundance is around 4 years. Finally, for parks close to populated areas (non-forests), the insect outbreaks seem to occur with a dominant periodicity of 8 years.
3.2.6 Discussion

There are numerous studies in the literature about periodicities in the gypsy moth populations (Bjørnstad et al., 2010; Johnson et al., 2006; Montgomery and Wallner, 1988). Inference in these studies, however, is based on binary defoliation data only (Bjørnstad et al., 2010; Johnson et al., 2006). If the gypsy moth populations are cycled with a periodicity, but does not always reach very high densities to cause forest defoliation at its every cycle, these methods therefore may fail to result in accurate estimation. Some studies infer the periodicities of the gypsy moth abundance from aggregated data for different types of forest (Johnson et al., 2006). It is possible that population cycles from different locations with out of phase dynamics may give rise to spurious periodicity estimates when analyzed as a combined region. Therefore, methods where data are aggregated by either forest types or larger regions based on distances may result in poor estimates as well.

In this section, we described a new approach to infer the unknown gypsy moth population
periodicities. The method is based on recovering the latent gypsy moth population sizes by exploiting the spatiotemporal dependencies in the defoliation data. We present simulated examples to demonstrate that our approach works very well even when data have a lot of noise. Using the approach, we infer the gypsy moth periodicities from the defoliation data for forests in Pennsylvania. The results show that the posterior densities of the periodicities of gypsy moth population fluctuations for all forest types are multimodal. We also find that the dominant peaks of these densities for different forest types are different, which is consistent with the previous scientific literature (Bess et al., 1947; Bjørnstad et al., 2010; Houston and Valentine, 1977; Johnson et al., 2006).

What may be of additional interest is addressing and explaining the possible causes in the observed differences of the gypsy moth periodicities by geographical locations or due to forest type. Common models proposed to explain interspecific differences in population dynamics of insects include models based on host-pathogen interactions, host-parasite in-
teractions, maternal effects, and induced host defences. In a recent study, Bjørnstad et al. (2010) investigate a model that describes both the interaction between the gypsy moth and the pathogen and the interaction between the insect and the generalist-predators. Using simulations from the model, the paper explains the differences in the gypsy moth dynamics in different forest types by possible changes in predator carrying capacity. To carry out a full-fledged statistical analysis along these lines, however, one would need additional sources of information that are currently unavailable.
Emulating a Gravity Model to Infer the Spatiotemporal Dynamics of an Infectious Disease

In the previous chapters, we described inferential approaches based on approximate Bayesian computation. These methods are useful when the model has an intractable likelihood, but when simulations from the model that corresponds to that likelihood are relatively cheap. When the model is complex and simulations from it are also time consuming, ABC-based inference is computationally infeasible. In this chapter, we describe a new inferential approach that can be used in problems with intractable likelihoods and expensive simulations. Our motivating example for this method is a challenging inference problem for measles dynamics.

4.1 Motivation

Infectious disease dynamics are of interest to modelers from a range of disciplines. The theory of disease dynamics provides a tractable system for investigating key questions in population and evolutionary biology. Understanding the disease dynamics helps in management and with pressing disease issues such as disease emergence and epidemic control strategies. Probabilistic models for disease dynamics are important as they help increase our understanding of the mechanism underlying the spread of the infection while also accounting for their inherent stochasticity. Observations on reported cases of the diseases, especially in the
form of space-time data, are becoming increasingly available, allowing for statistical inference for unknown parameters of these models. However, traditional likelihood-based inference for many disease dynamics models is often challenging because the likelihood function may be expensive to evaluate, making likelihood-based inference computationally intractable. Furthermore, traditional inference may lead to poor parameter estimates and the fitted model may not capture important biological characteristics of the observed data. Hence, an approach that simultaneously addresses the computational challenges as well as the inferential issues would be very useful for a number of interesting and important probabilistic models for dynamics of diseases. Inspired by work in the field of emulation and calibration for complex computer models (cf. Bayarri et al., 2007a; Craig et al., 2001; Kennedy and O’Hagan, 2001; Sacks et al., 1989a), we develop a novel approach for inference for such models. Our approach uses a Gaussian process approximation to the disease dynamics model using key biologically relevant summary statistics obtained from simulations of the model at differing parameter values. As we will demonstrate, this approach results in reliable parameter estimates and a good model fit, and is also computationally efficient.

The motivating example for our approach is the gravity time series susceptible-infectious-recovered (TSIR) model for measles dynamics. The spatiotemporal dynamics of measles have received a lot of attention in part due to the importance of the disease, the highly nonlinear outbreak dynamics and also because of the availability of rich data sets. Important aspects of local dynamics of measles are well studied. These include key issues like seasonality in transmission of the infection (Bjørnstad et al., 2002a; Dietz, 1976), effects of host demography on outbreak frequency (Finkenstädt et al., 1998; McLean and Anderson, 1988), and causes of local persistence and extinctions (Bartlett, 1956b; Grenfell et al., 2001; Grenfell and Harwood, 1997). During the course of outbreaks in well-mixed local populations, the epidemic trajectory of measles is virtually unaffected by infection that may enter from neighboring locations. However, spatial coupling is fundamental to the dynamics and management of measles for smaller communities where the infection may become locally extinct (Bartlett, 1956b; Grenfell and Harwood, 1997). Hence, ecologists have also studied the spatial spread of the disease using so-called metapopulation models (Earn et al., 1998; Grenfell and Harwood, 1997; Swinton and Grenfell, 1998).

In this work, we investigate inference for a model first proposed by Xia et al. (2004). The model represents a combination of the TSIR model (Bjørnstad et al., 2002a; Grenfell et al., 2002) with a term that allows for spatial transmission between different host commu-
nities modeled as a gravity process. Xia et al. (2004) demonstrate how this model captures scientifically important properties of measles dynamics. Since each likelihood evaluation is computationally very expensive, however, Xia et al. (2004) obtain only point estimates of the parameters minimizing ad hoc objective functions instead of using a likelihood-based approach. Here, we develop a more statistically rigorous approach to inferring model parameters, characterizing associated uncertainties and carefully studying parameter identifiability issues. First, in order to explain the issues that arise in inferring these parameters via a likelihood-based approach, we propose a partial discretization of the parameter space that allows us to perform Bayesian inference for the parameters using a fast MCMC algorithm. Using this approach we are able to study uncertainties about the parameter values. The method allows us to investigate parameter identifiability issues, showing which gravity model parameters can or cannot be inferred from a given data set. However, this approach to resolving the computational challenges of traditional likelihood-based inference is problematic, as is revealed by our simulated data examples. We find that the parameter estimates are poor and the forward simulations of the model at these parameter settings do not reproduce epidemiological features of the data deemed key in Xia et al. (2004).

In order to address the above issues, we propose a new approach that directly focuses on the aspects of the underlying process that are of scientific interest. We develop a Gaussian process approximation to the gravity model based on key summary statistics obtained from simulations of the model at different parameter values. These statistics are chosen by domain experts to capture the biologically important characteristics of the dynamics of the disease. The Gaussian process model ‘emulator’ is then used to develop a probability model for the observations, thereby permitting an efficient MCMC approach to Bayesian inference for the parameters. We demonstrate that the new method recovers the true parameters and the resultant fitted model captures biologically relevant features of the data.

When applied to the gravity TSIR model, our approach allows us to investigate several scientific questions that are of interest to the dynamics of measles. For example, we study changes in dynamics between school holiday periods versus non-holidays. This is particularly interesting because the local, age-structured transmission rate of the disease changes from holidays to non-holidays (Bjørnstad et al., 2002a; Dietz, 1976). Since our approach allows us to construct confidence regions easily, we also infer the amounts of exported and imported infected individuals for different cities during different time periods. More generally, the methodology we develop here may be useful for models where the likelihood is expensive to
evaluate or in situations where the likelihood is unable to capture characteristics of the model that are of scientific interest. We note that the computational cost of forward simulations for our model makes approaches based on approximate Bayesian computation (ABC) (cf. Beaumont et al., 2002; Marjoram et al., 2003; Pritchard et al., 1999) infeasible. Hence our approach is computationally efficient, while ABC is not a viable option here.

The rest of the section is organized as follows. Section 4.2 describes in detail the gravity TSIR model, which acts as our motivating example. Section 4.3 describes the inferential and computational challenges posed by the model and the large space-time data set. Section 4.4 describes our new emulation-based approach that is an alternative to traditional likelihood-based inference. Section 4.5 describes computational details and the application of our method to the gravity TSIR model in simulated data examples. Section 4.6 describes the application of our method to the England-Wales measles data. Finally, in Section 4.7, we summarize our results and discuss our statistical approach and scientific conclusions.

4.2 A Gravity Model for Disease Dynamics

A general goal of fitting metapopulation disease dynamics models is to describe spatiotemporal patterns of epidemics at the local scale and understand how these patterns are affected by the network of spatial spread of the disease (Cliff et al., 1993; Keeling et al., 2004). The gravity model we study is an extension of a discrete time-series susceptible-infectious-recovered model (Bjørnstad et al., 2002a; Grenfell et al., 2002) for local disease dynamics which includes an explicit formulation for the spatial transmission between different host cities (Xia et al., 2004).

The common theoretical framework used to describe the dynamics of infectious diseases is based on the division of the human host population into groups containing susceptible, infectious (infected) and recovered individuals. Let \( I_{kt} \) and \( S_{kt} \) denote the number of infected and susceptible individuals respectively in disease generation \( t \) in city \( k \) and variable \( L_{kt} \) be the number of infected people commuting to city \( k \) at time \( t \). The ‘commuting’ assumption reflects that movement of infection is mostly through transient movement of individuals. Denote the size and birth rate of city \( k \) at time \( t \) by \( N_{kt} \) and \( B_{kt} \), and let \( d_{kj} \) represent the distance between cities \( k \) and \( j \). The model can then be described as follows. First, the
model for the number of incidences of measles is

\[ I_{k(t+1)} \sim \text{Poisson}(\lambda_{k,t+1}), \quad \text{where} \quad \lambda_{k,t+1} = \beta_t S_k(t) (I_{kt} + L_{kt})^\alpha, \]  

(4.1)

with \( t = 1, ..., T, k = 1, ..., K \). The time-step is taken to be 2 weeks, roughly corresponding to the generation length (serial interval) of measles. The so-called transmission coefficient, \( \beta := \{ \beta_t \} \), is a parameter that represents the attack rate of measles at time \( t \). The parameter \( \alpha \) is a positive real number which usually takes a value slightly less than 1. This parameter adjusts for the fact that a discrete-time approximation to the underlying continuous-time epidemic process may reduce the force of infection (Glass et al., 2003). Additionally, inclusion of \( \alpha \) accounts for the possibility that if individuals are spatially or socially clustered within local communities, the force of infection cloud be smaller at high infectious densities (Fine and Clarkson, 1982; Liu et al., 1987). Since the parameters \( \alpha \) and \( \beta_t \) only affect the local dynamics of measles, henceforth we refer to these parameters as “local dynamics parameters”.

The indexing by \( t \) for \( \beta_t \) reflects how this parameter is taken to be a piece-wise constant taking 26 different values to accommodate seasonal variability of the transmission rate that is repeated every year (Bjørnstad et al., 2002a; Fine and Clarkson, 1982; Finkenstädt and Grenfell, 2000; Grenfell et al., 2002). From this, it can be seen that \( I_{k(t+1)} \) increases depending on the number of susceptibles and the number of moving infections coming to city \( k \) at the previous time step. Note that we use the Poisson distribution whereas Xia et al. (2004) use the Negative Binomial distribution; this is due to the greater computational stability of the Poisson distribution for small values of \( \lambda \). Our exploratory analysis show that a model fit from using the Poisson distribution is similar to a model fit obtained with the Negative Binomial distribution.

The susceptibles are modeled as follows

\[ S_{k(t+1)} = S_{kt} + B_{kt} - I_{k(t+1)}, \]  

(4.2)

reflecting how susceptibles are replenished by births and depleted by infection. Since case fatality from measles was very low for the period of time in this study and mean age of infection was small, mortalities are not included in this balance equation.
Finally, the gravity model describes the number of moving infected individuals by

\[ L_{kt} \sim \text{Gamma}(m_{kt}, 1), \text{ where } m_{kt} = \theta N_{kt}^\tau_1 \sum_{j=1, j\neq k}^K \frac{I_{jt}^{\tau_2}}{d_{kj}}, \tag{4.3} \]

where Gamma(a,b) represents the Gamma distribution with shape and scale parameters a and b respectively. Here, b is chosen to be equal to unity based on exploratory analysis of the fitted model (Xia et al., 2004). The reason to model immigrant infection as a continuous random variable lies in the assumption that the transient infectives do not remain for a full epidemic generation.

