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

LIST OF FIGURES

Figure Number Page

LIST OF TABLES

Table Number Page

GLOSSARY

AFSS: Automated factor slice sampler.

CLE: Chemical Langevin equation.

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.

SDE: Stochastic differential equation.

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

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 [7].

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 [16]. 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 [6, 15], or the need to use computationally

intensive particle filter methods for non–Gaussian emission distributions as in [11]. 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 P 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, $\mathbf{Y} = \{Y_1, \ldots, Y_L\}$, arise as increments of the numbers of new cases accumulated in a set of time intervals, $\mathcal{I} = \{\mathcal{I}_1, \ldots, \mathcal{I}_L : \mathcal{I}_\ell = (t_{\ell-1}, t_\ell]\}$. 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 $\mathbf{N}^c = (N^c_{SI}, N^c_{IR})$ denote the counting process for the cumulative numbers of infections ($S \to I$ transitions) and recoveries ($I \to R$ 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, $\Delta N^c_{SI}(t_\ell)$ 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 ρ. Thus,

$$Y_\ell | \Delta N^c_{SI}(t_\ell), \rho \sim \text{Pois}(\rho \Delta N^c_{SI}(t_\ell)). \tag{4.1}$$

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 neither a necessary assumption for any of the subsequent results, nor is it 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 SIR model is typically expressed in terms of compartment counts, $\mathbf{X}^c = \{S^c, I^c, R^c\}$, that evolve in continuous time on state space $\mathcal{S}^c_X = \{\mathcal{C}_{lmn} : l, m, n \in \{1, \ldots, P\},\ l + m + n = P\}$. We will make the (not particularly limiting) modeling choice to express the waiting times between disease state transitions as being exponentially distributed. Thus, $\mathbf{X}$ evolves according to a Markov jump process (MJP). If our data had consisted of prevalence counts, which arise as partial observations of infected individuals, we might have chosen to approximate transition densities of the MJP in the usual way that appears in [15, 6]. That is, we would write down the diffusion approximation for increments of the MJP for $\mathbf{X}$, which takes the form of a chemical Langevin equation (CLE), and Taylor expand the resulting stochastic differential equation (SDE) around its deterministic limit, discarding higher order terms, to

arrive at the LNA.

However, incidence data are discretely observed, partial realizations of the increments of counting processes that evolve continuously in time as individuals transition among disease states. The emission probabilities for incidence data, e.g., (4.1), depend on the change in N^c_{SI} over the time interval $(t_{\ell-1}, t_\ell]$, not on the change in I over the interval. It would be incorrect to treat incidence as simply the difference in prevalence since we could easily construct a scenario where there are positive numbers of infections, but where the prevalence does not change due to an equal number of recoveries. We need to construct the LNA that approximates transition densities of $\mathbf{N}$ if we are to write down correctly specified emission probabilities.

The cumulative incidence process for infections and recoveries, $\mathbf{N}^c$, is a Markov jump process state space $\mathcal{S}^c_N = \{\mathcal{C}_{jk} : j, k \in \{0, \ldots, P\}\}$. Let β denote the per–contact infection rate, and μ denote the rate at which each infected individual recovers. Suppressing the dependence on the rate parameters, the rate at which $\mathbf{N}^c$ transitions from state $\mathbf{n}$ to $\mathbf{n}'$ is

$$\boldsymbol{\lambda}_{\mathbf{n},\mathbf{n}'} = \begin{cases} \lambda_{SI} = \beta SI, & \mathbf{n} = (n_{SI}, n_{IR}),\ \mathbf{n}' = (n_{SI}+1, n_{IR}),\ \text{and } n_{SI}+1 \leq P, \\ \lambda_{IR} = \mu I, & \mathbf{n} = (n_{SI}, n_{IR}),\ \mathbf{n}' = (n_{SI}, n_{IR}+1),\ \text{and } n_{IR}+1 \leq P, \\ 0, & \text{for all other } \mathbf{n} \text{ and } \mathbf{n}'. \end{cases} \tag{4.2}$$

4.2.3 Diffusion Approximation — SIR Model

As outlined in Section 2.1.3, there are a variety of methods for arriving at a diffusion approximation for a Markov jump process, which under certain conditions yield equivalent results (for a comprehensive reference, see [8]). In the interest of clarity, we follow [6, 10, 11, 21] and appeal to an intuitive, though somewhat informal, construction of the CLE by matching its drift and diffusion with the approximate moments of increments of the MJP path in infinitesimal time intervals. For more detailed presentations see [8, 9, 20].

