Bayesian Nonparametric Inference of Effective Population Size Trajectories from Genomic Data

Julia Adela Palacios Roman

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

Doctor of Philosophy

University of Washington

2013

Reading Committee:

Dr. Vladimir N. Minin, Chair

Dr. Elizabeth Thompson

Dr. Peter Guttorp

Program Authorized to Offer Degree:
Statistics

University of Washington

Abstract

Bayesian Nonparametric Inference of Effective Population Size Trajectories from Genomic Data

Julia Adela Palacios Roman

Chair of the Supervisory Committee:
Dr. Vladimir N. Minin
Department of Statistics, University of Washington

Phylodynamics is an area at the intersection of phylogenetics and population genetics that aims to reconstruct population size trajectories from genetic data. Phylodynamic methods rely on a standard framework based on the coalescent, a stochastic process that generates genealogies connecting randomly sampled individuals from the population of interest. The shape of a genealogy is influenced by the effective population size trajectory and, under the coalescent framework, the times at which genealogical lineages coalesce contain information about population size dynamics. I show that these coalescent times can be viewed as realization of a point process and that estimation of population size trajectories is equivalent to estimating a conditional intensity of the coalescent point process. This thesis presents a Gaussian process-based Bayesian nonparametric approach to estimate effective population size trajectories. First, I summarize and discuss current approaches to statistical inference in phylodynamics. Next, I demonstrate how recent advances in Gaussian process-based nonparametric inference for Poisson processes can be extended to Bayesian nonparametric estimation of population size dynamics when the genealogy is assumed fixed. I compare our Gaussian process (GP) approach to one of the state of the art Gaussian Markov random field (GMRF) methods for estimating population trajectories. Next, I show that when a representative genealogy is available, perhaps estimated using one of the phylogenetic reconstruction methods, we can replace Markov chain Monte Carlo (MCMC) methods to

perform inference by integrated nested Laplace approximation (INLA). This approximation, actively used in spatial statistics, results in recovery of population size trajectories that is much faster than current MCMC-based methods. However, the INLA algorithm cannot be generalized to a more realistic setting, where one starts with molecular data instead of a genealogy. Therefore, I return to MCMC to extend the GP approach to infer population size trajectories from molecular data directly. I test the GP-based method on simulated and real data. For real data, I estimate effective number of infected individuals with Hepatits C virus in Egypt from 1700 to 1993, the effective number of individuals infected with human influenza A virus in New York between 2000 and 2005 and effective number of Bisons across Beringia from present time to 100,000 years ago.

TABLE OF CONTENTS

LIST OF FIGURES

ACKNOWLEDGMENTS

I would like to thank my advisor Vladimir Minin for introducing me to statistical inference in genetics and sharing with me the research questions that led to this thesis. I am very grateful to Vladimir Minin for all the guidance and constant encouragement that he has given me during my time at the University of Washington, giving me opportunities to attend conferences and workshops, and helping me plan my future career. I wish to thank my committee member Elizabeth Thompson for her useful comments in the presentations about my work, including my stochastic processes prelim, statistical genetics seminars, population genetics lunch seminars, general and final exams. I am very thankful for the mentoring she has provided during my graduate studies and her support in writing me letters of recommendation, including a letter to particpate in a long term program on mathematical and computational approaches in high-throughput genomics at IPAM-UCLA. My committee member Peter Guttorp, also provided thoughtful suggestions that continue to guide my learning on stochastic processes and point processes. I am very thankful for his encouragement during my stochastic processes prelim, general and final exams and for writting letters of recommendation. I have also had the honor of receiving constant feedback from Joe Felsenstein, providing invaluable guidance on my work in population genetics. I would like to thank Marc Suchard and Mandev Gill for helping me write a review chapter on phylodynamics. Thank you Marc for introducing me to the BEAST and for helping me do the BEAST implementation more efficient. In addition to my interactions with my committee members, the students in the Minin group have played an important role in my graduate studies. My friends Chris Glazner and Peter Chi even went beyond in reviewing my manuscripts and providing me helpful suggestions on my writing and oral presentations. My friend and officemate Amanda Allen for her constant support and tolerance. I have enjoyed all our fun experiences and stressful momments working on homeworks or for exams

in the middle of the night. Most of my knowledge in theoretical statistics and probability came from Prof. Jon Wellner. I am very thankful in having the opportunity of taking his classes and even let me adventure in exploring diffussion processes theory in genetics. Thank you Prof. Wellner for your trust and writing me letters of recommendation. I had a great opportunity to work with Adam Leaché on an evolutionary study of lizards. Thank you for your patience and for letting me join you on your very interesting research. I learned a great deal in working with you and I hope we can continue collaboration in the future. To my family, thank you for all your support. Thank you Iván for joining me in this incredible journey and your constant support and guidance. I would like to thank the Mexican council of science Conacyt for the fellowship that made my graduate studies possible.

DEDICATION

To my baby girl, Ximena

Chapter 1

INTRODUCTION

Changes in population size affect the variability of gene frequencies in natural populations, allowing genetic variation in present-day and recent-past molecular sequence data to help recover the more distant past demographic history of the population. This variability also enables researchers to examine the factors driving past population dynamics and to establish molecular surveillance of emerging infectious diseases. For example, Campos et al. (2010a) analyze ancient and modern musk ox mtDNA samples dated from $56,900$ radiocarbon years old to present and recover the population dynamics throughout the late Pleistocene to the present; 63 RNA sequences of hepatitis C virus (HCV) obtained in 1993 effectively reveal the dynamics of HCV infections in Egypt over the past century (Pybus et al., 2003); and human influenza A/H3N2 subtype sequences sampled over a 12 year period in New York state return estimates of the seasonal population dynamics of human influenza A/H3N2 (Rambaut et al., 2008).

In 1931, Sewall Wright introduced the concept of the *effective population size* of a population (Wright, 1931). The effective population size is the number of breeding individuals in an idealized population that is randomly mating and that has the same gene frequency changes as the population being studied. The study of effective population sizes has grown into a central theme in population genetics ever since. Molecular epidemiologists often employ estimates of effective population size to approximate census population size (number of infected individuals) by incorporating knowledge about the expected number of molecular sequence substitutions (e.g., DNA substitutions) per calendar time unit along the inferred genealogy, generation time in calendar units and the population variance in number of offspring (Wakeley and Sargsyan, 2008). However, even when such prior information is available, interpreting estimates of the effective population size remains challenging, especially in studies of infectious diseases(Frost and Volz, 2010).

Initially, most studies focused on two summary statistics of a multiple alignment of molecular sequences – the number of segregating sites (Watterson, 1975) and the mean number of nucleotide differences between two sequences in a sample (Tajima, 1983) – to quantify the effective population size. However, with the introduction of coalescent theory (Kingman, 1982) and the coalescent with variable population size (Slatkin and Hudson, 1991a; Griffiths and Tavare, 1994), the genealogical relationships among the sequences in a sample have begun to inform estimates of effective population size and its dynamics over time.

The coalescent provides a probability model that describes the relationship between the coalescent times in a gene genealogy and the effective population size (Nordborg, 2001; Hein et al., 2005). Some coalescent-based methods assume that a single genealogy is available (Fu, 1994; Pybus et al., 2000) and others produce estimates of effective population size trajectories directly from molecular sequence data through an unknown genealogy (Kuhner et al., 1995; Drummond et al., 2002, 2005; Minin et al., 2008). Methods that take into account the genealogical uncertainty more efficiently use the information present in the data (Felsenstein, 1992); however, the achieved efficiency comes at substantial computational cost of Monte Carlo methods needed to integrate over the space of genealogies. When molecular sequences contain sufficient phylogenetic information, the computationally expensive Monte Carlo can be omitted in favor of inferring the population size trajectory from a single estimated genealogy (Pybus et al., 2000; Minin et al., 2008). However, such a two step estimation procedure can lead to substantial underestimation of uncertainty in population size estimates and, therefore, should be used with caution (Minin et al., 2008).

When all molecular sequence data are sampled at the same time under an *isochronous sampling* scheme, it is possible to estimate $\theta = 4N_e\mu$, where the effective population size N_e is measured in units of generations in a diploid population and μ represents the substitution rate per site per generation. If one has access to an independent estimate of μ via previous studies or from other sources, one can estimate N_e directly, otherwise N_e and μ remain confounded. Felsenstein and Rodrigo (1999) extend the coalescent model to incorporate genealogies with sequence data sampled at different times under a *heterochronous sampling* scheme. Here, the sampling times of noncontemporary sequences can provide information

about μ and help to identify N_e and μ separately (Rambaut, 2000; Pybus et al., 2003). Such serially sampled data are common in studies of ancient DNA and of rapidly evolving viruses.

Scientific interest often lies in the changes of the effective population size over time or, in other words, in the effective population size trajectory, $N_e(t)$. Most inferential tools for estimating such a trajectory assume a simple *parametric* form of $N_e(t)$, such as exponential or logistic growth. Maximum likelihood (Griffiths and Tavare, 1994; Kuhner et al., 1998) and full Bayesian (Drummond et al., 2002) approaches provide estimates of the parameters that characterize these functional forms. However, for poorly studied populations, a simple parametric form remains difficult to justify and more flexible *nonparametric* methods are preferable.

Over the last 15 years, the development of nonparametric methods to infer $N_e(t)$ has blossomed. A common characteristic of most of these methods is an underlying assumption of a piece-wise constant or linear trajectory describing $N_e(t)$. Early methods, such as the *skyline plot* (Pybus et al., 2000) and its regularized version, the *generalized skyline plot* (Strimmer and Pybus, 2001), provide fast but noisy estimates of $N_e(t)$ from a fixed genealogy. Drummond et al. (2005), who call their method the *Bayesian skyline plot*, and Opgen-Rhein et al. (2005) introduced more sophisticated multiple change-point models to estimate population trajectories in a Bayesian framework. The most popular implementation of the Bayesian skyline plot, available in the Bayesian Evolutionary Analysis by Sampling Trees software package (BEAST) (Drummond et al., 2012), starts from the sequence data and models $N_e(t)$ as a piece-wise constant function with a fixed number of changes *a priori*. Opgen-Rhein et al. (2005) propose a similar model, but these authors infer the number of change-points simultaneously with other model parameters. However, in contrast to Drummond et al. (2012), Opgen-Rhein et al. (2005) condition on a single genealogy.

Recently, Minin et al. (2008) and Palacios and Minin (2013) have proposed and implemented Bayesian nonparametric approaches that rely on Gaussian processes (GPs) for prior specification of $N_e(t)$. GP-based nonparametric methods enjoy a long and successful history in spatial statistics (Cressie, 1993) and machine learning literature (Rasmussen

and Williams, 2006), but GP-based inference started to appear in the evolutionary genetics literature only recently. In the context of effective population size estimation, GP-based methods allow for more flexible prior specification than previous approaches based on piecewise continuous prior formulations. The Bayesian skyride model of Minin et al. (2008) *a priori* assumes that the population size trajectory follows a discretized log-Gaussian process, while the continuous time GP-based method of Palacios and Minin (2013) puts a continuously defined exponential Gaussian process prior on the population size trajectory.

Chapter 2

MODERN BAYESIAN NONPARAMETRIC METHODS FOR PHYLODYNAMICS

The main goal of this chapter is to provide a general overview of modern Bayesian nonparametric methods for inference of effective population size trajectories. First, we formulate the problem of effective population size estimation in general terms within a Bayesian framework. We then develop a notation that allows us to work with both isochronous and heterochronous sampling, and unifies the presentation of multiple change-point and Gaussian process models. Next, we show that coalescent-based estimation of population size trajectories can be thought of as estimation of an intensity function of a temporal point process. This point process representation allows us to make explicit connections between Bayesian nonparametric phylodynamics methods and Bayesian nonparametric estimation of an intensity of an inhomogeneous Poisson process. This connection is important, because borrowing statistical and computational techniques from the point process literature should be fruitful for extending Bayesian nonparametric phylodynamics models in the future. To illustrate the performance of Bayesian nonparametric phylodynamics, we analyze simulated and real data.

2.1 General Model Formulation

2.1.1 Likelihood for Sequence Alignment

Let $\mathbf{Y}$ denote an $n \times L$ sequence alignment matrix, where n represents the number of individuals randomly sampled from the population of interest and L refers to the length of the sequence alignment. The sequence alignment usually represents DNA or RNA sequences from a protein coding region or a "gene." We assume that sites within the alignment are fully linked, meaning that there is no recombination possible between the sequences. This last assumption implies that we can postulate an existence of a genealogy/phylogeny $\mathbf{g}$,

in the form of a rooted bifurcating tree, that represents ancestral relationships among the sampled individuals. We assume that sequence alignment $\mathbf{Y}$ is generated by a substitution process, defined by a parameter vector $\mathbf{m}$, acting on the genealogy $\mathbf{g}$. This construction yields the following likelihood function:

$$P(\mathbf{Y} \mid \mathbf{g}, \mathbf{m}). \tag{2.1}$$

Although specifics of the substitution process are not important for the general model formulation, it is commonly assumed that at each alignment site substitutions occur according to a continuous-time Markov chain. In such cases the likelihood function (2.1) is referred to as the Felsenstein likelihood (Felsenstein, 1981).

2.1.2 Coalescent Prior

We proceed in a hierarchical Bayesian framework by putting a coalescent prior distribution on the genealogy $\mathbf{g}$. When all sequences are sampled at effectively the same time (*isochronous sampling*), this prior becomes

$$P(\mathbf{g} \mid N_e(t)) \propto \prod_{k=2}^{n} \frac{C_k}{N_e(t_{k-1})} \exp\left[-C_k \int_{t_k}^{t_{k-1}} \frac{1}{N_e(t)} dt\right], \tag{2.2}$$

where $t_n = 0$ denotes the time of sampling, $0 < t_{n-1} < ... < t_1$ are *coalescent times*, times at which two lineages meet their most recent common ancestor, and $C_k = \binom{k}{2}$ is the coalescent factor that depends on the number of lineages $k = 2, ..., n$. The prior (2.2) is a product of $(n-1)$ conditional densities of coalescent times, where each density is quadratic in the number of lineages and inversely proportional to the effective population size. Figure 2.1 shows an example of a population that experiences growth and then decay in population size. To appreciate the effect of the effective population size on the distribution of the coalescent times, consider coalescent times t_5, t_4, t_3, and t_2 in Figure 2.1. The elapsed times since the last coalescent event are long for t_5, t_4, owing to relatively large population sizes in the vicinity of these times. In contrast, coalescences occur vary fast at times t_3 and t_2 when the population size becomes small.

The *heterochronous coalescent* or *serially sampled coalescent* arises when not all se-

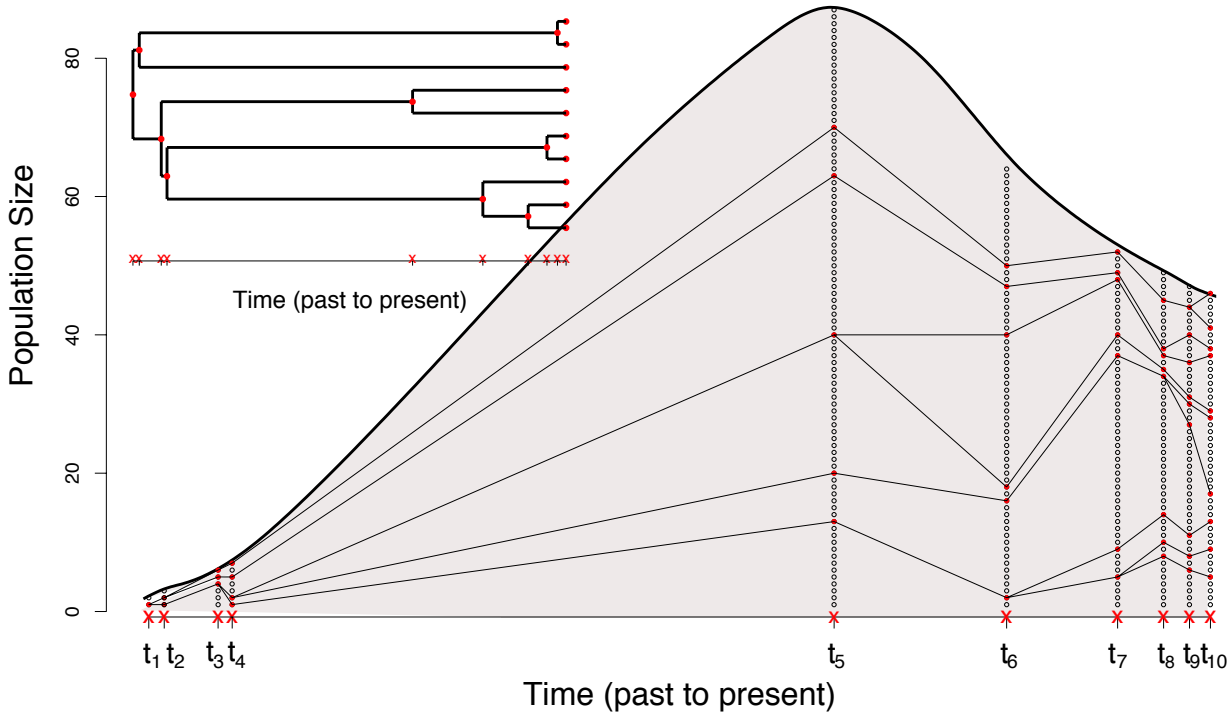

Figure 2.1: Example of a genealogy of 10 individuals randomly sampled at time t_{10} (red circles) from the population depicted as black circles at time t_{10} in the left plot. When we follow their ancestry back in time, two of the lineages coalesce at time t_9, the rest of the lineages continue to coalesce until the time to the most recent common ancestor of the sample at time t_1. The population size trajectory is shown as the solid black curve. When the population size is large (around t_5), any pair of lineages coming from time t_5 (red circles at t_5) take longer to meet a common ancestor at time t_4. The figure in the top left corner shows the genealogy reconstructed by following the ancestry of the 10 individuals. It is an aligned representation of the genealogy depicted in the main plot.

quences are sampled at the same time (Figure 2.2). In this case, the coalescent prior is

$$P(\mathbf{g} \mid N_e(t)) \propto \prod_{k=2}^{n} \frac{C_{0,k}}{N_e(t_{k-1})} \exp\left[-\int_{I_{0,k}} \frac{C_{0,k}}{N_e(t)} dt - \sum_{i=1}^{m} \int_{I_{i,k}} \frac{C_{i,k}}{N_e(t)} dt \right], \tag{2.3}$$

where $t_n = 0 < t_{n-1} < \cdots < t_1$ denote the coalescent times as before, but the coalescent factor $C_{i,k} = \binom{n_{i,k}}{2}$ depends on the number of lineages $n_{i,k}$ in the interval $I_{i,k}$ defined by coalescent times and sampling times $s_m = 0 < s_{m-1} < ... < s_1 < s_0$ of $n_m, ..., n_1$ sequences respectively, $\sum_{j=1}^{m} n_j = n$. Here, time is measured backwards. Present time is $t_1 = 0$ and the time to the most recent common ancestor is t_1 units ago. We denote intervals that end with a coalescent event by

$$I_{0,k} = (\max\{t_k, s_j\}, t_{k-1}], \text{ for } s_j < t_{k-1} \text{ and } k = 2, \ldots, n, \tag{2.4}$$

and intervals that end with a sampling event by

$$I_{i,k} = (\max\{t_k, s_{j+i}\}, s_{j+i-1}], \text{ for } s_{j+i-1} > t_k \text{ and } s_j < t_{k-1}, k = 2, \ldots, n. \tag{2.5}$$

The main difference between equations (2.2) and (2.3) is in that the conditional density for the next coalescent time t_{k-1} is the product of the density of the coalescent time $t_{k-1} \in I_{0,k}$ and the probability of not having a coalescent event during the period of time spanned by intervals $I_{1,k}, ..., I_{m_k,k}$, where m_k is the number of intervals that end with a sampling event in $(t_k, t_{k-1}]$ (Felsenstein and Rodrigo, 1999). Lastly, we point out the densities on the right-hand sides of equations (2.2) and (2.3) are in fact densities of the coalescent times. The corresponding genealogical densities are obtained by dropping the factors C_k and $C_{0,k}$, because for a given genealogy, we do not need to enumerate all possible orders in which lineages coalesce one pair at a time.

2.1.3 Posterior Inference

We have now defined the likelihood function (2.1) and two priors for genealogies (2.2) and (2.3), corresponding to contemporaneously and serially sampled data. The next step is to define a hyper-prior $P(N_e(t) \mid \boldsymbol{\theta})$ for the effective population size trajectory with hyper-parameters $\boldsymbol{\theta}$, accompanied by their own prior $P(\boldsymbol{\theta})$. Various approaches to this prior specification will be discussed in the next section. We also need a prior for the substitution

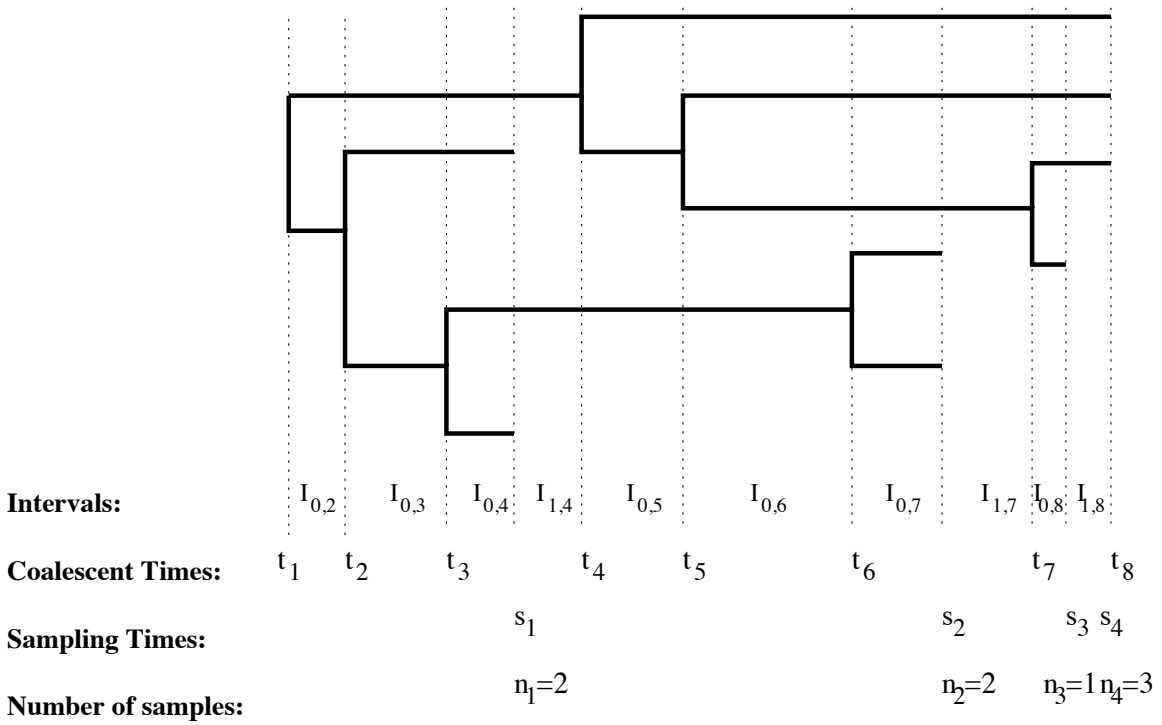

Figure 2.2: Example of a genealogy relating serially sampled sequences (heterochronous sampling). The number of lineages changes every time we move between intervals ($I_{i,k}$). Each endpoint of an interval is a coalescent time ($\{t_k\}_{k=1}^{n}$) or a sampling time ($\{s_j\}_{j=1}^{m}$). The number of sequences sampled at time s_j is denoted by n_j.

process parameters, $P(\mathbf{m})$. Such a prior heavily depends on the choice of the substitution model and usually follows standard practice in Bayesian phylogenetics (Suchard et al., 2001; Ronquist et al., 2012).

We are now ready to define the posterior distribution of all model parameters:

$$P(\mathbf{g}, \mathbf{m}, N_e(t), \boldsymbol{\theta} \mid \mathbf{Y}) \propto P(\mathbf{Y} \mid \mathbf{g}, \mathbf{m}) P(\mathbf{m}) P(\mathbf{g} \mid N_e(t)) P(N_e(t) \mid \boldsymbol{\theta}) P(\boldsymbol{\theta}), \tag{2.6}$$

where we assume that the substitution process parameters $\mathbf{m}$ and genealogy $\mathbf{g}$ are *a priori* independent. If the genealogy is assumed to be fixed, the posterior distribution simplifies as follows:

$$P(N_e(t), \boldsymbol{\theta} \mid \mathbf{g}) \propto P(\mathbf{g} \mid N_e(t)) P(N_e(t) | \boldsymbol{\theta}) P(\boldsymbol{\theta}). \tag{2.7}$$

2.2 Priors on Effective Population Size Trajectory

2.2.1 Multiple Change-Point Models

A Bayesian multiple change-point method, developed by Opgen-Rhein et al. (2005), is implemented for a fixed genealogy with contemporaneous data in the `R` package `APE` (Paradis et al., 2004). The authors assume a piece-wise constant trajectory

$$N_e(t) = \sum_{j=1}^{k+1} \gamma_j 1_{(a_{j-1}, a_j]}(t), \tag{2.8}$$

where the change-points $a_1, ..., a_k$ are *a priori* uniformly distributed in $[0, t_1]$, $a_{k+1} = t_1$ is the time to the most recent common ancestor, $a_0 = 0$ is the sampling time of the contemporaneous sequences and for an interval A

$$1_A(t) = \begin{cases} 1 & \text{if } t \in A, \\ 0 & \text{otherwise.} \end{cases}$$

The number of change-points, k, has a truncated Poisson prior with hyperparameter $\lambda \sim \text{Gamma}(a, b)$. *A priori*

$$\gamma_j \sim \text{Gamma}(\alpha_j, \beta_j) \text{ for } j = 1, \ldots, k+1. \tag{2.9}$$

The authors approximate the posterior distribution of $N_e(t)$ and other model parameters by Markov chain Monte Carlo (MCMC) sampling.

