

©Copyright 2018

Jonathan Fintzi

# Bayesian Modeling of Partially Observed Epidemic Count Data

Jonathan Fintzi

A dissertation  
submitted in partial fulfillment of the  
requirements for the degree of

Doctor of Philosophy

University of Washington

2018

Reading Committee:

Vladimir Minin, Chair

Jon Wakefield, Chair

M. Elizabeth Halloran

James Hughes

Program Authorized to Offer Degree:  
Biostatistics

University of Washington

**Abstract**

Bayesian Modeling of Partially Observed Epidemic Count Data

Jonathan Fintzi

Co-Chairs of the Supervisory Committee:

Co-chair Vladimir Minin

Co-chair Jon Wakefield

An incredible abstract with all the best words will appear here.

# TABLE OF CONTENTS

	Page
List of Figures . . . . .	iii
List of Tables . . . . .	iv
Glossary . . . . .	v
Chapter 1: Introduction and data setting . . . . .	1
1.1 Motivating examples . . . . .	1
1.1.1 Influenza in a British boarding school . . . . .	1
1.1.2 Ebola in West Africa . . . . .	1
1.1.3 Pandemic A(H1N1) influenza in Finland . . . . .	1
1.2 Organization of this dissertation . . . . .	1
Chapter 2: Background . . . . .	2
2.1 Models for the Spread of Infectious Disease . . . . .	2
2.1.1 Deterministic Representations . . . . .	2
2.1.2 Stochastic Representations . . . . .	2
2.1.3 Large-Population Approximations . . . . .	2
2.2 Computational Approaches to Fitting Stochastic Epidemic Models . . . . .	2
2.3 Bayesian Computation . . . . .	2
2.3.1 Markov Chain Monte Carlo . . . . .	2
2.3.2 Bayesian Data Augmentation . . . . .	2
Chapter 3: Agent-Based Data Augmentation for Fitting Stochastic Epidemic Models to Prevalence Data . . . . .	3
3.1 Overview . . . . .	3
3.2 The data augmentation algorithm for an SIR model . . . . .	3
3.3 Generalizing the algorithm to other models . . . . .	3

3.3.1	Data augmentation for SEIR dynamics . . . . .	3
3.3.2	Data augmentation for SIRS dynamics . . . . .	3
3.3.3	Data augmentation for arbitrary dynamics . . . . .	3
3.4	Simulation results . . . . .	3
3.5	Example: Influenza in a British boarding school . . . . .	3
3.6	Discussion . . . . .	3
Chapter 4:	Approximate Inference for Stochastic Epidemic Models of Outbreaks in Large Populations . . . . .	4
4.1	Overview . . . . .	4
4.2	Fitting Stochastic Epidemic Models via the Linear Noise Approximation . .	5
4.2.1	Measurement Process and Data . . . . .	5
4.2.2	Latent Epidemic Process . . . . .	6
Chapter 5:	Dynamic Transmission Modeling of Pandemic A(H1N1) Influenza in Finland . . . . .	7
Chapter 6:	Discussion and Future Work . . . . .	8
Bibliography	. . . . .	9
Appendix A:	Appendix A . . . . .	10

## LIST OF FIGURES

Figure Number

Page

## LIST OF TABLES

Table Number

Page

## GLOSSARY

AFSS: Automated factor slice sampler.

CTMC: Continuous-time Markov chain.

DA: Data augmentation.

ELIPTSS: Elliptical slice sampling.

ESS: Effective sample size.

ILI: Influenza-like illness.

LNA: Linear noise approximation.

MJP: Markov jump process.

SEM: Stochastic epidemic model.



## ACKNOWLEDGMENTS

Very grateful to many people.

## **DEDICATION**

Dedication to important people.