The local dynamics parameters in Equation (4.1) have been estimated previously (Bjørnstad et al., 2002a; Finkenstädt et al., 2002; Grenfell et al., 2002). In this study, we are interested in learning about the parameters \( \theta, \tau_1, \tau_2 \) and \( \rho \) in Equation (4.3) as these parameters control the spatial spread and regional behavior of the disease. Note, however, that for convenience and numerical stability, we use a reparametrization of \( \theta, \theta' = -\log_{10}(\theta)/5 \) throughout the chapter.

4.3 Parameter Inference for the Gravity Model

Reliable estimates of the local dynamics parameters \( \alpha \) and \( \beta \) are available for measles dynamics (Bjørnstad et al., 2002a; Finkenstädt et al., 2002; Grenfell et al., 2002; Xia et al., 2004). Therefore, since we are interested in spatial dynamics here, we assume that these parameters are known and use the previously obtained estimates as the true values. In particular, the local seasonal transmission parameters for biweeks 1 through 26, \( \beta_t \), is taken to be equal to \( \beta_t = (1.24, 1.14, 1.16, 1.31, 1.24, 1.12, 1.06, 1.02, 0.94, 0.98, 1.06, 1.08, 0.96, 0.92, 0.92, 0.86, 0.76, 0.63, 0.62, 0.83, 1.13, 1.20, 1.11, 1.02, 1.04, 1.08) \), and \( \alpha \) is assumed to be 0.97.

Given parameter identifiability issues, joint estimation of the spatial dynamics and all the local dynamics parameters (\( \alpha \) and \( \beta_t \)) is infeasible. Furthermore, assuming that the local dynamics parameters are known does not have an undue effect on the model fit as was shown in the literature (cf. Xia et al., 2004). This leaves us with four unknown parameters, \( \theta', \tau_1, \tau_2 \) and \( \rho \), that we call the gravity model parameters (in our Gaussian process based approach in Section 4.4 we will also introduce several other parameters). In this chapter our focus is on investigating the gravity model parameters and, when possible, obtaining the best estimates
of them with relevant descriptions of their variability.

As suggested by our domain experts, feasible values for the gravity parameters lie in the interval $[0, 2]$ (see also Xia et al., 2004). Therefore, we use uniform priors for $(\theta', \tau_1, \tau_2, \rho)$ in all the inferential approaches that follow.

The data are spatiotemporal and high-dimensional, $K \times T = 952 \times 546$ in the case of the England-Wales measles data. Here, $K = 952$ is the number of cities in the data and $T = 546$ is the number of bi-weeks from 1944 to 1965. To study whether our fitted model captures epidemiologically relevant features of the data, we focus on two important biological characteristics of the process as suggested by domain experts. These are:

1. Maximum number of incidences which we will denote by $\mathbf{M} = (M_1, \cdots, M_K)$, where $M_i$ is the maximum number of incidences for the $i$-th city.

2. Proportions of bi-weeks without any cases of infection denoted by $\mathbf{P} = (P_1, \cdots, P_K)$, where $P_i$ is the proportion of incidence free biweeks for the $i$-th city.

An important goal of our work is to find parameter settings (along with associated uncertainties and dependencies among them) that yield a model that produces disease dynamics that are as close as possible to the data in terms of capturing these key properties.

### 4.3.1 A Gridded MCMC Approach

In this section we demonstrate via simulated data examples how traditional likelihood-based approaches can be problematic for the gravity model. Because likelihood-based inference is computationally intractable, we develop a gridded MCMC approach instead; this approach involves a partial discretization of the parameter space. The gridded MCMC algorithm also requires certain simplifying assumptions and data imputation for unobservable susceptibles ($\{S_{kt}\}$) which are explained below. We would like to clarify, however, that the alternative inferential method we later recommend in Section 4.4 neither requires data imputation nor any of the simplifying assumptions used for the gridded MCMC approach.

It is easy to see why each evaluation of the likelihood for the gravity model is expensive. As in many population dynamic models, the major difficulty is in integrating over high-dimensional unobserved variables. For our model, $\{L_{kt}\}$ and $\{S_{kt}\}$ are of $K \times T$ dimensions each, which translates to $2 \times K \times T = 2 \times 952 \times 546 = 1,039,584$ in the case of measles data set considered in Section 4.6. Details of the likelihood function are given in Appendix A.
In order to avoid integrating over the latent \( \{ S_{kt} \} \), we first use a standard susceptible reconstruction algorithm (Bobashev et al., 2000; Ellner et al., 1998; Fine and Clarkson, 1982; Finkenstädt and Grenfell, 2000; Schenzle, 1984) that is based on the fact that in the prevaccination era almost all children were infected. This algorithm allows us to impute \( \{ S_{kt} \} \) from the observations for \( \{ I_{kt} \}, \{ N_{kt} \} \) and \( \{ B_{kt} \} \) by the following. We rewrite Equation (4.2) of the gravity model according to Finkenstädt and Grenfell (2000):

\[
S_{kt} = \overline{S}_k + D_{k0} + \sum_{j=0}^{t} B_{kj} - \sum_{j=0}^{t} I_{kj}/\omega
\]

(4.4)

where \( \overline{S}_k \) is the mean number of susceptibles in city \( k \), \( D_{k0} \) is the unknown deviations around the mean at time 0, and \( \omega \) is the reporting rate. We take this rate to be equal to 0.5 which is a common assumption for measles dynamics in the prevaccination era based on several other studies (cf. Bjørnstad et al., 2002a; Clarkson and Fine, 1985; Finkenstädt and Grenfell, 2000). In practice, there are variations in reporting rates with infection level for different locations. This does not, however, seem to be an issue for susceptible reconstruction for measles dynamics (cf. Clarkson and Fine, 1985).

We can reconstruct the time series \( D_{kt} \) of how the local susceptible numbers deviate from the local mean value, \( D_{kt} = S_{kt} - \overline{S}_k \), by rewriting Equation (4.4) as,

\[
\sum_{j=0}^{t} B_{kj} = D_{k0} + 1/\omega \sum_{j=0}^{t} I_{kj} + D_{kt}
\]

(4.5)

from which it is clear that \( D_{kt} \) is the residual from the regression of cumulative number of births on the cumulative number of cases. Note here that this algorithm works when \( D_{k0} \) and the reporting rate \( \omega \) are unknown since these are accommodated by the intercept and slope of the cumulative-cumulative regression. The method, however, does not allow the independent estimation of the mean number of susceptibles. Using the previous analysis, we assume that this number is equal to 4% of the population in city \( k \) (Bjørnstad et al., 2002a).

We avoid integrating over the unobserved \( \{ L_{kt} \} \) by setting \( L_{kt} = m_{kt} \) for all \( k \) and \( t \), that is, using the expectation instead of using the full Gamma distribution. Based on a study of this in several simulated examples (where we know the true values of \( \{ L_{kt} \} \)), this assumption does not seem to affect our likelihood-based inference and conclusions.

As expected, after these steps of simplification and data imputation, likelihood calcu-
lations are faster, but still expensive taking more than two minutes to evaluate for each parameter setting. This computational efficiency is sufficient if we want to make inference based only on maximization of the likelihood. It takes about 72 hours for a standard optimizer in R (R Development Core Team, 2011) to converge on an AMD Quad Core 2.6 GHz processor. However, a much more thorough exploration of inference for this model is desirable; we are interested in learning about parameter uncertainties, the joint distributions of the parameters, as well as any identifiability issues. We achieve the additional speedup necessary for MCMC by using a gridded MCMC approach as follows. Note that for each calculation of the likelihood, we calculate $K \times T$ values of $m_{kt}$ according to Equation 4.3. For each $k$ and $t$, this requires summing over $K$ quantities. We speed our calculations by selecting a grid on the range of possible values for $\tau_2$ and $\rho$. For each point of the grid, we then calculate and save a set of computationally expensive matrices $\{M_{kt}(\tau_2, \rho)\}$:

$$M_{kt}(\tau_2, \rho) = \sum_{j=1, j \neq k}^{K} \frac{(Ijt)^{\tau_2}}{d_{kj}^\rho}.$$  

Then, when evaluating the likelihoods, $L(\theta', \tau_1, \tau_2, \rho)$, we use the pre-calculated matrices $\{M_{kt}(\tau_2, \rho)\}$ and the fact that

$$m_{kt} = \theta' N_{kt}^{\tau_1} M_{kt}(\tau_2, \rho).$$

In this way, we reduce the number of arithmetic operations necessary to calculate the likelihood for each iteration from $O(K^2T)$ to $O(T + K)$. For the measles data we analyze in Section 4.6.2, this reduces the number of floating point operations involved in each likelihood evaluation from 499,009,056 to 1,502.

With discretized $\tau_2$ and $\rho$, we make inference based on the posterior distribution of the parameters using samples obtained via MCMC. Details of inference based on the approach and relevant plots of the inferred posterior surface are summarized in Sections 4.3.2 and 4.5.1.

### 4.3.2 Simulated Examples

We note that all simulated data sets we consider in this work are generated from the full gravity model described in Section 4.2 with initial points equal to the actual observations at $t = 1$. In these examples, the number of locations, their coordinates, demographic variables, and the number of time steps are the same as those in the measles data described in Section 4.6.1.
In our first example, we simulate a data set using values for the gravity parameters $\theta' = 0.71$, $\tau_1 = 0.3$, $\tau_2 = 0.7$ and $\rho = 1$. This parameter setting results in realistic data that resembles the observations. Figure 4.1 shows conditional and unconditional posterior likelihood surface plots for $\theta'$ and $\rho$ obtained by using the above gridded MCMC approach. From these plots, we can easily see that inference for $\theta'$ and $\rho$ is not possible because of the apparent issue with identifiability (Figure 4.1 (a)). In Figure 4.1 (b) we see that identifiability is reduced, but still exists when we fix one of the parameters, say $\tau_1$, at its known true value. In Figure 4.1 (c), we fix both of $\tau_1$ and $\tau_2$ at their true values and see that the obtained ridge contains the true values for $\theta'$ and $\rho$. Figure 4.1 (d) demonstrates that the ridge moves by changing the values of $\tau_1$ and $\tau_2$ away from their true values. Figure 4.2 is similar to Figure 4.1. The difference is that here we plot everything in $\tau_2$ and $\rho$ surface first without any assumptions (Figure 4.2 (a)), then by fixing $\theta'$ (Figure 4.2 (b)) and then by fixing both $\theta'$ and $\tau_1$ (Figure 4.2 (c)). Finally, Figure 4.2 (d) shows how the ridge in the posterior moves by changing the values of the other parameters.

Figure 4.1: Inferred posterior 2D likelihood surface obtained for data with known parameters ($\theta' = 0.71$, $\tau_1 = 0.3$, $\tau_2 = 0.7$ and $\rho = 1$): (a) Marginal 2D likelihood surface for ($\theta'$, $\rho$); (b) Marginal 2D likelihood surface for ($\theta'$, $\rho$) assuming $\tau_1 = 0.3$ (true); (c) 2D likelihood surface for ($\theta'$, $\rho$) assuming $\tau_1 = 0.3$ (true) and $\tau_2 = 0.7$ (true); (d) 2D likelihood surface for ($\theta'$, $\rho$) assuming $\tau_1 = 0.5$ (any value) and $\tau_2 = 1$ (any value).

In our second example, we simulate a data set using values for the gravity parameters $\theta' = 0.71$, $\tau_1 = 0.5$, $\tau_2 = 1$ and $\rho = 1$. Figure 4.3 is a plot of the two-dimensional likelihood
Figure 4.2: Inferred posterior 2D likelihood surface obtained for data with known parameters ($\theta' = 0.71$, $\tau_1 = 0.3$, $\tau_2 = 0.7$ and $\rho = 1$): (a) Marginal 2D likelihood surface for ($\tau_2$, $\rho$); (b) Marginal 2D likelihood surface for ($\tau_2$, $\rho$) assuming $\theta' = 0.71$ (true); (c) 2D likelihood surface for ($\tau_2$, $\rho$) assuming $\theta' = 0.71$ (true) and $\tau_1 = 0.3$ (true); (d) 2D likelihood surface for ($\tau_2$, $\rho$) assuming $\theta' = 1$ (any value) and $\tau_1 = 0.3$ (any value).