The *Bayesian skyline plot* (Drummond et al., 2005) is implemented in `BEAST` (Drummond et al., 2012) to estimate population size trajectories directly from serially sampled sequence data. This method considers the following piece-wise constant prior

$$N_e(t) = \sum_{j=1}^{k+1} \gamma_j 1_{(a_{j-1}, a_j]}(t), \tag{2.10}$$

with $\gamma_k > 0$ for $k \leq n-2$ change-points $a_1, ..., a_k$. These change-points are an ordered subset of the coalescent times $\{t_{n-1}, ..., t_2\}$, and $a_0 = t_n = 0$ and $a_{k+1} = t_1$. That is, the effective population size changes only at some coalescent events. The number of change-points is fixed by the user. The heights of step function (2.10) follow an auto-exponential prior:

$$\gamma_j \sim \text{Exponential}(\gamma_{j-1}), \text{ for } j = 1, \ldots, k+1. \tag{2.11}$$

The first step function height, γ_1 receives an improper scale-invariant prior with density $f(x) \propto 1/x$. Drummond et al. (2005) approximate the posterior distribution of $N_e(t)$ and other model parameters, including the genealogy, using MCMC.

2.2.2 Coalescent as a Point Process

Before proceeding to the GP-based priors on the effective population size trajectory, we take a moment to view the coalescent as a point process. First, we notice that equations (2.2) and (2.3) suggest that once we extract the coalescent times from a genealogy , the genealogy does not provide any further information about the effective population size dynamics. Next, recall that the coalescent and its companion ancestral process, that tracks the number of genealogical lineages, are continuous-time Markov chains and the coalescent times are transition times of these two chains (Tavaré, 2004). Therefore, the coalescent times form a Markov point process with a conditional intensity inversely proportional to $N_e(t)$ (Andersen et al., 1995). See (Palacios and Minin, 2013) for a more detailed exposition.

Benefits of the above reflection may not be immediately obvious, but viewing the coalescent as a point process allows us to recognize that the problem of reconstructing population size trajectory from coalescent times closely resembles the problem of estimating an intensity function of the inhomogeneous Poisson process. The latter task is well studied in the

statistical literature, providing opportunities to adapt point process estimation tools to the coalescent framework. In fact, the multiple change-points discussed above can be viewed as modifications of the change-point approach to Poisson process intensity estimation (Raftery and Akman, 1986; Green, 1995).

The aforementioned change-point modeling falls into the area of nonparametric statistics, in which the number of parameters can grow indefinitely with dimensionality of the data. One popular alternative to the change-point approaches is GP-based nonparametric inference. A GP is a stochastic process such that any finite sample taken from the realization of the process has a multivariate normal (MVN) distribution (Rasmussen and Williams, 2006). For example, Brownian motion and Ornstein-Uhlenbeck processes are GPs. GP-based models are very flexible and amenable to multivariate extensions, making these models dominant players in the point process and spatial statistics literature (Cressie, 1993). We now turn to describing the methods that use GP-based approaches to nonparametric inference of population dynamics.

2.2.3 Gaussian Process-Based Nonparametrics

There are two GP-based approaches to estimation of effective population size trajectories. The first approach – the *Bayesian skyride* (Minin et al., 2008) – assumes that given a genealogy, the effective population size trajectory is a piece-wise constant function with change-points coinciding with the coalescent times:

$$N_e(t) = \sum_{k=2}^{n} \exp(\gamma_k)\, 1_{(t_k, t_{k-1}]}(t). \tag{2.12}$$

In contrast to the *Bayesian skyline*, all elements in $\boldsymbol{\gamma} = (\gamma_2, \ldots, \gamma_n)$ are allowed to be distinct. Instead, Minin et al. (2008) place a potentially strong smoothing prior on $\boldsymbol{\gamma}$:

$$P(\boldsymbol{\gamma} \mid \tau) \propto \tau^{(n-2)/2} \exp\left[-\frac{\tau}{2} \sum_{k=2}^{n} \frac{(\gamma_k - \gamma_{k-1})^2}{\delta_k}\right], \tag{2.13}$$

where τ is a precision parameter that determines how much differences between adjacent γs are penalized and δs are chosen either to be all one or are set to midpoint distances between inter-coalescent intervals. The precision hyperparameter τ receives a Gamma prior

distribution. The prior specified by (2.13) can be thought as a first-order random walk with normal increments and with initial distribution unspecified. Alternatively, we can view $\boldsymbol{\gamma}$ *a priori* as a discretized GP, discretely observed Brownian motion to be specific, on an irregular grid. In light of our discussion of point processes, it is not surprising that a very similar random walk prior was used for Bayesian nonparametric estimation of Poisson process intensity (Arjas and Heikkinen, 1997). Minin et al. (2008) approximate the posterior distribution of $\boldsymbol{\gamma}$, τ, and other model parameters, including the genealogy, using MCMC, implemented in `BEAST` (Drummond et al., 2012).

To avoid artificial discretization of $N_e(t)$ and potential statistical problems associated with such discretization, Palacios and Minin (2013) propose a more flexible GP-based prior for $N_e(t)$. They model the effective population size trajectory $N_e(t)$ as a linear transformation of an exponential Brownian motion:

$$N_e(t) = \{1 + \exp[-\gamma(t)]\}/\lambda, \tag{2.14}$$

where $\gamma(t) \sim \mathcal{BM}(\tau)$ and $\mathcal{BM}(\tau)$ denotes a Brownian motion process with mean function $\mathbf{0}$ and precision parameter τ. Parameter λ is necessary for technical reasons we outline below.

Using a continuous stochastic process prior for $N_e(t)$ has its price: the densities (5.2) and (2.3) become intractable due to the stochastic integration involved in computing these densities. A similar problem occurs during estimation of the Poisson process intensity. Until recently, the only available solution involved discretization of the Poisson intensity (Møller et al., 1998), which is not too different from the *Bayesian skyride* solution above. For estimation of the Poisson process intensity, Adams et al. (2009) propose a data augmentation that bypasses stochastic integration and avoids discretization of the intensity. Palacios and Minin (2013) develop a similar solution for the effective population size estimation. Their data augmentation procedure requires that $N_e(t)$ has a lower bound denoted by λ^{-1} in Equation (2.14). Palacios and Minin (2013) place a Gamma prior distribution on the precision hyperparameter τ and a mixture of uniform and exponential distributions on λ as follows:

$$P(\lambda) = \epsilon\frac{1}{\hat{\lambda}}1_{\{\lambda<\hat{\lambda}\}} + (1-\epsilon)\frac{1}{\hat{\lambda}}e^{-\frac{1}{\hat{\lambda}}(\lambda-\hat{\lambda})}1_{\{\lambda\geq\hat{\lambda}\}}, \tag{2.15}$$

where $\epsilon > 0$ is a mixing proportion and $\hat{\lambda}$ is our best guess of the upper bound, possibly obtained from previous studies.

The discretization-free GP-based estimation of $N_e(t)$ is currently available only for a fixed genealogy. Palacios and Minin (2013) approximate the posterior distribution of $N_e(t)$, τ, λ, and latent variables, introduced by the data augmentation, by MCMC sampling. Of course, one cannot keep track of the infinite dimensional object $N_e(t)$ without discretization. During MCMC, $N_e(t)$ is sampled at a finite number of time points at each iteration. *After MCMC is finished*, a grid of points $\{s_1, ..., s_B\}$ is formed and for each $g = 1, \ldots, B$, the posterior distribution of $N_e(s_g)$ is obtained from the MCMC samples by drawing from the posterior predictive distribution of $N_e(s_g)$. We emphasize that the coarseness of the grid has no bearing on statistical inference, because the grid is not used during the MCMC sampling. The grid can be made as fine as necessary after the MCMC is finished.

2.3 Examples

2.3.1 Fixed Genealogy

Consider a population that experienced an expansion followed by a contraction under the following demographic scenario:

$$N_e(t) = \begin{cases} e^{4t} & t \in [0, 0.5], \\ e^{-2t+3} & t \in (0.5, \infty). \end{cases} \tag{2.16}$$

Starting with 100 contemporaneous samples from this population, we simulated coalescent time points according to the specified demographic scenario. These coalescent times are shown as crosses at the bottom of each plot in Figure 2.3. The first plot in Figure 2.3 shows the piece-wise constant nature of the reconstructed trajectory using the Bayesian skyride method for a fixed genealogy and the second plot shows a smoother reconstructed trajectory using the continuous-time GP-based method. In both plots, the truth (Equation 2.16) is represented by a dashed line, posterior medians by solid lines and 95% Bayesian credible intervals (BCIs) by shaded areas. Although both methods recover a trajectory of a population that experiences growth and then a decay, the methods have difficulty timing the

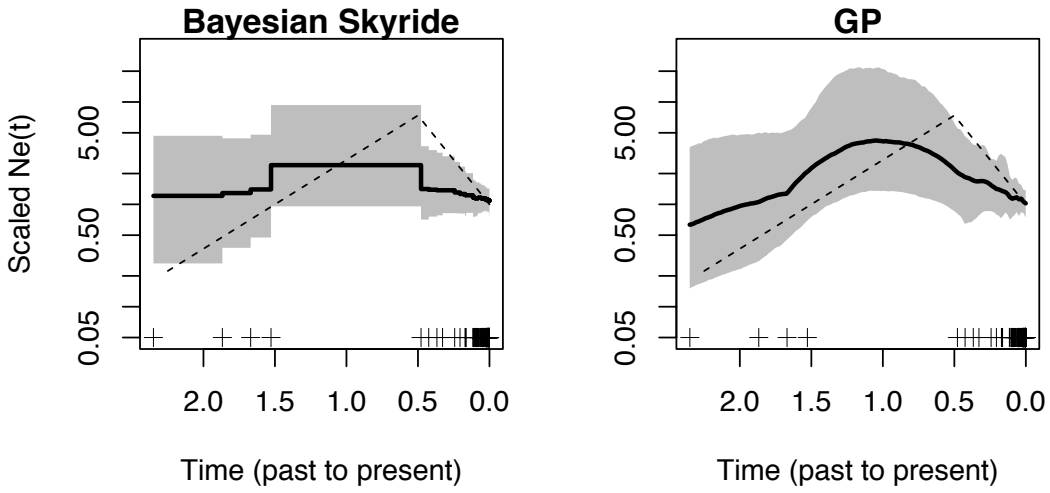

Figure 2.3: Example of a population that experienced expansion followed by a crash in population size. The true population size trajectory is depicted as dashed lines in the two plots. The simulated coalescent times are represented by the points at the bottom of each plot. We show the log of the effective population size trajectory estimated under the Bayesian Skyride (left plot) and continuous time GP inference of effective population size (right plot). Posterior medians are shown as solid black lines and 95% Bayesian credible intervals (BCIs) are represented by gray shaded areas.

peak of the population size. We will see further in the chapter how increasing the number of independent loci under analysis improves precision of the phylodynamic reconstruction.

We now consider real data examples. There are two major areas of application of phylodynamic methods. The first area corresponds to evolutionary studies of rapidly evolving infectious agents, such as RNA viruses. The second application area seeks to uncover past population size dynamics from ancient DNA. We showcase the usefulness of phylodynamic methods in these two areas by re-analyzing HCV in Egypt and bison across Beringia.

We first consider an estimated gene genealogy from 63 HCV E1 sequences sampled in 1993 in Egypt (Pybus et al., 2003). This genealogy is depicted by the 62 coalescent time points at the bottom of the plots in Figure 2.4. We apply the Bayesian skyride and GP-based phylodynamic reconstruction to this genealogy. Both methods recover a population size trajectory that increases exponentially after the 1920s and decreases after the 1970s. This reconstruction is consistent with a hypothesized role of parenteral antischistosomal therapy (PAT) in HCV spread in Egypt (Frank et al., 2000). The PAT campaign started

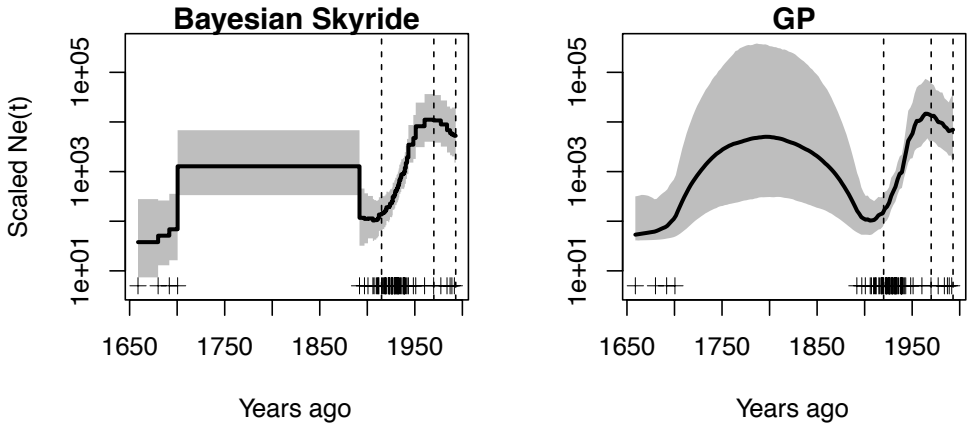

Figure 2.4: Egyptian HCV. The plots show the log of the scaled effective population size estimated using the Bayesian skyride (left) and the GP-based method (right) from the majority clade support genealogy with median node heights. Vertical dashed lines mark the years 1920, 1970, and 1993 from left to right. See the caption of Figure 2.3 for the rest of the legend.

in the 1920s, with the treatment administered intravenously, which together with a lack of sanitary practices is believed to have lead to a rapid increase of HCV infections in Egypt. The intravenous administration of the PAT was gradually replaced by oral administration in the 1970s, a change that is reflected in the decay in the effective number of HCV infections in our phylodynamic reconstructions.

To investigate the evolution and demographic history of Pleistocene bison, Shapiro et al. (2004) collected 152 mtDNA samples from bison fossils found in Alaska, Canada, Siberia, China, and the lower 48 United States with dates that spanned a period of more than 80,000 radiocarbon years before present. As with the HCV example, we used these DNA samples to reconstruct a genealogy of these samples under a molecular clock assumption. In Figure 2.5, we show bison population size trajectories reconstructed using the Bayesian skyride and the GP-based method from the estimated genealogy. Here, both methods agree in a recovered population size trajectory that reaches its maximum around 40 ka B.P., followed by a decay until it reaches a bottleneck around 10 ka B.P. This bottleneck occurs at the time of human settlement in Alaska, agreeing with previous analyses (Drummond et al., 2005).

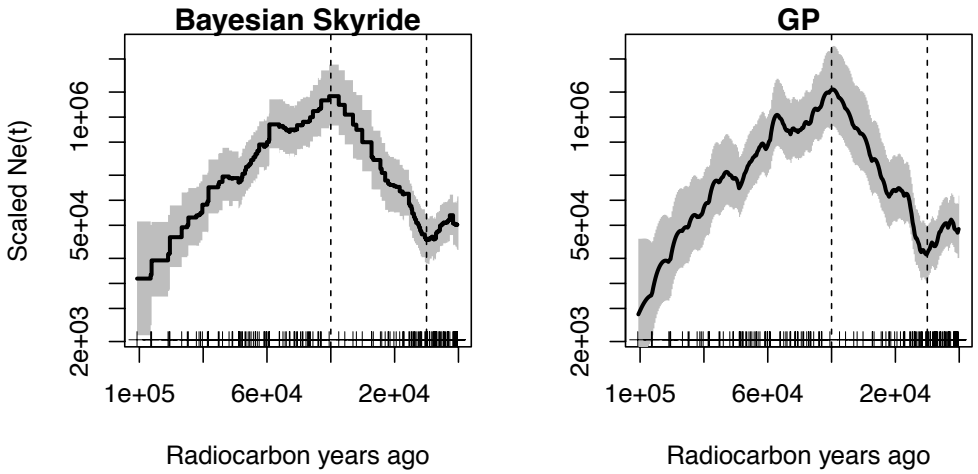

Figure 2.5: Berigian bison. The plots show the log of the scaled effective population size estimated using the Bayesian skyride (left) and the GP-based method (right) from the majority clade support genealogy with median node heights. Vertical dashed lines mark 40 ka B.P. and 10 ka B.P.. See the caption of Figure 2.3 for the rest of the legend.

2.3.2 Accounting for Genealogical Uncertainty

We now consider the same molecular data of HCV in Egypt and bison described earlier, but instead of conditioning on a reconstructed genealogy, proceed with the estimation of population size trajectories from the molecular data directly using the Bayesian skyride method. Figure 2.6 shows the recovered trajectories. In both cases, the key aspects of the population sizes are recovered from a fixed genealogy and from the sequence data directly, however, assessment of uncertainty and the degree of smoothness of the recovered trajectories differ substantially. These results reiterate that despite the attractive simplicity of the fixed genealogy methods, methods that properly account for genealogical uncertainty should be preferred to the fixed-genealogy approaches.

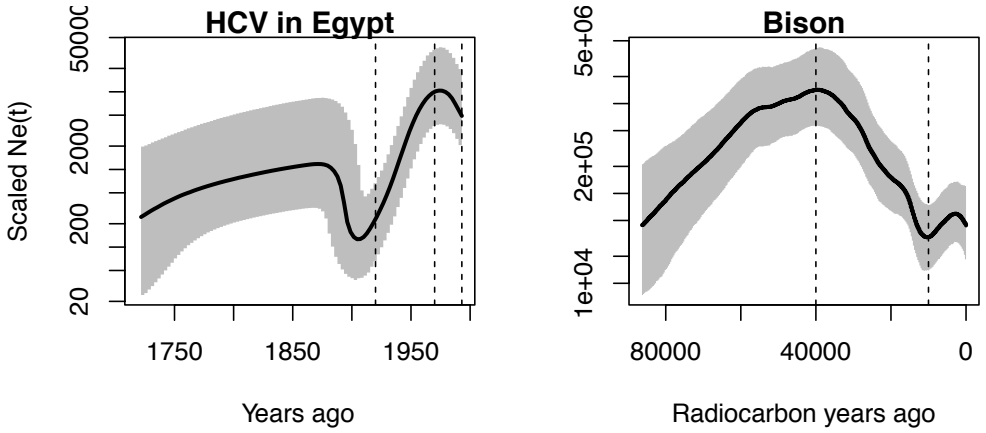

Figure 2.6: Analysis of molecular data from Egyptian HCV and bison examples. The left plot shows the log of the scaled effective population size of the Egyptian HCV and the right plot shows the log of the bison scaled effective population size, with both trajectories recovered by the Bayesian skyride method from the molecular data directly. See the captions of Figures 2.3, 2.4 2.5 for the rest of the legend.

Chapter 3

INFERENCE OF EFFECTIVE POPULATION SIZES FROM GENEALOGIES WITH GAUSSIAN PROCESSES

3.1 Introduction

Statistical inference in population genetics increasingly relies on the coalescent (Kingman, 1982), the probability model that describes the relationship between a gene genealogy of a random sample of molecular sequences and effective population size. This model provides a good approximation to the distribution of ancestral histories that arise from classical population genetics models (Rosenberg and Nordborg, 2002). More importantly, coalescent-based inference methods allow us to estimate population genetic parameters, including population size trajectories, directly from genomic sequences (Griffiths and Tavaré, 1994). Recent examples of coalescent-based population dynamics estimation include reconstructing demographic histories of musk ox (Campos et al., 2010b) from fossil DNA samples and elucidating patterns of genetic diversity of the dengue virus (Bennett et al., 2010a).

Here, we are interested in estimating effective population size trajectories from gene genealogies. The *effective population size* is an abstract parameter that for a real biological population is proportional to the rate at which genetic diversity is lost or gained. In the absence of natural selection, the effective population size can be used to approximate census population size by knowing the generation time in calendar units (e.g. years) and the population variability in number of offspring (Wakeley and Sargsyan, 2009). The latter quantity might be difficult to know; however, sometimes it suffices to analyze an arbitrarily rescaled population size trajectory, assuming the variability in number of offspring remains constant. The effective population size is equal to the census population size in an idealized Wright-Fisher model. The Wright-Fisher model is a simple and established model of neutral reproduction in population genetics that assumes random mating and non-overlapping generations. For some RNA viruses, for example human influenza A virus, the effective

population size rescaled by generation time (3 to 4 days) cannot be interpreted directly as the effective number of infections because of the presence of strong natural selection. However, one can always adopt a more cautious interpretation of the effective population size as a measure of relative genetic diversity (Rambaut et al., 2008; Frost and Volz, 2010).

Coalescent-based methods for estimation of population size dynamics have evolved from stringent parametric assumptions, such as constant population size or exponential growth (Griffiths and Tavaré, 1994; Kuhner et al., 1998; Drummond et al., 2002), to more flexible nonparametric approaches that assume piecewise linear population trajectories (Strimmer and Pybus, 2001; Opgen-Rhein et al., 2005; Drummond et al., 2005; Heled and Drummond, 2008; Minin et al., 2008). The latter class of methods is more appropriate in the absence of prior knowledge about the underlying demographic dynamics, allowing researchers to infer shapes of population size trajectories rather than to impose parametric constraints on these shapes. These nonparametric methods, however, model population dynamics by imposing *a priori* piecewise continuous functions which require regularization either by smoothing or by controlling the number of change points, also *a priori*. The former regularization – which works better in practice (Minin et al., 2008) – is inherently difficult because these piecewise continuous functions are defined on intervals of varying size. The piecewise nature of these methods creates further modeling problems if one wishes to incorporate covariates into the model or impose constraints on population size dynamics (Minin et al., 2008). In this chapter, we propose to solve these problems by bringing the coalescent-based estimation of population dynamics up to speed with modern Bayesian nonparametric methods. Making this leap in statistical methodology will allow us to avoid artificial discretization of population trajectories, to perform regularization without making arbitrary scale choices, and, in the future, to extend our method into a multivariate setting.

Our key insight stems from the fact that the coalescent with variable population size is an inhomogeneous continuous-time Markov chain (Tavaré, 2004) and, therefore, can be viewed as an inhomogeneous point process (Andersen et al., 1995). In fact, all current Bayesian nonparametric methods of estimation of population size dynamics resemble early Bayesian approaches to nonparametric estimation of the Poisson intensity function via piecewise continuous functions (Arjas and Heikkinen, 1997). Estimation of the intensity function of

an inhomogeneous Poisson process is a mature field that evolved from maximum likelihood estimation under parametric assumptions (Brillinger, 1979) to frequentist (Diggle, 1985) and, more recently, Bayesian nonparametric methods (Arjas and Heikkinen, 1997; Møller et al., 1998; Kottas and Sansó, 2007; Adams et al., 2009).

Following Adams et al. (2009), we *a priori* assume that population trajectories follow a transformed Gaussian process (GP), allowing us to model the population trajectory as a continuous function. This is a convenient way to specify prior beliefs without a particular functional form on the population trajectory. The drawback of such a flexible prior is that the likelihood function involves integration over an infinite-dimensional random object and, as a result, likelihood evaluation becomes intractable. Fortunately, we are able to avoid this intractability and perform inference exactly by adopting recent algorithmic developments proposed by Adams et al. (2009). We achieve tractability by a novel data augmentation for the coalescent process that relies on thinning algorithms for simulating the coalescent.

Thinning is an accept/reject algorithm that was first proposed by Lewis and Shedler (1979) for the simulation of inhomogeneous Poisson processes and was later extended to a more general class of point processes by Ogata (1981). In the spirit of Ogata (1981), we develop novel thinning algorithms for the simulation of the coalescent. These algorithms, interesting in their own right, open the door for latent variable representation of the coalescent. This representation leads to a new data augmentation that is computationally tractable and amenable to standard Markov chain Monte Carlo (MCMC) sampling from the posterior distribution of model parameters and latent variables.

We test our method on simulated data and compare its performance with a representative piecewise linear approach, a Gaussian Markov random field (GMRF) based method (Minin et al., 2008). We demonstrate that our method is more accurate and more precise than the GMRF method in all simulation scenarios. We also apply our method to two real data sets that have been previously analyzed in the literature: a hepatitis C virus (HCV) genealogy estimated from sequences sampled in 1993 in Egypt and a genealogy of the H3N2 human influenza A virus estimated from sequences sampled in New York state between 2002 and 2005. In the HCV analysis, we successfully recover all believed key aspects of the population size trajectory. Compared to the GMRF method, our GP method better reflects

the uncertainty inherent in the HCV data. In our second real data example, our GP method successfully reconstructs a population trajectory of the human influenza A virus with an expected seasonal series of peaks followed by population bottlenecks, while the GMRF method's reconstructed trajectory fails to recover some of the peaks and bottlenecks.

3.2 Methods

3.2.1 Coalescent Background

The coalescent model allows us to trace the ancestry of a random sample of n genomic sequences. These ancestral relationships are represented by a genealogy or tree; the times at which two sequences or lineages merge into a common ancestor are called coalescent times. The coalescent with variable population size can be viewed as a non-homogeneous Markov death process that starts with n lineages at present time $t_n = 0$ and decreases by one, with time running backwards, until reaching one lineage at t_1, at which point the samples have been traced to their most recent common ancestor (Griffiths and Tavaré, 1994). Here, we assume that a genealogy with time measured in units of generations is observed. The shape of the genealogy depends on the effective population size trajectory, $N_e(t)$, and the number of samples accumulated through time: the larger the effective population size, the longer two lineages need to wait to meet a common ancestor and the larger the sample size, the faster two lineages coalesce. Formally, let $t_n = 0$ denote the present time when all the n available sequences are sampled (*isochronous coalescent*) and let $t_n = 0 < t_{n-1} < \cdots < t_1$ denote the coalescent times of lineages in the genealogy with time going backwards. Then, the conditional density of the coalescent time t_{k-1} takes the following form:

$$P[t_{k-1}|t_k, N_e(t)] = \frac{C_k}{N_e(t_{k-1})} \exp\left\{-\int_{t_k}^{t_{k-1}} \frac{C_k}{N_e(t)} dt\right\}, \tag{3.1}$$

where $C_k = \binom{k}{2}$ is the coalescent factor that depends on the number of lineages $k = 2, \ldots, n$.

