Fitting Stochastic Epidemic Models to Multiple Data Types

Mingwei Tang

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

Doctor of Philosophy

University of Washington

2019

Reading Committee:

Volodymyr Minin, Chair

Jonathan Wakefield

Zaid Harchaoui

Program Authorized to Offer Degree:
Statistics

University of Washington

Abstract

Fitting Stochastic Epidemic Models to Multiple Data Types

Mingwei Tang

Chair of the Supervisory Committee:
Professor Volodymyr Minin
Department of Statistics

Traditional infectious disease epidemiology focuses on fitting deterministic and stochastic epidemics models to surveillance case count data. Recently, researchers began to make use of infectious disease agent genetic data to complement statistical analyses of case count data. Such genetic analyses rely on the field of phylodynamics — a set of population genetics tools that aim at reconstructing demographic history of a population based on molecular sequences of individuals sampled from the population of interest. In this thesis, we aim at designing a general framework that can fit stochastic epidemic models to surveillance count data and to genetic data separately, or to use both sources of information at the same time. Firstly, we propose a Bayesian model that combines phylodynamic inference and stochastic epidemic models. We bypass the current computationally intensive particle Markov chain Monte Carlo (MCMC) methods and achieve computational tractability by using a linear noise approximation (LNA) — a technique that allows us to approximate probability densities of stochastic epidemic model trajectories. LNA opens the door for using modern MCMC tools to approximate the joint posterior distribution of the disease transmission parameters and of high dimensional vectors describing unobserved changes in the stochastic epidemic model compartment sizes (e.g., numbers of infectious and susceptible individuals). Next, we propose a joint model that allows us to integrate incidence data and genetic data. Finally, we consider the dependency of genetic sequence sampling times on the latent prevalence of the

infectious disease and propose a preferential sampling phylodynamics model that improves performance of phylodynamic inference. In a series of simulation studies, we show that all our proposed estimation methods can successfully recover parameters of stochastic epidemic models. Moreover, we demonstrate that combining multiple data types helps resolve identifiability issues and improves estimation precision. Throughout the dissertation, we use the incidence and genetic data from the 2014 Ebola epidemic in Sierra Leone and Liberia to illustrate our methodological developments.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

GLOSSARY

LNA: Linear Noise Approximation.

SIR: Susceptible-Infectious-Removed model.

SEIR: Susceptible-Exposed-Infectious-Removed model.

MCMC: Markov chain Monte Carlo

ODE: Ordinary differential equation

SDE: Stochastic differential equation

MJP: Markov jump process

MH: Metropolis-Hastings algorithm.

CTMC: Continuous time Markov chain.

WHO: World Health Organization

PREVALENCE: Number of infected individuals, i.e $I(t)$.

GMRF: Gaussian Markov random field.

NOTATION USED THROUGHOUT THE THESIS

- $\mathbf{g}$: Genealogy estimated from molecular sequences, treated as observed data.
- $S(t)$: Counts of susceptible individuals at time t.
- $I(t)$: Counts of infected/infectious individuals at time t.
- $l(t)$: Number of lineages present at time t in a phylogenetic tree.
- $\mathbf{A}$: Effect matrix in stochastic epidemic model
- N: Total population size.
- $\beta(t)$: Infection rate at time t.
- $\gamma(t)$: Recovery rate at time t. γ denotes a constant removal rate.
- $\boldsymbol{\theta}$: Rate parameters for stochastic epidemic model.
- $R_0(t)$: Basic reproduction number at time t.
- $\mathbf{X}(t)$: Population trajectory at t.
- $Y_{1:T}$: Observed incidence at time $t_1, \ldots, t_T$.
- ρ: Reporting probability for incidence case count.
- ϕ: Over-dispersion parameter in negative binomial link function.

- $\mathcal{S}$: Genetic sequence of sampling times for molecular samples.
- $\boldsymbol{\kappa}$: Coefficients for preferential sampling. κ_0: intercept; κ_1: slope, preferential sampling power.
- $\boldsymbol{\xi}$: Independent standard normal random variables for LNA reparameterization
- $\boldsymbol{\eta}$: Independent standard normal random variables for $R_0(t)$ GRMF reparameterization

ACKNOWLEDGMENTS

First of all, I want to give my deepest gratitude to my PhD advisor Vladimir Minin. I could never earn this degree without the inspiration, guidance and support from him. Vladimir is an awesome adviser with a deep insight in methodology and a broad knowledge of real applied problems. His high taste in research was my major aspiration for finding challenging research projects, from phylogenetic to infectious disease modeling. Outside academics, he is an interesting and warm-hearted mentor. I benefit a lot from those random chats we have in our weekly meetings. His advice helps me to seek a balance between work and life, and the tremendous encouragement he gave to me helped me survive through the uncertainties during the job hunting season. I am more than honored and fortunate to be able to work with Vladimir in the past few years and I am looking forward to continuing our collaboration in the future.

I would also like to thank my committee members and collaborators. I want to thank Professor Jon Wakefield for his suggestions and comments during my final exam. I owe a debt of gratitude to Professor Zaid Harchaoui, who lead me to the wonderful field of optimization and state space models. Moreover, I feel lucky to have Professor Trevor Bedford as the GSR in my committee, his advice helped me apply my methodology to real Ebola data. I want to thank my collaborator, Gytis Dudas without whose support I could never completing my first paper. Finally, many thanks to my academic advisor Elizabeth Thompson for her help navigating life in and outside of academia in my first and second year.

I appreciate all the support from members of the Minin Group. Thanks to my labmate Jon Fintzi for many fruitful discussions on linear noise approximation and MCMC algorithms. I also want to thank Michael Karcher and Julia Palacios for explaining their work in

nonparametric phylodynamic inference, which inspired me to dive into molecular epidemiology and to choose my thesis topic. Also many thanks to Amrit Dhar, Jim Faulkner and Andy Magee for their feedback and advice on my research projects and practice exams.

As a student in Department of Statistics at University of Washington, I couldn't be more proud to be surrounded by many brilliant statisticians and warm-hearted staff members.I owe many thanks to Ellen Reynolds, who help me dealing with all the logistic issues such as TA arrangement and OPT application. I was honored to work with many faculty members as their teaching assistant. I want to thank Caren Marzban, Don Percival, Peter Hoff, Adrian Raftery and Galen Shorack for their patience and help. Also, I'm lucky to have made many friends in this department. I will definitely miss the time I spent with Qiyang Han having hotpots, drinking beers and chatting about research, careers, and gossip. I want to thank Bowen Wang for introducing me to the world of hiking and taking me to many wondering trails in his Honda Accord. I'm grateful to be accompanied by many classmates and friends in the past six years, including Zehang Li, Haoran Cai, Peiran Liu, Ning Li, Yali Wan, Xiang Cui, Anqi Cheng, Yunqi Bu, Yushi Tan, Pengjie Pan and Yicheng Li. My TA experience enabled me to get acquaintance with many outstanding students: Haoxun Luo, Jiayuan Shi, Yachen Jiang, Ge Fang, Tongfang Sun, Xiaoxiao Li, Xinrui Cao and Jiaxi Wang, who kept pushing me to improve my instructing and communication skills.

I'm very lucky to make a lot of friends outside UW in the past six years. I owe a debt of gratitude to my college friend Di Zhang, without whose help I'll never be able to experience a new life style in southern California. Besides, I'm grateful to make friends with UCI stats folks: Yannan Tang, Tong Shen, Lechuan Hu and Fan Yin. I want to thank them for the inviting me to their office and offering great assistance in both academic and life. I also want to express my gratefulness to Xuehan Zhou and other Ph.Ds from UCI education department including Huafeng Zhang, Linyan Ruan, Qiujie Li and Sirui Wan for feeding me with dumplings, hotpots and so many joys. Last but not least, my thank goes to Xingxu

Yan, Chengcen Sha, Xuexi Zhang and Megan Jiang for various happy hour events.

Alongside with classmates and colleagues in UW, there are many friends that I owe a lot thanks to. I'm grateful to have my friend Yitian Shao, with whom I shared many unforgettable road trips during these six years, from New York to San Francisco, Los Angeles to Phoenix. I'm also want to thank Bingqing Peng for those interesting chats about life and gossips, providing me with the much-needed distractions and happines that make me survive through the journey of Ph.D. In addition, I really appreciate the patience and encouragements from Xuejiao Guan , which helps me overcome the difficulties and anxieties in my Ph.D life. Furthermore, I feel lucky to have the emotion support from bunch of high school friends, they are Sitegeqi, Yuechen Zhang, Zhongxiu Wang, Wei Bai and Tingyu Gong. I'm glad our friendship stands as strong as 10 years ago even we are living in different cities now.

Finally I want to dedicate this thesis to my family, who always stand behind me and encourage me to pursue the dreams in my life. With their endless love and support, I am able to reach to the end and proudly present my PhD work.

DEDICATION

To my parents, Jinyue Tang and Mei Zhao

To my grandmother, Guilian Zuo

Chapter 1

INTRODUCTION

Infectious disease modeling studies the transmission of the disease on the subject level and the disease spread on the population level. On the subject level, researchers are interested in the transmission rate, length of the infection period, and another attributes of the disease natural history. On the population level, the host population dynamics reveals the trend of the disease spread, providing important information for disease control and prevention efforts. Though neither of the disease transmission rates nor population dynamics are observed directly, infectious disease modelers take many sources of observed information, construct mathematical epidemic models, and obtain estimates of the model parameters and latent variables of interest. In this thesis, we will consider two types of data used for infectious disease modeling: surveillance case count data and infectious disease agent genetic sequence data. The former source of information is directly related to the infected population and is widely used in traditional infectious disease epidemiolgy. The use of genetic data falls into the territory of *phylodynamics* — a relatively new field that combines tools from infectious disease epidemiology, population genetics, and phylogenetics.

1.1 Surveillance Data and Stochastic Epidemic Models

Surveillance data are usually cases counts related to disease incidence or prevalence reported over a sequence of times during the course of the epidemic. Such data provide disease population level information about the dynamics of the outbreak. The challenges for modeling the surveillance case count data are that the data are often observed at discrete times and complete subject level data consisting of the exact times at which individuals transition through disease states are rarely available. Moreover, surveillance case count data often

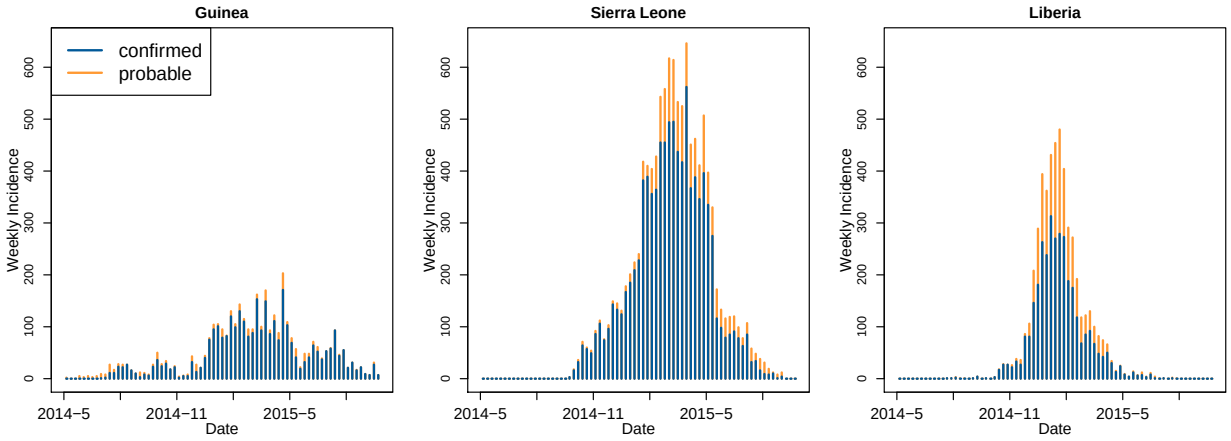

Figure 1.1: Weekly incidence report for the 2014-2015 West Africa Ebola outbreak from WHO. Left: Guinea, middle: Sierra Leone, right: Liberia. The confirmed cases are plotted in blue while the probable cases are drawn in orange.

suffer from the problem of under-reporting. A probability model is needed to establish the link between the true population dynamics and the observed data, making this problem as a complicated missing data/ latent variable state space model. Below we provide a concrete example illustrating these challenges.

Example: Ebola Weekly Incidence Data Figure 1.1 demonstrates an example of reported incidence data for the 2014-2015 Ebola outbreak in West Africa by World Health Organization [b]. During years 2014-01-05, the incidence was reported weekly in Guinea, Sierra Leone, and Liberia, with two categories: confirmed and probable cases. The term confirmed indicates an Ebola case with lab confirmation, while probable cases are those being suspected to have Ebola but without lab-confirmed diagnosis.

The population dynamics of infectious disease are often characterized using compartmental models, which divide the population into different subgroups/states and assume interactions and transitions between groups. For example, a Susceptible-Infectious-Removed (SIR) model divides the population into three groups: susceptible, infectious, and removed,

characterized by transitions: (1) from susceptible to infected and (2) infectious to removed.

Current methodologies for compartmental epidemic models fall into three categories: (1) deterministic ordinary differential equation (ODE) approach (2) Markov Jump process (MJP) and (3) stochastic differential equation (SDE) approach. The first approach describes a deterministic dynamic of the population evolution over times [Kermack and McKendrick, 1927]. However, ignoring the stochastic nature of the disease can lead to overconfident estimation, especially for disease dynamics with small initial epidemic size. The MJP models the dynamics as a continuous time Markov chain over discrete state space of counts in each compartment. MJP inference mainly focuses on epidemic with small population size [Ho et al., 2016, Fintzi et al., 2017], either by augmenting the latent state space or calculating/ approximating the transition likelihood. However, the transition probability becomes intractable for large population dynamics, making it difficult to perform likelihood based inference. The SDE method is another way to develop a stochastic epidemic model for large population dynamics. Since the transition likelihood for small time $[t, t + \Delta t]$ can be approximated using a Gaussian distribution, Golightly and Wilkinson [2005, 2008], for example, perform standard data augmentation approaches for SDE method. However, such method is not computationally efficient due to the high dimensional latent space for data augmentation. [Golightly and Wilkinson, 2011, King et al., 2010, Rasmussen et al., 2011, Koepke et al., 2016] use particle Markov chain Monte Carlo (MCMC) [Andrieu et al., 2010] to tackle this problem. However, article MCMCl is computationally costly and suffers from convergence problem for high dimensional parameter space, motivating development of new methods capable of fitting stochastic epidemic models to surveillance count data.

1.2 Genetic Data and Phylodynamics

Phylodynamics is an area at the intersection of phylogenetics and population genetics that studies how epidemiological, immunological, and evolutionary processes affect viral phylogenies constructed based on molecular sequences sampled from the population of interest [Grenfell et al., 2004, Volz et al., 2013]. Phylodynamics is especially useful in infectious dis-

ease modeling because genetic data provide a source of information that is complimentary to the traditional disease case count data.

Example: Ebola Genealogy in West Africa Figure 1.2 depicts an reconstructed genealogy using the 1610 aligned Ebola virus whole genomes, collected from Spring 2014 to fall 2015 in Guinea, Sierra Leone, and Liberia. The dataset represents over 5% of known cases of Ebola detected during that outbreak.

Currently, learning about population-level infectious disease dynamics from molecular sequences can be accomplished using three general strategies. The first strategy relies on the coalescent theory — a set of population genetics tools that specify probability models for genealogies relating individuals randomly sampled from the population of interest [Kingman, 1982, Griffiths and Tavare, 1994, Donnelly and Tavare, 1995]. Using a subset of these models [Griffiths and Tavaré, 1994], it is possible to estimate changes in effective population size — the number of breeding individuals in an idealized population that evolves according to a Wright-Fisher model [Wright, 1931]. Such reconstruction can be done assuming parametric [Kuhner et al., 1998, Drummond et al., 2002] or non-parametric [Drummond et al., 2002, 2005, Minin et al., 2008, Palacios and Minin, 2013, Gill et al., 2013] functional forms of the effective population size trajectory. In the context of infectious disease phylodynamics, non-parametric inference is the norm and the estimated effective population size is often interpreted as the effective number of infections or the effective number of infectious individuals. However, reconstructed effective population size trajectories are not easy to interpret and estimation of parameters of disease dynamics is difficult to accomplish if one wishes to maintain statistical rigor [Pybus et al., 2001, Frost and Volz, 2010].

Another way to learn about infectious disease dynamics from molecular sequences is to model explicitly events that occur during the infectious disease spread and to link these events to the genealogy/phylogeny of sampled individuals using birth-death processes. For example, a Susceptible-Infectious-Removed (SIR) model includes two possible events: infections and removals (e.g., recoveries and deaths), represented by births and deaths in the corresponding

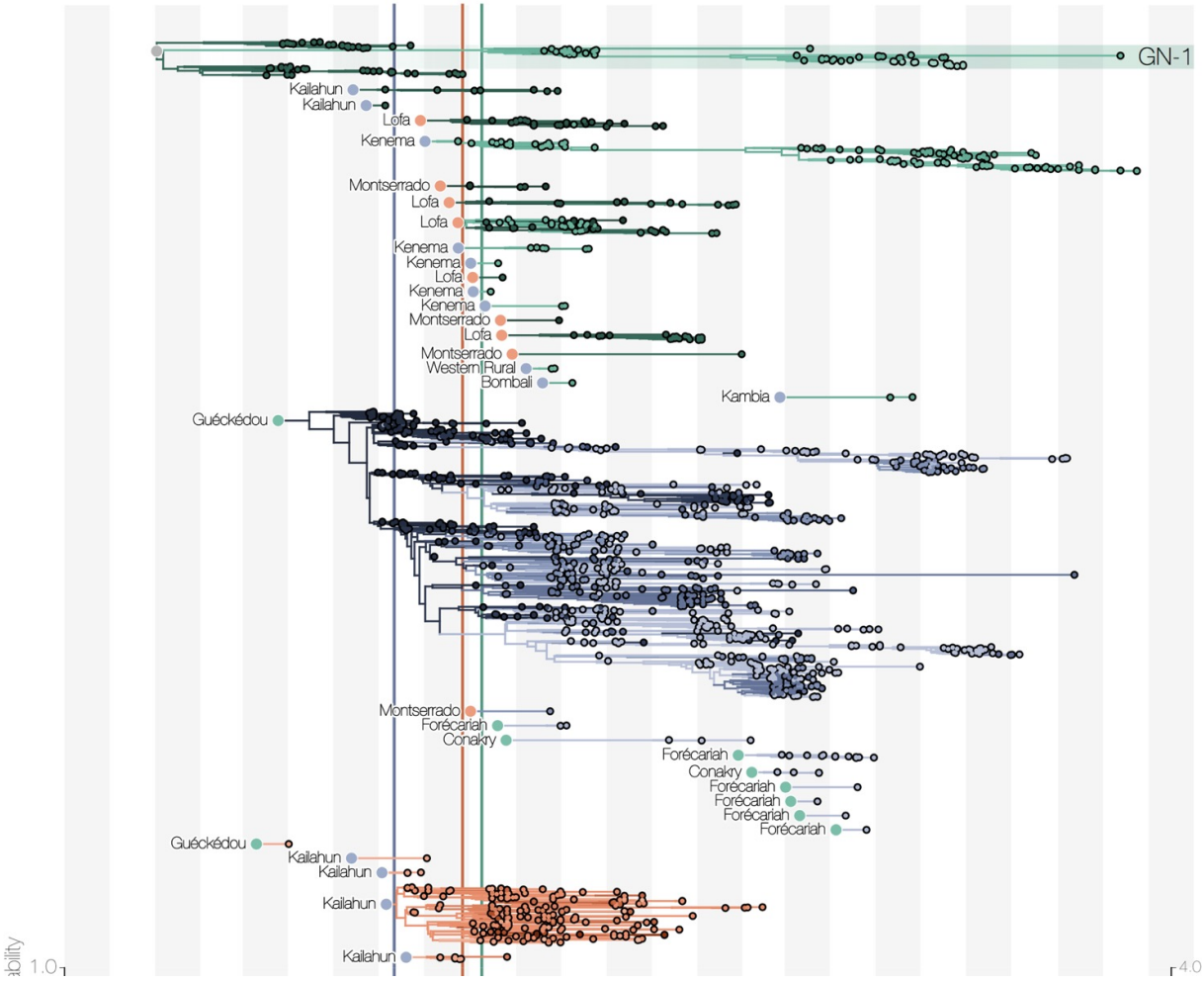

Figure 1.2: Estimated Ebola genealogies by Gytis Dudas (`https://github.com/ebov/space-time`). EBOV lineages by country (Guinea, green; Sierra Leone, blue; Liberia, red), tracked until the sampling date of their last known descendants.

birth-death model [Stadler et al., 2013, Kühnert et al., 2014]. Other SIR-like models (e.g., SI and SIS models) differ by the number and types of the events that are needed to accurately describe natural history of the infectious disease [Leventhal et al., 2013]. Although these methods are more principled than post-hoc processing of nonparametrically estimated disease dynamics, they are not easy to scale to large datasets and/or high dimensional models. For example, in order to fit phylodynamic birth-death models to genomics and epidemiological data Vaughan et al. [2018] use particle filter MCMC. However, computational burden of particle filter MCMC methods is usually very high. Moreover, these methods often struggle with convergence when the dimensional of statistical model parameters is even moderately high [Andrieu et al., 2010].

Structured coalescent models provide the third strategy of inferring parameters governing spread of an infectious disease [Volz et al., 2009, Volz, 2012, Dearlove and Wilson, 2013]. These models assume infectious disease agent genetic data have been obtained from a random sample of infected individuals, allowing for serial sampling over time. Although similar to the birth-death modeling framework, the structured coalescent models have two advantages. First, one does not have to keep track, analytically or computationally, of extinct and not sampled genetic lineages. Second, the density of the genealogy can be obtained given the population level information about status of individuals: for example, in the SIR model it is sufficient to know the numbers of susceptible, $(S(t))$, infectious, $(I(t))$, and recovered, $(R(t))$, individuals at each time point t. The second advantage comes with two caveats: 1) such densities can be obtained only approximately and 2) evaluating densities of genealogies is not straightforward and involves numerical solutions of differential equations. Even in cases when these caveats are manageable, the density of the assumed stochastic epidemic model population trajectory remains computationally intractable. One way around this intractability assumes a deterministic model of infectious disease dynamics [Volz et al., 2009, Volz, 2012, Volz and Pond, 2014], which potentially leads to overconfidence in estimation of model parameters. Particle filter MCMC offers another solution [Rasmussen et al., 2011, 2014], but, as we discussed already, these methods are difficult to use in practice, especially

in high dimensional parameter spaces.

1.3 Thesis Contributions

With the development of high throughput sequencing technologies, during many of infectious disease outbreaks both genetic sequence data and surveillance case count data are available to researchers. In spite of great methodological progress in analyzing surveillance case count data and in molecular epidemiology/phylodynamics, progress in taking advantage of integrating both sources of information to improve the estimation results has been much slower. Rasmussen et al. [2011] proposed a Bayesian framework for jointly modeling incidence data and genealogy data and demonstrated an improvement in estimation precision using simulated data. Inspired by this work, this thesis aims at constructing a flexible Bayesian stochastic epidemic model framework that can fit stochastic epidemic models to either genealogy or incidence data, or to use both data types simultaneously.

As discussed in Subsections 1.1 and 1.2, most of the state-of-the-art statistical methods for surveillance case count data and infectious disease agent genetic data rely on computational expensive PMCMC methods. In this thesis, we decided to bypass computationally unwieldy PMCMC by using a linear noise approximation (LNA) in order to fit stochastic models, such as a susceptible-infectious-removed (SIR) model, to data. The LNA can be viewed as a low order correction of the deterministic ordinary differential equation describing the asymptotic mean trajectory of a Markov jump process describing a kinetic model or chemical reaction [Van Kampen and Reinhardt, 1983]. Recent work shows that the LNA can also be considered as an approximation of the stochastic differential equation process of a kinetic model based on first order Taylor expansion, which makes it easier to understand [Wallace, 2010, Giagos, 2010]. A key feature of the LNA method is that it approximates the transition density of a stochastic kinetic model with a closed-form Gaussian distribution [Komorowski et al., 2009]. The LNA method has been recently applied to stochastic epidemic modeling in order to fit these models to data consisting of disease case counts time series [Fearnhead et al., 2014, Fintzi, 2019]. We make use of the non-restarting linear noise approximation [Komorowski

et al., 2009] to represent the latent disease trajectory as a locally Gaussian process. We provide an efficient Markov chain Monte Carlo (MCMC) algorithm by combining Metropolis-Hastings (MH) and elliptical slice sampler [Murray et al., 2010] kernels to obtain samples from the posterior distribution of rate parameters in the epidemic model, as well as the latent trajectories of different population compartments.

Here is the structure of the thesis. In Chapter 2, we give a brief review of mathematical and statistical tools used in the thesis, including stochastic epidemic models, coalescent likelihood formulation, Bayesian inference and computation. In Chapter 3, a Bayesian semi-parametric framework is developed to fit a stochastic SIR model to a phylogeny/genealogy estimated from molecular sequences. Next, we propose a method for integrating both genealogical and incidence data and demonstrate that this data integration provides more precise estimation of the infectious disease dynamic model parameters and latent variables. In Chapter 5, we show that ignoring the dependency of the distribution of molecular sequence sampling times and infectious disease prevalence can lead to systematic bias. Furthermore, we develop a preferential sampling model to reduce this bias and improve precision in of phylodynamic inference. Finally, we summarize our contributions and point out several prospective future research directions in Chapter 6.

Chapter 2

BACKGROUND

2.1 Motivation

The objective of infectious disease modeling-based data analysis is to recover the dynamics of infectious disease spread and to estimate parameters that governs the these dynamics. However, such reconstructions are challenging because the infectious disease spread is sparsely and indirectly observed. For example, when we observe a sequence of cases counts reported at discrete observation times, this offers us only noisy data on the number of infected individuals in the population. Even less directly, molecular sequences of an infectious disease agents with high mutation rates allow us to make an observation that changes in infectious disease agent genetic diversity are driven by the changes in the number of infected hosts. Let $\mathbf{Y}$ denote the observed data, $\mathbf{X}_{1:T}$ denote the latent population trajectory (the number of individuals in infectious, susceptible) at time $t_0, \ldots, t_T$ and $\boldsymbol{\theta}$ be the parameters of interest. Under Bayes rule, the posterior distribution of $\mathbf{X}_{0:T}$ and $\boldsymbol{\theta}$ given observation $\mathbf{Y}$ is proportional to the product of three following likelihood functions:

$$\Pr(\mathbf{X}_{0:T}, \boldsymbol{\theta} \mid \mathbf{Y}) \propto \Pr(\mathbf{Y} \mid \mathbf{X}_{0:T}, \boldsymbol{\theta}) \cdot \Pr(\mathbf{X}_{0:T} \mid \boldsymbol{\theta}) \cdot \pi(\boldsymbol{\theta}), \tag{2.1}$$

where the first and second terms on the right hand side of Equation (2.1) denote the observation (emission density) likelihood, population trajectory density respectively. The third component is a prior density of $\boldsymbol{\theta}$.

Equation (2.1) serves as the foundation for Bayesian infectious disease modeling in this thesis. In this chapter, we explain how to define each component of the product in Equaiton (2.1). In Section 2.2, we give an introduction to stochastic epidemic models and explain how to (approximately) calculate the population trajectory density $\Pr(\mathbf{X}_{0:T} \mid \theta)$. Section 2.3 will

focus on the observed likelihood $\Pr(\mathbf{Y} \mid \mathbf{X}_{0:T}, \boldsymbol{\theta})$ when $\mathbf{Y}$ represents the phylogeny/genealogy of sampled infectious disease sequences. Section 2.4 reviews Bayesian inference and MCMC algorithms that will be helpful in the subsequent Chapters.

2.2 Mathematical Models ofs the Spread of Infectious Disease

2.2.1 Infectious Disease Representation

Motivating Example: SIR Model

In epidemiology, the dynamics of infectious disease are often characterized by the nonlinear and rapid changes in sizes of different population groups/compartments. One of most widely used models is the Susceptible-Infectious-Removed (SIR) model [Bailey, 1975, Anderson and May, 1992], which is a closed compartment model that describes population dynamics over time. Under SIR model, the population consists of three groups/compartments categories: susceptible, infected/infectious and removed (recovered). The SIR model can be viewed as a stochastic mechanistic model or chemical reaction model characterized by following two reactions:

$$\text{Susceptible} + \text{Infected} \xrightarrow{\beta} 2 \cdot \text{Infected}, \tag{2.2}$$

$$\text{Infected} \xrightarrow{\gamma} \text{Recovered}, \tag{2.3}$$

where β denotes the per capita rate of the disease transmission to susceptible hosts (infection rate) and γ denote the removal rate for an infected individual. "Reaction" (2.2) is the process during which one infected individual infects a susceptible individual, ending up with two infected individuals, while "reaction" (2.3) is the process of recovery of an infected individual. Since the recovered individuals will never react with susceptible and infected individuals, for simplicity, we only model the susceptible and infected populations. Given

reactions (2.2) (2.3), we can define the effect matrix $\mathbf{A}$ for SIR model as

$$\mathbf{A} = \begin{array}{c} \begin{array}{cc} \text{susceptible} & \text{infected} \end{array} \\ \begin{pmatrix} -1 & 1 \\ 0 & -1 \end{pmatrix} \end{array} \begin{array}{l} \text{reaction (2.2)} \\ \text{reaction (2.3).} \end{array} \tag{2.4}$$

In matrix (2.4), each row represents a reaction and each column represents a change of susceptible individuals and recovered individuals in each reaction. The first row $(-1, 1)$ means that in (2.2) the number of susceptibles is decreased by 1 and the number of infectious individuals is increased by 1. Let $\mathbf{X}(t) = (S(t), I(t))^T$ denote the population state vector and $\boldsymbol{\theta} = (\beta, \gamma)^T$ denote the parameter vector. The rate vector $\mathbf{h}$ and rate matrix $\mathbf{H}$ for the two reactions are defined as

$$\mathbf{h}(\mathbf{X}(t), \boldsymbol{\theta}) = \begin{pmatrix} \beta S(t) I(t) \\ \gamma I(t) \end{pmatrix}, \text{ and } \mathbf{H} = \text{diag}\left(\mathbf{h}(\mathbf{X}(t), \boldsymbol{\theta})\right) = \begin{pmatrix} \beta S(t) I(t) & 0 \\ 0 & \gamma I(t) \end{pmatrix}. \tag{2.5}$$

.

Figure 2.1: SIR model under Markov process: Under the current state with S, I, R in each subgroup, the population transitions to state $S-1, I+1, R$ with rate βSI, the population transitions to state $S, I-1, R-1$ with rate γI.

2.2.2 A general reaction network framework for infectious disease spread

The above notation using the effect matrix $\mathbf{A}$, rate vector $\mathbf{h}$ and rate matrix $\mathbf{H}$ can be used not only in SIR model, but also generalized to other stochastic mechanistic models or chemical reaction models such as Lotka-Volterra model and auto-regulatory gene network model [Wilkinson, 2011]. In this section, we give a general representation of the reaction models for infectious disease dynamics. The notation is based on the work of Fearnhead

et al. [2014]. Let's start with a reaction system with d reactants $\mathcal{X}_1, \ldots, \mathcal{X}_d$ and q reactions. Without loss of generality, each reaction is assumed to have a constant rate parameter θ_i for $i = 1, \ldots, q$ and $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_q)$ denotes the rate vector of the system (this framework can be extended to handle stochastic kinetic models with time-varying rates as in Section 3.2). The transition event in the ith reaction ($i = 1, \ldots, q$) has the following form:

$$\tilde{a}_{i1}\mathcal{X}_1 + \cdots + \tilde{a}_{id}\mathcal{X}_d \xrightarrow{\theta_i} \bar{a}_{i1}\mathcal{X}_1 + \ldots + \bar{a}_{i1}\mathcal{X}_d, \tag{2.6}$$

where $\tilde{a}_{ij}$ and $\bar{a}_{ij}$ are non-negative integers representing the number of $\mathcal{X}_j$ in the ith reaction equation. For simplicity, we use a pre-reaction coefficient matrix $\tilde{\mathbf{A}} := \{\tilde{a}_{ij}\} \in \mathbf{R}^{q\times d}$ to represents the number of reactant j is the left hand side (LHS) of the reaction equations. Similar, post-reaction coefficient matrix $\bar{\mathbf{A}} := \{\bar{a}_{ij}\} \in \mathbf{R}^{q\times d}$ is used to represent the number of reactants on the right hand side (RHS) the reaction. The transitions in the reaction system can be encoded in an effect matrix,

$$\mathbf{A} := \{\tilde{a}_{ij} - \bar{a}_{ij}\} \in \mathbb{Z}^{q\times d}, \tag{2.7}$$

with each row corresponding to a certain type of reaction event and each column representing the change in the counts of reactants.

Reaction rates

Let $X_j(t)$ denote the number of of the $\mathcal{X}_j$ reactant at t, and the population state at time t can be tracked by vector $\mathbf{X}(t) := (X_1(t), \ldots, X_d(t))$. Let h_i denote the reaction rate of the ith reaction, where h_i can be written as

$$h_i(\mathbf{X}, \theta_i) = \theta_i \prod_{j=1}^{d} \binom{X_j}{\tilde{a}_{ij}}. \tag{2.8}$$

Hence, following the same notation as in Section 2.2.1, the rate vector $\mathbf{h}$ and the rate matrix $\mathbf{H}$ can be defined as

$$\mathbf{h}(\mathbf{X}, \boldsymbol{\theta}) = (h_1, \ldots, h_q)^T, \qquad \mathbf{H}(\mathbf{X}, \boldsymbol{\theta}) = \operatorname{diag}(\mathbf{h}(\mathbf{X}, \boldsymbol{\theta})). \tag{2.9}$$

the number of reactants on the LHS of i-th reaction equation $\sum_j^d \tilde{a}_{ij}$, as characterizes the order of the reaction. Higher order reactions that involve more than two reactants on the LHS of reaction equation are less likely to happen in infectious disease modeling. Here we give an introduction of low order reactions and relate them to infectious disease models.

Zero order reactions Zero order reaction defines a special family of reactions with an empty set on the LHS of the reaction equation, which can be written as

$$\emptyset \xrightarrow{\theta} \mathcal{X}_i. \tag{2.10}$$

Zero order reaction describes the introduction or birth of new reactant in the system. In infectious disease modeling, the birth process or the migration process from outside of the system can be represented using a zero order reaction with rate birth/migration rate θ.

First order reactions The first order reaction is a family of reactions that only have one reactant with coefficient 1 on the LHS of the reaction equation. First order reactions can be formulated as

$$\mathcal{X}_i \xrightarrow{\theta} \mathcal{X}_j \text{ or } \emptyset. \tag{2.11}$$

First order reactions has reaction rate $h(\mathbf{X}, \theta) = \theta X_j$. They can be used to describe the development of disease from one stage to another, without interacting with other individuals. One example of the first order reaction is the removal/recovery event in SIR model, where an infectious individual gets removed (e.g., recovered) with a certain removal rate. Another example is an event during which an individual moves from a latent stage of the infection to the infectious state.

Second order reactions One kind of a second order reaction has two reactants of different types on the LHS of the reaction equation. The first describes the reaction between two single reactant:

$$\mathcal{X}_{i_1} + \mathcal{X}_{i_2} \xrightarrow{\theta} \mathcal{X}_j \text{ or } \phi \tag{2.12}$$

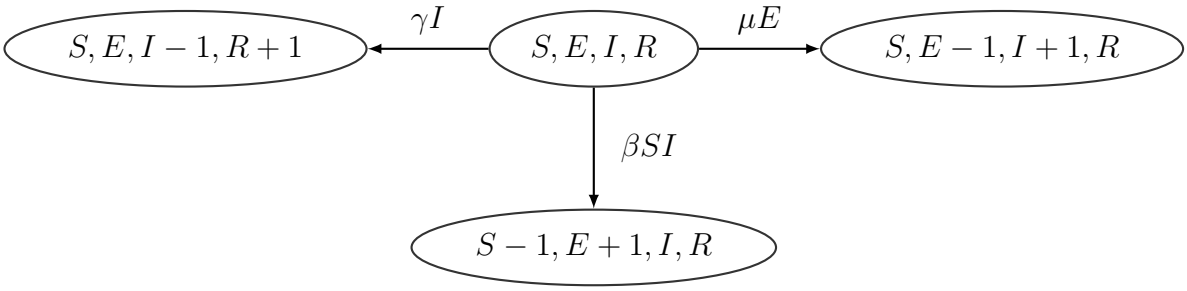

Figure 2.2: SEIR model under Markov jump process. From the current state with the counts S, E, I, R, the population can transition to (1) state $S-1, E+1, I, R$ (an infection event) with rate βSI or to (2) state $S, E-1, I+1, R$ (an event where infected individual becomes infectious) with rate μE or to (3) state $S, E, I-1, R+1$ (a removal event) with rate γI. No other instantaneous transitions are allowed.

where $i_1 \neq i_2$. An example of second order reaction is the infection event in SIR model, where a susceptible individual interacts with an infectious individual, resulting in two infected individuals. Such a reaction has rate $h(\mathbf{X}, \theta) = \theta X_{i_2} X_{i_1}$. The second kinds of a second order reaction has two reactants of the same type:

$$2\mathcal{X}_i \xrightarrow{\theta} \mathcal{X}_j \text{ or } \phi. \tag{2.13}$$

Such reactions rarely appear in infectious disease modeling.

Example: SEIR model

We use a Susceptible-Exposed-Infected-Recovery (SEIR) model to illustrate stochastic kinetic model representation. SEIR model is an extension of the SIR model that assumes a latent period corresponding to an"Exposed" compartment, in which an infected individual does not have the ability to infect others. The exposed individual becomes infectious with rate μ. As in the SIR model, an infectious individual has removal rate γ. The transition events between different states of the SEIR model are depicted in Figure 2.2.

Following the stochastic kinetic model representation, the SEIR model can be viewed as a reaction system of four reactants — susceptible, exposed, infectious, and recovered individuals — and the following three reactions:

$$\text{Susceptible} + \text{Infectious} \xrightarrow{\beta} \text{Exposed} + \text{Infectious}, \tag{2.14}$$

$$\text{Exposed} \xrightarrow{\mu} \text{Infectious}, \tag{2.15}$$

$$\text{Infectious} \xrightarrow{\gamma} \text{Recovered}. \tag{2.16}$$

Since removed individuals never interact with individuals in other compartments, we will only keep track of the counts of susceptible, exposed, and infectious individuals at time t, denoted by $S(t)$, $E(t)$, $I(t)$ respectively. The effect matrix $\mathbf{A}$ for the SEIR model can be written as:

$$\begin{array}{c} \quad\text{Susceptible}\quad\text{Exposed}\quad\text{Infectious} \\ \mathbf{A} = \begin{pmatrix} -1 & 1 & 0 \\ 0 & -1 & 1 \\ 0 & 0 & -1 \end{pmatrix} \begin{array}{l} \text{reaction (2.14)} \\ \text{reaction (2.15)} \\ \text{reaction (2.16)} \end{array} \end{array}, \tag{2.17}$$

with columns representing compartments and rows representing reactant changes during reaction events. If we let $\mathbf{X}(t) = (S(t), E(t), I(t))$ denote the state vector at time t, then the rate vector $\mathbf{h}$ for the SEIR model is

$$\mathbf{h}(\mathbf{X}(t), \boldsymbol{\theta}) = (\beta S(t) I(t), \mu E(t), \gamma I(t))^T.$$

2.2.3 Stochastic Epidemic model

The population trajectory probability/density $\Pr(\mathbf{X}_{0:T} \mid \boldsymbol{\theta})$ in (2.1) can be decomposed into the product of conditional probabilities/densities:

$$\Pr(\mathbf{X}_{0:T} \mid \boldsymbol{\theta}) = \pi(\mathbf{X}_0) \prod_{i=1}^{n} \Pr(\mathbf{X}_i \mid \mathbf{X}_{0:i-1}, \boldsymbol{\theta}),$$

where $\pi(\mathbf{X}_0)$ is the initial distribution of population at time t_0. Stochastic epidemics models assume Markov property so that the population probability/density can be written as the

product of transition probabilities/densities $\Pr(\mathbf{X}_i|\mathbf{X}_{i-1})$:

$$\Pr(\mathbf{X}_{0:T} \mid \boldsymbol{\theta}) = \pi(\mathbf{X}_0) \prod_{i=1}^{n} \Pr\left(\mathbf{X}_i \mid \mathbf{X}_{i-1}, \boldsymbol{\theta}\right). \tag{2.18}$$

Commonly used stochastic epidemic models are Markov Jump Process and stochastic differential equations, which will be reviewed in Section 2.2.4 and Section 2.2.6. Deterministic ordinary differential equation approach is also included in Section 2.2.5. Finally, we introduce the linear noise approximation method that yields closed-form transition densities in Section 2.2.7.

2.2.4 Markov Jump Process

One of the classic stochastic models for infectious disease population dynamics is the Markov Jump Process (MJP). The MJP assumes population dynamics follow a continuous time Markov chain (CTMC). A stochastic process $\{\mathbf{X}(t) :\geq 0\}$ with discrete state space $\mathcal{S}$ is a CTMC if for all $t \geq 0, s \geq 0$, $i \in \mathcal{S}, j \in \mathcal{S}$,

$$\Pr(\mathbf{X}(s+t) = j \mid \mathbf{X}(s) = i, \{\mathbf{X}(u) : 0 \leq u < s\}) = \Pr(\mathbf{X}(s+t) = j \mid \mathbf{X}(s) = i).$$

Furthermore, a CTMC is called time-homogeneous when

$$\Pr(\mathbf{X}(t+s) = j \mid \mathbf{X}(s) = i) = \Pr(\mathbf{X}(t) = j \mid \mathbf{X}(0) = i) = P_{ij}(t)$$

for any time $t > 0, s > 0$ and state $i, j \in \mathcal{S}$ where $P_{ij}(t)$ is the transition probability that the chain will be in state j, t time units from now, given it is in state i now. For finite state space $\mathcal{S}$ and $t \geq 0$, we can define a $|\mathcal{S}| \times |\mathcal{S}|$ transition matrix

$$\mathbf{P}(t) = \{P_{ij}(t)\},$$

with $\mathbf{P}(0) = \mathbf{I}$ since $P_{ij}(0) = \Pr(\mathbf{X}(0) = j \mid \mathbf{X}(0) = i) = \mathbf{1}_{\{i=j\}}$. The transition matrix $P(t)$ is element-wise non-negative, with each row summing to 1. The transition matrix satisfies Chapman-Kolmogorov equation, such that

$$\mathbf{P}(t+s) = \mathbf{P}(s)\mathbf{P}(t), \tag{2.19}$$

for all $t \geq 0, s \geq 0$, i.e

$$P_{ij}(t+s) = \sum_{k \in \mathcal{S}} P_{ik}(s) P_{kj}(t).$$

Infinitesimal generator If $\lim_{h \to 0} \mathbf{P}(h) = \mathbf{P}(0) = \mathbf{I}$, for $i \in \mathcal{S}$, there exists

$$\lambda_i = \lim_{h \to 0} \frac{1 - P_{ii}(h)}{h}$$

and for $i \neq j$, there exists

$$\lambda_{ij} = \lim_{h \to 0} P_{ij}(h)/h.$$

The infinitesimal generator for CTMC is defined as $\mathbf{\Lambda} = \{\lambda_{ij}\}, i, j \in \mathcal{S}$ with $\lambda_{ii} = -\lambda_i$. A CTMC $\mathbf{X}(t)$ is called stable when $\lambda_i < \infty$ and is conservative when $\lambda_i = \sum_{j \neq i} \lambda_{ij}$ for any $i \in \mathcal{S}$. In this thesis, we will work with stable and conservative continuous-time Markov chains.

Kolmogorov forward equation Based on Chapman-Kolmogrov equation (2.19),

$$\frac{\mathrm{d}\mathbf{P}(t)}{\mathrm{d}t} = \lim_{h \to 0} \frac{\mathbf{P}(t+h) - \mathbf{P}(t)}{h} = \lim_{h \to} \frac{\mathbf{P}(t)(\mathbf{P}(h) - \mathbf{I})}{h} = \mathbf{P}(t)\mathbf{\Lambda}. \tag{2.20}$$

By exchanging the order of the multiplication on the RHS of (2.20), we can derive the backward equation

$$\frac{\mathrm{d}\mathbf{P}(t)}{\mathrm{d}t} = \mathbf{\Lambda}\mathbf{P}(t).$$

Under the initial condition $\mathbf{P}(0) = \mathbf{I}$, by forward and backward equation, there's a unique solution for the transition matrix

$$\mathbf{P}(t) = \exp(\mathbf{\Lambda t}) = \sum_{k=0}^{\infty} \frac{(\mathbf{\Lambda}t)^k}{k!}.$$

MJP for infectious disease Assume an epidemic model with d population compartments and total population size N, the MJP for infectious disease model focus on state space

$\mathcal{S} = \left\{(x_1, \ldots, x_d) : x_j \geq 0, x_j \in \mathbb{Z}, \sum_{j=1}^{d} x_j = N\right\}$. The state change in the i-th reaction can be characterized by

$$\mathbf{X} \xrightarrow{\boldsymbol{\theta}} \mathbf{X} + \mathbf{A}^T \mathbf{e}_i,$$

where $\mathbf{A}$ is the effect matrix defined in Section 2.2.2 and $\mathbf{e}_i$ is a q-dimensional vector with the i-th element equal to one and the rest of the elements equal to zero. Hence, the transition rate in the elements of the infinitesimal generator for MJP can be written as

$$\lambda_{(\mathbf{X} \to \mathbf{X}')} = \begin{cases} h_i(\mathbf{X}, \boldsymbol{\theta}) & \mathbf{X}' = \mathbf{X} + \mathbf{A}^T \mathbf{e}_i, \text{for } i = 1, \ldots, q \\ -\sum_{i=1}^{q} h_i(\mathbf{X}, \theta_i) & \mathbf{X}' = \mathbf{X}, \\ 0 & \text{otherwise.} \end{cases}$$

Example: SIR model In the CTMC for the SIR model, the state space $\mathcal{S}$ is a set of all the combinations of S, I and R counts under the constraint that $S + I + R = N$. The transition rates in the infinitesimal generator of CTMC for state $\mathbf{X}$ are elements of vector $\mathbf{h}(\mathbf{X}, \boldsymbol{\theta})$. The infinitesimal generator is a $|\mathcal{S}| \times |\mathcal{S}|$ matrix $\Lambda = \{\lambda_{\{(S,I,R)\to(S',I',R')\}}\}$, where the transition rate $\lambda_{\{(S,I,R)\to(S',I',R')\}}$ is given by

$$\lambda_{\{(S,I,R)\to(S',I',R')\}} = \begin{cases} \beta SI & S' = S - 1, I' = I + 1, R' = R, \\ \gamma I & S' = S, I' = I - 1, R' = R + 1, \\ -(\beta SI + \gamma I) & S' = S, I' = I, R' = R, \\ 0 & \text{otherwise.} \end{cases}$$

Exact realizations of the MJP can be simulated through a Gillespie algorithm [Gillespie, 1977], see Algorithm 1 for more details. Keeling and Ross [2007] demonstrate that for compartment models with large population, calculating transition probability via matrix exponential can be computationally prohibitive. Even for the simple SIR model, the size of the state space $\mathcal{S}$ grows on the order of N^2. Furthermore, if the rates $\boldsymbol{\theta}$ are not constant across time, calculating transition probability by solving the above system of differential equations becomes even more challenging.

2.2.5 Ordinary Differential Equation

Infectious disease dynamics can also be modeled using a set of ordinary differential equations (ODE) [Anderson and May, 1992, Bailey, 1975, Wilkinson, 2011]. The ODE approach corresponds to the approximated mean process of the MJP as the system size N goes to infinity. This approach ignores stochastic in the infectious disease dynamics. Based on the notation in section 4.2.4, the ODE model can be formulated as

$$\frac{\mathrm{d}\mathbf{X}}{\mathrm{d}t} = \mathbf{A}^T\mathbf{h}(\mathbf{X}, \boldsymbol{\theta}). \tag{2.21}$$

Example: SIR model In SIR model, the population vector $\mathbf{X} = (S, I)$ and the rate vector for each reaction is $\mathbf{h} = (\beta SI, \gamma I)^T$. Following the representation (2.21), the ODE representation of SIR model consists of the following two ODEs:

$$\frac{\mathrm{d}S}{\mathrm{d}t} = -\beta SI, \qquad \frac{\mathrm{d}I}{\mathrm{d}t} = \beta SI - \gamma I.$$

An intuitive derivation of ODE method

Here we give an intuitive derivation of ODE method based on [Engblom, 2006, Ferm et al., 2008]. Let $p(\mathbf{x}, t) = \Pr(\mathbf{X}(t) = \mathbf{x})$, the following the definition of infinitesimal generator,

$$\frac{\mathrm{d}p(\mathbf{x}, t)}{\mathrm{d}t} = \sum_{i=1}^{q} p(\mathbf{x} - \mathbf{A}^T\mathbf{e}_i, t)h_i(\mathbf{x} - \mathbf{A}^T\mathbf{e}_i, \theta_i) - p(\mathbf{x}, t)\sum_{i=1}^{q} h_i(\mathbf{x} - \mathbf{A}^T\mathbf{e}_i, \theta_i). \tag{2.22}$$

After multiplying both sides of (2.22) by $\mathbf{x}$ and sum of all possible state $\mathbf{x} \in \mathcal{S}$, one gets

$$\begin{aligned}
\frac{\mathrm{d}\mathbf{E}\mathbf{X}(t)}{\mathrm{d}t} &= \sum_{i=1}^{q} \left\{ \mathbf{E}\left[h_i(\mathbf{X}, \theta_i)\left(\mathbf{X}(t) + \mathbf{A}^T\mathbf{e}_i\right)\right] - \mathbf{E}\left[h_i(\mathbf{X}, \theta_i)\mathbf{X}(t)\right] \right\} \\
&= \sum_{i=1}^{q} \mathbf{A}^T\mathbf{e}_i\mathbf{E}\left[h_i(\mathbf{X}(t), \boldsymbol{\theta})\right] \\
&= \mathbf{E}\left[\mathbf{A}^T\mathbf{h}(\mathbf{X}(t), \boldsymbol{\theta})\right].
\end{aligned}$$

By approximating the expression $\mathbf{E}\left[\mathbf{A}^T\mathbf{h}(\mathbf{X}(t), \boldsymbol{\theta})\right]$ with $\mathbf{A}^T\mathbf{H}(\mathbf{E}(\mathbf{X}(t)), \boldsymbol{\theta})$, we have ODE equation system for the expected population $\mathbf{E}[\mathbf{X}(t)]$,

$$\frac{\mathrm{d}\mathbf{E}\mathbf{X}(t)}{\mathrm{d}t} = \mathbf{A}^T\mathbf{h}(\mathbf{E}\mathbf{X}(t), \boldsymbol{\theta})$$

Theoretical perspective of ODE methods

System expansion The reaction network representation for stochastic mechanistic model satisfies the following system expansion assumption.