Figure 4.3: Inferred posterior 2D likelihood surface obtained for data with known parameters ($\theta' = 0.71$, $\tau_1 = 0.5$, $\tau_2 = 1$ and $\rho = 1$): Posterior 2D likelihood surface for ($\theta'$, $\rho$) assuming $\tau_1 = 0.5$ (true) and $\tau_2 = 1$ (true) has a shift and does not contain the true ($\theta'$, $\rho$) at its highest probability area.
in $\theta'$ and $\rho$ space obtained by fixing $\tau_1$ and $\tau_2$ at their true values 0.5 and 1 respectively. We can see here that the true values of the parameters of interest are not in the region where the likelihood is maximized. This, unfortunately, means that repeating the above with other simulated data with different true values for the gravity parameters reveals that the ridge analogous to the ridge in Figure 4.1 (c) does not always have to contain the true values for $\theta'$ and $\rho$. From our study of multiple simulated data, we also find that the likelihood ridge can have an intercept that is different from the ridge that we would intuitively think as the true ridge while having the same slope. This difference in intercepts creates a shift thereby resulting in poor parameter inference. Unfortunately the magnitude and direction of the shift depends on the true parameter values, so no simple bias correction is available. At first, one may think that the discretization of the parameters $\tau_2$ and $\rho$ may be causing some of these issues. We verify that this is not the case by simply computing the values of the true likelihood function at the top of the ridges obtained with the discretized likelihood. We are able to see that the likelihood surface using the discretization is similar to the true likelihood surface. The poor inference from our traditional Bayes approach is therefore clearly not a result of the discretization.

By generating additional simulations using a simpler model where we fix all the latent variables at their means we also find the full gravity model does not substantially differ from the simpler one in terms of capturing interesting biological characteristics of the underlying dynamics of the disease. In order to study the effect of this fixing on the likelihood surface, we save the true latent variables while simulating data and use them in our gridded MCMC in place of the expectations used in our gridded MCMC algorithm. The results show that using the true values of the latent variables does not change the traditional Bayes inference. This also confirms that the shifts that we observe in the traditional Bayes approach are not due to simplifying the model, but rather due to inherent problems with the likelihood function.

We note that our main interest is to examine whether the parameter estimates result in a model fit that is capable of reproducing important characteristics of the observations. In order to study the model fit from the gridded MCMC, we simulate a data set using the full gravity model with estimated values of the parameters, where here and throughout the chapter, we use modes of the corresponding posterior density functions as estimates of the parameters. These estimates for the measles data described in Section 4.6.1 are $(\theta', \tau_1, \tau_2, \rho) = (0.71, 0.5, 1, 1.48)$. For the simulated data set, we calculate the two 952 dimensional vectors (number of cities in the data) of summary characteristics and plot them against the summary
vectors for the observed measles data (Figure 4.4). We can see that the simulated data do not seem to match the actual data in terms of the maximums $M$ and the proportions of zeros $P$ (Figure 4.4 (a)-(b)). In Section 4.5.2, we compare the model fit obtained via the gridded MCMC to the model fit we obtain via our Gaussian process-based approach described in Section 4.4.

We summarize below our conclusions based on the gridded MCMC approach:

1. The confidence regions for the parameters are very wide, suggesting that there may be relatively little information even with a fairly rich data set. Hence we assume that $\tau_1 = 1, \tau_2 = 1$ as estimated in Xia et al. (2004) and study the joint distribution of $\theta'$ and $\rho$, which becomes well informed by the data.

2. The fitted gravity model, using the above inference about its parameters, does not capture important biological features of the data.

3. We find that the parameter estimates from the traditional Bayes approach are shifted and the direction of the shift varies as shown in Figure 4.3. For example, for a simulated data set using the parameters values $(\theta', \tau_1, \tau_2, \rho) = (0.71, 0.5, 1, 1)$, our attempt to infer $\rho$ assuming other parameters are known results in an estimate $\hat{\rho} = 1.5$ with a confidence region that does not contain the truth.
Figure 4.5: Approximation of a nonlinear function using a Gaussian process. Our “data” are denoted by black dots, the green curves are the independent error process without spatial dependence, red curves are linear Gaussian process model with exponential covariance, blue curves are with squared exponential covariance.

4.4 Gaussian Processes for Emulation-based Inference

Gaussian processes are used as emulators for complex computer models. Computer models describe and predict characteristics of physical processes. These models are usually very complicated and computationally expensive to simulate from. It also is generally difficult to explore and understand the connection between the input and output of computer models. A useful approach to investigate this connection is to approximate the computer model using a stochastic process (cf. Sacks et al., 1989b). The concept of using a simpler stochastic model to approximate a complex computer model is referred to as emulation. An approximating stochastic process is then called an emulator. Emulation allows to construct a tractable model that links the output of the computer model to its input. Gaussian processes are widely used as emulators due to the simplicity and flexibility they can offer.

We now illustrate how a Gaussian process-based emulator for a complex deterministic nonlinear functions works using simple examples. These examples are taken from Bhat (2010). First, consider a function $y = \sin(x)$ (shown in black dots in Figure 4.5), where we assume that we can evaluate the function at certain input locations. Using a linear Gaussian process with a constant mean, our objective is interpolate the function at locations between
Figure 4.6: Approximation of nonlinear functions using a Gaussian process. The black dots represent ‘data’ from functions \( f(x) = 2 \sin(x) \) and \( 5 \exp(-x/5) \sin(x) \). Both are fit with the same stochastic model: \( f(x) = c + \epsilon(x) \) where \( \epsilon(x) \) is a linear Gaussian process. Solid blue lines are predictions at new locations and the dotted blue lines reflect uncertainty.

In Figure 4.5, we plot our approximate function obtained via the emulator. Here, in the right plot, we use a model that accounts for spatial dependence (red curves and blue curves). In the left plot, we assume that there was no dependence between the locations (green curves). It is clear that the model that account for spatial dependence results in much more accurate approximation. We note that the approximation using the linear Gaussian process is accurate even though the mean function is misspecified. This is explained with the inclusion of spatial dependence that is sufficient for a good approximation.

In Figure 4.6, we demonstrate how a simple Gaussian process-based emulator can approximate two different non-linear functions, \( f(x) = 2 \sin(x) \) and \( f(x) = 5 \exp(-x/5) \sin(x) \). We fit both of these functions with the same Gaussian process emulator, and it can be seen that the predictions match the functions and the data points well with small uncertainty. This shows the flexibility of the linear Gaussian process-based emulator, which allows us to approximate different nonlinear functions accurately and explains the main idea of using Gaussian process-based emulators in the field of complex deterministic computer models.

Since a traditional Bayes approach for the gravity model suffers from the shortcomings described in previous sections, using ideas of computer model emulation via Gaussian pro-
cesses, we develop a method of inference that is directly linked to the characteristics of the infectious disease dynamics that are of most interest to biologists.

We describe a new two-stage approach for inferring the gravity parameters. In the first stage, we simulate the gravity model at several parameter settings. For each forward simulation of the model we can calculate the vector of summary statistics based on the simulated data set. This vector is high-dimensional, 952 dimensions in the case of measles data. Since Gaussian process-based emulation for high dimensions poses serious computational challenges, we emulate the model by fitting a Gaussian process to the Euclidean distances between the summary statistics of the simulated data at the chosen parameter settings and the summary statistics for the real data. In the second stage, we perform Bayesian inference for the observations using the GP emulator from the first stage. We also allow for additional sources of uncertainty such as observational error and model-data discrepancy as described below. We note that such two-stage approaches to parameter inference in complex models has been used to reduce computational challenges and alleviate identifiability issues (cf. Bhat et al., 2010a; Liu et al., 2009a).

We begin with some notation. Let \( Z \) denote the vector of summary statistics of interest (e.g. proportions of zeros) calculated using the observed space-time data set. Let \( \Theta \) be the gravity parameters and \( Y(\Theta) \) denote the vector of summary statistics obtained using a simulation from the gravity model with the parameter setting \( \Theta \). Let \( \Omega = (\Theta_1, \cdots, \Theta_p) \) be a grid on the parameter space. Our first goal is then to model \( D = (D_1, \cdots, D_p) \), where \( D_i \) is the Euclidean distance between \( Y(\Theta_i) \) and \( Z \) for \( i = 1, \cdots, p \). This is done in the first stage of our approach where we assume,

\[
D|\Omega, \beta_G, \xi_G \sim N(X\beta_G, \Sigma(\xi_G)) \tag{4.6}
\]

where \( X \) is a design matrix of dimension \( p \times 5 \) with \( i \)-th row equal to \( (1, \Theta_i^T) \). In other words, columns of \( X \) are the values the gravity parameters, \( (\theta, \tau_1, \tau_2, \rho) \), on the selected grid and an intercept. We use Gaussian covariance matrix, \( \Sigma(\xi_G) \), elements of which are given by,

\[
(\Sigma(\xi_G))_{ij} = \text{cov}(D_i, D_j) = \begin{cases} \sigma_G^2 \exp(-\phi_G ||\Theta_i - \Theta_j||^2), & \text{if } i \neq j \\ \sigma_G^2 + \tau_G^2, & \text{otherwise.} \end{cases}
\]

Here, \( ||a - b|| := d(a - b, a - b) \), where throughout the chapter, the function \( d(\cdot, \cdot) \) returns the
Euclidean distance between the argument vectors. \( \xi_G = (\sigma_G^2, \tau_G^2, \phi_G) \) is a vector of unknown parameters that specify the covariance matrix with a known parametric form, and \( \beta_G \) is a vector of unknown regression coefficients. Then, if we let the maximum likelihood estimate of \( (\beta_G, \xi_G) \) be \( (\hat{\beta}_G, \hat{\xi}_G) \), using standard multivariate normal theory (cf. Anderson, 1984), the normal predictive distribution for the simulated distance \( D \) at a new \( \Theta \) can be obtained by substituting \( (\hat{\beta}_G, \hat{\xi}_G) \) in place of \( (\beta_G, \xi_G) \) and conditioning on \( D \). We denote this predictive distribution by \( \eta(D; \Theta) \). Detailed version of constructing this predictive distribution (emulator) is given in Appendix B.

Consider a new space-time data set, and let the vector of summary statistics for these data be \( Y^* \). Let the distance between \( Y^* \) and \( Z \) be \( D^* \). The predictive distribution from the first stage provides a model for \( D^* \), \( \eta(D^*; \Theta^*) \), connecting it to some unknown parameter vector \( \Theta^* \).

Following Bayarri et al. (2007c), we model the discrepancy between the gravity model and the real data. Failing to account for data-model discrepancy can lead to poor inference as pointed out in Bayarri et al. (2007c) and Bhat et al. (2010b). We account for this by setting \( D^* = D^*_i := \delta \), where \( \delta > 0 \) is the discrepancy term. It is positive since it represents an Euclidean distance that is non-negative (in the unrealistic case that there is an exact match between the model for the data and the model used to fit the data, \( \delta \) would be identically equal to 0). We then infer the gravity parameters using \( \eta(D^*_i; \Theta^*) \) considering \( \delta \) to be another unknown parameter in the MCMC algorithm. In other words, the likelihood function we use for our MCMC algorithm is a function \( f(\delta, \Theta^*) := \eta(D^*_i; \Theta^*) \). We note that including a model discrepancy term results in more reliable parameter inference with narrower confidence regions since it adjusts for the fact that even the best model fit is not going to reduce the distance between the simulated and observed summary statistics to zero. In our simulated examples, where data are generated from the gravity model, the discrepancy term can be thought of as an adjustment parameter for the fact that two data sets simulated at the same parameter settings will always have small differences due to stochasticity. In these examples, as it is expected, estimate of the discrepancy is very small compared to the discrepancy term inferred from the original data. We also note that using negative values for \( \delta \) would mean an extrapolation in our emulator beyond the grid of the parameter space that may lead to unreliable inference. In many situations, having a well-defined discrepancy term with an informative prior helps to reduce problems with identifiability of the parameters as well (cf. Craig et al., 2001).
We can now summarize our inferential approach as follows:

1. Emulating the gravity model:
   
   (a) Select a grid \((\Theta_1, \ldots, \Theta_p)\) on the range of possible values for \(\Theta\).
   
