Preferential sampling and model checking in phylodynamic inference

Michael D Karcher

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

Doctor of Philosophy

University of Washington

2018

Reading Committee:

Volodymyr Minin, Chair

Jonathan C Wakefield

Marina Meila-Predoviciu

Program Authorized to Offer Degree:
Statistics

University of Washington

Abstract

Preferential sampling and model checking
in phylodynamic inference

Michael D Karcher

Chair of the Supervisory Committee:
Professor Volodymyr Minin
Statistics

Estimating population size fluctuations is one of the key tasks in Ecology. Traditional sampling based approaches to this task have limitations when populations of interest are extinct or are hard to reach, as is the case for individuals infected for a short time period by a pathogen. Phylodynamics combines coalescent theory from population genetics and statistical modeling to estimate fluctuations of effective population size—an idealized quantity that can be mapped to census population size with additional demographic information—from molecular sequences of individuals sampled from a population of interest. However, many methods implicitly assume that the samples' collection times do not depend on the effective population size. When sampling times do probabilistically depend on effective population size, estimation methods that do not account for this dependence may be systematically biased. We propose a model that accommodates preferentially sampled data by modeling the distribution of sampling times as an inhomogeneous Poisson process dependent on effective population size via a log-linear intensity function. We extend our model to include optional time-varying covariates into the intensity function. Via simulations and via recent influenza and Ebola datasets, we demonstrate that our model not only reduces bias, but also improves estimation precision. Finally, we propose and implement a posterior predictive diagnostic method to check the adequacy of the coalescent and sampling time models.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

GLOSSARY

COALESCENT MODEL: a population genetics model for relating genealogies of sampled individuals to the demographic properties of the population from which the samples are taken.

EFFECTIVE POPULATION SIZE: a measure of genetic diversity, equivalent to the population size if the population followed the Wright-Fisher model.

GENEALOGY: a rooted, bifurcating tree that tracks the mututal ancestry of a collection of samples from the population of interest.

HETEROCHRONOUS: a sample taken from more than one time point, or a genealogy based thereon.

ISOCHRONOUS: a sample where every member is taken from the same time point, or a genealogy based thereon.

PHYLOGENETICS: the study of the evolutionary relationships of a collection of species or organisms.

WRIGHT-FISHER MODEL: a discrete-generation-based model of genetic inheritance within a population. Its assumptions and implications are covered in more depth in §2.1.2.

ACKNOWLEDGMENTS

Heartfelt thanks to my advisor Vladimir Minin for years of wise guidance, patience, and good humor. Thanks to Julia Palacios, Shiwei Lan, and Arman Bilge for teaching me the right way to code. Thanks to Trevor Bedford for sharing his colors. And many thanks to my committee, Marina Meila, Jon Wakefield, and Lorenz Hauser, for their efforts and support.

DEDICATION

to my parents, Jack and Anne Karcher

Chapter 1

INTRODUCTION

Phylodynamics, as a subfield of population genetics and epidemiology, represents the study of how (most often viral) populations undergo changes in genetic variation due to evolutionary and epidemiological pressures, as well as how these changes show up in the populations' phylogenies (species' family trees) and genealogies (individuals' family trees). It is common to analyze population genetics data, and thereby phylodynamics data, using a probabilistic model called the *coalescent* [Kingman, 1982]—a time-reversed extension of the Wright-Fisher model which relates the shape of a population's phylogeny or *genealogy* to the amount of genetic diversity in the population, called the *effective population size* which is what the size of the population would be if it followed the Wright-Fisher model perfectly.

In many population genetics applications, such as studies of human ancestry, sequence data are sampled on a markedly different timescale to the the genealogies that bind their genetic samples together. This leads to the theoretical underpinnings of *isochronous* sequence data, where sampling times are treated as effectively simultaneous. In others, such as many infectious disease applications, sampling times and genealogical times will be on similar timescales, forcing sampling times to be considered non-simultaneous or *heterochronous*. Heterochronous data allow the possibility of estimating both the mutation rate affecting the sequence data and estimating important genealogy events in terms of calendar time. Phylodynamics uses both types of data to gain insight into how infectious diseases spread through the population, as well as the effectiveness of control efforts against this spread. We review the theory behind coalescent-based inference and its phylodynamic applications in chapter 2.

However, having heterochronous data also leaves open the possibility of a relationship

between sampling frequency and changes in the trajectory of the effective population size. In chapter 3, we explore the consequences of having model misspecification due to not accounting for a relationship between sampling times and effective population size. We suggest modifications to the coalescent model to improve Bayesian inference, including a simple sampling model for scenarios where genetic sequences are sampled with frequency proportional to a power (including zero) of the effective population size. We examine the effect of using different models on several influenza case studies from different parts of the world. (This chapter is a modified version of the paper by Karcher et al. [2016a].)