## Chapter 1

# INTRODUCTION AND DATA SETTING

### **1.1 Motivating examples**

*1.1.1 Influenza in a British boarding school*

*1.1.2 Ebola in West Africa*

*1.1.3 Pandemic A(H1N1) influenza in Finland*

### **1.2 Organization of this dissertation**

## Chapter 2

# BACKGROUND

### **2.1 Models for the Spread of Infectious Disease**

#### *2.1.1 Deterministic Representations*

#### *2.1.2 Stochastic Representations*

#### *Agent-based models*

#### *Population-level models*

#### *2.1.3 Large-Population Approximations*

#### *Diffusion approximations of Markov jump processes*

#### *Linear noise approximation*

### **2.2 Computational Approaches to Fitting Stochastic Epidemic Models**

### **2.3 Bayesian Computation**

#### *2.3.1 Markov Chain Monte Carlo*

#### *2.3.2 Bayesian Data Augmentation*

## Chapter 3

# AGENT-BASED DATA AUGMENTATION FOR FITTING STOCHASTIC EPIDEMIC MODELS TO PREVALENCE DATA

### **3.1 Overview**

### **3.2 *The data augmentation algorithm for an SIR model***

### **3.3 *Generalizing the algorithm to other models***

#### *3.3.1 Data augmentation for SEIR dynamics*

#### *3.3.2 Data augmentation for SIRS dynamics*

#### *3.3.3 Data augmentation for arbitrary dynamics*

### **3.4 Simulation results**

### **3.5 Example: Influenza in a British boarding school**

### **3.6 Discussion**

## Chapter 4

# APPROXIMATE INFERENCE FOR STOCHASTIC EPIDEMIC MODELS OF OUTBREAKS IN LARGE POPULATIONS

### 4.1 Overview

Surveillance and outbreak response systems often report incidence counts of new cases detected in each inter-observation time interval. Analyzing this type of time series data is challenging since we must overcome many of the same challenges that we face in modeling the transmission dynamics of infectious diseases in small population settings with prevalence data — discrete snapshots of a continuously evolving epidemic process, detecting a fraction of the new cases, and often directly observing only one aspect of the disease process. Furthermore, our task is made more difficult by the additional computational burden that results from repeated evaluation of CTMC likelihoods; the products of exponential waiting time distributions consist of polynomially increasing numbers of terms, and agent-based data augmentation MCMC algorithms become unwieldy as the numbers of subject-path proposals required to meaningfully perturb the CTMC likelihood get large [3].

In this chapter, we show how the LNA of Section 2.1.3 can be adapted to obtain approximate inference for SEMs fit to epidemic count data in large populations. Our contributions are threefold: First, we demonstrate how the SEM dynamics should be reparameterized so that the LNA can be used to approximate transition densities of the counting processes for disease state transition events. Second, we fold the LNA into a Bayesian data augmentation framework in which latent LNA paths are sampled using the elliptical slice sampling (EliptSS) algorithm of [6]. This provides us with general machinery for jointly updating the latent paths while absolving us of the *de facto* modeling choice that the data be Gaussian in order to efficiently perform inference as in [5, 2], or the need to use particle filter methods

for non-Gaussian emission distributions as in [4]. Finally, we introduce a non-centered parameterization for the latent LNA process that massively improves the efficiency of our DA MCMC framework and makes it tractable for fitting complex models.

## 4.2 Fitting Stochastic Epidemic Models via the Linear Noise Approximation

For clarity, we will present the algorithm for fitting SEMs via the LNA in the context of fitting the susceptible–infected–recovered (SIR) model to Poisson distributed incidence counts before proceeding to generalize the algorithm to more complex SEM dynamics and measurement processes. This simple SIR model is an abstraction of the transmission dynamics of an outbreak as a closed, homogeneously mixing population of exchangeable individuals who are either susceptible ( $S$ ), infected, and hence infectious, ( $I$ ), or recovered ( $R$ ). It is important to note that the model compartments refer to disease states as they relate to the transmission dynamics, not the disease process. Thus, an individual is considered to be recovered when she no longer has infectious contact with other individuals in the population, not when she clears disease carriage. As another example, in the susceptible–exposed–infected–recovered (SEIR) type models that we will consider later, the latent period in which an individual is exposed, but not yet infectious, should be understood as possibly varying in population with different contact dynamics, even when the incubation period of the pathogen should arguably be consistent across groups.