   (b) Calculate \(Y(\Theta_i)\) using a simulation from the gravity model with \(\Theta_i\) for all \(i\).
   
   (c) Calculate \(D = (D_1, \ldots, D_p)\), distances from \(Y_i\) to \(Z\) for all \(i\).
   
   (d) Find the maximum likelihood estimates of \((\beta_G, \xi_G)\), the parameters of the Gaussian process in Equation (4.6). Obtain the predictive distribution \(\eta(D; \Theta)\).

2. Bayesian inference for \(\delta\) and \(\Theta^\ast\) given the observations \(Z\):
   
   (a) Using the predictive distribution with a discrepancy term, \(\eta(D^\ast_\delta; \Theta^\ast)\), perform Bayesian inference for the parameters \((\Theta^\ast, \delta)\) from the posterior distribution via MCMC.

4.5 Emulation-based Inference for the Gravity TSIR Model

In this section we describe details of the application of the inferential approach described in Section 4.4 to the gravity TSIR model. By using simulated data examples, we show that the approach resolves the problems posed by traditional approaches. In order to contrast our approach to a traditional likelihood-based approach (carried out by gridded MCMC as described in Section 4.3.1), we also provide computational details from the application of both methods.

4.5.1 Computational Details of Gridded MCMC and Emulation-based Approaches

Inference for both the traditional Bayes and emulator-based approaches relies on sampling from the corresponding posterior distributions via MCMC. In both methods, we use univariate sequential slice sampling updates for the continuous parameters (Agarwal and Gelfand, 2005; Neal, 2003a). Parameters that are on the grid are updated via an analog of a simple
random walk for discrete variables. In all the MCMC algorithms that are used for the discretized MCMC approach, the chain is run until we obtain 200,000 samples. This takes about 3 days on a Intel Xeon E5472 Quad-Core 3.0 GHz processor. In all the MCMC algorithms for the Gaussian process-based method, all the updates are carried out using slice sampling since all the parameters here are continuous. Chain lengths are 200,000 again and it takes about 10 hours to generate them. The chain lengths in both methods are adequate for producing posterior estimates with small Monte Carlo standard errors (Flegal et al., 2008; Jones et al., 2006).

We emulate the gravity model with a Gaussian process using proportions of zeros as a summary statistic of interest. Using different summary statistics may, of course, lead to different inference. Inference based on the maximums, however, was identical to what is obtained here and therefore we do not include details of the analysis and the corresponding results. It is also possible to develop an emulator using these two summary statistics at the same time; this is computationally more demanding and based on our exploratory data analysis, it will not impact our conclusions. In general, summary statistics for emulation need to be chosen to capture the most scientifically important aspects of the disease dynamics process.

We use the priors for the gravity model parameters that are described in Section 4.3. Since the discrepancy term, $\delta$, is always positive, we use an exponential(1) as its prior distribution.

We use a uniform grid in the four-dimensional cube, each side of which is equal to the intervals $[0, 2]$. For each parameter, we use 20 different values on each axis of the cube; this grid size permits computationally expedient inference. Our analysis of simulated data sets also shows that 20 is sufficient for accurate inference. In addition, for each point on the grid, the average distances from multiple forward simulations can be used instead of the distances calculated from a single simulation. This may be important when model realizations are highly variable. For the parameters of the gravity model, however, our inference was insensitive to the number of repetitions as the model had relatively small amount of stochasticity.

4.5.2 Application to Simulated Data

In the simulated examples that follow, our goal is to compare inference based on the GP-approach to inference from the traditional Bayes approach. In Figure 4.7, we show a simulated
Figure 4.7: 95% C.I.’s for $(\theta', \rho)$ obtained via different methods (assuming that $\tau_1$ and $\tau_2$ are known): Solid line shows the 95% region obtained using the traditional Bayes method. Dashed line outlines the 95% region obtained via GP emulator: (a) Both regions contain the true parameter values; (b) Region obtained by the GP emulator contains the true values of the parameters, while the traditional Bayes region does not.

example where both the GP and traditional Bayes approaches yield the same inference, and another simulated example where the two approaches yield different answers. In both cases, the emulation-based approach provides inference that captures the true parameter values. In the first simulated data, the true parameters are $\theta' = 1$, $\tau_1 = 0.6$, $\tau_2 = 1$ and $\rho = 1$. In Figure 4.7 (a), we overlay two different 95% confidence regions obtained using the two different methods. Both of these regions are found by assuming $\tau_1 = 0.6$ and $\tau_2 = 1$. We can see that for this example, both solid (traditional Bayes) and dashed (GP emulator-based) regions contain the true values of $\theta'$ and $\rho$. This shows that inference based on the GP emulator is as good as inference based on the traditional Bayes method. To demonstrate that the new approach is better than the traditional Bayes approach, we choose a second set of values for the gravity parameters ($\theta' = 0.71$, $\tau_1 = 0.62$, $\tau_2 = 1$ and $\rho = 1.5$) for which we know inference based on the traditional Bayes approach to be poor (like in Figure 4.3). Figure 4.7 (b) shows how the 95% confidence region from the traditional Bayes method (outlined with a solid line) is shifted and does not contain the truth. The permissible region obtained using the GP emulator (outlined with a dashed line) have corrected the shift and contains the true values of the parameters.

We analyze the ability of the fitted gravity model to reproduce the key characteristics of
the process at these new parameter estimates. Using estimates obtained via the GP emulator-based approach, \((\theta', \tau_1, \tau_2, \rho) = (0.71, 0.5, 0.5, 1.48)\), we generate a data set to obtain plots similar to the ones in Figure 4.4. Plots on Figure 4.8 (a)-(b) show that the model now can fit the maximums \(M\) and the proportions of zeros \(P\) very well. Comparing the plots in Figures 4.4 and 4.8, we can now say that the new emulation-based approach improves the model fit substantially while the traditional Bayes parameter estimates from the gridded MCMC fail to provide a model that captures the key epidemiological features of the data.

![Figure 4.8: Characteristics of simulated data at the parameters chosen to minimize the discrepancy between the data and the simulation: (a) Simulated \(M\) vs \(M\) from the data; (b) Simulated \(P\) vs \(P\) from the data.](image)

In order to study the effect of a discrepancy term in our approach, we also tried to infer the gravity parameters using the emulation-based model with \(\delta = 0\) (no discrepancy). The resultant 95% confidence regions were much wider for the latter approach containing incorrect parameter settings, supporting the points made in Bayarri et al. (2007c) about the importance of adding a discrepancy term to approximate models. We also tried a few different priors; using the exponential(1) prior for the discrepancy term worked very well as was clear from the results: the posterior median for the discrepancy term was found to be around 2 which was close to the minimal distance from the simulated and the true vectors of summary statistics taken over all the points on the grid.
4.6 Results from Application to Measles Data

We apply our emulation-based approach to inference for the gravity TSIR model to a well known measles data set from the U.K. The purpose of this is twofold: to demonstrate the applicability of our approach to a real data set as well as to provide some insights into measles dynamics in the pre-vaccination era.

4.6.1 Description of Measles Data

The following description of the data closely follows Xia et al. (2004). We analyze weekly case reports of measles for the 952 locations in England and Wales from 1944 to 1967. The data represent an interesting case study of spatiotemporal epidemic dynamics (Grenfell et al., 2002) with well understood underreporting rate of 40%-55% (Bjørnstad et al., 2002a). Besides the under-reporting, the data are complete and reveal inter-annual outbreaks of infection. A critical feature of this data set is that, except for a few large cities, infection frequently goes locally extinct, so that overall persistence hinges on episodic reintroduction and spatial coupling. Before further analysis, we correct the reported data by a factor of 1/0.52, with 52% being the average reporting rate taken from previous analysis (Bjørnstad et al., 2002a; Clarkson and Fine, 1985; Finkenstäd and Grenfell, 2000). In addition, as in previous works, we use a timescale that represent the exposed and infectious period, which is known to be about 2 weeks for measles (Black, 1989).

Following a standard assumption in the literature (see, for instance, Bjørnstad et al., 2002a; Grenfell et al., 2002; Xia et al., 2004, and the references therein), the population sizes and per capita birth rates for all locations in this work are assumed to be approximately constant throughout the time period. These variables are taken as those in 1960 for each of the areas. This is a rough approximation, since most communities grew during the period we analyze. The force of infection is, therefore, on average slightly underestimated (overestimated) during the early (late) part of the study.

4.6.2 Some Implications for Measles Dynamics

Important biological questions we want to answer based on these data are: (i) do the gravity model parameters (and hence disease transmission) change for different time periods? Do they change for school holiday periods versus non-holiday periods? (ii) do movement rates
of infected people change in different time periods?

We first fit the model to the data from 1944-1955 and 1956-1967 separately. As demonstrated in our simulated examples in Section 4.3.2 and 4.5.2, it is not possible to infer all the gravity model parameters at once. Hence, we set the parameters $\tau_1$ and $\tau_2$ equal to 1 and study the remaining key gravity model parameters $\theta$ and $\rho$. The resulting 95% confidence regions for $\theta'$ and $\rho$ are provided in Figure 4.9. Based on this, we conclude that the change in parameter values is statistically insignificant for these two different time periods. Figure 4.10 shows confidence regions obtained by the GP emulator-based approach by fitting the model to the parts of the data corresponding to periods of holidays and non-holidays. As can be seen, the two regions are almost identical again, indicating that any change in the number of incidences for holidays and non-holidays is not due to the change in the way the infection spreads between cities of the metapopulation.

Since the matrix $M = \{m_{kj}\}$, where $m_{kj} = \theta' N_{kl}^\tau \sum_{t=1}^T \frac{(I_{jt})^{\tau_2}}{d_{kj}}$ is interpreted as a matrix of the amount of movement, sum of $k$-th row of $M$ represents the amount of infected individuals leaving city $k$ while sum of $k$-th column is the number of infected people coming to city $k$. Using samples for $\theta'$ and $\rho$, we easily obtain a sample for the spatial flux of infection for selected cities. In Table 1, we report our estimates with corresponding credible regions based on this analysis. We use the posterior median as point estimates. For example, we estimate the average number of emigrating infections during the holiday periods each week to be equal to 31.1 for London. Below the estimate, we report a 95% credible interval for it which is (4.4, 479.1). Based on these estimates, the mobility of the infection appears to be less during the periods of holidays. Table 2 shows estimates of the average amount of transit infections each week for years 1944-55 and 1956-67. We see here that the infection appears to move less during the later years. Note that none of the differences are statistically significant. Note, however, that these analysis are not meant to represent a comprehensive epidemiological treatment, rather they are meant as an illustration of how the GP-approach permits rigorous statistical inference for the gravity TSIR model.

4.7 Discussion

Complex models are very useful for representing physical phenomena, whether the phenomena is the spread of an infectious disease or the change in sea surface temperatures in the Atlantic.
Figure 4.9: 95% C.I. for \((\theta', \rho)\) obtained via fitting GP emulator to a part of the data: Solid line outlines the confidence region for parameters when data for years from 1944-55 is used; Dashed line outlines the confidence region for parameters when data for years from 1956-67 is used.