The *heterochronous coalescent* arises when samples of sequences are collected at different times (Figure 3.1). Such serially sampled data are common in studies of rapidly evolving viruses and analyses of ancient DNA (Campos et al., 2010b). Let $t_n = 0 < t_{n-1} < \cdots < t_1$ denote the coalescent times as before, but now let $s_m = 0 < s_{m-1} < \cdots < s_1 < s_0 =$ denote

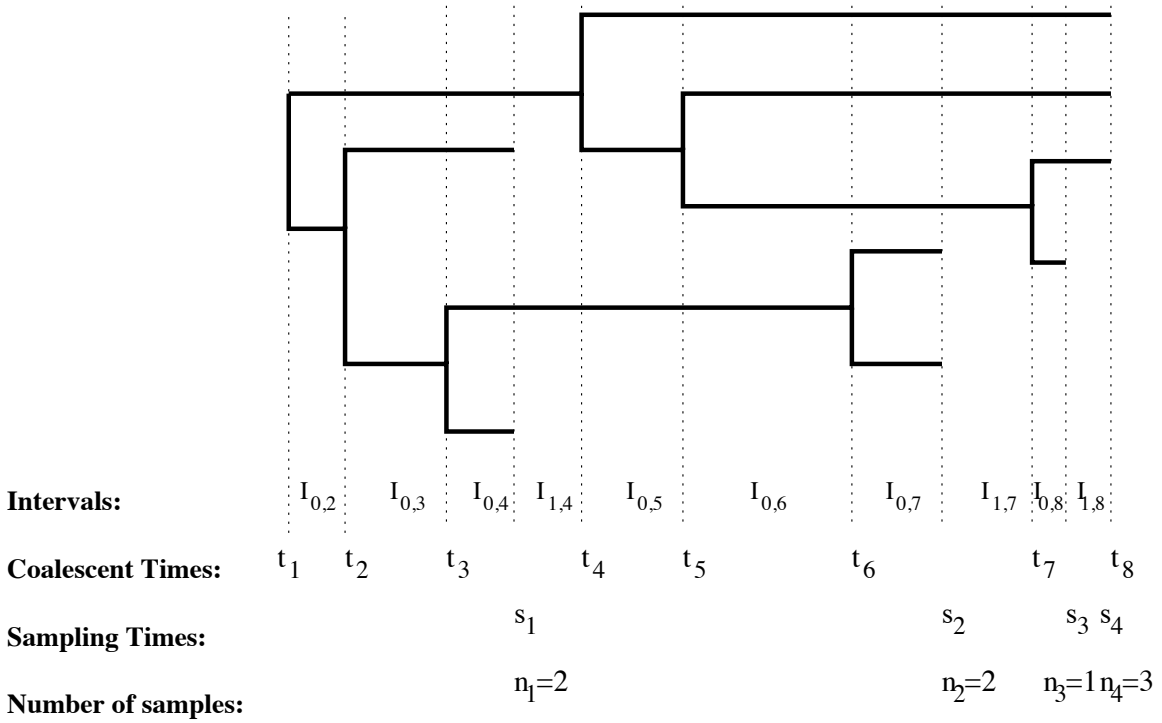

Figure 3.1: Example of a genealogy relating serially sampled sequences (heterochronous sampling). The number of lineages changes every time we move between intervals ($I_{i,k}$). Each endpoint of an interval is a coalescent time ($\{t_k\}_{k=1}^{n}$) or a sampling time ($\{s_j\}_{j=1}^{m}$). The number of sequences sampled at time s_j is denoted by n_j.

sampling times of $n_m, \ldots, n_1$ sequences respectively, $\sum_{j=1}^{m} n_j = n$. Further, let $\mathbf{s}$ and $\mathbf{n}$ denote the vectors of sampling times (time measured backwards) and numbers of sequences sampled at these times, respectively (Figure 3.1). Now, the coalescent factor changes not only at the coalescent events but also at the sampling times. Let

$$I_{0,k} = (max\{t_k, s_j\}, t_{k-1}], \text{ for } s_j < t_{k-1} \text{ and } k = 2, \ldots, n, \tag{3.2}$$

be the intervals that end with a coalescent event and

$$I_{i,k} = (max\{t_k, s_{j+i}\}, s_{j+i-1}], \text{ for } s_{j+i-1} > t_k \text{ and } s_j < t_{k-1}, k = 2, \ldots, n, \tag{3.3}$$

be the intervals that end with a sampling event. We denote the number of lineages in $I_{i,k}$ with $n_{i,k}$. Then, for $k = 2, ..., n$,

$$P[t_{k-1}|t_k, \mathbf{s}, \mathbf{n}, N_e(t)] = \frac{C_{0,k}}{N_e(t_{k-1})} \exp - \left\{ \int_{I_{0,k}} \frac{C_{0,k}}{N_e(t)} dt + \sum_{i=1}^{m} \int_{I_{i,k}} \frac{C_{i,k}}{N_e(t)} dt \right\}, \tag{3.4}$$

where $C_{i,k} = \binom{n_{i,k}}{2}$. That is, the density for the next coalescent time t_{k-1} is the product of the density of the coalescent time $t_{k-1} \in I_{0,k}$ and the probability of not having a coalescent event during the period of time spanned by intervals $I_{1,k}, \ldots, I_{m,k}$ (Felsenstein and Rodrigo, 1999).

3.2.2 Gaussian Process Prior for Population Size Trajectories

For both isochronous or heterochronous data, we place the same prior on $N_e(t)$:

$$N_e(t) = \left[\frac{\lambda}{1 + \exp\{-f(t)\}} \right]^{-1}, \tag{3.5}$$

where

$$f(t) \sim \mathcal{GP}(\mathbf{0}, \mathbf{C}(\boldsymbol{\theta})) \tag{3.6}$$

and $\mathcal{GP}(\mathbf{0}, \mathbf{C}(\boldsymbol{\theta}))$ denotes a Gaussian process with mean function $\mathbf{0}$ and covariance function $\mathbf{C}(\boldsymbol{\theta})$ with hyperparameters $\boldsymbol{\theta}$. *A priori*, $1/N_e(t)$ is a Sigmoidal Gaussian Process, a scaled logistic function of a Gaussian process whose range is restricted to lie in $[0, \lambda]$; λ is a positive constant hyperparameter, the inverse of which serves as a lower bound of $N_e(t)$ (Adams et al., 2009).

A *Gaussian process* is a stochastic process such that any finite sample from the process has a joint Gaussian distribution. The process is completely specified by its mean and covariance functions (Rasmussen and Williams, 2006). For computational convenience we use Brownian motion as our Gaussian process prior. Generating a finite sample from a Gaussian processes requires $\mathcal{O}(n^3)$ computations due to the inversion of the covariance matrix. However, when the precision matrix, the inverse of the covariance, is sparse, such simulations can be accomplished much faster (Rue and Held, 2005). For example, when we choose to work with a Brownian motion with covariance matrix elements $C(t,t^{'}) = \frac{1}{\theta}(min(t,t^{'}))$ and precision parameter θ, then the inverse of this matrix is tri-diagonal, which reduces the computational complexity of simulations from $\mathcal{O}(n^3)$ to $\mathcal{O}(n)$. In our MCMC algorithm, we need to generate realizations from the Gaussian processes at thousands of points, so the speed-up afforded by the Brownian motion becomes almost a necessity, prompting us to use this process as our prior in all our examples.

3.2.3 Priors for Hyperparameters

The precision parameter θ controls the degree of autocorrelation of our Brownian motion prior and influences the "smoothness" of the reconstructed population size trajectories. We place on θ a Gamma prior distribution with parameters α and β. The other hyperparameter in our model is the upper bound of $1/N_e(t)$, λ. When this upper bound λ is unknown, the model is unidentifiable (see equation (3.5)). However, in many circumstances it is possible to obtain an upper bound λ (or equivalently, a lower bound on $N(t)$) from previous studies and use this value to define the prior distribution of λ. We use the following strategy to construct an informative prior for λ. Let $\hat{\lambda}$ denote our best guess of the upper bound, possibly obtained from previous studies. Then, the prior on λ is a mixture of a uniform distribution for values to the left of $\hat{\lambda}$ and an exponential distribution to the right:

$$P(\lambda) = \epsilon\frac{1}{\hat{\lambda}}I_{\{\lambda<\hat{\lambda}\}} + (1-\epsilon)\frac{1}{\hat{\lambda}}e^{-\frac{1}{\hat{\lambda}}(\lambda-\hat{\lambda})}I_{\{\lambda\geq\hat{\lambda}\}}, \tag{3.7}$$

where $\epsilon > 0$ is a mixing proportion. When $\hat{\lambda}$ is considerably smaller than the unknown λ, the recovered curve will be visibly truncated around $\hat{\lambda}$, indicating that one needs to try higher values of $\hat{\lambda}$.

3.2.4 Doubly Intractable Posterior

Coalescent times $\mathcal{T} = \{t_n, t_{n-1}, \ldots, t_1\}$ of a given genealogy contain information needed to estimate $N_e(t)$ (see equations (3.1) and (3.4)). Given that $N_e(t)$ is a one-to-one function of $f(t)$ (equation (3.5)), we will focus the discussion on the inference of $f(t)$. The posterior distribution of $f(t)$ and hyperparameters θ and λ becomes

$$P(f(t), \theta, \lambda|\mathcal{T}) \propto P(\mathcal{T}|\lambda, f(t))P(f(t)|\theta)P(\theta)P(\lambda), \tag{3.8}$$

where $P(f(t)|\theta)$ is a Gaussian process prior with hyperparameter θ and

$$P(\mathcal{T}|\lambda, f(t)) = \prod_{k=2}^{n} \frac{C_k \lambda}{1 + \exp\{-f(t_{k-1})\}} \exp\left[-C_k \int_{t_k}^{t_{k-1}} \frac{\lambda}{1 + \exp\{-f(t)\}} dt\right] \tag{3.9}$$

is the likelihood function for the isochronous data (heterochronous data likelihood has a similar form). The integral in the exponent of equation (3.9) and the normalizing constant of equation (3.8) are computationally intractable, making the posterior doubly intractable (Murray et al., 2006).

Adams et al. (2009) faced a similar doubly intractable posterior distribution in the context of nonparametric estimation of intensity of the inhomogeneous Poisson process. These authors propose an introduction of latent variables so that the augmented data likelihood becomes tractable. This tractability makes the posterior distribution of latent variables and model parameters amenable to standard MCMC algorithms. Since Adams et al. (2009) based their data augmentation on the thinning algorithm for simulating inhomogeneous Poisson processes, we would like to devise a similar data augmentation based on a thinning algorithm for simulation of the coalescent with variable population size. In this simulation, we envision generating coalescent times assuming a constant population size and then thinning these times so that the distribution of the remaining (non-rejected) coalescent times follows the coalescent with variable population size. Since no thinning algorithm for simulating the coalescent process exists, we develop a series of such algorithms. In developing these algorithms, we find it useful to view the coalescent as a point process, a representation that we discuss below.

3.2.5 The Coalescent as a Point Process

The joint density of coalescent times is obtained by multiplying the conditional densities defined in equations (3.1) or (3.4). This density can be expressed as

$$P(t_1, \ldots, t_{n-1}|N_e(t)) = \prod_{k=2}^{n} \lambda^*(t_{k-1}|t_k) \exp\left\{ - \int_{t_k}^{t_{k-1}} \lambda^*(t|t_k) dt \right\}, \tag{3.10}$$

where $\lambda^*(t|t_k)$ denotes the conditional intensity function of a point process on the real line (Daley and Vere-Jones, 2002). For isochronous coalescent, the conditional intensity is defined by the step function:

$$\lambda^*(t|t_k) = \binom{k}{2} N_e(t)^{-1} 1_{\{t \in (t_k, t_{k-1}]\}}, \text{ for } k = 2, \ldots, n, \tag{3.11}$$

and the conditional intensity of the heterochronous coalescent point process is:

$$\lambda^*(t|\mathbf{n}, \mathbf{s}, t_k) = \sum_{i=1}^{m} \binom{n_{i,k}}{2} N_e(t)^{-1} 1_{\{t \in I_{i,k}\}}, \text{ for } k = 2, \ldots, n. \tag{3.12}$$

This novel representation allows us to reduce the task of estimating $N_e(t)$ to the estimation of the inhomogeneous intensity of the coalescent point process and to develop simulation algorithms based on thinning.

3.2.6 Coalescent Simulation via Thinning

To the best of our knowledge, the only method available for simulating the coalescent under the deterministic variable population size model is a time transformation method (Slatkin and Hudson, 1991b; Hein et al., 2005). This method is based on the random time-change theorem due to Papangelou (1972). Under the time transformation method, to simulate coalescent times, we proceed sequentially starting with $k = n$ and $t_n = 0$, generating t from an exponential distribution with unit mean, solving

$$t = \int_{t_k}^{t_{k-1}} \lambda^*(u|t_k) du \tag{3.13}$$

for t_{k-1} analytically or numerically and repeating the procedure until $k = 2$. For isochronous coalescent, $\lambda^*(u|t_k)$ is defined in equation (3.11) and for the heterochronous coalescent, $\lambda^*(u|t_k) = \lambda^*(u|\mathbf{n}, \mathbf{s}, t_k)$ is the piecewise function defined in equation (3.12). When $N_e(t)$ is

stochastic, the integral in equation (3.13) becomes intractable and the time transformation method is no longer practical. Instead, we propose to use *thinning*, a rejection-based method that does not require calculation of the integral in equation (3.13).

Lewis and Shedler (1979) proposed thinning a homogeneous Poisson process for the simulation of an inhomogeneous Poisson process with intensity $\lambda(t)$. The idea is to start with a realization of points from a homogeneous Poisson process with intensity λ and accept/reject each point with acceptance probability $\lambda(t)/\lambda$, where $\lambda(t) \leq \lambda$. The collection of accepted points forms a realization of the inhomogeneous Poisson process with conditional intensity $\lambda(t)$. Ogata (1981) extended Lewis and Shedler's thinning for the simulation of any point process that is absolutely continuous with respect to the standard Poisson process. We develop a series of thinning algorithms for the coalescent process that are similar to Ogata's algorithms, but not identical to them. Algorithm 1 outlines the simulation of n coalescent times under the isochronous sampling. Given t_k, we start generating and accumulating exponential random numbers E_i with rate $C_k\lambda$, until $t_{k-1} = t_k + E_1 + E_2 + \ldots$ is accepted with probability $1/N_e(t_{k-1})\lambda$. In order to ensure convergence of the algorithm, we require $\int_0^\infty \frac{du}{N_e(u)} = \infty$ a.s., which is equivalent to requiring that all sampled lineages can be traced back to their single common ancestor with probability 1. Notice that $N_e(t)$ can be either deterministic or stochastic. The latter case is considered in Algorithm 3, where we work with $f(t)$ instead of $N_e(t)$ for notational convenience. Algorithm 2 and 4 describe the equivalent algorithms for heterochronous data.

If $N_e(t)$ is deterministic and equation (3.13) can be solved analytically, the time transformation method is likely to be more efficient that thinning since the thinning algorithm is an accept/reject algorithm with the acceptance probability highly dependent on the definition of λ. However, efficiency of the thinning algorithm can be improved by replacing the constant upper bound λ on $1/N_e(t)$, by a piece-wise constant or a piece-wise linear function of local upper bounds in order to achieve higher acceptance probabilities, similarly to the adaptive rejection sampling of Gilks and Wild (1992).

To prove that Algorithm 1 generates coalescent times, we have the following proposition:

Algorithm 1 Simulation of isochronous coalescent times by thinning - $N_e(t)$ is a deterministic function

Input: $k = n$, $t_n = 0$, $t = 0$, $1/N_e(t) \leq \lambda$, $N_e(t)$
Output: $\mathcal{T} = \{t_k\}_{k=1}^{n}$

while $k > 1$ **do**
  Sample $E \sim Exponential(C_k\lambda)$ and $U \sim U(0,1)$
  t=t+E
  if $U \leq \frac{1}{N_e(t)\lambda}$ **then**
    $k \leftarrow k-1$, $t_k \leftarrow t$
  end if
end while

Algorithm 2 Simulation of heterochronous coalescent by thinning - $N_e(t)$ is a deterministic function

Input: $n_1, n_2, \ldots, n_m$, $s_1, \ldots, s_m$, $1/N_e(t) \leq \lambda$, $N_e(t)$, m
Output: for $n = \sum_{j=1}^{m} n_j$, $\mathcal{T} = \{t_k\}_{k=1}^{n}$

$i = 1$, $j = n-1$, $n = n_1$, $t = t_n = s_1$
while $i < m+1$ **do**
  Sample $E \sim Exp(\binom{n}{2}\lambda)$ and $U \sim U(0,1)$
  if $U \leq \frac{1}{N_e(t+E)\lambda}$ **then**
    if $t + E < s_{i+1}$ **then**
      $t_j \leftarrow t \leftarrow t + E$
      $j \leftarrow j-1$, $n \leftarrow n-1$
      if $n > 1$ **then**
        go to 2
      else
        go to 14
      end if
    else
      $i \leftarrow i+1$, $t \leftarrow s_i$, $n \leftarrow n + n_i$
    end if
  else
    $t \leftarrow t + E$
  end if
end while

Algorithm 2 Simulation of isochronous coalescent times by thinning with $f(t) \sim \mathcal{GP}(\mathbf{0}, \mathbf{C}(\theta))$

Input: $k = n$, $t_n = 0$, $t = 0$, $i_j = 0$, $m_j = 0$, $j = 2, \ldots, n$, λ
Output: $\mathcal{T} = \{t_k\}_{k=1}^{n}$, $\mathcal{N} = \{\{t_{k,i}\}_{i=1}^{m_k}\}_{k=2}^{n}$, $\mathbf{f}_{\mathcal{T},\mathcal{N}}$

while $k > 1$ **do**
  Sample $E \sim Exponential(C_k\lambda)$ and $U \sim U(0,1)$
  t=t+E
  Sample $f(t) \sim P(f(t)|\{f(t_l)\}_{l=k}^{n}, \{\{f(t_{l,i})\}_{i=1}^{m_l}\}_{l=k}^{n}; \theta)$
  if $U \leq \frac{1}{1+\exp(-f(t))}$ **then**
    $k \leftarrow k-1$, $t_k \leftarrow t$
  else
    $i_k \leftarrow i_k + 1$, $m_k \leftarrow m_k + 1$, $t_{k,i_k} \leftarrow t$
  end if
end while

Algorithm 3 Simulation of heterochronous coalescent by thinning with $f(t) \sim \mathcal{GP}(\mathbf{0}, \mathbf{C}(\theta))$

Input: $n_1, n_2, \ldots, n_m$, $s_1 = 0, \ldots, s_m$, $i_j = 0$, $m_j = 0$, $j = 2, \ldots, n$, λ, m
Output: for $n = \sum_{j=1}^{m} n_j$, $\mathcal{T} = \{t_k\}_{k=1}^{n}$, $\mathcal{N} = \{\{t_{k,i}\}_{i=1}^{m_k}\}_{k=2}^{n}$, $\mathbf{f}_{\mathcal{T},\mathcal{N}}$

$i = 1$, $j = n-1$, $n = n_1$, $t = t_n = s_1$
while $i < m+1$ **do**
  Sample $E \sim Exp(\binom{n}{2}\lambda)$ and $U \sim U(0,1)$
  Sample $f(t+E) \sim P(f(t+E)|\{f(t_l)\}_{l=k}^{n}, \{\{f(t_{l,i})\}_{i=1}^{m_l}\}_{l=k}^{n}; \theta)$
  if $U \leq \frac{1}{1+\exp(-f(t+E))}$ **then**
    if $t + E < s_{i+1}$ **then**
      $t_j \leftarrow t \leftarrow t + E$
      $j \leftarrow j-1$, $n \leftarrow n-1$
      if $n > 1$ **then**
        go to 2
      else
        go to 14
      end if
    else
      $i \leftarrow i+1$, $t \leftarrow s_i$, $n \leftarrow n + n_i$
    end if
  else
    if $t + E < s_{i+1}$ **then**
      $t_{j+1,i_{j+1}} \leftarrow t + E$, $i_{j+1} \leftarrow i_{j+1} + 1$
    end if
    $t \leftarrow t + E$
  end if
end while

Proposition 1. *Algorithm 1 generates $t_n < t_{n-1} < \cdots < t_1$, such that*

$$P(t_{k-1} > t | t_k) = \exp\left[-\int_{t_k}^{t} \frac{C_k dx}{N_e(x)}\right], \tag{3.14}$$

where $N_e(t)$ is known deterministically.

Proof. Let $T_i = t_k + E_1 + \ldots + E_i$, where $\{E_i\}_{i=1}^{\infty}$ are iid exponential $Exp(C_k\lambda)$ random numbers. Given t_k, Algorithm 1 generates and accumulates iid exponential random numbers until T_i is accepted with probability $1/\lambda N_e(T_i)$, in which case, T_i is labeled t_{k-1}. Let $N(t_k, t] = \#\{i \geq 1 : t_k < T_i \leq t\}$ denote the number of iid exponential random numbers generated in $(t_k, t]$. Then, $\{N(t_k, t], t > t_k\}$ constitutes a Poisson process with intensity $C_k\lambda$. Then, given $N(t_k, t] = 1$, the conditional density of a point x in $(t_k, t]$ is $1/(t - t_k)$ and the probability of accepting such a point as a coalescent time point with variable population size is $1/\lambda N_e(x)$. Hence

$$P(t_{k-1} \leq t | t_k, N(t_k, t] = 1) = \frac{1}{\lambda(t - t_k)} \int_{t_k}^{t} \frac{dx}{N_e(x)}, \tag{3.15}$$

and

$$P(t_{k-1} > t | t_k, N(t_k, t] = m) = \left(1 - \frac{1}{\lambda(t - t_k)} \int_{t_k}^{t} \frac{dx}{N_e(x)}\right)^m. \tag{3.16}$$

Then,

$$P(t_{k-1} > t | t_k) = \sum_{m=1}^{\infty} P(t_{k-1} > t | t_k, N(t_k, t] = m) P(N(t_k, t] = m)$$

$$= \sum_{m=1}^{\infty} \left(1 - \frac{1}{\lambda(t - t_k)} \int_{t_k}^{t} \frac{dx}{N_e(x)}\right)^m \frac{(C_k\lambda(t - t_k))^m \exp\left[-C_k\lambda(t - t_k)\right]}{m!}$$

$$= \exp\left[-C_k\lambda(t - t_k)\right] \sum_{m=1}^{\infty} \frac{\left(C_k\lambda(t - t_k) - C_k \int_{t_k}^{t} \frac{dt_{k-1}}{N_e(t_{k-1})}\right)^m}{m!}$$

$$= \exp\left[-\int_{t_k}^{t} \frac{C_k dx}{N_e(x)}\right].$$

□

3.2.7 Data Augmentation and Inference

As mentioned in the previous section, our thinning algorithm for the coalescent is motivated by our desire to construct a data augmentation scheme. We imagine that observed coalescent times $\mathcal{T}$ were generated by the thinning procedure described in Algorithm 1, so we augment $\mathcal{T}$ with rejected (thinned) points $\mathcal{N}$. If we keep track of the rejected points resulting from

Algorithm 1, then, given t_k, $f(t_k), \mathbf{f}_{\mathcal{N}_k} = \{f(t_{k,i})\}_{i=1}^{m_k}$ and λ, we start proposing new time points $\mathcal{N}_k = \{t_{k,1}, \ldots, t_{k,m_k}\}$ until t_{k-1} is accepted, so that

$$P(t_{k-1}, \mathcal{N}_k | t_k, f(t_{k-1}), \mathbf{f}_{\mathcal{N}_k}, \lambda) = (C_k\lambda)^{m_k+1} \exp\{-C_k\lambda(t_k - t_{k-1})\} \left[\frac{1}{1+\exp\{-f(t_{k-1})\}}\right] \times \prod_{i=1}^{m_k} \left[1 - \frac{1}{1+\exp\{-f(t_{k,i})\}}\right] \quad (3.17)$$

For the heterochronous coalescent, equation (3.17) is modified in the following way:

$$P(t_{k-1}, \mathcal{N}_k | t_k, f(t_{k-1}), \mathbf{f}_{\mathcal{N}_k}, \lambda, \mathbf{s}, \mathbf{n}) = (\lambda C_{0,k})^{1+m_{0,k}} \exp\{-\lambda C_{0,k} l(I_{0,k})\} \times \left[\left(\frac{1}{1+\exp\{-f(t_{k-1})\}}\right) \prod_{i=1}^{m_k} \frac{1}{1+\exp\{f(t_{k,i})\}}\right] \prod_{i=1}^{m} \left[(\lambda C_{i,k})^{m_{i,k}} \exp\{-\lambda C_{i,k} l(I_{i,k})\}\right], \quad (3.18)$$

where $l(I_{i,k})$ denotes the length of the interval $I_{i,k}$ and $m_{i,k} = \sum_{l=1}^{m_k} 1\{t_{k,l} \in I_{i,k}\}$ denotes the number of latent points in interval $I_{i,k}$. Let $\mathbf{f}_{\mathcal{T},\mathcal{N}} = \left\{\{f(t_k)\}_{k=1}^{n}, \left\{\{f(t_{k,i})\}_{i=1}^{m_k}\right\}_{k=2}^{n}\right\}$, then the augmented data likelihood of $\mathcal{T}$ and $\mathcal{N}$ becomes

$$P(\mathcal{T}, \mathcal{N} | \mathbf{f}_{\mathcal{T},\mathcal{N}}, \lambda) = \prod_{k=2}^{n} P(t_{k-1}, \mathcal{N}_k | t_k, f(t_{k-1}), \mathbf{f}_{\mathcal{N}_k}, \lambda). \quad (3.19)$$

Then, the posterior distribution of $f(t)$ and hyperparameters evaluated at the observed $\mathcal{T}$ and latent $\mathcal{N}$ time points is

$$P(\mathbf{f}_{\mathcal{T},\mathcal{N}}, \lambda, \theta | \mathcal{T}, \mathcal{N}) \propto P(\mathcal{T}, \mathcal{N} | \mathbf{f}_{\mathcal{T},\mathcal{N}}, \lambda) P(\mathbf{f}_{\mathcal{T},\mathcal{N}} | \theta) P(\lambda) P(\theta). \quad (3.20)$$

The augmented posterior can now be easily evaluated since it does not involve integration of infinite-dimensional random functions. We follow Adams et al. (2009) and develop a MCMC algorithm to sample from the posterior distribution (3.20). At each iteration of our MCMC, we update the following variables: (1) number of "rejected" points $\#\mathcal{N}$; (2) the locations of the rejected points $\mathcal{N}$; (3) the function values $\mathbf{f}_{\mathcal{T},\mathcal{N}}$ and (4) the hyperparameters θ and λ.

3.2.8 MCMC Sampling

Since the coalescent under isochronous sampling is a particular case of the coalescent model under heterochronous sampling, we employ here the notation of the heterochronous coalescent, understanding that $C_{0,k} = C_k$, $I_{0,k} = (t_k, t_{k-1}]$ and $i = 0$ for isochronous data.