ASSUMPTION 1 (System expansion assumption): The rate function $h_i(\mathbf{X}_t, \boldsymbol{\theta})$ has third derivative in space and time, and there exist function v such that, $h_i(\mathbf{x}, \boldsymbol{\theta})$ can be written as

$$h_i(\mathbf{x}, \boldsymbol{\theta}) = N v_i(\mathbf{x}/N). \tag{2.23}$$

As we rescale both the compartment size and reaction rate by applying the following transform,

$$\tilde{\boldsymbol{\theta}}_i = N^{m_i - 1} \boldsymbol{\theta}_i, \tag{2.24}$$

where $m_i = \sum_{j=1}^{d} a_{ij}$ denote the number of the reactants on the LHS of reaction equation, Assumption 1 for ODE (2.21) is satisfied by setting $v_i(\mathbf{x}, \theta_i) = h_i\left(\mathbf{x}/N, \tilde{\theta}_i\right)$.The ODE for the rescaled system can be written as

$$\frac{\mathrm{d}\tilde{\boldsymbol{\eta}}}{\mathrm{d}t} = \mathbf{A}^T \mathbf{h}(\tilde{\boldsymbol{\eta}}, \tilde{\boldsymbol{\theta}}), \tag{2.25}$$

with initial condition $\tilde{\boldsymbol{\eta}}_0 = \dfrac{X(0)}{N}$. The ODE can be viewed as a asymptotic mean of the MJP for rescaled population as $N \to +\infty$, which can be characterized by the following theorem from [Kurtz, 1970, 1971].

THEOREM 1: *Let* $\mathbf{X} = (X_1, \ldots, X_n)$ *be a vector of non-negative stochastic random variables with solving the master equation. Suppose the rate vector* $h(\mathbf{x}, \boldsymbol{\theta})$ *satisfies Assumption* (1) *with system size* N *and the initial conditions are such that* $\lim_{N\to\infty} N^{-1}\mathbf{X}(0) = \tilde{\boldsymbol{\eta}}_0$ *then for any* $\delta > 0$*,*

$$\lim_{N\to\infty} P\left(\sup_{t \le T} ||N^{-1}\mathbf{X}(t) - \tilde{\boldsymbol{\eta}}(t)|| > \delta\right) = 0, \tag{2.26}$$

where $\tilde{\boldsymbol{\eta}}(t)$ *is the solution of ODE* (2.25)*.*

The advantages of the ODE-based modeling is that there is no need to compute probabilities or densities of latent (S, I, R) counts. The trajectory can be obtained directly by integrating the ODE for given and initial condition and ODE parameters. In most of the reactions, the ODE is non-linear, which means there is no closed-form solution. For non-linear ODE systems, numerical methods such as Euler method and Runge-Kutta method can be used to obtain numerically ODE solutions.

Algorithm 1 Gillespie's Algorithm for MJP simulation

Input: Reaction effect matrix $\mathbf{A}$ rate parameter $\boldsymbol{\theta}$ and rate function $\mathbf{h}(\cdot, \cdot)$. Initial state $\mathbf{X}(t_0) = \mathbf{x}_0$ and a grid of interests $t_0 < t_1 < \ldots < t_T$.
Output: A sequence of state $\mathbf{X}_{1:T}$, where $\mathbf{X}_i := \mathbf{X}_0(t_i)$
Initialize $i \leftarrow 0$, $t \leftarrow t_0$, $x \leftarrow x_0$, $i \leftarrow 1$.
while $t < t_T$ **do**
    $\lambda \leftarrow \mathbf{1}^T \mathbf{h}(\mathbf{X}, \boldsymbol{\theta})$, $\mathbf{p} \leftarrow \mathbf{h}(\mathbf{X}, \boldsymbol{\theta})/\lambda$.
    Sample $\tau \sim \text{Exp}(\lambda)$ and $j \sim \text{Multinomial}(\{1, \ldots, q\}, \mathbf{p})$.
    $t \leftarrow t + \tau$ and $\mathbf{x} \leftarrow \mathbf{x} + \mathbf{A}^T \mathbf{e}_j$.
    while $t_i < t$ **do**
        $\mathbf{X}_i \leftarrow \mathbf{x}$.
        $i \leftarrow i + 1$.
Return: A sequence of states $\mathbf{X}_1 : T$ observed at $t_1, \ldots, t_T$.

2.2.6 Stochastic Differential Equation

Stochastic differential equation (SDE) method is a continuous Markov process that operates in a continuous state space. For random vector $\mathbf{X}(t) \in \mathbb{R}^d$, d-dimensional Wiener process $\mathbf{W}_t$, and a Cholesky triangle of the $d \times d$ covariance matrix, denoted by the square root $\sqrt{\cdot}$, the SDE of $\mathbf{X}(t)$ has the following form:

$$\mathrm{d}\mathbf{X}(t) = \boldsymbol{\mu}(t, \mathbf{X}(t))\mathrm{d}t + \sqrt{\boldsymbol{\Sigma}(t, \mathbf{X}(t))}\mathrm{d}\mathbf{W}_t, \quad \mathbf{X}(t_0) = \mathbf{x}_0, \tag{2.27}$$

where $\boldsymbol{\mu}(t, \mathbf{X}_t) \in \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^d$ and $\Sigma(t, \mathbf{X}_t) \in \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^{d\times d}$ denote the drift vector and diffusion matrix respectively. The solution of the SDE is the stochastic integral

$$\mathbf{X}(t) = x_0 + \int_{t_0}^{t} \boldsymbol{\mu}\left(\tau, \mathbf{X}(\tau)\right) \mathrm{d}\tau + \int_{t_0}^{t} \sqrt{\boldsymbol{\Sigma}(\tau, \mathbf{X}(\tau))}\mathrm{d}\mathbf{W}_\tau. \tag{2.28}$$

For a infinitely small time step Δt, the SDE can be interpreted as the limit of difference equation when $\Delta t \to 0$:

$$\mathbf{X}(t + \Delta t) = \mathbf{X}_t + \boldsymbol{\mu}(t, \mathbf{X}(t))\Delta t + \sqrt{\boldsymbol{\Sigma}(t, \mathbf{X}(t))}\Delta \mathbf{W}_t,$$

where $\sqrt{\boldsymbol{\Sigma}(t, \mathbf{X}(t))}\Delta \mathbf{W}_t$ denotes a $d-$dimensional multivariate normal random vector with mean $\mathbf{0}$ and covariance matrix $\boldsymbol{\Sigma}(t, \mathbf{X}_t)\Delta t$.

The transform of SDE satisfies Itô's formula [Øksendal, 2003],

THEOREM 2 (Itô's formula): *Let $\mathbf{X}(t)$ be a n-dimensional Itô process. Let $\mathbf{g}(t, x) = (g_1(t, \mathbf{x}), \ldots, g_k(t, \mathbf{x}))$ be a twice differentiable function map from $[0, +\infty) \times \mathbb{R}^n \to \mathbb{R}^p$, i.e $\mathbf{g}(\cdot, \cdot) \in C^2$. Then the process $\mathbf{Y}(t) = \mathbf{g}\left(t, \mathbf{X}(t)\right)$ is a Itô process given by the SDE formula*

$$\begin{aligned}
\mathrm{d}\mathbf{Y} &= \frac{\partial \mathbf{g}}{\partial t}\mathrm{d}t + \left(\frac{\partial \mathbf{g}}{\partial \mathbf{x}}\right)^T \mathrm{d}\mathbf{X}(t) + \frac{1}{2}(\mathrm{d}\mathbf{X}(t))^T \left(\frac{\partial^2 \mathbf{g}}{\partial \mathbf{x}\mathbf{x}^T}\right) \mathrm{d}\mathbf{X}(t) \\
&= \left\{ \frac{\partial \mathbf{g}}{\partial t} + \frac{\partial \mathbf{g}}{\partial \mathbf{x}}\boldsymbol{\mu}(t, \mathbf{X}(t)) + \frac{1}{2} Tr\left[\sqrt{\boldsymbol{\Sigma}(t, \mathbf{X}(t))} \left(\frac{\partial^2 f}{\partial \mathbf{x}\mathbf{x}^T}\right) \sqrt{\boldsymbol{\Sigma}(t, \mathbf{X}(t))} \right] \right\} \mathrm{d}t \\
&+ \left(\frac{\partial \mathbf{g}}{\partial \mathbf{x}}\right)^T \sqrt{\boldsymbol{\Sigma}(t, \mathbf{X}(t))}\mathrm{d}\mathbf{W}_t.
\end{aligned} \tag{2.29}$$

SDEs for Infectious Disease Modeling

A stochastic way to approximate the MJP infectious disease model is to use the SDE approximation, also known as the chemical Langevin equation (CLE) [Gillespie, 2000]. The SDE can be viewed an approximation of the MJP at time t by applying normal approximation to the Poisson distributed number of state transitions in a small interval of time $(t, t + \Delta t)$ [Gillespie, 2000, Wallace, 2010]. Here we provide an intuitive derivation of SDE approximations from the MJP of epidemic models.

Let $N_i(t)$ be the cumulative number of i-th reactive events happening up to time t and $\boldsymbol{N}(t) = (N_1(t), \ldots, N_q(t))^T$ be the cumulative reaction counts vector. Hence, $\mathbf{X}(t)$ —- the reactant population counts — and $\boldsymbol{N}(t)$ have the following linear relationship

$$\mathbf{X}(t) = \mathbf{X}(t_0) + \mathbf{A}^T \boldsymbol{N}(t). \tag{2.30}$$

For an infinitely small $\Delta t > 0$, the reaction is assumed to be constant on $[t, t + \Delta t]$, i.e $\mathbf{h}(\mathbf{X}(t'), \boldsymbol{\theta}) \approx \mathbf{h}(\mathbf{X}(t), \boldsymbol{\theta})\mathrm{d}t$, $t' \in [t + \Delta t]$. For reaction system with large population size N, the reaction rate $\mathbf{h}(\mathbf{X}(t), \boldsymbol{\theta})$ can be sufficiently large such that $\mathbf{h}(\mathbf{X}_t, \boldsymbol{\theta})\Delta t >> 1$. The system can be viewed as q independent Poisson process with rate $h_i(\mathbf{X}(t), \boldsymbol{\theta})$ for $i = 1, \ldots, n$ at $[t, t + \Delta]$. The difference in the cumulative number of reactions $\Delta \boldsymbol{N}(t)$ follows Poisson distribution and can be approximated using normal distribution,

$$\Delta \boldsymbol{N}_i(t) \sim \text{Poisson}\left(h_i(\mathbf{X}(t), \boldsymbol{\theta})\Delta t\right) \approx \mathcal{N}\left(h_i(\mathbf{X}(t), \boldsymbol{\theta})\Delta t, h_i(\mathbf{X}(t), \boldsymbol{\theta})\Delta t\right). \tag{2.31}$$

Following the normal approximation in 2.31 and the interpretation of SDE, the cumulative number reactions $\boldsymbol{N}(t)$ can be approximated with the following SDE,

$$\mathrm{d}\boldsymbol{N}(t) = \mathbf{h}(\mathbf{X}(t), \boldsymbol{\theta})\mathrm{d}t + \sqrt{\mathbf{H}(\mathbf{X}(t), \theta)\mathrm{d}}\mathbf{W}_t. \tag{2.32}$$

By applying Itô formula (2.32) using transform (2.30), we can derive the SDE approximation of the MJP for general stochastic kinetic models: .

$$\mathrm{d}\mathbf{X}(t) = \mathbf{A}^T \mathbf{h}(\mathbf{X}(t), \boldsymbol{\theta})\mathrm{d}t + \sqrt{\mathbf{A}^T \mathbf{H}(\mathbf{X}(t), \boldsymbol{\theta})\mathbf{A}} \cdot \mathrm{d}\mathbf{W}_t. \tag{2.33}$$

The deterministic part in SDE corresponds to the right hand side of ODE (2.34) and stochastic part is related to the variance of the system.

One of the advantages of SDE over MJP is that it is scalable for larger population sizes. Hence, computation and simulation complexity does not grow with the population size. Usually there is no closed-form solution for non-linear SDE systems (2.33). The numerical solution can be obtained by simulation via an Euler-Maryama method [Iacus, 2009] with small interval Δt. The SDE method is widely used in modeling infectious population

dynamics given observed incidence data or prevalence data [Golightly and Wilkinson, 2005, 2008, 2011]. However, the integration using the Eular-Maryama method requires dividing the time interval into small time intervals with length Δt. The transition density $\Pr(\mathbf{X}_i|\mathbf{X}_{i-1}, \boldsymbol{\theta})$ is non-tractable for large $t_i - t_{i-1}$. The implementation of SDE involves discretizing time between observation times over a set of grid points and using computationally-intensive methods that impute values of the state at both the observation times and the grid of times between observations. Golightly and Wilkinson [2005] used a vanilla data augmentation approach and Golightly and Wilkinson [2008, 2011] used sequential Monte Carlo method and particle MCMC to perform such imputation. Other methods include approximated Bayesian computation methods [Sisson et al., 2007]. Statistical methods for non-linear SDEs are always computationally intensive and involve high dimensional latent variables, resulting in slow MCMC mixing time and long running time per iteration.

2.2.7 Linear Noise Approximation

The linear noise approximation (LNA) is known as the lowest-order correction to the deterministic ODE in a Markov Jump process of chemical reactions. The original derivation is based on Taylor expansion of the Fokker-Planck equation for the process [Van Kampen and Reinhardt, 1983, Kurtz, 1971]. Recent work [Wallace, 2010, Ferm et al., 2008] shows a more intuitive derivation by considering the LNA as a first order approximation of the diffusion process related to the CLE of the stochastic kinetic model. Here we give a heuristic illustration of linear noise approximation.

The idea of LNA is to decompose the process into two parts: a deterministic time part $\boldsymbol{\eta}(t)$ related to the deterministic solution of the ODE and a stochastic part $\mathbf{M}(t)$ as a Gaussian process, that is $\mathbf{X}(t) = \boldsymbol{\eta}(t) + \mathbf{M}(t)$, where $\boldsymbol{\eta}(t)$ is the solution of

$$\mathrm{d}\boldsymbol{\eta}(t) = \mathbf{A}^T \mathbf{h}(\boldsymbol{\eta}(t), \boldsymbol{\theta})\mathrm{d}t. \tag{2.34}$$

By applying the first order Taylor expansion to (2.33) at $\boldsymbol{\eta}(t)$ and ignoring the higher order

terms, (2.33) yields the SDE for $\mathbf{M}(t)$:

$$\mathrm{d}\mathbf{M}(t) = \mathbf{F}(\boldsymbol{\eta}, \boldsymbol{\theta})\mathbf{M}(t)\mathrm{d}t + \sqrt{\mathbf{A}^T\mathbf{H}(\boldsymbol{\eta}, \boldsymbol{\theta})\mathbf{A}}\mathrm{d}\mathbf{W}_t, \tag{2.35}$$

where $\mathbf{F}(\boldsymbol{\eta}, \boldsymbol{\theta}) := \left.\dfrac{\partial \mathbf{A}^T\mathbf{h}(\mathbf{X}(t), \boldsymbol{\theta})}{\partial \mathbf{X}}\right|_{\mathbf{X}=\boldsymbol{\eta}}$ is the Jacobian matrix of the deterministic part $\mathbf{A}^T\mathbf{h}(\mathbf{X}, \boldsymbol{\theta})$ in (2.34) at $\boldsymbol{\eta}$. In SDE (2.35), both the deterministic part $\mathbf{F}(\boldsymbol{\eta}, \boldsymbol{\theta})$ and stochastic part $\sqrt{\mathbf{A}^T\mathbf{H}(\boldsymbol{\eta}, \boldsymbol{\theta})\mathbf{A}}$ do not depend on the residual process $\mathbf{M}(t)$ itself. Hence, (2.35) is a linear SDE, which solution is a Gaussian process. Furthermore, the solution $\mathbf{M}(t)$ can be recovered by solving two ordinary differential equations governing the mean process $\mathbf{m}(t) := \mathbf{E}[\mathbf{M}(t)]$ and variance process $\boldsymbol{\Phi}(t) := \mathbf{Var}(\mathbf{M}(t))$, such that

$$\mathrm{d}\mathbf{m}(t) = \mathbf{F}(\boldsymbol{\eta}, \boldsymbol{\theta})\mathbf{m}(t)\mathrm{d}t, \tag{2.36}$$

$$\mathrm{d}\boldsymbol{\Phi}(t) = \left(\mathbf{F}(\boldsymbol{\eta}, \boldsymbol{\theta})\boldsymbol{\Phi}(t) + \boldsymbol{\Phi}(t)(\mathbf{F}(\boldsymbol{\eta}, \boldsymbol{\theta}) + \mathbf{A}^T\mathbf{H}(\boldsymbol{\eta}, \boldsymbol{\theta})\mathbf{A}\right)\mathrm{d}t. \tag{2.37}$$

Let $\boldsymbol{\eta}_{t_{i-1}}, \mathbf{m}_{t_{i-1}}, \boldsymbol{\Phi}_{t_{i-1}}$ denote the initial values of $\boldsymbol{\eta}(t), \mathbf{m}(t), \boldsymbol{\Phi}(t)$ at time t_{i-1} in differential equations (2.34), (2.36), and (2.37) respectively. There are two options for choosing these initial conditions: the non-restarting LNA of Komorowski et al. [2009] and the restarting LNA of Fearnhead et al. [2014].

Non-restarting LNA The non-starting LNA solves the system of ODEs (2.34), (2.36), (2.37) with the following choice of initial conditions:

1. $\boldsymbol{\eta}_{t_{i-1}} = \boldsymbol{\eta}(t_{i-1})$, where $\boldsymbol{\eta}(t_{i-1})$ was obtained by solving the ODE (3.7) over the interval $[t_{i-2}, t_{i-1}]$,

2. $\mathbf{m}_{t_{i-1}} = \mathbf{X}_{i-1} - \boldsymbol{\eta}(t_{i-1})$,

3. $\boldsymbol{\Phi}_{t_{i-1}} = \mathbf{0}$,

and obtain $\boldsymbol{\eta}(t_i), \mathbf{m}(t_i)$ and $\boldsymbol{\Phi}(t_i)$. The solution $\mathbf{m}(t_i)$ will be a function of the initial value $\mathbf{X}_{i-1} - \boldsymbol{\eta}(t_{i-1})$, the interval length $\Delta t_i := t_i - t_{i-1}$. To make this dependence explicit, we write $\mathbf{m}(t_i) := \boldsymbol{\mu}\left(\mathbf{X}_{i-1} - \boldsymbol{\eta}(t_{i-1}), \Delta t_i, \boldsymbol{\theta}\right)$. Since (2.36) is a first order homogeneous linear

ODE, the solution $\boldsymbol{\mu}\left(\mathbf{X}_{i-1}-\boldsymbol{\eta}(t_{i-1}), \Delta t_i, \boldsymbol{\theta}\right)$ is a linear function of $\mathbf{X}_{i-1}-\boldsymbol{\eta}(t_{i-1})$. Hence, the transition from $\mathbf{X}_{i-1}$ to $\mathbf{X}_i$ follows the following Gaussian distribution:

$$\mathbf{X}_i \mid \mathbf{X}_{i-1}, \boldsymbol{\theta} \sim \mathcal{N}\left(\boldsymbol{\eta}(t_i)+\boldsymbol{\mu}\left(\mathbf{X}_{i-1}-\boldsymbol{\eta}(t_{i-1}), \Delta t_i, \boldsymbol{\theta}_{i-1}\right), \boldsymbol{\Phi}(t_i)\right). \tag{2.38}$$

Restarting LNA The restarting LNA solves ODEs (2.34), (2.36), (2.37) with the following choice of initial conditions:

1. $\boldsymbol{\eta}_{t_{i-1}} = \mathbf{X}_{i-1}$,
2. $\mathbf{m}_{t_{i-1}} = \mathbf{0}$,
3. $\boldsymbol{\Phi}_{t_{i-1}} = \mathbf{0}$.

The solution $\mathbf{m}(t_i)$ will be $\mathbf{0}$ and $\boldsymbol{\eta}(t_i)$ will be a non-linear function of $\mathbf{X}_{i-1}$, denoted by $\boldsymbol{\eta}(\mathbf{X}_i, \Delta t_i, \boldsymbol{\theta})$. The transition from $\mathbf{X}_{i-1}$ to $\mathbf{X}_i$ have the following Gaussian distribution:

$$\mathbf{X}_i \mid \mathbf{X}_{i-1}, \boldsymbol{\theta}_{i-1} \sim \mathcal{N}\left(\boldsymbol{\eta}\left(\mathbf{X}_{i-1}, \Delta t_i, \boldsymbol{\theta}_{i-1}\right), \boldsymbol{\Phi}(t_i)\right). \tag{2.39}$$

Derivation of LNA

Here we provide an intuitive derivation of LNA from SDE based on [Wallace, 2010].

THEOREM 3 (Linear Noise Approximation for SDE): *Let $\boldsymbol{\eta}_t$ be the solution of ordinary differential equation* (2.34) *with initial value $\boldsymbol{\eta}_0$. Let N be the system size, which is the total number of individuals in the system (In SIR model, N will be the total population, i.e $N = S + I + R$), $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_q)$ denote the vector of rate parameters in q reactions. Then the solution $\mathbf{X}_t$ of the CLE* (2.33) *satisfies,*

$$\frac{1}{\sqrt{N}}\mathrm{d}(\mathbf{X}_t - \boldsymbol{\eta}_t) = \frac{1}{\sqrt{N}}\left(\mathbf{F}(\boldsymbol{\eta}_t, \boldsymbol{\theta})(\mathbf{X}_t - \boldsymbol{\eta}_t) + o(1)\right)\mathrm{d}t + \left(\frac{1}{\sqrt{N}}\sqrt{\mathbf{A}^T\mathbf{H}(\boldsymbol{\eta}, \boldsymbol{\theta})\mathbf{A}} + o(1)\right)\mathrm{d}\mathbf{W}_t \tag{2.40}$$

as $N \to +\infty$.

Proof. We rescale both the compartment size and reaction rate by applying the following transform,

$$\tilde{\mathbf{X}} = N^{-1} \cdot \mathbf{X} \tag{2.41}$$

$$\tilde{\boldsymbol{\theta}}_i = N^{m_i - 1}\boldsymbol{\theta}_i, \tag{2.42}$$

where $m_i = \sum_{j=1}^{d} a_{ij}$ and $\boldsymbol{\theta}_i$ denote the number of the reactants and reaction rate in reaction i as in Section 2.2.2. The transformed $\tilde{\mathbf{X}}$ represents the proportion of each compartment with respect to the total population. Then we have $\mathbf{h}(\mathbf{X}(t), \boldsymbol{\theta}) = N\mathbf{h}(\tilde{\mathbf{X}}, \tilde{\boldsymbol{\theta}})$ and $\mathbf{F}(\boldsymbol{\eta}, \boldsymbol{\theta}) = \mathbf{F}(\tilde{\boldsymbol{\eta}}, \tilde{\boldsymbol{\theta}})$. Hence, the SDE (2.33) becomes

$$\mathrm{d}\tilde{\mathbf{X}} = \mathbf{A}^T\mathbf{h}(\tilde{\mathbf{X}}, \tilde{\boldsymbol{\theta}})\mathrm{d}t + \frac{1}{\sqrt{N}}\sqrt{\mathbf{A}^T\mathbf{H}(\tilde{\mathbf{X}}, \tilde{\boldsymbol{\theta}})\mathbf{A}} \cdot \mathrm{d}\mathbf{W}_t. \tag{2.43}$$

Let $\tilde{\boldsymbol{\eta}}(t)$ be the solution of the ODE

$$\mathrm{d}\tilde{\boldsymbol{\eta}} = \mathbf{A}^T\mathbf{h}(\tilde{\boldsymbol{\eta}}(t), \tilde{\boldsymbol{\theta}})\mathrm{d}t, \tag{2.44}$$

and we have $\boldsymbol{\eta} = N\tilde{\boldsymbol{\eta}}$, where $\boldsymbol{\eta}$ is the solution of the ODE (2.34). $\tilde{\boldsymbol{\eta}}_t$ can be viewed as a scaled version solution of master equation (2.34). Let $\boldsymbol{\xi}(t) = \sqrt{N}\left(\tilde{\mathbf{X}}(t) - \tilde{\boldsymbol{\eta}}(t)\right) = \frac{1}{\sqrt{N}}\left(\mathbf{X}(t) - \boldsymbol{\eta}(t)\right)$ denote the scaled residual, then $\tilde{\mathbf{X}}(t)$ can be written as

$$\tilde{\mathbf{X}} = \frac{1}{\sqrt{N}}\boldsymbol{\xi} + \tilde{\boldsymbol{\eta}}. \tag{2.45}$$

By applying first order Taylor expansion on $\boldsymbol{\eta}$, the SDE (2.43) becomes

$$\begin{aligned}
\mathrm{d}\tilde{\mathbf{X}} &= \mathbf{A}^T\mathbf{h}\left(\tilde{\boldsymbol{\eta}} + \frac{1}{\sqrt{N}}\boldsymbol{\xi}, \tilde{\boldsymbol{\theta}}\right)\mathrm{d}t + \sqrt{\mathbf{A}^T\mathbf{H}\left(\tilde{\boldsymbol{\eta}} + \frac{1}{\sqrt{N}}\boldsymbol{\xi}, \tilde{\boldsymbol{\theta}}\right)\mathbf{A}} \cdot \mathrm{d}\mathbf{W}_t \\
&\approx \left(\mathbf{A}^T\mathbf{h}(\tilde{\boldsymbol{\eta}}, \tilde{\boldsymbol{\theta}}) + \mathbf{F}(\tilde{\boldsymbol{\eta}}, \tilde{\boldsymbol{\theta}}) \cdot \frac{1}{\sqrt{N}}\boldsymbol{\xi} + \mathcal{O}(N^{-1})\right)\mathrm{d}t + \frac{1}{\sqrt{N}}\sqrt{\mathbf{A}^T\mathbf{H}(\tilde{\boldsymbol{\eta}}, \tilde{\boldsymbol{\theta}})\mathbf{A} + \mathcal{O}(\frac{1}{\sqrt{N}})} \cdot \mathrm{d}\mathbf{W}_t \\
&= \left(\mathbf{A}^T\mathbf{h}(\tilde{\boldsymbol{\eta}}, \tilde{\boldsymbol{\theta}}) + \frac{1}{\sqrt{N}}\mathbf{F}(\tilde{\boldsymbol{\eta}}, \tilde{\boldsymbol{\theta}}) \cdot \boldsymbol{\xi}\right)\mathrm{d}t + \frac{1}{\sqrt{N}}\sqrt{\mathbf{A}^T\mathbf{H}(\tilde{\boldsymbol{\eta}}, \tilde{\boldsymbol{\theta}})\mathbf{A}} \cdot \mathrm{d}\mathbf{W}_t + o(N^{-1/2})\mathrm{d}\mathbf{W}_t + o(N^{-1})\mathrm{d}t.
\end{aligned}$$

By subtracting (2.44) and multiplied $\sqrt{N}$ on the two ends, the above equation becomes a differential equation with respect to $\boldsymbol{\xi}$, we have

$$\mathrm{d}\boldsymbol{\xi} = \mathbf{F}(\tilde{\boldsymbol{\eta}}, \boldsymbol{\theta})\boldsymbol{\xi}\mathrm{d}t + \sqrt{\mathbf{A}^T\mathbf{H}(\tilde{\boldsymbol{\eta}}, \tilde{\boldsymbol{\theta}})\mathbf{A}} \cdot \mathrm{d}\mathbf{W}_t + o(N^{-\frac{1}{2}})\mathrm{d}\mathbf{W}_t + o(N^{-1})\mathrm{d}t. \tag{2.46}$$

After multiplied by $\sqrt{N}$, the above equation will end up with (2.40).

Recall that $\mathbf{M}$ is the solution of (2.35) with initial condition $\mathbf{M}_0 = \mathbf{X}_0 - \boldsymbol{\eta}_0$. We can expect $\boldsymbol{\eta} + \mathbf{M}$ to be an approximation of $\mathbf{X}$. Based on the local Lipschitz property of $\mathbf{F}(\boldsymbol{\eta}(t), \boldsymbol{\theta})$ on t and $\mathbf{A}^T\mathbf{H}(\boldsymbol{\eta}(t), \boldsymbol{\theta})$, $\mathbf{X}_t$ can be approximate by $\boldsymbol{\eta} + \mathbf{M}_t$ with

$$\mathbf{X}_t = \boldsymbol{\eta}_t + \mathbf{M}_t + o(N^{1/2}). \tag{2.47}$$

for fixed t as system size $N \to +\infty$.

Derivation of equation (2.36) *and* (2.37)

LEMMA 1 (Solution of Linear Ordinary equation system): *Let* $\mathbf{F}(t) \in \mathbb{R}^{d\times d}$ *and* $\mathbf{X}_t \in \mathbb{R}^d$ *be function of defined on* $\{t : t \geq 0\}$ *that satisfies the following linear ordinary equation*

$$\mathrm{d}\mathbf{X}_t = \mathbf{F}(t)\mathbf{X}_t\mathrm{d}t, \qquad \mathbf{X}_0 = \mathbf{x}_0. \tag{2.48}$$

For $t \geq 0$*, the solution of (2.48) can be represented as*

$$\mathbf{X}_t = \boldsymbol{\Sigma}(t, 0)\mathbf{x}_0,$$

where $\boldsymbol{\Sigma}(t, 0)$ *is the solution of ordinary differential equation in* $\mathbb{R}^{d\times d}$

$$\mathrm{d}\boldsymbol{\Sigma}(t, 0) = \mathbf{F}(t)\boldsymbol{\Sigma}(t, 0)\mathrm{d}t, \qquad \boldsymbol{\Sigma}(0, 0) = \mathbf{I}.$$

Lemma 1 gives the solution of linear ODE. Hence, the solution for the linear ODE 2.36 is on $[t_{i-1}, t]$ will be

$$\mathbf{m}(t) = \boldsymbol{\Sigma}(t, t_{i-1})\mathbf{m}_{i-1},$$

where $\boldsymbol{\Sigma}(t, t_{i-1})$ is the transition matrix by

$$\mathrm{d}\boldsymbol{\Sigma}(t, t_{i-1}) = \mathbf{F}(\boldsymbol{\eta}(t), \boldsymbol{\theta})\boldsymbol{\Sigma}(t, t_{i-1})\mathrm{d}t, \qquad \boldsymbol{\Sigma}(t_{i-1}; t_{i-1}) = \mathbf{I}, \tag{2.49}$$

and $\mathbf{m}_{i-1}$ is the initial value for $\mathbf{m}$ at time t_{i-1}.

THEOREM 4: *Let* $\{\mathbf{M}(t)\}_{t\geq 0} \in \mathbb{R}^d$ *be stochastic process that satisfies the following Stochastic differential equation,*

$$\mathrm{d}\mathbf{M}(t) = \mathbf{F}(\boldsymbol{\eta}, \boldsymbol{\theta})\mathbf{M}(t)dt + \sqrt{\mathbf{A}^T\mathbf{H}(\boldsymbol{\eta}, \boldsymbol{\theta})\mathbf{A}}\mathrm{d}\mathbf{W}_t. \tag{2.50}$$

Then the solution of (2.50) is the Gaussian process

$$\mathbf{M}(t) = \boldsymbol{\Sigma}(t, t_0)\left(\mathbf{M}(t_0) + \int_{t_0}^{t} \boldsymbol{\Sigma}^{-1}(s, t_0)\sqrt{\mathbf{A}^T\mathbf{H}(\boldsymbol{\eta}, \boldsymbol{\theta})\mathbf{A}}d\mathbf{W}_s\right),$$

with mean process $\mathbf{m}(t) := \mathbf{E}[\mathbf{M}_t]$ *satisfies* (2.36) *and variance process* $\boldsymbol{\Phi}(t) := Var(\mathbf{M}_t)$ *satisfies* (2.37)*.*

Proof. Define matrix function $\boldsymbol{\Sigma}(t, t_0)$ as (2.49). First we apply the linear transform $\tilde{\mathbf{M}}(t) = \boldsymbol{\Sigma}^{-1}(t; t_0)\mathbf{M}(t)$. Based on Itô's formula 2, (2.50) can be simplified as a SDE of $\tilde{\mathbf{M}}(t)$:

$$\mathrm{d}\tilde{\mathbf{M}}(t) = \boldsymbol{\Sigma}^{-1}(t; t_0)\sqrt{\mathbf{A}^T\mathbf{H}(\boldsymbol{\eta}, \boldsymbol{\theta})\mathbf{A}}\mathrm{d}\mathbf{W}_t,$$

with solution.

$$\tilde{\mathbf{M}}(t) = \tilde{\mathbf{M}}(t_0) + \int_{t_0}^{t} \boldsymbol{\Sigma}^{-1}(s; t_0)\sqrt{\mathbf{A}^T\mathbf{H}(\boldsymbol{\eta}, \boldsymbol{\theta})\mathbf{A}} \cdot d\mathbf{W}_s$$

Then the solution of $\mathbf{M}(t)$ is

$$\mathbf{M}(t) = \boldsymbol{\Sigma}(t, t_0)\left(\mathbf{M}(t_0) + \int_{t_0}^{t} \boldsymbol{\Sigma}^{-1}(s, t_0)\sqrt{\mathbf{A}^T\mathbf{H}(\boldsymbol{\eta}, \boldsymbol{\theta})\mathbf{A}}d\mathbf{W}_s\right). \tag{2.51}$$

$\boldsymbol{\Sigma}(t, t_0)\mathbf{M}(t_0)$ in (2.51) is a deterministic function of t. In integral with $\int_{t_0}^{t} \boldsymbol{\Sigma}^{-1}(s, t_0)\sqrt{\mathbf{A}^T\mathbf{H}(\eta)\mathbf{A}}d\mathbf{W}_s$ in (2.51) should be Gaussian random variable with mean $\mathbf{0}$ since it is a linear combination of the increments of Brownian motion with different variance. Hence, the $\mathbf{M}(t)$ should be a Gaussian process. By taking the expectation of (2.51), the mean of $\mathbf{m}(t) = \mathbf{E}[\mathbf{M}_t]$ satisfies

$$\mathbf{m}(t) = \boldsymbol{\Sigma}(t, t_0)\mathbf{m}(t_0),$$

which corresponds to the solution of ODE (3.9).

For the variance process, from (2.51),

$$\boldsymbol{\Phi}(t) = \boldsymbol{\Sigma}(t, t_0)\int_{t_0}^{t} \boldsymbol{\Sigma}^{-1}(s, t_0)\mathbf{A}^T\mathbf{H}(\boldsymbol{\eta}, \boldsymbol{\theta})\mathbf{A}\boldsymbol{\Sigma}^{-1}(s, t_0)ds\boldsymbol{\Sigma}(t, t_0). \tag{2.52}$$

By differentiation with respect to t, (2.52) becomes

$$\begin{aligned}
\mathrm{d}\boldsymbol{\Phi}(t) &= \mathrm{d}\boldsymbol{\Sigma}(t,t_0)\cdot\int_{t_0}^{t}\boldsymbol{\Sigma}^{-1}(s,t_0)\mathbf{A}^T\mathbf{H}(\boldsymbol{\eta},\boldsymbol{\theta})\mathbf{A}\boldsymbol{\Sigma}^{-T}(s,t_0)\mathrm{d}s\cdot\boldsymbol{\Sigma}^T(t,t_0)\\
&+ \boldsymbol{\Sigma}(t,t_0)\cdot\mathrm{d}\left[\int_{t_0}^{t}\boldsymbol{\Sigma}^{-1}(s,t_0)\mathbf{A}^T\mathbf{H}(\boldsymbol{\eta},\boldsymbol{\theta})\mathbf{A}\boldsymbol{\Sigma}^{-T}(s,t_0)\mathrm{d}s\right]\cdot\boldsymbol{\Sigma}^T(t,t_0)\\
&+ \boldsymbol{\Sigma}(t,t_0)\cdot\int_{t_0}^{t}\boldsymbol{\Sigma}^{-1}(s,t_0)\mathbf{A}^T\mathbf{H}(\boldsymbol{\eta},\boldsymbol{\theta})\mathbf{A}\boldsymbol{\Sigma}^{-T}(s,t_0)\mathrm{d}s\cdot\mathrm{d}\boldsymbol{\Sigma}^T(t,t_0)\\
&= \mathbf{F}(\boldsymbol{\eta}(t),\boldsymbol{\theta})\cdot\boldsymbol{\Sigma}(t,t_0)\int_{t_0}^{t}\boldsymbol{\Sigma}^{-1}(s,t_0)\mathbf{A}^T\mathbf{H}(\boldsymbol{\eta},\boldsymbol{\theta})\mathbf{A}\boldsymbol{\Sigma}^{-T}(s,t_0)\mathrm{d}s\cdot\boldsymbol{\Sigma}^T(t,t_0)\mathrm{d}t\\
&+ \boldsymbol{\Sigma}(t,t_0)\cdot\boldsymbol{\Sigma}^{-1}(t,t_0)\mathbf{A}^T\mathbf{H}(\boldsymbol{\eta},\boldsymbol{\theta})\mathbf{A}\boldsymbol{\Sigma}^{-T}(t,t_0)\cdot\boldsymbol{\Sigma}^T(t,t_0)\cdot\mathrm{d}t\\
&+ \boldsymbol{\Sigma}(t,t_0)\cdot\int_{t_0}^{t}\boldsymbol{\Sigma}^{-1}(s,t_0)\mathbf{A}^T\mathbf{H}(\boldsymbol{\eta},\boldsymbol{\theta})\mathbf{A}\boldsymbol{\Sigma}^{-T}(s,t_0)\mathrm{d}s\cdot\boldsymbol{\Sigma}^T(t,t_0)\mathbf{F}^T(\boldsymbol{\eta}(t),\boldsymbol{\theta})\cdot\mathrm{d}t\\
&= \left(\mathbf{F}(\boldsymbol{\eta}(t),\boldsymbol{\theta})\boldsymbol{\Phi}(t)+\boldsymbol{\Phi}(t)\mathbf{F}^T(\boldsymbol{\eta}(t),\boldsymbol{\theta})+\mathbf{A}^T\mathbf{H}(\boldsymbol{\eta}(t),\boldsymbol{\theta})\mathbf{A}\right)\mathrm{d}t,
\end{aligned}$$

which is the result in (2.37).

2.2.8 Relationship between LNA and other methods

The SDE approach can be viewed as a normal approximation based on a τ-leaping step for the MJP. The LNA can be derived either directly from Taylor expansion of the transition probability of the MJP or the Taylor expansion of the transition density of the SDE. The ODE solution can be considered as a limit of the mean trajectory of the MJP when system size N goes to infinity. ODE solution can also be viewed as the deterministic part for SDE (2.33) and the mean process for LNA based on (A.7). Figure 2.2.8 depicts relationships between different dynamical system representations as a diagram.

2.3 Coalescent Modeling

Phylodynamic studies start from n molecular sequences obtained from individuals sampled uniformly at random from the total population. In this thesis, we assume that a phylogenetic tree, or genealogy, $\mathbf{g}$ relating these sequences has been estimated in such a way that the tree branch lengths respect the known sequence sampling times. The genealogy is a tree graph showing the inferred ancestral relationship between individuals or other entities based on

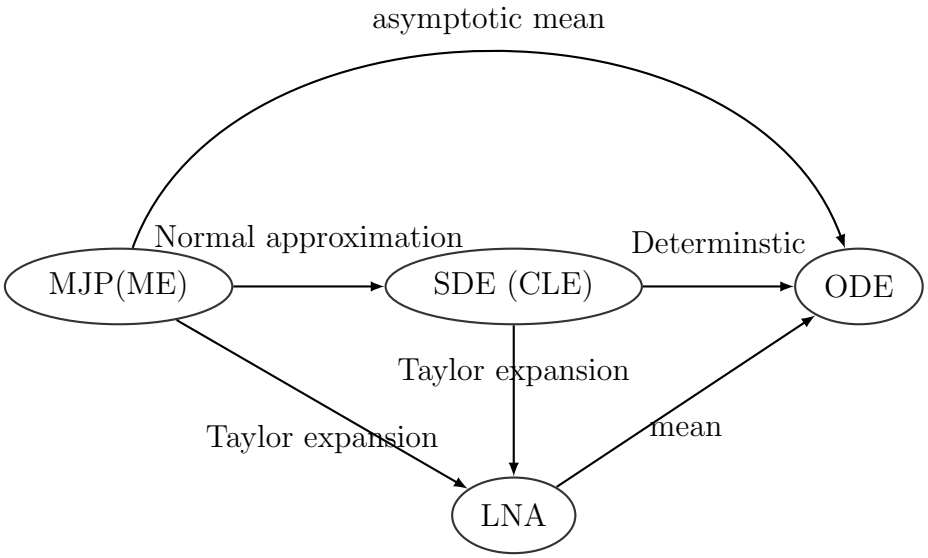

Figure 2.3: The relationship between different models

similarities and distinction in their genetic characteristics. The tip (leave) node of the tree correspond to a directly observed sampled member in the population. The internal nodes represent the estimated unobserved mutual ancestors, while the root node of the tree is the most recent common ancestor of the entire sample. Time along the tree follows chronological order from root to leaves/tips.

Sampling and coalescent times The nodes of genealogy contain two types of attributes: the coalescent times and the sampling times. The coalescent times correspond to the internal nodes of the tree, which are defined as the times at which two lineages in the tree are merged into a common ancestry. The sampling time is the time at which we sample one or multiple molecular sequences, which correspond to the tips of the tree. Since the sequences can be sampled at different times, for a genealogy constructed by n samples, we denote the sampling times by $s_1, \ldots, s_m$ and denote by n_i the number of sequences sampled at time s_i, with $\sum_{i=1}^{n} n_i = n$. We denote the $n-1$ coalescent times as $\mathcal{T} = \{c_1, \ldots, c_n\}$, with $c_1 < c_2 < \cdots < c_{n-1}$ respectively. The coalescent events happen in the reverse time order

from c_{n-1} to c_1. The sampling times are also ordered, so $s_1 < s_2 < \cdots < s_m$ and $s_m > c_{n-1}$, for notational convenience, we create a dummy coalescent time c_n to be the most recent sample time, i.e $c_n = s_m$.

Number lineages present By sorting the sampling times $s_1, \ldots, s_m$ along with the coalescent times $c_1, \ldots, c_{n-1}$, the interval $(c_{k-1}, c_k]$ is partitioned by the sampling events into i_k sub-intervals: $I_{0,k}, \ldots, I_{i_k-1,k}$. The interval that ends with a coalescent event is

$$I_{0,k} = (c_{k-1}, \min\{c_k, s_j\}], \text{for } s_j > c_{k-1} \text{ and } k = 2, \ldots, n.$$

Based on the partition, the number of lineages presents, denoted by $l_{i,k}$, remains constant in each interval $I_{i,k}$. If the interval $I_{i,k}$ ends with a coalescent time, the lineages in the next interval will be decreased by 1. If the interval ends with a sampling event s_i, then the lineages in the next interval is increased by n_i — the number of sequences sampled at time s_i. Figure 2.3 shows an example of a genealogy with labeled coalescent times, sampling times, number of lineages, and the corresponding intervals. The number of lineages present at time t, denoted by $l(t)$, is a piecewise constant function with the following form

$$l(t) = \sum_{k=2}^{n} \sum_{i=0}^{i_{k-1}} \mathbf{1}_{\{t \in I_{i,k}\}} l_{i,k}. \tag{2.53}$$

Coalescent rate Coalescent theory [Kingman, 1982] establishes a probabilistic model of the coalescent events that depends on the population dynamic. The coalescent events can be viewed as a realization of an inhomogeneous Markov death process, with time flowing backwards from the present time until all the samples coalesce into one common ancestor. Each death event corresponds to a coalescent event when two lineages merge. Let the rate of the coalescent process at time t, known as the coalescent rate, be $\lambda(t)$. Under general heterochronous sampling, the coalescent likelihood is

$$\Pr(\mathbf{g} \mid \lambda(\cdot)) \quad \propto \quad \prod_{k=2}^{n} \lambda(c_{k-1}) \exp\left(- \sum_{i=0}^{i_k - 1} \int_{I_{i,k}} \lambda(s) \mathrm{d}s \right). \tag{2.54}$$

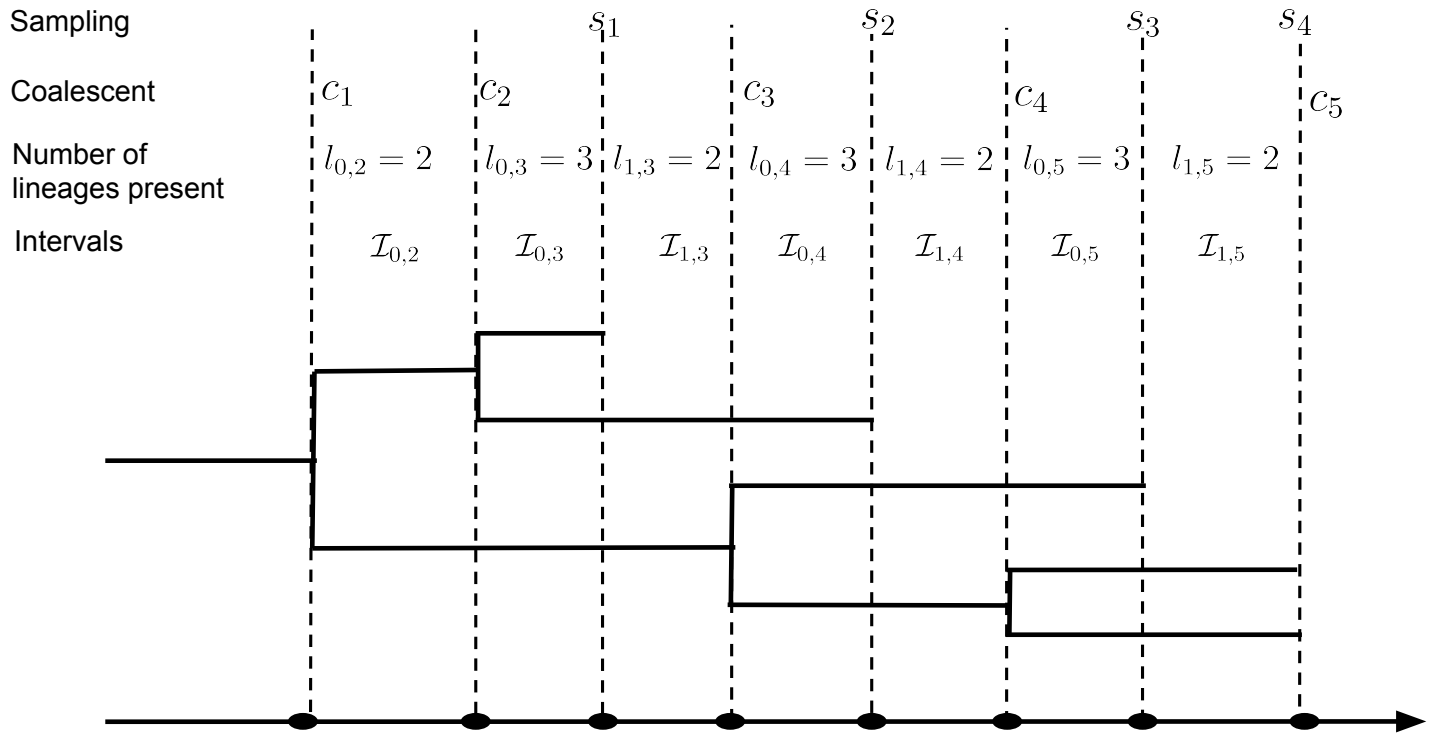

Figure 2.4: Example of a genealogy. Black solid lines show the genealogy structure. The coalescent times $c_1, \ldots, c_4$ and sampling times $s_1, \ldots, s_4$ are labeled with vertical dashed lines. The number of lineages $l_{i,k}$ is given in each intervals $\mathcal{I}_{i,k}$.

The coalescent rate $\lambda(t)$ not only depends on the number of present lineages, but also relates to the population size. The coalescent rate has different mathematical forms under different model assumptions. In Section 2.3.1, we will introduce the classic Kingman's coalescent model, which is widely used for nonparametric phylodynamic inference. Section 2.3.2 introduces a structured coalescent model for general compartmental epidemic models. The SIR structured coalescent model, as a special case with closed-form coalescent rate function, is explained in Section 2.3.3.

2.3.1 Kingman coalescent model

One of the key advances in early phylodynamics methodology is estimation of changes in the effective population size based on coalescent theory. The notion of effective population size, first introduced by [Wright, 1931], is defined as the number of breeding individuals in an idealized population that evolves according to a Wright-Fisher model. The effective population size assumes that there is no recombination, no natural selection, and no population structure. Early coalescent models are directly based on the Wright-Fisher model, assuming overlapping generations [Kingman, 1982]. Kingman [1982] showed that coalescent model can also be derived from a Moran model of genetic drift that assumes non-overlapping generation [Moran, 1958, 1962]. The Kingman's coalescent model was later generalized to accommodate time varying effective population sizes [Griffiths and Tavare, 1994, Donnelly and Tavare, 1995].

Here we give an intuitive derivation of Kingman coalescent model based on Wright-Fisher model. Starting from a population with constant effective population size N_e over generations, suppose we observe l lineages in the current generation. Each individual is assumed to have probability $\frac{1}{N_e}$ to choose an ancestor from the previous generation. The probability that none of the l samples in the current generation will have common ancestors in the previous generation is

$$\frac{N_e}{N_e} \cdot \left(\frac{N_e - 1}{N_e}\right) \cdots \left(\frac{N_e - l + 1}{N_e}\right) = 1 - \binom{l}{2}\frac{1}{N_e} + \mathcal{O}\left(\frac{1}{N_e^2}\right).$$

Let random variable U be the waiting time for l lineages to coalesce. For large population $N_e >> l$, the probability of no coalescent events over t generations will be

$$\Pr(U > t) = \left(1 - \binom{l}{2}\frac{1}{N_e} + \mathcal{O}\left(\frac{1}{N_e^2}\right)\right)^t \to \exp\left(-\binom{l}{2}\frac{t}{N_e}\right), \tag{2.55}$$

which gives the coalescent rate for Kingman's coalescent model to be $\binom{l}{2}/N_e$. The above derivation can be easily generalized to a time-varying population dynamic $N_e(t)$ with heterochronous/serial lineage sampling [Rodrigo and Felsenstein, 1999]. With $l(t)$ lineages present at time t, the coalescent rate for Kingman's coalescent model yields the following coalescent rate:

$$\lambda(t) = \binom{l(t)}{2}\frac{1}{N_e(t)}. \tag{2.56}$$

Together with the assumption that pairs of lineages coalesce uniformly at random, the Kingman coalescent model implies the following density of genealogy $\mathbf{g}$:

$$\Pr(\mathbf{g}|N_e(t)) \propto \prod_{k=2}^{n}\frac{l_{0,k}}{N_e(c_{k-1})}\exp\left(-\sum_{i=0}^{i_k-1}\int_{I_{i,k}}\frac{\binom{l_{i,k}}{2}}{N_e(s)}\mathrm{d}s\right). \tag{2.57}$$

In infectious disease modeling, somewhat heuristically, the effective population size $N_e(t)$ is associated with the number of infectious individuals $I(t)$. We will see how this interpretation can be made more precise below.