Table 4.1: Estimated amount of average movement in two weeks

<table>
<thead>
<tr>
<th>City</th>
<th>From</th>
<th>To</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Holiday</td>
<td>Non-Holiday</td>
</tr>
<tr>
<td>London</td>
<td>31.1</td>
<td>46.6</td>
</tr>
<tr>
<td></td>
<td>(4.4, 479.1)</td>
<td>(6.6, 744.7)</td>
</tr>
<tr>
<td>Birmingham</td>
<td>7.5</td>
<td>10.8</td>
</tr>
<tr>
<td></td>
<td>(1.2, 72.9)</td>
<td>(1.8, 110.6)</td>
</tr>
<tr>
<td>Manchester</td>
<td>7.8</td>
<td>10.3</td>
</tr>
<tr>
<td></td>
<td>(1.0, 151.4)</td>
<td>(1.4, 180.9)</td>
</tr>
<tr>
<td>Blackpool</td>
<td>0.8</td>
<td>1.1</td>
</tr>
<tr>
<td></td>
<td>(0.1, 6.7)</td>
<td>(0.2, 8.8)</td>
</tr>
</tbody>
</table>

As is well known, it is not always possible for every aspect of such complicated phenomena to be modeled accurately; certain key characteristics of the process necessarily have to be focal points of the modeling effort. However, these key characteristics are not typically the focus of a statistical inferential procedure that uses a traditional likelihood-based approach. The approach we have described in this chapter addresses this point by providing a flexible inferential method that directly takes into account the characteristics of the process that are most important to scientists. Even though focusing on different summary statistics can lead to different estimates, parameter inference based on our approach produces an improved model fit to the biologically interesting features of the infectious disease dynamics. In addition
Figure 4.10: 95% C.I. for $(\theta', \rho)$ obtained via fitting GP emulator to a part of the data: Solid line outlines the confidence region for parameters when data only holidays periods is used; Dashed line outlines the confidence region for parameters when data for only non-holiday periods is used.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>London</td>
<td>49.3</td>
<td>42.3</td>
<td>53.6</td>
<td>45.4</td>
</tr>
<tr>
<td></td>
<td>(8.3, 532.4)</td>
<td>(5.4, 679.9)</td>
<td>(8.9, 587.5)</td>
<td>(5.6, 779.8)</td>
</tr>
<tr>
<td>Birmingham</td>
<td>11.6</td>
<td>10.3</td>
<td>12.8</td>
<td>10.4</td>
</tr>
<tr>
<td></td>
<td>(2.1, 80.2)</td>
<td>(1.5, 109.7)</td>
<td>(2.3, 88.4)</td>
<td>(1.5, 109.5)</td>
</tr>
<tr>
<td>Manchester</td>
<td>11.0</td>
<td>9.9</td>
<td>12.9</td>
<td>9.7</td>
</tr>
<tr>
<td></td>
<td>(1.8, 139.1)</td>
<td>(1.2, 189.2)</td>
<td>(2.1, 157.8)</td>
<td>(1.2, 184.6)</td>
</tr>
<tr>
<td>Blackpool</td>
<td>1.2</td>
<td>1.0</td>
<td>1.0</td>
<td>0.65</td>
</tr>
<tr>
<td></td>
<td>(0.2, 7.8)</td>
<td>(0.1, 9.4)</td>
<td>(0.2, 6.4)</td>
<td>(0.1, 5.9)</td>
</tr>
</tbody>
</table>

to the flexibility this provides, we find that our approach is also computationally tractable in situations where traditional likelihood-based inference is not.

Computer model emulation and calibration is an active area of research (cf. Bayarri et al., 2007a; Bhat et al., 2010b; Bhattacharya, 2007; Conti and O’Hagan, 2010; Higdon et al., 2008; Rougier and Beven, 2009; Rougier et al., 2009; Sansó and Forest, 2009) but most of this work has focused on deterministic models. In this chapter, we describe an emulation-calibration approach for probabilistic models. In our view, therefore, our work makes the following main contributions: (1) a general inferential approach that focuses on characteristics (summary statistics) of a process; (2) a method for statistical inference when the likelihood is intractable
and simulation from the probability model is expensive; (3) a study of a particular model for measles dynamics, the gravity TSIR model, using the approach we have developed.

In the context of measles in the pre-vaccination era, our method allows us to study some interesting aspects of the dynamics based on the gravity TSIR model. For instance, we find that there does not appear to be a significant change in the gravity parameters for the school holiday periods versus non-holidays which means that we do not have enough evidence of a change in the dynamics of measles between these different periods. The gravity parameter estimates for years 1944-55 and 1956-67 were also not statistically different.

The methodology we have described in this chapter is particularly useful in cases where simulation from a probability model might be too expensive to allow the use of other popular inferential approaches like ABC. It is worth noting that our approach works well when the parameter dimensionality is small, but is generally infeasible for parameter dimensions greater than around five to eight depending on the model complexity. Our approach is widely applicable for inference in computationally expensive but biologically realistic models. In principle, whenever a likelihood is expensive to evaluate or when traditional Bayes approaches does not capture the most scientifically relevant features of the model, our method provides a way to incorporate important characteristics in a computationally tractable inferential approach.
Chapter 5

Emulator-Based Inference: Generalizations and Extensions

In this chapter, I will first summarize the inferential approach introduced in Chapter 4 in a general setting. I will then demonstrate the performance of the ABC and emulator-based approaches on toy examples. I will also be undertaking a brief overview of the methods of inference that involve “indirect likelihoods”, composite likelihoods, and methods where Gaussian processes can be used for density estimation. Finally, I will provide an interesting application of our emulator-based approach to an inference problem for a mixture random graph model which is a useful model for metabolic reactions networks.

5.1 Emulator-Based Inference: General Overview

First, let us recall the notation introduced in Chapter 2. We have a probabilistic model defined by a set of probability distributions \( \{\pi(y|\theta)\} \) that depend on some unknown \( d \)-dimensional parameter \( \theta = (\theta_1, \cdots, \theta_d) \). We let \( D \) denote the data and \( U(\theta) \) is a simulation from \( \pi(y|\theta) \). If \( S(\cdot) \) is a function that returns some quantitative summary characteristics of the data (possibly multidimensional), let \( Z = S(D) \) and \( Y(\theta) = S(U(\theta)) \). Let \( \Theta = \{\theta_1, \cdots, \theta_p\} \) be a set of parameters selected from the range of possible values for \( \theta \) and \( U = \{U_1, \cdots, U_p\} \), where \( U_i = U(\theta_i) \). Finally, let \( \mathbf{Y} = (Y_1, \cdots, Y_p) \), where \( Y_i = S(U(\theta_i)) \).

The objective is then to infer the posterior distribution of \( \theta \) given the observed \( Z \) and simulated summaries statistics \( \mathbf{Y} \).
In general, our method used in Chapter 4 is a two stage procedure to infer $\theta$. It is based on using Gaussian processes. The fact that we split the method into two can be thought of as a way of “cutting feedback” (cf. Rougier, 2008) or modularization (Liu et al., 2009a). In the first stage, we approximate the model by assuming that $Y$ can be modeled as a Gaussian process, $Y|\beta, \Theta, \xi \sim N(\mu_\beta(\Theta), \Sigma_Y(\xi_Y))$, where we assume a linear mean function, $\mu_\beta = X\beta$, with $X$, a covariates matrix of dimension $p \times (1 + d)$ with rows $X_i = (1, \theta_{i1}, \ldots, \theta_{id})$, where $i = 1 \cdots p$. $\xi_Y$ is a vector of covariance parameters that specify the covariance matrix $\Sigma_Y(\xi_Y)$ with a known form, and $\beta$ is a vector of regression coefficients. Then, if we let the maximum likelihood estimate of $(\xi_Y, \beta)$ be $(\hat{\xi}_Y, \hat{\beta})$, using standard multivariate normal theory (cf. Anderson, 1984), the normal predictive distribution for any $Y$ at a new $\theta$ can be obtained by substituting $(\hat{\xi}_Y, \hat{\beta})$ in place of $(\xi_Y, \beta)$ and conditioning on $Y$ (this is known as kriging in geostatistics). We denote this predictive distribution by $\eta(Y, \theta)$. In the second stage, since the predictive distribution $\eta(Y, \theta)$ allows us to connect any new $Y$ with $\theta$, we set $Y$ to be equal to $Z$ and use the resulting likelihood function to infer the unknown parameters of the model via MCMC. Note that in the gravity model example in Chapter 4, $S(U) = d(P(U), P(D))$, where $d(\cdot, \cdot)$ is the Euclidean distance and $P(\cdot)$ returns the proportions of zeros in the data.

5.2 Simulated Examples

In this section we illustrate, via simulated examples, how the ABC and emulator-based inference methods described in the previous chapters work. The goal is to explore how different algorithms compare to each other and how various characteristics of the algorithms affect their performance. At the end of the section, we list our conclusions based on the analysis of different simulated scenarios.

In our example, we generate a sample of $n = 100$ numbers from the Gamma($\alpha = 5, \beta = 2$), where $\alpha$ and $\beta$ are the shape and scale parameters of the distribution respectively. Pretending that we do not know the true values of $\theta = (\alpha, \beta)$, we use ABC and emulator-based approaches to infer the unknown parameters from simulated data.

Figure 5.1 shows the true likelihood surface of the parameters for the data. As can be seen, any point inside the blue area is as good as the truth in maximizing the likelihood
indicating that we have an issue with identifiability in the parameters. Figure 5.2 shows the area for the possible values of $\theta$ obtained via using an ABC algorithm based on rejection sampler (Algorithm 1 described in Chapter 2). Here, different values of $\epsilon$ are used. These values are 60, 70, 80 and 90. The corresponding acceptance rates of the algorithm are in decreasing order: 0.0167, 0.0726, 0.1437 and 0.2245. We note that values of tolerance levels smaller than 50 result in all the proposals being rejected and therefore are not considered.

![Figure 5.1: True likelihood surface of the data.](image)

Figure 5.3 shows the posterior density of the parameters via an ABC algorithm based on MCMC (Algorithm 2 described in Chapter 2) with a fixed $\epsilon$ and different values of $R$ (repetitions). These values are 10, 50, 100 and 500 with the acceptance rates of the resulting algorithms equal to 0.204, 0.228, 0.182 and 0.195 respectively.

Finally, Figure 5.4 shows the approximate posterior likelihood surface of the parameters obtained via our emulator-based approach. Here, we use different grid sizes ($g = 5$ and 10) and different numbers of repeated simulations at each point of the grid ($r = 1$ and 5).

As a summary, based on this and other various simulations from different one and two dimensional probability distributions, one can draw the following conclusions about the ABC and emulator-based approaches:

- As can be seen from Figure 5.2, the tolerance level $\epsilon$ is a very important parameter
Figure 5.2: Inference based on Algorithm 1 with different tolerance levels $\epsilon$: (a) $\epsilon = 60$; (b) $\epsilon = 70$; (c) $\epsilon = 80$; (d) $\epsilon = 90$.

Figure 5.3: Inference based on modified Algorithm 2 with fixed tolerance levels $\epsilon = 80$ and different values of $R$: (a) $R = 10$; (b) $R = 50$; (c) $R = 100$; (d) $R = 500$. 
in the ABC algorithms. Small $\epsilon$ results in a very low acceptance rate. Increasing of $\epsilon$ increases the acceptance rate, but may produce a poor approximation to the likelihood. Therefore, the optimal choice of $\epsilon$ needs to be studied very carefully. This is an active area of current research in the literature of ABC. In a recent paper, Ratmann et al. (2009) suggests to incorporate $\epsilon$ in the algorithm and concludes that inferred value of optimal $\epsilon$ can be used as a model selection criteria.

- The choice of $R$ affects the smoothness of the ABC posterior, but does not have an obvious effect on the acceptance rate of the algorithm. The increase in $R$ does create a smoother posterior surface, but adds an additional computational complexity to the algorithm.

- When simulations from the model cannot be generated quickly, direct ABC algorithms are very inefficient.

- Given that the unknown likelihood function is reasonably smooth, the GP emulator-based approach works well even for a small number of points on the grid. Depending on the variability of the unknown parameters and the complexity of the posterior density.
function, repetitions can result in better approximation of the true likelihood function. After one reaches a certain level, however, this effect seem to become insignificant. Similarly, the “density” and “concentration” of the grid at certain areas of the parameter space may impact the accuracy of emulator-based inference.

- Since all simulations from the model and calculations of the related summary statistics are done at each point of the parameter grid in parallel, GP-based approach can be used in situations where ABC methods are not effective due to a slow simulation process.

Consequently, several interesting and important questions about the inference approaches based on an emulator and ABC that may emerge from our simulation analysis are:

(i) How can one choose a grid on the parameter space?

(ii) How can we determine the necessary number of repetitions for each point on the grid?

(iii) What are the properties of the GP emulator-based approach as the size of the dataset increases?

(iv) For a fixed amount of data, does the approximate likelihood function obtained via the GP emulator converge to the true likelihood function as we increase the number points on the grid and the number of forward simulations at each point of the grid?

(v) What summary statistics can we choose for the convergence in (iv), if any?

(vi) Can we analyze the appropriate error?

We return to a more detailed discussion of these and some other related theoretical questions in Chapter 6.

5.3 Indirect Inference

When the model is too complicated and the likelihood functions are not readily available, there is another class of methods based on so-called “indirect inference”. Methods of “indirect inference” have been developed and used in the field of econometrics where they have proved valuable for parameter estimation in highly complex models. In indirect inference methods, statistical inference is based on intermediate statistics, which often follow an asymptotic
normal distribution, but may not necessarily be consistent estimators of the parameters of interest. These intermediate statistics can be simple estimators based on convenient, but possible misspecified models, sample moments or solutions to estimating equations. The objective functions of these methods are interpreted as “indirect likelihoods” based on the intermediate statistics (cf. Jiang and Turnbull, 2003, 2004). The simplest example of indirect inference methods is the popular method of moments.