Sampling the number of latent points. A reversible jump algorithm is constructed for the number of "rejected" points. We propose to add or remove points with equal probability in each intercoalescent interval. When adding a point in a particular interval, we propose a location uniformly from the interval and its predicted function value $f(t^*) \sim P(f(t^*)|\mathbf{f}_{\mathcal{T},\mathcal{N}}, \theta)$. When removing a point, we propose to remove a point selected uniformly from the pool of rejected points in that particular interval. We add points with proposal distributions $q_{up}^{i,k}$ and remove points with proposal distributions $q_{down}^{i,k}$. Then,

$$q_{up}^{k} = \frac{P(f(t^*)|\mathcal{T},\mathcal{N})}{2l(I_k)}, \tag{3.21}$$

$$q_{down}^{k} = \frac{1}{2\#(\mathcal{N}_k)}, \tag{3.22}$$

and the acceptance probabilities are:

$$a_{up}^{k} = \frac{l(I_k)\lambda C_k}{(\#\mathcal{N}_k + 1)(1 + e^{f(t^*)})}, \tag{3.23}$$

$$a_{down}^{k} = \frac{(\#\mathcal{N}_k)(1 + e^{f(t^*)})}{l(I_k)\lambda C_k}, \tag{3.24}$$

Sampling locations of latent points. We use a Metropolis-Hastings algorithm to update the locations of latent points. We first choose an intercoalescent interval with latent points with probability proportional to its length and we then propose point locations uniformly at random in that interval together with their predictive function values $\mathbf{f}_{\mathbf{t}^*} \sim P(\mathbf{f}_{\mathbf{t}^*}|\mathbf{f}_{\mathcal{T},\mathcal{N}}, \theta)$.

$$a^{k} = \frac{C_{i,k}\left[1 + e^{f(t)}\right]}{C_{i^*,k}\left[1 + e^{f(t^*)}\right]}. \tag{3.25}$$

Sampling transformed effective population size values. We use an elliptical slice sampling proposal described in (Murray et al., 2010). In both cases, isochronous or heterochronous, the full conditional distribution of the function values $\mathbf{f}_{\mathcal{T},\mathcal{N}}$ is proportional to the product of a Gaussian density and the thinning acceptance and rejection probabilities:

$$P(\mathbf{f}_{\mathcal{T},\mathcal{N}}|\mathcal{T},\mathcal{N},\lambda,\theta) \propto P(\mathbf{f}_{\mathcal{T},\mathcal{N}}|\theta)L(\mathbf{f}_{\mathcal{T},\mathcal{N}}), \tag{3.26}$$

where

$$L(\mathbf{f}_{\mathcal{T},\mathcal{N}}) = \prod_{k=2}^{n}\left(\frac{1}{1+e^{-f(t_{k-1})}}\right)\prod_{i=1}^{m_k}\frac{1}{1+e^{f(t_{k,i})}}. \tag{3.27}$$

Sampling hyperparameters. The full conditional of the precision parameter θ is a Gamma distribution. Therefore, we update θ by drawing from its full conditional:

$$\theta|\mathbf{f}_{\mathcal{T},\mathcal{N}},\mathcal{T},\mathcal{N} \sim Gamma\left(\alpha^* = \alpha + \frac{\#\{\mathcal{N}\cup\mathcal{T}\}}{2}, \beta^* = \beta + \frac{\mathbf{f}^t_{\mathcal{T},\mathcal{N}}Q\mathbf{f}_{\mathcal{T},\mathcal{N}}}{2}\right), \tag{3.28}$$

where $Q = \frac{1}{\theta}C^{-1}$.

For the upper bound λ on $N_e(t)^{-1}$, we use the Metropolis-Hastings update by proposing new values using a uniform proposal reflected at 0. That is, we propose λ^* from $U(\lambda-a, \lambda+a)$. If the proposed value λ^* is negative, we flip its sign. Since the proposal distribution is symmetric, the acceptance probability is:

$$a = \frac{P(\lambda^*)}{P(\lambda)}\left(\frac{\lambda^*}{\lambda}\right)^{\#\{\mathcal{N}\cup\mathcal{T}\}}\exp\left[-(\lambda^*-\lambda)\sum_{k=2}^{n}\sum_{i=1}^{m_k}C_{i,k}l(I_{i,k})\right], \tag{3.29}$$

where $P(\lambda)$ is defined in equation (3.7)

We summarize the posterior distribution of $N_e(t)$ by its empirical median and 95% Bayesian credible intervals (BCIs) evaluated at a grid of points. This grid can be made as fine as necessary after the MCMC is finished. That is, given the function values $\mathbf{f}_{\mathcal{T},\mathcal{N}}$ at coalescent and latent time points, and the value of the precision parameter θ at each iteration, we sample the function values at a grid of points $\mathbf{g} = \{g_1, ..., g_B\}$ from its predictive distribution $\mathbf{f}_\mathbf{g} \sim P(\mathbf{f}_\mathbf{g}|\mathbf{f}_{\mathcal{T},\mathcal{N}}, \theta)$, and evaluate $\{N_e(g_i)\}_{i=1}^B$.

3.3 Results

3.3.1 Simulated Data

We simulate three genealogies relating 100 individuals, sampled at the same time $t = 0$ (isochronous sampling) under the following demographic scenarios: 1) constant population size trajectory: $N_e(t) = 1$; 2) exponential growth: $N_e(t) = 25e^{-5t}$; and 3) population expansion followed by a crash: $N_e(t) = e^{4t}1_{\{t\in[0,0.5]\}} + e^{-2t+3}1_{\{t\in(0.5,\infty)\}}$. We compare the posterior median with the truth by the sum of relative errors (SRE):

$$SRE = \sum_{i=1}^{K}\frac{|\hat{N}_e(s_i) - N_e(s_i)|}{N_e(s_i)}, \tag{3.30}$$

where $\hat{N}_e(s_i)$ is the estimated trajectory at time s_i with $s_1 = t_1$, the time to the most recent common ancestor and $s_K = t_n = 0$ for any finite K. In our examples, we use 150 for the number of points K to calculate the SRE. Similarly, we compute the mean relative width (MRW) of the 95% BCIs defined in the following way:

$$MRW = \sum_{i=1}^{K} \frac{|\hat{N}_{97.5}(s_i) - \hat{N}_{2.5}(s_i)|}{KN_e(s_i)}. \tag{3.31}$$

We also compute the percentage of time, the 95% BCIs cover the truth (envelope) in the following way:

$$envelope = \frac{\sum_{i=1}^{K} I(\hat{N}_{2.5}(s_i) \leq N(s_i) \leq \hat{N}_{97.5}(s_i))}{K}. \tag{3.32}$$

As a measure of the frequentist coverage, we calculate the percentage of times the truth is completely covered by the 95% BCIs (envelope = 1), by simulating each demographic scenario and performing Bayesian estimation of each such simulation 100 times.

We compute the three statistics for the three simulation scenarios for $K = 150$ at equally spaced time points (Table 3.1). These statistics do not change significantly when we use higher values of K. Additionally, we compute the variation of $\hat{N}_e(t)$ over a regular grid of $K = 150$ points as follows:

$$variation = \sum_{i=1}^{K-1} |\hat{N}_e(s_{i+1}) - \hat{N}_e(s_i)|, \tag{3.33}$$

For all simulations, we set the mixing parameter ϵ of the prior density for λ (equation (3.7)) to $\epsilon = 0.01$. The parameters of the Gamma prior on the GP precision parameter θ were set to $\alpha = \beta = 0.001$. We summarize our posterior inference in Figure 3.2 and compare our GP method to the GMRF smoothing method (Minin et al., 2008). The effective population trajectory is log transformed and time is measured in units of generations.

For the constant population scenario (first row in Figure 3.2), the truth (dashed lines) is almost perfectly recovered by the GP method (solid black line) and the 95% BCIs shown as gray shaded areas are remarkably tight. For the exponential growth simulation (second row), the GMRF method recovers the truth better in the right tail, while our GP method recovers it much better in the left tail. The higher variation of the GP reconstruction in the right tail makes this measure higher than for the GMRF reconstruction. Overall, our

Table 3.1: Summary of Simulation Results Depicted in Figure 3.2. SRE is the sum of relative errors as defined in equation (3.30), MRW is the mean relative width of the 95% BCI as defined in equation (3.31), envelope is calculated as in equation (3.32) and variation is calculated as in equation (3.33).

Simulations	SRE		MRW		Envelope		Variation		
	GMRF	GP	GMRF	GP	GMRF	GP	GMRF	GP	TRUTH
Constant	50.41	**4.15**	4.21	**0.72**	**100.0%**	**100.0%**	2.27	**0.08**	0.00
Exp. growth	47.65	**33.60**	2.55	**2.35**	**100.0%**	**100.0%**	**30.19**	52.41	24.80
Expansion/crash	181.88	**140.88**	10.7	**7.26**	77.33%	**92.0%**	5.69	**6.94**	13.46

GP method better recovers the truth in the exponential growth scenario, as evidenced by SREs and MRWs in Table 3.1. The last row in Figure 3.2 shows the results for a population that experiences expansion followed by a crash in effective population size. In this case, 95% BCIs of the two methods do not completely cover the true trajectory. While an area near the bottleneck is particularly problematic, the GP method's envelope is much higher (92%) than the envelope produced by the GMRF method (77.3%), the variation recovered by the GP method is closer to the true variation in all simulation scenarios and in general, in terms of the four statistics employed here, the GP method shows better performance. Results for the GMRF method were obtained using the `BEAST` software (Drummond et al., 2012) with running times ranging from 25 to 40 minutes, while results for the GP method were obtained using `R` with running times ranging from 60 to 180 minutes. Although our GP implementation takes longer, we obtain better performance in a still reasonable amount of time. Figure 3.3 shows the trace plot of loglikehood and the effective sample size per grid point for the exponential simulation. In all cases, the effective sample size is larger than 120.

Next, we simulate each of the three scenarios 100 times and compute the four statistics described before for both methods. The distributions of these statistics are represented by the boxplots depicted in Figure 3.4. In general, our GP method has smaller SREs, except in the constant case, where the distributions look very similar; smaller MRWs, larger envelopes and variation closer to the truth. Additionally, we calculate the percentage of times, the envelope is 1 as a proxy for frequentist coverage of the 95% BCIs. Since the 95% BCIs are calculated pointwise at 150 equally spaced points, we do not necessarily expect frequentist

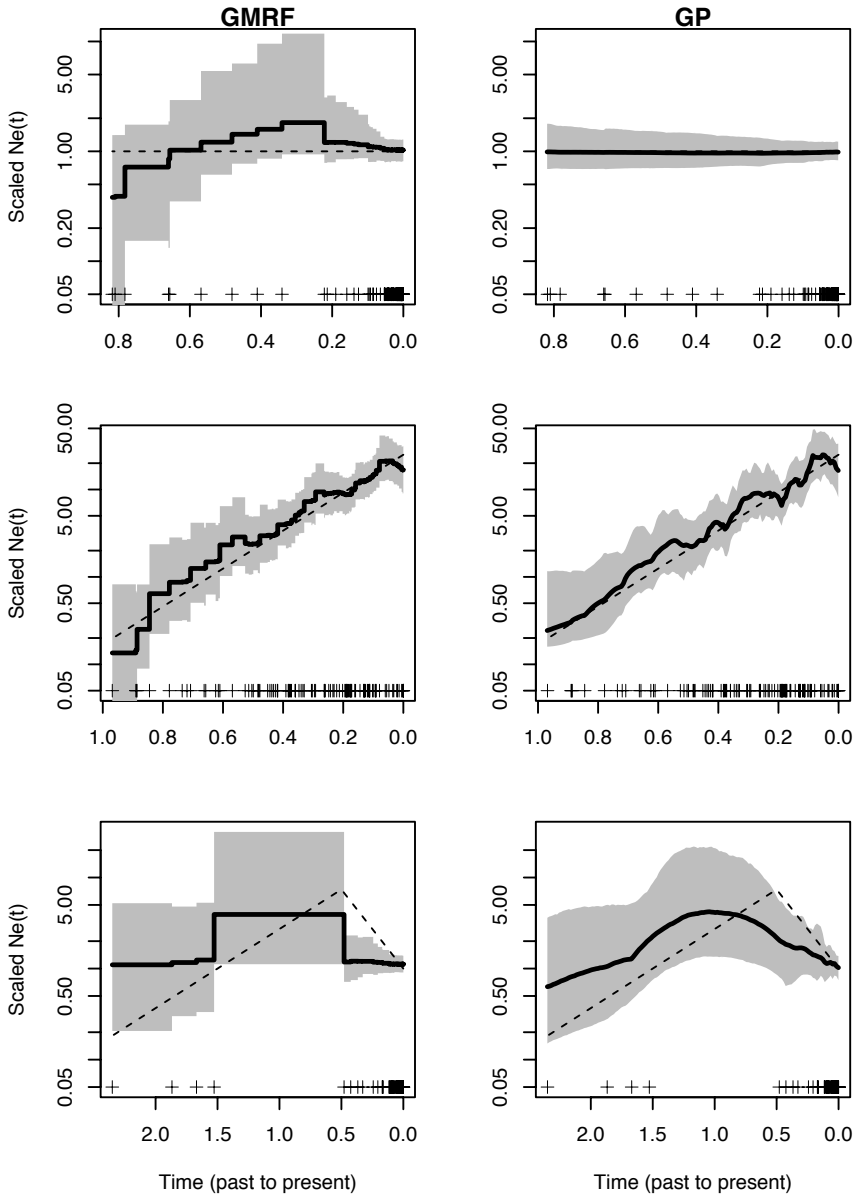

Figure 3.2: Simulated data under the constant population size (first row), exponential growth (second row) and expansion followed by a crash (third row). The simulated points are represented by the points at the bottom of each plot. We show the log of the effective population size trajectory estimated under the Gaussian Markov random field smoothing (GMRF) method and our method: Gaussian process-based nonparametric inference of effective population size (GP). We show the true trajectories as dashed lines, posterior medians as solid black lines and 95% BCIs by gray shaded areas.

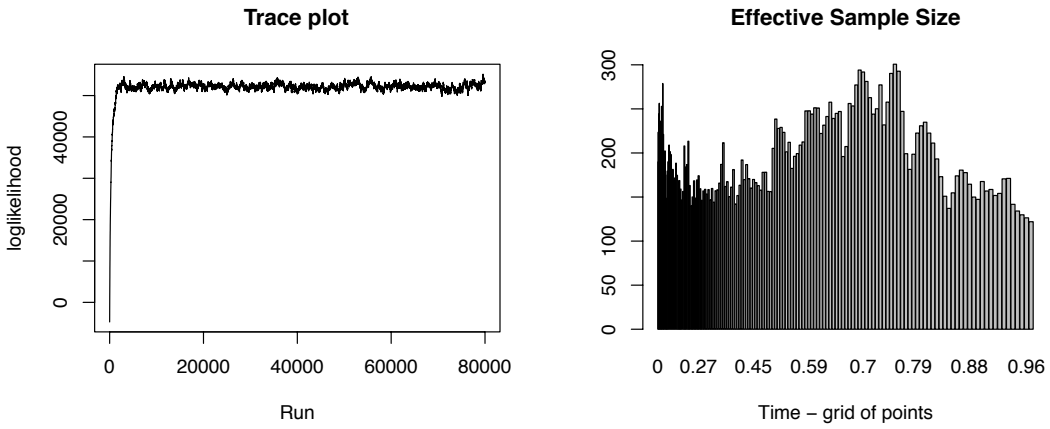

Figure 3.3: Trace plot of loglikehood (left plot) and effective sample sizes of $N_e(t)$ evaluated at a grid of points (right plot) for the recovered exponential growth trajectory using the GP method.

coverage to be close to 95%. The results are shown as the numbers at the top of the right plot in Figure 3.4. The coverage levels obtained using the GP method are larger than those obtained using the GMRF method.

3.3.2 Egyptian HCV

Hepatitis C virus was first identified in 1989. By 1992, when HCV antibody testing became widely available, the prevalence of HCV in Egypt was about 10.8%. Today, Egypt is the country with the highest HCV prevalence (Miller and Abu-Raddad, 2010). A widely held hypothesis that can explain the epidemic is the role of a parenteral antischistosomal therapy (PAT) campaign, that started in the 1920s, combined with lack of sanitary practices as the means for transmission. The campaign was discontinued in the 1970s when the intravenous treatment was gradually replaced by oral administration of the treatment (Ray et al., 2000). Coalescent demographic methods developed over the last 10 years demonstrated evidence in favor of this hypothesis (Pybus et al., 2003; Drummond et al., 2005; Minin et al., 2008). Therefore, this example is well suited for testing our method. We analyze the genealogy estimated by Minin et al. (2008), based on 63 HCV sequences sampled in Egypt in 1993, and compare our method to the GMRF smoothing method (Minin et al., 2008). The results are depicted in Figure 3.5, with time scaled in units of years. In line with previous

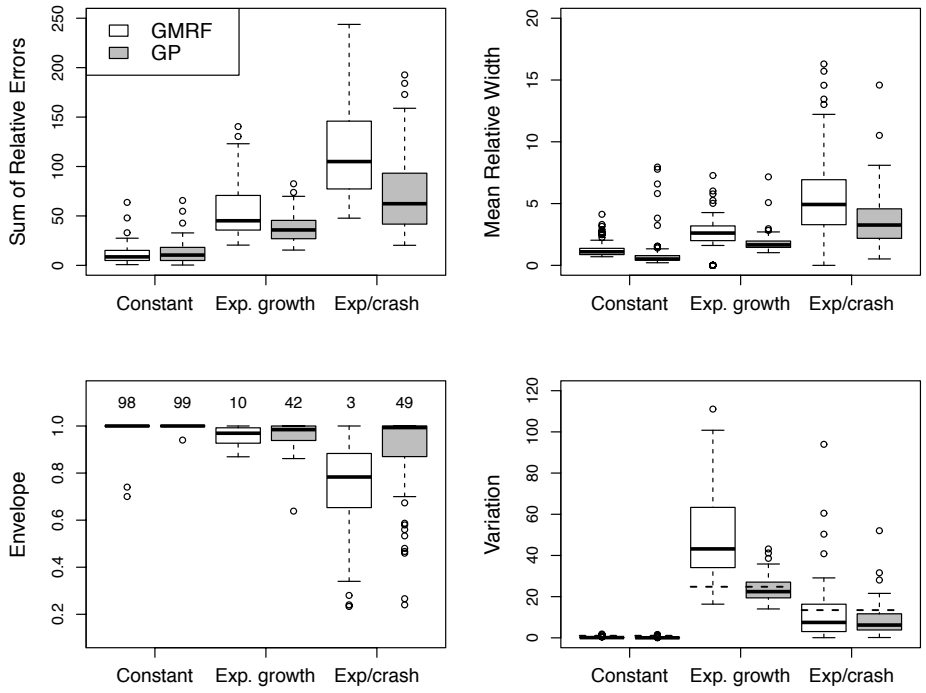

Figure 3.4: Boxplots of SRE (top left), MRW (top right), envelope (bottom left) and variation (bottom right) based on 100 simulations for a constant trajectory, exponential growth and expansion followed by crash. The numbers above the boxplots of the bottom left plot represent the estimated frequentist coverage of the 95% BCIs, and the dashed lines in the bottom right plot indicate variations of the true simulated trajectories.

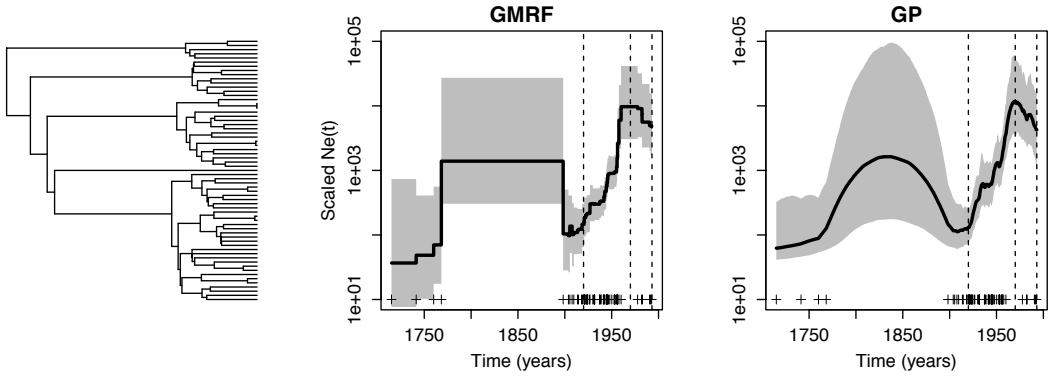

Figure 3.5: Egyptian HCV. The first plot (left to right) is one possible genealogy reconstructed by Minin et al. (2008). The next two plots represent the log of scaled effective population trajectory estimated using the GMRF smoothing method and our GP method. The posterior medians for the last two plots are represented by solid black lines and the 95% BCI's are represented by the gray shaded areas. The vertical dashed lines mark the years 1920 (the start of intravenous PAT) , 1970 (the end of intravenous PAT) and 1993 (sampling time of sequences).

results (Pybus et al., 2003; Drummond et al., 2005; Minin et al., 2008), our estimation recovers the exponential growth of the HCV population size starting from the 1920s when the intravenously administered PAT was introduced. Both Pybus et al. (2003) and Minin et al. (2008) hypothesize that the population trajectory remained constant before the start of the exponential growth. The GMRF and GP approaches disagree the most on the HCV population size reconstruction prior to 1920s. The GP method produces narrower BCIs near the root of the genealogy (1710-1770) than the GMRF approach. In contrast, GP BCIs are inflated in the time period from 1770 to 1900 in comparison to the GMRF results. We believe that the uncertainty estimates produced by the GP approach are more reasonable than the GMRF result, because there are multiple coalescent events during 1710- 1770, providing information about the population size, while the time interval 1770 - 1900 has no coalescent events, a data pattern that should result in substantial uncertainty about the HCV population size. Another notable difference between the GMRF and GP methods is in estimation of the HCV population trajectory after 1970. The GP approach suggests a sharper decline in population size during this time interval.

3.3.3 Seasonal Human Influenza

Here, we estimate population dynamics of human influenza A, based on 288 H3N2 sequences sampled in New York state from January, 2001 to March, 2005. Sequences from the coding region of the influenza hemagglutinin (HA) gene of H3N2 influenza A virus from New York state were collected from the NCBI Influenza Database (Influenza Genome Sequencing Project, 2011), incorporating the exact dates of viral sampling in weeks (heterochronous sampling) and aligned using the software package MUSCLE (Edgar, 2004). These sequences form a subset of sequences analyzed in (Rambaut et al., 2008). We carried out a phylogenetic analysis using the software package `BEAST` (Drummond et al., 2012) to generate a majority clade support genealogy with median node heights as our genealogical reconstruction. The reconstructed genealogy is depicted in the left plot of Figure 3.6. Demography of H3N2 influenza A virus in temperate regions, such as New York, is characterized by epidemic peaks during winters followed by strong bottlenecks at the end of epidemic seasons. As expected, our method recovers the peaks in the effective number of infections during all seasons starting from the 2001-2002 flu season (flu seasons are represented as dotted rectangles in Figure 3.6). The GMRF method fails to recover the peak during the 2002-2003 season. The large uncertainty in population size estimation during the 1999-2000, 2000-2001, and at the beginning of 2005-2006 seasons is explained by the small number of coalescent events during those time periods, however, this uncertainty is larger in the GMRF recovered trajectory. During the 2001-2002 flu season, the GMRF method fails to recover the expected trajectory of a peak followed by a bottleneck and instead, this method recovers an epidemic that started during the end of 2001, increased and remained "at peak" until the end of the following winter. The GMRF recovered trajectory during the winter season of 2003 exhibits a steep decrease. In contrast, the GP method detects a late peak during the 2001-2002 season, followed by a decline in the number of infections. There is a small bump in the effective population size of influenza in the winter of 2003, which is more realistic than a steady decline in the number of infections estimated by the GMRF method. Overall, we believe that the GP reconstructed trajectory is more plausible from an epidemiological point of view than the GMRF population size reconstruction.

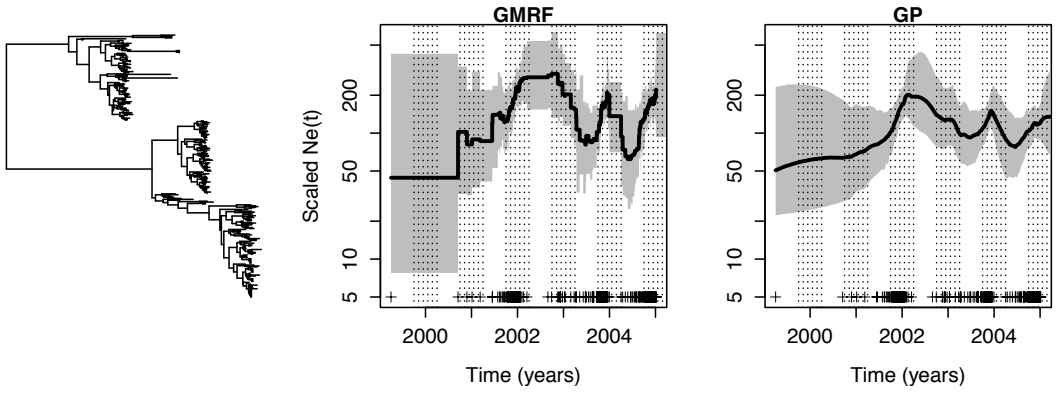

Figure 3.6: H3N2 Influenza A virus in New York state. The first plot (left) is the estimated genealogy. The second and third plots are the GMRF and GP estimations of log scaled effective population trajectories. Winter seasons are represented by the dotted shaded areas. Posterior medians are represented by solid black lines and 95% BCIs are represented by gray shaded areas.

3.4 Prior Sensitivity

In all our examples, we placed a Gamma prior on the precision parameter θ with parameters $\alpha = 0.001$ and $\beta = 0.001$. This precision parameter, unknown to us *a priori*, controls the smoothness of the GP prior. We investigate the sensitivity of our results to the Gamma prior specification using the Egyptian HCV data. In the first plot of Figure 3.7, we show the prior and posterior distributions of θ under our default prior. The difference in densities suggests that prior choices do not have an impact on the posterior distribution. Since the mean of a Gamma distributed random variable is α/β, we investigate the sensitivity by fixing $\beta = .001$ and setting the value of α to 0.001, 0.002, 0.005, 0.01 and 0.1, corresponding to prior means 1, 2, 5, 10 and 100 and variances 1000, 2000, 5000, 10000 and 100000, and by trying two extremes: $\alpha = 1$, $\beta = .0001$ and $\alpha = .001$, $\beta = 1$, to examine the posterior distribution of θ under these priors. The posterior sample boxplots displayed in Figure 3.7 demonstrate that our results are fairly robust to different choices of α.