2.3.2 Structured coalescent model

The non-structured population assumption in the Wright-Fisher model is no longer valid for compartmental models of infectious disease spread. For example, the infection rate can be different among different age groups. Besides, an infected individual may go through multi-stages of infections with different strength to spread the virus. Hence, Kingman's coalescent model is no longer valid for compartmental models operating in a structured population. Volz [2012] propose a structured coalescent model that works in structured populations and has direct applications to infectious disease agent evolution during its spread through the population.

Birth and Migration matrices

Let $1, \dots d'$ denote the index of d' possible population groups ($d' \leq d$) that a lineage in the genealogy can belong to. To formalize the structured coalescent model, we first specify all possible types of events associated with coalescent (or split) events in the genealogy. We adopt Volz [2012]'s terminology and define such events as birth processes, denoted by a $d' \times d'$ matrix $\mathbf{B}(t) = \{b_{kr}(t)\}$, with $b_{kr}(t)$ being the rate at which lineages currently in state k give birth to lineages in state r. The birth rate b_{kr} corresponds to the rate of the reactions that ends with $\mathcal{X}_k + \mathcal{X}_r$. We also specify the events during which an individual changes from one infection state to another, defining a migration matrix $\mathbf{G}(t) = \{g_{kr}(t)\}$, as $g_{kr}(t)$ being the rate at which a lineage currently in state k migrate to state r, which is associated with the reaction $\mathcal{X}_k \to \mathcal{X}_r$.

Example: SEIR model Here we give an example of birth matrix and migration matrices for the structured coalescent model under SIER model. Since the removed individuals never interact without other population group, here we only consider the susceptible, exposed and infectious states/compartments, denoted by S, E and I respectively. The infection event associated with coalescent in the SEIR model is the reaction $S + I \xrightarrow{\beta} E + I$ with reaction rate βSI. Then birth matrix is

$$B(t) = \begin{array}{c} \begin{matrix} \text{Exposed} & \text{Infectious} \end{matrix} \\ \begin{pmatrix} 0 & 0 \\ \beta S(t) I(t) & 0 \end{pmatrix} \end{array} \begin{array}{l} \text{Exposed} \\ \text{Infectious.} \end{array} \tag{2.58}$$

The migration event corresponds to the event during which an exposed individual becomes infectious, i.e $E \xrightarrow{\mu} I$. Hence, the migration matrix in SEIR model is defined as

$$G(t) = \begin{array}{c} \begin{matrix} \text{Exposed} & \text{Infectious} \end{matrix} \\ \begin{pmatrix} 0 & \mu E(t) \\ 0 & 0 \end{pmatrix} \end{array} \begin{array}{l} \text{Exposed} \\ \text{Infectious.} \end{array} \tag{2.59}$$

Structured coalescent formulation

Under the notation in Section 2.2.2, let $\mathbf{X} = (X_1, \ldots, X_d)$ be the population vector. The coalescent rate for two lineages in states i and j, $\lambda_{ij}(t)$, at time t, can be formulated as

$$\lambda_{ij}(t) = \sum_{k}^{d'} \sum_{r}^{d'} \frac{b_{kr}(t)}{X_k(t)X_r(t)} \left(p_{ik}(t)p_{jr}(t) + p_{ir}(t)p_{jk}(t)\right), \tag{2.60}$$

where $p_{ik}(t)$ is the probability that the lineage i is in state k at time t. Then $\lambda(t)$ —- the coalescent rate at time t, is the summation of all possible coalescent rates among the present lineage pairs at time t:

$$\lambda(t) = \sum_{i,j \in \mathcal{A}_t, i \neq j} \lambda_{ij}(t), \tag{2.61}$$

where $\mathcal{A}_t$ denotes a set of active lineages that presents in the genealogy at time t. Unfortunately, most of the compartmental models do not have directly observed lineage state probabilities $p_{ij}(t)$s. The changes of lineage state probabilities can be tracked using a set of ODEs going backward through time. For lineage i and state k

$$\begin{aligned} \frac{\mathrm{d}p_{ik}(t)}{\mathrm{d}t} = \sum_{r}^{d'} \Bigg\{ & p_{ir}(t)\frac{g_{kr}(t)}{X_r(t)} - p_{ik}(t)\frac{g_{lk}(t)}{X_k(t)} + p_{ir}(t)\frac{b_{kr}(t)(X_k(t) - l_k(t))}{X_l(t)X_k(t)} \\ & - p_{ik}(t)\frac{b_{rk}(t)(X_r(t) - l_r(t))}{X_k(t)X_r(t)} \Bigg\}, \end{aligned} \tag{2.62}$$

where $l_k(t) = \sum_{i \in \mathcal{A}_t} p_{ik}(t)$, which is the expected number of lineages in state k in the genealogy at time t. The first two terms in the RHS of (2.62) show the probability mass gained or lost from lineages transitioning in or out of state k through a migration event. The second two terms give the probability change from lineage transitioning between states through a transmission event that was not observed as a coalescent event in the genealogy.

The lineage probability will get updated after a coalescent event. The computation of parent lineage probability $p_{hk}(t)$ will be updated by

$$p_{hk}(t) = \frac{1}{\lambda_{ij}(t)} \sum_{r}^{d'} \frac{b_{kr}(t)}{X_k(t)X_r(t)} \left(p_{ik}(t)p_{jr}(t) + p_{ir}(t)p_{jk}(t)\right). \tag{2.63}$$

Given (2.62) and (2.63), all the lineage probabilities can be specified. Most structured coalescent models do not have closed-form expression for coalescent rate given population compartment sizes $\mathbf{X}(t)$ and rate parameters $\boldsymbol{\theta}$. The computational cost for structured coalescent rates is quite high, because such computations involve solving m ODEs for each lineage and calculating the coalescent rate for each pair of present lineages.

2.3.3 SIR Structured Coalescent Model

A special case of structured coalescent is the SIR structured coalescent model [Volz et al., 2009]. In a SIR model, the only possible state of the lineage is the infected/infectious state (i.e $d' = 1$) and all lineage probabilities will simply reduce to $p_{i1}(t) = 1$ for $i \in \mathcal{A}_t$. The birth matrix in SIR model is $\mathbf{B}(t) = \{\beta S(t)I(t)\} \in \mathbb{R}^{1\times 1}$, which is the birth rate for new infection at time t, where $S(t)$ and $I(t)$ corresponds to the number of susceptibles and the number of infectious individuals at time t. Since SIR model only have one infected/infectious compartment and one infection stage, the migration matrix is $\mathbf{G}(t) = \{0\} \in \mathbb{R}^{1\times 1}$. For any $i, j \in \mathcal{A}_t$, the between lineage coalescent rate in (2.60) will be

$$\lambda_{ij} = \frac{\beta S(t)I(t)}{I^2(t)} = \frac{\beta S(t)}{I(t)}.$$

Hence the coalescent rate in structured coalescent model will be reduced to

$$\lambda(t) = \binom{l(t)}{2} \frac{2\beta S(t)}{I(t)}. \tag{2.64}$$

The SIR structured coalescent density of genealogy $\mathbf{g}$ given rate parameter $\boldsymbol{\theta}$ and population state vector $\mathbf{X}(t) = (S(t), I(t))$ is

$$\Pr(\mathbf{g} \mid \mathbf{X}_{0:T}, \boldsymbol{\theta}_{0:T}) \propto \prod_{k=2}^{n} \binom{l(c_{k-1})}{2} \frac{2\beta S(c_{k-1})}{I(c_{k-1})} \exp\left(- \sum_{i=0}^{i_k - 1} \int_{\mathcal{I}_{i,k}} \binom{l_{i,k}}{2} \frac{2\beta S(\tau)}{I(\tau)} \mathrm{d}\tau \right). \tag{2.65}$$

Another way to look at the SIR structured coalescent rate is to approximate rate (2.64) as $\lambda(t) \approx \frac{\binom{l(t)}{2}}{\binom{I(t)}{2}} \beta S(t)I(t)$. The fraction $\binom{l(t)}{2} / \binom{I(t)}{2}$ can be interpreted as the probability that the two selected coalescing lineages are both in the sample and $\beta S(t)I(t)$ is the birth rate

infection event in the population. The SIR structured coalescent model yields a closed-form expression of the coalescent rate and establishes a connection between the coalescent rate and the true population dynamics with respect to the numbers of infectious and susceptible individuals, $I(t)$ and $S(t)$. Furthermore, the SIR structured coalescent model can also be reduced the Kingman's coalescent likelihood by assuming a constant susceptible population over time, i.e $S_t \simeq N$, and constant infection rate β. Then the effective population size is written as $N_e(t) = I(t)/(2\beta N)$, which is proportional to the the number of infectious individuals $I(t)$, as was mentioned in Section 2.3.1. This means that under the SIR model with negligible depletion of the susceptible compartment, we can define $\tau = 2\beta N$ and written the coalescent rate as

$$\lambda(t) = \binom{l(t)}{2} \frac{1}{I(t)\tau},$$

or equivalently, write $N_e(t) = \tau I(t)$, justifying interpretation of $N_e(t)$ in the Kingman's coalescent as a quantity proportional to the true population prevalence, $I(t)$.

2.4 Bayesian Inference and Markov Chain Monte Carlo

2.4.1 Bayesian Inference

Let $\mathbf{Y}$ denote the observed data and $\boldsymbol{\theta}$ be the parameters governing the sampling distribution of $\mathbf{Y}$. Bayesian inference treats the unknow parameter $\boldsymbol{\theta}$ as random variable and requires us to use a prior distribution $\pi(\boldsymbol{\theta})$. The prior distribution encapsulates our prior knowledge of the model parameters without seeing the data. The objective of Bayesian inference is to quantify the uncertainty of our knowledge of the parameters $\boldsymbol{\theta}$ using the posterior distribution $\pi(\boldsymbol{\theta}|\mathbf{Y})$. The posterior distribution, based on Bayes rule, can be written as

$$\pi(\boldsymbol{\theta} \mid \mathbf{Y}) = \frac{\pi(\mathbf{Y} \mid \boldsymbol{\theta})\pi(\boldsymbol{\theta})}{\pi(\mathbf{Y})}, \tag{2.66}$$

where $\pi(\mathbf{Y} \mid \boldsymbol{\theta})$ is the likelihood function originating from the sampling distribution of $\mathbf{Y}$ and $\pi(\boldsymbol{\theta})$ is the prior distribution for $\boldsymbol{\theta}$. The marginal distribution of $\mathbf{Y}$ s constant with respect to $\boldsymbol{\theta}$ and can be calculated as follows: $\pi(\mathbf{Y}) = \int \pi(\mathbf{Y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}$.

Algorithm 2 Simulation of heterochoronus sampling coalescent under SIR structured coalescent model by thinning.

Input: Sampling time: $s_1 \leq s_2 \leq \cdots \leq s_m = T$ and correponding number of samples $n_1, \ldots, n_m$. Deterministic population trajectory $I(t), S(t)$ and infection rate $\beta(t)$. An envelope parameter $\lambda \geq 2\beta(t)S(t)/I(t)$.

Output Coalescent times $\mathcal{T} = \{c_1, \ldots, c_{n-1}\}$, where $n = \sum_{i=1}^{m} n_i$.

$i \leftarrow m$, $j \leftarrow n-1$, $t \leftarrow s_m$, $l \leftarrow n_m$

while $i \geq 0$ **do**

    Sample $\tau \sim \text{Exp}(\binom{l}{2}\lambda)$ and $U \sim \text{Unif}(0,1)$

    if $U \leq \dfrac{2\beta(t)S(t)}{\lambda I(t)}$ **then**

        if $t + \tau \leq s_{i-1}$ **then**

            $c_j \leftarrow t - \tau$, $t \leftarrow c_j$

            $j \leftarrow j - 1$, $l \leftarrow l - 1$

            if $n \geq 1$ **then**

                go to 4

            else

                go to 14

        else

            $i \leftarrow i - 1, t \leftarrow s_i, l \leftarrow l + n_i$

    else

        $t \leftarrow t - \tau$

Return: coalescent time $\mathcal{T} = \{c_1, \ldots, c_{n-1}\}$.

Different choices of prior usually leave to different posterior distributions. Based on the amount of information specified by the prior distribution, the priors can be loosely divided into two categories: informative prior and weakly informative prior. The informative prior contains definite information about the parameter. An informative prior usually makes the inference result less sensitive to the outliers in the observed data. However, the price for this robustness is diminished ability of the data to override incorrectly specified priors. In many cases, informative prior also plays a helpful role in resolving identifiability issues in the model itself, leading to more reasonable and interpretable posterior distributions. A weakly informative prior, in contrast to an informative prior, expresses less information and has more *a priori* uncertainty about model parameters. Using weakly informative prior makes Bayesian inference lean on the observed data more than on our prior knowledge, which is beneficial when such knowledge is incomplete. However, if the model suffers from identifiability issues, putting a weakly informative prior may leads to systematic biases.

Statistical models with latent variables benefit form additional notation to accommodate these variables. The observed data $\mathbf{Y}$ depends on the latent variable (process) $\mathbf{X}$ through likelihood function $\Pr(\mathbf{Y} \mid \mathbf{X}, \boldsymbol{\theta})$. The latent variable is governed by parameter $\boldsymbol{\theta}$ through $\Pr(\mathbf{X} \mid \boldsymbol{\theta})$. The marginal likelihood $\Pr(\mathbf{Y} \mid \boldsymbol{\theta})$, calculated via integral $\Pr(\mathbf{Y} \mid \boldsymbol{\theta}) = \int \Pr(\mathbf{Y} \mid \mathbf{X}, \boldsymbol{\theta}) \Pr(\mathbf{X} \mid \boldsymbol{\theta}) \mathrm{d}\mathbf{X}$, usually is not available in closed-form and often is computationally intractable. Besides, the latent variable $\mathbf{X}$ itself can be of interest. Hence, for models with latent variables, Bayesian inference often focuses on the joint posterior of $\mathbf{X}$ and $\boldsymbol{\theta}$ given $\mathbf{Y}$:

$$\Pr(\mathbf{X}, \boldsymbol{\theta} \mid \mathbf{Y}) \propto \Pr(\mathbf{Y} \mid \boldsymbol{\theta}, \mathbf{X}) \Pr(\mathbf{X} \mid \boldsymbol{\theta}) \pi(\boldsymbol{\theta}). \tag{2.67}$$

Usually neither $\Pr(\boldsymbol{\theta} \mid \mathbf{Y})$ nor $\Pr(\mathbf{X}, \boldsymbol{\theta} | \mathbf{Y})$ has a closed-form expression, so we have to rely on numerical approximations of the posterior distribution.

2.4.2 Markov chain Monte Carlo

Markov chain Monte Carlo (MCMC) is a framework for numerical integration that is especially useful for high dimensional integration. In Bayesian inference, MCMC is applied

to approximate posterior distributions $\pi(\boldsymbol{\theta}|\mathbf{Y})$ and obtain posterior summaries. The idea behind MCMC is to construct an ergodic discrete time Markov chain with some transition kernel $\mathcal{K}(\boldsymbol{\theta}, \boldsymbol{\theta}')$, with the stationary distribution equal to the posterior distribution of interest. The transition kernel in $\mathcal{K}(\boldsymbol{\theta}, \boldsymbol{\theta}')$ in MCMC gives the transition density from the current state $\boldsymbol{\theta}$ to the next state $\boldsymbol{\theta}'$ and needs to satisfy the global balance condition:

$$\pi(\boldsymbol{\theta} \mid \mathbf{Y}) = \int \mathcal{K}(\boldsymbol{\theta}', \boldsymbol{\theta}) \pi(\boldsymbol{\theta}' \mid \mathbf{Y}) \mathrm{d}\boldsymbol{\theta}'.$$

Generally it is hard to verify whether global balance condition holds. A sufficient condition for global balance equation is the detailed balance equation:

$$\pi(\boldsymbol{\theta} \mid \mathbf{Y}) \mathcal{K}(\boldsymbol{\theta}', \boldsymbol{\theta}) = \pi(\boldsymbol{\theta}' \mid \mathbf{Y}) \mathcal{K}(\boldsymbol{\theta}, \boldsymbol{\theta}').$$

The MCMC algorithm yields a sequences of correlated random samples $\boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(n)}$. Based on the ergodic theorem, the mean of a integrable function will converge almost surely to the target expectation:

$$\lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^{n} h(\boldsymbol{\theta}^{(i)}) \xrightarrow{a.s.} \mathbb{E}_{\boldsymbol{\theta}|\mathbf{Y}} \left[h(\boldsymbol{\theta}) \right]. \tag{2.68}$$

2.4.3 Metropolis-Hastings Algorithm

The Metropolis-Hastings (MH) algorithm [Metropolis et al., 1953, Hastings, 1970] is a recipe for constructing Markov chains described above. The MH algorithm is widely used in Bayesian inference when directly sampling from posterior distribution $\pi(\boldsymbol{\theta} \mid \mathbf{Y})$ is difficult. The MH algorithm requires a proposal density $q(\boldsymbol{\theta}_1 \mid \boldsymbol{\theta}_2)$. In each step a MH algorithm, given the currently state $\boldsymbol{\theta}$, a new state $\boldsymbol{\theta}'$ is proposed by a random draw from the distribution with density $q(\cdot \mid \boldsymbol{\theta})$. The newly proposed value $\boldsymbol{\theta}'$ is accepted with probability

$$\begin{aligned} a(\boldsymbol{\theta}, \boldsymbol{\theta}') &= \min \left\{ 1, \frac{\pi\left(\boldsymbol{\theta}' \mid \mathbf{Y}\right) q(\boldsymbol{\theta} \mid \boldsymbol{\theta}')}{\pi\left(\boldsymbol{\theta} \mid \mathbf{Y}\right) q(\boldsymbol{\theta}' \mid \boldsymbol{\theta})} \right\} \\ &= \min \left\{ 1, \frac{\pi\left(\mathbf{Y} \mid \boldsymbol{\theta}'\right) \pi(\boldsymbol{\theta}') q(\boldsymbol{\theta} \mid \boldsymbol{\theta}')}{\pi\left(\mathbf{Y} \mid \boldsymbol{\theta}\right) \pi(\boldsymbol{\theta}) q(\boldsymbol{\theta}' \mid \boldsymbol{\theta})} \right\}. \end{aligned}$$

The ratio $r(\boldsymbol{\theta}, \boldsymbol{\theta}') = \dfrac{\pi(\boldsymbol{\theta}' \mid \mathbf{Y})\, q(\boldsymbol{\theta} \mid \boldsymbol{\theta}')}{\pi(\boldsymbol{\theta} \mid \mathbf{Y})\, q(\boldsymbol{\theta}' \mid \boldsymbol{\theta})}$ is called MH ratio. The Metropolis-Hasting algorithm defines a transition kernel with transition density $K(\boldsymbol{\theta}, \boldsymbol{\theta}') = q(\boldsymbol{\theta}' \mid \boldsymbol{\theta}) a(\boldsymbol{\theta}, \boldsymbol{\theta}')$. It can be shown that the detailed balance equation is satisfied:

$$\pi(\boldsymbol{\theta}|\mathbf{Y}) q(\boldsymbol{\theta}' \mid \boldsymbol{\theta}) a(\boldsymbol{\theta}, \boldsymbol{\theta}') = \pi(\boldsymbol{\theta}'|\mathbf{Y}) q(\boldsymbol{\theta} \mid \boldsymbol{\theta}') a(\boldsymbol{\theta}', \boldsymbol{\theta}),$$

and the stationary distribution of the resulting Markov chain is $\pi(\boldsymbol{\theta} \mid \mathbf{Y})$. Pseudo code for the MH algorithm is given in Algorithm 3.

A special case for MH update, with $q(\theta_1|\theta_2) = q(\theta_2|\theta_1)$, is called a Metropolis algorithm. The acceptance rate in such case will reduce to $\min\left\{1, \dfrac{\pi(\boldsymbol{\theta}'|\mathbf{Y})}{\pi(\boldsymbol{\theta}|\mathbf{Y})}\right\}$, which depends on the ration of posterior likelihood ratio of posterior densities. One way to achieve this symmetric proposal density property is by using the random walk proposal, i.e

$$\boldsymbol{\theta}' = \boldsymbol{\theta} + \boldsymbol{\epsilon},$$

where $\boldsymbol{\epsilon}$ is symmetrically distributed around $\mathbf{0}$ and stochastically independent of $\boldsymbol{\theta}$. One of the most commonly used random walk proposal is the normal distribution with mean $\mathbf{0}$.

The acceptance probability in the MH algorithm depends not only on the posterior density function $\pi(\boldsymbol{\theta} \mid \mathbf{Y})$, but also on the proposal density $q(\cdot \mid \cdot)$. A proposal with potentially large jumps can explore the parameter space more quickly, but draws from such a proposal are more likely to be rejected. A proposal density with small jumps often leads to higher acceptance probability, but leads to samples with high autocorrelation and slow exploration of the posterior.

2.4.4 Gibbs Sampler

A Gibbs sampler is a MCMC method that obtains samples from a multivariate distribution based on iteratively sampling each parameter. Suppose $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_d)$ is vector and the target posterior distribution is $\pi(\boldsymbol{\theta} \mid \mathbf{Y})$. A Gibbs sampler consists steps of sampling θ_j, for $j = 1, \ldots, d$, from conditional distribution $\pi(\theta_j \mid \mathbf{Y}, \boldsymbol{\theta}_{-j})$. Gibbs sampler is applicable when the conditional distribution $\pi(\theta_i \mid \mathbf{Y}, \boldsymbol{\theta}_{-j})$ can be directly sampled from. Gibbs sampling can

Algorithm 3 Metropolis-Hastings algorithm for posterior distribution $\pi(\cdot \mid \mathbf{Y})$

1: **Input:** Parameter from the previous iteration $\boldsymbol{\theta}$. Proposal density $q(\cdot|\cdot)$. Observed data $\mathbf{Y}$.

2: **Output** Updated parameters values

3: Proposal new value $\boldsymbol{\theta}' \sim q\left(\boldsymbol{\theta}' \mid \boldsymbol{\theta}\right)$.

4: Compute acceptance probability:

$$a(\boldsymbol{\theta}, \boldsymbol{\theta}') = \min\left\{1, \frac{\pi\left(\mathbf{Y} \mid \boldsymbol{\theta}'\right)\pi(\boldsymbol{\theta}')q(\boldsymbol{\theta} \mid \boldsymbol{\theta}')}{\pi\left(\mathbf{Y} \mid \boldsymbol{\theta}\right)\pi(\boldsymbol{\theta})q(\boldsymbol{\theta}' \mid \boldsymbol{\theta})}\right\}.$$

5: Sample $u \sim \text{Unif}(0, 1)$

6: Accept/ reject proposal

$$\boldsymbol{\theta} = \begin{cases} \boldsymbol{\theta}' & u < a(\boldsymbol{\theta}, \boldsymbol{\theta}') \\ \boldsymbol{\theta} & \text{Otherwise.} \end{cases}$$

be generalized to update $\boldsymbol{\theta}$ blockwisely, which is called blocked Gibbs sampler. By splitting the multivariable $\boldsymbol{\theta}$ into blocks $J_1, \ldots, J_p$, the blocked Gibbs sampler update parameters within block J_i jointly conditioning on a the most recent value in other blocks. Gibbs sampler pseudo code is given in Algorithm 4.

The Gibbs sampler can be considered as a special case of the MH algorithm, with proposal density q satisfying

$$q(\theta_j' \mid \theta_j) = \pi(\theta_j' \mid \mathbf{Y}, \boldsymbol{\theta}_{-j}) \propto \pi(\theta_j', \boldsymbol{\theta}_{-j} \mid \mathbf{Y}). \tag{2.69}$$

The acceptance rate for Gibbs sampler is 1 since

$$a(\boldsymbol{\theta}', \theta) = \min\left\{1, \frac{\pi(\mathbf{Y} \mid \theta_j', \boldsymbol{\theta}_{-j})q(\theta_j \mid \theta_j')}{\pi\left(\mathbf{Y} \mid \theta_j, \boldsymbol{\theta}_{-j}\right)q(\theta_j' \mid \theta_j)}\right\} = 1 \tag{2.70}$$

The idea of Gibbs sampling can be extended to a general framework that incorporates with MH algorithm and other sophisticated sampling frameworks.

Algorithm 4 Gibbs sampler for posterior distribution $\pi(\cdot \mid \mathbf{Y})$

1: **Input:** Parameter from the previous iteration $\boldsymbol{\theta} \in \mathbb{R}^d$. Observed data $\mathbf{Y}$.

2: **Output** Updated parameters values

3: **for** $i = 1, \ldots, p$: **do**

4: Sample $\boldsymbol{\theta}'_{J_i} | \mathbf{Y}, \boldsymbol{\theta}'_{J_1}, \ldots, \boldsymbol{\theta}'_{J_{i-1}}, \boldsymbol{\theta}_{J_{i+1}}, \ldots, \boldsymbol{\theta}_{J_p} \sim \pi\left(\boldsymbol{\theta}'_{J_i} | \mathbf{Y}, \boldsymbol{\theta}'_{J_1}, \ldots, \boldsymbol{\theta}'_{J_{i-1}}, \boldsymbol{\theta}_{J_{i+1}}, \ldots, \boldsymbol{\theta}_{J_p}\right)$

5: Return $\boldsymbol{\theta} = (\theta'_1, \ldots, \theta'_d)$

2.4.5 Elliptical Slice Sampler

An elliptical slice sampler is proposed by Murray et al. [2010] for targeting posterior distributions resulting from models with latent Gaussian random field. The latent variable is assumed to be a random vector $\mathbf{X} \in \mathbb{R}^d$ following zero-mean Gaussian distribution with covariance $\Sigma(\boldsymbol{\theta})$ as prior distribution, i.e $\mathbf{X} \sim \mathcal{N}(\mathbf{0}, \Sigma)$. We use $\Pr(\mathbf{Y}|\mathbf{X}, \boldsymbol{\theta})$ to denote the likelihood function for observed data $\mathbf{Y}$ given latent variable $\mathbf{X}$ and parameter vector $\boldsymbol{\theta}$. Hence, the target posterior distribution for $\mathbf{X}$ given is

$$\Pr(\mathbf{X}|\mathbf{Y}, \boldsymbol{\theta}) \propto \Pr(\mathbf{Y}|\mathbf{X}, \boldsymbol{\theta}) \mathcal{N}(\mathbf{X}|\mathbf{0}, \Sigma(\boldsymbol{\theta})) \pi(\boldsymbol{\theta}),$$

where $\pi(\boldsymbol{\theta})$ is the prior distribution for parameter $\boldsymbol{\theta}$. The goal of elliptical slice sampler is to obtain posterior samples of latent variable $\mathbf{X}$ from $\mathbf{X}|\mathbf{Y}, \boldsymbol{\theta}$. The proposing step in elliptical sampling consists of two parts: (1) proposing an auxiliary random vector $\mathbf{Z} \in \mathbb{R}^d$ from distribution $\mathcal{N}(\mathbf{0}, \Sigma)$, (2) proposing a variable $\alpha \in [0, 2\pi]$ as an angle parameter. In elliptical slice sampler, a new state $(\mathbf{X}', \mathbf{Z}')$ is proposed by rotating the previous state $(\mathbf{X}, \mathbf{Z})$ with angle α,

$$\mathbf{X}' = \mathbf{X} \cos(\alpha) + \mathbf{X} \sin(\alpha) \tag{2.71}$$

$$\mathbf{Z}' = \mathbf{Z} \sin(\alpha) - \mathbf{Z} \cos(\alpha). \tag{2.72}$$

For any given α, this transition leaves the joint prior probability invariant, i.e,

$$\mathcal{N}(\mathbf{X}|\mathbf{0}, \Sigma) \mathcal{N}(\mathbf{Z}|\mathbf{0}, \Sigma) = \mathcal{N}(\mathbf{X}'|\mathbf{0}, \Sigma) \mathcal{N}(\mathbf{Z}'|\mathbf{0}, \Sigma).$$

Hence, $(\mathbf{X}', \mathbf{Z}')$ are considered as the proposed state and the ratio and the propose transition probability from $(\mathbf{X}, \mathbf{Z})$ to $(\mathbf{X}', \mathbf{Z}')$ should equal that from $(\mathbf{X}', \mathbf{Z}')$ to $(\mathbf{X}, \mathbf{Z})$, i.e

$$\frac{\Pr\left((\mathbf{X}', \mathbf{Z}') \to (\mathbf{X}, \mathbf{Z})\right)}{\Pr\left((\mathbf{X}, \mathbf{Z}) \to (\mathbf{X}', \mathbf{Z}')\right)} = 1.$$

Then such idea of proposing new state is plugged into a slice sampler framework [Neal, 2003], where we firstly proposing a random variable U from $\text{Uniform}(0, \Pr(Y|\mathbf{X}, \boldsymbol{\theta}))$ and then find the updated variable $\mathbf{X}'$ by exploring a elliptical path on the surface $\left\{\mathbf{X}' \in \mathbb{R}^d \mid \Pr(Y|\mathbf{X}', \boldsymbol{\theta}) > U\right\}$.

Pseudo code for elliptical slice sampler is given in Algorithm 5. Notice that iterations stops only when a new sample is accepted. Hence, the elliptical slice sampler has acceptance ratio 1, meaning that it will always update the latent random field $\mathbf{X}$ is the sampling step.

Algorithm 5 Elliptical slice sampler for posterior distribution $\pi(\cdot \mid \mathbf{Y}, \boldsymbol{\theta})$

Input: Latent variable from the previous iteration $\mathbf{X} \in \mathbb{R}^d$. Observed data $\mathbf{Y}$, previous updated parameter $\mathbf{X}$.

Output Updated Latent variable value $\mathbf{X}'$

Sample ellipse $\mathbf{Z} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}(\boldsymbol{\theta}))$

Compute log-likelihood threshold: sample $U \sim \text{Uniform}(0, 1)$ and let

$$\tau \leftarrow \log \Pr(\mathbf{Y}|\mathbf{X}, \boldsymbol{\theta}) + \log(U)$$

Sample angle parameter $\alpha \sim \text{Uniform}[0, 2\pi]$ and $[\alpha_{\min}, \alpha_{\max}] \leftarrow [\alpha - 2\pi, \alpha]$

$\mathbf{X}' \leftarrow \mathbf{X} \cdot \cos\alpha + \mathbf{Z} \cdot \sin\alpha$

while $\log(\Pr(\mathbf{Y}|\mathbf{X}', \boldsymbol{\theta})) < \tau$ **do**

    if $\alpha < 0$ **then**

        $\alpha_{\min} \leftarrow \alpha$

    else

        $\alpha_{\max} \leftarrow \alpha$

    Sample $\alpha \sim \text{Uniform}\left(\alpha_{\min}, \alpha_{\max}\right)$.

    Make new proposal:

$$\mathbf{X}' \leftarrow \mathbf{X} \cdot \cos\alpha + \mathbf{Z} \cdot \sin\alpha$$

Return $\mathbf{X}'$.

Chapter 3

FITTING STOCHASTIC EPIDEMIC MODEL TO GENE GENEALOGY

3.1 Introduction

Phylodynamics is an area at the intersection of phylogenetics and population genetics that studies how epidemiological, immunological, and evolutionary processes affect viral phylogenies constructed based on molecular sequences sampled from the population of interest [Grenfell et al., 2004, Volz et al., 2013]. Phylodynamics is especially useful in infectious disease modeling because genetic data provide a source of information that is complimentary to the traditional disease case count data. Here we are interested in inferring parameters governing infectious disease dynamics from the genealogy/phylogeny estimated from infectious disease agent molecular sequences collected during the disease outbreak. Working in a Bayesian framework, we develop an efficient Markov chain Monte Carlo (MCMC) algorithm that allows us to work with stochastic models of infectious disease dynamics, properly accounting for stochastic nature of the dynamics.

Currently, learning about population-level infectious disease dynamics from molecular sequences can be accomplished using three general strategies. The first strategy relies on the coalescent theory — a set of population genetics tools that specify probability models for genealogies relating individuals randomly sampled from the population of interest [Kingman, 1982, Griffiths and Tavare, 1994, Donnelly and Tavare, 1995]. Using a subset of these models [Griffiths and Tavaré, 1994], it is possible to estimate changes in effective population size — the number of breeding individuals in an idealized population that evolves according to a Wright-Fisher model [Wright, 1931]. Such reconstruction can be done assuming parametric [Kuhner et al., 1998, Drummond et al., 2002] or nonparametric [Drummond et al., 2002, 2005,

Minin et al., 2008, Palacios and Minin, 2013, Gill et al., 2013] functional forms of the effective population size trajectory. In the context of infectious disease phylodynamics, nonparametric inference is the norm and the estimated effective population size is often interpreted as the effective number of infections or the effective number of infectious individuals. However, reconstructed effective population size trajectories are not easy to interpret and estimation of parameters of disease dynamics is difficult to accomplish if one wishes to maintain statistical rigor [Pybus et al., 2001, Frost and Volz, 2010].

Another way to learn about infectious disease dynamics from molecular sequences is to model explicitly events that occur during the infectious disease spread and to link these events to the genealogy/phylogeny of sampled individuals using birth-death processes. For example, a Susceptible-Infectious-Removed (SIR) model includes two possible events: infections and removals (e.g., recoveries and deaths), represented by births and deaths in the corresponding birth-death model [Stadler et al., 2013, Kühnert et al., 2014]. Other SIR-like models (e.g., SI and SIS models) differ by the number and types of the events that are needed to accurately describe natural history of the infectious disease [Leventhal et al., 2013]. Although these methods are more principled than post-hoc processing of nonparametrically estimated disease dynamics, they are not easy to scale to large datasets and/or high dimensional models. For example, in order to fit phylodynamic birth-death models to genomic and epidemiological data Vaughan et al. [2018] use particle filter MCMC. However, computational burden of particle filter MCMC methods is usually very high. Moreover, these methods often struggle with convergence when the dimensionality of statistical model parameters is even moderately high [Andrieu et al., 2010].

Structured coalescent models provide the third strategy of inferring parameters governing spread of an infectious disease [Volz et al., 2009, Volz, 2012, Dearlove and Wilson, 2013]. These models assume infectious disease agent genetic data have been obtained from a random sample of infected individuals, allowing for serial sampling over time. Although similar to the birth-death modeling framework, the structured coalescent models have two advantages. First, one does not have to keep track, analytically or computationally, of extinct and not

sampled genetic lineages. Second, the density of the genealogy can be obtained given the population level information about status of individuals: for example, in the SIR model it is sufficient to know the numbers of susceptible, $(S(t))$, infectious, $(I(t))$, and recovered, $(R(t))$, individuals at each time point t. The second advantage comes with two caveats: 1) such densities can be obtained only approximately and 2) evaluating densities of genealogies is not straightforward and involves numerical solutions of differential equations. Even in cases when these caveats are manageable, the density of the assumed stochastic epidemic model population trajectory remains computationally intractable. One way around this intractability assumes a deterministic model of infectious disease dynamics [Volz et al., 2009, Volz, 2012, Volz and Pond, 2014], which potentially leads to overconfidence in estimation of model parameters. Particle filter MCMC offers another solution [Rasmussen et al., 2011, 2014], but, as we discussed already, these methods are difficult to use in practice, especially in high dimensional parameter spaces.

In this chapter, we develop methods that allow us to bypass computationally unwieldy particle filter MCMC with the help of a linear noise approximation (LNA). LNA is a low order correction of the deterministic ordinary differential equation describing the asymptotic mean trajectories of compartmental models of population dynamics defined as Markov jump processes (e.g., chemical reaction models and SIR-like models of infectious disease dynamics) [Kurtz, 1970, 1971, Van Kampen and Reinhardt, 1983]. LNA can also be viewed as a first order Taylor approximation of Markov population dynamics models represented by stochastic differential equations [Giagos, 2010, Wallace, 2010]. A key feature of the LNA method is that it approximates the transition density of a stochastic population model with a Gaussian density [Komorowski et al., 2009].

Inspired by recent applications of LNA to analysis of Google Flu Trends data [Fearnhead et al., 2014] and disease case counts [Buckingham-Jeffery et al., 2018], we develop a Bayesian framework that combines LNA for stochastic models of infectious disease dynamics with structured coalescent models for genealogies of infectious disease agent genetic samples. Our approach yields a latent Gaussian Markov model that closely resembles a Gaussian state-

space model. We use this resemblance to develop an efficient MCMC algorithm that combines high dimensional elliptical slice sampler updates [Murray et al., 2010] with low dimensional Metropolis-Hastings (MH) moves. Using simulations, we demonstrate that this algorithm can handle reasonably complex models, including an SIR model with a time-varying infection rate. We apply this SIR model to a recent Ebola outbreak in West Africa. Our analysis of data from Liberia and Sierra Leone illuminates significant changes in the Ebola infection rate over time, likely caused by the public health response measures and increased awareness of the outbreak in the population.

3.2 Methodology

3.2.1 Genealogy as data

We start with n infectious disease agent molecular sequences obtained from infected individuals sampled uniformly at random from the total infected population. Further, we assume that a phylogenetic tree, or genealogy, $\mathbf{g}$ relating these sequences has been estimated in such a way that the tree branch lengths respect the known sequence sampling times. Such estimation can be performed with, for example, `BEAST` — a leading software package for Bayesian phylogenetic studies, particularly popular among molecular epidemiologists who collect and analyze viral genetic sequences [Suchard et al., 2018]. The genealogy is represented by a tree structure with its nodes containing two sources of temporal information: coalescent and sampling times. The coalescent times correspond to the internal nodes of the tree, which are defined as the times at which two lineages in the tree are merged into a common ancestor. The sampling times, corresponding to the tips of the tree, are the times at which molecular sequences were sampled. Note that sampling times are observed directly, while coalescent times are estimated from molecular sequences during phylogenetic reconstruction.

To perform inference about infectious disease dynamics using the above genealogy we need a probability model that relates the genealogy and infectious disease dynamics model parameters.

Figure 3.1: SIR Markov jump process. From the current state with the counts S, I, R, the population can transition to state $S-1, I+1, R$ (an infection event) with rate $\beta(t)SI$ or to state $S, I-1, R+1$ (a removal event) with rate $\gamma(t)I$. No other instantaneous transitions are allowed.

Without too much loss of generality, we assume that the infectious disease is spreading through the population according to the SIR model — a compartmental model that at each time point t tracks the number of susceptible individuals $S(t)$, number of infected/infectious individuals $I(t)$, and number of removed individuals $R(t)$ [Bailey, 1975, Anderson and May, 1992]. We assume that the population is closed so $S(t) + I(t) + R(t) = N$ for all times t, where N is the population size that we assume to be known. This constraint implies that vector $\mathbf{X}(t) = (S(t), I(t))$ is sufficient to keep track of the population state at time t. We follow common practice and model $\mathbf{X}(t)$ as a Markov jump process (MJP) with allowable instantaneous jumps shown in Figure 3.1 [O'Neill and Roberts, 1999]. The assumed MJP process $\mathbf{X}(t)$ is inhomogeneous, because we allow the infection rate $\beta(t)$ and removal rate $\gamma(t)$ to be time-varying.

The structured coalescent models assume that only coalescent times $c_1 < c_2 < \cdots < c_{n-1}$ provide information about the population dynamics. These times are modeled as jumps of an inhomogeneous pure death process with rate $\lambda(t)$, where each "death" event corresponds to coalescence of two lineages and $\lambda(t)$ is called a coalescent rate. Then the density of the genealogy, which serves as a likelihood in our work, is written as

$$\Pr(\mathbf{g}) \propto \prod_{k=2}^{n} \lambda(c_{k-1}) \exp\left(- \int_{c_{k-1}}^{c_k} \lambda(\tau) \mathrm{d}\tau \right),$$

where c_n denotes the most recent sequence sampling time. The dependence of coalescent rate on the assumed population dynamics can be complicated and mathematically intractable,

but luckily approximations exist for some specific cases. For the SIR model the approximate coalescent rate can be obtained via the following formula:

$$\lambda(t) = \lambda(l(t), \beta(t), \mathbf{X}(t)) = \binom{l(t)}{2} \frac{2\beta(t)S(t)}{I(t)}, \tag{3.1}$$

where $l(t)$ is the number of lineages present at time t. Note that when the number of susceptibles is not changing significantly relative to the total population size (i.e., $S(t) \approx N$) and infection rate is constant (i.e., $\beta(t) = \beta$), the structured coalescent reduces to the classical Kingman's coalescent, where we interpret $I(t)/(2\beta N)$ as the effective population size trajectory [Kingman, 1982]. It is possible to find approximate coalescence rate for general compartmental models, but closed-form expressions exist only for a few models with a low number of compartments (e.g., SI, SIR) [Volz et al., 2009, Volz, 2012, Dearlove and Wilson, 2013].

Since we allow sequences to be sampled at different times $s_1 < s_2 < \cdots < s_m = c_n$, some inter-coalescent times are censored. To deal with this censoring algebraically, each inter-coalescent interval $[c_{k-1}, c_k)$ is partitioned by the sampling events into i_k sub-intervals: $\mathcal{I}_{0,k}, \ldots, \mathcal{I}_{i_k-1,k}$. The intervals that end with a coalescent event are defined as $\mathcal{I}_{0,k} = [c_{k-1}, \min\{c_k, s_j\})$, for $s_j > c_{k-1}$ and $k = 2, \ldots, n$. Let the number of lineages in each interval $\mathcal{I}_{i,k}$ be $l_{i,k}$. Then the number of lineages at each time point t can be written as $l(t) = \sum_{k=2}^{n} \sum_{i=0}^{i_{k-1}} 1_{\{t \in I_{i,k}\}} l_{i,k}$.

If the interval $\mathcal{I}_{i,k}$ ends with a coalescent time, the number of lineages in the next interval will be decreased by 1. If the interval ends with a sampling event s_i, then the number of lineages in the next interval is increased by n_i — the number of sequences sampled at time s_i. Figure 3.2.1 shows an example of a genealogy with labeled coalescent times, sampling times, number of lineages, and the corresponding intervals.

We are now ready to connect the SIR model and a genealogy with serially sampled tips with the help of a structured coalescent density/likelihood. First we discretize the time interval between the time to most recent common ancestor c_1 (time corresponding to the root of the tree) and the most recent sampling time s_m using a regular grid $t_0 < t_1 < \cdots < t_T$ ($t_0 <$

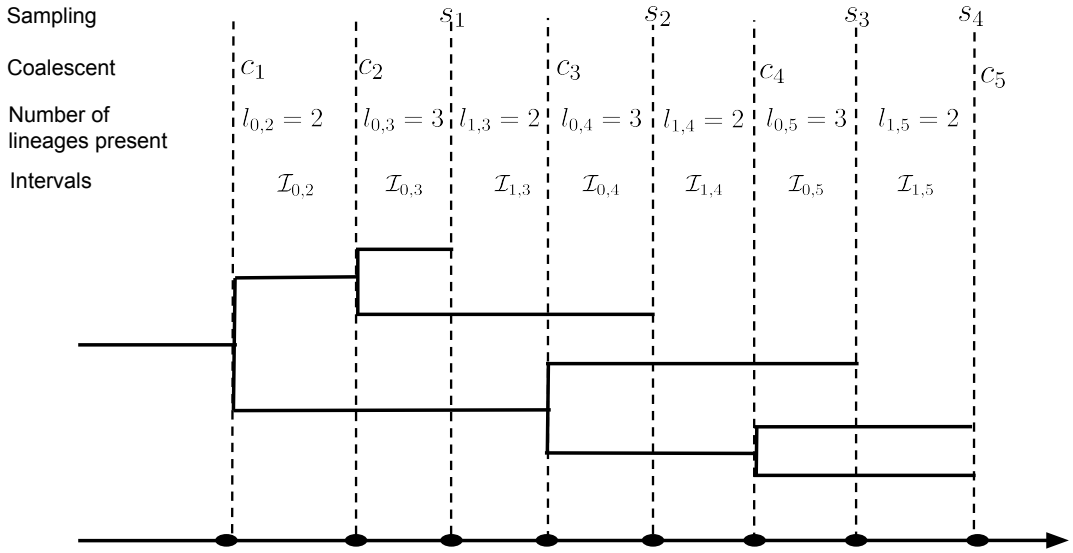

Figure 3.2: Example of a genealogy. Black solid lines show the genealogy structure. The coalescent times $c_1, \ldots, c_4$ and sampling times $s_1, \ldots, s_4$ are labeled with vertical dashed lines. The number of lineages $l_{i,k}$ is given in each intervals $\mathcal{I}_{i,k}$.

c_1 and $t_T > s_m$). Using this grid, we discretize the latent epidemic trajectory by assuming that $\mathbf{X}(t) = \sum_{j=1}^{T} \mathbf{X}_{j-1} \mathbf{1}_{[t_{j-1}, t_j)}(t)$, where $\mathbf{X}_j = (S_j, I_j)$ is a column vector. Similarly, we discretize the infectious disease dynamics parameter vector trajectory $\boldsymbol{\theta}(t) = (\beta(t), \gamma(t))$ so that $\boldsymbol{\theta}(t) = \sum_{j=1}^{T} \boldsymbol{\theta}_{j-1} \mathbf{1}_{[t_{j-1}, t_j)}(t)$, where $\boldsymbol{\theta}_j = (\beta_j, \gamma_j)$ is also a column vector. We collect latent variables $\mathbf{X}_j$s and parameters $\boldsymbol{\theta}_j$s into matrices $\mathbf{X}_{0:T}$ and $\boldsymbol{\theta}_{0:T}$ respectively. The SIR structured coalescent density/likelihood then becomes

$$\Pr(\mathbf{g} \mid \mathbf{X}_{0:T}, \boldsymbol{\theta}_{0:T}) \propto \prod_{k=2}^{n} \binom{l(c_{k-1})}{2} \frac{2\beta(c_{k-1})S(c_{k-1})}{I(c_{k-1})} \exp\left(- \sum_{i=0}^{i_k - 1} \int_{\mathcal{I}_{i,k}} \binom{l_{i,k}}{2} \frac{2\beta(\tau)S(\tau)}{I(\tau)} \mathrm{d}\tau \right) \quad (3.2)$$

Since $S(t)$, $I(t)$, and $\beta(t)$ are piecewise constant functions, the integrals in the above formula are readily available in closed-form and are fast to compute.

3.2.2 Bayesian data augmentation

Posterior distribution

Given genealogy $\mathbf{g}$, our goal is to infer the latent SIR population dynamic $\mathbf{X}_{0:T}$ and rate parameters $\boldsymbol{\theta}_{0:T}$ over time grid $t_0 < t_1 < \cdots < t_T$. Let $\Pr(\mathbf{X}_0)$ and $\Pr(\boldsymbol{\theta}_{0:T})$ denote the prior densities for the initial compartment states and the SIR parameters respectively. The posterior distribution for the population trajectory $\mathbf{X}_{0:T}$ and parameters $\boldsymbol{\theta}_{0:T}$ given observed genealogy $\mathbf{g}$ is

$$\Pr\left(\mathbf{X}_{0:T}, \boldsymbol{\theta}_{0:T} \mid \mathbf{g}\right) \propto \Pr\left(\mathbf{g} \mid \mathbf{X}_{0:T}, \boldsymbol{\theta}_{0:T}\right) \Pr\left(\mathbf{X}_{1:T} \mid \mathbf{X}_0, \boldsymbol{\theta}_{0:T}\right) \Pr\left(\boldsymbol{\theta}_{0:T}\right) \Pr\left(\mathbf{X}_0\right), \tag{3.3}$$

where $\Pr\left(\mathbf{g} \mid \mathbf{X}_{0:T}, \boldsymbol{\theta}_{0:T}\right)$ is the structured coalescent likelihood introduced in Section 3.2.1 and $\Pr(\mathbf{X}_{1:T} \mid \mathbf{X}_0, \boldsymbol{\theta}_{0:T})$ is the likelihood function for discrete observations of trajectory $\mathbf{X}_{1:T}$ given the initial value $\mathbf{X}_0$:

$$\Pr(\mathbf{X}_{1:T} \mid \mathbf{X}_0, \boldsymbol{\theta}_{0:T}) = \prod_{i=1}^{T} \Pr(\mathbf{X}_i \mid \mathbf{X}_{i-1}, \boldsymbol{\theta}_{i-1}), \tag{3.4}$$

where the factorization comes from the assumed Markov property of the disease dynamics. However, the SIR transition density $\Pr(\mathbf{X}_i \mid \mathbf{X}_{i-1}, \boldsymbol{\theta}_{i-1})$ becomes intractable as population size N grows large, making it difficult to perform likelihood-based inference for outbreaks in large populations.

Linear noise approximation

To furnish a feasible computation strategy for large populations, we use a linear noise approximation (LNA) method, in which the computationally intractable transition probability $\Pr\left(\mathbf{X}_i \mid \mathbf{X}_{i-1}, \boldsymbol{\theta}_{i-1}\right)$ is approximated using a closed-form Gaussian transition density.

The LNA method replaces the MJP discrete state space with a continuous state space of $\mathbf{X}(t)$ to approximate the counts of at time t, under the following constraints: $S(t) > 0, I(t) > 0$ and $S(t) + I(t) \leq N$. To briefly explain how this approximation is obtained, we will need additional notation.

The SIR MJP instantaneous transitions, depicted in Figure 3.1, are encoded in an effect matrix

$$\mathbf{A} = \begin{array}{c} \begin{array}{cc} \text{susceptible} & \text{infected} \end{array} \\ \begin{pmatrix} -1 & 1 \\ 0 & -1 \end{pmatrix} \end{array} \begin{array}{l} \text{infection} \\ \text{removal.} \end{array} \tag{3.5}$$

Each row in matrix (3.5) represents a type of transition event and each column corresponds to a change in the susceptible and infected populations. Next, we define a rate vector $\mathbf{h}$ and a rate matrix $\mathbf{H}$:

$$\mathbf{h}(\mathbf{X}(t), \boldsymbol{\theta}(t)) = \begin{pmatrix} \beta(t)S(t)I(t) \\ \gamma(t)I(t) \end{pmatrix}, \mathbf{H} = \text{diag}\left(\mathbf{h}(\mathbf{X}(t), \boldsymbol{\theta}(t))\right) = \begin{pmatrix} \beta(t)S(t)I(t) & 0 \\ 0 & \gamma(t)I(t) \end{pmatrix} \tag{3.6}$$

The above notation, as well as subsequent developments based on it, can be generalized to other epidemic models and, more generally, to a large class of density dependent stochastic processes, such as chemical reaction and gene regulation models [Wilkinson, 2011]. See Section 2.2.2 for more details on this generalization.