Suppose that, at the current time, the compartment counts are given by $\mathbf{X}^c(t) = \mathbf{x}^c_t$. We are interested in approximating the numbers of infections and recoveries in a small time interval, $(t, t + \mathrm{d}t]$, i.e., $\mathbf{N}^c(t + \mathrm{d}t) - \mathbf{N}(t)$. Now, suppose that we can choose $\mathrm{d}t$ such that

the following two *leap* conditions hold:

1. dt is sufficiently *small* that the $\mathbf{X}^c$ is essentially unchanged over $(t, t + \mathrm{d}t]$, so that the rates of infections and recoveries are approximately constant:

$$\boldsymbol{\lambda}(\mathbf{X}^c(t')) \approx \boldsymbol{\lambda}(\mathbf{x}^c(t)),\ \forall t' \in (t, t + \mathrm{d}t]. \tag{4.3}$$

2. dt is sufficiently *large* that we can expect many disease state transitions of each type:

$$\boldsymbol{\lambda}(\mathbf{x}^c(t)) \gg \mathbf{1}. \tag{4.4}$$

Condition (4.3), which can be trivially satisfied just by choosing dt to be small, implies that the numbers of infections and recoveries in $(t, t + \mathrm{d}t]$ are essentially independent of one another since the rates at which they occur are approximately constant within the interval [9]. This condition also carries the stronger implication that the numbers of infections and recoveries in the interval are independent Poisson random variables with rates $\boldsymbol{\lambda}(\mathbf{x}^c(t)\mathrm{d}t)$, i.e., $N^c_{SI}(\mathrm{d}t) \sim \text{Poisson}(\beta S(t)I(t)\mathrm{d}t)$ and $N^c_{IR}(t + \mathrm{d}t) \sim \text{Poisson}(\mu I(t)\mathrm{d}t)$. Condition (4.4), which we can reasonably expect to be satisfied in large populations where transmission dynamics are near their deterministic ODE limits [20], implies that the Poisson distributed increments can be well approximated by independent Gaussian random variables.

If (4.3) and (4.4) are satisfied, the change in the cumulative incidence of infections and recoveries can be approximated as

$$\mathbf{N}(t + \mathrm{d}t) - \mathbf{N}(t) \approx \boldsymbol{\lambda}(\mathbf{X}(t))\mathrm{d}t + \boldsymbol{\Lambda}(\mathbf{X}(t))^{1/2}\mathrm{d}t^{1/2}\mathbf{Z}, \tag{4.5}$$

where $\boldsymbol{\Lambda} = \text{diag}\,(\boldsymbol{\lambda}(\mathbf{X}))$ and $\mathbf{Z} \sim MVN(\mathbf{0}, \mathbf{I})$. This implies the equivalent CLE,

$$\mathrm{d}\mathbf{N}(t) = \boldsymbol{\lambda}(\mathbf{X}(t))\mathrm{d}t + \boldsymbol{\Lambda}(\mathbf{X}(t))^{1/2}\mathrm{d}\mathbf{W}_t, \tag{4.6}$$

where $\mathbf{W}_t$ are independent Brownian motion and $\boldsymbol{\Lambda}(\mathbf{X}(t))^{1/2}$ is the matrix square root.

Reparameterization the CLE in terms of comparment counts

We should reparameterize (4.6) so that we need not make reference to $\mathbf{X}$ explicitly, thus ensuring that the CLE is self–consistent with respect to $\mathbf{N}$. To this end, we borrow a reparameterization from [3, 12] for expressing $\mathbf{X}(t)$ in terms of $\mathbf{N}(t)$, conditional on the initial conditions $\mathbf{X}(t) = \mathbf{x}_0$ and $\mathbf{N}(t) = \mathbf{0}$. Let $\mathbf{A}$ denote the matrix whose rows specify the changes in the numbers of susceptible, infected, and recovered individuals corresponding to one infection or recovery event:

$$\mathbf{A} = \begin{array}{c} \\ S \to I \\ I \to R \end{array} \begin{array}{c} \begin{array}{ccc} S & I & R \end{array} \\ \begin{pmatrix} -1 & 1 & 0 \\ 0 & -1 & 1 \end{pmatrix} \end{array}. \tag{4.7}$$

Now, $\mathbf{X}$ is coupled to $\mathbf{N}$ via $\mathbf{X}(t) = \mathbf{x}_0 + \mathbf{A}^T\mathbf{N}(t)$. For the SIR model,

$$\begin{pmatrix} S(t) \\ I(t) \\ R(t) \end{pmatrix} = \begin{pmatrix} S_0 - N_{SI}(t) \\ I_0 + N_{SI}(t) - N_{IR}(t) \\ R_0 + N_{IR}(t) \end{pmatrix}, \tag{4.8}$$

which enables us to rewrite (4.6) as

$$\begin{aligned} \mathrm{d}\mathbf{N}(t) &= \boldsymbol{\lambda}(\mathbf{N}(t))\mathrm{d}t + \boldsymbol{\Lambda}(\mathbf{N}(t))^{1/2}\mathrm{d}\mathbf{W}_t \qquad (4.9) \\ &= \begin{pmatrix} \beta(S_0 - N_{SI}(t))(I_0 + N_{SI}(t) - N_{IR}(t)) \\ \mu(I_0 + N_{IR}(t))\mathrm{d}t \end{pmatrix} \mathrm{d}t + \\ &\qquad \begin{pmatrix} \beta(S_0 - N_{SI}(t))(I_0 + N_{SI}(t) - N_{IR}(t)) & 0 \\ 0 & \mu(I_0 + N_{IR}(t)) \end{pmatrix}^{1/2} \mathrm{d}\mathbf{W}_t. \end{aligned}$$