3.5 Sensitivity to the Order of the Gaussian Process

We evaluate our GP-based method for three different Gaussian Process priors for the Egyptian HCV genealogy. In Figure 3.8, we show the recovered trajectories for Brownian Motion

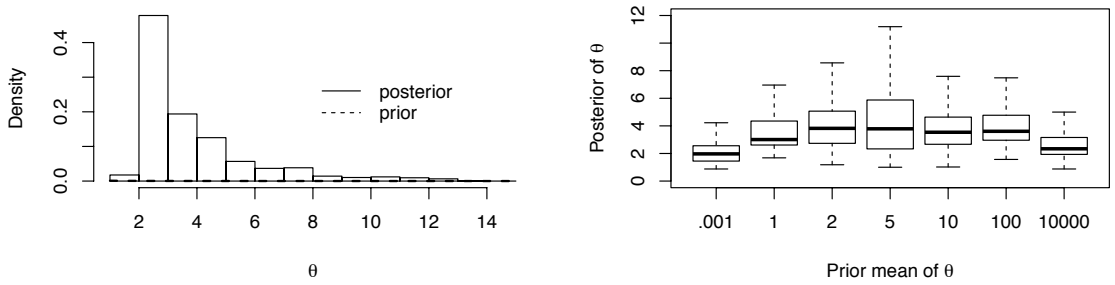

Figure 3.7: Prior sensitivity on the GP precision parameter. Left plot shows the prior and posterior distributions represented by dashed line and vertical bars respectively. Right plot shows the boxplots of the posterior distributions of the precision parameter when the prior distributions differ in mean and variance of the precision parameter θ. These plots are based on the Egyptian HCV data.

(BM), Ornstein-Uhlenbeck (OU) and approximated Integrated Brownian motion (IBM) (Lindgren and Rue, 2008). The common characteristic of these three priors is the sparsity of their precision matrices (inverse covariance matrix), allowing for computational tractability. Figure 3.8 shows that the order of the process does make a difference, but only in regions with large posterior uncertainty, where prior influence is more pronounced.

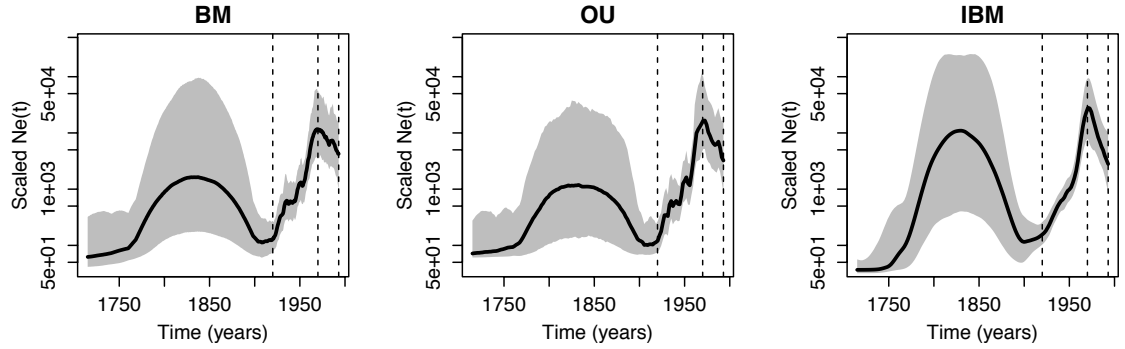

Figure 3.8: Egyptian HCV recoverd by placing three different Gaussian process priors. The first plot (left to right) corresponds to a Brownian motion (BM), the second – to Ornstein-Uhlenbeck (OU) and the last one – to the approximated integrated Brownian motion (IBM).

3.6 Discussion

We propose a new nonparametric method for estimating population size dynamics from gene genealogies. To the best of our knowledge, we are the first to solve this inferential problem using modern tools from Bayesian nonparametrics. In our approach, we assume that the population size trajectory *a priori* follows a transformed Gaussian process. This flexible prior allows us to model population size trajectory as a continuous function without specifying its parametric form and without resorting to artificial discretization methods. We tested our method on simulated and real data and compared it with the competing GMRF method. On simulated data, our method recovers the truth with better accuracy and precision. On real data, where true population trajectories are unknown, our method recovers known epidemiological aspects of the population dynamics and produces more reasonable estimates of uncertainty than the competing GMRF method.

We bring Bayesian nonparametrics into the coalescent framework by viewing the coalescent as a point process. This representation allows us to adapt Bayesian nonparametric methods originally developed for Poisson processes to the coalescent modeling. In particular, it allows us to adapt the thinning-based data augmentation for Poisson processes developed by Adams et al. (2009). We devise an analogous data augmentation for the coalescent by developing a series of new thinning algorithms for the coalescent. Although we use these algorithms in a very narrow context, our novel coalescent simulation protocols should be of interest to a wide range of users of the coalescent. For example, we are not aware of any competitors of our Algorithms 2 and 4 that allow one to simulate coalescent times with a continuously and *stochastically* varying population size.

Our method works with any Gaussian process with mean $\mathbf{0}$ and covariance matrix $\mathbf{C}$, where the latter controls the level of smoothness and autocorrelation. For computational tractability however, sparsity in the precision matrix (inverse covariance matrix) may be necessary for complex trajectories with a high number of latent points. One way to achieve sparse matrix computations and computational tractability is to use GP that is also Markov. In all our examples, we use Brownian motion with precision parameter θ; however, the non-differentiability characteristic of the Brownian motion is compensated by the fact that our

estimate of effective population trajectory is the posterior median evaluated pointwise, which is smoother than any of the sampled posterior curves. Additionally, we compared Brownian motion, Ornstein-Uhlenbeck and a higher order integrated Brownian motion for one of our examples and obtained very similar results under all three priors. Finite sample distributions under these three priors enjoy sparse precision matrices that yield computational tractability comparable to the GMRF method. In our Brownian motion prior, the precision parameter controls the level of smoothness of the estimated population size trajectory. We find that this important parameter shows little sensitivity to prior perturbations.

Our method assumes that a genealogy or tree is given to the researcher. However, genealogies are themselves inferred from molecular sequences, so we need to incorporate genealogical uncertainty into our estimation. Our framework can be extended to inference from molecular sequences instead of genealogies by introducing another level of hierarchical modeling into our Bayesian framework, similar to the work of Drummond et al. (2005) and Minin et al. (2008). Further, we plan to extend our method to handle molecular sequence data from multiple loci as in (Heled and Drummond, 2008). Finally, we would like to extend our nonparametric estimation into a multivariate setting, so that we can estimate cross correlations between population size trajectories and external time series. Estimating such correlations is a critical problem in molecular epidemiology.

We deliberately adapted the work of Adams et al. (2009) on estimating the intensity function of an inhomogeneous Poisson process to the coalescent, as opposed to alternative ways to attack this estimation problem (Møller et al., 1998; Kottas and Sansó, 2007). We believe that among the state-of-the-art Bayesian nonparametric methods, our adopted GP-based framework is the most suitable for developing the aforementioned extensions. First, it is straightforward to incorporate external time series data into our method by replacing a univariate GP prior with a multivariate process that evolves the population size trajectory and another variable of interest in a correlated fashion (Teh et al., 2005). Second, the fact that our method does not operate on a fixed grid is critical for relaxing the assumption of a fixed genealogy, because fixing the grid *a priori* is problematic when one starts sampling genealogies, including coalescent times, during MCMC iterations.

Finally, since the coalescent model with varying population size can be viewed as a

particular example of an inhomogeneous continuous-time Markov chain, all our mathematical and computational developments are almost directly transferable to this larger class of models. Therefore, our developments potentially have implications for nonparametric estimation of inhomogeneous continuous-time Markov chains with numerous applications.

Chapter 4

INTEGRATED NESTED LAPLACE APPROXIMATION IN PHYLODYNAMICS

4.1 Introduction

Estimation of population size dynamics from molecular data is a fundamental task in ecology and public health. Since population size fluctuations affect the variability of population gene frequencies, current molecular sequence data provide information about the past population size trajectory. Such indirect inference is particularly useful in retrospective studies, where assessing past population sizes via sampling or fossil records is impossible. For example, RNA samples of hepatitis C virus (HCV) obtained in 1993 were sufficient to estimate the dynamics of HCV infections in Egypt from 1895 to 1993 (Pybus et al., 2003); and ancient and modern musk ox mitochondrial DNA samples, dated from 56,900 radiocarbon years old to contemporaneous, allowed for estimation of musk ox population dynamics throughout the late Pleistocene to the present (Campos et al., 2010a).

Molecular sequence data of individuals sampled at a single time point (*isochronous* sampling) or at different points in time (*heterochronous* sampling) are related to each other via, a usually unknown, genealogical relationship. A genealogy is a rooted bifurcating tree that describes the ancestral relationships of the sampled individuals. In the genealogy, each internal node indicates that the two lineages met a common ancestor. Such events are called *coalescent events*, and these events occur at *coalescent times*.

Kingman's coalescent (Kingman, 1982) is a probability model that describes a stochastic process of generating a genealogy of a random sample of molecular sequences given the effective population size (Nordborg, 2001; Hein et al., 2005). The original formulation, that considered only a constant population size, was later generalized to a variable population size (Slatkin and Hudson, 1991b; Griffiths and Tavaré, 1994). Statistically, the coalescent model was an important advance, because it allowed for likelihood-based inference of population dynamics.

Many coalescent-based methods for estimation of effective population size trajectories have been developed over the last 10 years. Some methods may or may not consider the genealogical uncertainty and can produce estimates of population size trajectories from a fixed genealogy or directly from molecular data (Kuhner et al., 1995; Drummond et al., 2002, 2005; Minin et al., 2008). Felsenstein (1992) showed that likelihood-based methods that account for genealogical uncertainty are statistically the most efficient. However, all methods that incorporate genealogical uncertainty in population size dynamics reconstruction integrate over the space of genealogies using Markov chain Monte Carlo (MCMC). Such MCMC sampling of genealogies is computationally expensive. Sometimes, a single genealogy estimated from sequences that contain sufficient phylogenetic information is enough to estimate population trajectories accurately (Pybus et al., 2000; Minin et al., 2008). In this chapter, we are interested in providing a fast estimation of population size trajectories from a fixed genealogy.

Some coalescent-based methods assume a simple parametric form of the population size trajectory (e.g., exponential or logistic growth), allowing the model parameters to be estimated by maximum likelihood or Bayesian methods. However, more flexible nonparametric methods are preferable for populations with poorly understood population dynamics, where it may be difficult to justify a simple parametric form of the population size trajectory. In fact, all recently developed methods rely on Bayesian nonparametric techniques to perform inference (Opgen-Rhein et al., 2005; Drummond et al., 2005; Heled and Drummond, 2008; Minin et al., 2008; Palacios and Minin, 2013). A common characteristic of most of these methods is the assumption of a piece-wise linear trajectory of effective population sizes and the possibility of the number of parameters growing with the number of samples. Bayesian skyline methods (Drummond et al., 2005; Heled and Drummond, 2008) and Opgen-Rhein et al. (2005) use multiple change point models to estimate population trajectories in a Bayesian framework. The method of Opgen-Rhein et al. (2005) is implemented only for a fixed genealogy. Recently, Bayesian nonparametric approaches that rely on Gaussian processes have been successfully implemented (Minin et al., 2008; Palacios and Minin, 2013). These methods model the effective population size as a function of a Gaussian process (GP) *a priori*, providing more flexible priors than previous Bayesian nonparametric methods.

GP-based models use MCMC methods to perform Bayesian inference. We show that when the genealogy remains fixed, these models fall into a general class of latent Gaussian models, for which integrated nested Laplace approximation (INLA) can be used to perform computationally efficient approximate Bayesian inference (Rue et al., 2009; Illian et al., 2012). Here, we adapt the INLA methodology to the estimation of population size trajectories and replace MCMC entirely. Our approximation is accurate and much faster than MCMC, while still providing the benefits of the Gaussian process-based Bayesian nonparametric approach. We illustrate the performance of our method with simulated and two real data sets.

4.2 Coalescent Background

We assume that a genealogy with time measured in units of generations is available. Let $t_n = 0$ denote the present time when all n available sequences are sampled (*isochronous*) and let $t_n = 0 < t_{n-1} < ... < t_1$ denote the coalescent times of lineages in the genealogy. Let $N_e(t)$ denote the time evolution of the effective population size as we move into the past. Then, the conditional density of the coalescent time t_{k-1}, given the previous coalescent time t_k, takes the following form:

$$P[t_{k-1}|t_k, N_e(t)] = \frac{C_k \exp\left[-\int_{t_k}^{t_{k-1}} \frac{C_k}{N_e(t)} dt\right]}{N_e(t_{k-1})}, \tag{4.1}$$

where $C_k = \binom{k}{2}$ is the coalescent factor that depends on the number of lineages $k = 2, ..., n$, meaning that the density for the next coalescent time is quadratic in the number of lineages and inversely proportional to the effective population size. The larger the population size, the more genetic variability is in the population and hence, the longer it takes for two lineages to coalesce. The larger the number of lineages, the faster two of them meet their common ancestor.

The *heterochronous* coalescent arises when samples of sequences are collected at different times. The conditional density of a coalescent time t_{k-1} is slightly different than equation 4.1 since it takes into account the fact that the number of lineages at each time point depends not only on the number of coalescent events (in which case, the number of lineages decreases by one each time), but also on the new samples incorporated into the analysis at any time

after the last coalescent time t_k. See Felsenstein and Rodrigo (1999) and Drummond et al. (2002) for a more detailed account of heterochronous sampling.

Under this coalescent-based framework, we ignore the effects of population structure, recombination and selection (Nordborg, 2001). The parameter of interest, the effective population size, can be used to approximate census population size by knowing the generation time in calendar units and the population variability in the number of offspring. The latter quantity might be difficult to know *a priori*, however, sometimes it suffices to analyze an arbitrarily rescaled population size trajectory, assuming the variability in the number of offspring remains constant.

4.2.1 Estimation of $N_e(t)$ using a discrete-time GMRF

There are two approaches to estimation of effective population size trajectories that use Gaussian processes. The first approach, developed by Minin et al. (2008), assumes *a priori* that given a genealogy, the effective population size trajectory is a piecewise constant trajectory with change points (knots) placed at coalescent times. That is,

$$N_e(t) = \sum_{k=2}^{n} \exp(\gamma_k)\, 1_{(t_k, t_{k-1}]}(t), \tag{4.2}$$

where

$$\boldsymbol{\gamma} = (\gamma_2, ..., \gamma_k) \sim MVN\left(0, (\tau \mathbf{Q})^{-1}\right) \text{ and}$$

$$1_{(t_k, t_{k-1}]}(t) = \begin{cases} 1 & \text{if } t \in (t_k, t_{k-1}], \\ 0 & \text{otherwise.} \end{cases}$$

More specifically, *a priori* $\boldsymbol{\gamma}$ is assumed to be an intrinsic Gaussian Markov random field (GMRF) on a chain graph connecting nodes 2 through n. Minin et al. (2008) used a random walk of the first order (rw1) on an irregular grid of mid-points of inter-coalescent time intervals. For this reason, we refer to this method here as the coalescent grid Gaussian process (CGGP). The random walk construction implies that matrix $\mathbf{Q}$ is tridiagonal and positive semidefinite (hence the intrinsic GMRF). See (Rue and Held, 2005) for background on GMRFs.

The precision parameter τ has a Gamma prior distribution with $\alpha = \beta = 0.001$. The authors estimate $\boldsymbol{\gamma}$ and τ by MCMC sampling from the posterior distribution of these

parameters. The estimated trajectory and the corresponding uncertainty are reported in the form of pointwise posterior medians and 95% Bayesian credible intervals (BCIs) obtained from the MCMC samples.

4.2.2 Estimation of $N_e(t)$ using a continuous-time GP

Instead of modelling $N_e(t)$ as a piecewise continuous function *a priori*, Palacios and Minin (2013) propose a more flexible prior specification and place a transformed Gaussian process prior on $N_e(t)$. The transformation is a sigmoidal function with a lower bound. This particular transformation is required in order to perform exact posterior inference via a data augmentation scheme, which is similar to the work of Adams et al. (2009). However, a log-Gaussian transformation using a finely discretized Gaussian process, in principle, would produce similar results (Møller et al., 1998; Adams et al., 2009).

Exact posterior inference with GP

Palacios and Minin (2013) place the following prior on $N_e(t)$:

$$N_e(t) = \left(\frac{\lambda}{1 + \exp[-\gamma(t)]} \right)^{-1}, \tag{4.3}$$

where

$$\gamma(t) \sim \mathcal{GP}(0, C) \tag{4.4}$$

and $\mathcal{GP}(0, C)$ denotes a Gaussian process with mean function 0 and covariance function C. A Gaussian process restricted to finite data is a multivariate Gaussian distribution. That is, $\gamma(t_1), ..., \gamma(t_B) \sim MVN(\mathbf{0}, \Sigma)$. *A priori*, $1/N_e(t)$ is a sigmoidal Gaussian process, a scaled logistic function of a Gaussian process which range is restricted to lie in $[0, \lambda]$; λ is a positive constant hyperparameter, inverse of which serves as a lower bound of $N_e(t)$ (Adams et al., 2009). The likelihood function is the product of the conditional densities in equation (4.1) and involves integration of $N_e(t)$, that under the $\mathcal{GP}$ assumption, is intractable. Following earlier work by Adams et al. (2009) on Poisson processes, we performed inference assuming an augmented data likelihood which allows to bypass intractability in the likelihood in previous work. We implemented our method for the Brownian motion $\mathcal{GP}$ with a precision parameter τ. We placed a Gamma prior distribution on the precision hyperparameter τ

with $\alpha = \beta = 0.001$ and a mixture of uniform and exponential distributions on an upper bound of $1/N_e(t)$ (or equivalently, a lower bound on $N_e(t)$) as follows:

$$P(\lambda) = \epsilon \frac{1}{\hat{\lambda}} I_{\{\lambda < \hat{\lambda}\}} + (1-\epsilon) \frac{1}{\hat{\lambda}} e^{-\frac{1}{\hat{\lambda}}(\lambda - \hat{\lambda})} I_{\{\lambda \geq \hat{\lambda}\}}, \tag{4.5}$$

where $\epsilon > 0$ is a mixing proportion and $\hat{\lambda}$ is our best guess of the upper bound, possibly obtained from previous studies. We estimated τ and $N_e(t)$, or equivalently, τ, $\gamma(t)$ and λ by MCMC sampling from the posterior distribution of these parameters. The estimated trajectory and the corresponding uncertainty are reported in the form of the pointwise posterior medians and 95% BCIs evaluated at a grid of points $\{s_1, ..., s_B\}$ obtained from the MCMC samples. This grid can be made as fine as necessary after the MCMC is finished. The values of $\{\gamma(s_1), \gamma(s_2), .., .\gamma(s_B)\}$ are obtained via the $\mathcal{GP}$ predictive distribution conditioning on the values of each iteration. This method will be referred to as exact Gaussian process (EGP).

Discretized continuous-time GP

The continuous-time version of the prior specified in Eq. 4.2, is

$$N_e(t) = \exp\left[\gamma(t)\right], \tag{4.6}$$

where $\gamma(t)$ is the Gaussian process described in equation (4.4). However, for the same reason described in the previous section, the likelihood function becomes intractable. Palacios and Minin (2013) showed that estimation of the effective population size is analogous to the estimation of an inhomogeneous intensity of a point process. In this context, and under the prior described in Eq. 6, estimation of $N_e(t)$ is computationally equivalent to the estimation of the intensity function of a Log-Gaussian Cox process (Møller et al., 1998). In a Log-Gaussian Cox process, the likelihood is commonly approximated by discretization. The approximation method proceeds by constructing a fine regular grid $\{s_1, ..., s_B\}$ over the observation window and approximate

$$\int \frac{dt}{N_e(t)} = \int \exp\left[-\gamma(t)\right] dt, \tag{4.7}$$

by

$$\sum_{j=2}^{B} \exp\left(-\gamma_j^*\right) \Delta, \tag{4.8}$$

where Δ is the distance between grid points, and γ_j^* is a representative value of $\gamma(t)$ in the interval (s_{j-1}, s_j), usually at the midpoint. Note that if the Gaussian process is a Brownian motion process, this approximation is similar to the CGGP method described in section 4.2.1. The difference is in the construction of the grid. In the CGGP method, the grid is irregular and determined by the coalescent times. For this reason, we call approximation (4.8) a regular grid Gaussian process (RGGP).

4.3 Integrated Nested Laplace Approximation

INLA provides fast and accurate Bayesian approximation to posterior marginals in *latent Gaussian models* (Rue et al., 2009). Latent Gaussian models are a wide class of hierarchical models in which the response variables $\mathbf{y} = (y_1, \ldots, y_n)$ are assumed to be conditionally independent given some latent parameters $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_n)$ and other parameters $\boldsymbol{\theta}_1$. The second hierarchical level corresponds to specifying $\boldsymbol{\eta}$ as a function of a GMRF $\mathbf{x} = (x_1, \ldots, x_n)$ with a precision matrix $\mathbf{Q}$ and hyperparameters $\boldsymbol{\theta}_2$, and the third and last hierarchical stage corresponds to prior specifications for the hyperparameters $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ Formally,

$$\pi(\mathbf{y}|\boldsymbol{\eta}, \boldsymbol{\theta}_1) = \prod_j \pi(y_j|\eta_j(x_j), \boldsymbol{\theta}_1), \tag{4.9}$$

$$\mathbf{x} \sim MVN(\mathbf{0}, \mathbf{Q}^{-1}(\boldsymbol{\theta}_2)), \tag{4.10}$$

and

$$\boldsymbol{\theta} \sim P(\boldsymbol{\theta}). \tag{4.11}$$

An interface in `R`, called `INLA`, implements a wide variety of likelihoods (equation (4.9)), link functions ($\boldsymbol{\eta}$) and GMRFs (equation (4.10)), including the Poisson likelihood model for each observed value of y_j (not necessarily the same for every y_j) with a logarithmic additive link function and random walk of first order as a GMRF. See `www.r-inla.org` for documentation.

The coalescent with variable population size (equation (4.1)), together with the GMRF prior specification (equation (4.2)) falls into the latent Gaussian model class, so INLA can be implemented for these coalescent models. In the case of the continuously specified $\mathcal{GP}$

(section 4.2.2), the approximate posterior method described in Section 4.2.2.2 (RGGP) also falls into the latent Gaussian model class.

4.3.1 INLA for Phylodynamics

Although INLA is implemented for a wide variety of latent Gaussian models, we will only describe the main steps of the approximation for posterior inference of effective population size trajectories. A typical summary of the posterior distribution of the effective population size trajectory, $N_e(t)$, is described by posterior medians and 95% BCIs evaluated pointwise on a grid of time points. These values can be obtained from the posterior marginals of the population trajectory on the grid. For the CGGP model described in section 4.2.1, we then wish to obtain the posterior marginals

$$\Pr(\gamma_i|\mathbf{t}) = \int_0^\infty \Pr(\gamma_i|\tau,\mathbf{t})\Pr(\tau|\mathbf{t})d\tau, i = 2,..,n \tag{4.12}$$

and

$$\Pr(\tau|\mathbf{t}), \tag{4.13}$$

where $\mathbf{t}$ denotes the vector of coalescent times. A nested procedure is used to construct approximations of $\Pr(\gamma_i|\tau,\mathbf{t})$ and $\Pr(\tau|\mathbf{t})$ first and then numerically integrate out τ to arrive at $\Pr(\gamma_i|\mathbf{t})$. The approximation of the marginal of τ is

$$\tilde{\Pr}(\tau|\mathbf{t}) \propto \left.\frac{\Pr(\boldsymbol{\gamma},\tau,\mathbf{t})}{\tilde{\Pr}_G(\boldsymbol{\gamma}|\tau,\mathbf{t})}\right|_{\boldsymbol{\gamma}^*(\tau)}, \tag{4.14}$$

where $\boldsymbol{\gamma}^*(\tau)$ is the mode of the full conditional $\Pr(\boldsymbol{\gamma}|\tau,\mathbf{t})$, obtained using the Newton-Raphson algorithm, and $\tilde{\Pr}_G(\boldsymbol{\gamma}|\tau,\mathbf{t})$ is the Gaussian approximation of this full conditional constructed via a Taylor expansion around $\boldsymbol{\gamma}^*(\tau)$. The resulting $\tilde{\Pr}_G(\boldsymbol{\gamma}|\tau,\mathbf{t})$ is a Gaussian distribution with mean $\boldsymbol{\gamma}^*$ and precision matrix $\mathbf{Q}$ +diag($\mathbf{c}$), where $\mathbf{Q}$ is the prior precision matrix of the GMRF $\boldsymbol{\gamma}$ and a vector $\mathbf{c}$ consists of the second order Taylor series coefficients.

The approximation to the full conditional $\Pr(\gamma_i|\tau,\mathbf{t})$ is the following:

$$\tilde{\Pr}(\gamma_i|\tau,\mathbf{t}) \propto \left.\frac{\Pr(\boldsymbol{\gamma},\tau,\mathbf{t})}{\tilde{\Pr}_G(\boldsymbol{\gamma}_{-i}|\tau,\mathbf{t})}\right|_{\boldsymbol{\gamma}^*_{-i}}, \tag{4.15}$$

where $\boldsymbol{\gamma}^*_{-i} = \mathrm{E}_G(\boldsymbol{\gamma}_{-i}|\gamma_i,\tau,\mathbf{t})$ and $\tilde{\Pr}_G(\boldsymbol{\gamma}_{-i}|\tau,\mathbf{t})$ are derived from $\tilde{\Pr}_G(\boldsymbol{\gamma}|\tau,\mathbf{t})$.

For the continuously specified GP approximation described in section 4.2.2, the INLA approximation is, in essence the same, but the GMRF is placed at the mid-points of a finer and regular grid. In this case, there are two levels of approximation, one level corresponding to the likelihood discretization and another level corresponding to the approximation of marginal posterior distributions of model parameters.

4.4 Results

4.4.1 Simulated Data

We compare INLA and MCMC approaches for the models described in sections 4.2.1 and 4.2.2. With time measured in expected mutations per site, we simulate three genealogies relating $n = 100$ individuals under the following demographic scenarios:

1. Constant population size trajectory: $N_e(t) = 1$.

2. Exponential growth: $N_e(t) = 25e^{-5t}$.

3. Population expansion followed by a crash:

$$N_e(t) = \begin{cases} e^{4t} & t \in [0, 0.5], \\ e^{-2t+3} & t \in (0.5, \infty). \end{cases} \tag{4.16}$$

Figure 4.1 shows the log effective population size trajectories recovered for the three scenarios under the CGGP model using the MCMC approach (black lines and gray shaded areas) and the INLA approach (blue dark lines and blue dashed lines). In all the cases, the INLA approximation is very close to the results obtained using MCMC.