Consider a transition from $\mathbf{X}_{i-1}$ at time t_{i-1} to $\mathbf{X}_i$ at t_i. Recall that we assume that the SIR rates $\boldsymbol{\theta}(t)$ take constant values $\boldsymbol{\theta}_{i-1}$ in $[t_{i-1}, t_i)$. The LNA represents the value of the next state $\mathbf{X}_i$ as $\mathbf{X}_i = \boldsymbol{\eta}(t_i) + \mathbf{M}(t_i)$, where $\boldsymbol{\eta}(t_i)$ is a deterministic component and $\mathbf{M}(t_i)$ is a stochastic component. The deterministic component $\boldsymbol{\eta}(t_i)$ can be obtained by solving the standard SIR ODE that in our notation can be written as

$$\mathrm{d}\boldsymbol{\eta}(t) = \mathbf{A}^T \mathbf{h}(\boldsymbol{\eta}(t), \boldsymbol{\theta}_{i-1})\mathrm{d}t, \quad t \in [t_{i-1}, t_i]. \tag{3.7}$$

The stochastic part $\mathbf{M}(t_i)$ corresponds to the solution of the following SDE at time t_i:

$$\mathrm{d}\mathbf{M}(t) = \mathbf{F}(\boldsymbol{\eta}(t), \boldsymbol{\theta}_{i-1})\mathbf{M}(t)\mathrm{d}t + \sqrt{\mathbf{A}^T \mathbf{H}(\boldsymbol{\eta}(t), \boldsymbol{\theta}_{i-1})\mathbf{A}}\mathrm{d}\mathbf{W}_t, \quad t \in [t_{i-1}, t_i], \tag{3.8}$$

where $\mathbf{F}(\boldsymbol{\eta}(t), \boldsymbol{\theta}_{i-1}) := \dfrac{\partial \mathbf{A}^T \mathbf{h}(\mathbf{X}(t), \boldsymbol{\theta}_{i-1})}{\partial \mathbf{X}}\Big|_{\mathbf{X}=\boldsymbol{\eta}(t)}$ is the Jacobian matrix of the deterministic part $\mathbf{A}^T \mathbf{h}(\mathbf{X}(t), \boldsymbol{\theta}_{i-1})$ in (3.7) evaluated at $\boldsymbol{\eta}(t)$. The solution of SDE (3.8), $\mathbf{M}(t)$, is a Gaussian process and can be recovered by solving two ordinary differential equations governing

the mean function $\mathbf{m}(t) := \mathbf{E}[\mathbf{M}(t)]$ and covariance function $\mathbf{\Phi}(t) := \mathbf{Var}(\mathbf{M}(t))$:

$$\begin{aligned} \mathrm{d}\mathbf{m}(t) &= \mathbf{F}(\boldsymbol{\eta}(t), \boldsymbol{\theta}_{i-1})\mathbf{m}(t)\mathrm{d}t, & (3.9) \\ \mathrm{d}\mathbf{\Phi}(t) &= \left(\mathbf{F}\left(\boldsymbol{\eta}(t), \boldsymbol{\theta}_{i-1}\right)\mathbf{\Phi}(t) + \mathbf{\Phi}(t)\mathbf{F}^T(\boldsymbol{\eta}(t), \boldsymbol{\theta}_{i-1}) + \mathbf{A}^T\mathbf{H}\left(\boldsymbol{\eta}(t), \boldsymbol{\theta}_{i-1}\right)\mathbf{A}\right)\mathrm{d}t, & (3.10) \end{aligned}$$

for $t \in [t_{i-1}, t_i]$. A heuristic derivation of LNA, based on Wallace [2010], is given in Section 2.2.7. Let $\boldsymbol{\eta}_{t_{i-1}}, \mathbf{m}_{t_{i-1}}, \mathbf{\Phi}_{t_{i-1}}$ denote the initial values of $\boldsymbol{\eta}(t), \mathbf{m}(t), \mathbf{\Phi}(t)$ at time t_{i-1} in differential equations (3.7), (3.9), and (3.10) respectively. There are two options for choosing these initial conditions: the non-restarting LNA of Komorowski et al. [2009] and the restarting LNA of Fearnhead et al. [2014]. In this chapter, we will use the non-restarting LNA by Komorowski et al. [2009] with the following choice of initial conditions:

1. $\boldsymbol{\eta}_{t_{i-1}} = \boldsymbol{\eta}(t_{i-1})$, where $\boldsymbol{\eta}(t_{i-1})$ was obtained by solving the ODE (3.7) using parameter $\boldsymbol{\theta}_{i-2}$ over the interval $[t_{i-2}, t_{i-1}]$,

2. $\mathbf{m}_{t_{i-1}} = \mathbf{X}_{i-1} - \boldsymbol{\eta}(t_{i-1})$,

3. $\mathbf{\Phi}_{t_{i-1}} = \mathbf{0}$.

Solving the system of ODEs (3.7), (3.9), (3.10), we obtain $\boldsymbol{\eta}(t_i)$, $\mathbf{m}(t_i)$, and $\mathbf{\Phi}(t_i)$. The solution $\mathbf{m}(t_i)$ will be a function of the initial value $\mathbf{X}_{i-1} - \boldsymbol{\eta}(t_{i-1})$, the interval length $\Delta t_i := t_i - t_{i-1}$ and the SIR rates $\boldsymbol{\theta}_{i-1}$. To make this dependence explicit, we write $\mathbf{m}(t_i) := \boldsymbol{\mu}\left(\mathbf{X}_{i-1} - \boldsymbol{\eta}(t_{i-1}), \Delta t_i, \boldsymbol{\theta}_{i-1}\right)$. Since (3.9) is a first order homogeneous linear ODE, the solution $\boldsymbol{\mu}\left(\mathbf{X}_{i-1} - \boldsymbol{\eta}(t_{i-1}), \Delta t_i, \boldsymbol{\theta}_{i-1}\right)$ is a linear function of $\mathbf{X}_{i-1} - \boldsymbol{\eta}(t_{i-1})$. Hence, the transition from $\mathbf{X}_{i-1}$ to $\mathbf{X}_i$ follows the following Gaussian distribution:

$$\mathbf{X}_i \mid \mathbf{X}_{i-1}, \boldsymbol{\theta}_{i-1} \sim \mathcal{N}\left(\boldsymbol{\eta}(t_i) + \boldsymbol{\mu}\left(\mathbf{X}_{i-1} - \boldsymbol{\eta}(t_{i-1}), \Delta t_i, \boldsymbol{\theta}_{i-1}\right), \mathbf{\Phi}(t_i)\right). \tag{3.11}$$

To summarize, the derived conditional Gaussian densities $\Pr(\mathbf{X}_i \mid \mathbf{X}_{i-1}, \boldsymbol{\theta}_{i-1})$ allow us to compute the density of the latent SIR trajectory (3.4). As a result, our augmented posterior distribution of $\mathbf{X}_{0:T}$ and $\boldsymbol{\theta}_{0:T}$, shown in equation (3.3), can be computed up to proportionality constant and approximated via "standard" (not particle filter) MCMC approaches.

3.2.3 Reparameterization, priors, and MCMC algorithm

Reparameterizing SIR rates

We have experimented with multiple parameterizations of our inhomogeneous SIR model and found that the following parameterization works best with our proposed MCMC algorithm for approximating the posterior distribution (3.3). First, recall that we allow SIR rates to vary with time. Since it is much more likely for the infection rate to be time variable, we are going to assume a constant removal/removal rate γ. This leaves us with the following parameters: infection rates on a grid $\boldsymbol{\beta}$, removal rate γ, and initial SIR state $\mathbf{X}_0 = (S_0, I_0)$. Since we are interested in modeling an emerging infectious disease outbreak, we set the initial counts of susceptible to $S_0 = N - I_0$. Initial counts of infected individuals, I_0, is assumed to be low and treated as an unknown parameter with a lognormal prior distribution. Instead of the time-varying infection rate $\beta(t)$, we parameterize our SIR model with a time-varying basic reproduction number $R_0(t) = [\beta(t)N]/\gamma$. The reproduction number is interpreted as the average number of cases that one case generates over its infection period in a completely susceptible population. Since our infection rate changes in a piecewise manner, the basic reproduction number varies over time in a piecewise manner too:

$$R_0(t) = \sum_{i=1}^{T} R_{0_{i-1}} \mathbf{1}_{[t_{i-1}, t_i)}(t), \tag{3.12}$$

where $R_{0_i} = [\beta_i N]/\gamma$ is the reproduction number corresponding to the time interval $[t_{i-1}, t_i)$. Let $R_0 = R_{0_0}$ be the initial basic reproductive number and $\delta_i = \log\left(R_{0_i}/R_{0_{i-1}}\right)/\sigma$ be a normalized log ratio of $R_0(t)$ over two successive time intervals. Then, interval-specific basic reproduction numbers can be written as

$$R_{0_i} = R_0(t, \boldsymbol{\delta}_{1:T}, \sigma) = R_0 \exp\left(\sum_{k=1}^{i} \sigma\delta_k\right), \text{for } i = 1, \ldots, T, \tag{3.13}$$

where we assume *a priori* that δ_is are independent standard normal random variables.

This construction implies that log-transformed piecewise constant reproduction numbers, $\log(R_{0_i})$s, *a priori* follow a first order Gaussian Markov random field (GMRF) with

standard deviation σ that controls the *a priori* smoothness of $R_0(t)$ trajectory [Rue, 2001, Rue and Held, 2005]. In addition to speeding MCMC convergence, working with $R_0(t)$ is convenient, because this trajectory is dimensionless and retains its interpretation when one changes the population size N. The initial R_0 is assigned a lognormal(a_1, b_1) prior. We use a lognormal(a_2, b_2) prior for the inverse of standard deviation $1/\sigma$.

Reparameterizing SIR latent trajectories

We reparameterize the latent SIR trajectory $\mathbf{X}_{1:T}$ with a sequence of independent Gaussian random variables $\boldsymbol{\xi}_{1:T}$, following a non-centered parameterization framework of Papaspiliopoulos et al. [2007]. According to formula (3.11), conditional on $\mathbf{X}_{i-1}$, $\mathbf{X}_i$ can be written as

$$\mathbf{X}_i = \boldsymbol{\eta}(t_i) + \boldsymbol{\mu}(\mathbf{X}_{i-1} - \boldsymbol{\eta}(t_{i-1}), \Delta t_i, \boldsymbol{\theta}_{i-1}) + \boldsymbol{\Phi}_i^{1/2}\boldsymbol{\xi}_i, \tag{3.14}$$

where $\boldsymbol{\xi}_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \mathbf{I})$ for $i = 1, \ldots, T$ and $\mathbf{I}$ is a 2×2 identity matrix. In our parameterization, we will treat $\boldsymbol{\xi}_{1:T}$ as random latent variables and the SIR latent trajectory $\mathbf{X}_{1:T}$ as a deterministic transformation of $\boldsymbol{\xi}_{1:T}$. More details about our non-centered parameterization of $\mathbf{X}_{1:T}$ can be found in Section A.1 of the Appendix.

MCMC algorithm

Using our new parameterization, we are now interested in the posterior distribution of the initial number of infected individuals, I_0, removal rate, γ, the initial basic reproduction number, R_0, standardized vectors, $\boldsymbol{\delta}_{1:T}$ and $\boldsymbol{\xi}_{1:T}$, and GMRF standard deviation, σ:

$$\begin{aligned}\Pr(I_0, R_0, \gamma, \boldsymbol{\delta}_{1:T}, \boldsymbol{\xi}_{1:T}, \sigma|\mathbf{g}) &\propto \Pr(\mathbf{g}|I_0, R_0, \gamma, \boldsymbol{\delta}_{1:T}, \boldsymbol{\xi}_{1:T}, \sigma)\Pr(I_0)\Pr(R_0)\Pr(\gamma)\Pr(\boldsymbol{\delta}_{1:T})\Pr(\boldsymbol{\xi}_{1:T})\Pr(\sigma)\\ &\propto \Pr(\mathbf{g}|\mathbf{X}_{0:T}, \boldsymbol{\theta}_{0:T})\Pr(I_0)\Pr(R_0)\Pr(\gamma)\Pr(\boldsymbol{\delta}_{1:T})\Pr(\boldsymbol{\xi}_{1:T})\Pr(\sigma).\end{aligned}$$

The latent variables $\mathbf{X}_{0:T}$ and parameter vector $\boldsymbol{\theta}_{0:T}$ are deterministic functions of new parameters I_0, γ, R_0, $\boldsymbol{\delta}_{1:T}$, $\boldsymbol{\xi}_{1:T}$, and σ. We use the following MCMC with block updates to approximate this posterior distribution. We update high dimensional vector $\mathbf{U} =$

$(\log(R_0), \boldsymbol{\delta}_{1:T}, \log(\sigma))$ using the efficient elliptical slice sampler [Murray et al., 2010]. Vector $\boldsymbol{\xi}_{1:T}$ is updated the same way in a separate step. Initial number of infected individuals I_0 and removal rate γ are updated using univariate Metropolis steps. The full procedure is described in Algorithm 6, which together with details of the elliptical slice sampler can be found in Section 2.4.5. After MCMC is done, we report posterior summaries using natural parameterization. For example, we report posterior medians and 95% Bayesian credible intervals (BCIs) of the piecewise latent reproduction number trajectory, R_{0_i} for $i = 0, \ldots, T$, and latent trajectory $\mathbf{X}_{0:T}$.

Implementation

Our R package called `LNAPhylodyn` provides an implementation of our MCMC algorithm. The package code is publicly available at `https://github.com/MingweiWilliamTang/LNAphyloDyn`. This repository also contains scripts that should allow one to reproduce key numerical results in this manuscript.

3.3 Simulation experiments

3.3.1 Simulations based on single genealogy realizations

In this section, we use simulated genealogies to assess performance of our LNA-based method and to compare it with an ODE-based method, where we replace equation (3.14) with its simplified version: $\mathbf{X}_i = \boldsymbol{\eta}(t_i)$. Under our assumption of a fixed and known genealogy and constant R_0, our ODE-based method closely resembles previously developed methods by Volz et al. [2009] and Volz and Siveroni [2018]. To compare ODE-based and LNA-based models in a Bayesian nonparametric setting, we equip the ODE model with the GMRF prior for time-varying $R_0(t)$, described in Section 3.2.3. We use the same MCMC algorithm for both LNA-based and ODE-based models, except we do not have a separate step to update latent vector $\boldsymbol{\xi}_{1:T}$ (equivalently, $\mathbf{X}_{0:T}$) in the ODE-based inference. See Algorithm 7 in the Appendix for a more detailed description of the ODE-based MCMC.

Algorithm 6 Updating rule in the LNA-based MCMC algorithm

1: **Input:** Parameter values from the previous interation $I_0, R_0, \gamma, \boldsymbol{\delta}_{1:T}, \sigma, \boldsymbol{\xi}_{1:T}$, genealogy $\mathbf{g}$. Proposal density $q_1(\cdot|\cdot)$, $q_2(\cdot|\cdot)$ for updating the initial number of infected individuals and the removal rate.

2: **Output** Updated parameters values

3: Calculate $\mathbf{X}_{0:T}$, $\boldsymbol{\theta}_{0:T}$ based on $I_0, R_0, \gamma, \boldsymbol{\delta}_{1:T}, \sigma, \boldsymbol{\xi}_{1:T}$.

4: Propose I_0' based on $q_1(\cdot|I_0)$, then $\mathbf{X}_{0:T}$ will be deterministically updated to $\mathbf{X}_{0:T}'$ according to I_0', $R_0, \gamma, \boldsymbol{\delta}_{1:T}, \sigma, \boldsymbol{\xi}_{1:T}$.

5: Accept $(I_0', \mathbf{X}_{0:T}')$ with acceptance probability

$$a \leftarrow \min\left(1, \frac{\Pr(\mathbf{g}|\boldsymbol{\theta}_{0:T}', \mathbf{X}_{0:T}')\Pr(I_0')q_1(I_0|I_0')}{\Pr(\mathbf{g}|\boldsymbol{\theta}_{0:T}, \mathbf{X}_{0:T})\Pr(I_0)q_1(I_0'|I_0)}\right).$$

6: Propose γ' based on $q_2(\cdot|\gamma)$, then $\mathbf{X}_{0:T}, \boldsymbol{\theta}_{0:T}$ will be deterministically updated to $\mathbf{X}_{0:T}', \boldsymbol{\theta}_{0:T}'$ according to I_0, $R_0, \gamma', \boldsymbol{\delta}_{1:T}, \sigma, \boldsymbol{\xi}_{1:T}$.

7: Accept $(\gamma', \mathbf{X}_{0:T}', \boldsymbol{\theta}_{0:T}')$ with acceptance probability

$$a \leftarrow \min\left(1, \frac{\Pr(\mathbf{g}|\boldsymbol{\theta}_{0:T}', \mathbf{X}_{0:T}')\Pr(\gamma')q_2(\gamma|\gamma')}{\Pr(\mathbf{g}|\boldsymbol{\theta}_{0:T}, \mathbf{X}_{0:T})\Pr(\gamma)q_2(\gamma'|\gamma)}\right).$$

8: Let $\mathbf{U} = (\log(R_0), \boldsymbol{\delta}_{1:T}, \log(\sigma))$, then $\mathbf{U}$ *a priori* follows a multivariate normal distribution. Use elliptical slice sampler to obtain $\mathbf{U}'$ and get the updated $R_0', \boldsymbol{\delta}_{1:T}'$ and σ'. $\mathbf{X}_{0:T}$ will be deterministically updated to $\mathbf{X}_{0:T}'$ according to $I_0, R_0', \gamma, \boldsymbol{\delta}_{1:T}', \sigma'$.

9: Since $\boldsymbol{\xi}_{1:T}$ *a priori* follows a multivariate normal distribution, we use the elliptical slice sampler to obtain $\boldsymbol{\xi}_{1:T}'$. $\mathbf{X}_{0:T}$ will be deterministically updated to $\mathbf{X}_{0:T}'$ according to $I_0, R_0, \gamma, \boldsymbol{\delta}_{1:T}, \sigma, \boldsymbol{\xi}_{1:T}'$.

The simulation protocol consists of two steps. First, given the population size N and pre-specified parameters γ, I_0, and $R_0(t)$, we simulate one realization of the SIR population trajectory based on the MJP using the Gillespie algorithm [Gillespie, 1977]. Next, we generate realistic lineage sampling times and simulate coalescent times from the distribution specified by density (3.2) using a thinning algorithm by Palacios and Minin [2013].

We test LNA-based and ODE-based methods under three "true" $R_0(t)$ trajectories over the time interval $[0, 90]$:

1. Constant (CONST) $R_0(t)$. $R_0(t) = 2.2$ for $t \in [0, 90]$. Recovery rate $\gamma = 0.2$. Initial counts of infected individuals $I_0 = 1$. Total population size is $N = 100{,}000$.

2. Stepwise decreasing (SD) $R_0(t)$. $R_0(t) = 2, t \in [0, 30)$, $R_0(t) = 1, t \in [30, 60)$ and $R_0(t) = 0.6, t \in [60, 90]$. Recovery rate $\gamma = 0.2$. Initial counts of infected individuals $I_0 = 1$. Total population size $N = 1{,}000{,}000$.

3. Non-monotonic (NM) $R_0(t)$. $R_0(t) = 1.4 \times 1.015^{0.5t}, t \in [0, 30]$, $R_0(t) = 1.750 \times 0.975^{t-30}, t \in [30, 80]$ and $R_0(t) = 0.4583, t \in [80, 90]$. Recovery rate $\gamma = 0.3$. Initial counts of infected individuals $I_0 = 3$. Total population size $N = 1{,}000{,}000$.

For all simulations, we use lognormal$(1, 1)$ prior for I_0. The parameters of the lognormal priors for the initial R_0 and inverse standard deviation $1/\sigma$ are set to $a_1 = 0.7, b_1 = 0.5$ and $a_2 = 3, b_2 = 0.2$ respectively, in such a way that *a priori* $R_0(t)$ trajectory stayed within a reasonable range of $[0, 5]$ with 0.9 probability. We assign an informative prior for γ in each simulation scenario, because prior information about this parameter is usually available: (1) CONST: $\gamma \sim$ lognormal$(-1.7, 0.1)$, (2) SD: $\gamma \sim$ lognormal$(-1.7, 0.1)$, (3) NM: $\gamma \sim$ lognormal$(-1.2, 0.1)$. We set the grid size to $T = 36$, with $t_i - t_{i-1} = 2.5$ for $i = 1, \ldots, 36$. For both LNA-based and ODE-based methods, we use 300,000 MCMC iterations. All MCMC chains appeared to converge (trace plots are shown in Section B.1.3 of the Appendix). The effective sample sizes of all unknown quantities were above 100.

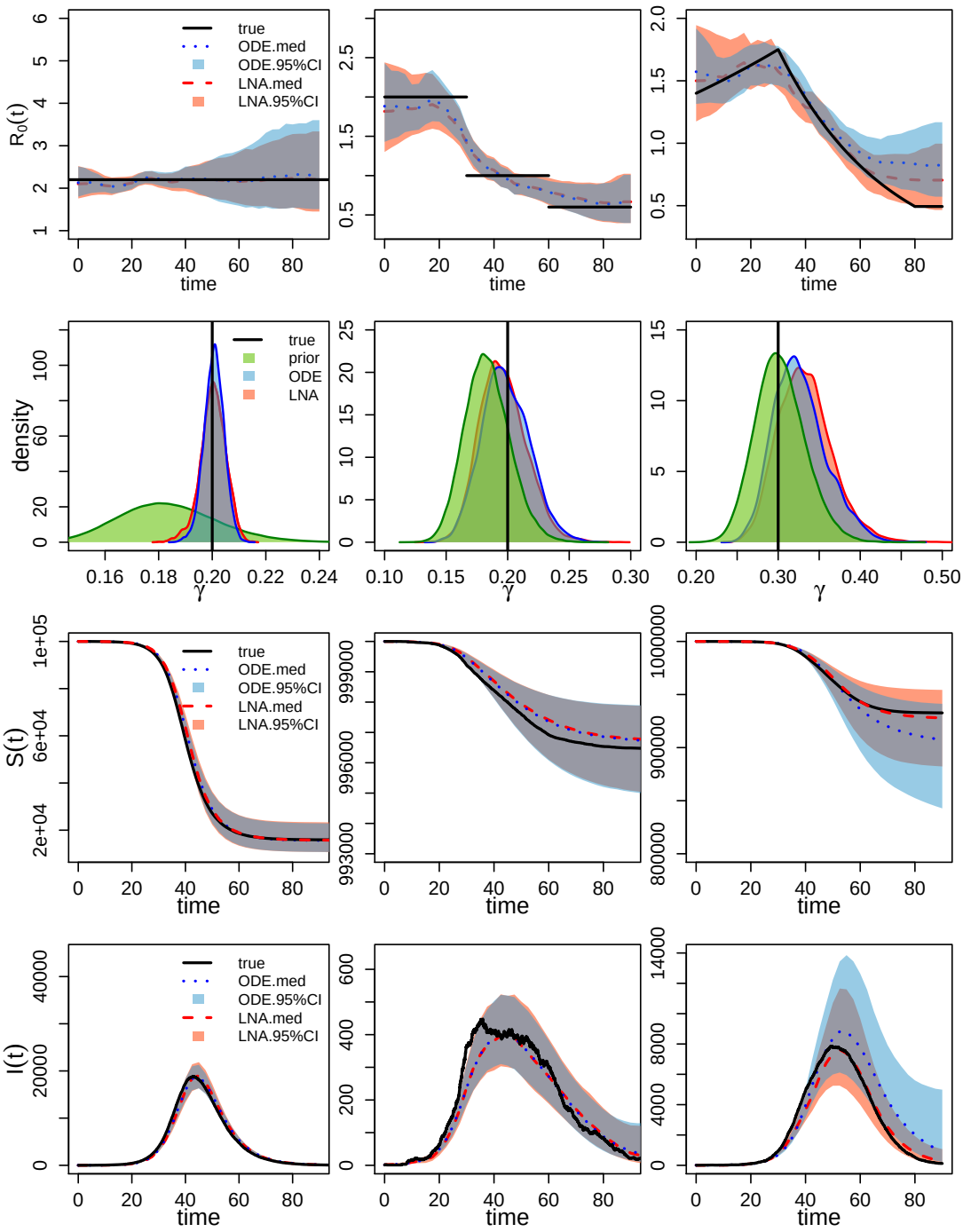

Figure 3.3: Analysis of 3 simulation scenarios. Columns correspond to CONST, The first row shows the estimated $R_0(t)$ trajectories for the 3 scenarios, with the black solid lines representing the truth, the red dashed lines depicting the posterior median and the red-shaded area showing the 95% BCIs for the LNA-based method. For the ODE-based method, the posterior median is plotted in blue dotted lines, with blue shading showing the 95% BCIs. The second row corresponds to the estimation for the removal rate γ. Posterior density curves from the LNA are shown in red lines and the posterior density for ODE is plotted in blue lines, compared with prior density curve in green lines. The bottom two figures shows the estimated trajectory of $S(t)$ and $I(t)$ respectively.

The first row of Figure 3.3 shows point-wise posterior medians and 95% BCIs for the basic reproduction number trajectory, $R_0(t)$. Our LNA-based method performs well in capturing the continuous dynamics of $R_0(t)$. Though our approach may not perfectly catch the discontinuous changes in R_0 in the SD scenario, the method provides BCIs that are able to capture most of the $R_0(t)$ trajectory. The ODE-based method yields similar results in the CONST case and the SD case, but fails to capture the decreasing trends in the NM scenario.

The second row in Figure 3.3 shows posterior summaries of removal rate γ. Both LNA-based and ODE-based methods provide good estimates in the CONST scenario, with posterior modes centered at the true value and higher posterior densities at truth when compared with the prior. In the SD and NM scenarios with the time varying $R_0(t)$, the posterior estimates from the LNA-based method and ODE-based method, though still centered at the truth, do not differ much from the prior distribution.

Posterior summaries of $S(t)$ and $I(t)$ are depicted in the third and fourth rows of Figure 3.3. The two methods produce similar results in the CONST and SD scenario, as both of them have narrow BCIs covering the true trajectories. However, in the NM case, while the LNA-based method manages to recover the latent SIR trajectory trend, the BCIs from the ODE-based method fail to cover the true prevalence trajectory in the middle and at the end of the epidemic. Somewhat counterintuitively, LNA-based method produces BCIs for the latent trajectories, $S(t)$ and $I(t)$, that are narrower than its ODE counterparts. We suspect this is a result of the ODE-based method poor estimation of the basic reproduction number trajectory at the end of the epidemic.

3.3.2 Frequentist properties of posterior summaries

In this Section, we design a simulation study based on repeatedly simulating SIR trajectories using MJP with pre-specified parameters. The simulations are based on the non-monotonic $R_0(t)$ trajectory scenario in Section 3.3.1 with the same parameter setup, except the parameters of the lognormal prior for the initial R_0 are set to $a_1 = 0.7, b_1 = 0.3$. Simulating SIR dynamics under low initial number of infected individuals I_0 can end up with low prevalence

trajectories that end at the beginning of the epidemic, or trajectories having unrealistically high prevalence, which are less likely to be observed during real infectious disease outbreaks. Therefore, while simulating SIR trajectories we reject such "unreasonable" realizations to arrive at 100 simulated trajectories. The details of the rejection criteria are given in Section B.1.2 of the Appendix. For each simulated SIR trajectory, a realization of a genealogy is generated using the structured coalescent process. We use both LNA-based and ODE-based model to approximate the posterior distribution of model parameters and latent variables for each genealogy.

We use three metrics to evaluate models based on their estimates of $R_0(t)$ and $I(t)$: average error of point estimates (posterior medians), width of credible intervals, and frequentist coverage of credible intervals. Since the value of $R_0(t)$ is greater than 0 and usually upper-bounded by 20 (i.e, it stays within the same order of magnitude), we will measure accuracy using an unnormalized mean absolute error (MAE):

$$\text{MAE} = \frac{1}{T+1}\sum_{i=0}^{T} |\hat{R}_{0_i} - R_0(t_i)|, \tag{3.15}$$

where $\hat{R}_{0_i}$ is the posterior median of $R_0(t_i)$. In contrast, $I(t)$ varies from one at the beginning of the epidemic to thousands at the peak, so to evaluate accuracy of prevalence estimation we use the mean relative absolute error (MRAE):

$$\text{MRAE} = \frac{1}{T+1}\sum_{i=0}^{T} \frac{|\hat{I}_i - I(t_i)|}{I(t_i)+1}, \tag{3.16}$$

where $\hat{I}_i$ is the posterior median of $\text{I}(t_i)$. We assess precision of $R_0(t)$ estimation based on the mean credible interval width (MCIW):

$$\text{MCIW} = \frac{1}{T+1}\sum_{i=0}^{T} \left[\hat{R}_{0_i}^{0.975} - \hat{R}_{0_i}^{0.025}\right], \tag{3.17}$$

where $\hat{R}_{0_i}^{0.025}$ and $\hat{R}_{0_i}^{0.975}$ denote the lower and upper bounds of the 95% BCI for R_{0_i}. Similar as our measure of accuracy, precision of $I(t)$ estimation is quantified via mean relative credible

interval width (MRCIW):

$$\text{MRCIW} = \frac{1}{T+1}\sum_{i=0}^{T}\frac{\hat{I}_i^{0.975} - \hat{I}_i^{0.025}}{I(t_i)+1}, \tag{3.18}$$

where $\hat{I}_i^{0.025}$ and $\hat{I}_i^{0.975}$ specify the lower and upper bounds of the 95% BCI of $I(t_i)$. In addition, we compute the "envelope" (ENV) — a measure of coverage of BCIs the true trajectory — for $R_0(t)$ and $I(t)$ as follows:

$$\text{ENV-R}_0 = \frac{1}{T+1}\sum_{i=0}^{T}\mathbb{1}\left(\hat{R}_{0_i}^{0.025} \le R_0(t_i) \le \hat{R}_{0_i}^{0.975}\right), \text{ENV-I} = \frac{1}{T+1}\sum_{i=0}^{T}\mathbb{1}\left(\hat{I}_i^{0.025} \le I(t_i) \le \hat{I}_i^{0.975}\right) \tag{3.19}$$

Sampling distribution boxplots of $R_0(t)$ posterior summaries are depicted in the left three plots of Figure 3.4. The LNA-based method yields significantly lower MAE compared with the ODE-based method. As a trade-off, the MCIWs produced by the LNA-method are generally higher, as expected since the LNA-based method incorporates the stochasticity in the population dynamics. With less bias and wider BCIs, the LNA-based method BCIs result in better coverage than ODE-based BCIs, as shown by the envelope boxplots.

Sampling distribution boxplots of $I(t)$ posterior summaries, shown in Figure 3.4, are similar to the $R_0(t)$ results, with the LNA-based method generally having lower MRAEs, higher MRCIWs and a better coverage/envelope than the ODE-based method. Again, somewhat counterintuitively, the MRCIWs for the LNA-based method are smaller than the ODE counterparts. This is likely caused by significant bias in $R_0(t)$ estimation by the ODE-based method.

We also report the absolute error (AE) and 95% BCI widths for removal rate γ in Figure 3.4. We note that an informative prior has been chosen for γ, because this parameter is weakly identifiability from genetic data alone. The LNA-based method yields a slightly higher AE than the ODE method. Both methods produce similar BCI widths.

3.4 Analysis of Ebola outbreak in West Africa

We apply our LNA-based method to the Ebola genealogies reconstructed from molecular data collected in Sierra Leone and Liberia during the 2014–2015 epidemic in West Africa

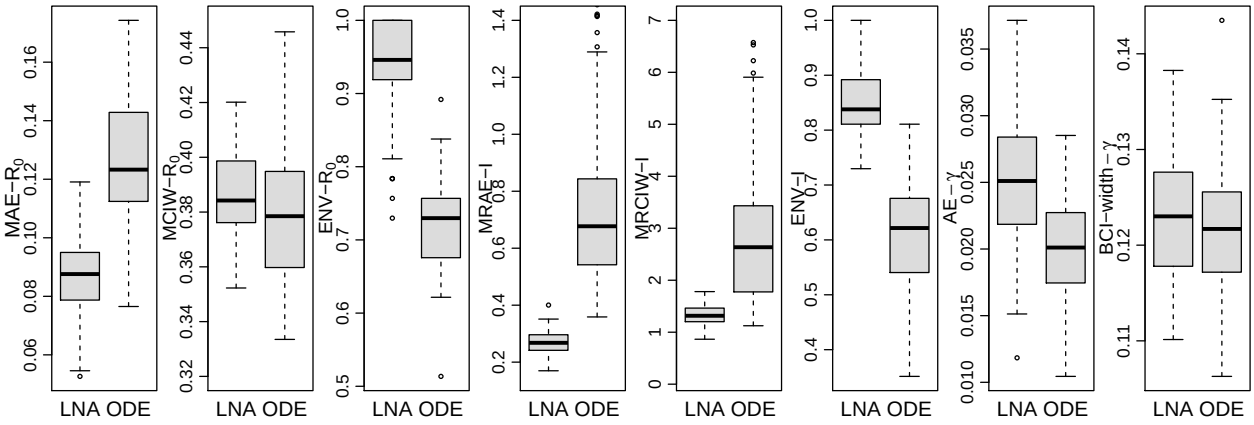

Figure 3.4: Boxplots comparing performance of LNA-based and ODE-based methods using 100 simulated genealogies. The first three plots show mean absolute error (MAE), mean credible interval width (MCIW) and envelope for $R_0(t)$ trajectory. The next three plots depict mean relative absolute error (MRAE), mean relative credible interval width (MRCIW), and envelope for $I(t)$ (prevalence) trajectory. The last two plots show the absolute error (AE) and Bayesian credible interval (BCI) width for γ.

[Dudas et al., 2017]. We use a Sierra Leone genealogy, depicted in the top left plot of Figure 3.5, which was estimated from 1010 Ebola virus full genomes sampled from 2014-05-25 to 2015-09-12 in 15 cities. The Liberia genealogy, shown in the top left plot of Figure 3.6, was estimated from a smaller number of samples: 205 Ebola virus full genomes sampled from 2014-06-20 to 2015-02-14. The original sequence data and the reconstructed genealogies are publicly available at `https://github.com/ebov/space-time`.

When Ebola virus infections were detected in West Africa in mid-Spring of 2014, various intervention measures were proposed and implemented to change behavior of individuals in the populations through which Ebola was spreading. Border closures, encouragement to reduce individual day-to-day mobility, and recommendations on changing burial practices were among the broad spectrum of interventions attempted by multiple countries. It is reasonable to expect that these intervention measures resulted in lowering the contact rates

among members of the populations, which in turn reduced the infection rate, or equivalently the basic reproduction number.

When analyzing the Sierra Leone and Liberia genealogies, we rely on conclusions of Dudas et al. [2017] and assume the population in each country to be well mixed. Furthermore, we assume Ebola spread to follow SIR dynamics. For each country, the population size is specified based on its census population size in 2014, with $N = 7{,}000{,}000$ for Sierra Leone and $N = 4{,}400{,}000$ for Liberia. As in our simulation study, we use the lognormal prior for R_0 with $a_1 = 0.7$ and $b_1 = 0.5$ and the lognormal prior for the inverse standard deviation $1/\sigma$ with $a_2 = 3, b_2 = 0.2$. Recall that this prior setting ensures that *a priori* $R_0(t)$ stays within a reasonable range of $[0, 5]$ with probability 0.9. For removal rate γ, we use an informative lognormal prior with mean 3.4 and variance 0.2 based on previous studies [Towers et al., 2014]. The parameter $1/\gamma$, interpreted as the length of the infection period, is expected to be 8-18 days for each country *a priori*. The total time span for each genealogy is divided evenly into 40 intervals, which results in grid interval lengths, Δt_is, to be 12.41 days for Sierra Leone and 6.9 days for Liberia. We run the MCMC algorithm in Section 3.2.3 for 3,000,000 iterations for Sierra Leone data and 750,000 iterations for Liberia data. The posterior samples are obtained by discarding the first 100,000 iterations and saving every 30th iteration afterward. The trace plots in Section B.1.3 of the Appendix indicate the MCMC algorithm has converged and achieved good mixing in each case.

Figures 3.5 and 3.6 show results for Sierra Leone and Liberia respectively, with intervention events mapped onto the calendar time on the x-axis. Our LNA-based method estimates the initial R_0 in Sierra Leone during 2014–2015 to be 1.68, with 95% BCI of $(1.33, 2.23)$. Similarly, R_0 in Liberia during 2014–2015 has a point estimate 1.67 and a 95% BCI $(1.29, 2.24)$. Our estimate of initial R_0 in Sierra Leone is consistent with the estimates of Stadler et al. [2014], who fitted multiple birth-death models to 72 sequences at the early stages of the outbreak. Volz and Pond [2014] used a susceptible-exposed-infectious-recovered (SEIR) model with a constant R_0 and estimated it to be 2.40 (CI: $(1.54, 3.87)$). Althaus [2014] assumed an exponentially decaying $R_0(t)$ with an estimated initial R_0 of 2.52 (CI: $(2.41, 2.67)$). The dis-

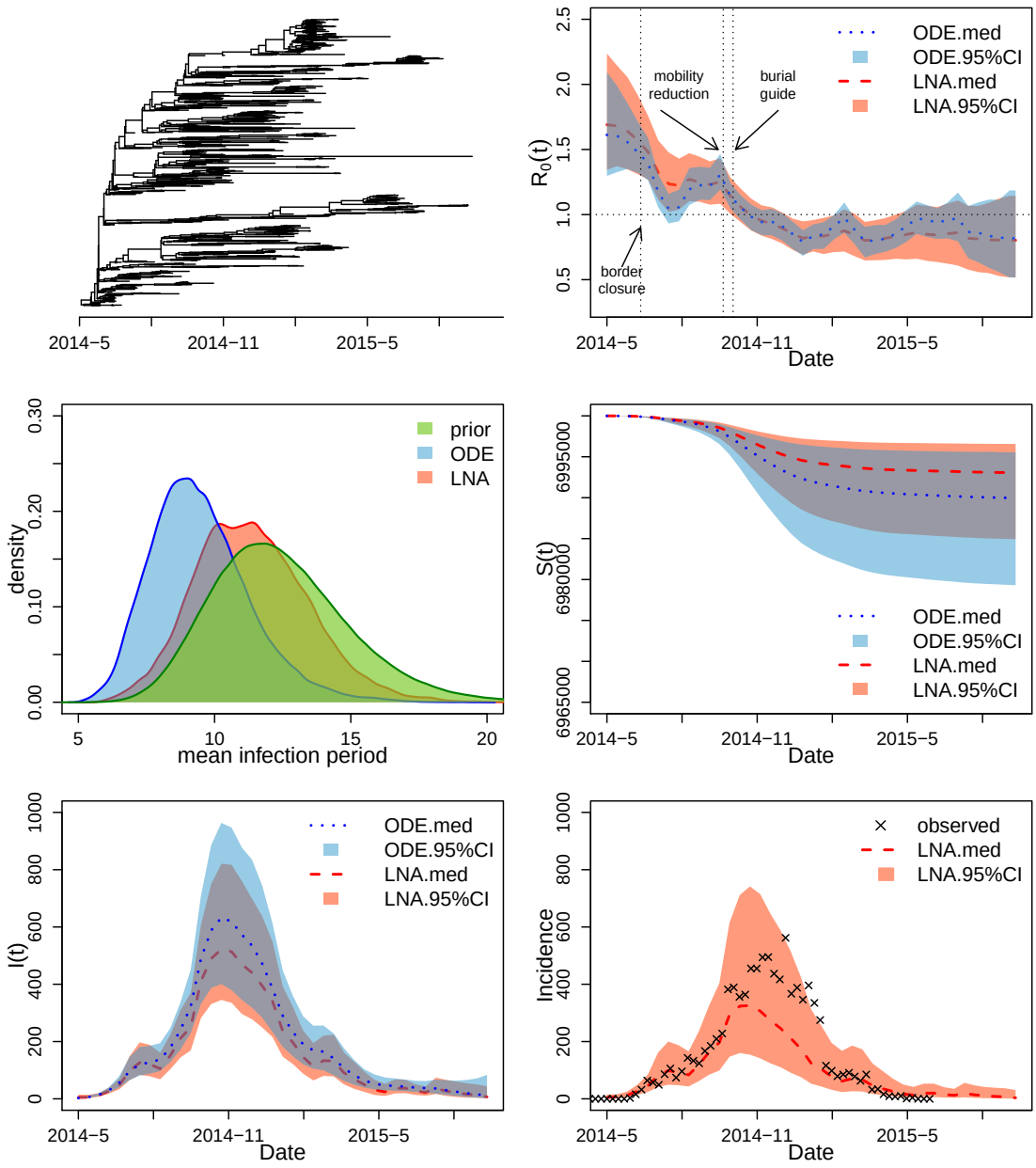

Figure 3.5: Analysis of the genealogy relating Ebola virus sequences collected in Sierra Leone. Top top left plot depicts the Ebola genealogy. The top right plot shows the estimated $R_0(t)$, with the red dashed line showing the posterior median and the salmon shaded area showing the 95% BCIs of the LNA-based method. The posterior median based on the ODE-based method is plotted as the blue dotted line with blue shading corresponding to the 95% BCIs. The medium left figure shows prior and posterior densities of the mean infection period $1/\gamma$. The prior density is shown in green, while the posterior densities based on LNA and ODE are plotted in red and blue respectively. The medium right and the bottom left figures show the estimated trajectory of $S(t)$ and $I(t)$, using the same legend as in top right plot. The bottom right plot shows the predicted median and 95% BCIs for weekly reported incidence together with the reported incidence from WHO shown as crosses.

crepancies between our and SEIR-based estimates are not unexpected, because SEIR models generally yield higher R_0 estimates than SIR models when applied to the same dataset [Wearing et al., 2005, Keeling and Rohani, 2011]. Our estimated R_0 for Liberia is in agreement with results of Althaus [2014], who fitted a SEIR model to incidence data and arrived at an estimated R_0 of 1.59 (CI: $(1.57, 1.60)$).

The $R_0(t)$ dynamics in the two countries share a similar pattern: with (1) a decreasing trend that starts in Spring/Summer of 2014, (2) a stable/constant period until the end of September 2014 and (3) a final decrease below 1.0 (epidemic is contained) around November 2014. Since the number of susceptible individuals did not change significantly over the course of the epidemic, relative to the total population size, the basic and effective reproduction numbers, $R_0(t) = \beta(t)N/\gamma$ and $R_{\text{eff}}(t) = \beta(t)S(t)/\gamma$, are approximately equal. This allows us to compare our $R_0(t)$ estimation results with previously estimated changes in $R_{\text{eff}}(t)$. Our estimation of early $R_0(t)$ dynamics in Sierra Leone agrees with results of Stadler et al. [2013], who concluded that the effective reproduction number did not significantly decrease until mid June. Our estimated $R_0(t)$ trajectory suggests that later interventions, such as border closures and release of burial guides, may have been helpful in controlling the spread of the disease. The infection period for Sierra Leone epidemic is estimated to be 11.2 days with a 95% BCI (7.6,16). For Liberia, the infection period has a point estimate of 9.8, with a 95% BCI $(6.87, 14.05)$. The posterior median of the total number of infected individuals (final epidemic size) is 7,284 and its 95% BCI is $(3397, 14870)$ for Sierra Leone, which is close to 8,706 total confirmed number of cases reported by Centers for Disease Control (CDC). Liberia had a smaller epidemic than Sierra Leone, with estimated total infected individuals being 2,842 and a 95% BCI of $(1296, 6173)$. These results are also in agreement with 3,163 total confirmed cases from CDC.

We perform an out-of-sample validation by comparing our results with weekly reported confirmed incidence in Sierra Leone and Liberia from the World Health Organization [b] (WHO). The posterior predictive weekly incidence at time t, denoted by $\hat{N}(t)$, is approxi-

mated by

$$\hat{N}(t) = \hat{\beta}(t)\hat{S}(t)\hat{I}(t) \cdot \Delta t, \tag{3.20}$$

where $\hat{\beta}(t), \hat{S}(t)$ and $\hat{I}(t)$ are the posterior estimates of the infection rate, number of susceptible and number of infected individuals at time t respectively, and $\Delta t := 7/365$ corresponds the time interval of one week. We plot the posterior predictive estimates of weekly incidence together with the corresponding weekly reported confirmed incidence. For both countries, our model-based incidence 95% BCIs cover the reported incidence counts from WHO, suggesting that our time varying SIR model can estimate incidence well from genetic data alone. We note that our estimated latent incidence should be greater than the reported incidence, because not all Ebola cases were reported and recorded. However, the discrepancy between latent and reported incidence should not be large, because Ebola reporting rate was high. For example, Scarpino et al. [2014] estimated that 83% of Ebola cases were reported.

We also report results from the ODE-based method and superimpose these results over LNA-based results on Figures 3.5 and 3.6. For the relatively small Liberia genealogy, the ODE-based and LNA-based methods yield similar parameter estimates. However, the larger Sierra Leone genealogy produces substantial differences between ODE-based and LNA-based estimates of the $R_0(t)$. The ODE-based method captures the decreasing trend of $R_0(t)$ in Spring and Summer of 2014, but provides narrow BCIs with unrealistic short term fluctuations in the basic reproduction number trajectory.

3.5 Discussion

In this chapter, we propose a Bayesian phylodynamic inference method that can fit a stochastic epidemic model to an observed genealogy estimated from infectious disease genetic sequences sampled during an outbreak. Our statistical model can be viewed as semi-parametric: with (1) a parametric SIR model describing the infectious disease dynamics and (2) a non-parametric GMRF-based estimation of the time varying basic reproduction number. To the best of our knowledge, this is the first method combining a Bayesian nonpara-

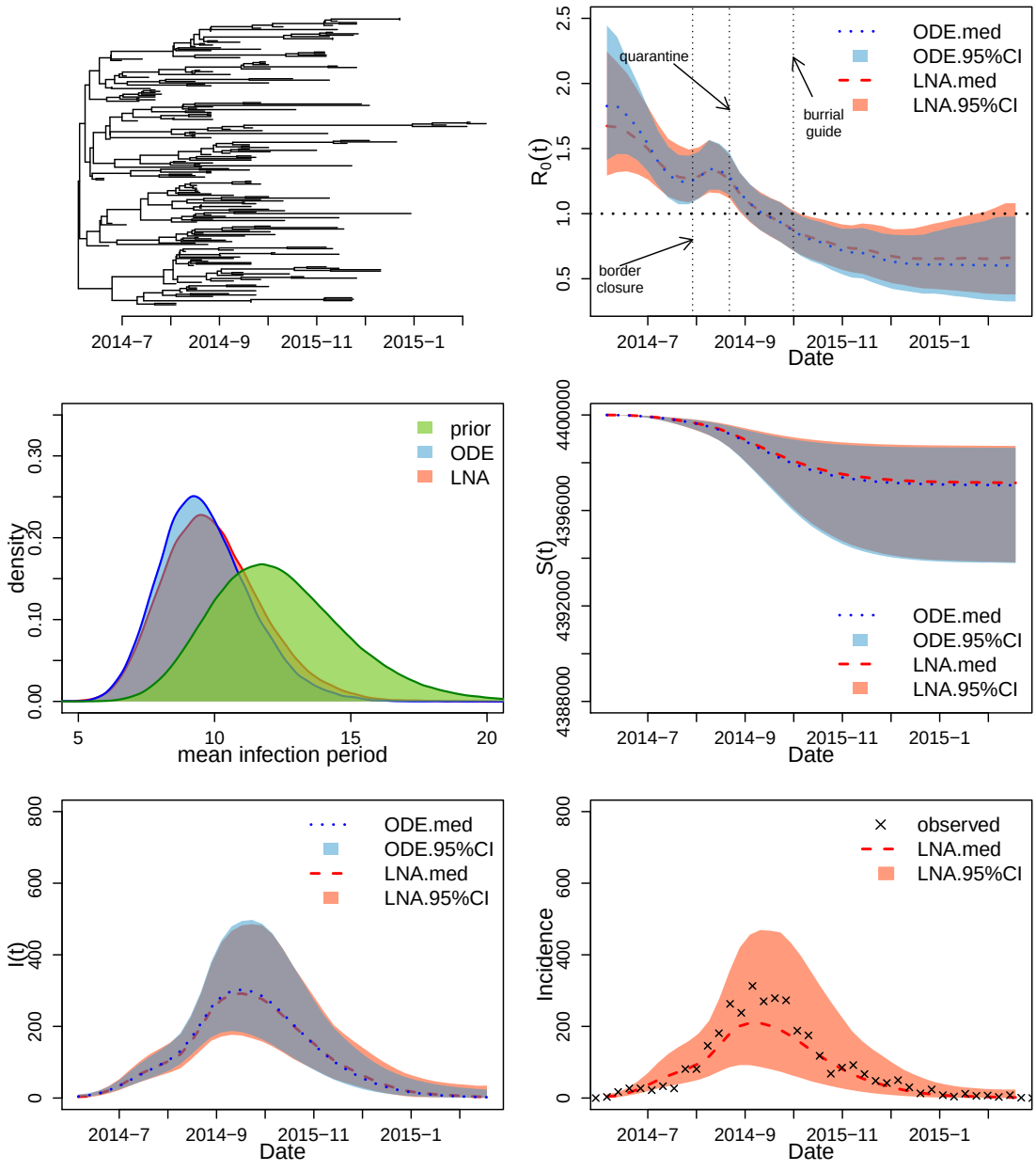

Figure 3.6: Analysis of the genealogy relating Ebola virus sequences collected in Liberia. See caption in Figure 3.5 for the explanation of the plots.

metric approach with a deterministic or stochastic SIR model for phylodynamic inference. Our use of LNA allows us to devise an efficient MCMC algorithm to approximate high dimensional posterior distribution of model parameters and latent variables. Our LNA-based method produces posterior summaries with better frequentist properties than the state-of-the-art ODE-based method, underscoring the importance of working with stochastic models even in large populations. We showcase our method by applying it to the Ebola genealogies estimated from viral sequences collected in Sierra Leone and Liberia during the 2014–2015 outbreak. Our nonparametric estimates of $R_0(t)$ in Sierra Lione and Liberia suggest that the basic reproduction number decreased in two-stages, where the second stage brought it below 1.0 — a sign of epidemic containment. The second stage of $R_0(t)$ decrease closely follows in time implementation of interventions, pointing to their effectiveness.

The experiments in Section 3.3.1 and Appendix Section B.2 indicate that one has to pay close attention to parameter identifiability when fitting SIR models to genealogies or to sequence data directly. Identifiability may not be a problem under an assumption of a constant $R_0(t)$. However, the removal rate tends to be only weakly identifiable in the scenarios with a time-varying basic reproduction number, in which the estimation can be sensitive to the choice of priors. In Section B.2 of the Appendix, we demonstrate that putting a weakly informative prior on the removal rate can cause bias not only in the estimation for removal rate, but also can lead to a failure in recovering the reproduction number and latent population dynamics. Therefore, successful inference of SIR model parameters using genealogical data should rely on a sound informative prior for the removal rate. This constraint is not a big shortcoming in practice, since prior information about the removal rate, or mean length of the infection period, is usually readily available from patient hospitalization data [WHO Ebola Response Team, 2014].

Since parameter identifiability is a recurring problem in infectious disease modeling, integration of multiple sources of information is of great interest. Using particle filter MCMC, Rasmussen et al. [2011] demonstrated that jointly analyzing genealogy and incidence case counts considerably reduces the uncertainty in both estimation of latent population tra-

jectory and SIR model parameters, compared with estimation based on a single source of information. We plan to use our LNA-based framework to perform similar integration of genealogical data and incidence time series. Another possible source of information is the distribution of genetic sequence sampling times. Karcher et al. [2016] proposed a preferential sampling approach that explicitly models dependence of the sampling times distribution on the effective population size. The authors demonstrated that accounting for preferential sampling helps decrease bias and results in more precise effective population size estimation. It would be interesting to incorporate preferential sampling into our LNA-based framework by assuming a probabilistic dependency between sampling times and latent prevalence $I(t)$.