More generally, if we have data of $n$ independent samples, the basic steps of indirect inference methods are as follows (Jiang and Turnbull, 2004):

1. Suppose that our data come from a model with a distribution $P(\theta)$, which is indexed by an unknown $d$ dimensional parameter of interest $\theta$.

2. In indirect inference, one first computes an intermediate or auxiliary statistic $\hat{s} = S(P(n))$, which is a functional of the empirical distribution function $P(n)$.

3. A bridge relationship $s = S(P^\theta)$ is then defined. The unknown quantity $s$ is called the auxiliary parameter.

4. With the auxiliary estimate $\hat{s}$ replacing $s$, the bridge relationship above is used to compute an adjusted estimate $\hat{\theta}(\hat{s})$ for $\theta$.

We can compare the methods of indirect inference to our emulator-based approach by noting that since the methods that utilize the steps above are based on constructing an alternative likelihood function via auxiliary statistics, in some sense, our emulator-based inferential approach developed in Chapter 4 can also be classified as an indirect likelihood-based approach. However, there are advantages of our method over the classical indirect inference methods: (i) it is very flexible and scientists can choose to focus on the most important aspects of the data; (ii) it is more general and does not impose any restrictions on auxiliary statistics; (iii) does not require solving bridging equations; and (iv) easily accounts for model-data discrepancy.

5.4 Other Approaches to Inference with Complicated Likelihoods

In this section, we aim to overview several other approaches proposed in the literature that are based on using alternative or approximate likelihood functions. At the end of the section, we describe an application of Gaussian processes in problems of density estimation.
Composite Likelihood

A popular classic approach to deal with large datasets and very complex models is based on using composite likelihood-based methods (cf. Lindsay, 1988). These methods assume that computing the likelihood for certain subsets of the data is possible, and therefore one may construct a pseudo-likelihood by combining such likelihood objects and use this as a surrogate for the true likelihood. Using our notation, composite likelihood has the following form (Lindsay, 1988),

\[ \text{CL}(\theta; D) = \prod_{k \in K} L_k(\theta, D)^{\omega_k}, \]

where \( L_k(\theta, D) = P(D \in A_k) \), \( \{A_k\}_{k \in K} \) are some subsets of the set of all values of \( D \) and \( \{\omega_k\}_{k \in K} \) are a set of suitable weights. Some of more recent applications of composite likelihood-based methods include inference for longitudinal data (cf. Molenberghs and Verbeke, 2005), survival analysis (Fiocco et al., 2009), geostatistics (cf. Curriero and Lele, 1999; Heagerty and Lele, 1998) and genetics/bioinformatics (Mardia et al., 2009; Tamura et al., 2007). In the context of the infectious disease model in Section 4.2, the major computational expense comes from calculating the gravity matrix. In this problem, composite likelihood ideas could be useful for reducing the size of this matrix by taking products of the approximate likelihoods over smaller regions.

Alternatives to ABC

Wood (2010) proposes an alternative to ABC for statistical inference in nonlinear ecological dynamical systems. The first step in this approach is to reduce the data \( D \) to a summary statistics \( s \) designed to capture the dynamic structure of the model. Then, it is assumed that

\[ s \sim N(\mu_{\theta}, \Sigma_{\theta}). \]

The unknown mean vector, \( \mu_{\theta} \), and the covariance matrix \( \Sigma_{\theta} \), are generally intractable functions of the unknown parameter \( \theta \). These functions, for any \( \theta \), are estimated by simulations from the model of \( R \) datasets for each \( \theta \).

Henderson et al. (2009) describes an emulator-based approach for a DNA deletions model. In order to explain this method, let us assume that we have data for \( \{z_i\} \), measurements of \( \{y_i\} \). Let \( y(u) \) denote a sample of latent data from the simulation model using inputs \( u \). In Henderson et al. (2009), authors propose to emulate \( \rho(u) := \text{logit}(p(u)) \), where \( p(u) \) and \( y(u) \)
are connected via $y(u) = -\log(1 - p(u))$. In their emulation, they first assume that

$$\rho(u) = N(\eta(u), \exp(2\xi(u)))$$

where $\eta()$ and $\xi()$ are some unknown functions. Further, the authors assume that

$$\eta(\cdot) \sim \text{GP}(m_\eta(\cdot), c_\eta(\cdot, \cdot)) \quad \text{and} \quad \xi(\cdot) \sim \text{GP}(m_\xi(\cdot), c_\xi(\cdot, \cdot))$$

Finally, these functions are estimated by fitting two independent Gaussian processes to summary statistics obtained from $y(u)$. Even though it is explained that these summary statistics can be empirical estimates of $\eta$ and $\xi$, we note that it is not clear that they have scientific interpretations.

In contrast to Henderson et al. (2009), in our methods in Chapter 4, we propose to emulate the scientifically important summary statistics calculated directly from $\{z_i\}$ (or $\{z(u)\}$). This makes our approach very useful when the goal is to obtain parameter estimates with their associated uncertainty that results in a model fit that captures the important aspects of the biological process well. In our emulator, we account for the fact that there might be a discrepancy between the output of the biological model and the data. Additionally, in order to fit our emulator, we do not need to transform the latent $y$’s and verify the normality assumptions. Therefore, even though the method described in Henderson et al. (2009) shares some ideas with our approach in solving an inferential problem, we believe that our method is more generally applicable and computationally easier to implement.

**Gaussian Processes for Density Estimation**

In addition the applications described in the previous chapters, Gaussian processes are also used in constructing generative models for unknown probability density functions (cf. Adams et al., 2009). These Gaussian process-based density samplers can be useful in estimating unknown normalized/unnormalized density functions. The main idea of Gaussian process-based density estimation methods is to allow nonparametric Bayesian estimation where the prior is specified via a Gaussian process covariance function that encodes the intuition that “similar data should have similar probabilities”. Under certain conditions, it is possible to establish posterior consistency properties of Gaussian process priors (cf. Tokdar and Ghosh, 2007). We note, once again, that the goal of these methods is to learn about the unknown density functions nonparametrically. In contrast to these methods, when we use Gaussian
processes in our emulator, our goal is to construct an approximate probability model that focuses on capturing the aspects of the data that is of scientific interest. This approximate model is then allows us to infer the unknown parameters of the original model.

5.5 Inference for a Mixture Random Graphs Model

In this section, I demonstrate the applicability of the emulator-based approach in solving a challenging inferential problem for a mixture random graph model.

5.5.1 Background

Let us suppose we have relational data that consists of pairwise measurements, such as presence or absence of links (edges) between pairs of objects (nodes or vertices). These type of data can arise in the analysis of protein-protein interactions, gene regulatory networks, Web-pages networks, and social networks. A popular approach to model such data is based on using random graph models. In random graph models, the nodes of the graph are assumed to be fixed and known \((N)\), while the edges in the graph are determined by a set of random variables. Defining the graph in terms of these random variables then makes it possible to use the tools of probability and statistics in order to perform analysis, simulation, and inference for random graphs. The history of development of complex random graph models dates back to at least 1959 (Erdős and Rényi, 1959; Gilbert, 1959). In Gilbert (1959) and Erdős and Rényi (1959), the authors develop a random graph model in which the edges between nodes exist with a common probability that is independent of all the other edges. This model is commonly referred to as the Gilbert - Erdős - Rényi model. The Gilbert - Erdős - Rényi model falls into the class of random graph models called the exponential-family random graph models (ERGM). For a detailed overview of the history and development of the ERGM framework, readers are referred to (cf. Carrington et al., 2005, Chapter 8, and references therein).

As can be seen, the major advantage of the Gilbert - Erdős - Rényi model is its simplicity. This model provides a dyadic independence model for relational data controlled with only one parameter \((p)\). It is often relatively easy to study this model and various theoretical properties of the Gilbert - Erdős - Rényi model are well known (cf. Janson et al., 2000). The biggest disadvantage of the model is its lack of flexibility and realism. In relational data, the
underlying networks are often structured in unknown classes (functionally related proteins or social communities) with different connectivity properties. The Gilbert - Erdős - Rényi model for such type of data does not always offer a good model fit. In particular, the Gilbert - Erdős - Rényi model may fail to capture important properties of the networks such as degree distributions and clustering coefficients (cf. Daudin et al., 2008). For example, the Gilbert - Erdős - Rényi model cannot be made to produce an arbitrary degree distribution; for given values of $N$ and $p$, the distribution for the degree of any node is binomial with parameters $N - 1$ and $p$. In this chapter, we consider a class of models to fit the relational data that are called mixture random graph models (cf. Nowicki and Snijders, 2001). These models assume that the nodes of the network are partitioned into several unobserved classes and that the probability distribution of the relation between two vertices depends only on the classes to which they belong.

Mixture random graph models provide a proper probabilistic framework and a good model fit to data. However, inference for these models is extremely difficult. When the number of nodes of the network is small ($< 200$), one can study the unknown parameters of the mixture graph models using a Bayesian approach via data augmented MCMC algorithms (cf. Britton and O’neill, 2002; Nowicki and Snijders, 2001). When the number of nodes of the network is large, Bayesian inference via Monte Carlo methods is infeasible and only existing approaches for inference for mixture graph models are based on approximation techniques like variational methods (cf. Airoldi et al., 2008; Daudin et al., 2008; Quang Vu et al., 2012). These methods require problem-specific adjustments and, in general, do not have a straightforward way to quantify uncertainty in the parameters. In the following, I demonstrate the potential of our emulator-based approach in solving this challenging inference problem in mixture random graph models.

The key idea of the mixture graph models is in constructing a mixing matrix which specifies the probability of connection between two classes. The inference of the mixing parameters can be easy if the classes are defined using external information. However, the inference is extremely computationally expensive when classes and mixing parameters have to be inferred from the data for the network without any covariances.
5.5.2 Mixture Graphs Models

We now describe the stochastic mixture model for graphs, a model which explicitly describes the way edges link the nodes of the graph accounting for possible heterogeneities among the nodes (cf. Nowicki and Snijders, 2001). The mixture model assumes the nodes are spread into \( Q \) classes with probabilities \( \pi_1, \ldots, \pi_Q \). For each node \( i \), we let \( Z_{i,k} \) denote the indicator functions, such that
\[
\pi_k = Pr(Z_{i,k} = 1) = Pr(i \in k\text{-th class}),
\]
where \( \sum_k \pi_k = 1 \) and \( \sum_k Z_{i,k} = 1 \) for each \( i \). Then we let \( \pi_{q,k} \) denote the probability for a node from class \( q \) to be connected to a node from class \( k \). When we assume that the graph is undirected, these probabilities are symmetric:
\[
\pi_{q,k} = \pi_{k,q}.
\]
Finally, we assume that edges \( X_{i,j} \) are conditionally independent given the classes of nodes \( i \) and \( j \):
\[
\begin{align*}
X_{i,j} | i \in q, j \in k & \sim \text{Bernoulli}(\pi_{q,k}), & \text{if } i \neq j, \\
X_{i,j} & = 0, & \text{otherwise}.
\end{align*}
\]
We note therefore that this model describes the edges and the structure of the nodes of the network via the connectivity matrix \( P = (\pi_{q,k}) \). As a demonstration of the model, Figure 5.5 shows a graph simulated from a mixture model with 7 nodes from 3 different classes with varying between class connectivity probabilities.