Figure 4.2 shows the log effective population size trajectories recovered for the same three scenarios for the continuously specified GP. In this case, the comparison is not entirely fair because we are comparing the exact MCMC method (EGP) with the doubly approximated INLA on the RGGP model. Nevertheless, both estimations look very similar for the last two cases (exponential growth and expansion followed by crash). In all cases, INLA results are very similar to the results for the CGGP model and the difference between the MCMC method and INLA methods in the constant trajectory example could be an artifact of the

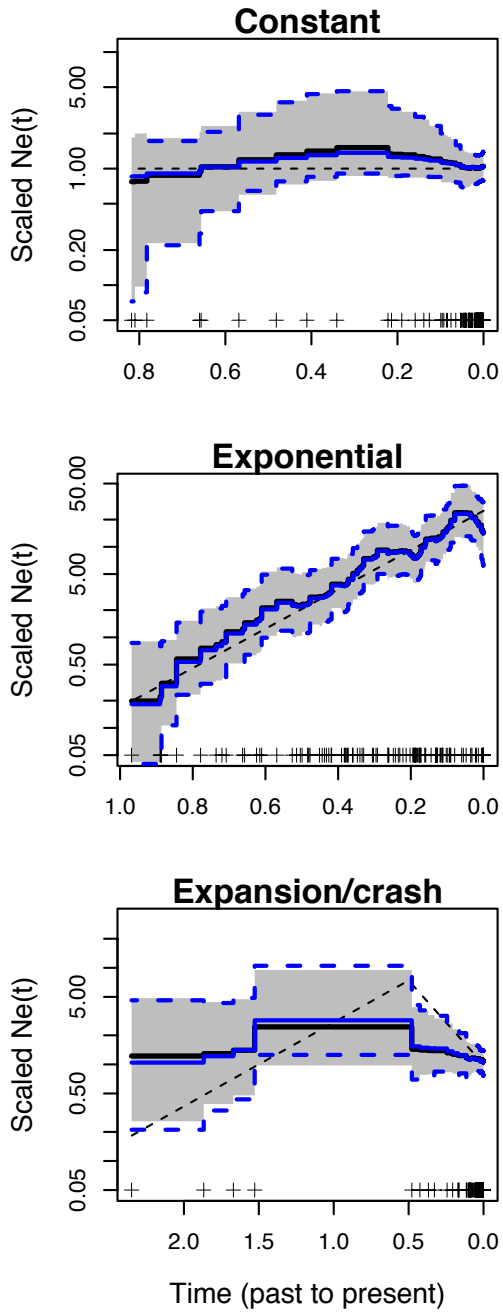

Figure 4.1: INLA vs MCMC for CGGP: Simulated data under the constant population size (first row), exponential growth (second row) and expansion followed by a crash (third row). The true trajectories are represented by black dashed lines. We show posterior medians estimated with MCMC sampling (solid black lines) and 95% BCIs estimated with MCMC (gray shaded areas). Posterior medians obtained using INLA are denoted by solid blue lines and INLA 95% BCIs are shown as dashed blue lines.

likelihood approximation and the convergence of the MCMC method. However, a more likely explanation is poor approximation of the marginal posterior of the Brownian motion precision, τ, by INLA. Indeed, when we examined MCMC-based and INLA-based marginal posteriors of τ, we found that the two marginals did not agree at all. We recommend to use the INLA approximation for rapid estimation, however one needs to be cautious about the precision estimated by this approximation.

4.4.2 Hepatitis C virus in Egypt

We analyze a genealogy estimated from 63 HCV E1 sequences sampled in 1993 in Egypt. This is perhaps the most commonly used dataset for evaluating different methodologies for estimation of population size trajectories. Minin et al. (2008) compared population size trajectories recovered from a single fixed genealogy and from the sequence data directly. The authors show that there is little difference between these two estimation protocols. They argue that in this case genealogical uncertainty does not play a significant role in the estimation of the Egyptian HCV population dynamics.

Figure 4.3 shows the recovered effective population sizes as black lines and uncertainty as gray shaded areas for the CGGP (left plot) and the EGP (right plot) using MCMC and as blue solid lines and blue dashed lines for the INLA approximation for CGGP (left plot) and RGGP (right plot). In this case, it is remarkable how similar the INLA approximations are to the MCMC results, even for the continuously specified model with the double approximation (INLA-RGGP). In all cases, the known aspects of the HCV epidemic in Egypt are recovered: an exponential growth starting around 1920s and a decline in population size after 1970s (Pybus et al., 2003).

4.4.3 Influenza A virus in New York

We analyze a genealogy estimated from 288 H3N2 sequences sampled in New York state from January, 2001 to March, 2005 to estimate population size dynamics of human influenza A in New York. This genealogy has also been analyzed before (Palacios and Minin, 2013) and can be obtained from the authors. The key aspects of the influenza A virus epidemic in temperate regions like New York are the epidemic peaks during winters followed by

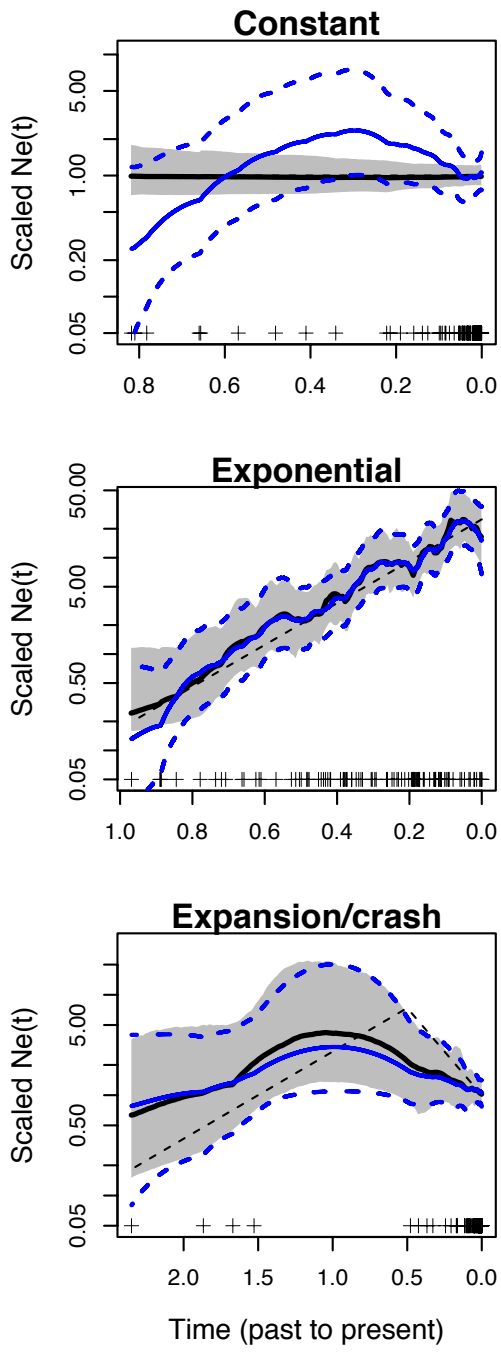

Figure 4.2: INLA vs MCMC for RGGP and EGP respectively: see Figure 4.1 for the legend.

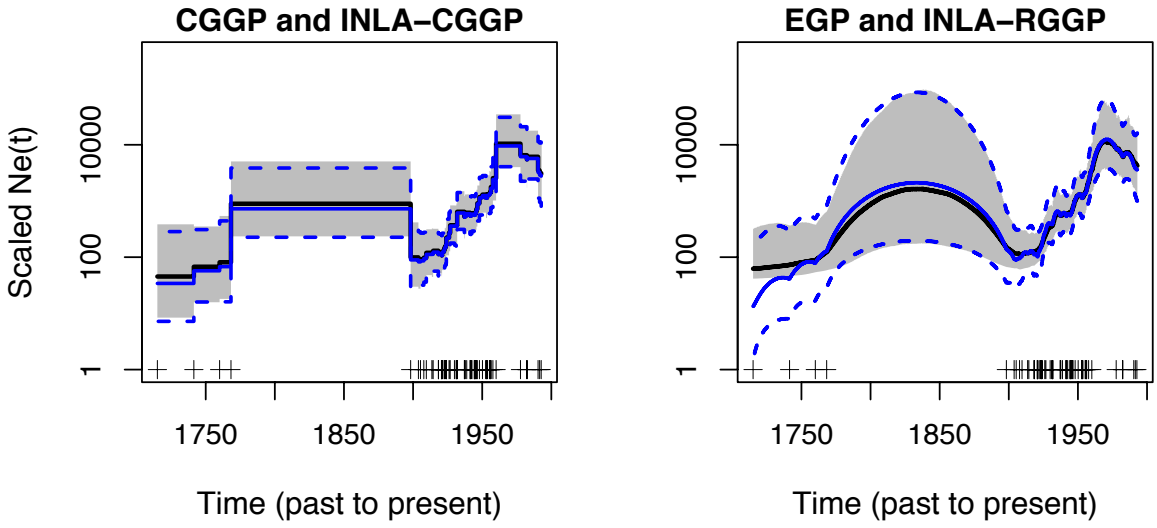

Figure 4.3: HCV in Egypt. Estimation of the log effective population size trajectories. In both plots, INLA approximations to posterior medians and 95% BCIs are represented by blue solid lines and blue dashed lines respectively. Approximations using MCMC sampling are represented by black solid lines and shaded areas. The left plot shows the results assuming the CGGP model and the right plot shows the result assuming the EGP for the MCMC sampling results and the RGGP model for the INLA approximation.

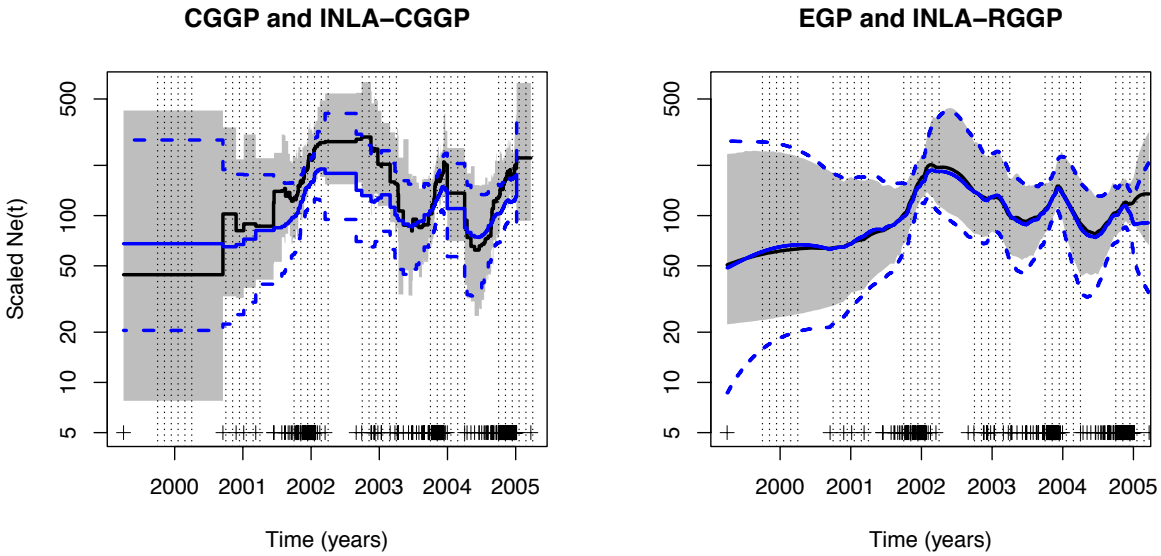

Figure 4.4: Influenza A in New York. Estimation of the log effective population size trajectories. In both plots, INLA approximations to posterior medians and 95% BCIs are represented by blue solid lines and blue dashed lines respectively. Approximations using MCMC sampling are represented by black solid lines and shaded areas. The left plot shows the results assuming the CGGP model and the right plot shows the result assuming the EGP for the MCMC sampling results and the RGGP model for the INLA approximation.

strong bottlenecks at the end of the winter season. The first plot in Figure 5 shows the recovered population size trajectories assuming the CGGP model. In this case, the MCMC and the INLA approximation deviate from each other substantially, however, the expected peaks during the winter seasons in 2002, 2004 and 2005 are recovered by both methods. The MCMC approach does not recover a peak in the 2003 season, while the INLA approximation resemble more the results from the continuously specified model. INLA and MCMC results are very similar for the continuously specified model (right plot of Figure 4.4) with the notable differences in 95% BCIs near the time to the most recent common ancestor. This difference again may be an artifact of the double approximation involved.

4.4.4 Running times

The MCMC chains used for the CGGP model have length 1,000,000 with 100,000 of burn-in and generated using the BEAST software (Drummond and Rambaut, 2007; Minin et al.,

2008) on a desktop PC. The running times range from 20 minutes to a couple of hours depending on the data. For the INLA approach, results were generated using the `R` interface `INLA` on the same computer in less than 2 seconds for all scenarios.

For the continuously specified GP model described in section 4.2.2, MCMC times are at best as fast as MCMC for the CGGP approach, while the results obtained using `INLA`, were generated in less than 5 seconds on a grid of size 1000.

4.5 Discussion

We show that recent Gaussian process-based Bayesian nonparametric approaches to estimation of effective population size trajectories fall into a larger class of latent Gaussian models, allowing us to perform approximate Bayesian inference using INLA. We show that it is possible to estimate population size trajectories from fixed genealogies in seconds without sacrificing any modeling advantages of recently developed Bayesian nonparametric methods.

We did observe a significant discrepancy between the INLA approximation and MCMC inference for the continuously specified GP model in the case of constant population size. We want to point out that in this case, we are not comparing apples to apples. We should be comparing INLA approximation to the MCMC for the regular grid approximation of the continuously specified GP. However, we did not have access to approximate GP-based MCMC for phylodynamics. In the absence of a better option, we are comparing INLA to the *exact* MCMC for this GP model (Palacios and Minin, 2013). Therefore, we remain uncertain whether the grid approximation or the INLA approximation is to blame for the discrepancy observed in the top plot of Figure 4.2. The discrepancy between the marginal posterior distributions estimated by INLA and MCMC and the fact that the precision of the RGGP likelihood discretization did not have any effect on our results suggest that INLA approximation indeed fails in this simulation scenario. This assertion is supported by another disagreement of INLA and MCMC for the CGGP model in the influenza A example, where we are comparing them under the same discretization.

A natural extension of the methods presented here is the incorporation of genealogical uncertainty into the model. This extension can be accomplished by introducing another level of hierarchical modeling and analyzing molecular data directly (Drummond et al., 2005;

Minin et al., 2008). Even though the full posterior distribution of population trajectories from molecular sequence data no longer falls into the latent Gaussian model class, we believe that the extension is possible using Metropolis independence sampler (Rue et al., 2004). Nevertheless, the ability to obtain fast estimates of population size trajectories from a fixed genealogy (as with INLA) should be a boon for biological researchers who need to screen multiple populations of interest quickly or to provide an online analysis of epidemic outbreaks with enormous flow of molecular data in real time (Fraser et al., 2009).

There are other approaches to the estimation of effective population sizes under more complicated coalescent models that include recombination (McVean and Cardin, 2005; Li and Durbin, 2011). These methods assume a simple change point model for the effective population size trajectory. In principle, Bayesian nonparametric approaches similar to the approaches discussed here can be applied in this setting. However, presence of recombination makes such extensions potentially challenging.

Other approximate Bayesian methods could be applied to Bayesian nonparametric phylodynamics, such as variational Bayes (VB) (Bishop, 2006) and expectation propagation (EP) (Cseke and Heskes, 2010). For our particular application with a sparse GP prior, such as Brownian motion, Cseke and Heskes (2010) show that INLA should be faster than EP methods.

Chapter 5

GAUSSIAN PROCESS-BASED BAYESIAN NONPARAMETRIC INFERENCE OF PHYLODYNAMICS DIRECTLY FROM MOLECULAR SEQUENCES

5.1 Introduction

The effective population size is one of the most important parameters that affect genetic variation. For example, a population with no migration and a small number of individuals sustained over several generations, will lose most of its genetic variation. Estimation of effective population sizes is therefore of paramount importance in areas such as conservation biology and epidemiology, with the latter aiming at containing viral prevalence. Genetic variation present in genetic sequence data allows us to recover past population dynamics under some simplifying assumptions. A popular neutral model used to understand genetic variation present in a random sample of individuals from a single population assumes that the observed genetic variability is the consequence of mutation forces acting at random on the genealogical tree of the samples. A genealogical tree or gene genealogy is a bifurcating tree with tips representing individuals' genetic data at the time when the genetic sequences are sampled and the lineages back in time represent the ancestry of the sample. The root of the tree denotes the time when all lineages meet their most recent common ancestor. Researchers are able to observe genetic sequences and possibly gain information about substitution rates from a combination of *in vivo* and *in vitro* experiments, but gene genealogies are hardly available. Kingman's coalescent (Kingman, 1982) allows us to model gene genealogies in a probabilistic framework where the rate at which two lineages meet a common ancestor in the past (coalescent event) depends on the number of lineages and the effective population size. The coalescent model has proven to be a good model to approximate genealogical relationships (Wakeley and Ramachandran, 2012).

Several coalescent-based methods to infer effective population size trajectories have been proposed in the last 15 years. Most of these methods model the effective population size trajectory as a piece-wise linear trajectory (Heled and Drummond, 2008; Minin et al., 2008) or as a specific functional form, such as exponential growth (Griffiths and Tavaré, 1994; Kuhner et al., 1998; Drummond et al., 2002). Recently, a Gaussian process (GP) based Bayesian nonparametric method has been proposed (Palacios and Minin, 2013). This method models the effective population size trajectory as a continuous function in time in a nonparametric framework. More specifically, the method of Palacios and Minin (2013) places a transformed Brownian motion prior on the effective population size trajectory. However, this method is only implemented for a fixed genealogy. We extend this methodology to perform inference of effective population size dynamics from genetic sequence data directly. We equip modern Bayesian methods implemented in `BEAST` (Drummond et al., 2012) with the GP-based prior on effective population size dynamics. We test our method on a real data set of hepatitis virus C in Egypt.

5.2 Methods

5.2.1 Likelihood for Sequence Alignment

Let $\mathbf{Y}$ denote an $n \times L$ sequence alignment matrix of n individuals randomly sampled from the population of interest. The sequence alignment of length L usually represents DNA or RNA sequences from a protein coding region or a "gene." We assume that there is no recombination possible between the sequences, so that we can assume the existence of one genealogical tree $\mathbf{g}$ relating all sequences. We assume that the sequence alignment $\mathbf{Y}$ is generated by a substitution process, defined by a parameter vector $\mathbf{m}$, acting on the genealogy $\mathbf{g}$. This construction yields the following likelihood function:

$$P(\mathbf{Y} \mid \mathbf{g}, \mathbf{m}). \tag{5.1}$$

We assume standard finite-sites neutral substitution models that randomly substitute one nucleotide for another at individual sites according to a continuous-time Markov chain (Tavaré, 2004). The parameter vector $\mathbf{m}$ contains parameters such as the transition/transversion ratio and a global mutation rate μ that scales the genealogy from units of mean number

of substitutions to calendar units. In principle, we can use any of the substitution models implemented in `BEAST` (Drummond et al., 2012), denoting by $p(\mathbf{m})$ the prior distribution on $\mathbf{m}$. The likelihood function equation (5.1) is referred to as the Felsenstein likelihood and computed via a recursive method, known as the *peeling algorithm* (Felsenstein, 1981).

5.2.2 Coalescent Prior

We proceed in a hierarchical Bayesian framework by putting a coalescent prior distribution on the genealogy $\mathbf{g}$. Let $t_n = 0$ be the time of sampling (*isochronous sampling*) and let $0 < t_{n-1} < ... < t_1$ denote the *coalescent times*. The coalescent prior is

$$P[\mathbf{g} \mid N_e(t)] = \prod_{k=2}^{n} \frac{1}{C_k} P[t_{k-1} \mid t_k, N_e(t)], \tag{5.2}$$

where $C_k = \binom{k}{2}$ is the coalescent factor that depends on the number of lineages $k = 2, ..., n$ and

$$P[t_{k-1} \mid t_k, N_e(t)] = \frac{C_k}{N_e(t_{k-1})} \exp\left\{ - \int_{t_k}^{t_{k-1}} \frac{C_k}{N_e(t)} dt \right\}, \tag{5.3}$$

denotes the conditional density of the coalescent time t_{k-1}. Since the genealogy contains both coalescent times and topology, we need to take the combinatorial factor C_k out of expression (5.3) to get expression (5.2).

For the *heterochronous coalescent*, when samples of sequences are collected at different times, let $t_n = 0 < t_{n-1} < \cdots < t_1$ denote the coalescent times as before, but now let $s_m = 0 < s_{m-1} < \cdots < s_1 < s_0 = t_1$ denote sampling times of $n_m, \ldots, n_1$ sequences respectively, $\sum_{j=1}^{m} n_j = n$. Further, let $\mathbf{s}$ and $\mathbf{n}$ denote the vectors of sampling times and numbers of sequences sampled at these times, respectively. Then the coalescent prior becomes

$$P[\mathbf{g} \mid \mathbf{s}, \mathbf{n}, N_e(t)] = \prod_{k=2}^{n} \frac{1}{C_{0,k}} P[t_{k-1} \mid t_k, \mathbf{s}, \mathbf{n}, N_e(t)], \tag{5.4}$$

where

$$P[t_{k-1} | t_k, \mathbf{s}, \mathbf{n}, N_e(t)] = \frac{C_{0,k}}{N_e(t_{k-1})} \exp - \left\{ \int_{I_{0,k}} \frac{C_{0,k}}{N_e(t)} dt + \sum_{i=1}^{m} \int_{I_{i,k}} \frac{C_{i,k}}{N_e(t)} dt \right\}, \tag{5.5}$$

and the coalescent factor $C_{i,k} = \binom{n_{i,k}}{2}$ depends on the number of lineages $n_{i,k}$ in the interval $I_{i,k}$ defined by coalescent times and sampling times. We denote half-open intervals that end

with a coalescent event by

$$I_{0,k} = (max\{t_k, s_j\}, t_{k-1}], \tag{5.6}$$

for $s_j < t_{k-1}$ and $k = 2, \ldots, n$, and half-open intervals that end with a sampling event by

$$I_{i,k} = (max\{t_k, s_{j+i}\}, s_{j+i-1}], \tag{5.7}$$

for $s_{j+i-1} > t_k$ and $s_j < t_{k-1}, k = 2, \ldots, n,$.

5.2.3 Gaussian Process Prior for Population Size Trajectories

For both isochronous or heterochronous data, we place the same prior on $N_e(t)$:

$$N_e(t) = \left[\frac{\lambda}{1 + \exp\{-f(t)\}}\right]^{-1}, \tag{5.8}$$

where

$$f(t) \sim \mathcal{GP}(\mathbf{0}, \mathbf{C}(\boldsymbol{\theta})) \tag{5.9}$$

and $\mathcal{GP}(\mathbf{0}, \mathbf{C}(\boldsymbol{\theta}))$ denotes a Gaussian process with mean function $\mathbf{0}$ and covariance function $\mathbf{C}(\boldsymbol{\theta})$ with hyperparameters $\boldsymbol{\theta}$. *A priori*, $1/N_e(t)$ is a Sigmoidal Gaussian Process, a scaled logistic function of a Gaussian process whose range is restricted to lie in $[0, \lambda]$; λ is a positive constant hyperparameter, the inverse of which serves as a lower bound of $N_e(t)$ (Adams et al., 2009). For computational convenience, we use Brownian motion with precision parameter θ as our Gaussian process prior.

5.2.4 Priors for Hyperparameters

We place a Gamma prior distribution with parameters α and β on the precision parameter θ. The other hyperparameter in our model is the upper bound of $1/N_e(t)$, λ. We will assume that it is possible to obtain an upper bound λ (or equivalently, a lower bound on $N(t)$) from previous studies. This assumption can be relaxed by placing a prior on λ as described in Chapter 3.