Our method assumes a genealogy/phylogenetic tree is given to us. In reality, genealogies are not directly observed and need to be inferred from molecular sequences. Ideally, uncertainty in the genealogy should be handled by building a Bayesian hierarchical model and integrating over the space of genealogies using MCMC. In fact, implementations of such Bayesian hierarchical modeling already exist for nonparametric, birth-death, and ODE-based phylodynamic approaches [Drummond et al., 2005, Minin et al., 2008, Gill et al., 2013, Stadler et al., 2013, Volz and Siveroni, 2018]. Therefore, an important future direction will be to extend our LNA framework to fitting stochastic epidemic models to molecular sequences instead of genealogies. Similarly to the structured coalescent model implementation of Volz and Siveroni [2018], the easiest way to achieve this will be integration of our LNA MCMC algorithm into popular open source phylogenetic/phylodynamic software packages, such as `BEAST`, `BEAST2`, and `RevBayes` [Suchard et al., 2018, Bouckaert et al., 2014a, Höhna et al., 2016a].

Chapter 4

FITTING STOCHASTIC EPIDEMIC MODEL USING THE INTEGRATION OF INCIDENCE DATA AND GENEALOGY

4.1 Background

Surveillance data are an important source of information for researchers to study and spread of infectious disease on the population level [Anderson and May, 1992, Vynnycky and White, 2010, Keeling and Rohani, 2011]. The most common type of surveillance data is a sequence of incidence counts over a set of time intervals. The incidence data are defined as reported case counts of newly infected individuals over a certain period, for example, weekly-based or monthly-based [Centers for Disease Control, WHO Ebola Response Team, 2014]. One of the challenges of fitting incidence data is that the observed case counts are often recorded at discrete times. Another difficulty is that the incidence data are often subjected to under-reporting [Atkins et al., 2015, Chowell, 2017], since many infectious are never reported (some are asymptomatic), leading to only a fraction of infections being detected and reported by surveillance programs.

One of the most challenging parts of fitting infectious disease models to incidence lies in the time series nature of the data, case counts are not independent across observation time points and depend on the latent disease transmission process governed by the postulated model parameters. The unobserved true population dynamic is often characterized by the rapid change and interactions between different population groups, for example, susceptible and infectious individuals. The SIR model introduce Chapter 2 is a classic example of modeling such interactions [Kermack and McKendrick, 1927].

In recent years, with the development of molecular epidemiology, researchers started producing large amounts of genetic data to complement surveillance case count data. Ge-

nealogies of infectious disease agent samples can be reconstructed from sampled molecular sequences. These genealogies encode coalescent times that provide information about disease transmission events in the population. The coalescent likelihood, linking the genealogy data with population dynamics, opens a door for understanding the disease dynamic, as we explained in Section 2.3.

Both incidence-based and molecular epidemiology studies have been successfully applied to analyze infectious disease dynamics, yet models using only one type of data can suffer from identifiability issues. Despite abundant literature describing either count or molecular data, few researchers take advantages of integrating both types of data to improve inference. Rasmussen et al. [2011] proposed estimating infectious disease dynamics using molecular sequence-based genealogy data along with times-series incidence data and demonstrated an improvement of estimation precision on simulated data. Inspired by the Rasmussen et al. [2011], we modify our LNA-based framework in Chapter 3 to incorporate time series of reported incidence data. We propose a Bayesian framework to fit stochastic epidemic models jointly to genealogy data and surveillance case count data. We also demonstrate that our method can be applied to sequentially obtained/streaming incidence data and molecular sequence data collected during the course of an epidemic. The sequential updating of model parameter estimates allows researchers to continuously improve estimation of stochastic epidemic model parameters, such as the basic reproduction number and prevalence, which is important for timely evaluation of infectious disease control strategies. Moreover, we propose a forecasting framework to predict future incidence and prevalence based on past data. Through simulation studies, we compare incidence-based estimation, genealogy-based estimation, and our method that takes into account both sources of information and show that using genealogy and incidence data leads to higher estimation and forecasting precision and improved parameter identifiability.

Finally, we apply our method for the 2014-2015 Ebola outbreak in Sierra Leone and Liberia, using the WHO weekly incidence data [World Health Organization, b] depicted in Figure 1.1 as well as estimated genealogy from [Dudas et al., 2017].

4.2 Methodology

4.2.1 Modeling Observed Incidence for SIR Model

Throughout this chapter, we assume the epidemic follows a SIR model and the population is well-mixed. We adopt the same notation for SIR model as in Chapter 3. We use a vector $\mathbf{X}(t) = (S(t), I(t))$ to represent the compartmental counts/population trajectory at t, with $S(\cdot)$ and $I(\cdot)$ denoting the counts of susceptible and recovered individuals respectively.

Let Y_i denote the reported incidence over time interval $\mathcal{R}_i = (r_{i-1}, r_i]$. The incidence vector $Y_{1:k}$ represents a sequence of reported incidence counts collected during the disease epidemic over consecutive time intervals $\mathcal{R}_1, \ldots, \mathcal{R}_k$ respectively. Let $\mathbf{N}(t) = \{N_{SI}(t), N_{IR}(t)\}$ denote a counting process for the cumulative events at interval $(r_0, t]$, with $N_{SI}(\cdot)$ recording the cumulative infectious events ($S \to I$ transitions) and $N_{IR}(\cdot)$ representing the cumulative recovery events ($I \to R$ transitions). For SIR model, Ho et al. [2016] summarize a one-to-one linear transform from the cumulative events $\mathbf{N}(t)$ to the population trajectory $\mathbf{X}(t)$:

$$\begin{pmatrix} S(t) \\ I(t) \end{pmatrix} = \begin{pmatrix} S_0 \\ I_0 \end{pmatrix} + \begin{pmatrix} -1 & 0 \\ 1 & -1 \end{pmatrix} \cdot \begin{pmatrix} N_{SI}(t) \\ N_{IR}(t) \end{pmatrix}. \tag{4.1}$$

Then the true incidence over interval $\mathcal{R}_i$ can be represented by $Z_i = N_{SI}(r_i) - N_{SI}(r_{i-1})$. For notation simplicity, we use $\mathbf{X}_i$ to denote and the population counts at r_i, i.e. $\mathbf{X}(r_i)$. By (4.1), the true incidence Z_i in SIR model can be further reduced to

$$Z_i \;\; = \;\; S(r_{i-1}) - S(r_i).$$

Since the observed incidence counts are usually under-reported, we model the observed incidence Y_i as a negative binomial sample from the true incidence Z_i with (under)reporting rate $\rho \in (0, 1)$ and parameter $\phi > 0$ adjusting for over-dispersion, i.e.

$$Y_i \sim \text{NegativeBinomial}\left(\mu_i = \rho \cdot Z_i, \sigma_i^2 = \mu_i + \frac{\mu_i^2}{\phi}\right). \tag{4.2}$$

The relationship between the population trajectory $\mathbf{X}_{0:T}$, true incidence $Z_{1:T}$ and observed incidence $Y_{1:T}$ is depicted in Figure 4.1. In the above setup, the expected value of Y_i is ρZ_i.

The variance of Y_i is $\rho Z_i + (\rho Z_i)^2/\phi$, with ϕ controlling over-dispersion in the mean and variance relationship. As $\phi \to +\infty$, the variance for Y_i approaches ρZ_i, which is the same as its mean, ending up with the Poisson mean variance relationship.

Given the regular grid $t_0 < t_1 < \ldots < t_T$ ($t_0 \leq r_0$ and $r_k \leq t_T$) defined in Section 3.2.1, without loss of generality, we assume the incidence observation times $r_0, \ldots, r_k$ are aligned with the grid points such that $r_i = t_i$ for $i = 1, \ldots, T$ and $T = k$. The joint likelihood of observed incidence $\mathbf{Y}_{1:T}$ given population trajectory $\mathbf{X}_{0:T}$, reporting rate ρ and over-dispersion parameter ϕ can be written as

$$\begin{aligned} L(\mathbf{Y}_{1:T} \mid \mathbf{X}_{0:T}, \rho, \phi) &= \prod_{i=1}^{T} \Pr(Y_i \mid \mathbf{X}_{i-1}, \mathbf{X}_i, \rho, \phi) \\ &= \prod_{i=1}^{T} \frac{\Gamma\left(Y_i + \phi\right)}{\Gamma\left(\phi\right) Y_i!} \frac{\left(\rho(S_i - S_{i-1})\right)^{Y_i} \phi^{\phi}}{\left(\rho(S_i - S_{i-1}) + \phi\right)^{Y_i + \phi}} \cdot \mathbf{1}_{[0,+\infty)}(S_i - S_{i-1}). \end{aligned} \tag{4.3}$$

4.2.2 Coalescent modeling of genealogy

Since we are interested in integrating incidence and genealogy data, we briefly repeat the coalescent model formulation that forms the basis of our probabilistic model of the genealogy. Let $\boldsymbol{\theta}(t) = (\beta(t), \gamma(t))$ be a vector of SIR rates at time t, where $\beta(t)$ represents for the per capita infection rate and $\gamma(t)$ is the removal/ removal rate. For a given genealogy $\mathbf{g}$ defined as in Section 3.2.1, recall that the SIR structured coalescent density/likelihood can be written as

$$\Pr(\mathbf{g} \mid \mathbf{X}_{0:T}, \boldsymbol{\theta}_{0:T}) \propto \prod_{k=2}^{n} \binom{l(c_{k-1})}{2} \frac{2\beta(c_{k-1})S(c_{k-1})}{I(c_{k-1})} \exp\left(- \sum_{i=0}^{i_k - 1} \int_{\mathcal{I}_{i,k}} \binom{l_{i,k}}{2} \frac{2\beta(\tau)S(\tau)}{I(\tau)} \mathrm{d}\tau \right) \tag{4.4}$$

where $c_1, \ldots, c_n$ are coalescent times, $\mathcal{I}_{i,k}$ are intervals at which the $l(t)$ — the number of lineages present at time t takes a constant value. See Section 1.2 for more details. For simplicity, we let population dynamics $\mathbf{X}(t)$ and the rate parameters $\boldsymbol{\theta}(t)$ to vary in a piecewise constant manner, i.e

$$\mathbf{X}(t) = \sum_{i=1}^{T} \mathbf{1}_{[t_{i-1}, t_i)}(t) \mathbf{X}_{i-1} \quad \boldsymbol{\theta}(t) = \sum_{i=1}^{T} \mathbf{1}_{[t_{i-1}, t_i)}(t) \boldsymbol{\theta}_{i-1}.$$

Hence, the integrals in the (4.4) are readily available in closed-form and are fast to compute.

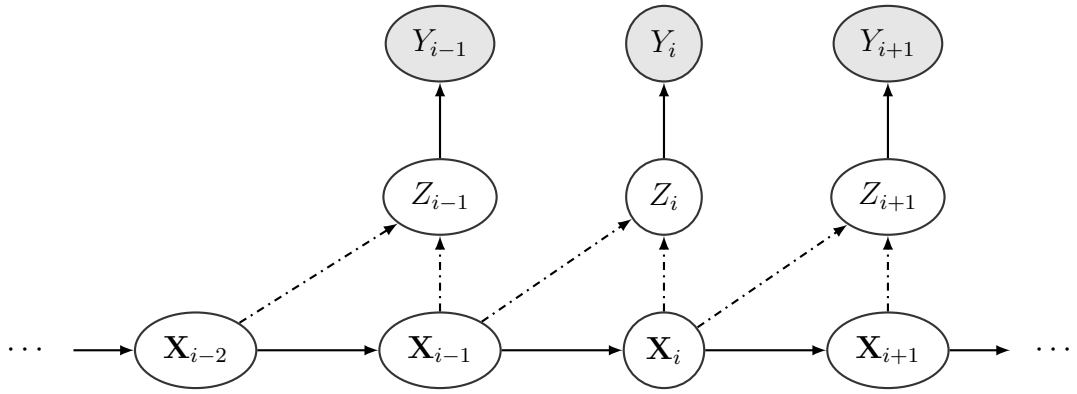

Figure 4.1: Dependency relationship between population trajectory and observed incidence. $\mathbf{X}_i$s denote the population trajectory, which can be considered as a discrete time Markov chain. The true incidence $Z_i = S_{i-1} - S_i$ deterministically depends on the previous state $\mathbf{X}_{i-1}$ and current state $\mathbf{X}_i$. The observed reported incidence Y_i depends on the actual incidence Z_i through negative binomial likelihood function.

4.2.3 SIR dynamics using LNA

Under the assumption of Markov property, the likelihood function for discrete observation of $\mathbf{X}_{1:T}$ given initial value $\mathbf{X}_0$ satisfies the following factorization:

$$\Pr\left(\mathbf{X}_{1:T} \mid \mathbf{X}_0, \boldsymbol{\theta}_{0:T}\right) = \prod_{i=1}^{T} \Pr(\mathbf{X}_i \mid \mathbf{X}_{i-1}, \boldsymbol{\theta}_{i-1}). \tag{4.5}$$

Following the LNA in Section 3.2.2, the intractable SIR transition density $\Pr(\mathbf{X}_i \mid \mathbf{X}_{i-1}, \boldsymbol{\theta})$ will be approximated by a closed-form Gaussian distribution. The closed-form transition densities enable us to compute compute the density of the latent SIR trajectory (4.5). As a result, our augmented posterior distribution of $\mathbf{X}_{0:T}$ and $\boldsymbol{\theta}_{0:T}$ can be computed up to proportionality constant and approximated via "standard" (not particle filter) MCMC approaches.

4.2.4 Bayesian data augmentation

Posterior distribution

Given genealogy $\mathbf{g}$ and reported incidence sequence $\mathbf{Y}_{1:T}$, our aim is to infer the latent SIR population dynamic $\mathbf{X}_{0:T}$, rate parameters $\boldsymbol{\theta}_{0:T}$ over time grid $t_0 < t_1 < \cdots < t_T$ as well as parameters ρ, ϕ in the negative binomial incidence observation likelihood. We make an assumption that the genealogy $\mathbf{g}$ and incidence $\mathbf{Y}_{1:k}$ are conditionally independent, given the SIR population trajectories $\mathbf{X}_{0:T}$ and parameters $\boldsymbol{\theta}_{0:T}, \rho, \phi$, i.e.

$$\Pr(\mathbf{g}, \mathbf{Y}_{1:k} \mid \boldsymbol{\theta}_{0:T}, \mathbf{X}_{0:T}, \rho, \phi) = \Pr\left(\mathbf{g} \mid \mathbf{X}_{0:T}, \boldsymbol{\theta}_{0:T}\right) \Pr\left(Y_{1:T} \mid \mathbf{X}_{0:T}, \rho, \phi\right), \tag{4.6}$$

where $\Pr\left(\mathbf{g} \mid \mathbf{X}_{0:T}, \boldsymbol{\theta}_{0:T}\right)$ is the structured coalescent likelihood defined in Section 3.2.1 and $\Pr\left(Y_{1:T} \mid \mathbf{X}_{0:T}, \rho, \phi\right)$ denotes the observed incidence negative binomial likelihood (4.3). Let $\pi(\mathbf{X}_0)$, $\pi(\boldsymbol{\theta}_{0:T})$, $\pi(\rho)$ and $\pi(\phi)$ denote the prior densities for the initial compartment states, the SIR parameters, the incidence reporting rate, and the over-dispersion parameter respectively. The posterior distribution for the population trajectory $\mathbf{X}_{0:T}$ and parameters $\boldsymbol{\theta}_{0:T}$ given observed genealogy $\mathbf{g}$ is

$$\begin{aligned}\Pr\left(\boldsymbol{\theta}_{0:T}, \mathbf{X}_{0:T}, \rho, \phi \mid \mathbf{g}, Y_{1:T}\right) \propto\ & \underbrace{\Pr\left(\mathbf{g} \mid \mathbf{X}_{0:T}, \boldsymbol{\theta}_{0:T}\right)}_{\text{coalescent likelihood}} \cdot \underbrace{\Pr\left(Y_{1:T} \mid \mathbf{X}_{0:T}, \rho, \phi\right)}_{\text{incidence likelihood}} \cdot \\ & \underbrace{\Pr\left(\mathbf{X}_{1:T} \mid \mathbf{X}_0, \boldsymbol{\theta}_{0:T}\right)}_{\text{trajectory density}} \cdot \underbrace{\pi\left(\boldsymbol{\theta}_{0:T}\right) \pi\left(\mathbf{X}_0\right) \pi\left(\rho\right) \pi\left(\phi\right)}_{\text{prior density}},\end{aligned} \tag{4.7}$$

where $\Pr(\mathbf{X}_{1:T} \mid \mathbf{X}_0, \boldsymbol{\theta}_{0:T})$ is the joint density of discrete observations of trajectory $\mathbf{X}_{1:T}$ given the initial value $\mathbf{X}_0$, defined in Equation (4.5) .

Reparameterization

Since it is much more likely for the infection rate to be time variable, we are going to assume a constant remove/ removal rate γ. We assume an emerging infectious disease outbreak at the beginning and set the initial counts of susceptible to be $S_0 = N - I_0$.

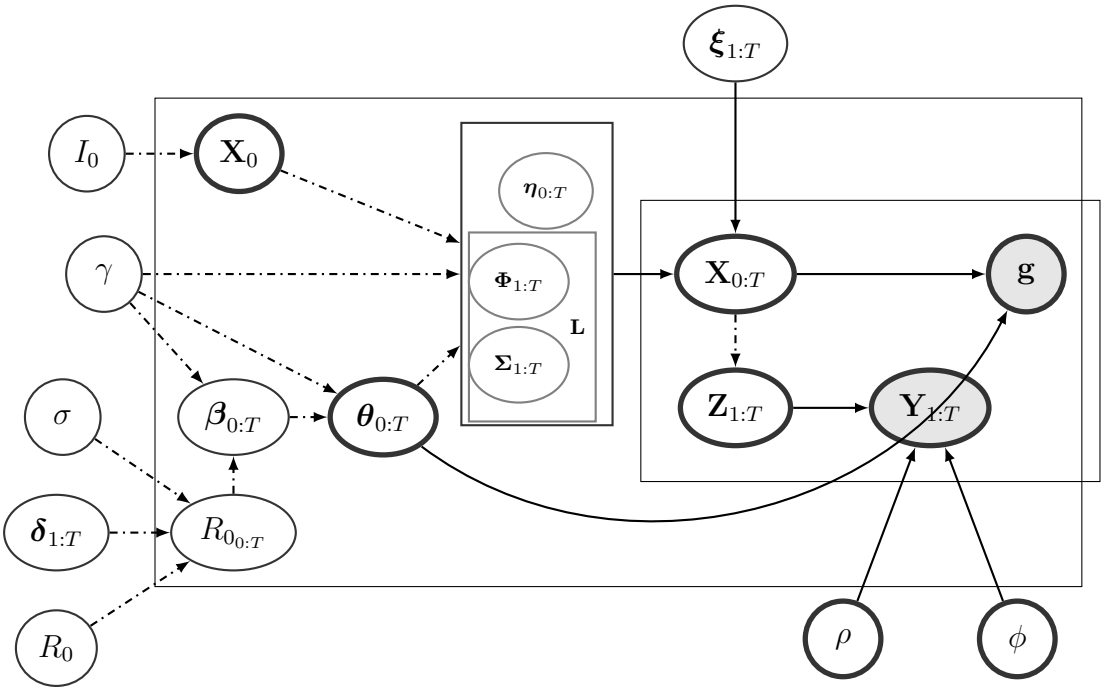

Figure 4.2: Parameter dependency graph for the data integration method after reparameterization. The observed genealogy data $\mathbf{g}$ and incidence data $\mathbf{Y}_{1:T}$ are in the gray nodes. The root nodes $I_0, \gamma, \sigma, \boldsymbol{\delta}_{1:T}, R_0, \sigma$ outside the large box are parameters after reparameterization, for which we assigned prior distributions. The dashed-dotted lines show deterministic relationships and the solid lines show stochastic dependencies. The gray nodes denote the observed data. The figure shows the dependency structure among the transformed parameters and original parameters $\boldsymbol{\theta}_{0:T}, \mathbf{X}_0, \mathbf{X}_{0:T}$ and $\mathbf{Z}_{1:T}$.

As in Section 3.2.3, $\mathbf{X}_{1:T}$ is parameterized with a sequence of independent Gaussian random variables $\boldsymbol{\xi}_{1:T}$, following a non-centered parameterization according to formula (3.11). In such parameterization, we will treat $\boldsymbol{\xi}_{1:T}$ as random latent variables and the SIR latent trajectory $\mathbf{X}_{1:T}$ as a deterministic transformation of $\boldsymbol{\xi}_{1:T}$.

We also reparameterize the infection rate vector $\boldsymbol{\beta}_{0:T}$ using reproduction number vector $\mathbf{R}_{0_{0:T}}$ based on (3.12). Like Section 3.2.3, we use a GMRF to *a priori* model the time-varying reproduction numbers and parameterize $\mathbf{R}_{0_{0:T}}$ in terms of independent normal random variables $\boldsymbol{\delta}_{1:T}$, representing log increments of $R_0(t)$, and smoothing hyperparameter σ.

Hence, after the reparameterization, the parameters $\mathbf{X}_{0:T}, \boldsymbol{\theta}_{0:T}$ will be replaced by $I_0, R_0, \gamma, \boldsymbol{\delta}_{1:T}, \sigma$ and $\boldsymbol{\xi}_{1:T}$. Figure 4.2 shows the parameter dependency graph after reparameterization. More details on the parameterization is provided in Section 3.2.3.

parameter	prior	parameter	prior
I_0	lognormal$(1,1)$	R_0	lognormal(a_1,b_1)
σ	lognormal(a_2,b_2)	γ	lognormal(a_3,b_3)
$\boldsymbol{\delta}_{1:T}$	$\mathcal{N}(\mathbf{0},\mathbf{I})$	$\boldsymbol{\xi}_{1:T}$	$\mathcal{N}(\mathbf{0},\mathbf{I})$
logit(ρ)	$\mathcal{N}(a_4,b_4)$	ϕ	lognormal (3,1.5)

Table 4.1: Table for prior distributions of parameters and latent variables.

Table 4.1 shows the prior specification for parameters and latent variables. We use the same family of prior distribution of $R_0, I_0, \sigma, \gamma, \delta_{1:T}$ and $\xi_{1:T}$ as in Chapter 3. For the reporting rate ρ, we assume its logit transform follows a normal distribution with mean a_3 and variance b_3. The over-dispersion parameters ϕ in the negative binomial incidence likelihood function has prior lognormal$(3, 1.5)$.

MCMC algorithm

Using our new parameterization, we are now interested in the posterior distribution of the initial number of infected individuals, I_0, removal rate, γ, the initial basic reproduction number, R_0, standardized vectors, $\boldsymbol{\delta}_{1:T}$ and $\boldsymbol{\xi}_{1:T}$, GMRF standard deviation, σ, reporting rate, ρ and over-dispersion parameter ϕ:

$$\begin{aligned}
&\Pr(I_0, R_0, \gamma, \boldsymbol{\delta}_{1:T}, \boldsymbol{\xi}_{1:T}, \sigma, \rho, \phi \mid \mathbf{g}) \\
&\propto \Pr(\mathbf{g} \mid I_0, R_0, \gamma, \boldsymbol{\delta}_{1:T}, \boldsymbol{\xi}_{1:T}, \sigma) \Pr(\boldsymbol{\delta}_{1:T}) \Pr(\boldsymbol{\xi}_{1:T}) \pi(I_0) \pi(R_0) \pi(\gamma) \pi(\sigma) \pi(\rho) \pi(\phi) \\
&\propto \Pr(\mathbf{g} \mid \mathbf{X}_{0:T}, \boldsymbol{\theta}_{0:T}) \Pr(\boldsymbol{\delta}_{1:T}) \Pr(\boldsymbol{\xi}_{1:T}) \pi(I_0) \pi(R_0) \pi(\gamma) \pi(\sigma) \pi(\rho) \pi(\phi).
\end{aligned}$$

The latent variables $\mathbf{X}_{0:T}$ and parameter vector $\boldsymbol{\theta}_{0:T}$ are deterministic functions of new parameters I_0, γ, R_0, $\boldsymbol{\delta}_{1:T}$, $\boldsymbol{\xi}_{1:T}$, and σ.

We use the same sampling strategy to update $\log(R_0), \boldsymbol{\delta}_{1:T}, \log(\sigma), \boldsymbol{\xi}_{1:T}, I_0, \gamma$ as in Section 3.2.3. Reporting rate ρ and over-dispersion parameters ϕ are updated using univariate Metropolis steps in logit scale and log scale respectively. The full procedure is described in Algorithm 8 in Appendix and details of the elliptical slice sampler can be found in Section 2.4.5. After MCMC is done, we report posterior summaries using natural parameterization. For example, we report posterior medians and 95% Bayesian credible intervals (BCIs) of the piecewise latent reproduction number trajectory, $\mathbf{R}_{0_i}$ for $i = 0, \ldots, T$, and latent trajectory $\mathbf{X}_{0:T}$.

4.2.5 Implementation

Our R package called `phyloInt` provides an implementation of our MCMC algorithm, which contains functions to run genealogy-based, incidence-based, and data integration-based methods. The package code is publicly available at `https://github.com/MingweiWilliamTang/phyloInt`. This repository also contains scripts that should allow one to reproduce key numerical results in this chapter.

4.3 Sequential parameter estimation and forecasting

Until now, our simulation examples and real data analyses focused on epidemic models using data collected from the whole epidemic. In reality, the incidence cases counts and genetic data are collected sequentially through time and understanding the currently disease dynamic helps researchers to implement interventions, such as vaccination, quarantine, and restriction of mobility in the population. We adapt our framework to sequentially fit streaming epidemic data, providing an estimate of the reproduction number and disease prevalence at the current stage. Moreover, besides fitting the observations to current, we also produce posterior predictive forecasting to make short-term predictions about the future course of the epidemic.

Starting with observed data before time $t_{T'}$, we have incidence observation $\mathbf{Y}_{1:T'}$ and genealogy observation $\mathbf{g}(T')$. Let $\widehat{\boldsymbol{\theta}_{T'}}$, $\widehat{\mathbf{X}_{T'}}$, $\widehat{\rho}$ and $\widehat{\phi}$ be the posterior sample of rate SIR rate parameters, population size, reporting rate and over-dispersion parameter through MCMC. We use $\widetilde{\mathbf{X}}_{T'+j}$ to denote the predicted population at a future time point for $j = q, \dots q$, where $t_{T'} < t_{T'+1} < t_{T'+2} < \ldots < t_{T'+q}$ is a sequences of time steps of forecasting interest and $\widetilde{\mathbf{X}}_{T'} = \mathbf{X}_{T'}$. Our short term forecasting model assumes the SIR rates will not change drastically in the short-term, i.e., the rate $\boldsymbol{\theta}(t)$ will be fixed to $\widehat{\boldsymbol{\theta}}_{T'}$ for $t \geq t_{T'}$. For $j = 1, \dots, q$, the future population size $\widetilde{\mathbf{X}}_{T'+j}$ and the corresponding predicted reported incidence $\widetilde{Y}_{T'+j}$ can be estimated by running the posterior predictive model below:

$$\begin{aligned} \widetilde{\mathbf{X}}_{T'+j} &\sim \text{doLNA}\left(\cdot \mid \widetilde{\mathbf{X}}_{T'+j-1}, \widetilde{\boldsymbol{\theta}}_{T'}\right) \text{ on } [t_{T'+j-1}, t_{T'+j}], \\ \widetilde{Y}_{T'+j} &\sim \text{NegativeBinomial}\left(\mu = \widehat{\rho}\widetilde{\mathbf{X}}_{T'+j}, \sigma^2 = \mu + \mu^2/\widehat{\phi}\right), \end{aligned} \tag{4.8}$$

where function `doLNA` is the procedure of simulating SIR trajectory one step forward using non-restarting LNA method, see Section 3.2.2 for more details. By running (4.8) and obtaining the corresponding posterior predictive sample of $\widetilde{\mathbf{X}}_{T'+1:T'+q}, \widetilde{Y}_{T'+1:T'+q}$, we therefore have an empirical posterior probability distribution of future incidence.

Evaluating forecasts We apply logarithmic score metric used by Reich et al. [2019] to evaluate the probabilistic forecasting accuracy of future observed incidence. The log-score metric is a proper scoring rule for evaluating probabilistic forecasts of continuous variables [Gneiting and Raftery, 2007]. Here we follow a similar procedure from Reich et al. [2019] and computed the modified log-score. Let y_j^* denote the true value at t_j and $p(\cdot|\mathbf{Y}, t_j)$ be a posterior predictive distribution for forecasting y_j^* using observed data $\mathbf{Y}$. Let $q > 0$ such that $(y^* - q, y^* + q)$ characterizes an interval that helps to specify a critical region around the truth, then the logarithmic score is calculated by

$$ls(y^*_{T'+j}) = \log\left(\int_{y^*_{T'+j}-q}^{y^*_{T'+j}+q} p(y_{T'+j}|\mathbf{Y}_{1:T'})\mathrm{d}y_{T'+j}\right) \approx \log\left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{1}_{(y^*_{T'+j}-q, y^*_{T'+j}+q)}\left(\widetilde{Y}^{(i)}_{T'+j}\right)\right), \quad (4.9)$$

where $Y^{(i)}(f_j)$ for $i = 1, \ldots, n$ is a sequence of MCMC posterior predictive samples obtained using formula (4.8). Reich et al. [2019] choose q be the 0.5% of the total population size. Since the epidemic size is relatively small compared to the total population in our simulation examples and real data analyses, we choose $q = 15$ cases by considering an error within ± 15 as a good forecast for future reported incidence events. We assess the overall performance of j-steps ahead forecasting result by taking the exponential of the average score over a collection of cutoff times $\mathcal{C}$,

$$j\text{-step-score} = \exp\left(\frac{1}{\mathcal{C}}\sum_{T'\in\mathcal{C}} ls(y^*_{T'+j})\right). \quad (4.10)$$

4.4 Simulation study

The extension of our LNA-based method in this chapter now provides us with three methods that fit stochastic epidemic models to three different sources of data: (1) Genealogy (Gen-based) method that only uses a genealogy $\mathbf{g}$ introduced in Chapter 3 (2) Incidence-based (Incid-based) method that only uses incidence observations $\mathbf{Y}_{1:T}$ (3) a joint model that integrates incidence counts and genealogy (Joint-based). In this section, we simulate genealogies and incidences counts to assess the performance of the three methods.

4.4.1 Simulation based on a single genealogy and incidence realization

We take a similar simulation strategy as in Chapter 3, where one realization of SIR population trajectories based on the MJP is simulated using pre-specified population size N and parameters γ, I_0 and $R_0(t)$. Secondly, we simulate coalescent times with pre-specified sampling times. Finally, a sequence of observed incidence is simulated based on the population trajectory from negative binomial distribution (4.2).

To address to the sensitivity to prior specification of removal rate, discussed in Chapter 3, we compare the performance of joint inference versus Gen-based method and Incid-based method under weakly informative priors of removal and reporting rates. We compare the three methods under three "true" $R_0(t)$ trajectories over the time interval $[0, 90]$:

1. Constant (CONST) $R_0(t)$. $R_0(t) = 1.8$ for $t \in [0, 90]$. Recovery rate $\gamma = 0.2$. Initial counts of infected individuals $I_0 = 3$. Total population size $N = 100{,}000$. Reporting rate $\rho = 0.7$. Over-dispersion parameter $\phi = 15$.

2. Stepwise decreasing (SD) $R_0(t)$. $R_0(t) = 1.8, t \in [0, 30)$, $R_0(t) = 1.26, t \in [30, 60)$ and $R_0(t) = 0.756, t \in [60, 90]$. Recovery rate $\gamma = 0.2$. Initial counts of infected individuals $I_0 = 3$. Total population size $N = 1{,}000{,}000$. Reporting rate $\rho = 0.7$. Over-dispersion parameter $\phi = 15$.

3. Non-monotonic (NM) $R_0(t)$. $R_0(t) = 1.4 \times 1.015^{0.5t}, t \in [0, 30]$, $R_0(t) = 1.750 \times 0.975^{t-30}, t \in [30, 80]$ and $R_0(t) = 0.4583, t \in [80, 90]$. Recovery rate $\gamma = 0.3$. Initial counts of infected individuals $I_0 = 3$. Total population size $N = 1{,}000{,}000$. Reporting rate $\rho = 0.8$. Over-dispersion parameter $\phi = 15$.

We use lognormal$(1, 1)$ prior for I_0 in all simulations. The parameters of the lognormal priors for the initial R_0 and inverse standard deviation $1/\sigma$ are set to $a_1 = 0.7, b_1 = 0.3$ and $a_2 = 3, b_2 = 0.2$ respectively. In contrast to the informative prior setup for γ in Section 3.3.1, here we assign a weakly informative prior to see if the prior sensitivity issues in Appendix B.2 can be resolved in Incid-based method or Joint-based method: (1)

CONST: $\gamma \sim \text{lognormal}(-1.7, 0.25)$, (2) SD: $\gamma \sim \text{lognormal}(-1.7, 0.25)$, (3) NM: $\gamma \sim \text{lognormal}(-1.2, 0.25)$. The grid size is set to be $T = 36$, with $t_i - t_{i-1} = 2.5$ for $i = 1, \ldots, 36$.

As for the reporting rate, we set the prior to be $\text{logit}(\rho) \sim \mathcal{N}(0, 1.5)$, which provides a nearly uniform distribution on $(0, 1)$. The prior over-dispersion parameter ϕ is assigned to be $\text{lognormal}(3, 1.5)$. For each of the method in each scenario, we use 400,000 MCMC iterations. The effective sample sizes of all unknown quantities were above 100.

The first row of Figure 3.3 shows point-wise posterior medians and 95% BCIs for the basic reproduction number trajectory, $R_0(t)$. Our Joint-based method performs well in capturing the continuous dynamics of $R_0(t)$ and produces less biased point estimate and narrower BCIs than the other two methods. Though our approach may not perfectly catch the discontinuous changes in R_0 in the SD scenario and the sharp increase in NM scenario, the method provides BCIs that are able to capture most of the $R_0(t)$ trajectory. The Gen-based and Incid-based method yield similar results in the CONST case and the SD case with wider BCIs, but fail to capture the decreasing trend at the end of the NM scenario.

The second row in Figure 3.3 shows posterior summaries of removal rate γ. Both Joint-based and Gen-based methods perform well in the CONST scenario, with posterior modes centered at the true value and higher posterior densities at truth when compared with the prior. In the SD and NM scenarios with the time varying $R_0(t)$, while our Joint-based method also provides a good estimate, the posterior estimates from the Gen-based method and Joint-based method, though still centered at the truth, do not differ much from the prior distribution, hinting a identifiability problems when fitting stochastic epidemic models to single data type.

Posterior summaries of $S(t)$ and $I(t)$ are depicted in the third and fourth rows of Figure 3.3. The Joint-based and Gen-based methods produce similar results in all scenarios, as both of them have BCIs covering the true trajectories, but the Joint-based method produces narrower credible intervals. The Incid-based method provides similar estimate in the CONST scenarios, but provides really large BCIs in the SD scenario and significantly overestimates the disease prevalence in the NM case.

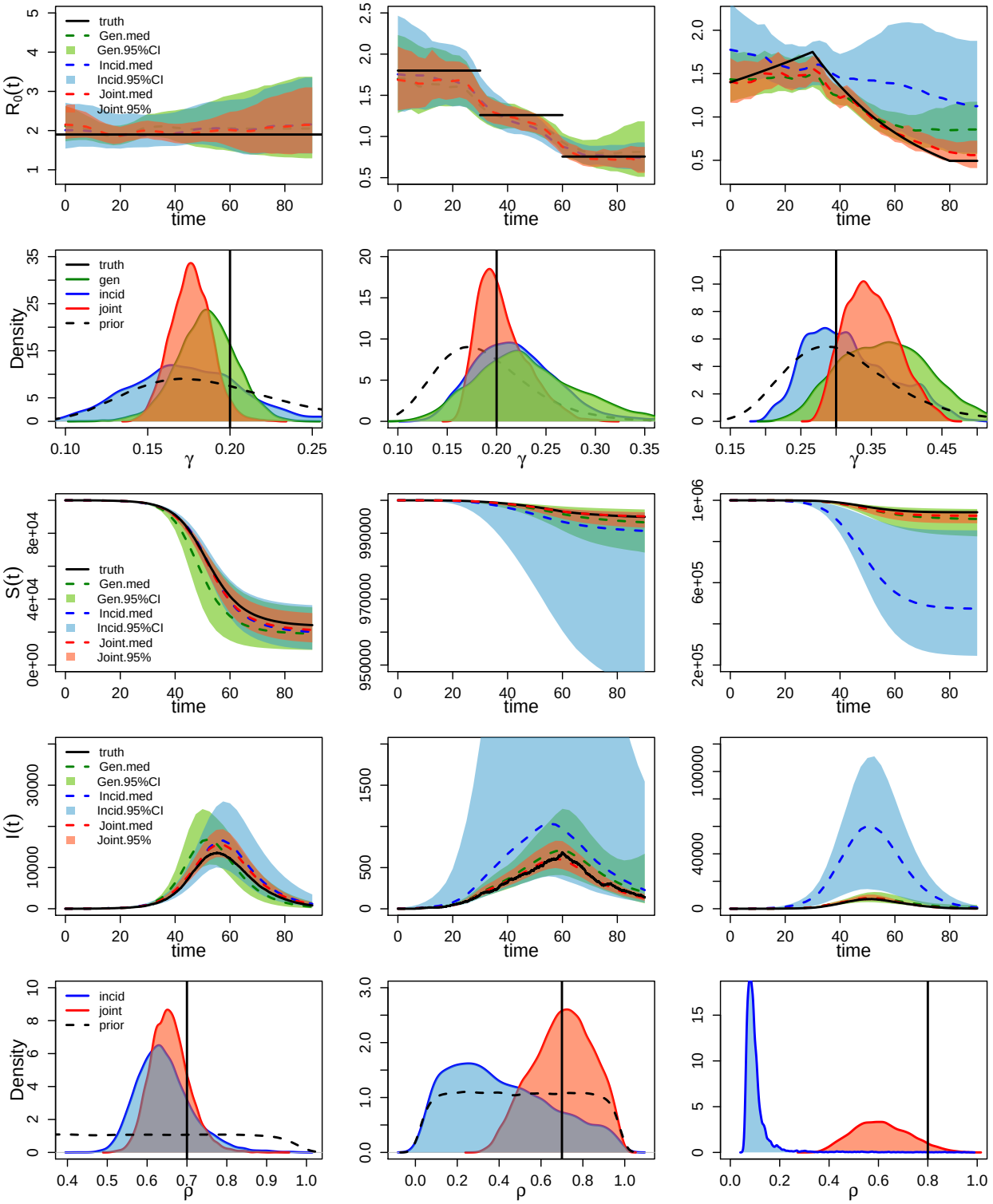

Figure 4.3: Analysis of 3 simulation scenarios. Columns correspond to CONST, SD and NM simulation scenarios.

The last row in Figure 4.3 shows the posterior estimate for reporting rate ρ in the Joint-based and Incid-based methods. Our Joint-based method has the posterior density centered near the truth in all scenarios. The Incid-based method, though successfully recovers the truth in CONST case, significantly underestimates the reporting rate in SD and NM scenarios. The huge bias in reporting rate estimates can be an explanation the discrepancies in the population trajectory estimate produced by the Incid-based method.

4.4.2 Repeated simulations

We also design a simulation study by repeatedly simulating SIR trajectories, based on which the corresponding genealogies and incidence data are simulated. The trajectories are simulated based on the non-monotonic $R_0(t)$ trajectory scenario in Section 4.4.1 with the same parameter setup, except the sampling times are set to be same as in Section B.2.2 to address the prior sensitivity issues discussed in the previous Chapter. See Appendix B.1.2 for details of the sampling times generation. Moreover, we use another weakly informative prior lognormal$(0.4, 1.2)$ for reporting rate ρ. We use the same rejection criterion as in Section 3.3.2 for "unreasonable" realizations and finally arrive at 100 simulated data sets. The details of the rejection criteria are given in Section B.1.2 of the Appendix.

Comparisons between three models are evaluated according to their estimation of reproduction number $R_0(t)$, disease prevalence $I(t)$, removal rate γ, and reporting rate ρ. As in Section 3.3.2, $R_0(t)$ estimation is evaluated using MAE, MCIW, and envelop. The performance of estimating $I(t)$ is assessed based on relative measures MRAE and MRCIW. We also report the envelop. See Section 3.3.2 for details about measurement criteria. We also report the absolute error (AE) and BCI width for γ and ρ respectively to access the model performance under weakly informative priors.

The left three plots of Figure 4.4 in the first row depict the sampling distribution of $R_0(t)$ posterior estimation summaries. The Joint-based method outperforms the Gen-based method and Incid-based method by producing significantly lower MAE, MCIW and higher envelope. The next two plots in the first row show the AE and the BCIs for removal rate γ,

demonstrating that our Joint-based method has lower absolute error and narrow BCI width. Moreover, under our weakly informative prior setup, the Incid-based method yields better γ estimate than the Gen-based method. Sampling distribution of $I(t)$ posterior summaries is given in the left three plots of Figure 4.4, with Joint-based method having lower MRAE and MRCIW than the other two methods. The Incid-based method has incredible huge larger bias and uncertainty than the other two methods. As for envelope, our Joint-based method is a little overconfident, with the envelope lower than that from the Gen-based method. The last two plots illustrate the estimation summaries for the reporting rate ρ from the Joint-Based and Incid-based methods. The Joint-based method produces much lower AE and BCIs than the Incid-based method. The large bias and uncertainty in the reporting rate estimate from Incid-based method is consistent with the result in Section 4.4.1. The poor estimation of the reporting rate goes hand in hand with bias in estimating prevalence by Incid-based method in repeated experiments.

Along side with the weakly informative prior setup, we also conduct three other experiments based on different priors for removal and reporting rates. These simulation results are given in Appendix B.3.

4.5 Real data

We apply our Joint-based method to study the 2014–2015 Ebola outbreak in West Africa. Besides the reconstructed genealogies from molecular data collected in Sierra Leone and Liberia [Dudas et al., 2017] used in Chapter 3, we also obtain weekly Sierra Leone and Liberia incidence data from World Health Organization [b]. The incidence data are publicly available at `http://apps.who.int/gho/data/node.ebola-sitrep.quick-downloads`.

The incidence case counts were collected weekly from 2014-01-05 to 2015-05-24, with cases falling into two categories — confirmed and probable. According to World Health Organization [a], the confirmed cases are defined as patients with laboratory Ebola confirmation, such as positive IgM antibodies, positive PCR or viral isolation. The probable cases are patients who had fever, but did not respond to usual fever treatments or show

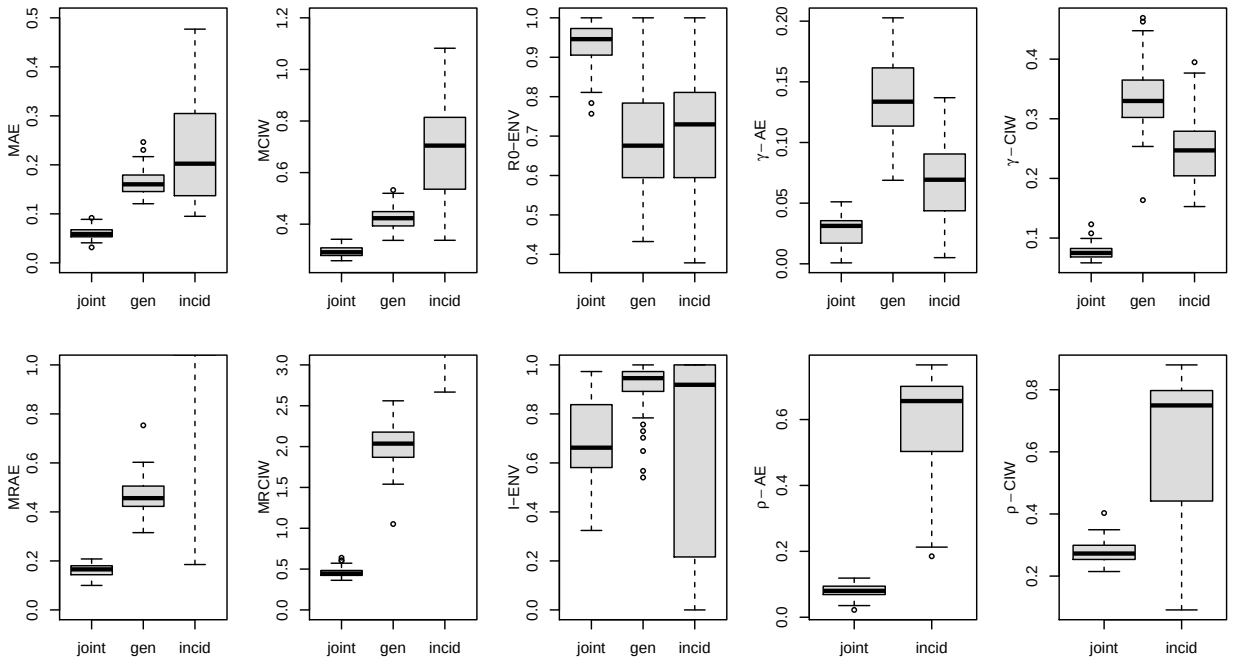

Figure 4.4: Boxplot comparing the performance of Joint-based, Gen-based and Incid-based methods using 100 simulated genealogies. The first three plots in the first row shows the mean absolute error (MAE), mean credible interval width (MCIW) and envelope for basic reproduction number $R_0(t)$ trajectory. The next two plot shows the absolute error (AE) and Bayesian credible interval (BCI) width for removal rate γ. The first three plots in the second row shows the mean relative absolute error (MRAE), mean relative credible interval width (MRCIW) and envelope for prevalence $I(t)$. The last two plots depict the AE and BCI for reporting rate ρ for comparing Joint-based and Incid-based method.

other symptoms of Ebola without lab diagnosis. Here we use the summation of combined confirmed and probable cases as our observed incidence data. As literature suggests, Ebola patients can be asymptomatic, so it is reasonable to assume the incidence cases from WHO are under-reported [Glynn et al., 2017].

4.5.1 Batch data from the whole epidemic

As in Chapter 3, we assume that each country's infectious disease dynamics follow a SIR model with time-varying basic reproduction number $R_0(t)$, possibly due to implementation of intervention measures (e.g., border closures). The total population size in each country is set to the census population size, with N =7,000,000 for Sierra Leone and N =4,400,000 for Liberia. We use a lognormal prior for R_0 with $a_1 = 0.7$ and $b_1 = 0.3$ and a lognormal prior for σ with $a_2 = 3$, $b_2 = 0.2$. For the removal rate, instead of the informative prior lognormal$(3.4, 0.2)$ used in Section 3.4, here we use a weakly informative prior lognormal$(3.4, 0.45)$. For Liberia model, we focus on the dynamics from 2014-06-02 to 2015-02-16 with change points and grid interval length set to be 7 days. In Sierra Leone, the time span is selected to be 2014-04-28 to 2015-09-14 with grid interval lengths (Δt_is) to be 7 days and the distance between change points set to 14 days. We run the MCMC algorithm for 1,200,000 iterations for Sierra Leone and 2,000,000 iterations for Liberia. After discarding the first 200,000 iterations results, we save posterior samples at every 30th iteration afterward.

Figure 4.5 and 4.6 depict posterior summary results for Sierra Lone and Liberia respectively, with the corresponding intervention events mapped onto the calendar time on the x-axis. Our Joint-based method estimates the initial R_0 in Sierra Leone to be 1.49 with 95% BCI of (1.26, 1.82). Similarly, the R_0 in Liberia has a point estimate 1.42 with 95% BCI to be (1.19,1.74). Under the weakly informative prior set up, both Sierra Leone and Liberia estimates are consistent with the R_0 estimate in Section 3.4, which was based on genealogy data with an informative prior on removal rate. Our Joint-based method provides narrower BCIs than BCIs reported in Section 3.4:(1.33, 2.23) in Sierra Leone and (1.29, 2.24). Our estimate of initial R_0 in Sierra Leone also agrees with the estimate from Stadler et al. [2014]'s

birth-death model based on 72 Ebola genomes collected early in the epidemic. Volz and Pond [2014], Althaus [2014] estimate the Sierra Leone dynamics using a SEIR model, ending up with R_0 of 2.40 (CI: (1.57,3.87)) and 2.52 (CI: $(2.41, 2.67)$). Such discrepancies between our SIR model and SEIR-based estimate are not unexpected, as the SEIR model often yield a higher reproduction number estimate than SIR model on the same dataset [Wearing et al., 2005, Keeling and Rohani, 2011].

The $R_0(t)$ dynamics in the two countries share a similar pattern: with (1) a decreasing trend that starts in Spring/Summer of 2014, (2) a steady period until the end of September 2014 and (3) a final decrease below 1.0 (epidemic is contained) around November 2014. Our estimation of early $R_0(t)$ dynamics in Sierra Leone agrees with results of Stadler et al. [2013], who concluded that the effective reproduction number did not significantly decrease until mid June. Our Joint-based model also provides strong evidence that $R_0(t)$ stopped decreasing and followed by a steady period in the summer of 2014. We suspect this can be attributed to the rain season in West Africa, as Ebola virus is able to survive for longer times during such weather conditions [Ng and Cowling, 2014, Schmidt et al., 2017]. The final decrease of $R_0(t)$ suggests later interventions, including border control and release of burial guides, may have been helpful in controlling the spread of disease.