5.5.3 An Example: Affiliation Network

A mixture model for graphs can be used to model different network structures. These include random networks, product connectivity networks, star pattern networks, and affiliation networks. In this section, we aim to give more details about a type of networks called the affiliation networks. The affiliation networks are networks in which actors are joined by a common participation, for example, in social events, or companies boards or scientists’ co-authoring papers. Another interesting application of these networks includes modeling protein metabolic reactions where each node is a protein reaction and two reactions are said
to be connected if they share common chemicals. Affiliation networks can be modeled as a specific case of a mixture model. The connectivity matrix of affiliation networks has the following form:

\[
P = \begin{pmatrix}
1 & \xi & \cdots & \xi \\
\xi & 1 & \cdots & \xi \\
\vdots & \vdots & \ddots & \vdots \\
\xi & \xi & \cdots & 1
\end{pmatrix},
\]

where \(\xi\) is the probability of connection between different classes. The effect of this probability on the graph can be seen in Figure 5.6. When \(\xi\) increases, there tend to be more edges between the nodes from different classes. We note that this model can also be thought of as a generalization of the Gilbert - Erdős - Rényi model: if the number of nodes is the same as the dimension of the connectivity matrix \(P\), the affiliation network becomes the Gilbert - Erdős - Rényi model with a probability of an edge equal to \(\xi\). However, for realistic datasets, the number of nodes in the affiliation networks is often greater than the number of classes \((N >> Q)\).
5.5.4 Emulator-Based Inference for Affiliation Networks

It is easy to see that inference for affiliation networks is problematic. This is due the dimensionality of the data and the amount of latent indicators $Z_{i,j}$. As was mentioned above, when the number of nodes in the network is large, only existing approaches to study these networks are based on approximations of the likelihood function via variational methods. In this section, our goal is to demonstrate that the Gaussian process-based inferential approach can be generalized to learn about the unknown parameters of an affiliation network.

Assuming that we have data, we now describe the application of our two-stage inferential approach to learn about unknown parameters of an affiliation network discussed in Section 5.5.3. In the first stage of the approach, we simulate data from the affiliation network model at several parameter settings. Each one of these simulations results in high dimensional matrices that indicate the presence of an edge between corresponding nodes. The dimensionality of these matrices is equal to the number of nodes in the actual data. Then we emulate the model by fitting a Gaussian process to the Euclidean distances between these matrices and the matrix that corresponds to our data. In the case when we have more than one simulation...
at each parameter setting, we fit our emulator to the average distances between the matrices simulated with the same parameter values and the data. Our study of multiple simulated data examples shows that this is a reasonable approach to reduce computational costs. In the second stage, we perform Bayesian inference for the observations using the GP emulator from the first stage. In this stage, we also allow for additional sources of uncertainty such as observational error and model-data discrepancy as was described Section 4.4 (with summary statistics function equal to the identity). For detailed version of our approach, readers are referred to Section 4.4 and Appendix B.

In the following, we simulate a graph with \( Q = 6 \) classes. In this graph, we use \( N = 600 \) and \( N = 1000 \) number of nodes. The true class probabilities used for the simulation are \( 0.25, 0.35, 0.12, 0.1, 0.1 \) and \( 0.08 \). The between class connectivity probability is equal to \( 0.2 \). Assuming that these parameters are unknown, our goal is to recover them using an emulator.

In the first stage, in order to create a grid for the class probabilities, we sampled 500 points from a six dimensional Dirichlet distribution with concentration parameters equal to 1. This guarantied that sum of these parameters was always equal to 1. As an illustration, two dimensional slices the obtained grid is plotted in Figure 5.7. For the between class connectivity probability, we used 10 equally spaced points on the interval \([0, 1]\). In the second stage of our approach, we assume that class probabilities have a Dirichlet distribution with equal concentration parameters as its prior distribution. Additionally, we use a uniform prior on \([0, 1]\) for the between class connectivity probability. In order to sample from the resulting approximate posterior distribution of the parameters, we use MCMC with univariate sequential slice sampling updates for the parameters (Agarwal and Gelfand, 2005; Neal, 2003a).

Figures 5.8 and 5.9 show the form of the approximate posterior surfaces calculated based on samples obtained via MCMC with an emulator. Here, the number of samples used was equal to 200,000 to guarantee small Monte Carlo standard errors (Flegal et al., 2008; Jones et al., 2006). In Figure 5.8, we see that our inferential approach recovers the true parameters for some of the class probabilities and for the \( \xi \). Here, the vertical lines represent the true values of the parameters that were used to generate these data. Also, in this figure, we see the posterior surface for the discrepancy term, which, in this case, can be interpreted as the inherent variability in the output of the true model. Figure 5.9, however, shows that certain class probabilities are not identified from the data. This is a common problem for mixture random graphs.
As a short summary, based on these and other simulated examples of inference for affiliation networks using an emulator-based approach, one can make the following conclusions. First, our emulator-based approach seem to work well in inferring the identifiable parameters of the affiliation network as was shown in Figure 5.8. However, we find that inference is sensitive to the number of repetitions used to generate data at each point of the grid. We recall that this was not the case in Chapter 4, when we used an emulator-based inference for the gravity model for measles. This strong sensitivity is related to greater variability of the output of the affiliation network. Second, we notice that the non-identifiability in the parameters increases by the size of $\xi$. This is intuitively clear and explainable as increased $\xi$ reduces the distinction between clusters formed by the nodes belonging to same classes. The problem of identifiability in mixture models is very common (cf. Nowicki and Snijders, 2001; Snijders and Nowicki, 1997), as was mentioned above. We are currently working on develop-
Figure 5.8: Inference for identifiable parameters

Inference methods to resolve these issues by selecting appropriate summary statistics of the graph and considering only the posterior distributions of those functions of the parameters that are invariant with respect to relabeling the classes. Alternative and potentially useful approach to increase identifiability in parameters would be to construct strong priors to identify class labels based on external data.

We note that when inferring the unknown parameters of the mixture graph model using our emulator-based approach, we assume that the number of different clusters in the model is known. Some possible approaches to choosing the number of component of the mixture model are discussed in Daudin et al. (2008). These methods are based on so-called ICL criterion and may not result in reliable inference for large networks. For details about this and the related issues about a more general model selection problem in the context of random graphs, readers are referred to Quang Vu et al. (2012). In addition to the issues discussed above, therefore, it would be interesting to develop an emulator-based method that incorporates a way to simultaneously learn about the best model (in the simplest case, the number of clusters in the model) from available data in a more general setting.
Figure 5.9: Flat posterior surface for $\pi_5$
6 Discussion and Future Work

In this chapter, I summarize the dissertation and discuss some of the possible avenues for future work.

6.1 Summary and Contributions

In my dissertation, I develop a general inferential approach for models when the likelihood functions are not available and/or when likelihood-based inference may not result in a fitted model that captures the most important aspects of the biological process. This approach is particularly useful in cases where simulations from a probability model might be too expensive to allow the use of other popular inferential approaches like approximate Bayesian computation (ABC).

As an overview of existing methods for models with intractable likelihoods, in Chapter 2, I provide a short discussion and outline of the methods based on approximate Bayesian computation. In this chapter, to highlight the advantages and disadvantages of ABC, I describe an application of the ABC-based inference for a new model for meningitis dynamics. This model, which we have developed in our paper titled “A compartmental model for meningitis: separating transmission from climate effects on disease incidence”, along with the ABC inferential approach to learn about its parameters, allows scientists answer important epidemiological questions that may help to guide efficient vaccination policies.

Since our new inferential approach is based on using Gaussian processes, in Chapter 3, I give a brief overview of the basics of Gaussian processes. In this chapter, I try to explain
the applicability of Gaussian processes in solving various challenging problems. Among these problems, I demonstrate how a Gaussian process can be used in describing the gypsy moth population dynamics in a new model that we have recently developed. The model, exploiting underlying dependencies in the proxy data for the gypsy moth, allows for very accurate inference of insect population periodicities as shown in our simulated examples.

Chapter 4 is based on a study of the parameters of the gravity model for measles dynamics in the pre-vaccination era. In this chapter, I show that traditional likelihood-based inference for the gravity model leads to biased parameter estimates and a poor model fit. I then demonstrate that our new approach provides a flexible and accurate inferential method for the gravity model and the resultant fitted model captures the biologically important features of the measles dynamics. In the context of measles in the pre-vaccination era, our method also allows us to study some interesting aspects of the dynamical system. For instance, we find that there does not appear to be a significant change in the gravity parameters for the school holiday periods versus non-holidays.

In Chapter 5, I discuss different properties of ABC and our emulator-based approaches to inference based on a number of toy examples. In this chapter, I also describe and propose a solution using our emulator-based approach to a challenging inferential problem for a mixed random graphs model. For this model, when data are large, the only existing approaches were based on approximation techniques like variational methods.

6.2 Caveats and Avenues for Future Research

In infectious disease dynamics modeling, since incomplete and noisy epidemiological measurements often lack the necessary information in order to accurately estimate the model parameters, the issue of parameter identifiability is a common problem. The problem was seen in our study of the gravity model for measles, in the model for meningitis dynamics and in the context of a mixture random graphs model for affiliation networks.

There are two types of identifiability issues. The first type of parameter identifiability issues is related to the inherent non-identifiability of the parameters of the model given data. For example, such identifiability issues were apparent from the flat or ridge-like posterior distributions of the gravity parameters in Chapter 4. Similarly, in the model for meningitis dynamics, posterior distributions of some parameters were flat and we could only find weak evidence in favor of environment-related increase in the number of cases of meningitis in
Nigeria. This type of parameter identifiability issues can sometimes be solved by modifying the model and/or obtaining additional relevant information to assist with parameter inference. In a simple SIR model with demography discussed in Section 1.2, for example, a good understanding of the host individual’s biology may provide accurate estimates of the birth and death rates. In our model for meningitis discussed in Chapter 2, the timing parameters of the model can be learned from data about different climate/environmental correlates of meningitis outbreaks (cf. Molesworth et al., 2003; Mueller and Gessner, 2010; Sultan et al., 2005; Thomson et al., 2006). In Chapter 5, there were apparent identifiability issues between the parameters of the mixture model as well. We found that some of the class probabilities of the graph could not be learned from data based on multiple simulated examples. Similar phenomenon for mixture random graphs was observed in Snijders and Nowicki (1997) and Nowicki and Snijders (2001). One way to resolve these issues in the context of an affiliation network would be to select appropriate summary statistics of the graph and consider only the posterior distributions of those functions of parameters that are invariant with respect to relabeling the classes. Another approach to reduce the identifiability here would be to construct strong priors to identify class labels based on external data.

The second type of parameter identifiability issues is related to the non-identifiability of the parameters introduced by the inferential approach. For example, our emulator-based approach allows for the fact that there might be a discrepancy between the true model and the data. Accounting for a model-data discrepancy, however, may introduce additional identifiability issues between the parameters of interest and the model discrepancy term (cf. Bayarri et al., 2007b; Liu et al., 2009b). In our approach, we propose to model the discrepancy term as a one dimensional non-negative variable which does not seem to increase the identifiability issues based on our simulation study. In general, however, when emulation requires a more complex model discrepancy term, the inferential results about the parameters should be considered with some caution.

When using Gaussian processes for constructing an emulator for a complex disease dynamics model, or when a latent Gaussian process-based model is used to describe the population dynamics of the gypsy moth, we make certain simplifying assumptions. For example, in our emulation-based approach, based on our domain experts, we assume that output of the model and observations are a smooth and stationary process that can be modeled via a Gaussian covariance function. In general, if the underlying process is not smooth and/or non-stationary, our emulator may fail to result in a good approximation to the model.
the process is stationary, but a smooth covariance function is not appropriate, one possible way to solve this problem would be to consider exponential or Matérn covariance functions. In our model for the gypsy moth population dynamics, we assume that space and time are separable and the underlying process has a spherical covariance function. We also divide the data into independent blocks. We verify that these assumptions do not have any adverse effects on our inference based on simulated data examples. In general, however, when applying some of these techniques to data in a different problem, it may be useful to develop and use more rigorous approaches to verify some of the assumptions made and ensure that the model strikes the right balance between complexity and computability.

Based on our approach, we are able to answer some important questions about measles dynamics in pre-vaccination era. Since similar questions can be asked about modern measles dynamics, it would be interesting to study the data for measles from 1966 to 2012. This, however, requires some changes to the gravity model, since, for example, a new model needs to incorporate the data for vaccination. The new model also needs to account for the fact that modern population growth rates are different from that of pre-vaccination era. A work on a new model and inference for measles data with vaccination is currently in progress.

### 6.3 Potential Avenues for Theoretical Research

In addition to the specific issues raised in the previous section, there are also some general avenues for future research into the theoretical underpinnings of the methods developed in this dissertation. I will now briefly discuss some ideas for future work.