Log transforming the CLE

Note that changes in compartment volumes affect the rates, and hence increments in the incidence process, multiplicatively. Therefore, from a scientific perspective, we would like for perturbations about the drift in (4.9) to be symmetric on a multiplicative, not an additive

scale. Hence, we log transform (4.9). Let $\widetilde{\mathbf{N}} = \log(\mathbf{N} + \mathbf{1}) \implies \mathbf{N} = \exp(\widetilde{\mathbf{N}}) - \mathbf{1}$. By Itô's lemma [18], the corresponding SDE for $\widetilde{\mathbf{N}}$ is

$$\begin{aligned} \mathrm{d}\widetilde{\mathbf{N}} = \operatorname{diag}\left(\exp(-\widetilde{\mathbf{N}}(t) - 0.5\exp(-2\widetilde{\mathbf{N}})\right)\boldsymbol{\lambda}\left(\exp(\widetilde{\mathbf{N}}) - \mathbf{1}\right)\mathrm{d}t + \\ \operatorname{diag}\left(\exp(-\widetilde{\mathbf{N}})\right)\boldsymbol{\Lambda}\left(\exp(\widetilde{\mathbf{N}}) - 1\right)^{1/2}\mathrm{d}\mathbf{W}_t. \end{aligned} \tag{4.10}$$

4.2.4 Linear Noise Approximation — SIR Model

- Diffusion approximation and reparameterization - Resume with reparameterization.
- LNA
- Non–centered parameterization
- MCMC

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] H. Andersson and T. Britton. *Stochastic Epidemic Models and Their Statistical Analysis*. Lecture Notes in Statistics. Springer, New York, 2000.

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

[4] T. Britton. Basic stochastic transmission models and their inference. *ArXiv e-prints*, 2018.

[5] E. Buckingham-Jeffery, V. Isham, and T. House. Gaussian process approximations for fast inference from infectious disease data. *Mathematical biosciences*, 2018.

[6] P. Fearnhead, V. Giagos, and C. Sherlock. Inference for reaction networks using the linear noise approximation. *Biometrics*, 70:457–466, 2014.

[7] 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.

[8] C. Fuchs. *Inference for Diffusion Processes: With Applications in Life Sciences*. Springer Science & Business Media, New York, 2013.

[9] D.T. Gillespie. The chemical Langevin equation. *The Journal of Chemical Physics*, 113:297–306, 2000.

[10] A. Golightly and C.S. Gillespie. Simulation of stochastic kinetic models. In *In Silico Systems Biology*, pages 169–187. Springer, 2013.

[11] 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.

[12] L.S.T. Ho, F.W. Crawford, and M.A. Suchard. Direct likelihood-based inference for discretely observed stochastic compartmental models of infectious disease. *arXiv preprint arXiv:1608.06769*, 2016.

[13] M.J. Keeling and P. Rohani. *Modeling Infectious Diseases in Humans and Animals*. Princeton University Press, Princeton, 2008.

[14] A.A. King, M.D. de Celles, F.M.G. Magpantay, and P. Rohani. Avoidable errors in the modeling of outbreaks of emerging pathogens, with special reference to Ebola. *Proceedings of the Royal Society, Series B*, 282:20150347, 2015.

[15] 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.

[16] I. Murray, R.P. Adams, and D.J.C. MacKay. Elliptical slice sampling. *JMLR: W&CP*, 9:541–548, 2010.

[17] P. Neal and G.O. Roberts. A case study in non-centering for data augmentation: stochastic epidemics. *Statistics and Computing*, 15:315–327, 2005.

[18] B. Øksendal. *Stochastic Differential Equations*. Springer, New York, 2003.

[19] P.D. O'Neill. Introduction and snapshot review: relating infectious disease transmission models to data. *Statistics in Medicine*, 29:2069–2077, 2010.

[20] E.W.J. Wallace, D.T. Gillespie, K.R. Sanft, and L.R. Petzold. Linear noise approximation is valid over limited times for any chemical system that is sufficiently large. *IET systems biology*, 6:102–115, 2012.

[21] D.J. Wilkinson. *Stochastic Modelling for Systems Biology*. CRC Press, Boca Raton, 2011.

Appendix A

APPENDIX A