In chapter 4, we extend our model to relax the fixed-genealogy assumption to allow inference from genetic sequence data. We also extend the sampling model to be able to include time-varying covariates as additional sources of information in the sampling model. We implement our preferential sampling model in a widely used and versatile software BEAST [Drummond et al., 2012] to make our methodology readily accessible to research groups working on evolutionary biology and/or epidemiology. We continue our analysis of regional influenza data and examine a recent ebola dataset.

Finally, in chapter 5, we develop tools for checking the validity of the coalescent and sampling models. We propose a series of posterior predictive checks for the coalescent model and the sampling model and explore what results we tend to see under different model specifications and misspecifications. We quantitatively compare the suitability of different models on our datasets.

Chapter 2

AN OVERVIEW OF COALESCENT BASED INFERENCE

2.1 Introduction

2.1.1 Genealogies

The use of tree graphs to describe a shared ancestry is as old as the concept of a "family tree." More recently, scientists have applied the idea of an ancestral tree to the study of phylogenetics based on genetic data, among other applications in evolutionary biology. In the context of the relationship between species, the ancestral tree takes the form of a *phylogeny*, with extant (or extinct, but witnessed) species represented by the leaves of the tree, and unobserved mutual ancestral species represented by the nodes. Similarly, in the context of the relationship between members of a population *within* a species, we refer to the ancestral tree as a *genealogy*. Here the leaves/tips represent sampled members of the population, and the nodes represent (unobserved) mutual ancestors, up to and including the *most recent common ancestor* (MRCA) of the entire sample.

In practice, we rarely observe a phylogeny or genealogy directly, instead we rely on other sources of data to estimate our phylogeny or genealogy. The earliest phylogenies were reliant on expert opinion based on observations of the species' shared traits. For most of history the only genealogical data available were family trees and pedigrees recorded by family members and animal owners. However, as genetic sequencing technology has advanced and the amount of genetic data available has grown exponentially, significant new opportunities and data sources for genealogical inference have become available. Genetic sequence data for homologous sites provides information about the sampled organisms' shared ancestral history via noting which mutations are shared and which mutations are not shared between which samples. In order to relate and analyze the genetic, genealogical, and population data, we

require a probabilistic generating model capable of producing genealogies given a population, as well as a model of how observed genetic data are generated given a genealogy. We can think of a tree generating process as running forward in time from the most recent common ancestor splitting up lineages along the way until they end at a sample. Alternately, we can view it as a process running backwards in time, traveling back from the present (or the most recent sample) joining up, or *coalescing*, lineages until it reaches the most recent common ancestor. We begin with the forward process to lay a foundation for the genealogical and genetic models, then we formalize the backward process into an important model called *the Coalescent*.

2.1.2 The Wright-Fisher Model

Wright [1931] proposed a simple model of evolution. Consider a population of N members of a haploid (having unpaired genetic sequences) population. Suppose that the population has discrete and non-overlapping generations, and that each individual's number of offspring does not depend on any heritable fitness traits, differences in geographic location, or differences in local social structure. We call the lack of fitness traits an absence of *selection*, and we refer to the lack of geographic and social structure as the population being *panmictic*. Finally, suppose that the next generation forward in time also has N members, and each member randomly "chooses" a parent (with replacement) from the preceding generation. We illustrate this process in Figure 2.1.

It is most often impossible to observe the full ancestral history of a population. Instead, we are constrained to taking samples from the population and inferring the genealogy that links them together. If we select a sample from the population at the present time and trace back their shared ancestry through the past, we see something like the right plot in Figure 2.1. As we will explore in the following sections, if we base our inference on genetic sequence data we only have information about the genealogy of our sample and not the rest of the population. However, as we will see, the coalescent times and the waiting times between coalescences tell us about the population and any population dynamics that are occurring.