### 4.2.1 Measurement Process and Data

Incidence data,  $\mathbf{Y} = \{Y_1, \dots, Y_L\}$ , arise as increments of the cumulative numbers of new cases accumulated over a set of time intervals,  $\mathcal{I} = \{\mathcal{I}_1, \dots, \mathcal{I}_L : \mathcal{I}_\ell = (t_{\ell-1}, t_\ell]\}$ . In outbreak or surveillance settings, we do not typically believe that every case is detected since individuals may be asymptomatic or may escape detection. Let  $\mathbf{N}^c = (N_{SI}^c, N_{IR}^c)$  denote the counting process for the cumulative numbers of infections ( $S \rightarrow I$  transitions) and recoveries ( $I \rightarrow R$  transitions), and let  $\Delta \mathbf{N}^c(t_\ell) = \mathbf{N}^c(t_\ell) - \mathbf{N}^c(t_{\ell-1})$  denote the change in cumulative numbers of transitions over  $\mathcal{I}_\ell$ ; so,  $\Delta N_{SI}^c(t_\ell)$  is the incidence over  $(t_{\ell-1}, t_\ell]$ . We might choose to model

the number of observed cases as a Poisson sample of the true incidence with detection rate  $\rho$ . Thus,

$$Y_\ell | \Delta N_{SI}^c(t_\ell), \rho \sim \text{Pois}(\rho \Delta N_{SI}^c(t_\ell)). \quad (4.1)$$

There are two minor points that we wish to make note of before proceeding. First, we have allowed for the possibility that cases are over-reported. This is not a necessary assumption for any of the subsequent results and is also not unreasonable when studying outbreaks in large populations where the “fog of war” might lead to inflation of reported incidence or misclassification of individuals whose symptoms are similar to the disease of interest. This modeling choice is also not particularly problematic when the detection probability is low since the emission densities will have negligible mass above the true incidence. Second, we are also making this modeling choice with an eye on the compatibility of the measurement distribution with the eventual LNA approximation, which takes real, not integer, values. The Poisson distribution, along with the negative binomial distribution that we will use in subsequent sections, are well defined for non-integer values of the mean parameter.

#### 4.2.2 Latent Epidemic Process

The latent incidence process,  $\mathbf{N}$ , is driven by the



## Chapter 5

# **DYNAMIC TRANSMISSION MODELING OF PANDEMIC A(H1N1) INFLUENZA IN FINLAND**

## Chapter 6

# **DISCUSSION AND FUTURE WORK**

## BIBLIOGRAPHY

- [1] L.J.S. Allen. An introduction to stochastic epidemic models. In *Mathematical Epidemiology*, pages 81–130. Springer, New York, 2008.
- [2] P. Fearnhead, V. Giagos, and C. Sherlock. Inference for reaction networks using the linear noise approximation. *Biometrics*, 70:457–466, 2014.
- [3] J. Fintzi, X. Cui, J. Wakefield, and V.N. Minin. Efficient data augmentation for fitting stochastic epidemic models to prevalence data. *Journal of Computational and Graphical Statistics*, 26:918–929, 2017.
- [4] A. Golightly, D.A. Henderson, and C. Sherlock. Delayed acceptance particle MCMC for exact inference in stochastic kinetic models. *Statistics and Computing*, 25(5):1039–1055, 2015.
- [5] M. Komorowski, B. Finkenstädt, C.V. Harper, and D.A. Rand. Bayesian inference of biochemical kinetic parameters using the linear noise approximation. *BMC Bioinformatics*, 10:343, 2009.
- [6] I. Murray, R.P. Adams, and D.J.C. MacKay. Elliptical slice sampling. *JMLR: W&CP*, 9:541–548, 2010.
- [7] P. Neal and G.O. Roberts. A case study in non-centering for data augmentation: stochastic epidemics. *Statistics and Computing*, 15:315–327, 2005.
- [8] P.D. O’Neill. Introduction and snapshot review: relating infectious disease transmission models to data. *Statistics in Medicine*, 29:2069–2077, 2010.

Appendix A

**APPENDIX A**