Our approach for inference for infectious disease dynamics models focuses on only one or two important summary statistics that are of interest to scientists. It is also worth noting that our approach works well when the parameter dimensionality is small, but is generally infeasible for parameter dimensions greater than around five to eight depending on the model complexity. For problems with high dimensional unknown parameter spaces, the only existing approaches of parameter inference are based on approximation techniques like variational methods (cf. Ambrosetti and Rabinowitz, 1973; Jordan et al., 1999; Struwe, 2008). Even though the variational methods may offer a reasonable solution for complex and high dimensional models, we note that the estimates obtained via these methods can be hard to interpret. For variational methods, it is also not always straightforward to characterize parameter uncertainty and investigate possible identifiability issues between the parameters.
In theory, it might be possible to construct an emulator based on more than two summary statistics and/or when the dimensionality of the parameter space is large. In practice, however, this poses huge computational challenges. For example, if we have ten summary statistics and five model parameters, a lattice of four settings per parameter will result in \(10 \times 4^5 = 10240\) dimensional vectors. In order to improve the applicability of our approach to such problems, one way would be to develop methods to find computationally efficient design points for our emulator. For disease dynamics models, these methods should be focused on increasing the accuracy of the estimators, rather than improving the predictions from an approximate model.

In the literature about theoretical properties of Gaussian processes for spatial data, Stein (1999b) studies the effect of using so-called fixed domain asymptotics, in which one considers an increasing number of observations in a fixed and bounded observational domain. In this work, authors argue that fixed domain asymptotics approach offers a nice way to understand the relationship between the spatial process and the predictors. Building upon the ideas of fixed domain asymptotics, it would be interesting to learn about the properties of the parameter estimates based on our emulation-based approach. From multiple simulated examples, it seems that the “density” and “concentration” of the grid has an effect on the uncertainty until a certain level. After one reaches that level, this effect becomes insignificant.

More general and more difficult theoretical problem in the context of inference with an emulator would be to explore the properties of the resultant posterior distribution. For simplicity, let us assume that summary statistics used for constructing an emulator are sufficient and the true model is “close” to a Gaussian process. When constructing an emulator using a Gaussian process, say, on a uniform grid, we use the following parameters: \(p\) - number of grid points, \(r\) - number of repetitions for each grid point, and \(n\) - dimensionality of our data. Using our previous notation, let \(\theta\) be an unknown parameter and \(D\) be an \(n\) dimensional data whose true joint distribution is \(P_{\theta}^{(n)}\). Let \(\Pi\) and \(\Pi(\cdot|D_n)\) be prior and true posterior distributions for \(\theta\), respectively. If we let \(\Pi(\cdot|D_n, p, r)\) be an approximate posterior distribution for \(\theta\) obtained via an emulator, our interest will be on studying the asymptotic consistency of \(\Pi(\cdot|D_n, p, r)\) and its closeness to \(\Pi(\cdot|D_n)\). In other word, assuming that it is possible to define a basis for convergence (asymptotics), interesting question would be to explore whether the sequence of posteriors \(\Pi(\cdot|D_n, p, r)\) converges, in a suitable sense, to the degenerate measure at \(\theta\) (cf. Choi and Ramamoorthi, 2008).

There are methods that show posterior consistency of the true posterior distribution for
certain classes of problems (cf. Choi and Ramamoorthi, 2008; Choi and Schervish, 2007). These methods are based on using Schwartz’s theorem (Schwartz, 1965) for convergence of density functions. For these classes of problems where the true posterior functions are asymptotically consistent, one way to approach the problem of consistency of the emulator-based posterior distribution would be to investigate distances between the true and approximate posterior functions. This research may involve concepts like affinity, Hellinger and Kullback-Leibler divergences of density functions.
Appendix A

Likelihood Function for the Gravity Model

We provide below details of the likelihood function for the gravity model. For \( t >= 2 \),

\[
L(I_{k(t+1)}, S_{k(t+1)}, L_{kt}|I_{kt}, S_{kt}) = \\
= L(I_{k(t+1)}, S_{k(t+1)}|I_{kt}, S_{kt}, L_{kt})L(L_{kt}|I_{kt}, S_{kt}) = \\
= L(I_{k(t+1)}, S_{k(t+1)}|I_{kt}, S_{kt}, L_{kt})L_{kt}^{m_{kt}-1}\frac{\exp(-L_{kt})}{\Gamma(m_{kt})}
\]

where, as defined before in Equation 3,

\[
m_{kt} = \theta N_{kt}^{\tau_1} \sum_{j=1,j\neq k}^{K} \frac{I_{jt}^{\tau_2}}{d_{kj}}
\]

\[
L(I_{k(t+1)}, S_{k(t+1)}|I_{kt}, S_{kt}, L_{kt}) = \\
= L(S_{k(t+1)}|L_{k(t+1)}, I_{kt}, S_{kt}, L_{kt})L(L_{k(t+1)}|I_{kt}, S_{kt}, L_{kt}) \\
= L(S_{k(t+1)}|L_{k(t+1)}, S_{kt})L(L_{k(t+1)}|I_{kt}, S_{kt}, L_{kt}) \\
= L(S_{k(t+1)}|L_{k(t+1)}, S_{kt})\lambda_{k,t+1}^{I_{k(t+1)}}\frac{\exp(-\lambda_{k,t+1})}{(I_{k(t+1)})!}
\]
where $\lambda_{k,t+1} = \beta_t S_{kt}(I_{kt} + L_{kt})^\alpha$ as in Equation 1. Here,

$$L(S_{k(t+1)}|L_{k(t+1)}, I_{kt}) = \begin{cases} 
1, & \text{if } S_{k(t+1)} = S_{kt} + B_{kt} - I_{k(t+1)} \\
0, & \text{otherwise.} 
\end{cases}$$

Finally, the likelihood function of the model (conditional on the first observations) at the parameters $\theta, \tau_1, \tau_2, \rho$ can be written as

$$L(\{I_{kt}\}_{k \geq 1, t \geq 2}, \{S_{kt}\}_{k \geq 1, t \geq 2}|\{I_{k1}\}_{k \geq 1}, \{S_{k1}\}_{k \geq 1}) =$$

$$= \int \int \prod_{t=2}^{T} \prod_{k=1}^{K} L(I_{k(t+1)}, S_{k(t+1)}, L_{kt}|I_{kt}, S_{kt}) dS dL.$$
Appendix B

Details of Constructing an Emulator

We first review our notation. Let $Z = (z_1, \cdots, z_K)$ denote the vector of summary statistics calculated using the observations. In our examples, these summary statistics are the proportions of zeros for each community and $K$ is the total number of communities in our space-time data. We then let $\Theta = (\theta, \tau_1, \tau_2, \rho)$ denote the vector of gravity parameters, $Y(\Theta) = (Y_1(\Theta), \cdots, Y_K(\Theta))$ denote the vector of summary statistics obtained using a simulation from the gravity model with the parameters $\Theta$. We choose a fixed grid $\Omega = (\Theta_1, \cdots, \Theta_p)$ on the parameter space. We recall that we used a uniform grid in the four dimensional cube $[0, 2]^4$ with 20 values of each parameter. In our examples, $p$, size of the grid, was therefore equal to $20^4$. For each $i$, we can calculate $D_i = d(Z, Y(\Theta_i))$, where, as was introduced before, the function $d(\cdot, \cdot)$ returns the Euclidean distance between the vectors. In the first stage of our approach, we model $D = (D_1, \cdots, D_p)$ as a Gaussian process:

$$D|\Omega, \beta_G, \xi_G \sim N(X\beta_G, \Sigma(\xi_G)),$$

where $\beta_G$ and $\xi_G$ are the parameters of the Gaussian process. The design matrix $X$ has a form,

$$X = \begin{pmatrix}
1 & \Theta_1^T \\
1 & \Theta_2^T \\
\vdots & \vdots \\
1 & \Theta_p^T
\end{pmatrix} = \begin{pmatrix}
1 & \theta_1 & \tau_{11} & \tau_{21} & \rho_1 \\
1 & \theta_2 & \tau_{12} & \tau_{22} & \rho_2 \\
\vdots & \vdots & \vdots & \vdots & \vdots \\
1 & \theta_p & \tau_{1p} & \tau_{2p} & \rho_p
\end{pmatrix}.$$
Elements of the covariance matrix $\Sigma(\xi_G)$ are given by

$$(\Sigma(\xi_G))_{ij} = \text{cov}(D_i, D_j) =
\begin{cases}
\sigma_G^2 \exp(-\phi_G^2 \|\Theta_i - \Theta_j\|^2), & \text{if } i \neq j \\
\sigma_G^2 + \tau_G^2, & \text{otherwise}.
\end{cases}$$

Note, therefore, that $\beta_G = (\beta_{0G}, \beta_{1G}, \cdots, \beta_{4G})$ are the regression parameters and $\xi_G = (\sigma_G^2, \tau_G^2, \phi_G)$ specify the covariance function of the Gaussian process. Since $(D_1, \cdots, D_p)$ and $(\Theta_1, \cdots, \Theta_p)$ are known, we can use maximum likelihood inference to obtain estimates $(\hat{\xi}_G, \hat{\beta}_G)$.

For any new $D^*$ obtained at an unknown $\Theta^*$, following a standard “plug-in” approach,

$$(\begin{array}{c}
D^* \\
D \\
\end{array}) =
\begin{pmatrix}
D^* \\
D_1 \\
D_2 \\
\vdots \\
D_p
\end{pmatrix}
\sim N(X_e \hat{\beta}_G, \Sigma_e(\hat{\xi}_G)),
$$ (B.1)

where

$$X_e =
\begin{pmatrix}
1 & \Theta^*^T \\
X
\end{pmatrix}
= \begin{pmatrix}
1 & \theta^* & \tau_1^* & \tau_2^* & \rho^*
\end{pmatrix}^T
\begin{pmatrix}
X
\end{pmatrix},$$

and

$$\Sigma_e(\xi_G) = \begin{pmatrix}
\sigma_G^2 + \tau_G^2 & v^T \\
v & \Sigma(\xi_G)
\end{pmatrix},$$

with $v = (v_1, \cdots, v_p)$, $v_i = \text{cov}(D^*, D_i)$ for $i = 1, \cdots, p$.

From Equation B.1, using standard multivariate normal theory (cf. Anderson, 1984), conditional distribution of $(D^*|D, \Omega, \Theta^*)$ can easily be derived,

$$D^*|D, \Omega, \Theta^* \sim N(\bar{\mu}, \bar{\Sigma}),$$

where $\bar{\mu} = \hat{\beta}_{0G} + \theta^* \hat{\beta}_{1G} + \tau_1^* \hat{\beta}_{2G} + \tau_2^* \hat{\beta}_{4G} + \rho^* \hat{\beta}_{4G} + v^T \Sigma(\hat{\xi}_G)^{-1}(D - X \hat{\beta}_G)$ and $\bar{\Sigma} = \sigma_G^2 + \tau_G^2 - v^T \Sigma(\xi_G)^{-1} v$. We denote this predictive distribution by $\eta(D^*; \Theta^*)$ and use as an emulator for the gravity model.


Bayarri, M., Berger, J., Higdon, D., Kennedy, M., Kottas, A., Paulo, R., Sacks, J., Cafeo, J.,

Bayarri, M., Berger, J., Paulo, R., Sacks, J., Cafeo, J., Cavendish, J., Lin, C., and Tu, J.


Bernoulli, D. (1760). Essai dune nouvelle analyse de la mortalit´ e caus´ ee par la petite v´ erole


Data with Applications to Climate Science. Dissertation, Penn State University.

climate system characteristics from climate models and multiple tracers. Journal of the
American Statistical Association, in review.

spatial output. In Chen, M.-H., Dey, D. K., Müller, P., Sun, D., and Ye, K., editors,
Frontiers of Statistical Decision Making and Bayesian Analysis: In Honor of James O.


grey-sided voles, clethrionomys rufocanus, of hokkaido: spectral and principal components
analyses. Philosophical Transactions: Biological Sciences, pages 867–875.


Vita
Roman A. Jandarov

CONTACT INFORMATION
Cell: (814) 441-0868
E-mail: raj153@psu.edu

RESEARCH INTERESTS
Space-time processes; infectious disease dynamics; complex computer experiments; statistical computing; Bayesian inference; inference for network models.

EDUCATION
The Pennsylvania State University, University Park, Pennsylvania, USA
Department of Statistics
Ph.D. Student, Statistics, GPA = 4.0 (expected graduation date: Summer 2012)

Lomonosov Moscow State University, Moscow, Russia
Department of Mathematics and Mechanics
Specialist (cum laude, GPA = 4.0), Mathematics: Functional Analysis (Equivalent of M.S in USA), 2007

PUBLICATIONS AND PROCEEDINGS


CITIZENSHIP
Republic of Kazakhstan