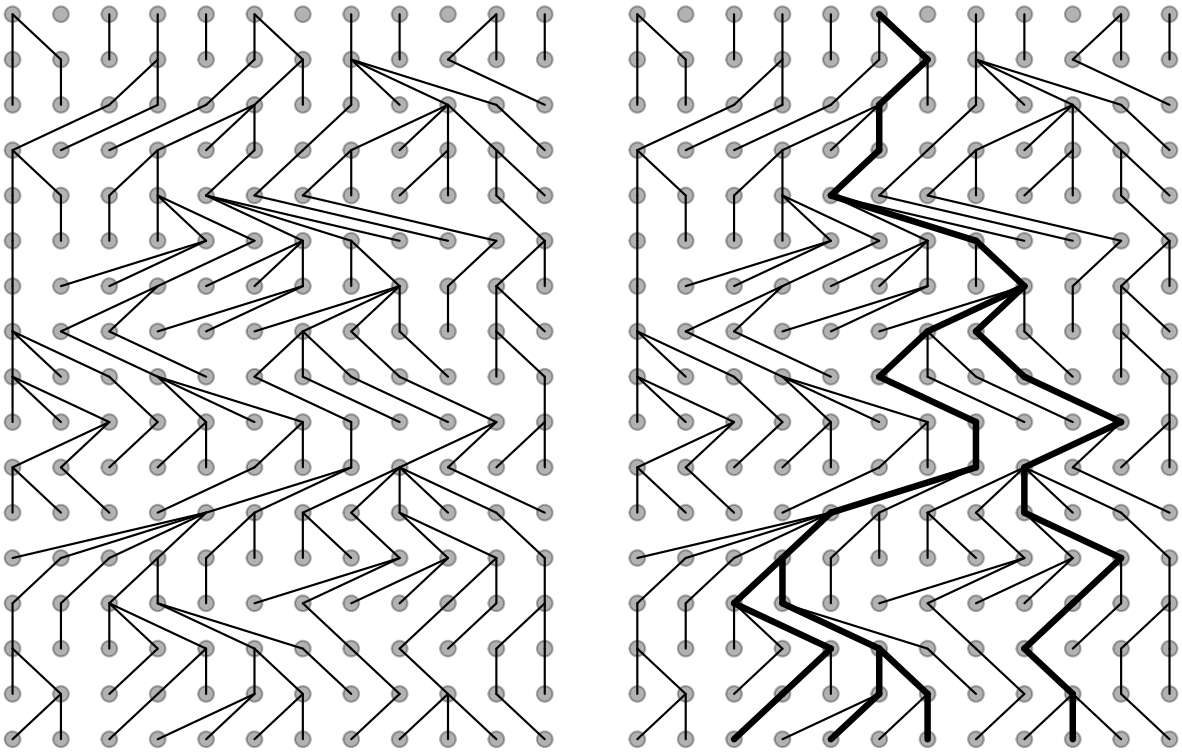

Figure 2.1: **Illustration of an example Wright-Fisher population.** Here the population size $N = 12$, and we see 17 generations, or 16 iterations of the Wright-Fisher process. The left plot shows the entire ancestral history of the population. The right plot shows the same, but with a random sample and its genealogy highlighted. Time runs forward from the top of the plot to the bottom of the plot (the present).

2.2 The Coalescent

2.2.1 Discrete-Time Coalescent

We now change perspective from a prospective model, iterating the entire population forward through offspring generations, to a retrospective model, iterating a sample from the population backwards via ancestral links. If we restrict ourselves to a model that matches the Wright-Fisher model when iterated forward in time, we induce a backwards-time model called the *discrete-time coalescent*. Suppose we select two members of a size-N Wright-Fisher population at the present time. If we iterate back one generation, we apply the Wright-Fisher uniform ancestor selection criterion once, and there is a $1/N$ probability that both members select the same ancestor or coalesce (the first member can choose any ancestor, but the second must choose that same ancestor out of N options, resulting in a $1/N$ probability). If they do not coalesce in that generation, the process can be repeated independently and identically until they do, some random number of generations back in time. If we label this random variable U_2, it follows directly that U_2 has a geometric distribution with event probability $1/N$ and expected value N.

Suppose instead we select k members of a size-N Wright-Fisher population. Then the probability that *none* of the members coalesce in the previous generation is

$$\left(1-\frac{1}{N}\right)\left(1-\frac{2}{N}\right)\ldots\left(1-\frac{k-1}{N}\right) \approx 1-\sum_{i=1}^{k-1}\frac{i}{N} = 1-\binom{k}{2}\frac{1}{N}$$

for N sufficiently larger than k. Then the probability of a coalescent event occurring among k lineages in a population of size N is approximately $\binom{k}{2}/N$. This leads to the random variable U_k, the waiting time (in generations) before k lineages experience a coalescent event, having a geometric distribution with event probability $\binom{k}{2}/N$. Then U_k has cumulative distribution function (CDF),

$$\Pr(U_k \leq n) \approx \left[1-\binom{k}{2}\frac{1}{N}\right]^{n-1}$$

and expected value $N/\binom{k}{2}$. We highlight one possible realization of this process within a Wright-Fisher population in the right-side plot in Figure 2.1.