5.2.5 Data Augmentation and Inference

Let coalescent times be denoted by $\mathcal{T} = \{t_n, t_{n-1}, \ldots, t_1\}$. Then, equations (5.2) and (5.4) can be expressed as

$$P(\mathbf{g} \mid N_e(t)) = \frac{P(\mathcal{T} \mid N_e(t))}{\prod_{k=2}^{n} C_k}, \tag{5.10}$$

and

$$P(\mathbf{g} \mid \mathbf{s}, \mathbf{n}, N_e(t)) = \frac{P(\mathcal{T} \mid \mathbf{s}, \mathbf{n}, N_e(t))}{\prod_{k=2}^{n} C_{0,k}}, \tag{5.11}$$

respectively. We are now ready to define the posterior distribution of all model parameters:

$$P(\mathbf{g}, \mathbf{m}, N_e(t), \boldsymbol{\theta} \mid \mathbf{Y}) \propto P(\mathbf{Y} \mid \mathbf{g}, \mathbf{m}) P(\mathbf{m}) P(\mathcal{T} \mid N_e(t)) P(N_e(t) \mid \boldsymbol{\theta}) P(\boldsymbol{\theta}), \tag{5.12}$$

for isochronous data and

$$P(\mathbf{g}, \mathbf{m}, N_e(t), \boldsymbol{\theta} \mid \mathbf{Y}, \mathbf{s}, \mathbf{n}) \propto P(\mathbf{Y} \mid \mathbf{g}, \mathbf{m}) P(\mathbf{m}) P(\mathcal{T} \mid \mathbf{s}, \mathbf{n}, N_e(t)) P(N_e(t) \mid \boldsymbol{\theta}) P(\boldsymbol{\theta}), \tag{5.13}$$

for heterochronous data. However, placing the GP-based prior on $N_e(t)$ yields an intractable prior for coalescent times,

$$P(\mathcal{T} \mid N_e(t)) = P(\mathcal{T} \mid \lambda, f(t)) = \prod_{k=2}^{n} \frac{C_k \lambda}{1 + \exp\{-f(t_{k-1})\}} \exp\left[-C_k \int_{t_k}^{t_{k-1}} \frac{\lambda}{1 + \exp\{-f(t)\}} dt \right], \tag{5.14}$$

because the integral in the exponent of equation (5.14) is computationally intractable (similarly for the heterochronous case). Therefore, we perform posterior inference in an augmented data space, as in Chapter 3, with posterior distribution:

$$P(\mathbf{g}, \mathcal{N}, \mathbf{m}, N_e(t), \boldsymbol{\theta} \mid \mathbf{Y}) \propto P(\mathbf{Y} \mid \mathbf{g}, \mathbf{m}) P(\mathbf{m}) P(\mathcal{T}, \mathcal{N} \mid \mathbf{f}_{\mathcal{T},\mathcal{N}}, \lambda) P(\mathbf{f}_{\mathcal{T},\mathcal{N}} \mid \boldsymbol{\theta}) P(\boldsymbol{\theta}), \tag{5.15}$$

where $\mathcal{N}$ is a set of $\mathcal{N}_k = \{t_{k,1}, \ldots, t_{k,m_k}\}$ sets formed by m_k latent points $t_{k,i}$, such that

$$P(t_{k-1}, \mathcal{N}_k \mid t_k, f(t_{k-1}), \mathbf{f}_{\mathcal{N}_k}, \lambda) = (C_k \lambda)^{m_k+1} \times \exp\{-C_k \lambda (t_k - t_{k-1})\} \left[\frac{1}{1 + \exp\{-f(t_{k-1})\}} \right] \prod_{i=1}^{m_k} \left[1 - \frac{1}{1 + \exp\{-f(t_{k,i})\}} \right], \tag{5.16}$$

and

$$P(\mathcal{T}, \mathcal{N} \mid \mathbf{f}_{\mathcal{T},\mathcal{N}}, \lambda) = \prod_{k=2}^{n} P(t_{k-1}, \mathcal{N}_k \mid t_k, f(t_{k-1}), \mathbf{f}_{\mathcal{N}_k}, \lambda). \tag{5.17}$$

Here, equation (5.17) is referred to as the augmented coalescent prior. The augmented model represented in Figure 5.1 allows for tractable inference given that the set of coalescent times $\mathcal{T}$ and the set of (non-coalescent) latent points $\mathcal{N}$ have a tractable joint density as shown in equation (5.16). The derivation of such a density is based on the thinning algorithm for

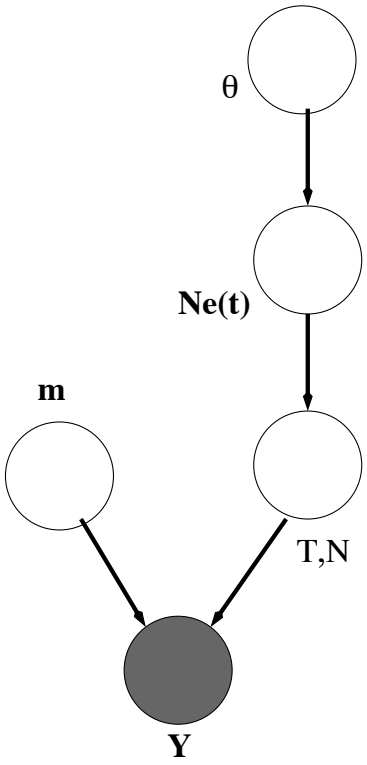

Figure 5.1: Graphical model representation of the augmented model. We assume that sequence data $\mathbf{Y}$ are observed and depend on the substitution process with parameters $\mathbf{m}$ and gene genealogy with coalescent times $\mathcal{T}$. The augmented coalescent process that generates $\mathcal{T}$ and $\mathcal{N}$ jointly depends on $N_e(t)$ and is conditionally independent of the substitution process with parameters m given $\mathbf{Y}$. The effective population size trajectory $N_e(t)$ depends on the Brownian motion precision parameter θ.

non-homogeneous Poisson processes (Lewis and Shedler, 1979) and inferential framework of Adams et al. (2009) for Poisson processes. The derivation of equation (5.16) is detailed in Chapter 3.

Although the fact that sequence data $\mathbf{Y}$ are independent of latent points $\mathcal{N}$ is not evident from the representation in Figure 5.1, our construction implies that $P(\mathbf{Y} \mid \mathbf{g}, \mathbf{m}, \mathcal{N}) = P(\mathbf{Y} \mid \mathbf{g}, \mathbf{m})$.

5.3 MCMC Sampling

We approximate the posterior distribution of model parameters via a MCMC sampling scheme to approximate the posterior given in equation (5.15). Parameters and latent points are sampled in blocks within a random scan Metropolis-within-Gibbs framework.

5.3.1 Latent Points Sampling

We construct a reversible jump algorithm for the number of latent points $m = \sum_{k=2}^{k} m_k$. Since conditional on everything else, the latent points $\mathcal{N}$ only depend on $N_e(t)$ and $\mathcal{T}$, we use the same proposals and acceptance probabilities defined in Chapter 3 for the number of latent points in each coalescent interval. Additionally, we use a Metropolis-Hastings algorithm to update the locations of latent points. We first choose an intercoalescent interval with latent points with probability proportional to its length and then propose point locations uniformly at random in that interval together with their predictive function values $\mathbf{f}_{\mathbf{t}^*} \sim P(\mathbf{f}_{\mathbf{t}^*}|\mathbf{f}_{\mathcal{T},\mathcal{N}}, \theta)$.

5.3.2 Sampling Transformed Effective Population Size Values

As shown in Figure 5.1, the effective population size trajectory is separated from the data $\mathbf{Y}$ by one level in the hierarchical model. This implies that given $\mathcal{T},\mathcal{N}$, N_e is independent of $\mathbf{Y}$. We then can use the same elliptical slice sampling proposal as described in Chapter 3, with the acceptance probabilities remaining the same.

5.3.3 Sampling GP precision parameter

As described in Chapter 3, the full conditional of the precision parameter θ is a Gamma distribution. Therefore, we update θ by drawing from its full conditional $\theta \mid \mathbf{f}_{\mathcal{T},\mathcal{N}}, \mathcal{T}, \mathcal{N} \sim Gamma\,(\alpha^*, \beta^*)$, with

$$\alpha^* = \alpha + \frac{\#\{\mathcal{N} \cup \mathcal{T}\}}{2}, \tag{5.18}$$

and

$$\beta^* = \beta + \frac{\mathbf{f}'_{\mathcal{T},\mathcal{N}}\mathbf{Q}\mathbf{f}_{\mathcal{T},\mathcal{N}}}{2}, \tag{5.19}$$

where $\mathbf{Q} = \frac{1}{\theta}\mathbf{C}^{-1}$ is the inverse covariance matrix.

5.3.4 Tree Sampling

We use a Metropolis-Hastings algorithm to update the genealogical trees. We use the same tree proposal moves defined in `BEAST` (Drummond et al., 2012). The tree proposal moves are *Wilson-Balding move, subtree-exchange, and node age move (uniform).* See (Drummond et al., 2002) for details. Since a change in topology alone does not affect the definition of coalescent and latent points, we only define new acceptance probabilities for moves that involve a change in coalescent times $\mathcal{T}$. Given a new coalescent time t^*, we sample $f(t^*) \sim P(f(t^*) \mid \mathbf{f}_{\mathcal{T},\mathcal{N}}, \theta)$ to obtain the new value of $N_e(t^*)$.

If the height of the tree remains constant, the definition of the latent points involved in the adjacent inter-coalescent intervals and the number of latent points in those inter-coalescent intervals change. Figure 5.2 (a) shows an example of a possible proposed change in coalescent times. In the top left plot, there are no latent points between t_3 and t_2, however, after t_3 is replaced by t_3^* in the bottom left plot, there are now 2 latent points between t_3^* and t_2. The locations of the latent points remain the same, but their labels and contribution to the augmented coalescent prior change. Let $q(\mathcal{T}^* \mid \mathcal{T})$ be the proposal distribution of a new set of coalescent times $\mathcal{T}^*$ given the current state $\mathcal{T}$ and let

$$H_1 = \frac{q(\mathcal{T} \mid \mathcal{T}^*)}{q(\mathcal{T}^* \mid \mathcal{T})} \tag{5.20}$$

be the ratio of tree backward and forward proposal densities. The specific expressions of H_1 can be obtained from Drummond et al. (2002). Let

$$H_2 = \frac{P(\mathbf{f}_{\mathcal{T}} \mid \mathbf{f}_{\mathcal{T}^*,\mathcal{N}}, \theta)}{P(\mathbf{f}_{\mathcal{T}^*} \mid \mathbf{f}_{\mathcal{T},\mathcal{N}}, \theta)} \tag{5.21}$$

be the ratio of backward and forward proposal densities of effective population sizes. Then, when the tree height remains unchanged, the joint acceptance probability of accepting the new genealogy, latent points and appropriate values of the effective population size is:

$$a_{\mathbf{g}} = min\left\{\frac{P(\mathbf{Y} \mid \mathbf{g}^*, \mathbf{m})}{P(\mathbf{Y} \mid \mathbf{g}, \mathbf{m})} \frac{P(\mathcal{T}^*, \mathcal{N}^* \mid \mathbf{f}_{\mathcal{T}^*,\mathcal{N}^*}, \lambda) P(\mathbf{f}_{\mathcal{T}^*,\mathcal{N}^*} \mid \boldsymbol{\theta})}{P(\mathcal{T}, \mathcal{N} \mid \mathbf{f}_{\mathcal{T},\mathcal{N}}, \lambda) P(\mathbf{f}_{\mathcal{T},\mathcal{N}} \mid \boldsymbol{\theta})} H_1 H_2, 1\right\} \tag{5.22}$$

When the tree height is changed we additionally propose new locations for the latent points that are affected by the change of coalescent times. The new location for each

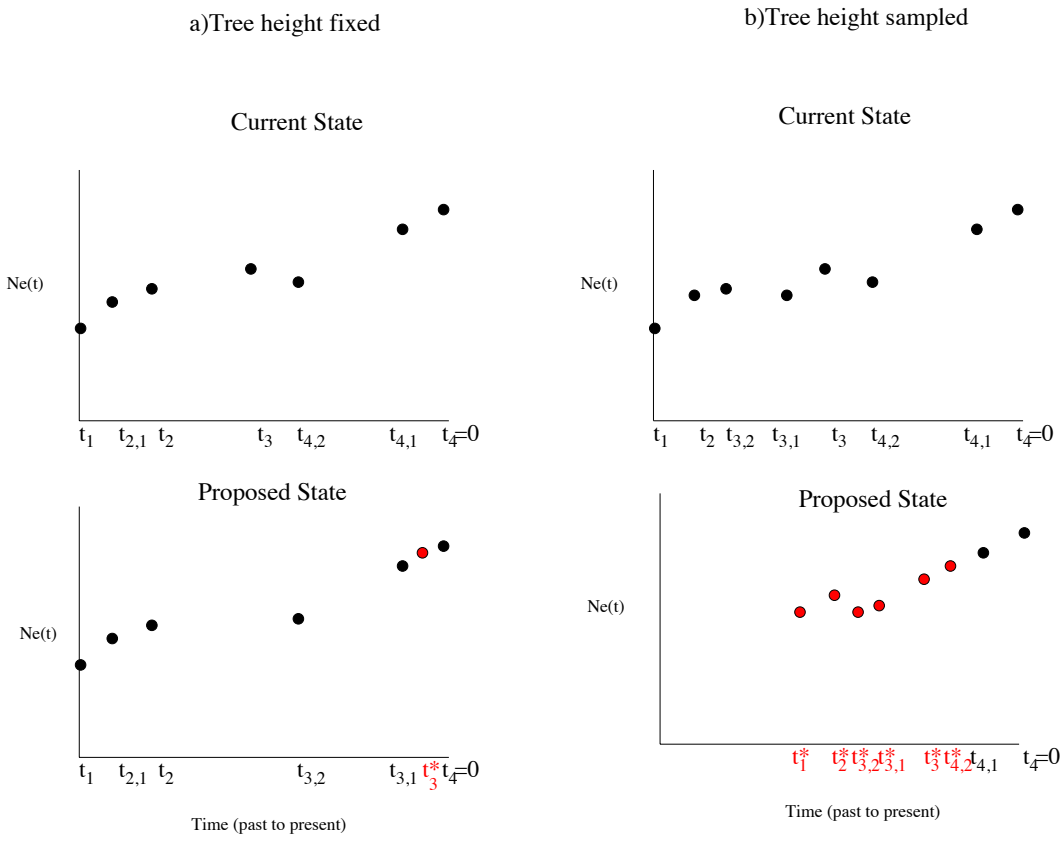

Figure 5.2: (a) Tree height fixed. The coalescent time t_3 is replaced by t_3^*; the labels of the latent points $t_{4,2}$ and $t_{4,1}$ change to $t_{3,2}$ and $t_{3,1}$, but not their values. Given the new coalescent time t_3^*, a new value $N_e(t_3^*)$ (red circle) replaces $N_e(t_3)$. (b) Tree height sampled. All the coalescent times are sampled except the sampling time t_4. In this case, all the latent points that change definition after the tree move are sampled within the new intercoalescent interval. For example, $t_{4,2}$ is replaced by $t_{4,2}^* \sim U(t_3^*, t_4^*)$, while $t_{4,1}$ remains in the same location. Given a new location t^*, its corresponding $N_e(t^*)$ is also sampled (red circles).

affected latent point is sampled uniformly at random in the new intercoalescent interval. For example, $t_{3,2}$ in the top right plot in Figure 5.2 is replaced by $t^*_{3,2} \sim U(t^*_2, t^*_3)$ in the right bottom plot in Figure 5.2, while the location of $t_{4,1}$ remains unchanged. For each new latent or coalescent time location t^*, we sample $f(t^*) \sim P[f(t^*)|\mathbf{f}_{\mathcal{T},\mathcal{N}}, \theta]$ to obtain the new value of $N_e(t^*)$. The ratio of backward and forward proposal densities is:

$$H_3 = \frac{P(\mathbf{f}_{\mathcal{T},\mathcal{N}}|\mathbf{f}_{\mathcal{T}^*,\mathcal{N}^*}, \theta)}{P(\mathbf{f}_{\mathcal{T}^*,\mathcal{N}^*}|\mathbf{f}_{\mathcal{T},\mathcal{N}}, \theta)} \prod_{k=2}^{n} \left[\frac{t^*_{k-1} - t^*_k}{t_{k-1} - t_k}\right]^{m^*_k}, \tag{5.23}$$

where $m^*_k \leq m_k$ is the number of latent point moves between t_{k-1} and t_k. The acceptance probability is then defined by equation (5.22) with a replacement of H_2 by H_3.

5.3.5 Sampling Substitution Parameters

Most of the acceptance probabilities for sampling the parameters $\mathbf{m}$ remain the same as the current implementation of `BEAST` (Drummond et al., 2012), however, there is a scaling move that proposes a new set of re-scaled coalescent times $\mathcal{T}^*$ and $\mu^* \in \mathbf{m}^*$ jointly. In this case, given a new set of coalescent times $\mathcal{T}^*$, we use the same proposals as when the height of tree is changed. The corresponding acceptance probability is then

$$a_{\mathbf{g}} = min\left\{\frac{P(\mathbf{Y} \mid \mathbf{g}^*, \mathbf{m})P(\mathbf{m}^*)}{P(\mathbf{Y} \mid \mathbf{g}, \mathbf{m})P(\mathbf{m}^*)} \frac{P(\mathcal{T}^*, \mathcal{N}^* \mid \mathbf{f}_{\mathcal{T}^*,\mathcal{N}^*}, \lambda)P(\mathbf{f}_{\mathcal{T}^*,\mathcal{N}^*} \mid \boldsymbol{\theta})}{P(\mathcal{T}, \mathcal{N} \mid \mathbf{f}_{\mathcal{T},\mathcal{N}}, \lambda)P(\mathbf{f}_{\mathcal{T},\mathcal{N}} \mid \boldsymbol{\theta})} H_4 H_3, 1\right\} \tag{5.24}$$

where

$$H_4 = \frac{q(\mathcal{T}, \mathbf{m} \mid \mathcal{T}^*, \mathbf{m}^*)}{q(\mathcal{T}^*, \mathbf{m}^* \mid \mathcal{T}, \mathbf{m})} \tag{5.25}$$

is the ratio for the scaling move proposal densities.

5.3.6 Egyptian HCV

We analyze 63 HCV sequences gathered in 1993 in Egypt. We reanalized the data using the Bayesian skyride (Minin et al., 2008) as well as our GP model. For both cases, we assumed the HKY CTMC mutation model (Hasegawa M, 1985) using the `BEAST` (Drummond et al., 2012) implementation of both methods. The Bayesian Skyride method was run for 50 million iterations with the first 5000 discarded as burn-in. Genealogies and model parameters were sampled every 5000 iterations. The effective population size trajectory estimated using Bayesian Skyride is a piece-wise linear trajectory of estimated effective population sizes at

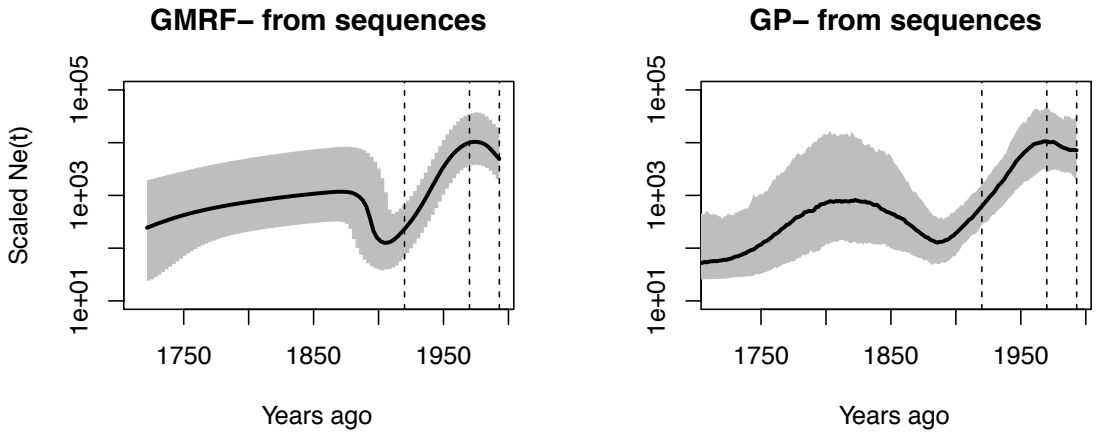

Figure 5.3: Egyptian HCV. Log of scaled effective population trajectory estimated using the GMRF method and our GP method. The posterior medians are represented by solid black lines and the 95% BCI's are represented by the gray shaded areas. The vertical dashed lines mark the years 1920 (the start of intravenous PAT) , 1970 (the end of intravenous PAT) and 1993 (sampling time of sequences).

a grid of 62 points. Our method was run for 550,000 iterations with the first 2000 discarded as burn-in and thinned every 1000 iterations. We estimated effective population sizes at a regular grid of 150 points. Log-likelihood trace plot and effective sample sizes at the grid of 150 points are displayed in Figure 5.4. Figure 5.3 shows the recovered effective population size trajectories, with time scaled in units of years, using both methods. In both cases, we recover the previously found pattern: An exponential growth phase starting around the 1920s and a decay around the 1970s (Palacios and Minin, 2013). The GP recovered trajectory agrees with previous result in Chapter 3 with a fixed genealogy while the GMRF results disagrees in having a local maximum around 1870. This result might be an artifact of the artificial discretization nature of the GMRF method.

5.4 Conclusions

Bayesian estimation of population size trajectories from gene sequence data usually involves two levels of hierarchical modeling. Given gene sequence data and a substitution model, one

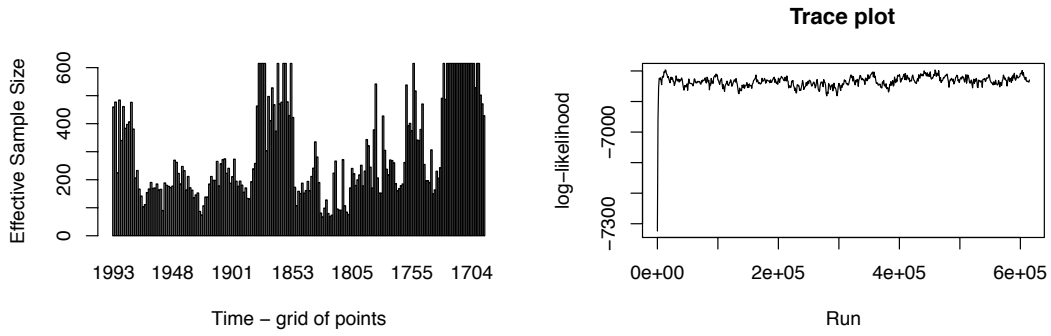

Figure 5.4: Effective sample sizes of $N_e(t)$ evaluated at a grid of points (left plot) and trace plot of loglikehood (right plot) for the recovered hcv effective population size trajectory using the GP method.

can estimate a genealogical tree. Given a genealogical tree and the coalescent model, one can estimate the effective population size. In order to add as much flexibility as possible in terms of the shape of the population trajectory, a GP-based Bayesian nonparametric method has been recently developed for a fixed genealogy. Here, we extend the GP method to sample genealogies and estimate population trajectories directly from molecular sequence data.

The GP-based method uses a transformed Gaussian process prior on population size trajectories and importantly, this prior is independent of the genealogy in contrast to Bayesian Skyride (Minin et al., 2008). This flexibility allows for easy extension to incorporate multiple loci data.

We tested our method on real data and compared it with the competing GMRF method. Our method allows to estimate population size trajectories further back in time and to better assess uncertainty. Further, our GP-based Bayesian nonparametric model can be extended to incorporate other temporal processes correlated to effective population sizes into the framework.

Chapter 6

DISCUSSION AND FUTURE DIRECTIONS

6.1 Summary of Contributions

The effective population size is one of the most important parameters in evolutionary genetics. The introduction of the coalescent model on genealogies (Kingman, 1982) allowed for estimation of effective population size trajectories from genetic data. In this thesis, I develop a series of Gaussian process-based Bayesian methods to infer effective population size trajectories over time. My work has been motivated by applications in public health, in particular, in infectious diseases of rapid evolution such as human influenza virus and hepatis C, as well as by analysis of ancient DNA.

There has been a recent increase in interest in models for estimation and prediction of disease dynamics that are able to combine molecular data with other sources of information (Kühnert et al., 2011). With this goal in mind, I focus my dissertation on developing a flexible model that would allow, in principle, for integration of temporal data pertinent to the population in question and molecular data. By placing a transformed Gaussian process prior on the effective population size trajectory, I explicitlly model the effective population size as a continuously varying stochastic process as opposed to the other methods developed for phylodynamics. The advantage of modeling population dynamics continuously is not only because such modeling is more intuitive, but also because it would allow for correlating population dynamics with other temporal processes observed at certain times.

In Chapter 2, I discuss modern Bayesian nonparametric methods for phylodynamics. I show that the coalescent process on genealogies is a point process comprised of coalescent times and that estimation of effective population size trajectories is equivalent to the problem of estimation of the intensity function of this point process. This connection is crucial for developing the methods of Chapters 3, 4 and 5. In Chapter 3, I develop a GP-based

method to infer population size trajectories from a fixed genealogy. The challenge of this method is to overcome intractability of the likelihood function. I demonstrate how recent advances in GP-based nonparametric inference for Poisson processes can be extended to solve intractability of the coalescent likelihood function when the effective population size varies stochastically. While developing the inferential framework, I introduce new algorithms for simulation of genealogies for stochastically varying effective population sizes. I successfully recover the effective population size trajectories from simulated and real data. In Chapter 4, I introduce an integrated nested Laplace approximation to the method of Chapter 3 and a popular GMRF method for estimation of effective population size trajectories from gene genealogies. This approximation replaces MCMC entirely and produces accurate approximations in a prompt manner. My method implemented in `R` has already been applied to infer dynamics of the number of infected individuals on simulated data (Frost and Volz, 2013). However, a fixed genealogy is rarely available and we need to infer population size trajectories from genomic data directly. The INLA approach approximates posterior marginals of a particular class of models called latent Gaussian models. When genealogies are unknow, the posterior distribution of substitution parameters, genealogy, and effective population size trajectory no longer fall into the latent Gaussian models class. I then developed an MCMC approach to infer effective population size trajectories from sequence data direclty. In Chapter 5, I extended the GP-based method of Chapter 3 to infer population size trajectories and other parameters from genomic data. This method has been successfully implemented in BEAST (Drummond et al., 2012).

The coalescent prior on genealogies used in all the models described here assumes a random sample of orthologous, nonrecombining and neutrally evolving sequences from a panmictic population. Any violation of these assumptions will result in an estimated effective population size trajectory that is not directly proportional to census population size. *Selection* effectively shifts the distribution of mutations on the genealogy; therefore, interpretation of effective population size estimation under selection needs to be done with caution (Pybus and Rambaut, 2009; Ho and Shapiro, 2011). Recombination is clearly a problem for all the methods described so far, since these methods assume a single genealogy. In

principle, it should be possible to extend Bayesian nonparametric phylodynamic methods to include a possibility of recombination similarly to the work of Kuhner and Smith (2007), but software development and computational costs will be significant. Alternatively, one can use a sequential Markov approximation of the coalescent process (McVean and Cardin, 2005), as was done in a recent attempt to leverage whole-genome sequence data of a single individual to estimate population size dynamics (Li and Durbin, 2011). Although the inferential framework in this case is very different from the one described in this chapter, Bayesian nonparametric approaches similar to those detailed here, can also be applied in this setting. In the rest of this section we discuss further possible extensions of the currently available phylodynamic methods.

6.1.1 Multiple Loci

Data from multiple unlinked genetic loci are rapidly becoming the norm in the era of next-generation sequencing. Evolutionary dynamics of such independently evolving loci are governed by the same demographic history of the population under study, enabling straightforward estimation of effective population size trajectories based on multilocus genetic data. Increasing the number of loci improves precision of the phylodynamic estimation, which is critical for these nonparametric procedures that often suffer from large BCIs. One of the primary difficulties in estimating population dynamics is that most of the coalescent events in the reconstructed genealogy usually occur in a short time span. During the long periods of time in which few coalescent events occur, there is not much data to infer the population dynamics. Increasing the sample size mitigates this problem to a certain extent, but the additional coalescent events also tend to occur in a small stretch of time. It is more advantageous to increase the number of loci since this provides extra information during the long time frames with few coalescent events (Heled and Drummond, 2008).

To allow for inference of effective population size trajectories from multiple loci, Heled and Drummond (2008) implemented the *extended Bayesian skyline plot* in `BEAST` (Drummond et al., 2012). This new version of the Bayesian skyline places a different prior on $N_e(t)$. Instead of a piece-wise constant function with jumps at some coalescent points, the estimated trajectory is piece-wise linear with straight lines connecting "heights" γ_j at

change-points $a_1, ..., a_k$, that is,

$$N_e(t) = \sum_{j=1}^{k+1} \left(\gamma_{j-1} + (\gamma_j - \gamma_{j-1}) \frac{t - a_{j-1}}{a_j - a_{j-1}} \right) 1_{(a_{j-1}, a_j]}(t). \tag{6.1}$$

The change-points are an ordered subset of the coalescent times $\{t_n, \ldots, t_1\}$, which now include coalescent times from genealogies at all loci, and $a_0 = t_n = 0$ and $a_{k+1} = t_1$. Here, *a priori*

$$\gamma_j \sim \text{Exponential}(\theta), \text{ for } j = 1, \ldots, k+1, \tag{6.2}$$

where θ may be either fixed or have a prior $P(\theta)$. The number of change-points k has a truncated Poisson distribution with mean $ln(2)$.