The mean infection period for Sierra Leone epidemic is estimated to be 6.74 days with a 95% BCI (5.30,8.33). Liberia data yields a similar mean infection period estimate, with the point estimate of 6.21 and a 95% BCI of $(3.96, 7.91)$. Our Joint-based method yields a lower estimated value of the mean infection period than genealogy based method in Chapter 3. However, such inconsistency is not unrealistic since the actual infection period can be shortened due to quarantine and hospitalization, resulting in less time for patients to infect others. Moreover, according to equation $\beta = [R_0\gamma]/N$, a shorter infection period suggests a higher removal rate, leading to a higher estimate of infection rate and ending up with larger epidemic size. The final epidemic size in Sierra Leone has point estimate 17,769 with 95% BCI $(11572, 28738)$, which is close to 14,124 — the total suspected number of cases reported by Centers for Disease Control. Liberia has a smaller final epidemic size, with point estimate of

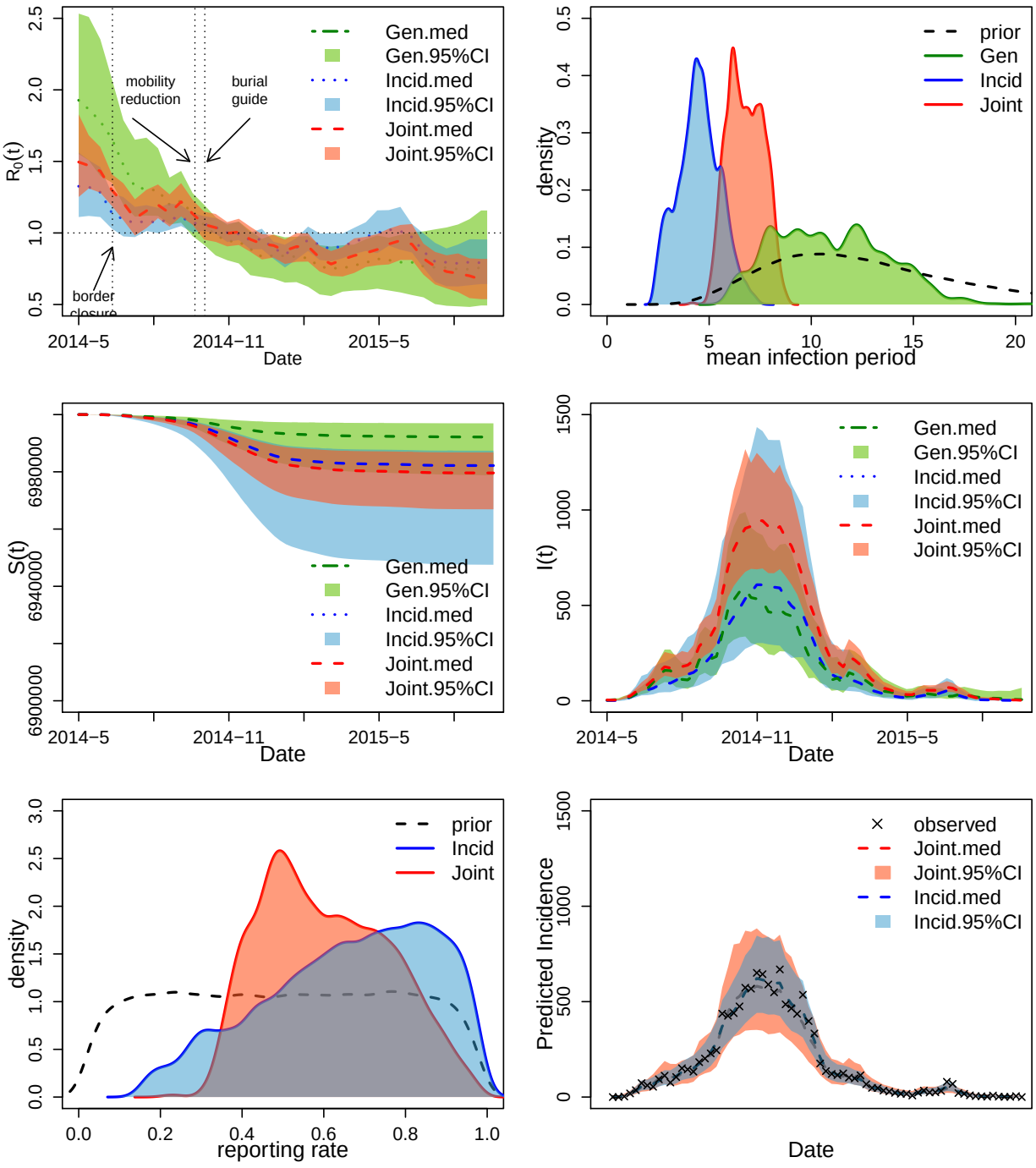

Figure 4.5: Analysis of the genealogy relating Ebola virus sequences collected in Sierra Leone. See caption in Figure 3.5 for the explanation of the plots.

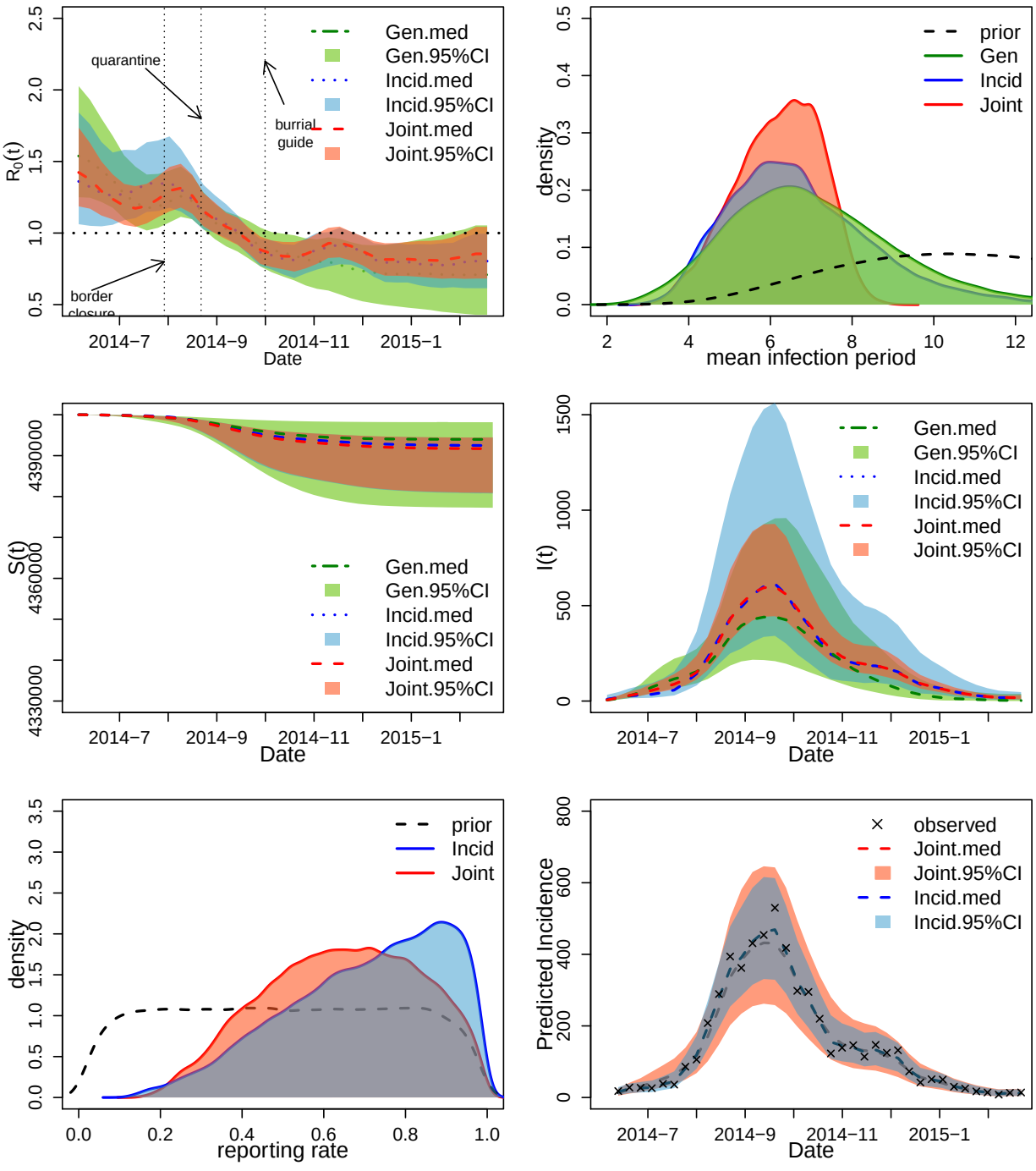

Figure 4.6: Analysis of the genealogy relating Ebola virus sequences collected in Liberia. See caption in Figure 3.5 for the explanation of the plots.

8,831 and 95% BCI $(5591, 19066)$, which is consistent with the 10,678 total suspected cases. The reporting probability for weekly incidence in Sierra Leone is estimated to be 0.59, and its 95% BCI is $(0.36, 0.91)$. Similarly, Liberia has reporting probability estimated to be 0.65 with 95% BCI to be $(0.29, 0.96)$. Our estimate is lower than the 83% reporting probability in [Scarpino et al., 2014]. Since recent studies [Glynn et al., 2017] suggest Ebola virus hosts can be asymptomatic, we expect a higher total epidemic size and lower reporting rate estimates may be more accurate than previous estimates.

We plot the posterior predictive incidence along with the observed weekly reported incidence in Sierra Leone and Liberia from World Health Organization [b] to exam the goodness-of-fit of our methods. The posterior predictive weekly incidence at time t_i, denoted by $\hat{Y}(t)$, is sampled using

$$\hat{Y}_i \sim \text{NegativeBinomial}\left(\mu = \hat{\rho}(\hat{S}_{i-1} - \hat{S}_i), \sigma^2 = \mu + \mu^2/\phi^2\right), \tag{4.11}$$

where $\hat{\rho}$, $\hat{\phi}$ and $\hat{S}_i$ are samples of the reporting rate, over-dispersion parameter and the number of infected individuals at time t_i respectively. For both countries, our model-based incidence 95% BCIs contain the reported incidence counts from WHO, suggesting that both Incid-based and Joint-based method fit the data well.

We also report results from the Gen-based and Incid-based methods, and superimpose these results over Joint-based results on Figures 4.5 and 4.6. For the relatively small Liberia genealogy, the Gen-based and Incid-based method methods yield similar parameter estimates as the Joint-based method. However, the larger Sierra Leone genealogy produces substantial differences, especially for the infection period, where a point estimate 4.5 days from the Incid-based method and 10 days from Gen-based method suggest that using any single source of information may encounter some identifiability issues. Generally, Gen-based method yields higher estimates of initial $R_0(t)$ and produces wider BCIs than the Incid-based method. The $R_0(t)$ estimate from Joint-based method lies in between the results if the Gen-based and Incid-based methods, but has the narrowest BCI widths. The Incid-based method provides larger estimated reporting rate than our Joint-based method, which has point es-

timate around 0.8 for each country. Somewhat counterintuitively, the Joint-based method has slightly wider posterior predictive credible intervals than the ones from the Incid-based method. The narrower posterior predictive credible interval from Incid-based method can be explained by the reporting rate estimation, which often has posterior mode close to 1.0 and result in less posterior predictive variance.

4.5.2 Sequentially updating parameters using streaming data

We implemented an algorithm for sequentially updating model parameters and used this algorithm to analyze Sierra Leone and Liberia data. Posterior distribution summaries obtained using subsets of the data depicted in Figure 4.7 and 4.8 respectively. The columns show the results obtained using incidences case counts and a genealogy constructed by pretending that we observed all data only up to a pre-specified time. For Sierra Leone data, we choose 2014-09-29, 2014-11-03, 2014-12-01, 2015-01-05 and 2015-02-02 as these times. For Liberia data, we take another four cutoff times: 2014-08-18, 2014-09-15, 2014-10-13 and 2014-11-17. The last row shows the reported incidence data before cutoff times, which give us a rough idea of the trend of the epidemic. The top two rows show the comparison between Joint-based method, Gen-based method and Incid-based method. For Sierra Leone and Liberia data, the Joint-based method quickly captures the decreasing trend of $R_0(t)$ in mid October and November, while the Gen-based method does not provide enough evidence about the decrease in the infection rate. The Incid-based method, though has similar prediction results, yields a wider BCI than the Joint-based method. The next two rows demonstrate the comparisons prevalence ($I(t)$) estimation, with the Joint-based and Incid-based methods successfully detecting the epidemic peak in early October using the observation before November and hte Incid-based method method producing wider BCIs and a larger final epidemic size estimate. However, the Gen-based method seems to identify a peak of infections before the true peak in Sierra Leone and fails to estimate the epidemic peak correctly even with more data in November in the Liberia case.

For both the Joint-based and Incid-based methods, based on the data observed before

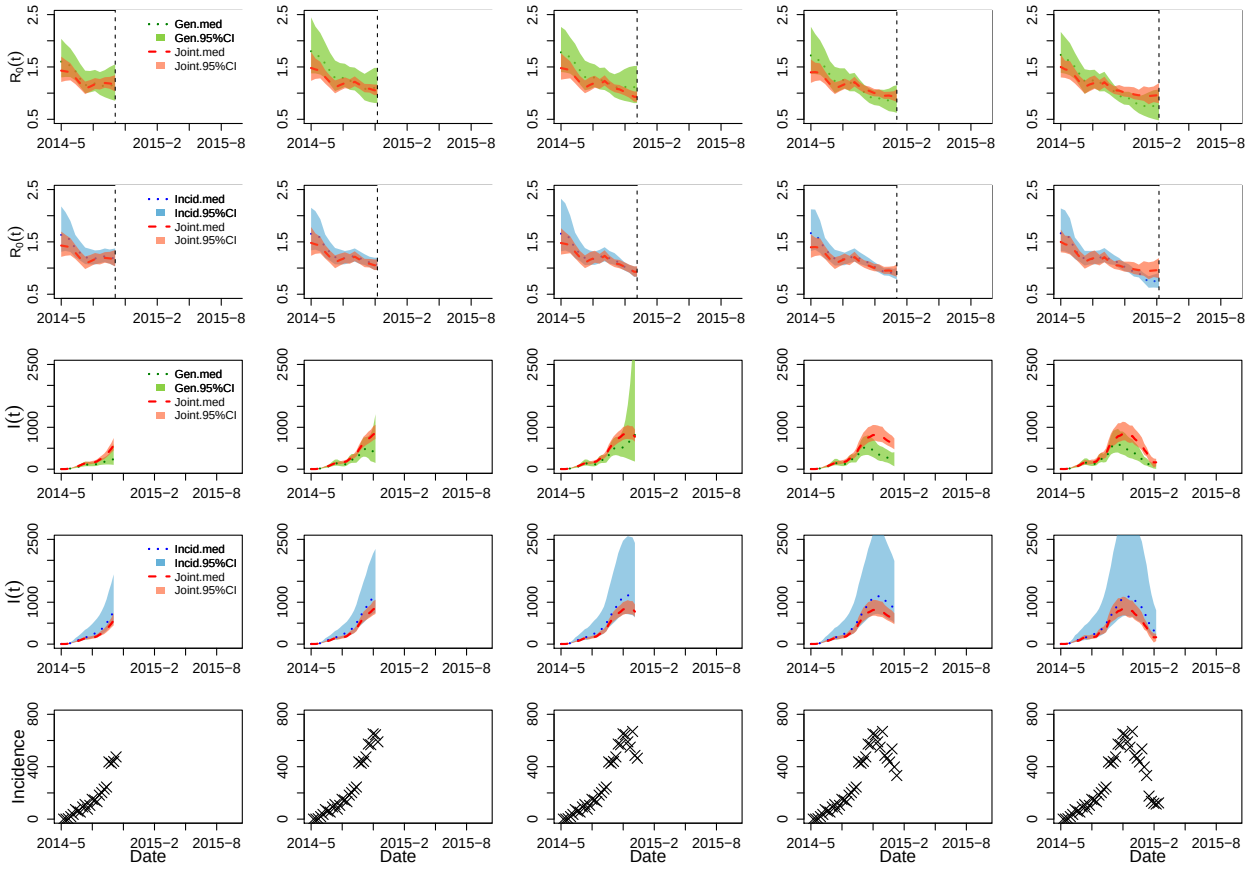

Figure 4.7: Analysis of sequentially updated parameters using genealogy and incidence data in Sierra Leone. The first two rows show the estimated $R_0(t)$ trajectories using the observations before the black dash lines, with the green shaded area showing the results for the Joint-based method, blue-shaded area for the Incid-based method and red shaded area for the Joint-based method. The third and forth rows show the posterior summaries of prevalence $I(t)$, using the same color legend as for $R_0(t)$.

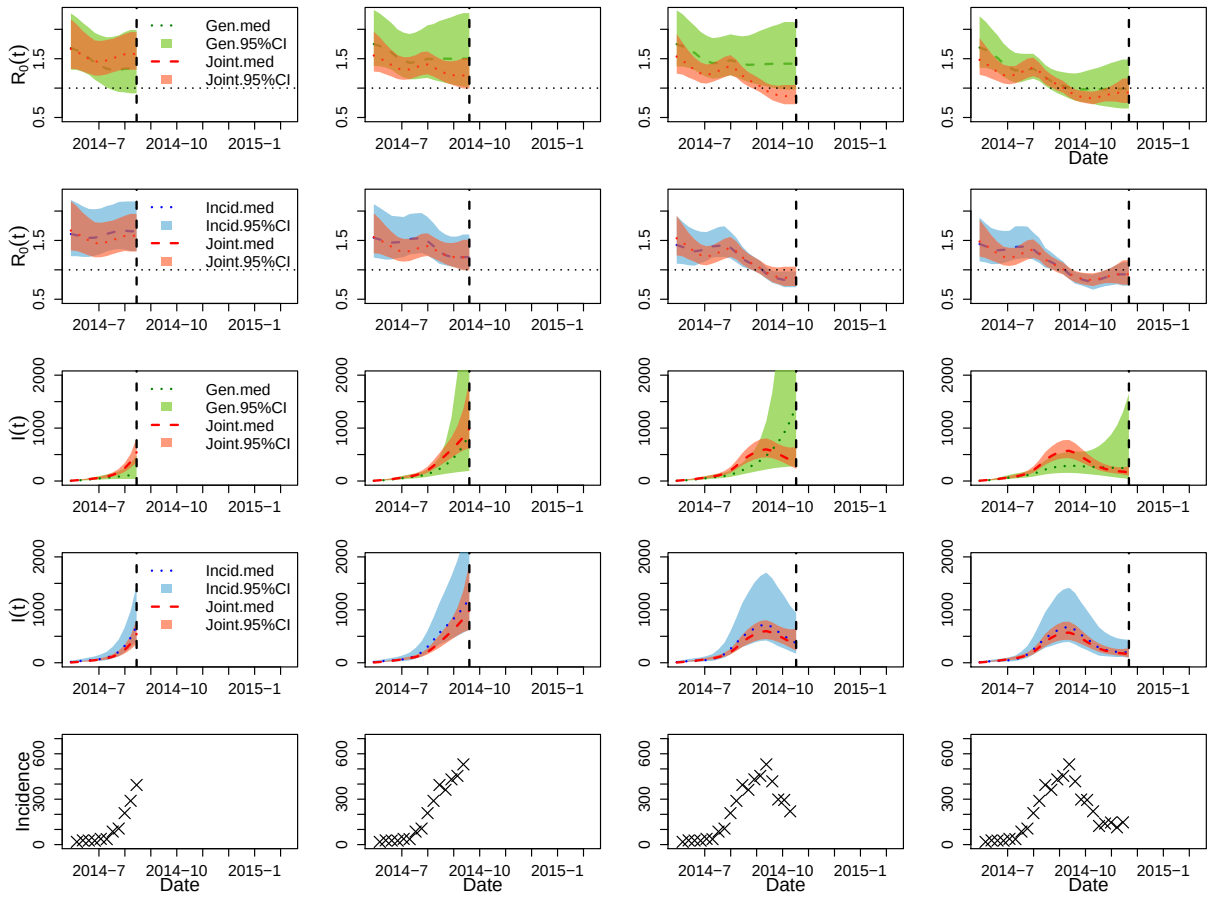

Figure 4.8: Analysis of sequentially updated parameters using genealogy and incidence data in Liberia. See caption in Figure 4.7 for the explanation of the plots.

	Sierra Leone		Liberia	
	Joint-based	Incid-based	Joint-based	Incid-based
1 wk ahead	**0.0428**	0.0405	**0.0554**	0.0451
2 wk ahead	0.0290	**0.0302**	**0.0357**	0.0238
3 wk ahead	0.0278	**0.0284**	**0.0279**	0.0207
4 wk ahead	**0.0308**	0.0294	**0.0199**	0.0099

Table 4.2: Logarithm scores of short term Ebola incidence forecast in Sierra Leone and Liberia. The rows represent the time period. The columns compare the results between Joint-based method and Incid-based method in Sierra Leone and Liberia.

each cutoff time points, we produce 1–4 week ahead incidence forecasts using Equations (4.8). The forecasting score is calculated using Equation (4.10). We report the score from 1 to 4 weeks in Table 4.2. For both countries, our Joint-based method outperforms the Incid-based method in most of the forecasting tasks. For each method, the score decreases as the prediction horizon becomes longer.

4.6 Discussion

In this Chapter, we extend our Bayesian stochastic SIR modeling framework to jointly using time series incidence and genealogy data for infectious disease dynamics inference and prediction. Computationally, our combination of LNA approximation and elliptical slice sampler help us devise an efficient MCMC algorithm for infectious disease state space model with latent SIR dynamics. Through simulation studies, we compare Incid-based, Gen-based, and Joint-based methods. Our Joint-based method turns out to resolve problems with prior sensitivity and identifiability of the removal rate that we have seen in Chapter 3 when using the Gen-based method. Moreover, the Joint-based method gives a better estimate of the reporting rate than the Incid-based method. By combining the two sources of information, the Joint-based method provides less biased and more precise estimation of the basic repro-

duction number and latent SIR population trajectory than the other two methods using a single data type. Finally, we successfully apply our method to incidence and genealogy data from the 2014-2015 Ebola outbreak in Sierra Leone and Liberia.

We demonstrate that our framework can be applied to streaming surveillance data to produce sequentially updated stochastic epidemic model parameters, which helps researchers understand changes in the disease dynamics and implement/evaluate interventions in a timely manner. When applying to streaming data, our Joint-based method shows the ability to quickly capture past disease dynamics and to provide a precise estimates of the reproduction number changes as well as the disease prevalence. Additionally, our methods can be used for a short-term disease forecasting, for example, predicting reported incidence in the next two weeks.

Similar to Chapter 3 developments, our Joint-based method takes a fixed estimated genealogy. In reality, the genealogies are inferred from sequence data with uncertainty and directly fitting stochastic epidemic models to an estimated genealogy may result in bias and overconfidence. Hence, a future direction is to directly fit stochastic epidemic models to sequence data, which can be achieved by adding a hierarchical layer of MCMC sampler that updates a genealogy based. It would be interesting to see if the Joint-based method still outperforms the Incid-based method after accounting for genealogical uncertainty.

Our current model assumes the epidemic follows a SIR dynamics and we can successfully fit this model to genealogy data and incidence data with the help of SIR-structured coalescent model and reparametrization based on Equation (4.1). A possible future direction is to generalize the Joint-based inference framework to arbitrary stochastic epidemic models, for example, an epidemic model with multi-stage infections or models accounting for population structure. Though the coalescent likelihood may not have a closed from solution, as in the SIR case, the structured coalescent method in Section 2.3.2 provides an approximate coalescent likelihood. In terms of the incidence data, some stochastic models may not have one-to-one transformation between the population trajectory and cumulative events as in Equation (4.1). Fortunately, Fintzi [2019] provides a way to reparameterize the population

trajectory with cumulative reaction events and use log-transform via Itô's formula to ensure a non-decreasing cumulative events trajectory. Through such a reparameterization, one can easily plugin a close-form incidence observation likelihood into the LNA framework.

Chapter 5

PREFERENTIAL SAMPLING IN MECHANISTIC MODELS OF PHYLODYNAMICS

5.1 Background

In phylodynamics inference, the disease agent sequence data are collected at different times during the disease outbreak. Such data collecting approach is called heterochronous or serial sampling [Rodrigo and Felsenstein, 1999]. When analyzing heterochornous sampled sequences, researchers implicitly assume the sampling times are either fixed or functionally independent of the disease population dynamics. However, in many of the cases, epidemiologists may unintentionally collect more samples when the disease prevalence is high and collect data less frequently during the time when fewer infected individuals are present. The dependency between sampling time distribution and population dynamics is called *preferential sampling*. This issue was addressed in [Volz and Frost, 2014, Karcher et al., 2016], where the authors found that ignoring the probabilistic dependence between population dynamics and the temporal frequency of collecting data samples can lead to biased estimation of effective population size.

Within a nonparametric phylodynamic inference setup, Karcher et al. [2016] found that sampling protocol implicitly depending on the effective population size can cause model misspecification bias in the model that does not take preferential sampling into account. In addition to reducing bias, incorporating such dependency into the phylodynamic model increases inference precision. The framework was later extended by incorporating additional predictors into the model and also equipped with an additional hierarchical sampling layer to integrate over genealogies while performing inference based directly on sequence data [Karcher, 2019]. However, nonparametric results are hard to interpret since the effec-

tive population size is not equivalent to the true disease prevalence. Volz and Frost [2014] proposed a preferential sampling framework by associating the sampling intensity with the birth rate in a deterministic birth-and-death disease model. Although the idea of preferential sampling has been applied in nonparametric models and deterministic ODE models, to our best knowledge, none of the current models connect preferential sampling with stochastic epidemic models under structured coalescent phylodynamic model framework.

Inspired by the idea of fitting stochastic epidemic model to genealogy data in Chapter 3, here, we propose a preferential sampling model based on a stochastic SIR model under structured coalescent model. The model assumes an explicit probabilistic connection between the sampling intensity and the prevalence. We perform several simulation studies in Section 5.3 to test the effect of preferential sampling on statistical inference under the stochastic SIR model. Our experiments show that adding preferential sampling is helpful in reducing estimation bias in infection rate and population trajectories and also increasing the precision of the phylodynamic inference. Finally, we apply our preferential sampling based inference to the Ebola genealogy estimated from sequences collected in Liberia during the 2014-2015 outbreak.

5.2 Methodology

5.2.1 SIR structured coalescent

We start with n molecular sequences from disease hosts sampled at m different times $s_1 < \cdots < s_m$, with n_i sequences obtained at s_i, for $i = 1, \ldots, m$. A genealogy $\mathbf{g}$ relating to those sequences has been estimated and treated as our observation. The genealogy has a tree structure containing two sources of temporal information: coalescent times and sampling times. Let $c_1 < \cdots < c_{n-1}$ denote $n-1$ coalescent times, corresponding to the times at which there's a branching event on the tree. The sampling times, denoted by $\mathcal{S} = \{s_1, \ldots, s_m\}$, are the times at which the sequence data are obtained, corresponding to the tips of the genealogy. More details about the genealogy is given in Section 2.3.

Recall that in Section 2.3.3, the SIR structured coalescent model divides the population into three subgroups: susceptible, infected and removed. The coalescent rate SIR structured coalescent likelihood is a function of the number of susceptible $S(t)$, number of infected $I(t)$ and the infection rate $\beta(t)$. The SIR structured coalescent likelihood, according to (2.64), is formulated as

$$\Pr(\mathbf{g} \mid \mathbf{X}_{0:T}, \boldsymbol{\theta}_{0:T}, \mathcal{S}) \propto \prod_{k=2}^{n} \binom{l(c_{k-1})}{2} \frac{2\beta(c_{k-1})S(c_{k-1})}{I(c_{k-1})} \exp\left(- \sum_{i=0}^{i_k-1} \int_{\mathcal{I}_{i,k}} \binom{l_{i,k}}{2} \frac{2\beta(\tau)S(\tau)}{I(\tau)} \mathrm{d}\tau \right) \quad (5.1)$$

where $l(t)$ is the piecewise constant function for the number of lineages presents at time t, see Section 2.3 for more details. The above coalescent likelihood (5.1)is the equivalent to the coalescent likelihoods in Equations (3.2) and (4.4) in previous chapters despite an additional $\mathcal{S}$ that appears on the right of the conditioning sign. The only difference that now we condition sampling times explicitly rather than implicitly.

Furthermore, as before, we model the population trajectory $\mathbf{X}(t) = (S(t), I(t))$ and infection rate $\beta(t)$ as piecewise constant functions on a regular grid $t_0 < \cdots < t_T$ such that $t_0 < c_1$ and $t_T > s_m$:

$$\mathbf{X}(t) = \sum_{i=1}^{T} \mathbf{1}_{[t_{i-1}, t_i)}(t)\mathbf{X}_i, \quad \boldsymbol{\theta}(t) = \sum_{i=1}^{T} \mathbf{1}_{[t_{i-1}, t_i)}(t)\boldsymbol{\theta}_i,$$

where $\mathbf{X}_i$ and $\boldsymbol{\theta}_i$ denote the population compartment sizes and the SIR rates at time t_i. Piecewise constant functions above make the integral in (5.1) tractable, so the structured coalescent likelihood is available in closed form.

5.2.2 SIR dynamics

As in Section 3.2.2, we assume the population is well-mixed and the disease dynamics follow a SIR model, with a time-varying infection rate $\beta(t)$ and a constant removal rate γ. The population trajectory $\mathbf{X}(t)$ is assumed to follow Markov property and the prior desnity of $\mathrm{d}\mathbf{X}_{1:T}$, given initial value state $\mathbf{X}_0$, can be written as

$$\Pr\left(\mathbf{X}_{1:T} \mid \mathbf{X}_0, \boldsymbol{\theta}_{0:T}\right) = \prod_{i=1}^{T} \Pr\left(\mathbf{X}_i \mid \mathbf{X}_{i-1}, \boldsymbol{\theta}_{i-1}\right), \quad (5.2)$$

where $\Pr(\mathbf{X}_i \mid \mathbf{X}_{i-1}, \boldsymbol{\theta}_{i-1})$ is the transition density of the SIR model. As in Section 3.2.2, the intractable SIR transition density will be approximated using a Gaussian distribution (3.11) using the LNA. Hence, the closed-form transition densities $\Pr(\mathbf{X}_i \mid \mathbf{X}_{i-1}, \boldsymbol{\theta}_{i-1})$ enable us to calculate the trajectory likelihood and allows us to perform standard MCMC approaches and bypass the likelihood-free based method like not particle filter or approximate Bayesian computation to obtain posterior estimates.

5.2.3 Preferential sampling model

In previous chapters, the sampling times are treated as fixed and its dependency on the population dynamic is ignored during the inference. In this section, we relax this assumption by modeling the sampling times as a an inhomogeneous point process in a fixed sampling window $[s_0, s_m]$. Let $\mu(t)$ denote the rate of sampling molecular sequences (sampling intensity) at time t. We assume the sampling intensity $\mu(t)$ to be proportional to the corresponding infected population (prevalence) at time t, $I(t)$, in the following log-linear fashion:

$$\log(\mu(t)) = \kappa_0 + \kappa_1 \cdot \log(I(t) + 1), \tag{5.3}$$

where the intercept κ_0 is the baseline of log intensity and the slope κ_1 is interpreted as the preferential sampling power that controls the strength of dependency between the sampling intensity and disease prevalence. Sampling with $\kappa_1 = 1$ will result in collecting genetic sequences directly proportional to the prevalence, while higher κ_1 values will result in more clustered in time samples. Another special case is when $\kappa_1 = 0$, where the sampling times of molecular sequences do not depend on population dynamics and instead follow a homogeneous Poission process with rate κ_0.

The probability density function for the sampling times given preferential sampling coefficients $\boldsymbol{\kappa} = (\kappa_0, \kappa_1)$ is

$$\Pr(\mathcal{S} \mid \mathbf{X}_{0:T}, \boldsymbol{\kappa}) = \exp\left(n\kappa_0 + \sum_{j=1}^{n} \kappa_1 \log\left(I(s_j) + 1\right) - \sum_{j=1}^{n} \int_{s_{j-1}}^{s_j} e^{\kappa_0} \left(I(\tau) + 1\right)^{\kappa_1} \mathrm{d}\tau \right). \tag{5.4}$$

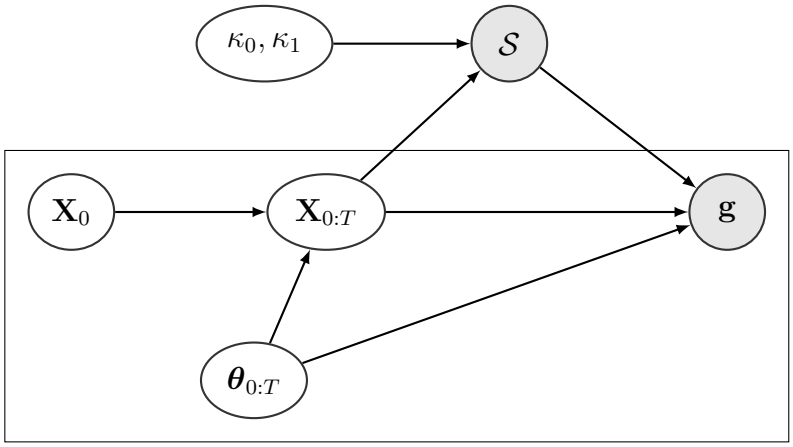

Figure 5.1: Variable dependency graph in preferential sampling model.

Since the prevalence trajectory $I(t)$ is piecewise constant, the integral in Equation (5.4) is available in closed-form.

5.2.4 Posterior approximation via MCMC

The coalescent likelihood conditions on the sampling times $\mathcal{S}$ and population dynamic $\mathbf{X}_{0:T}$. The sampling times, treated as random, depend on the population dynamic $\mathbf{X}_{0:T}$ and the coefficients κ_0, κ_1 through the sampling density (5.4). As in Section (3.3), we use $\pi(\boldsymbol{\theta}_{0:T})$ and $\pi(\mathbf{X}_0)$ to denote priors of SIR rates and initial population counts respectively. Additionally, we assign two independent normally distributed priors $\mathcal{N}(u_i, v_i)$ to the preferential sampling coefficients κ_i, for $i = 0, 1$. The parameter dependency relationship is given in Figure 5.1. The posterior distribution of all unknown quantities can be written as

$$\begin{aligned}\Pr(\mathbf{X}_{0:T}, \boldsymbol{\theta}_{0:T}, \boldsymbol{\kappa} \mid \mathbf{g}, \mathcal{S}) \propto & \underbrace{\Pr(\mathbf{g} \mid \mathbf{X}_{0:T}, \boldsymbol{\theta}_{0:T}, \mathcal{S})}_{\text{coalescent likelihood}} \cdot \underbrace{\Pr(\mathcal{S} \mid \mathbf{X}_{0:T}, \boldsymbol{\kappa})}_{\text{sampling likelihood}} \cdot \\ & \underbrace{\Pr\left(\mathbf{X}_{1:T} \mid \boldsymbol{\theta}_{0:T}, \mathbf{X}_0\right)}_{\text{trajectory likelihood}} \cdot \underbrace{\pi(\mathbf{X}_0)\pi(\boldsymbol{\theta}_{0:T})\pi(\boldsymbol{\kappa})}_{\text{priors}}.\end{aligned} \tag{5.5}$$

For computation convenience, like in Section 3.2.3, we reparameterize the SIR trajectory with independent normally distributed random variables $\boldsymbol{\xi}_{1:T}$ using the non-centered repa-

rameterization [Papaspiliopoulos et al., 2007]. The SIR rates $\boldsymbol{\theta}(t)$ are replaced with the removal rate γ and a GMRF representation of the basic reproduction number trajectory $R_0(t)$. Additionally, $R_0(t)$ is further reparameterized with a initial R_0, a vector of independent standard normal random variables $\boldsymbol{\delta}_{1:T}$, and a hyperparameters σ that controls the smoothness of GMRF. The initial R_0 is assigned a lognormal(a_1, b_1) prior. We use a lognormal(a_2, b_2) prior the inverse of standard deviation $1/\sigma$. More details of the reparameterization is provided in Section 3.2.3.

Figure 5.2 shows the variable dependency relationships after the reparameterization. The posterior distribution of new parameters is

$$\begin{aligned}\Pr(I_0, R_0, \gamma, \boldsymbol{\delta}_{1:T}, \boldsymbol{\xi}_{1:T}, \sigma, \boldsymbol{\kappa} \mid \mathbf{g}, \mathcal{S}) \propto & \underbrace{\Pr(\mathbf{g} \mid I_0, R_0, \gamma, \boldsymbol{\delta}_{1:T}, \boldsymbol{\xi}_{1:T}, \sigma, \mathcal{S})}_{\Pr(\mathbf{g}|\mathbf{X}_{0:T}, \boldsymbol{\theta}_{0:T}, \mathcal{S})} \\ & \underbrace{\Pr(\mathcal{S} \mid I_0, R_0, \gamma, \boldsymbol{\delta}_{1:T}, \boldsymbol{\xi}_{1:T}, \sigma, \boldsymbol{\kappa})}_{\Pr(\mathcal{S}|\mathbf{X}_{0:T}, \boldsymbol{\kappa})} \\ & \pi(I_0)\pi(R_0)\pi(\gamma)\pi(\boldsymbol{\delta}_{1:T})\pi(\boldsymbol{\xi}_{1:T})\pi(\sigma)\pi(\boldsymbol{\kappa}).\end{aligned}$$

We use the following MCMC with block updates to approximate this posterior distribution. We update high dimensional vector $\mathbf{U} = (\log(R_0), \boldsymbol{\delta}_{1:T}, \log(\sigma))$ using the efficient elliptical slice sampler [Murray et al., 2010]. Vector $\boldsymbol{\xi}_{1:T}$ is updated the same way in a separate step. Initial number of infected individuals I_0, removal rate γ, and preferential sampling coefficients $\boldsymbol{\kappa}$ are updated using univariate Metropolis steps.

5.3 Simulation studies

5.3.1 Simulations based on single genealogy realizations

In this section, we use simulated genealogies to evaluate the performance of our preferential sampling (Pref-based) method and to compare it with a baseline model Gen-based (LNA-based) method from Chapter 3 that uses only coalescent times an no other source of information. The two methods are compared in the context of a time-varying $R_0(t)$ and we use the same MCMC algorithm for both models. A more detailed description of the Gen-based (LNA-based) MCMC algorithm is given in Section 2.2.7 .

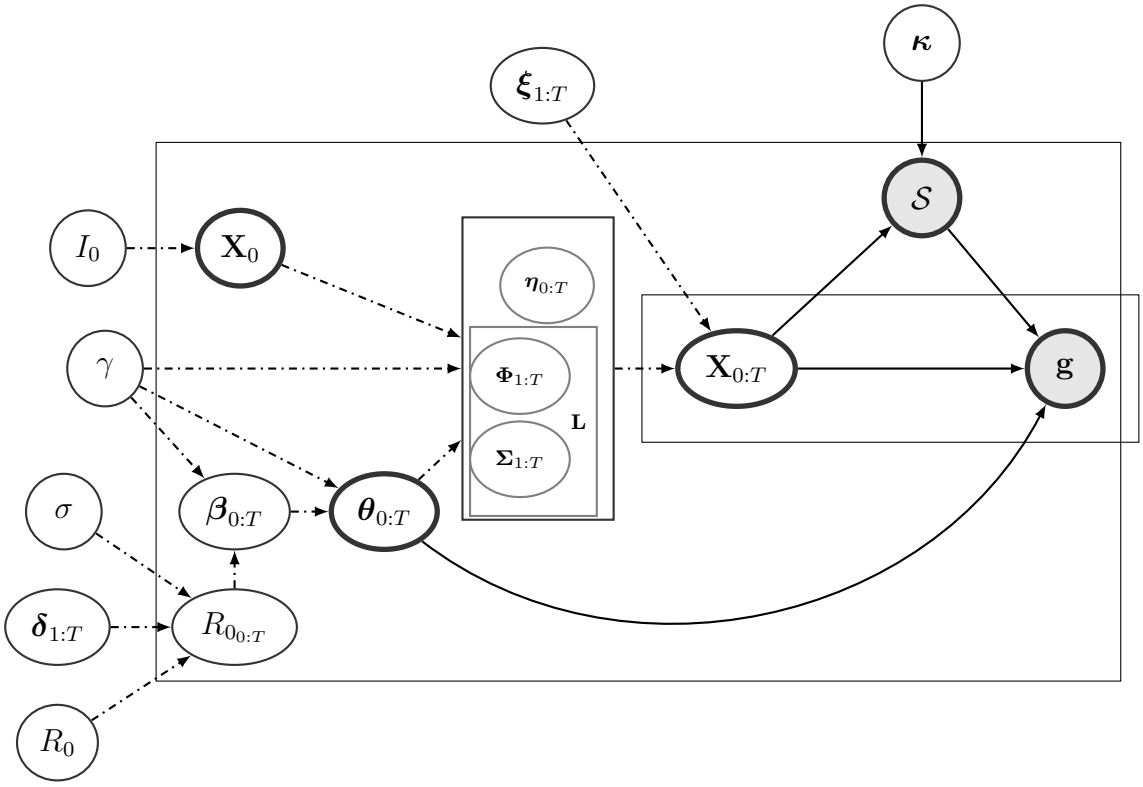

Figure 5.2: Parameter dependency graph after reparameterization. The root nodes $I_0, \gamma, \sigma, \boldsymbol{\delta}_{1:T}, R_0, \sigma$ outside the large box are parameters and latent variables after reparameterization, for which we assign prior distributions. The dash dotted lines show deterministic relationships and the solid lines show the stochastic dependencies. The grey nodes denote the observed data, which — the genealogy $\mathbf{g}$ and corresponding sampling times $\mathcal{S}$. The figure also shows the dependency structure between the transformed parameters and original parameters $\boldsymbol{\theta}_{0:T}, \mathbf{X}_0$ and $\mathbf{X}_{0:T}$.

The simulation protocol is similar to the one described in Section 3.3.1, where we first simulate one realization of the SIR population trajectory with pre-specified parameters γ, I_0, and $R_0(t)$. Next, we use a given set of preferential sampling coefficients $\boldsymbol{\kappa}$ to simulate Poisson-distributed lineage sampling times. Finally, the coalescent times are generated from the distribution specified by density (3.2) using a thinning algorithm by Palacios and Minin [2013].

The two methods are compared under two "true" $R_0(t)$ scenarios over the time interval $[0, 90]$:

1. Exponential decreasing (ED) $R_0(t)$. $R_0(t) = 1.8 \times 0.975^{t/2}$ for $t \in [0, 90]$. Recovery rate is set to $\gamma = 0.35$. Initial counts of infected individuals is $I_0 = 3$. Total population size is $N = 1{,}000{,}000$. Preferential sampling coefficients are set to $\kappa_0 = -2$, $\kappa_1 = 0.8$.

2. Non-monotonic (NM) $R_0(t)$. $R_0(t) = 1.4 \times 1.015^{0.5t}, t \in [0, 30]$, $R_0(t) = 1.750 \times 0.975^{t-30}, t \in [30, 80]$ and $R_0(t) = 0.4583, t \in [80, 90]$. Recovery rate is set to $\gamma = 0.3$. Initial counts of infected individuals is set to $I_0 = 3$. Total population size is set o $N = 1{,}000{,}000$. Preferential sampling coefficients are set to $\kappa_0 = -2$, $\kappa_1 = 0.8$.

For all simulations, we use lognormal$(1, 1)$ prior for I_0. The parameters of the lognormal priors for the initial R_0 and inverse standard deviation $1/\sigma$ are set to $a_1 = 0.7, b_1 = 0.3$ and $a_2 = 3, b_2 = 0.2$ respectively. As in Chapter 4, we are interested in whether incorporating information about sampling times helps resolve sensitivity to the choice of prior for the removal rate. Hence, we put weakly a informative prior on γ in each simulation scenario: (1) ED: $\gamma \sim$ lognormal$(-1.1, 0.25)$, (2) NM: $\gamma \sim$ lognormal$(-1.2, 0.2)$. We set the grid size to $T = 36$, with $t_i - t_{i-1} = 2.5$ for $i = 1, \ldots, 36$.

The first row of Figure 3.3 shows pointwise posterior medians and 95% BCIs for the basic reproduction number trajectory, $R_0(t)$. The Pref-based method performs well in capturing the continuous dynamics of $R_0(t)$ in all scenarios. The Gen-based method provides similar $R_0(t)$ estimates at the beginning of and in the middle of the epidemic. However, overesti-

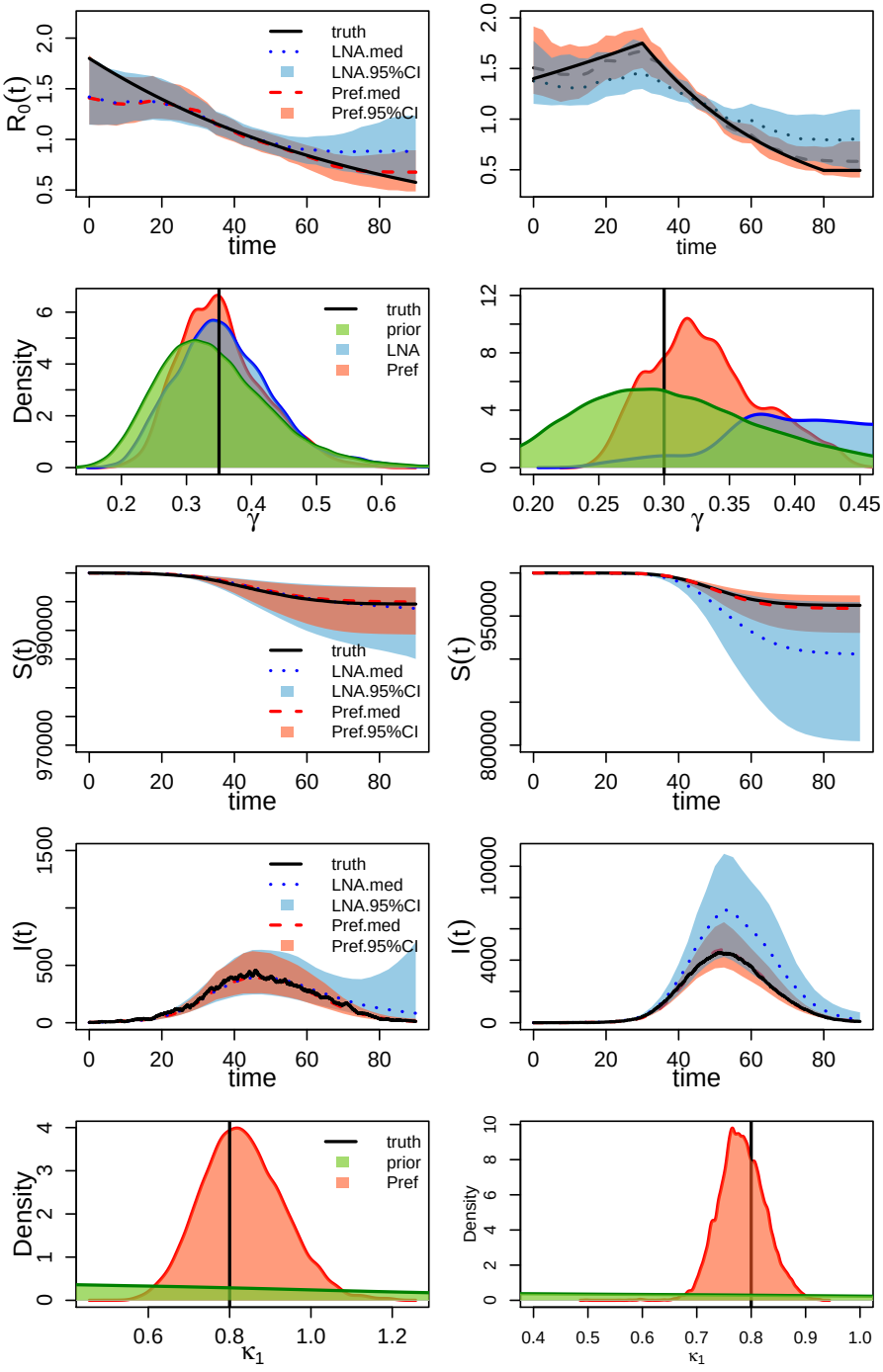

Figure 5.3: Analysis of 2 simulation scenarios. Columns correspond to EXPD and NM simulated $R_0(t)$ trajectories. The first row shows the estimated $R_0(t)$ trajectories for the 2 scenarios, with the black solid lines representing the truth, the red dashed lines depicting the posterior medians and the red-shaded area showing the 95% BCIs for the Pref-based method. For the Gen-based method, the posterior medians are plotted in blue dotted lines, with blue shading showing the 95% BCIs. The second row corresponds to the estimation for the removal rate γ. Posterior density curves from the preferential sampling model are shown in red lines and the posterior density for LNA is plotted in blue lines, with prior densities shown in green. The bottom two figures shows the estimated trajectories of $S(t)$ and $I(t)$ respectively.

mated $R_0(t)$ toward the end of the epidemic suggests that the Gen-based method may have some bias when estimating the decrease of the epidemic.

The second row in Figure 3.3 shows posterior summaries of removal rate γ. Both Pref-based and Gen-based methods yield posterior modes centered near the true value, with the Pref-based method yielding slightly higher posterior densities at the truth than the Gen-based method and the prior density. In the NM scenario, the Pref-based method still has higher density centered at the truth when compared with the prior, while the Gen-based method has posterior distribution shifted to right of the truth.

Posterior summaries of $S(t)$ and $I(t)$ are given in the third and fourth rows of Figure 3.3. Both methods successfully recover population dynamics at the beginning of epidemic, with the Pref-based method generating narrower credible interval for the population trajectories. However, while the Pref-based method manages to recover the latent decreasing prevalence trend at the end of the epidemic, the Gen-based method overestimates prevalence, with the BCIs failing to contain the truth.

The last rows demonstrates the posterior summaries of the preferential sampling rate κ_1 for Pref-based method. For both scenarios, we can see the intensity parameter κ_1 has been successfully estimated with significantly higher density centered at the truth than the prior.

5.3.2 Frequentist properties of posterior summaries

Similarly to previous chapters, we design a simulation study based on repeatedly simulating SIR trajectories using MJP with pre-specified parameters. The simulations are based on the exponential decreasing and non-monotonic scenarios in Section 5.3.1. Like in the simulation setting in Section 3.3.2, we reject "unreasonable" realizations with unrealistically low and high prevalence to arrive at 100 simulated SIR trajectories [1]. For each simulated SIR trajectory, we first generate a realization of sampling times under preferential sampling. Then a realization of the genealogy is simulated using the structured coalescent process. We

[1]See appendix for details of the rejection criterion

use both Pref-based and Gen-based methods to arrive at the posterior distribution of the parameters and latent variables for each simulated genealogy.

We use three metrics to evaluate models based on their estimates of $R_0(t)$ and $I(t)$: average error of point estimates, width of credible intervals, and frequentist coverage of credible intervals. The three evaluation metrics for the basic reproduction number trajectory $R_0(t)$ estimation are MAE, MCIW and ENV. In terms of the prevalence $I(t)$, we assess the accuracy, precision and coverage based on MRAE, MRCIW and ENV[2]. The posterior summary results for the ED and NM $R_0(t)$ scenarios are shown in Figure 5.4 and Figure 5.5 respectively.

Sampling distribution boxplots of $R_0(t)$ posterior summaries are depicted in the left three plots of Figure 5.4 and Figure 5.5. For both ED and NM scenario, the Pref-based method yields significantly lower MAE than the Gen-based method. In addition, the MCIWs produced by the Pref-method are generally lower than the ones from the Gen-based method. As for the envelope, both methods yield similarly high R_0-ENV for ED scenario. However, when it comes to the more complicated NM $R_0(t)$ trajectory, the Gen-based method loses some coverage while the Pref-based model still have ENV close to 1.0.

Sampling distribution boxplots of $I(t)$ posterior summaries, shown in Figure 5.4 and 5.5, are similar to the $R_0(t)$ results, with the Pref-based method generally having lower MRAEs, lower MRCIWs, and better coverage/envelope than the Gen-based method.

We also report the absolute error (AE) and 95% BCI widths for the removal rate γ in the last two plots of Figures 5.4 and 5.5. We note that a weakly informative prior has been chosen for γ to study the identifiability issue under preferential sampling. Like in the sensitivity analysis results from Appendix Section B.2.2, the Gen-based method has large bias and yields wide BCIs. After taking into account the information provided by the sampling times, the Pref-based method yields asignificantly lower AE and BCI widths than the Gen-based method. This suggests that the Pref-based method is helpful in addressing removal rate lack

[2]See Section 3.3.2 for more details on the definitions of evaluation metrics.