2.2.2 Continuous-Time Coalescent

The Wright-Fisher model relies on the assumption of discrete, non-overlapping generations. Suppose we divide each generation into s equally smaller time steps. We want to retain as much equivalence to the Wright-Fisher induced model as possible, so we divide the probability of k lineages coalescing by s. We label the discrete, but not integer, waiting time for k lineages to coalesce (in units of generations) $U_k^{(s)}$. Then its CDF is,

$$\Pr(U_k^{(s)} \leq t) \approx \left[1 - \frac{1}{s}\binom{k}{2}\frac{1}{N}\right]^{st-1},$$

which has the limit

$$\Pr(U_k^c \leq t) = 1 - e^{t\binom{k}{2}\frac{1}{N}},$$

as $s \to \infty$. We label the now continuous, exponentially distributed waiting time U_k^c, still in units of generations. If we run the waiting time for a coalescent event for k lineages, then the waiting time for a coalescent event for $k-1$ lineages, and so on until we have only one lineage remaining (the MRCA), then we have generated one realization of the *continuous-time coalescent* on k lineages from a population of size N.

2.2.3 Extensions to the Coalescent

The coalescent as we have explored it so far makes some stringent assumptions. The population size is assumed to remain constant at every generation in the discrete-time coalescent and constant at every time point in the continuous-time coalescent. We also assume that all samples are taken at the present (or most recent) time. Here we explore relaxations of these assumptions that lead to useful extensions to the coalescent framework in the context of population dynamics.

First, we relax the Wright-Fisher assumption of constant population size. Suppose the population size follows a fixed trajectory $N(t)$, with $t = 0$ representing the present and t increasing backwards in time. Then in the discrete-time coalescent, the probability of two lineages coalescing at time t is $1/N(t)$, and probability of k lineages having a coalescent event

at time t is approximately $\binom{k}{2}/N(t)$. This results in the waiting times having CDF,

$$\Pr(U_k \leq n|T_k) \approx \prod_{i=1}^{n}\left[1 - \binom{k}{2}\frac{1}{N(T_k+i)}\right],$$

conditioning on T_k, where T_k is the cumulative waiting time for there to remain only k lineages. Note that the waiting times are not independent, nor geometrically distributed, but do have the Markov property. Via a similar argument as the last section, the continuous-time coalescent waiting time, conditioning on T_k, has CDF

$$\Pr(U_k^c \leq t|T_k) = 1 - e^{\binom{k}{2}\int_{T_k}^{T_k+t}\frac{1}{N(t)}dt},$$

and density

$$\Pr(U_k^c = t|T_k) = \binom{k}{2}\frac{1}{N(T_k+t)}e^{\binom{k}{2}\int_{T_k}^{T_k+t}\frac{1}{N(t)}dt}.$$

This is also related to distributions induced by hazard rates, with hazard function $\binom{k}{2}\frac{1}{N(t)}$. Also note that the waiting times are not independent, nor exponentially distributed, but do again have the Markov property. From this point onward, we consider only the continuous-time coalescent due to the unrealistic nature of discrete, nonoverlapping generations, as well as the fact that the distribution functions for the continuous-time coalescent are exact as opposed to the approximations shown for the discrete-time coalescent.

Notice that so far we have been considering samples consisting entirely of members taken at $t = 0$. We refer to samples like this, and the genealogies that link them, as *isochronous* samples and genealogies. If we instead select n samples from different, but potentially nonunique, times $s_n \leq s_{n-1} \leq \ldots s_1$, we refer to the sample and genealogy as *heterochronous*. We illustrate a heterochronous genealogy in Figure 2.2. We partition the interval $[s_n, t_1]$ into subintervals $\{I_i\}$ wherein there are a constant number n_i of *active lineages*. Suppose we define $n(t) = \sum_i n_i \delta_{t\in I_i}$, where δ is the indicator function. Then the hazard of coalescence is $\lambda_c(t) = \binom{n(t)}{2}/N(t)$, and the likelihood is,

$$\Pr(\mathbf{g}|N(t), \mathbf{s}) \sim \prod_{k=2}^{n}\left[\lambda_c(t_{k-1})\exp\left(-\int_{I_{i,k}}\lambda_c(t)dt\right)\right], \tag{2.1}$$