Gill et al. (2013) have developed a model, called the *Bayesian skygrid* that generalizes and improves the *Bayesian skyride*. It differs from the *Bayesian skyride* not only in its ability to incorporate data from several loci, but also in that the estimated piecewise constant trajectory has change-points at fixed user-specified times rather than coalescent times. Given ordered times $s_B < \cdots < s_1$, with $s_B = 0$ and $s_0 = \infty$, we have

$$N_e(t) = \sum_{k=1}^{B} \exp(\gamma_k) \, 1_{[s_k, s_{k-1})}(t), \tag{6.3}$$

where $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_B)$ is *a priori* a Gaussian Markov random field.

To illustrate the benefits of estimating population dynamics from multilocus data, we simulate genealogies under the same demographic scenario as in Example 1.4. We estimate the effective population size from 1, 2, 5, and 10 genealogies using the *Bayesian skygrid* with $n = 99$ and $s_{100}, \ldots, s_1$ equally spaced times between 0 and 2.5. Figure 6.1 demonstrates that increasing the number of loci even modestly leads to appreciable gain of estimation precision.

6.1.2 Effect of Population Structure

The coalescent with variable population size assumes that there is random mixing in the population and that the samples are taken randomly from the population. This former assumption is clearly violated for many real populations, because individuals tend to mate with other individuals in geographic or social proximity. In presence of a well defined population structure, a simple random sample of the whole population will not efficiently capture

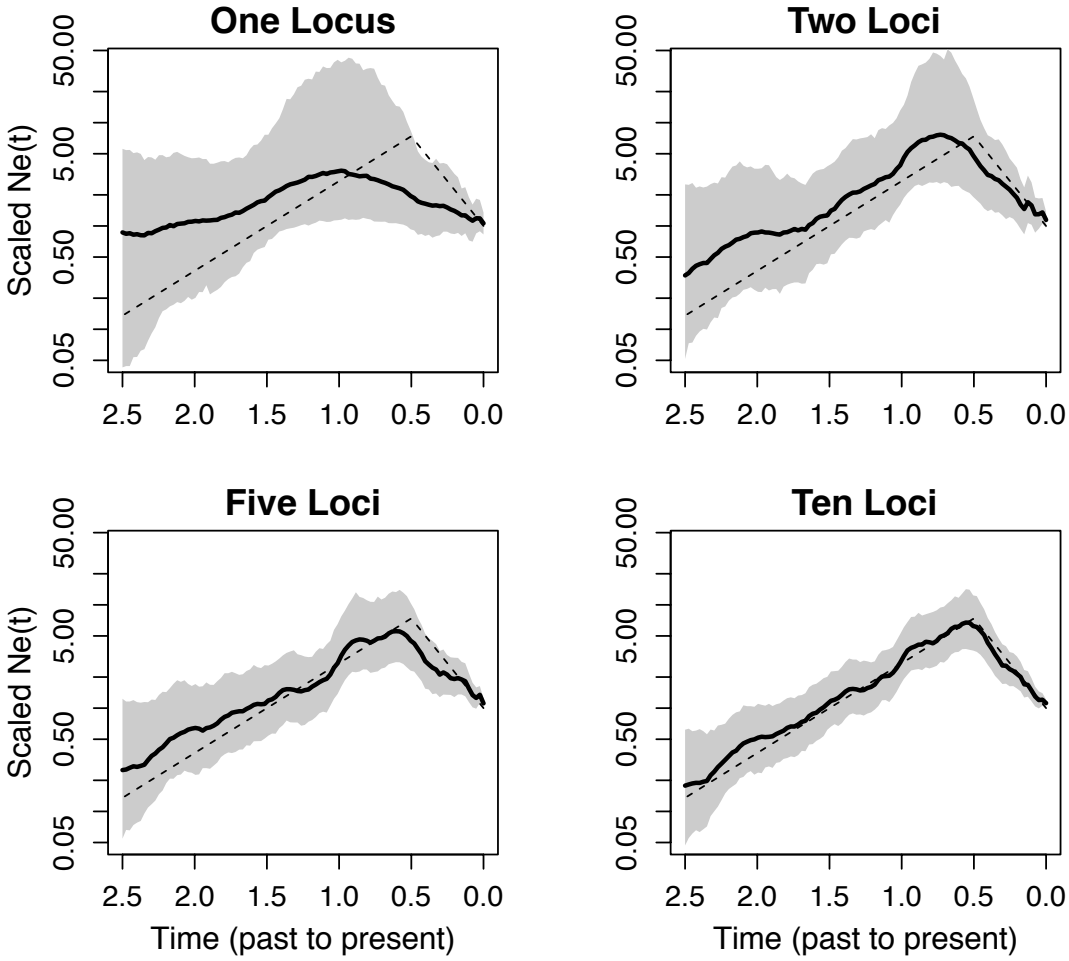

Figure 6.1: The log effective population size trajectories estimated under the *Bayesian skygrid* from 1, 2, 5, and 10 simulated genealogies.

the diversity of each subpopulation and the coalescent with variable population size applied directly to the whole sample will not account for the "blocking/clustering" of sampled lineages in the genealogy. Instead, a better strategy would be to consider gathering a stratified sample by subpopulation and estimate the total population size trajectory under the structured coalescent prior on genealogies (Hudson, 1990; Beerli and Felsenstein, 2001). The structured coalescent accounts for subpopulations with different population sizes and allows for migration between subpopulations. The structured coalescent has been used for parametric estimation of population size trajectories and migration parameters from isochronous and heterochronous data in Bayesian and maximum likelihood frameworks (Ewing and Rodrigo, 2006a,b; Beerli and Felsenstein, 2001). To our knowledge, no implementation of the structured coalescent equipped with Bayesian nonparametric estimation of subpopulation size trajectories is available.

A more complex violation of random mixing occurs when individuals in the population are a connected by a social or contact network (Welch et al., 2011). The standard structured coalescent assumes random mating within subpopulations and constant variability in number of offspring, making this modeling framework incapable of handling network-based population structure. In order to account for contact heterogeneity and hence, a variable reproductive variance, attempts to equip the coalescent with social network modeling are emerging. If we are interested in a population of infected individuals with a certain rapidly evolving infectious disease, knowing the social or contact network of the sampled individuals should help in reducing uncertainty in the genealogical reconstruction (transmission network) among the samples. However, knowledge about the social network and, moreover, knowledge of the dynamics of the social network are not readily available to us in practice. Network-based coalescent presents further challenges in formulating a model for the social network of the sample and the population size affect the shape of the genealogy in a statistical (likelihood-based) framework. In light of these difficulties, it is not surprising that so far progress on merging the coalescent and network-based approaches advanced mainly through simulations with an aim to measure the impact of network structure on the estimation of effective population sizes (Goodreau, 2006; O'Dea and Wilke, 2010; Leventhal et al., 2012).

6.1.3 Coalescent and Infectious Disease Dynamics

Phylodynamic methods have been widely applied to study the evolution of rapidly evolving diseases. Here, inference is based on sampled disease agent molecular sequences from infected hosts. If superinfection is rare and mutations accumulate fast relative to epidemic growth, each coalescent time in a genealogy of sampled consensus viral isolates from infected individuals corresponds to a transmission event (Volz et al., 2009). Estimation of effective population size under the coalescent with variable population size prior on genealogies assumes that generation length and variability in number of transmissions are constant through time. For some pathogens, this may be unrealistic and interpretability of the effective population size as the number of infections becomes imprecise. In order to gain interpretability of epidemiologically relevant parameters, there has been a growing interest in formalizing the integration of phylodynamic methods and standard epidemiological and ecological models (Grenfell et al., 2004; Pybus and Rambaut, 2009; Volz et al., 2009; Bennett et al., 2010b; Frost and Volz, 2010; Kühnert et al., 2011). In the most recent effort on this front, Volz (2012) considered a population under a continuous time birth-death process with varying birth-death rates and expressed the coalescence rate in terms of the birth rate (incidence) and the number of infected individuals (prevalence). Further, the dynamic population model was extended to include migration and two stages of infection to accommodate cases where the transmission probability per contact changes over the course of infection. In addition to incorporating disease dynamics into the coalescent, there is a growing need to integrate molecular data with clinical, socio-demographic, and other relevant data in order to measure the correlation between the population size and the environment (Rasmussen et al., 2011). We hope that more sophisticated Bayesian modeling will be able to solve these challenging problems.

BIBLIOGRAPHY

Adams, R., Murray, I., and MacKay, D. (2009). Tractable nonparametric Bayesian inference in Poisson processes with Gaussian process intensities. In *Proceedings of the 26th International Conference on Machine Learning*, pages 9–16.

Andersen, P. K., Borgan, O., Gill, R. D., and Keiding, N. (1995). *Statistical Models Based on Counting Processes*. Springer, 2nd edition.

Arjas, E. and Heikkinen, J. (1997). An algorithm for nonparametric Bayesian estimation of a Poisson intensity. *Computational Statistics*, 12:385–402.

Beerli, P. and Felsenstein, J. (2001). Maximum likelihood estimation of a migration matrix and effective population sizes in n subpopulations by using a coalescent approach. *Proceedings of the National Academy of Sciences*, 98(8):4563–4568.

Bennett, S., Drummond, A., Kapan, D., Suchard, M., Muñoz-Jordán, J., Pybus, O., Holmes, E., and Gubler, D. (2010a). Epidemic dynamics revealed in dengue evolution. *Molecular Biology and Evolution*, 27(4):811–818.

Bennett, S. N., Drummond, A. J., Kapan, D. D., Suchard, M. A., Muñoz Jordán, J. L., Pybus, O. G., Holmes, E. C., and Gubler, D. J. (2010b). Epidemic dynamics revealed in dengue evolution. *Molecular Biology and Evolution*, 27:811–818.

Bishop, C. (2006). *Pattern Recognition and Machine Learning*. Springer New York.

Brillinger, D. R. (1979). Analyzing point processes subjected to random deletions. *Canadian Journal of Statistics*, 7(1):21–27.

Campos, P. F., Willerslev, E., Sher, A., Orlando, L., Axelsson, E., Tikhonov, A., Aaris-Sørensen, K., Greenwood, A. D., Kahlke, R., Kosintsev, P., Krakhmalnaya, T., Kuznetsova, T., Lemey, P., MacPhee, R., Norris, C. A., Shepherd, K., Suchard, M. A.,

Zazula, G. D., Shapiro, B., and Gilbert, M. T. P. (2010a). Ancient DNA analyses exclude humans as the driving force behind late pleistocene musk ox (*Ovibos moschatus*) population dynamics. *Proceedings of the National Academy of Sciences*, 107(12):5675–5680.

Campos, P. F., Willerslev, E., Sher, A., Orlando, L., Axelsson, E., Tikhonov, A., Aaris Sørensen, K., Greenwood, A. D., Ralf-Dietrich Kahlke, Kosintsev, P., Krakhmalnaya, T., Kuznetsova, T., Lemey, P., MacPhee, R., Norris, C. A., Shepherd, K., and Suchard, M. A. (2010b). Ancient DNA analyses exclude humans as the driving force behind late Pleistocene usk ox (*Ovibos moschatus*) population dynamics. *Proceedings of the National Academy of Sciences*, 107(12):5675–5680.

Cressie, N. (1993). *Statistics for Spatial Data*. Wiley-Interscience, New York, USA.

Cseke, B. and Heskes, T. (2010). Improving posterior marginal approximations in latent Gaussian models. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, volume 9, pages 121–128.

Daley, D. and Vere-Jones, D. (2002). *An Introduction to the Theory of Point Processes*, volume 1. Springer, 2nd edition.

Diggle, P. (1985). A kernel method for smoothing point process data. *Journal of the Royal Statistical Society, Series C (Applied Statistics)*, 34(2):138–147.

Drummond, A. and Rambaut, A. (2007). BEAST: Bayesian Evolutionary Analysis by Sampling Trees. *BMC Evolutionary Biology*, 7(1):214.

Drummond, A., Suchard, M., Xie, D., and Rambaut, A. (2012). Bayesian phylogenetics with BEAUti and the BEAST 1.7. *Molecular Biology and Evolution*, 29:1969–1973.

Drummond, A. J., Nicholls, G. K., Rodrigo, A. G., and Solomon, W. (2002). Estimating mutation parameters, population history and genealogy simultaneously from temporally spaced sequence data. *Genetics*, 161(3):1307–1320.

Drummond, A. J., Rambaut, A., Shapiro, B., and Pybus, O. G. (2005). Bayesian coalescent

inference of past population dynamics from molecular sequences. *Molecular Biology and Evolution*, 22(5):1185–1192.

Edgar, R. C. (2004). MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, 32(5):1792–1797.

Ewing, G. and Rodrigo, A. (2006a). Coalescent-based estimation of population parameters when the number of demes changes over time. *Molecular Biology and Evolution*, 23:988–996.

Ewing, G. and Rodrigo, A. (2006b). Estimating population parameters using the structured serial coalescent with Bayesian MCMC inference when some demes are hidden. *Evolutionary Bioinformatics*, 2:227–235.

Felsenstein, J. (1981). Evolutionary trees from DNA sequences: a maximum likelihood approach. *Journal of Molecular Evolution*, 13:93–104.

Felsenstein, J. (1992). Estimating effective population size from samples of sequences: inefficiency of pairwise and segregating sites as compared to phylogenetic estimates. *Genetical Research*, 59(2):139–147.

Felsenstein, J. and Rodrigo, A. G. (1999). Coalescent Approaches to HIV Population Genetics. In *The Evolution of HIV*, pages 233–272. Johns Hopkins University Press.

Frank, C., Mohamed, M., Strickland, G., Lavanchy, D., Arthur, R., Magder, L., Khoby, T., Abdel-Wahab, Y., Ohn, E., Anwar, W., and Sallam, I. (2000). The role of parenteral antischistosomal therapy in the spread of hepatitis C virus in Egypt. *Lancet*, 355:887–891.

Fraser, C., Donnelly, C. A., Cauchemez, S., Hanage, W. P., Van Kerkhove, M. D., Hollingsworth, T. D., Griffin, J., Baggaley, R. F., Jenkins, H. E., Lyons, E. J., Jombart, T., Hinsley, W. R., Grassly, N. C., Balloux, F., Ghani, A. C., Ferguson, N. M., Rambaut, A., Pybus, O. G., Lopez-Gatell, H., Alpuche-Aranda, C. M., Chapela, I. B., Zavala, E. P., Guevara, D. M. E., Checchi, F., Garcia, E., Hugonnet, S., Roth, C., and Collaboration, T. W. R. P. A. (2009). Pandemic potential of a strain of influenza A (H1N1): Early findings. *Science*, 324(5934):1557–1561.

Frost, S. D. W. and Volz, E. M. (2010). Viral phylodynamics and the search for an 'effective number of infections'. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 365(1548):1879–1890.

Frost, S. D. W. and Volz, E. M. (2013). Modelling tree shape and structure in viral phylodynamics. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 368(1614).

Fu, Y. (1994). Estimating effective population size or mutation rate using the frequencies of mutations of various classes in a sample of DNA sequences. *Genetics*, 138(4):1375–1386.

Gilks, W. R. and Wild, P. (1992). Adaptive rejection sampling for Gibbs sampling. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 41(2):pp. 337–348.

Gill, M., Lemey, P., Faria, N., Rambaut, A., Shapiro, B., and Suchard, M. (2013). Improving bayesian population dynamics inference: A coalescent-based model for multiple loci. *Molecular Biology and Evolution*, 30:713–724.

Goodreau, S. M. (2006). Assessing the effects of human mixing patterns on human immunodeficiency virus-1 interhost phylogenetics through social network simulation. *Genetics*, 172:2033–2045.

Green, P. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, 82:711–732.

Grenfell, B. T., Pybus, O. G., Gog, J. R., Wood, J. L. N., Daly, J. M., Mumford, J. A., and Holmes, E. C. (2004). Unifying the epidemiological and evolutionary dynamics of pathogens. *Science*, 303(5656):327–332.

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

Griffiths, R. C. and Tavaré, S. (1994). Sampling theory for neutral alleles in a varying envi-

ronment. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, 344(1310):403–410.

Hasegawa M, Kishino H, Y. T. (1985). Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *Journal of Molecular Evolution*, 2:160–164.

Hein, J., Schierup, M. H., and Wiuf, C. (2005). *Gene Genealogies, Variation and Evolution: A Primer in Coalescent Theory*. Oxford University Press, USA, 1st edition.

Heled, J. and Drummond, A. (2008). Bayesian inference of population size history from multiple loci. *BMC Evolutionary Biology*, 8(1):1–289.

Ho, S. Y. W. and Shapiro, B. (2011). Skyline-plot methods for estimating demographic history from nucleotide sequences. *Molecular Ecology Resources*, 11(3):423–434.

Hudson, R. R. (1990). Gene genealogies and the coalescent process. *Oxford Surveys in Evolutionary Biology*, 7:1–44.

Illian, J., Sorbye, S. H., and Rue, H. (2012). A toolbox for fitting complex spatial point process models using integrated nested Laplace approximation (INLA). *Annals of Applied Statistics*, 1(2).

Influenza Genome Sequencing Project (2011). `http://www.niaid.nih.gov/labsandresources/resources/dmid/gsc/influenza/Pages/default.aspx`.

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

Kottas, A. and Sansó, B. (2007). Bayesian mixture modeling for spatial Poisson process intensities, with applications to extreme value analysis. *Journal of Statistical Planning and Inference*, 137:3151–3163.

Kuhner, M. and Smith, L. (2007). Comparing likelihood and Bayesian coalescent estimation of population parameters. *Genetics*, 175(1):155–165.

Kuhner, M., Yamato, J., and Felsenstein, J. (1995). Estimating effective population size and mutation rate from sequence data using Metropolis-Hastings sampling. *Genetics*, 140(4):1421–1430.

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

Kühnert, D., Wu, C.-H., and Drummond, A. J. (2011). Phylogenetic and epidemic modeling of rapidly evolving infectious diseases. *Infection, Genetics and Evolution*, 11(8):1825–1841.

Leventhal, G. E., Kouyos, R., Stadler, T., von Wyl, V., Yerly, S., Böni, J., Cellerai, C., Klimkait, T., Günthard, H. F., and Bonhoeffer, S. (2012). Inferring epidemic contact structure from phylogenetic trees. *PLoS Computational Biology*, 8(3):1–10.

Lewis, P. and Shedler, G. (1979). Simulation of nonhomogeneous Poisson processes by thinning. *Naval Research Logistics Quarterly*, 26(3):403–413.

Li, H. and Durbin, R. (2011). Inference of human population history from individual whole-genome sequences. *Nature*, 475(7357):493–496.

Lindgren, F. and Rue, H. (2008). On the second-order random walk model for irregular locations. *Scandinavian Journal of Statistics*, 35(4):691–700.

McVean, G. and Cardin, N. (2005). Approximating the coalescent with recombination. *Philos Trans R Soc Lond B Biol Sci*, 360(1459):1387–1393.

Miller, F. D. and Abu-Raddad, L. J. (2010). Evidence of intense ongoing endemic transmission of hepatitis C virus in Egypt. *Proceedings of the National Academy of Sciences, USA*, 107(33):14757–14762.

Minin, V. N., Bloomquist, E. W., and Suchard, M. A. (2008). Smooth skyride through a rough skyline: Bayesian coalescent-based inference of population dynamics. *Molecular Biology and Evolution*, 25(7):1459–1471.

Møller, J., Syversveen, A. R., and Waagepetersen, R. P. (1998). Log Gaussian Cox processes. *Scandinavian Journal of Statistics*, 25(3):451–482.

Murray, I., Adams, R. P., and MacKay, D. J. (2010). Elliptical slice sampling. In *Proceedings of the 13th International Conference on Artificial Intelligence and Statistics*, volume 9, pages 541–548.

Murray, I., Ghahramani, Z., and MacKay, D. (2006). MCMC for doubly-intractable distributions. In *Proceedings of the 22nd Annual Conference on Uncertainty in Artificial Intelligence*, pages 359–366.

Nordborg, M. (2001). Coalescent theory. In *Handbook of Statistical Genetics*, chapter Coalescent Theory, pages 179–212. Chichester, U.K., john wiley & sons edition.

O'Dea, E. B. and Wilke, C. O. (2010). Contact heterogeneity and phylodynamics: How contact networks shape parasite evolutionary trees. *Interdisciplinary Perspectives on Infectious Diseases*, 2011.

Ogata, Y. (1981). On Lewis' simulation method for point processes. *IEEE Transactions on Information Theory*, 27(1):23–31.

Opgen-Rhein, R., Fahrmeir, L., and Strimmer, K. (2005). Inference of demographic history from genealogical trees using reversible jump Markov chain Monte Carlo. *BMC Evolutionary Biology*, 5:1–6.

Palacios, J. A. and Minin, V. N. (2013). Gaussian process-based Bayesian nonparametric inference of population trajectories from gene genealogies. *Biometrics*, 63:8–18.

Papangelou, F. (1972). Integrability of expected increments of point processes and a related random change of scale. *Transactions of the American Mathematical Society*, 165:483–506.

Paradis, E., Claude, J., and Strimmer, K. (2004). APE: analyses of phylogenetics and evolution in R language. *Bioinformatics*, 20:289–290.

Pybus, O. G., Drummond, A. J., Nakano, T., Robertson, B. H., and Rambaut, A. (2003). The epidemiology and iatrogenic transmission of hepatitis C virus in Egypt: A Bayesian coalescent approach. *Molecular Biology and Evolution*, 20(3):381–387.

Pybus, O. G. and Rambaut, A. (2009). Evolutionary analysis of the dynamics of viral infectious disease. *Nature Reviews Genetics*, 10:540–550.

Pybus, O. G., Rambaut, A., and Harvey, P. H. (2000). An Integrated Framework for the Inference of Viral Population History From Reconstructed Genealogies. *Genetics*, 155(3):1429–1437.

Raftery, A. and Akman, V. (1986). Bayesian analysis of a Poisson process with a change-point. *Biometrika*, 73:85–89.

Rambaut, A. (2000). Estimating the rate of molecular evolution: incorporating non-contemporaneous sequences into maximum likelihood phylogenies. *Bioinformatics*, 16(4):395–399.

Rambaut, A., Pybus, O. G., Nelson, M. I., Viboud, C., Taubenberger, J. K., and Holmes, E. C. (2008). The genomic and epidemiological dynamics of human influenza A virus. *Nature*, 453(7195):615–619.

Rasmussen, C. E. and Williams, C. K. I. (2006). *Gaussian Processes for Machine Learning*. The MIT Press, Cambridge, MA.

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

Ray, S. C., Arthur, R. R., Carella, A., Bukh, J., and Thomas, D. L. (2000). Genetic epidemiology of hepatitis C virus throughout Egypt. *The Journal of Infectious Diseases*, 182(3):pp. 698–707.

Ronquist, F., Teslenko, M., van der Mark, P., Ayres, D., Darling, A., Höhna, S., Larget, B., Liu, L., Suchard, M., and Huelsenbeck, J. (2012). MrBayes 3.2: Efficient Bayesian

phylogenetic inference and model choice across a large model space. *Systematic Biology*, 61:539–542.

Rosenberg, N. A. and Nordborg, M. (2002). Genealogical trees, coalescent theory and the analysis of genetic polymorphisms. *Nature Reviews Genetics*, 3(5):380–390.

Rue, H. and Held, L. (2005). *Gaussian Markov Random Fields: Theory and Applications*. Chapman and Hall, Boca Raton, FL.

Rue, H., Martino, S., and Chopin, N. (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society: Series B*, 71(2):319–392.

Rue, H., Steinsland, I., and Erland, S. (2004). Approximating hidden Gaussian Markov random fields. *Journal of the Royal Statistical Society. Series B*, 66(4):pp. 877–892.

Shapiro, B., Drummond, A. J., Rambaut, A., Wilson, M. C., Matheus, P. E., Sher, A. V., Pybus, O. G., Gilbert, M. T. P., Barnes, I., Binladen, J., Willerslev, E., Hansen, A. J., Baryshnikov, G. F., Burns, J. A., Davydov, S., Driver, J. C., Froese, D. G., Harington, C. R., Keddie, G., and Kosintsev, P. (2004). Rise and fall of the Beringian steppe bison. *Science*, 306(5701):1561–1565.

Slatkin, M. and Hudson, R. (1991a). Pairwise comparisons of mitochondrial DNA sequences in stable and exponentially growing populations. *Genetics*, 129(2):555–562.

Slatkin, M. and Hudson, R. R. (1991b). Pairwise comparisons of mitochondrial DNA sequences in stable and exponentially growing populations. *Genetics*, 129(2):555–562.

Strimmer, K. and Pybus, O. G. (2001). Exploring the demographic history of DNA sequences using the generalized skyline plot. *Molecular Biology and Evolution*, 18(12):2298 –2305.

Suchard, M., Weiss, R., and Sinsheimer, J. (2001). Bayesian selection of continuous-time markov chain evolutionary models. *Molecular Biology and Evolution*, 18(6):1001–1013.

Tajima, F. (1983). Evolutionary relationship of DNA sequences in finite populations. *Genetics*, 105(2):437–460.

Tavaré, S. (2004). Part I: Ancestral inference in population genetics. In *Lectures on Probability Theory and Statistics*, volume 1837 of *Lecture Notes in Mathematics*, pages 1–188. Springer Verlag, New York.

Teh, Y. W., Seeger, M., and Jordan, M. I. (2005). Semiparametric latent factor models. In *Proceedings of the International Workshop on Artificial Intelligence and Statistics*, volume 10, pages 333–340.

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

Volz, E. M., Pond, S. L. K., Ward, M. J., Brown, A. J. K., and Frost, S. D. W. (2009). Phylodynamics of infectious disease epidemics. *Genetics*, 183(4):1421–1430.

Wakeley, J. and Sargsyan, O. (2008). Extensions of the coalescent effective population size. *Genetics*, 181(1):341–345.

Wakeley, J. and Sargsyan, O. (2009). Extensions of the coalescent effective population size. *Genetics*, 181:341–345.

Wakeley, J., K. L. L. B. S. and Ramachandran, S. (2012). Gene genealogies within a fixed pedigree, and the robustness of Kingman's coalescent. *Genetics*, 190:1433–1445.

Watterson, G. (1975). On the number of segregating sites in genetical models without recombination. *Theoretical Population Biology*, 7(2):256–276.

Welch, D., Bansal, S., and Hunter, D. R. (2011). Statistical inference to advance network models in epidemiology. *Epidemics*, 3:38–45.

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