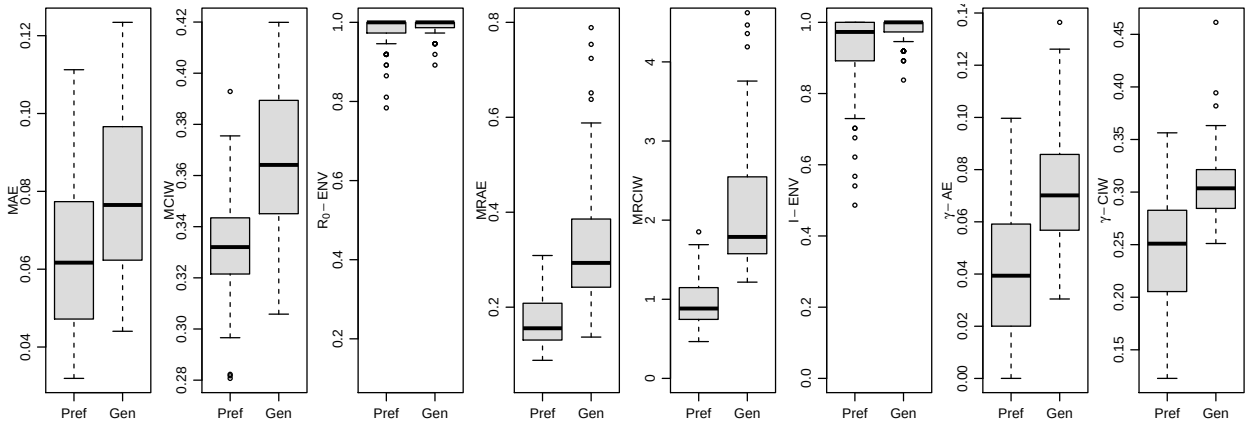

Figure 5.4: Boxplots comparing performance of preferential sampling-based and genealogy-based methods using 100 simulated genealogies under under exponential decay (ED) $R_0(t)$ trajectory scenario. The first three plots show mean absolute error (MAE), mean credible interval width (MCIW) and envelope for the $R_0(t)$ trajectory. The next three plots depict mean relative absolute error (MRAE), mean relative credible interval width (MRCIW), and envelope for $I(t)$(prevalence) trajectory. The last two plots show the absolute error (AE) and Bayesian credible interval (BCI) width for γ.

of identifiability from the genealogy alone.

5.4 Preferential sampling during Ebola outbreak in Liberia

We apply our Pref-based method to the Liberia Ebola genealogy in the 2014–2015 West Africa epidemic that we used in previous chapters. The Liberia genealogy, shown in the top left plot of Figure 5.6, was estimated from 205 Ebola virus full genomes sampled from 2014-06-20 to 2015-02-14 by Dudas et al. [2017].

When analyzing the genealogy, we assume the population in Liberia to be well-mixed and the disease spread according to the SIR dynamics with a time-varying reproduction number. The population size is specified based on the census population size in 2014 for Liberia, with $N = 4{,}400{,}000$. As in our simulation study, we use the lognormal prior for R_0

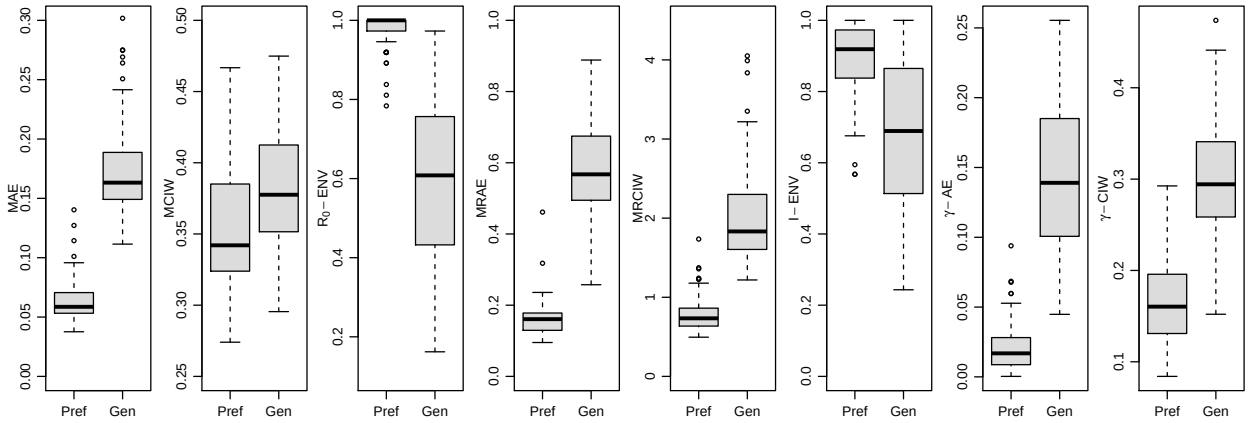

Figure 5.5: Boxplots comparing performance of preferential sampling-based and genealogy-based methods using 100 simulated genealogies under NM $R_0(t)$ trajectory scenario. See caption in Figure 5.4 for the explanation of the plots.

with $a_1 = 0.7$ and $b_1 = 0.3$ and the lognormal prior for the inverse standard deviation $1/\sigma$ with $a_2 = 3, b_2 = 0.2$. The prior distribution for the preferential sampling coefficients are set to be normal, with $u_0 = 0$, $v_0 = 2$ for intercept κ_0 and $u_1 = 0$, $v_1 = 1$ for preferential sampling power κ_1. For the removal rate γ, we relax the informative prior in Chapter 3 and choose a weakly informative lognormal prior with exponential scale mean 3.4 and variance 0.4. Parameter $1/\gamma$, interpreted as the length of the infection period, is expected to be 8-18 days for each country *a priori*. The total time span of the Liberia genealogy is divided evenly into 37 intervals from 2014-06-01 to 2015-02-15, which results in grid interval lengths, Δt_is, to be 7 days.

We run the MCMC algorithm described in Section 3.2.3 for 1,000,000 iterations and obtain posterior samples by discarding the first 100,000 iterations and saving every 20th iteration afterward. The trace plots in Section B.1.3 of the Appendix indicate the MCMC algorithm has converged and achieved good mixing.

Figure 3.6 show results of analyzing the Liberia data, with intervention events mapped onto the calendar time on the x-axis. Our Pref-based method estimates the initial R_0 in

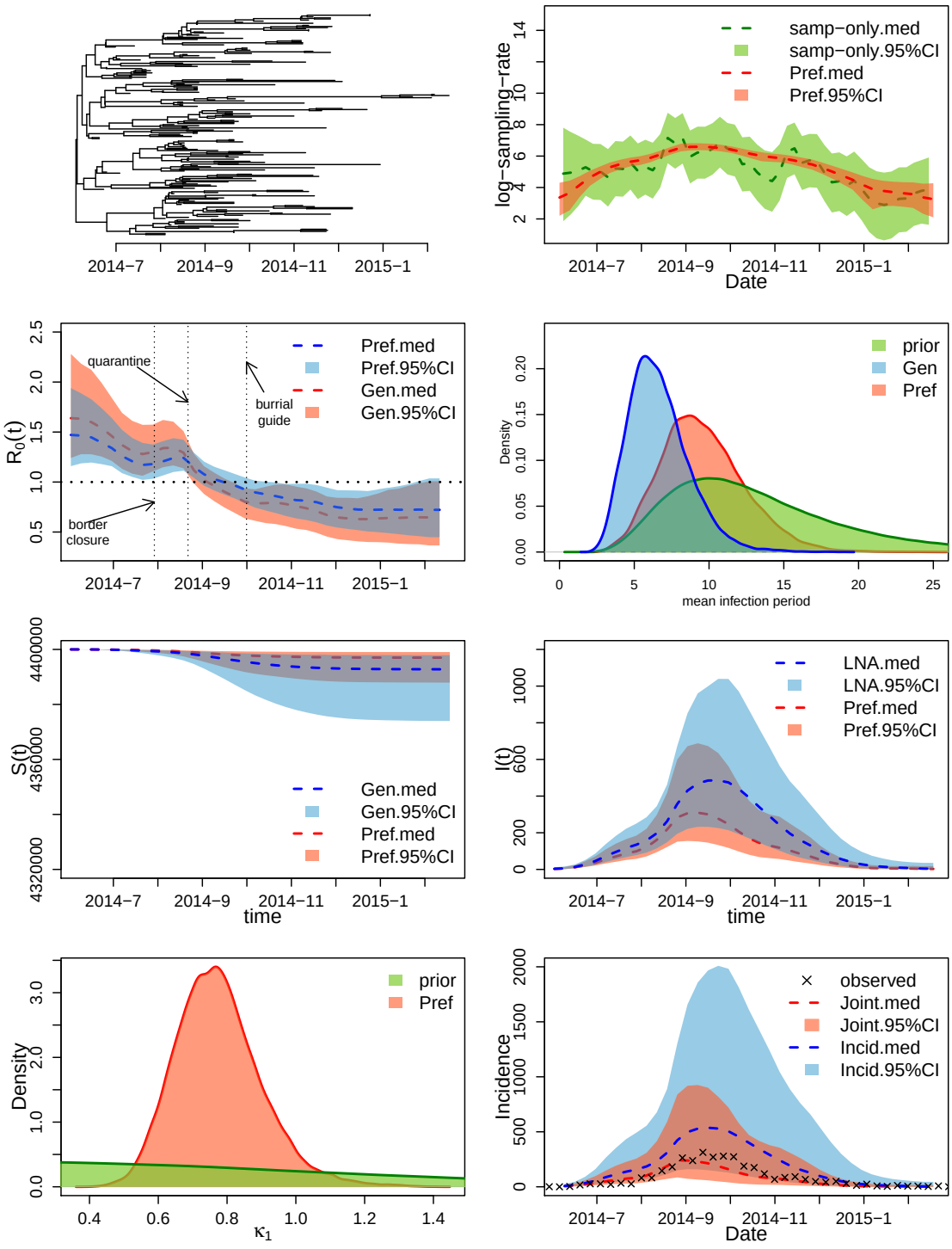

Figure 5.6: Analysis of the genealogy relating Ebola virus sequences collected in Liberia.

Liberia during 2014–2015 to be 1.64, with 95% BCI of $[1.24, 2.28]$. Our estimation is consistent with the initial R_0 results in Chapter 5 that relies on the informative prior of the removal rate. The estimation is also in agreement with results of Althaus [2014], who fitted a SEIR model to incidence data and arrived at an estimated R_0 of 1.59 (CI: $(1.57, 1.60)$).

The infection period in Liberia has a point estimate of 9.08, with a 95% BCI $(4.75, 15.27)$. Although we are using preferential sampling a weakly informative prior, the result is similar as that from Chapter 3, which estimate infection period to be 9.8 days (95% BCI: (6.87, 14.05)) under informative prior. estimated total infected individuals being 3000 and a 95% BCI of $(958, 12070)$. These results are in agreement with 3,163 total confirmed cases from WHO and the BCI also converges the 4,994 total cases (probable + confirmed).

We report the posterior density of the preferential sampling power κ_1 in the bottom left plot of Figure 5.6. The preferential sampling power for Liberia is estimated to be 0.77, with 95% BCIs to be $[0.56, 1.06]$, which indicates presence of preferential sampling in Liberia during the sequence collecting stage. Our estimate is also consistent with the estimated preferential sampling power by Karcher [2019], who fitted a non-parametric preferential sampling model directly to the same 205 sequence data and estimated the sampling power to be 0.78 (BCI: $(0.53, 1.2)$).

To check the adequacy of our preferential sampling model, we compare the posterior distributions of sampling intensities obtained via our Pref-based method with a nonparametric INLA-based estimate of the sampling rate (without information about coalescent times) in the top right plot of Figure 5.6. Our Pref-based method has sampling intensity estimation agreeing with the nonparametric result, indicating that the sampling time distribution can be well explained by out preferential sampling model. Our Pref-based method producing a smoother and narrower estimate of the sampling intensity due to incorporating additional information from the coalescent times.

We perform an out-of-sample validation by comparing our results with weekly reported confirmed incidence in Liberia from the World Health Organization [b] (WHO). Since our time grid is in line with the incidence reporting date in WHO, the weekly incidence in our

SIR model can be directly estimated using the transform

$$Z_i = S_{i-1} - S_i.$$

We plot the posterior estimates of weekly incidence together with the corresponding weekly reported confirmed incidence. Our model-based incidence 95% BCIs cover the reported incidence counts from WHO, suggesting that our time varying SIR model can estimate incidence well from genetic data alone.

We also report results from the Gen-based method and superimpose these results over Pref-based results on 5.6. The Gen-based method yields higher $R_0(t)$ at the end of the epidemic. The posterior distribution of the recovery has shifted to the right from the prior, leading the a smaller estimation of infection period than the estimate from the Pref-based method. Similarly to the simulation study, the Gen-based method has wider credible intervals when estimating Ebola prevalence.

5.5 Discussion

Currently, few phylodynamic models take into account information provided by the molecular sequence sampling times to reduce bias and improve the precision in the population dynamics inference. Moreover, none of the existing preferential sampling models focus on relating sampling intensity with a stochastic epidemic population dynamics. In this Chapter, we incorporate the sampling times as an additional source of information into our modeling framework to improve the performance of phylodynamic inference. Our method can be also considered as an extension of sampling-aware preferential sampling method by [Karcher et al., 2016] to a SIR stochastic epidemic model framework. We proposed an efficient MCMC algorithm that, with the help of LNA, allows us to approximate the posterior distribution of the disease population dynamics, the SIR rates as well as coefficients for the preferential sampling rate log-linear model. Through simulation studies, we have found our previous SIR based phylodynamic model have bias in the population dynamic estimation when the number of infected individuals desclines at the end of the epidemic. Such observation is

consistent with conclusions in previous previous work [Karcher et al., 2016, Karcher, 2019]. Furthermore, our simulation studies show that the information of sampling times helps improve identifiability of the removal rate γ. We test our preferential sampling model on real Ebola genealogy from the 2014 Ebola outbreak in West Africa. Our preferential sampling model yields more precise estimates of Ebola population dynamics and provide better agreement with the out-of-sample incidence reports than the model that does not take preferential sampling into account.

Like the previous preferential sampling models [Karcher et al., 2016, Volz and Frost, 2014, Karcher, 2019], a big concern for our SIR preferential sampling framework is the model misspecification issue. Misspecifying preferential sampling model can be dangerous and can lead to systematic bias in parameter estimation. One solution is to include other predictors that influence the sampling intensity into the preferential sampling likelihood. For example, Karcher [2019] proposed an extension of preferential sampling by including time dependent covariates as predictors. Furthermore, Bayesian posterior predictive check [Gelman et al., 1996] can be used to detect model misspecification and discrepancies by comparing the posterior generated data with the observed data.

One potential use of preferential sampling model is to design better molecular sequence sampling strategies for phylodynamic inference during infectious disease outbreaks. Researchers can intentionally guarantee preferential sampling during the sequence data collecting stage. For example, if the epidemic study contains noisy incidence data, the collected sequences can be further subsampled proportional to the observed incidence during each time period. Such preferential sampling procedure can be considered as an alternative to the joint model proposed in Chapter 4, which explicitly combines the incidence data and genealogy data to get more precise estimate on the population dynamics and disease transmission rates.

Our current implementation of the preferential sampling model fits this model to a point estimate of the genealogy obtained from molecular sequences. However, such inference ignores inherent uncertainty in the genealogy. The ideal framework should start from genetic sequences directly. For example, Karcher [2019] extends their sampling aware model to fit it

directly to sequence data with the help of `BEAST`, [Drummond and Rambaut, 2007]. Other software packages like `BEAST2` [Bouckaert et al., 2014a] and `RevBayes` [Höhna et al., 2016a] also provide functionality for sampling trees using MCMC. It would be interesting to add our SIR-based preferential sampling functionality to these packages.

Chapter 6

DISCUSSION AND FUTURE DIRECTIONS

6.1 Conclusion

The main contribution of this dissertation is development of a Bayesian framework that can fit stochastic epidemic models to multiple infectious disease data types, including an observed genealogy estimated from infectious disease genetic sequences sampled during an outbreak and times-series incidence data. Our statistical model can be viewed as semi-parametric: with (1) a parametric SIR model describing the infectious disease dynamics and (2) a non-parametric GMRF-based component modeling the time varying basic reproduction number. To the best of our knowledge, this is the first method combining a Bayesian nonparametric approach with deterministic or stochastic SIR models for phylodynamic inference. The parametric SIR part makes the model more interpretable for researchers and the non-parametric modeling of the basic reproduction number adds more flexibility, allowing the model to capture complicated disease dynamics with time varying infection rate that can change, for example, due to implementation of intervention measures. Our application of the LNA method to approximate the stochastic epidemic dynamic process with a closed-form Gaussian distribution allows us to devise an efficient MCMC algorithm to approximate the high dimensional posterior distribution of model parameters and latent variables.

In Chapter 3, we fit our Gen-based method to a fixed genealogy and demonstrate that our method produces posterior summaries with better frequentist properties than the state-of-the-art ODE-based method, underscoring the importance of working with stochastic models even in large populations. We successfully apply our method to Ebola genealogies estimated from sequences in Sierra Leone and Liberia during the 2014-2015 outbreak in west Africa. Our nonparametric estimates of $R_0(t)$ in Sierra Leone and Liberia suggest that the basic

reproduction number decreased in two-stages, where the second stage brought it below 1.0 — a sign of epidemic containment. The second stage of $R_0(t)$ decrease closely follows in time implementation of interventions, pointing to their effectiveness.

One problem that we noticed in Chapter 3 is posterior sensitivity to changes in prior distribution of the removal rate (i.e., an identifiability problem), especially in scenarios with a time-varying reproduction number. We also demonstrated that a similar issue appears when fitting stochastic epidemic models to surveillance case count data, where the reporting rate suffers from prior sensitivity. Driven by these identifiability problems, we proposed a Joint-based model using the integration of genealogy data and incidence time series in Chapter 4. Through simulation studies, we showed that combining multiple data types not only resolves parameter identifiability problems for removal and reporting rates, but also results in more precise and less biased estimating of the basic reproduction number and disease population trajectories. Moreover, our method performs well in fitting genealogy data and incidence data when these data are sequentially collected during the outbreak. Using these streaming data our Joint-based method is able to quickly detect the change in the basic reproduction number and the trend of prevalence. We also tested the performance of our model for short-term forecasting of reported incidence. Such forecasting should be helpful to researchers in disease control department when implementing interventions and allocating outbreak response resources.

As few of the current phylodynamic models take into account the information of sampling times to mitigate the estimation bias in population dynamics, we also extend our Gen-based method to incorporate the distribution of sampling times during sequence collection stage. Our simulation studies show the bias previously reported in traditional Gen-based phylodynamic inference also appears in our SIR framework. We propose a preferential sampling model that explicitly takes into account dependency between molecular sequence sampling intensity and disease prevalence. The Pref-based inference decreases the bias and provides more precise estimates for the parameters and latent disease population dynamics. Moreover, like Joint-based method, the Pref-based method is also helpful in addressing the identifiabil-

ity problem for removal rate in the Gen-based method.

6.2 Future Directions

6.2.1 Fitting stochastic epidemic models to genetic sequences

The first possible extension of our methodology is to add ability to perform statistical inference from sequence data directly by sampling over possible genealogies via MCMC. Our current model assumes the observed data are a point estimate of the genealogy estimated using genetic sequences. However, starting from point estimate ignores the randomness and uncertainty in the genealogy. Hence, the ideal phylodynamic model should directly start from the genetic sequences. Consider an *alignment* of $\mathbf{A} = \{\mathbf{a}_1, \ldots, \mathbf{a}_n\}$ of L sites and n infectious disease genetic sequences obtained from hosts in a well-mixed population at sampling times $s_1, \ldots, s_n$. Each sequence $\mathbf{a}_i$ is consist of vector of nucleotide bases, for example,

$$\mathbf{a}_i = ATGAACTAAC \cdots TCA.$$

We assume the sequences data $\mathbf{A}$ are generated on a given genealogy $\mathbf{g}$ through a CTMC substitution model. The model assumes alignment sites to be independent and identically distributed, with a transition matrix $\boldsymbol{\omega}$ controlling the CTMC substitution rates between the different nucleotide bases. For a fixed genealogy $\mathbf{g}$ and substitution rates $\boldsymbol{\omega}$, the phylogenetic likelihood, denoted by $\Pr(\mathbf{A} \mid \mathbf{g}, \boldsymbol{\omega})$, can be efficiently computed using Felsenstein's pruning algorithm [Felsenstein, 1973, 1981].

The posterior distribution for the population trajectory $\mathbf{X}_{0:T}$, genealogy $\mathbf{g}$, substitution rate $\boldsymbol{\omega}$ and SIR rates $\boldsymbol{\theta}$ given observed sequence alignment is

$$\begin{aligned}\Pr(\boldsymbol{\omega}, \mathbf{g}, \mathbf{X}_{0:T}, \boldsymbol{\theta} \mid \mathbf{A}) \propto & \underbrace{\Pr(\mathbf{a} \mid \mathbf{g}, \boldsymbol{\omega})}_{\text{phylogenetic likelihood}} \cdot \underbrace{\Pr(\mathbf{g} \mid \mathbf{X}_{0:T}, \boldsymbol{\theta})}_{\text{coalescent likelihood}} \cdot \underbrace{\Pr(\mathbf{X}_{1:T} \mid \boldsymbol{\theta}, \mathbf{X}_0)}_{\text{trajectory likelihood}} \cdot \\ & \underbrace{\pi(\mathbf{X}_0)\pi(\boldsymbol{\theta})\pi(\boldsymbol{\omega})}_{\text{priors}},\end{aligned} \tag{6.1}$$

where $\pi(\boldsymbol{\omega})$ is a prior distribution for substitution rates $\boldsymbol{\omega}$. Posterior approximation via MCMC requires adding an additional MCMC kernel to update genealogies. Incorporating

our LNA-based framework to existing phylodynamic/phylogenetic software packages, such as `BEAST` [Drummond and Rambaut, 2007], `BEAST2` [Bouckaert et al., 2014b] and `RevBayes` [Höhna et al., 2016b] should allow us to combine our MCMC kernels with existing MCMC for sampling trees and performing inference directly from genetic sequences.

6.2.2 More complicated stochastic epidemic and structured coalescent models

The methods proposed in this thesis mainly rely on the assumption that population is well-mixed and the disease dynamics follow a SIR model. In reality, the disease transmission process can be more complicated and may requires more sophisticated stochastic epidemic models. In Section 2.2.7, we demonstrate the LNA framework can be easily generalized to other compartmental models such as SEIR model. The structured coalescent likelihood for most of the compartment modes does not have a closed form and the calculation requires solving a system of ODEs. See Section 2.3.2 for more details. There are some existing work focusing on fitting complex structured coalescent model on genetic data. For instance, Volz and Pond [2014], Volz and Siveroni [2018] used a structured phylodynamic model and deterministic epidemic models to analyze Ebola genetic sequences. Rasmussen et al. [2014] adopted particle MCMC approach to fit a stochastic epidemic model to HIV genealogies. However, there is no general framework that allows one to fit arbitrary stochastic epidemic models to molecular sequence data.

Given the shortcomings of deterministic method and the computational challenge in PMCMC, one interesting future direction of this work is to generalize the LNA method to fit complicated compartmental epidemic models, including models with multi-stage infections like SEIR model and models with the population stratified by sex, age, geographic location, or other demographic variables, to molecular seqeunces. Here we present some possible future models in the following subsections.

Multi-stage infections

For Ebola, it is more realistic to use a SEIR model that assumes a latent period during which an infected individual is not infectious [Althaus, 2014, Volz and Siveroni, 2018]. More details about SEIR model can be seen in Chapter 2. Adding multiple infectious stages/compartments is also possible to relax the assumption of exponentially distributed infectious period.

Stratified populations

We are also interested in devising stochastic epidemic models that include different strata in each population group. For example, age can be a factor that affects the infection rate, since juvenile hosts have higher contact rates and are more susceptible to infection with many diseases. Volz [2012] demonstrated a SIR model dividing the population into juvenile (S_J for susceptible, I_J for infected, and R_J for removed) and adult states (S_A for susceptible, I_A for infected, and R_A for removed). Each of the age groups has distinct transition rates between states. Let β_{ij} denote the rate at which an individual from age group i infects an individual from age group j. The two age groups are assumed to have the same rate for recovery. Figure (6.1) demonstrate an epidemic model from (citepvolz2012complex. The "reaction" equations can be written as

$$I_i + S_j \xrightarrow{\beta_{ij}} I_i + I_j \quad i, j \in \{J, A\}, \tag{6.2}$$

$$I_i \xrightarrow{\gamma} R_i \quad i \in \{J, A\}. \tag{6.3}$$

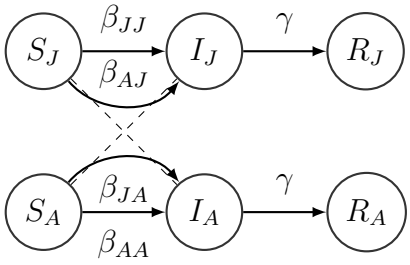

Figure 6.1: An illustration of the SIR model with two age strata. Vector (S_i, I_i, R_i) denotes the numbers of susceptible, infected and removed individuals in age group i ($i \in \{J, A\}$). Scalars β_{ij}s denote the infection rates within and between age groups. The two population strata share the same removal rate γ.

Spatial heterogeneity and migration

Another possible future direction is to develop a model that accounts for migration between and heterogeneity in disease transition within different geographical locations. This could be particularly helpful when not only incidence, but also molecular sequence data are collected in multiple locations. For example, Fintzi [2019] fitted a SEIR model with migration events between three countries using the incidence data collected from Sierra Leone, Liberia and Guinea during the 2014-2015 Ebola outbreak. Let (S_i, E_i, I_i, R_i) denote the number of susceptible, exposed, infected and recovered in country i, where $i \in \{S, L, G\}$ (S, L, G are used as abbreviations for Sierra Leone, Liberia and Guinea respectively). The infection rate and recovery rate are assumed to be different in each location, denoted by β_i and γ_i. The author also allows for migrations of infected individuals between three countries, with the migration rate from location i to j is denoted by λ_{ij}. Figure 6.2 depicts the transition graph

of the three country migration model. The "reactions" can be written as

$$I_i + S_i \xrightarrow{\beta_i(t)} E_i + I_i \quad i \in \{S, L, G\}, \tag{6.4}$$

$$E_i \xrightarrow{\mu_i} I_i \quad i \in \{S, L, G\}, \tag{6.5}$$

$$I_i \xrightarrow{\gamma_i} R_i \quad i \in \{S, L, G\}, \tag{6.6}$$

$$I_i \xrightarrow{\lambda_{ij}} I_j \quad i \neq j \text{ and } i, j \in \{S, L, G\}. \tag{6.7}$$

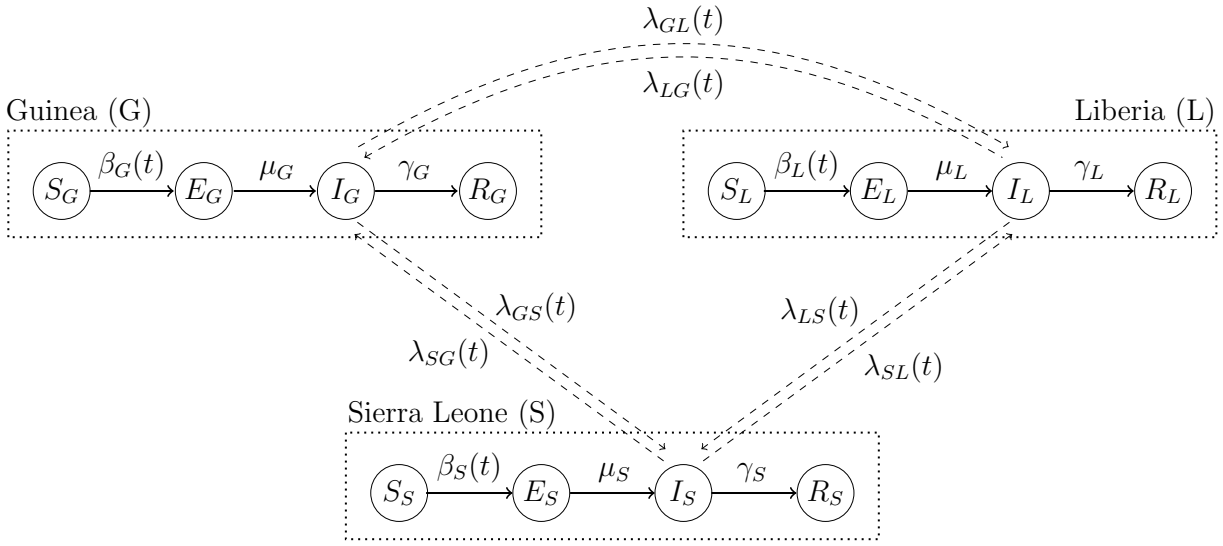

Figure 6.2: An illustration of three country Ebola compartmental model from Fintzi [2019]. Diagram of state transitions for a stratified SEIR model for an Ebola outbreak in Guinea, Liberia, and Sierra Leone. Dashed boxes denote countries, nodes in circles denote the model compartments: susceptible (S), exposed (E), infectious (I), recovered (R). Compartments are subscripted with country indicators. Solid lines with arrows indicate stochastic transitions between model compartments, which occur continuously in time. Dashed lines indicate migrations.

BIBLIOGRAPHY

CL Althaus. Estimating the reproduction number of Ebola virus (EBOV) during the 2014 outbreak in West Africa. PLoS Currents, 6, 2014.

RM Anderson and RM May. Infectious Diseases of Humans: Dynamics and Control, volume 28. Wiley Online Library, 1992.

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

KE Atkins, NS Wenzel, M Ndeffo-Mbah, FL Altice, JP Townsend, and AP Galvani. Under-reporting and case fatality estimates for emerging epidemics. bmj, 350:h1115, 2015.

NTJ Bailey. The Mathematical Theory of Infectious Diseases and Its Applications. Hafner Press/MacMillian Pub. Co, 1975.

JM Bernardo, MJ Bayarri, JO Berger, AP Dawid, D Heckerman, AFM Smith, and M West. Non-centered parameterisations for hierarchical models and data augmentation. In Bayesian Statistics 7: Proceedings of the Seventh Valencia International Meeting, page 307. Oxford University Press, USA, 2003.

R Bouckaert, J Heled, D Kühnert, T Vaughan, CH Wu, D Xie, MA Suchard, A Rambaut, and AJ Drummond. BEAST 2: A software platform for Bayesian evolutionary analysis. PLoS Computational Biology, 10(4):1–6, 04 2014a.

R Bouckaert, J Heled, D Kühnert, T Vaughan, CH Wu, D Xie, MA Suchard, A Rambaut, and AJ Drummond. Beast 2: a software platform for Bayesian evolutionary analysis. PLoS computational biology, 10(4):e1003537, 2014b.

E Buckingham-Jeffery, V Isham, and T House. Gaussian process approximations for fast inference from infectious disease data. Mathematical Biosciences, 301, 2018.

Centers for Disease Control. Centers for Disease Control and Prevention. 2014–2016 Ebola outbreak in West Africa. `https://www.cdc.gov/vhf/ebola/history/2014-2016-outbreak/index.html`. Last accessed: Dec, 15, 2018.

G Chowell. Fitting dynamic models to epidemic outbreaks with quantified uncertainty: A primer for parameter uncertainty, identifiability, and forecasts. Infectious Disease Modelling, 2(3):379–398, 2017.

B Dearlove and DJ Wilson. Coalescent inference for infectious disease: meta-analysis of hepatitis C. Philosophical Transactions of the Royal Society, Series B, 368(1614):20120314, 2013.

P Donnelly and S Tavare. Coalescents and genealogical structure under neutrality. Annual Review of Genetics, 29(1):401–421, 1995.

AJ Drummond and A Rambaut. BEAST: Bayesian evolutionary analysis by sampling trees. BMC evolutionary biology, 7(1):214, 2007.

AJ Drummond, GK Nicholls, AlG Rodrigo, and W Solomon. Estimating mutation parameters, population history and genealogy simultaneously from temporally spaced sequence data. Genetics, 161(3):1307–1320, 2002.

AJ Drummond, A Rambaut, B Shapiro, and OG Pybus. Bayesian coalescent inference of past population dynamics from molecular sequences. Molecular Biology and Evolution, 22 (5):1185–1192, 2005.

G Dudas, LM Carvalho, T Bedford, AJ Tatem, G Baele, NR Faria, DJ Park, JT Ladner, A Arias, D Asogun, et al. Virus genomes reveal factors that spread and sustained the Ebola epidemic. Nature, 544(7650):309–315, 2017.

S Engblom. Computing the moments of high dimensional solutions of the master equation. Applied Mathematics and Computation, 180(2):498–515, 2006.

P Fearnhead, V Giagos, and C Sherlock. Inference for reaction networks using the linear noise approximation. Biometrics, 70(2):457–466, 2014.

J Felsenstein. Maximum likelihood and minimum-steps methods for estimating evolutionary trees from data on discrete characters. Systematic Biology, 22(3):240–249, 1973.

J Felsenstein. Evolutionary trees from DNA sequences: a maximum likelihood approach. Journal of molecular evolution, 17(6):368–376, 1981.

L Ferm, P Lötstedt, and A Hellander. A hierarchy of approximations of the master equation scaled by a size parameter. Journal of Scientific Computing, 34(2):127–151, 2008.

J Fintzi, X Cui, J Wakefield, and VN Minin. Efficient data augmentation for fitting stochastic epidemic models to prevalence data. Journal of Computational and Graphical Statistics, 26(4):918–929, 2017.

JR Fintzi. Bayesian Modeling of Partially Observed Epidemic Count Data. PhD thesis, 2019.

SDW Frost and EM Volz. Viral phylodynamics and the search for an "effective number of infections". Philosophical Transactions of the Royal Society B: Biological Sciences, 365 (1548):1879–1890, 2010.

A Gelman, XL Meng, and H Stern. Posterior predictive assessment of model fitness via realized discrepancies. Statistica sinica, pages 733–760, 1996.

V Giagos. Inference for Auto-Regulatory Genetic Networks Using Diffusion Process Approximations. PhD thesis, Lancaster University, 2010.

MS Gill, P Lemey, NR Faria, A Rambaut, B Shapiro, and MA Suchard. Improving Bayesian population dynamics inference: a coalescent-based model for multiple loci. Molecular Biology and Evolution, 30(3):713–724, 2013.

DT Gillespie. Exact stochastic simulation of coupled chemical reactions. The Journal of Physical Chemistry, 81(25):2340–2361, 1977.

DT Gillespie. The chemical Langevin equation. The Journal of Chemical Physics, 113(1): 297–306, 2000.

JR Glynn, H Bower, S Johnson, CF Houlihan, C Montesano, JT Scott, MG Semple, MS Bangura, AJ Kamara, O Kamara, et al. Asymptomatic infection and unrecognised Ebola virus disease in Ebola-affected households in Sierra Leone: a cross-sectional study using a new non-invasive assay for antibodies to Ebola virus. The Lancet infectious diseases, 17(6): 645–653, 2017.

T Gneiting and AE Raftery. Strictly proper scoring rules, prediction, and estimation. Journal of the American Statistical Association, 102(477):359–378, 2007.

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

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

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

BT Grenfell, OG Pybus, JR Gog, JLN Wood, JM Daly, JA Mumford, and EC Holmes. Unifying the epidemiological and evolutionary dynamics of pathogens. Science, 303(5656): 327–332, 2004.

RC Griffiths and S Tavaré. Sampling theory for neutral alleles in a varying environment. Philosophical Transactions of the Royal Society of London B: Biological Sciences, 344 (1310):403–410, 1994.

RC Griffiths and S Tavare. Ancestral inference in population genetics. Statistical Science, pages 307–319, 1994.

WK Hastings. Monte Carlo sampling methods using Markov chains and their applications. Biometrika, 57(1):97–109, 1970.

Lam Si Tung Ho, Forrest W Crawford, and Marc A Suchard. Direct likelihood-based inference for discretely observed stochastic compartmental models of infectious disease. arXiv preprint arXiv:1608.06769, 2016.

S Höhna, MJ Landis, TA Heath, B Boussau, N Lartillot, BR Moore, JP Huelsenbeck, and F Ronquist. RevBayes: Bayesian phylogenetic inference using graphical models and an interactive model-specification language. Systematic Biology, 65(4):726–736, 2016a.

S Höhna, MJ Landis, TA Heath, B Boussau, N Lartillot, BR Moore, JP Huelsenbeck, and F Ronquist. Revbayes: Bayesian phylogenetic inference using graphical models and an interactive model-specification language. Systematic Biology, 65(4):726–736, 2016b.

SM Iacus. Simulation and Inference for Stochastic Differential Equations: with R Examples. Springer Science & Business Media, 2009.

MD Karcher. Preferential Sampling and Model Checking in Phylodynamic Inference. PhD thesis, 2019.

MD Karcher, JA Palacios, T Bedford, MA Suchard, and VN Minin. Quantifying and mitigating the effect of preferential sampling on phylodynamic inference. PLoS Computational Biology, 12(3):e1004789, 2016.

MJ Keeling and P Rohani. Modeling Infectious Diseases in Humans and Animals. Princeton University Press, 2011.

MJ Keeling and JV Ross. On methods for studying stochastic disease dynamics. Journal of The Royal Society Interface, 5(19):171–181, 2007.

WO Kermack and AG McKendrick. A contribution to the mathematical theory of epidemics. Proceedings of the royal society of london. Series A, Containing papers of a mathematical and physical character, 115(772):700–721, 1927.

AA King, EL Ionides, CM Bretó, S Ellner, B Kendall, H Wearing, MJ Ferrari, M Lavine, and DC Reuman. pomp: Statistical inference for partially observed Markov processes (R package). URL http://pomp. r-forge. r-rproject. org, 2010.

JFC Kingman. The coalescent. Stochastic Processes and their Applications, 13(3):235–248, 1982.

AA Koepke, IM Longini Jr, ME Halloran, J Wakefield, and VN Minin. Predictive modeling of cholera outbreaks in Bangladesh. The annals of applied statistics, 10(2):575, 2016.

M Komorowski, B Finkenstädt, CV Harper, and DA Rand. Bayesian inference of biochemical kinetic parameters using the linear noise approximation. BMC Bioinformatics, 10(1):343, 2009.

MK Kuhner, J Yamato, and J Felsenstein. Maximum likelihood estimation of population growth rates based on the coalescent. Genetics, 149(1):429–434, 1998.

D Kühnert, T Stadler, TG Vaughan, and AJ Drummond. Simultaneous reconstruction of evolutionary history and epidemiological dynamics from viral sequences with the birth–death SIR model. Journal of the Royal Society Interface, 11(94):20131106, 2014.

TG Kurtz. Solutions of ordinary differential equations as limits of pure jump Markov processes. Journal of Applied Probability, 7(1):49–58, 1970.

TG Kurtz. Limit theorems for sequences of jump Markov processes. Journal of Applied Probability, 8(2):344–356, 1971.

GE Leventhal, HF Günthard, S Bonhoeffer, and T Stadler. Using an epidemiological model for phylogenetic inference reveals density dependence in HIV transmission. Molecular Biology and Evolution, 31(1):6–17, 2013.

N Metropolis, AW Rosenbluth, MN Rosenbluth, AH Teller, and E Teller. Equation of state calculations by fast computing machines. The Journal of Chemical Physics, 21(6):1087–1092, 1953.

VN Minin, EW Bloomquist, and MA Suchard. Smooth skyride through a rough skyline: Bayesian coalescent-based inference of population dynamics. Molecular Biology and Evolution, 25(7):1459–1471, 2008.

PAP Moran. Random processes in genetics. In Mathematical Proceedings of the Cambridge Philosophical Society, volume 54, pages 60–71. Cambridge University Press, 1958.

PAP Moran. The Statistical Process of Evolutionary Theory. Clarendon Press, 1962.

I Murray, RP Adams, and DJC MacKay. Elliptical slice sampling. In AISTATS, volume 13, pages 541–548, 2010.

RM Neal. Slice sampling. The annals of statistics, 31(3):705–767, 2003.

Sophia Ng and Ben J Cowling. Association between temperature, humidity and ebolavirus disease outbreaks in Africa, 1976 to 2014. Eurosurveillance, 19(35):20892, 2014.

B Øksendal. Stochastic differential equations. In Stochastic Differential Equations, pages 65–84. Springer, 2003.

PD O'Neill and GO Roberts. Bayesian inference for partially observed stochastic epidemics. Journal of the Royal Statistical Society: Series A (Statistics in Society), 162(1):121–129, 1999.

JA Palacios and VN Minin. Gaussian process-based Bayesian nonparametric inference of population size trajectories from gene genealogies. Biometrics, 69(1):8–18, 2013.

O Papaspiliopoulos, GO Roberts, and M Sköld. A general framework for the parametrization of hierarchical models. Statistical Science, pages 59–73, 2007.

OG Pybus, MA Charleston, S Gupta, A Rambaut, EC Holmes, and PH Harvey. The epidemic behavior of the hepatitis C virus. Science, 292(5525):2323–2325, 2001.

DA Rasmussen, O Ratmann, and K Koelle. Inference for nonlinear epidemiological models using genealogies and time series. PLoS Computational Biology, 7(8):e1002136, 2011.

DA Rasmussen, EM Volz, and K Koelle. Phylodynamic inference for structured epidemiological models. PLoS Computational Biology, 10(4):e1003570, 2014.

NG Reich, LC Brooks, SJ Fox, S Kandula, CJ McGowan, E Moore, D Osthus, EL Ray, A Tushar, TK Yamana, et al. A collaborative multiyear, multimodel assessment of seasonal influenza forecasting in the United States. Proceedings of the National Academy of Sciences, page 201812594, 2019.

AG Rodrigo and J Felsenstein. Coalescent Approaches. The Evolution of HIV, page 233, 1999.

H Rue. Fast sampling of Gaussian Markov random fields. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 63(2):325–338, 2001.

H Rue and L Held. Gaussian Markov Random Fields: Theory and Applications. CRC press, 2005.

SV Scarpino, A Iamarino, C Wells, D Yamin, M Ndeffo-Mbah, NS Wenzel, SJ Fox, T Nyenswah, FL Altice, AP Galvani, et al. Epidemiological and viral genomic sequence analysis of the 2014 Ebola outbreak reveals clustered transmission. Clinical Infectious Diseases, 60(7):1079–1082, 2014.

JP Schmidt, AW Park, AM Kramer, BA Han, LW Alexander, and JM Drake. Spatiotemporal fluctuations and triggers of Ebola virus spillover. Emerging infectious diseases, 23(3):415, 2017.

Scott A Sisson, Yanan Fan, and Mark M Tanaka. Sequential Monte Carlo without likelihoods. Proceedings of the National Academy of Sciences, 104(6):1760–1765, 2007.

T Stadler, D Kühnert, S Bonhoeffer, and AJ Drummond. Birth–death skyline plot reveals temporal changes of epidemic spread in HIV and hepatitis C virus (HCV). Proceedings of the National Academy of Sciences, 110(1):228–233, 2013.

T Stadler, D Kühnert, DA Rasmussen, and L du Plessis. Insights into the early epidemic spread of Ebola in Sierra Leone provided by viral sequence data. PLoS Currents, 6, 2014.

MA Suchard, P Lemey, G Baele, DL Ayres, AJ Drummond, and A Rambaut. Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. Virus Evolution, 4 (1):vey016, 2018.

S Towers, O Patterson-Lomba, and C Castillo-Chavez. Temporal variations in the effective reproduction number of the 2014 West Africa Ebola outbreak. PLoS Currents, 6, 2014.

NG Van Kampen and WP Reinhardt. Stochastic Processes in Physics and Chemistry, 1983.

TG Vaughan, GE Leventhal, DA Rasmussen, AJ Drummond, D Welch, and T Stadler. Estimating epidemic incidence and prevalence from genomic data. BioRxiv, page 142570, 2018. doi: https://doi.org/10.1101/142570.

E Volz and S Pond. Phylodynamic analysis of Ebola virus in the 2014 Sierra Leone epidemic. PLoS Currents, 6, 2014.

E Volz and I Siveroni. Bayesian phylodynamic inference with complex models. BioRxiv, 2018. doi: 10.1101/268052.

EM Volz. Complex population dynamics and the coalescent under neutrality. Genetics, 190 (1):187–201, 2012.

EM Volz and SDW Frost. Sampling through time and phylodynamic inference with coalescent and birth–death models. Journal of The Royal Society Interface, 11(101):20140945, 2014.

EM Volz, SLK Pond, MJ Ward, AJL Brown, and SDW Frost. Phylodynamics of infectious disease epidemics. Genetics, 183(4):1421–1430, 2009.

EM Volz, K Koelle, and T Bedford. Viral phylodynamics. PLoS Computational Biology, 9 (3):e1002947, 2013.

E Vynnycky and R White. An Introduction to Infectious Disease Modelling. OUP oxford, 2010.

EWJ Wallace. A simplified derivation of the linear noise approximation. Arxiv preprint arXiv:1004.4280, 2010.

HJ Wearing, P Rohani, and MJ Keeling. Appropriate models for the management of infectious diseases. PLoS Medicine, 2(7):e174, 2005.

WHO Ebola Response Team. Ebola virus disease in West Africa —- the first 9 months of the epidemic and forward projections. New England Journal of Medicine, 371(16):1481–1495, 2014.

DJ Wilkinson. Stochastic Modelling for Systems Biology. CRC press, 2011.

World Health Organization. Case definition recommendtions for Ebola or Marburg virus diseases: interim guideline. Technical report, World Health Organization, 2014a.

World Health Organization. World Health Organization. Ebola data and statistics. `http://apps.who.int/gho/data/node.ebola-sitrep.quick-downloads?lang=en`, May 11, 2016b. Last accessed: February 28, 2018.

S Wright. Evolution in Mendelian populations. Genetics, 16(2):97–159, 1931.

Appendix A

METHODOLOGY DETAILS

A.1 Reparameterization Details

In LNA, the latent trajectory $\mathbf{X}(t)$ is decomposed into the deterministic part $\boldsymbol{\eta}(t)$ plus a stochastic part $\mathbf{M}(t)$ that follows a multivariate Gaussian distribution with mean $\mathbf{0}$. However, the population size at the i-th time interval $\mathbf{X}_i$ depends on rate parameter $\boldsymbol{\theta}$ and is correlated with other population sizes $\mathbf{X}_j$s in the trajectory, leading to mixing issues for the MCMC chain, especially when we introduce multiple change points for reproduction number R_0.

Here we take the idea of non-centered parameterization from Papaspiliopoulos et al. [2007], Bernardo et al. [2003] and reparameterize the latent trajectory in terms of residuals $\mathbf{X}_i - \boldsymbol{\eta}_i$ for $i = 1, \ldots, T$. Given rate parameters $\boldsymbol{\theta}_{i-1}$, ODE solution $\boldsymbol{\eta}_{0:T}$, fundamental matrix $\boldsymbol{\Sigma}(t_i, t_{i-1})$ and variance matrix $\boldsymbol{\Phi}_i$ in (3.10), the trajectory $\mathbf{X}_{0:T}$ can be parameterized using standard Gaussian noise $\boldsymbol{\xi}_{1:T}$ based on the following iterative equations:

$$\begin{aligned} \mathbf{X}_0 &= \boldsymbol{\eta}_0, && \text{(A.1)} \\ \mathbf{X}_i &= \boldsymbol{\mu}\left(\mathbf{X}_{i-1} - \boldsymbol{\eta}(t_{i-1}), \Delta t_i, \boldsymbol{\theta}_{i-1}\right) + \boldsymbol{\eta}_i + \boldsymbol{\Phi}_i^{1/2}\boldsymbol{\xi}_i, && \text{(A.2)} \\ &= \boldsymbol{\Sigma}(t_i, t_{i-1})\left(\mathbf{X}_{i-1} - \boldsymbol{\eta}_{i-1}\right) + \boldsymbol{\eta}_i + \boldsymbol{\Phi}_i^{1/2}\boldsymbol{\xi}_i, \quad \text{for } i = 1, \ldots, T. && \text{(A.3)} \end{aligned}$$

Let $\mathbf{M}_i := \mathbf{X}_i - \boldsymbol{\eta}_i$ denote the residual in grid cell i. Based on (A.3), the residual process satisfies

$$\begin{aligned} \mathbf{M}_1 &= \boldsymbol{\Phi}_1^{1/2}\boldsymbol{\xi}_1 && \text{(A.4)} \\ \mathbf{M}_i &= \boldsymbol{\Sigma}(t_{i-1}, t_i)\mathbf{M}_{i-1} + \boldsymbol{\Phi}_i^{1/2}\boldsymbol{\xi}_i, \qquad i = 2, \ldots, T. && \text{(A.5)} \end{aligned}$$

Then $\mathbf{M}_{0:T}$ can be viewed as a Gaussian Markov random field with mean $\mathbf{0}$ that follows the Markov property on a chain graph. Let $\boldsymbol{\Sigma}_i$ be the abbreviated notation of $\boldsymbol{\Sigma}(t_i, t_{i-1})$ and

$\mathbf{P}_i = \mathbf{\Phi}_i^{1/2}$. From (A.5) , $\mathbf{M}_i$ can be written as

$$\begin{aligned}
\mathbf{M}_i &= \mathbf{\Sigma}_{i-1}\mathbf{M}_{i-1} + \mathbf{P}_i\boldsymbol{\xi}_i \\
&= \mathbf{\Sigma}_{i-1}\left(\mathbf{\Sigma}_{i-2}\mathbf{M}_{i-2} + \mathbf{P}_{i-1}\boldsymbol{\xi}_{i-1}\right) + \mathbf{P}_i\boldsymbol{\xi}_i \\
&= \mathbf{\Sigma}_{i-1}\mathbf{\Sigma}_{i-2}\mathbf{M}_{i-2} + \mathbf{\Sigma}_{i-1}\mathbf{P}_{i-1}\boldsymbol{\xi}_{i-1} + \mathbf{P}_i\boldsymbol{\xi}_i \\
&= \mathbf{\Sigma}_{i-1}\mathbf{\Sigma}_{i-2}\cdots\mathbf{\Sigma}_1\mathbf{P}_1\boldsymbol{\xi}_1 + \cdots + \mathbf{P}_i\boldsymbol{\xi}_i \\
&= \sum_{k=1}^{i}(\prod_{j=k}^{i-1}\mathbf{\Sigma}_j)\mathbf{P}_k\boldsymbol{\xi}_k.
\end{aligned}$$

Since $\mathbf{\Sigma}_i$ and $\mathbf{P}_i$ are governed by rate parameters $\boldsymbol{\theta}_{i-1}$ and initial value $\mathbf{X}_0$, then we define the transform matrix $\mathbf{L}(\mathbf{X}_0, \boldsymbol{\theta}_{0:T}) \in \mathbb{R}^{2T\times 2T}$,

$$\mathbf{L}(\mathbf{X}_0, \boldsymbol{\theta}_{0:T}) = \begin{pmatrix}
\mathbf{P}_1 & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} \\
\mathbf{\Sigma}_1\mathbf{P}_1 & \mathbf{P}_2 & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} \\
\mathbf{\Sigma}_2\mathbf{\Sigma}_1\mathbf{P}_1 & \mathbf{\Sigma}_2\mathbf{P}_2 & \mathbf{P}_3 & \cdots & \mathbf{0} & \mathbf{0} \\
\cdots & \cdots & \cdots & \ddots & \cdots & \cdots \\
\mathbf{\Sigma}_{T-2}\cdots\mathbf{\Sigma}_1\mathbf{P}_1 & \mathbf{\Sigma}_{T-2}\cdots\mathbf{\Sigma}_2\mathbf{P}_2 & \mathbf{\Sigma}_{T-2}\cdots\mathbf{\Sigma}_2\mathbf{P}_3 & \cdots & \mathbf{P}_{T-1} & \mathbf{0} \\
\mathbf{\Sigma}_{T-1}\cdots\mathbf{\Sigma}_1\mathbf{P}_1 & \mathbf{\Sigma}_{T-1}\cdots\mathbf{\Sigma}_2\mathbf{P}_2 & \mathbf{\Sigma}_{T-1}\cdots\mathbf{\Sigma}_3\mathbf{P}_3 & \cdots & \mathbf{\Sigma}_{T-1}\mathbf{P}_{T-1} & \mathbf{P}_T
\end{pmatrix} \tag{A.6}$$

A linear relationship between $\mathbf{X}_{1:T}$ and the reparameterized noise $\boldsymbol{\xi}_{1:T}$ can be established with the help of the above transform matrix $\mathbf{L}$,

$$\begin{pmatrix} \mathbf{X}_1 \\ \vdots \\ \mathbf{X}_T \end{pmatrix} = \begin{pmatrix} \boldsymbol{\eta}_1 \\ \vdots \\ \boldsymbol{\eta}_T \end{pmatrix} + \mathbf{L}(\mathbf{X}_0, \boldsymbol{\theta}_{0:T}) \begin{pmatrix} \boldsymbol{\xi}_1 \\ \vdots \\ \boldsymbol{\xi}_T \end{pmatrix}. \tag{A.7}$$

Instead of directly updating $\mathbf{X}_{1:T}$, we will apply the above transform and update the Gaussian noise $\boldsymbol{\xi}_{1:T}$ instead. The MCMC approach will focus on sampling parameter $I_0, R_0, \gamma, \boldsymbol{\delta}_{1:T}$, $\boldsymbol{\xi}_{1:T}, \sigma$ with the posterior likelihood

$$\begin{aligned}
& \Pr(I_0, R_0, \gamma, \boldsymbol{\delta}_{1:T}, \boldsymbol{\xi}_{1:T}, \sigma|\mathbf{g}) \\
\propto\ & \Pr(\mathbf{g}|I_0, R_0, \gamma, \boldsymbol{\delta}_{1:T}, \boldsymbol{\xi}_{1:T}, \sigma)\Pr(I_0)\Pr(R_0)\Pr(\gamma)\Pr(\boldsymbol{\delta}_{1:T})\Pr(\boldsymbol{\xi}_{1:T})\Pr(\sigma) \\
\propto\ & \Pr(\mathbf{g}|\mathbf{X}_{0:T}, \boldsymbol{\theta}_{0:T})\Pr(I_0)\Pr(R_0)\Pr(\gamma)\Pr(\boldsymbol{\delta}_{1:T})\Pr(\boldsymbol{\xi}_{1:T})\Pr(\sigma).
\end{aligned}$$

In summary,the transformation that allows us to move from parameterization in terms of $\mathbf{X}_{0:T}, \theta_{0:T}$ to the parameterization in terms of $I_0, R_0, \gamma, \boldsymbol{\delta}_{1:T}, \boldsymbol{\xi}_{1:T}, \sigma$ are based on the following equations:

1. $R_{0_i} := R_0(t_i) = R_0 \cdot \exp(\prod_{j=1}^{i} \delta_j \sigma)$ — a function of R_0, $\boldsymbol{\delta}_{1:i}$ and σ.

2. $\beta_i := \beta(t_i) = \dfrac{NR_0(t_i)}{\gamma}$ — a function of R_0, $\boldsymbol{\delta}_{1:i}$, σ and γ.

3. $\boldsymbol{\theta}_i = (\beta_i, \gamma)$ — a function of R_0, $\boldsymbol{\delta}_{1:i}$, σ and γ.

4. $\boldsymbol{\theta}_{0:T}$ — a function of R_0, $\boldsymbol{\delta}_{1:T}$, σ and γ.

5. $\mathbf{X}_0 = (N, I_0)^T$.

6. $\mathbf{X}_{1:T} = \boldsymbol{\eta}_{1:T} + \mathbf{L}(\mathbf{X}_0, \boldsymbol{\theta}_{0:T})\boldsymbol{\xi}_{1:T}$ — a function of R_0, $\boldsymbol{\delta}_{1:T}$, σ, γ, I_0 and $\boldsymbol{\xi}_{1:T}$.

A.1.1 MCMC algorithm for the ODE-based model

The MCMC algorithm for ODE-based method is similar to the LNA-based MCMC except there is no need to update the Gaussian noise $\boldsymbol{\xi}_{1:T}$ in the population trajectory. The MCMC updates of parameters and latent variables is given in Algorithm 7.

A.1.2 MCMC Algorithm for Joint-based Model

Algorithm 7 Updating rule in the ODE-based MCMC algorithm

1: **Input:** Parameter values from the previous interation $I_0, R_0, \gamma, \boldsymbol{\delta}_{1:T}, \sigma$, geneology $\mathbf{g}$. Proposal density $q_1(\cdot|\cdot)$, $q_2(\cdot|\cdot)$ for updating the initial number of infected individuals and the removal rate.

2: **Output** Updated parameters values

3: Calculate $\mathbf{X}_{0:T}$, $\boldsymbol{\theta}_{0:T}$ based on $I_0, R_0, \gamma, \boldsymbol{\delta}_{1:T}, \sigma$.

4: Propose I_0' based on $q_1(\cdot|I_0)$, then $\mathbf{X}_{0:T}$ will be deterministically updated to $\mathbf{X}'_{0:T}$ according to I_0', $R_0, \gamma, \boldsymbol{\delta}_{1:T}, \sigma$ based on ODE integration.

5: Accept $(I_0', \mathbf{X}'_{0:T})$ with acceptance probability

$$a \leftarrow \min\left(1, \frac{\Pr(\mathbf{g}|\boldsymbol{\theta}'_{0:T}, \mathbf{X}'_{0:T}) \Pr(I_0') q_1(I_0|I_0')}{\Pr(\mathbf{g}|\boldsymbol{\theta}_{0:T}, \mathbf{X}_{0:T}) \Pr(I_0) q_1(I_0'|I_0)}\right).$$

6: Propose γ' based on $q_2(\cdot|\gamma)$, then $\mathbf{X}_{0:T}, \boldsymbol{\theta}_{0:T}$ will be deterministically updated to $\mathbf{X}'_{0:T}, \boldsymbol{\theta}'_{0:T}$ according to I_0, $R_0, \gamma', \boldsymbol{\delta}_{1:T}, \sigma$

7: Accept $(\gamma', \mathbf{X}'_{0:T}, \boldsymbol{\theta}'_{0:T})$ with acceptance probability

$$a \leftarrow \min\left(1, \frac{\Pr(\mathbf{g}|\boldsymbol{\theta}'_{0:T}, \mathbf{X}'_{0:T}) \Pr(\gamma') q_2(\gamma|\gamma')}{\Pr(\mathbf{g}|\boldsymbol{\theta}_{0:T}, \mathbf{X}_{0:T}) \Pr(\gamma) q_2(\gamma'|\gamma)}\right).$$

8: Let $\mathbf{U} = (\log(R_0), \boldsymbol{\delta}_{1:T}, \log(\sigma))$, then $\mathbf{U}$ *a priori* follows a multivariate normal distribution. Use elliptical slice sampler of obtain $\mathbf{U}'$ and get the updated $R_0', \boldsymbol{\delta}'_{1:T}$ and σ'. $\mathbf{X}_{0:T}$ will be deterministically updated to $\mathbf{X}'_{0:T}$ according to $I_0, R_0', \gamma, \boldsymbol{\delta}'_{1:T}, \sigma'$ based on ODE integration.

Algorithm 8 Updating rule in the MCMC algorithm

1: **Input:** Parameter values from the previous iteration $I_0, R_0, \gamma, \boldsymbol{\delta}_{1:T}, \sigma, \boldsymbol{\xi}_{1:T}$, genealogy $\mathbf{g}$. Proposal function $q_1(\cdot|\cdot)$, $q_2(\cdot|\cdot)$ for updating initial infected and removal rate. Proposal function $q_3(\cdot|\cdot)$, $q_4(\cdot|\cdot)$ for updating the reporting rate and over-dispersion parameter.

2: **Output** Updated parameters values

3: Calculate $\mathbf{X}_{0:T}$, $\boldsymbol{\theta}_{0:T}$ based on $I_0, R_0, \gamma, \boldsymbol{\delta}_{1:T}, \sigma, \boldsymbol{\xi}_{1:T}$.

4: Propose I_0' based on $q_1(\cdot|I_0)$, then $\mathbf{X}_{0:T}$ will be automatically updated to $\mathbf{X}_{0:T}'$ according to I_0', $R_0, \gamma, \boldsymbol{\delta}_{1:T}, \sigma, \boldsymbol{\xi}_{1:T}$. Accept $(I_0', \mathbf{X}_{0:T}')$ with acceptance rate

$$a \leftarrow \frac{L(\mathbf{g}|\boldsymbol{\theta}_{0:T}', \mathbf{X}_{0:T}')L\left(\mathbf{Y}_{1:t}|\mathbf{X}_{0:T}', \rho, \phi\right)\Pr(I_0')q_1(I_0|I_0')}{L(\mathbf{g}|\boldsymbol{\theta}_{0:T}, \mathbf{X}_{0:T})L\left(\mathbf{Y}_{1:t}|\mathbf{X}_{0:T}, \rho, \phi\right)\Pr(I_0)q_1(I_0'|I_0)}.$$

5: Propose γ' based on $q_2(\cdot|\gamma)$, then $\mathbf{X}_{0:T}, \boldsymbol{\theta}_{0:T}$ will be automatically updated to $\mathbf{X}_{0:T}', \boldsymbol{\theta}_{0:T}'$ according to I_0, $R_0, \gamma', \boldsymbol{\delta}_{1:T}, \sigma, \boldsymbol{\xi}_{1:T}$. Accept $(\gamma', \mathbf{X}_{0:T}', \boldsymbol{\theta}_{0:T}')$ with acceptance rate

$$a \leftarrow \frac{L(\mathbf{g}|\boldsymbol{\theta}_{0:T}', \mathbf{X}_{0:T}')L\left(\mathbf{Y}_{1:t}|\mathbf{X}_{0:T}', \rho, \phi\right)\Pr(\gamma')q_2(\gamma|\gamma')}{L(\mathbf{g}|\boldsymbol{\theta}_{0:T}, \mathbf{X}_{0:T})L\left(\mathbf{Y}_{1:t}|\mathbf{X}_{0:T}, \rho, \phi\right)\Pr(\gamma)q_2(\gamma'|\gamma)}.$$

6: Let $\mathbf{U} = (\log(R_0), \boldsymbol{\delta}_{1:T}, \log(\sigma))$, then $\mathbf{U}$ follows multivariate normal distribution. Use elliptical slice sampler of obtain $\mathbf{U}'$ and get the updated $R_0', \boldsymbol{\delta}_{1:T}'$ and σ'. $\mathbf{X}_{0:T}$ will be automatically updated to $\mathbf{X}_{0:T}'$ according to $I_0, R_0', \gamma, \boldsymbol{\delta}_{1:T}', \sigma'$.

7: Let $\mathbf{V} = \boldsymbol{\xi}_{1:T}$ then $\mathbf{V}$ follows multivariate normal distribution.. Using elliptical slice sampler of obtain $\mathbf{V}'$ and get the updated $\boldsymbol{\xi}_{1:T}'$. $\mathbf{X}_{0:T}$ will be automatically updated to $\mathbf{X}_{0:T}'$ according to $I_0, R_0, \gamma, \boldsymbol{\delta}_{1:T}, \sigma, \boldsymbol{\xi}_{1:T}'$.

8: Propose ρ' based on $q_3(\cdot|\rho)$. Accept ρ' with acceptance rate

$$a \leftarrow \frac{L\left(\mathbf{Y}_{1:t}|\mathbf{X}_{0:T}', \rho', \phi\right)\Pr(\rho')q_3(\rho|\rho')}{L\left(\mathbf{Y}_{1:t}|\mathbf{X}_{0:T}, \rho, \phi\right)\Pr(\rho)q_3(\rho'|\rho)}.$$

9: Propose ϕ' based on $q_4(\cdot|\phi)$. Accept ϕ' with acceptance rate

$$a \leftarrow \frac{L\left(\mathbf{Y}_{1:t}|\mathbf{X}_{0:T}, \rho, \phi'\right)\Pr(\phi')q_4(\phi|\phi')}{L\left(\mathbf{Y}_{1:t}|\mathbf{X}_{0:T}, \rho, \phi\right)\Pr(\phi)q_4(\phi'|\phi)}.$$

Appendix B

SIMULATION DETAILS

B.1 Details of the simulation study

B.1.1 Simulation details for Section 3.3.1 of the main text

Here we provide details for the specified sequence/lineage sampling times and number of samples in each simulation scenario:

1. CONST $R_0(t)$: Sampling times: $\mathcal{S} = \{5, 10, 50, 70, 80, 90\}$, number of samples at each time: $\{2, 20, 300, 300, 200, 200\}$.

2. SD $R_0(t)$: Sampling times: $\mathcal{S} = \{5, 10, 50, 70, 80, 90\}$, number of samples at each time: $\{2, 20, 200, 80, 20, 20\}$.

3. NM $R_0(t)$: Sampling times: $\mathcal{S} = \{5, 30, 50, 70, 80, 90\}$, number of samples at each time: $\{2, 50, 250, 100, 20, 20\}$.

B.1.2 Simulation details for Section 3.3.2 of the main text

The $R_0(t)$ trajectory in the simulations is set to

$$R_0(t) = \begin{cases} 1.4 \times 1.015^{t/2}, t \in [0, 30] & t \in [0, 30), \\ 1.750 \times 0.975^{t-30} & t \in [30, 80], \\ 0.494 & t \in (80, 90] \end{cases} \tag{B.1}$$

which is depicted in the left plot of Figure B.1. The initial number of infected individuals is $I_0 = 3$ and the removal rate is set to $\gamma = 0.3$. The population size is fixed to $N = 1,000,000$. Epidemic trajectories are simulated using the SIR Markov jump process (MJP) and are accepted/rejected based on the following criteria:

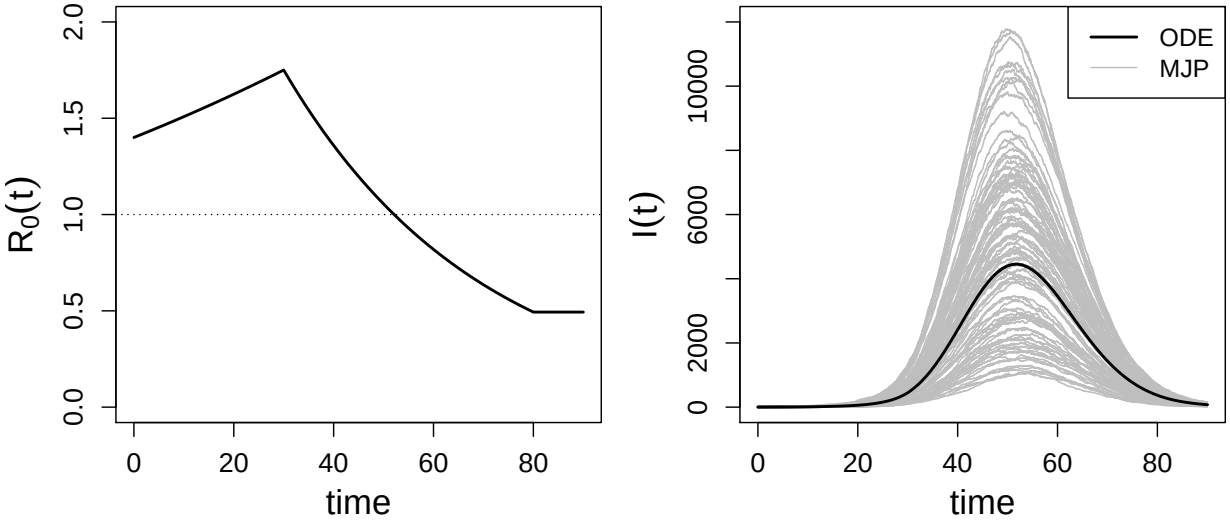

Figure B.1: Repeated simulation setup. Left: $R_0(t)$ trajectory under which the population trajectories are simulated. Right: The 100 simulated prevalence trajectories using MJP and the ODE trajectory under the same parameter setup.

1. Reject the SIR trajectories that ends before time 90. The number of infected individuals should never drop to 0 for $t \in [0, 90]$, i.e. $\min_{t\in[0,90]} I(t) > 0$.

2. Reject the SIR trajectories with extremely high maximum prevalence: the maximum prevalence should be less or equal than 12,000, i.e., $\max_{t\in[0,90]} I(t) \leq 12000$.

3. Reject SIR trajectories with extremely low maximum prevalence. The maximum prevalence should be greater or equal than 600, i.e., $\max_{t\in[0,90]} I(t) \geq 600$.

The 100 simulated SIR prevalence trajectories are shown in the right plot of Figure B.1. We also plot the ODE trajectory under the same parameter setting.

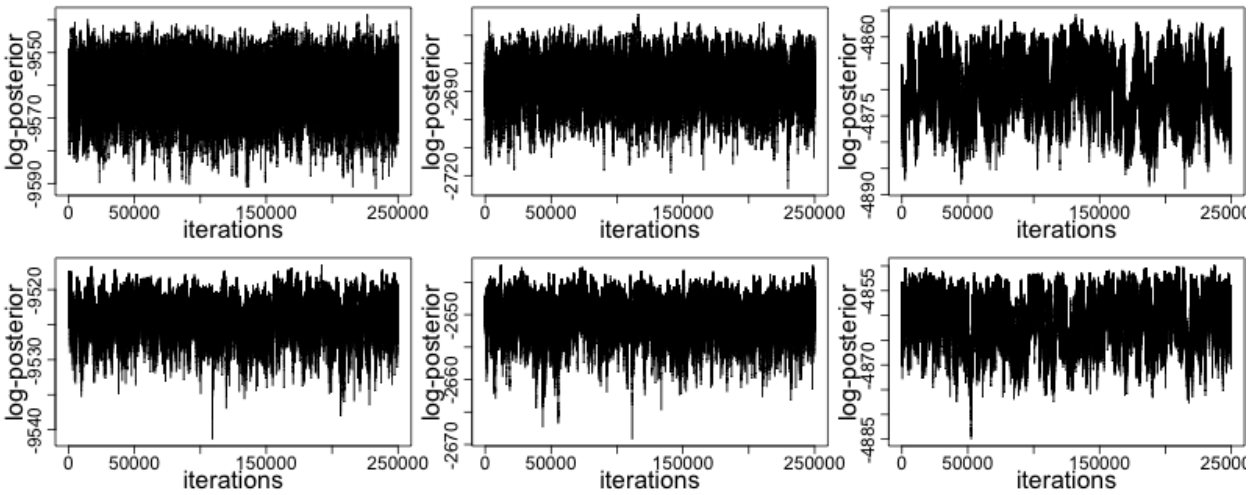

Figure B.2: MCMC trace plots of the log-posterior in the 3 simulation scenarios. Columns correspond to CONST, SD, and NM simulated $R_0(t)$ trajectories. The first row shows the LNA-based results and the second row shows the ODE-based results.

B.1.3 Trace plots

Trace plots for simulations from Section 3.3.1 of the main text

Figure B.2 shows the trace plots of the log-posterior for the LNA-based method and ODE-based method in the three simulation scenarios from Section 3.3.1. The effective sample sizes (ESSs) for all parameters range between 100 to 400.

Trace plots for Ebola data

Figures B.3 and B.4 show trace plots of parameters R_0, I_0, γ, σ for the LNA-based model and ODE-based model respectively applied to the Sierra Leone genealogy. Figures B.5 and B.6 show the analogous trace plots for the analysis of the Liberia genealogy. We also list posterior medians, 95% BCIs, and ESSs for each parameter in the MCMC algorithm in Table B.1.

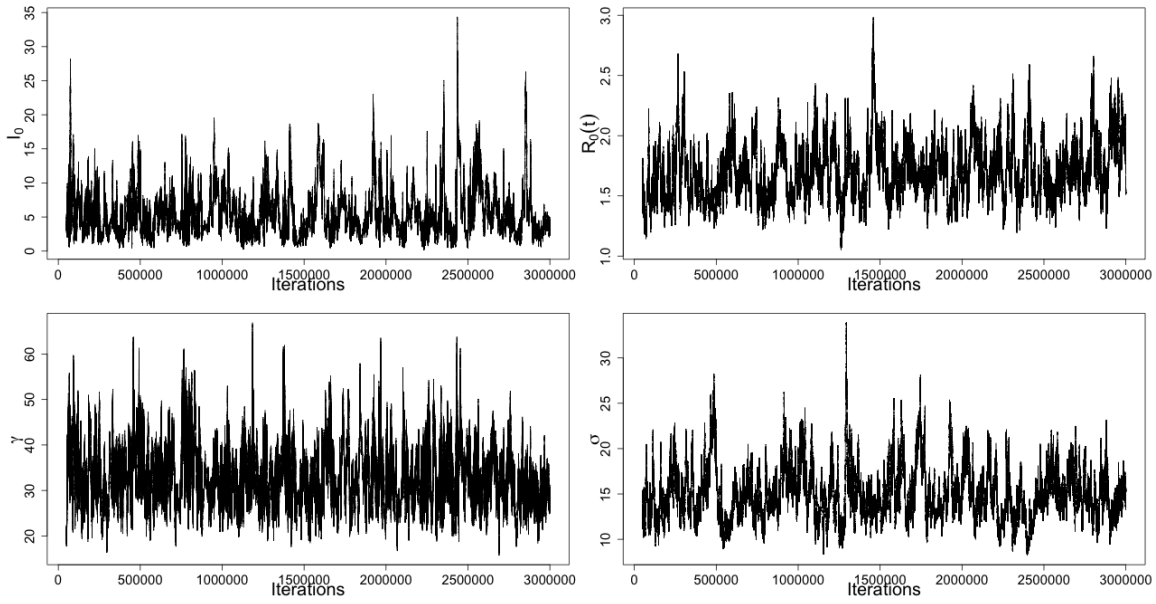

Figure B.3: Trace plots for the LNA-based MCMC algorithm applied to the Ebola genealogy in Sierra Leone. Top right: initial number of infected I_0. Top right: initial basic reproduction number R_0. Bottom left: removal rate γ. Bottom right: precision parameter σ.

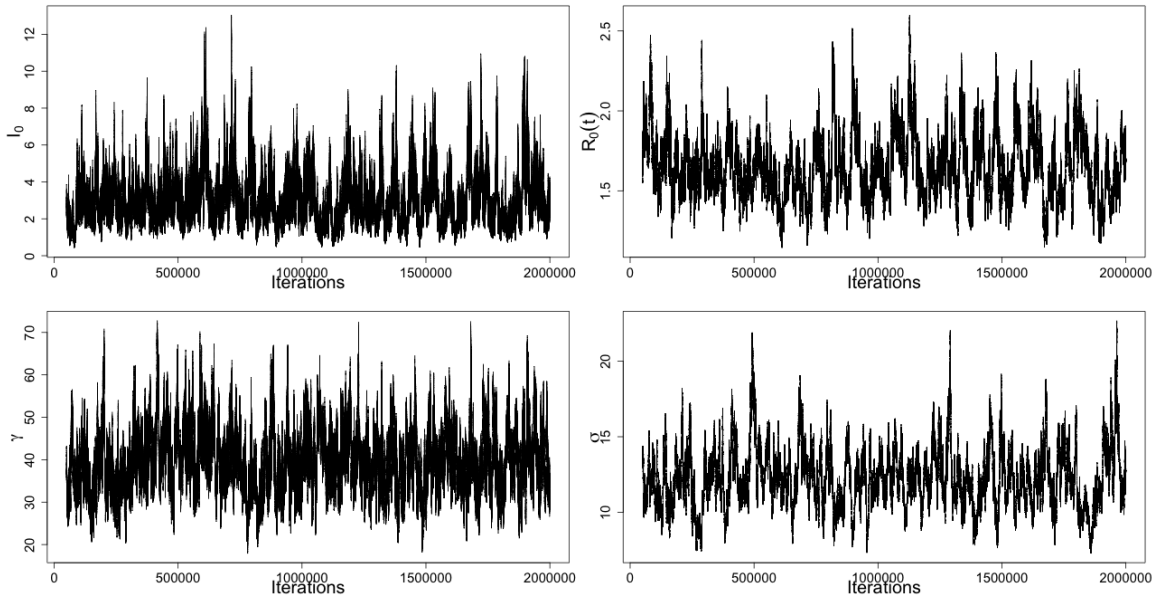

Figure B.4: Trace plots for the ODE-based MCMC algorithm applied to the Ebola genealogy in Sierra Leone. See caption in Figure B.3 for the explanation of the plots.

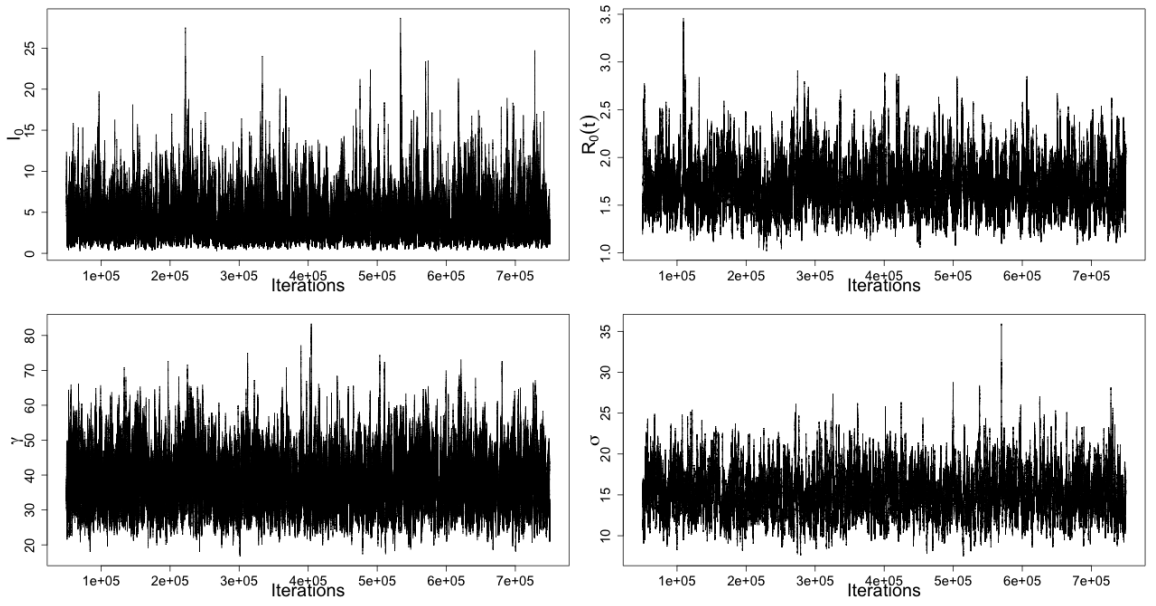

Figure B.5: Trace plots for the LNA-based MCMC algorithm applied to the Ebola genealogy in Liberia. See caption in Figure B.3 for the explanation of the plots.

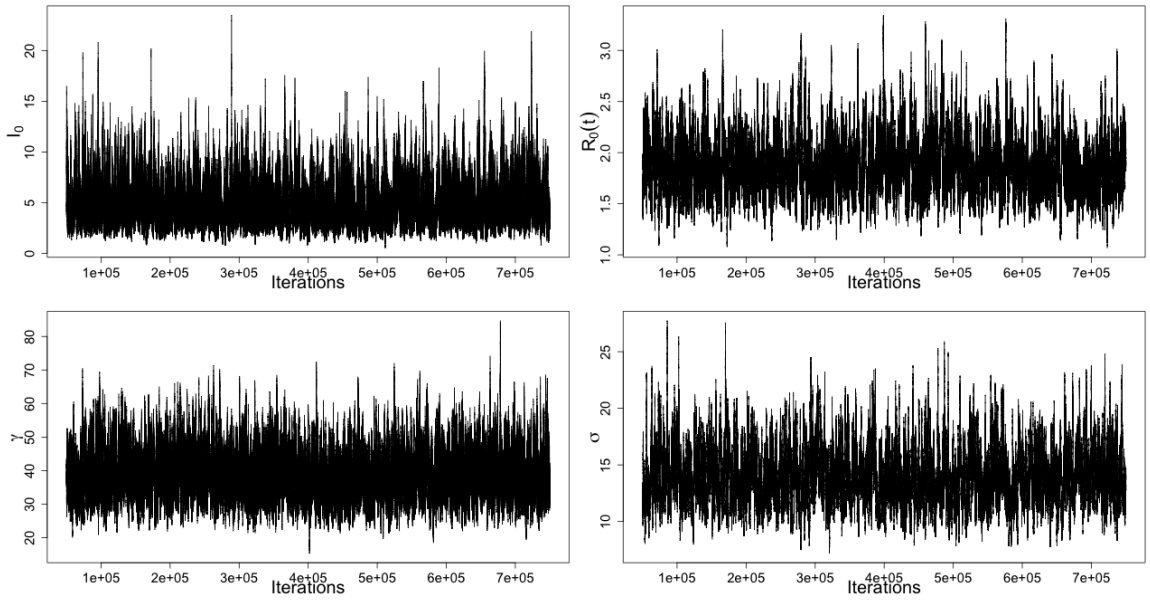

Figure B.6: Trace plots for the ODE-based MCMC algorithm applied to the Ebola data in Liberia. See caption in Figure B.3 for the explanation of the plots.

	Sierra Leone				Liberia			
		post med	95%BCI	ESS		post med	95%BCI	ESS
LNA	I_0	4.63	[1.28,14.41]	151	I_0	3.49	[1.03, 9.95]	1630
	R_0	1.69	[1.33.2.23]	167	R_0	1.67	[1.29,2.24]	942
	γ	32.47	[23.08,47.65]	345	γ	37.21	[25.98,53.13]	1704
	σ	14.71	[10.33,21.84]	141	σ	14.83	[10.41,20.70]	870
ODE	I_0	2.71	[1.10,6.37]	249	I_0	4.31	[1.89,9.27]	1236
	R_0	1.612	[1.30,2.09]	141	R_0	1.83	[1.41.2.44]	796
	γ	39.32	[26.63.55.82]	368	γ	38.31	[27.27.53.43]	1608
	σ	12.13	[8.61,16.98]	113	σ	13.67	[9.78,19.20]	879

Table B.1: Table for posterior median, 95% BCIs and ESSs for MCMC algorithms applied to Ebola data in Sierra Leone and Liberia.

B.2 Prior sensitivity analysis

B.2.1 Simulations based on single genealogy realizations

In Section 3.3.1, we put informative priors on the removal rate γ and explore three different simulation scenarios. Although our LNA-based model successfully recovers the $R(t)$ dynamics and SIR trajectories, the posterior density of the removal rate is not too different from its prior in the SD and NM scenarios. In this section, we investigate sensitivity of our inferences to changes in the prior of the removal rate γ. For the same genealogies and parameter settings as in Section 3.3.1, we assign weakly informative priors to the removal rate γ:

1. CONST $R_0(t)$ scenario: $\gamma \sim \text{lognormal}(-1.7, 0.25)$,

2. SD $R_0(t)$ scenario: $\gamma \sim \text{lognormal}(-1.7, 0.25)$,

3. NM $R_0(t)$ scenario: $\gamma \sim \text{lognormal}(-1.2, 0.25)$.

For each scenario, we fit a LNA-based model using 300,000 MCMC iterations. The first row in Figure B.7 shows the point-wise posterior medians and 95% BCIs for the basic reproduction number trajectories, $R_0(t)$. Our LNA-based method performs well in the CONST and SD scenario. However, for NM scenario, the method fails to fully capture the increase and decrease trend at the beginning and the end of the epidemic. The second row in Figure B.7 depicts the prior and posterior densities of the removal rate γ. The LNA-based method estimates the removal rate with good precision in the CONST scenario. However, for SD and NM scenario, the removal rate posterior densities are similar to the prior densities, but shift to the right from the truth. Posterior summaries of $S(t)$ and $I(t)$ are given in the third and fourth row of Figure B.7. The LNA-based method performs well in recovering the truth in the CONST and SD scenarios. In the NM scenario, the true trajectories are still covered by the wide BCIs, but the model seems to underestimate the $S(t)$ and overestimate $I(t)$.

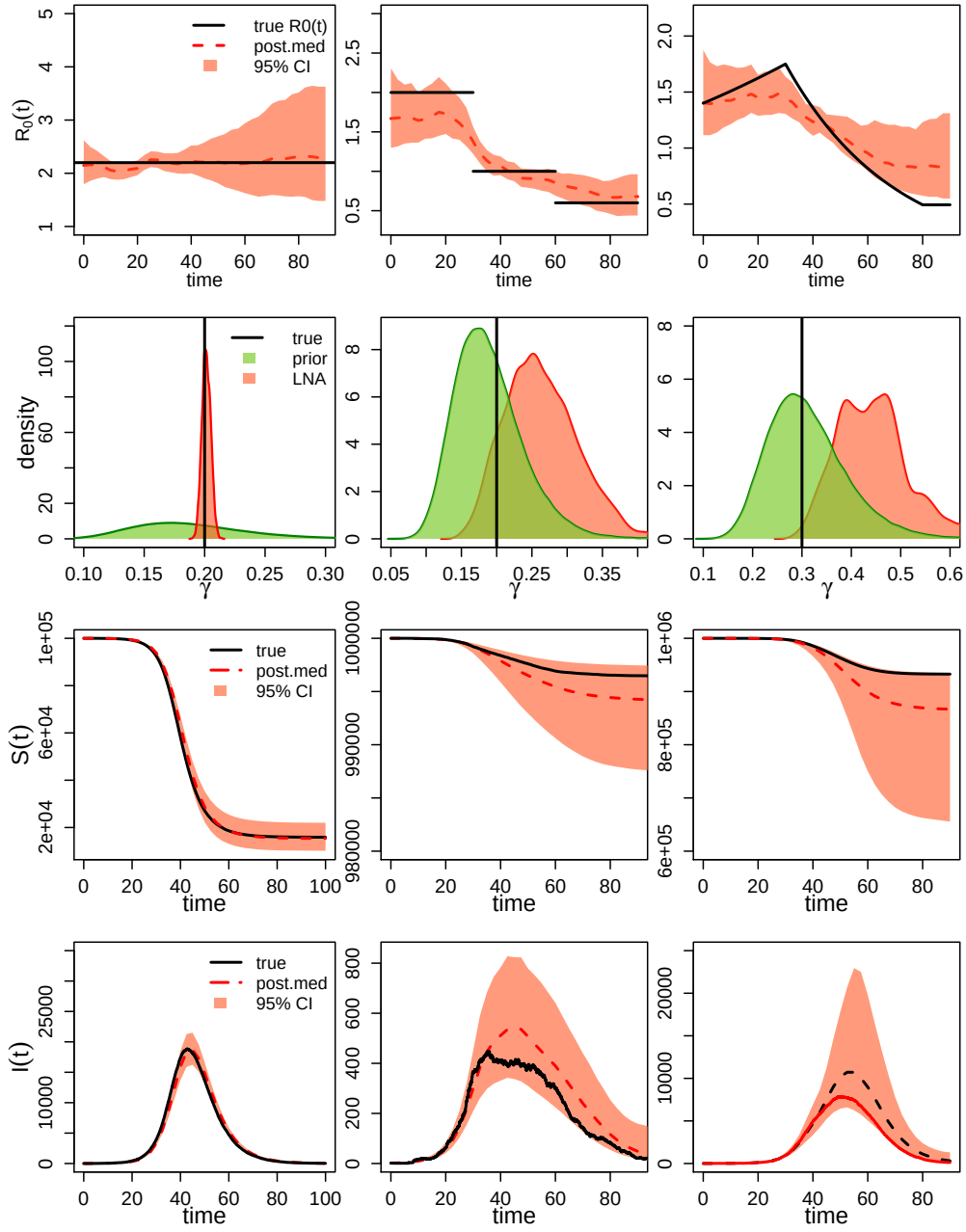

Figure B.7: Analysis of 3 simulation scenarios using the LNA-based method with weakly informative priors. Columns correspond to CONST, SD, and NM simulated $R_0(t)$ trajectories. The first row shows the estimated $R_0(t)$ trajectories for the 3 scenarios, with the black solid lines representing the truth, the red depicting the posterior medians and the red-shaded area showing the 95% BCIs for the LNA-based method. The second row corresponds to the estimation of the removal rate γ. Posterior density curves from the LNA-base method are shown in red lines compared with prior density curve in green lines. The bottom two rows show the estimated trajectories of $S(t)$ and $I(t)$ respectively.

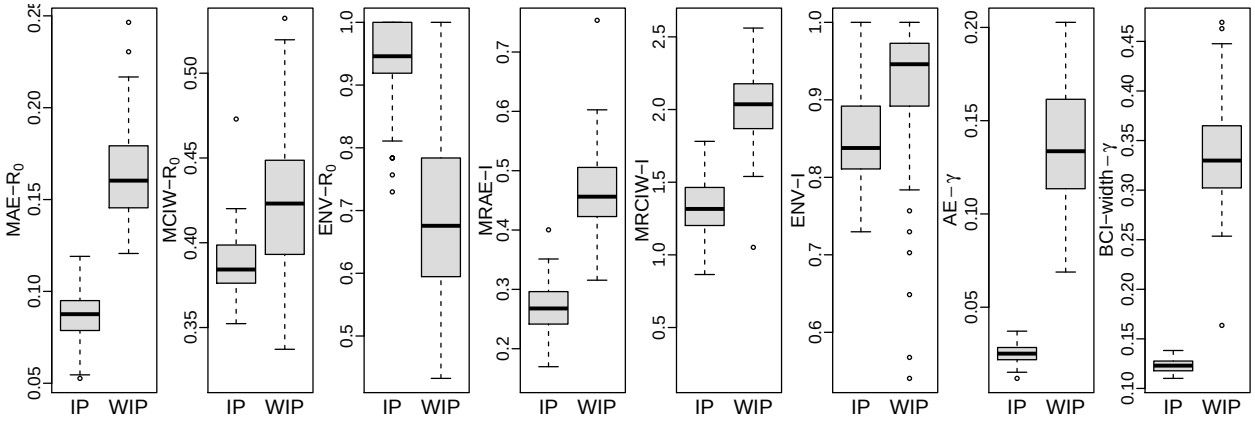

Figure B.8: Boxplots comparing performance LNA-based methods under informative prior (IP) and weakly informative prior (WIP) using 100 simulated genealogies. The first three plots show mean absolute error (MAE), mean credible interval width (MCIW) and envelope for $R_0(t)$ trajectory. The next three plots depict mean relative absolute error (MRAE), mean relative credible interval width (MRCIW), and envelope for $I(t)$ (prevalence) trajectory. The last two plots show the absolute error (AE) and Bayesian credible interval (BCI) width for γ.

B.2.2 Frequentist properties of posterior summaries

In this section, we repeat the simulation study in Section 3.3.2 with a weakly informative prior distribution on recovery: $\gamma \sim \text{lognormal}(-1.2, 0.25)$. We fit LNA-based models to approximate the posterior distribtuion of parameters and latent variables for each genealogy, and compare that with the estimation in Section 3.3.2 with informative prior on γ ($\gamma \sim \text{lognormal}(-1.2, 0.1)$). To evaluate the performance, we use same metric defined in Section 3.3.2 and generate posterior summary boxplots in Figure B.8. Sampling distribution boxplots of $R_0(t)$ posterior summaries are depicted in the left three plots of Figure B.8. The LNA-based model with informative prior (IP) on γ yields significantly lower MAE and MCIW than that with weakly informative prior (WIP). Compared with IP, the LNA-based

model with WIP has really poor envelope for the $R_0(t)$ trajectory.

Sampling distribution boxplot of $I(t)$ posterior summaries, shown in Figure B.8, are similar to $R_0(t)$ results, with IP yields significantly lower MRAE and lower MRCIW. Somewhat counter intuitively, the WIP cases end up with higher coverage for $I(t)$ trajectory. This is likely caused by the wide BCI under WIP.

We also report the absolute error (AE) and 95% BCI widths for the removal rate γ in Figure B.8. Though the WIP prior still centered at the truth of γ, we can see really large absolute error in the removal rate estimation. The 95% BCI coverage for γ under IP is 1. Though WIP yields wider BCIs, the coverage for γ is only 0.65.

B.3 Simulation under other prior settings

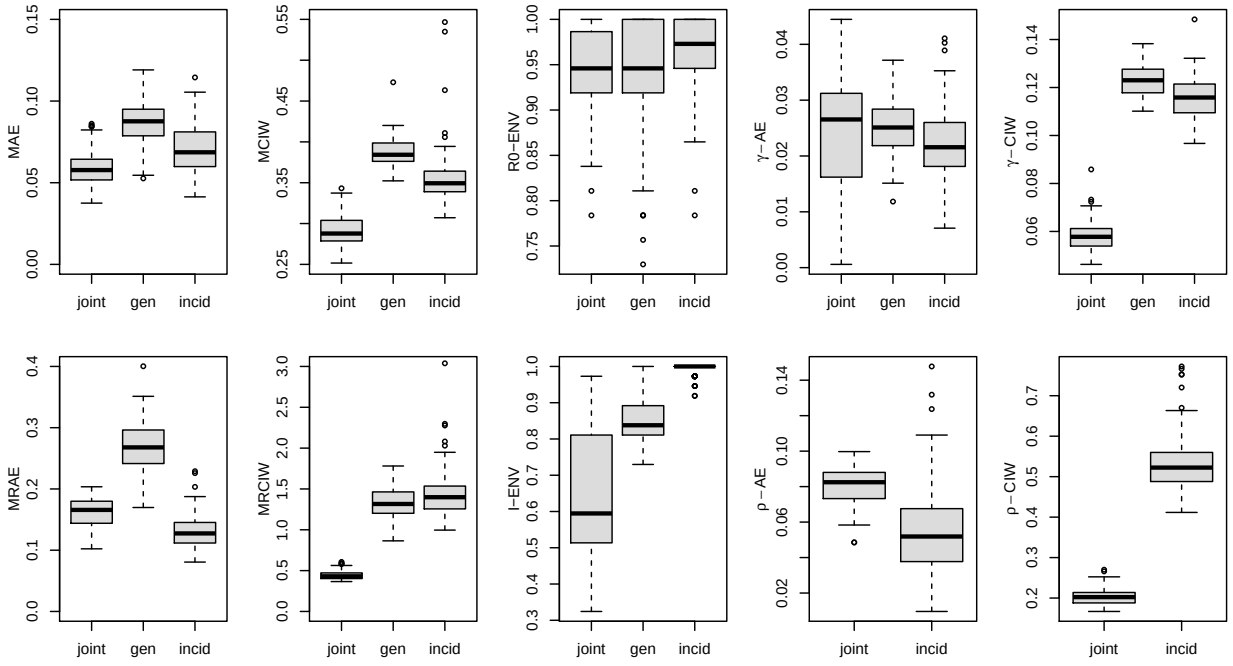

Figure B.9: Boxplot comparing the performance of Joint-based, Gen-based and Incid-based methods using 100 simulated genealogies.

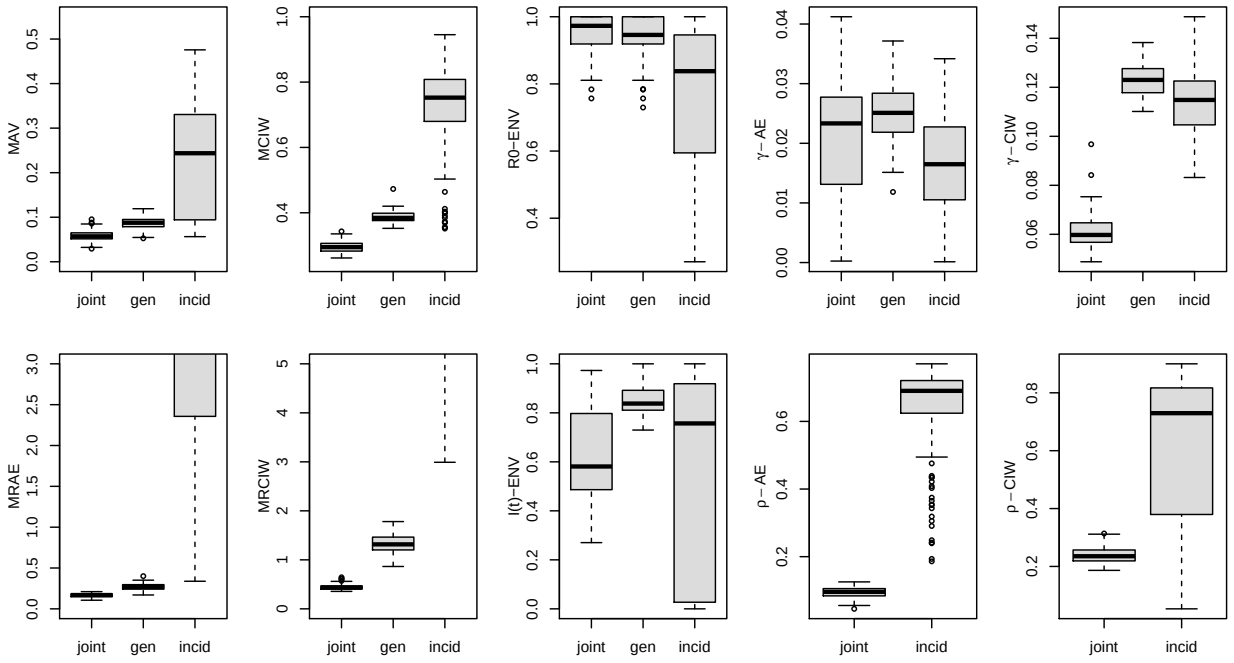

Figure B.10: Boxplot comparing the performance of Joint-based, Gen-based and Incid-based methods using 100 simulated genealogies.

B.4 Simulation Details for Chapter 5

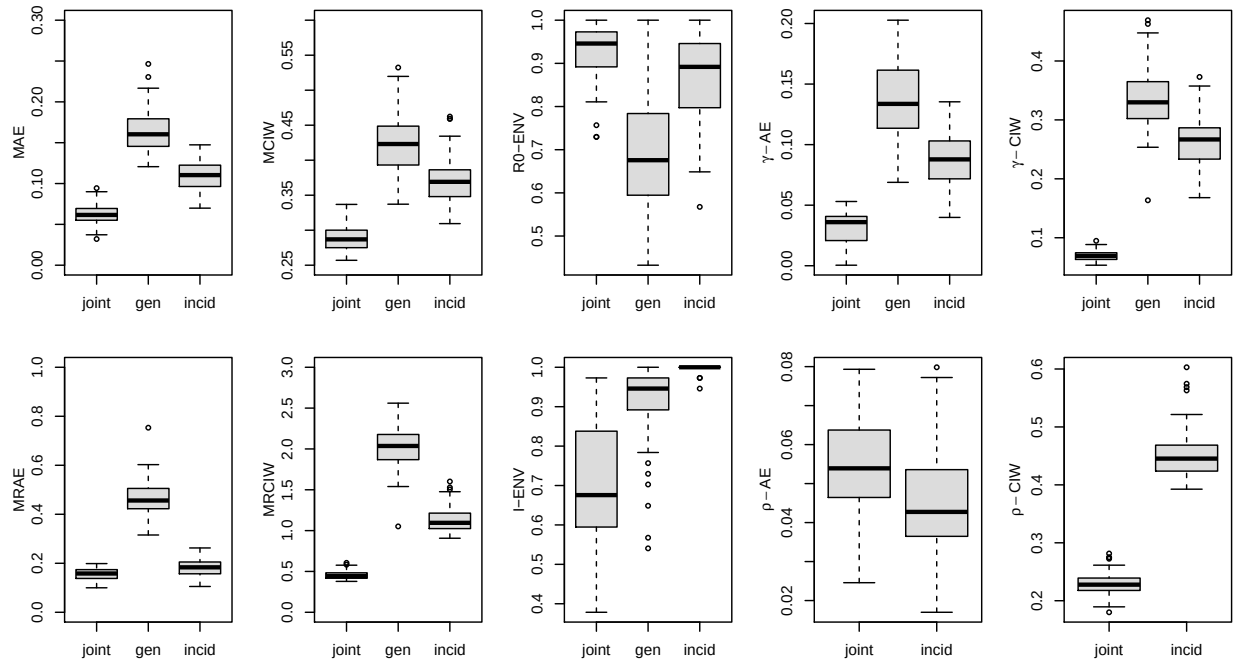

Figure B.11: Boxplot comparing the performance of Joint-based, Gen-based and Incid-based methods using 100 simulated genealogies.

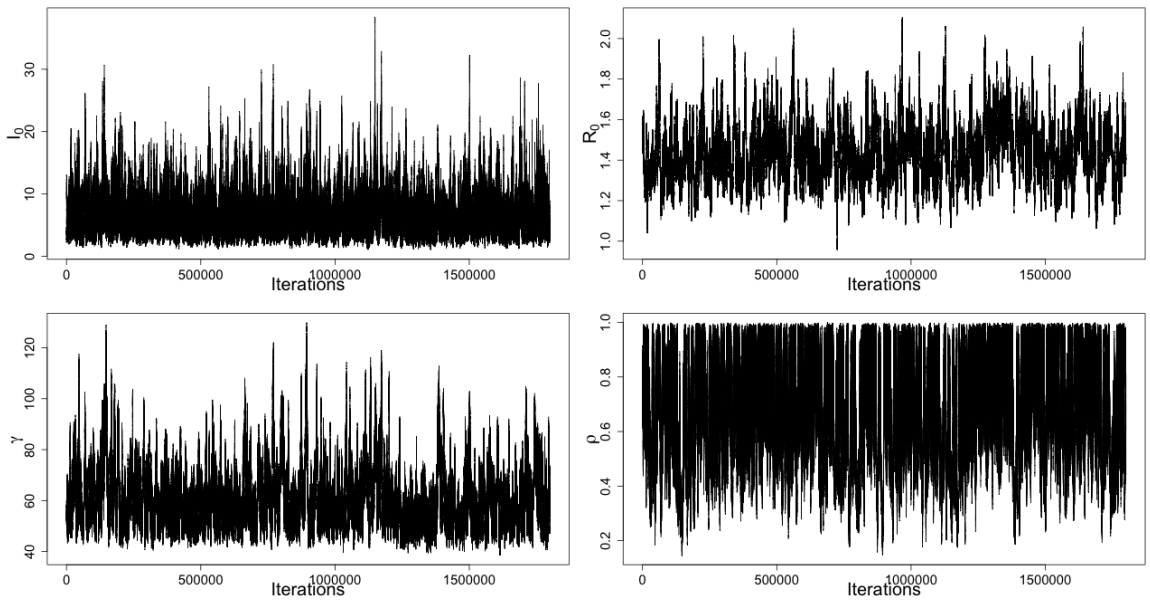

Figure B.12: Trace plots for the ODE-based MCMC algorithm applied to the Ebola data in Liberia. See caption in Figure B.3 for the explanation of the plots.

VITA

Mingwei Tang was born and grew up in Beijing, China. After the national college entrance exam, he moved to Nanjing and studied mathematics and Statistics at Nanjing University. He was an exchange student at University of Melbourne in 2012. Mingwei obtained his bachelor degree in June 2013. After graduation, he went to United States and joined the Ph.D program in the department of Statistics at University of Washington. Mingwei obtained his master degree in Statistics at University of Washington in November 2017.