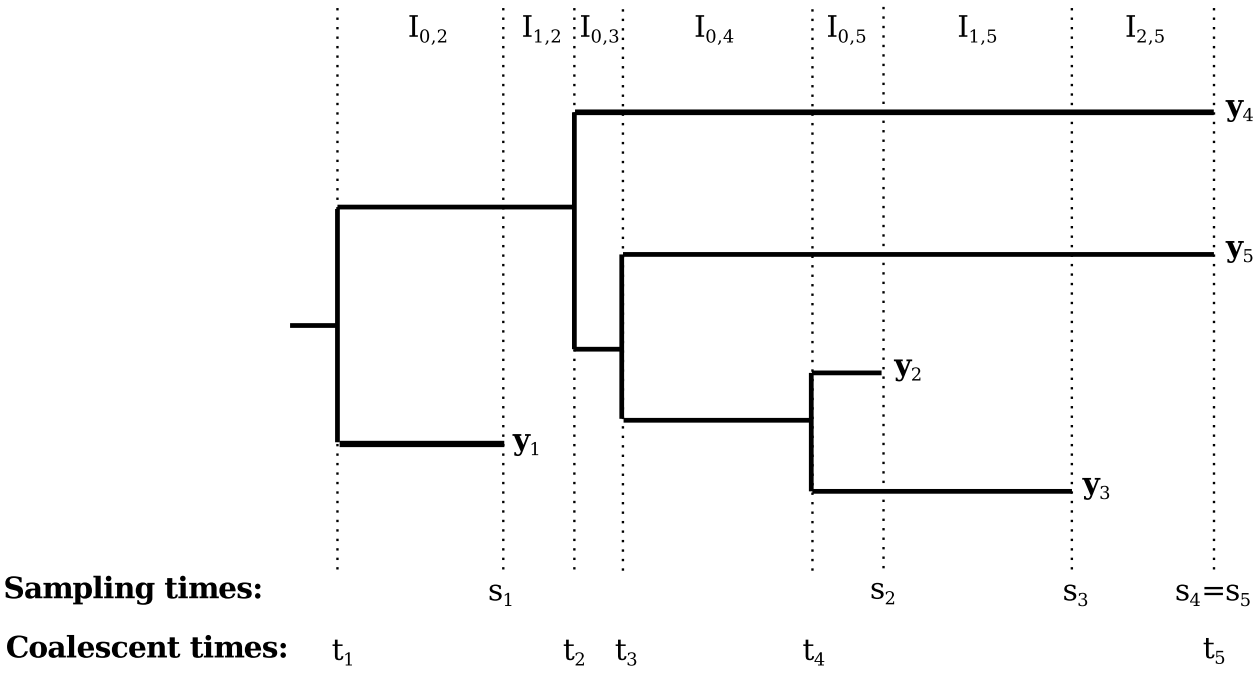

Figure 2.2: **Illustration of an example heterochronous genealogy with $n = 5$ lineages.** Sampling times $s_1, \ldots, s_5$ and coalescent times $t_1, \ldots, t_5$ are marked below the genealogy.

where $\mathbf{g} = t_n, \ldots, t_1$.

Finally, the assumptions of the Wright-Fisher model most often do not hold in nature, leading to estimates of N or $N(t)$ that do not reflect the true population size. However, Wright-Fisher and coalescent dynamics are often a good approximation, with one conceptual adjustment. Suppose instead of trying to estimate N, the census population size, we try to estimate N_e, the population size that makes the Wright-Fisher model best fit the data, which we refer to as the *effective population size*. There exist several often incompatible definitions of the effective population size. In this dissertation, we use the *inbreeding effective size*, a measure of genetic diversity which seeks consistency between the Wright-Fisher model and the genealogical and genetic sequence data.

2.3 Sequence Data

2.3.1 DNA Evolution

Consider a collection of samples of genetic sequence data from a population of interest. Each individual's data takes the form of a string of *nucleotides*, the smallest chemical building blocks of DNA or RNA, sampled from the individual's *genome*. Individual nucleotides are abbreviated as A, C, G, and T (DNA) or U (RNA). Many of our later examples will involve RNA viruses, but for consistency of notation we will only use DNA notation. Locations in the genome are called *sites*. Due to the often imprecise nature of sampling specific sites from a genome, care is required to align the sequence samples. A dataset where corresponding sites have been matched up is called an *alignment*.

Where the previous section modeled the relationship between (effective) population size and genealogies, here we consider the relationship between genealogies and observed sequence data via models of DNA evolution. The most common models are based on continuous-time Markov chains (CTMCs). Given a genealogy relating the samples to each other, each site is modeled to independently mutate according to a CTMC, however, with the restriction that mutations that accumulate earlier in the genealogy while two samples have not yet diverged

must be shared by those two samples. Each CTMC has a transition rate matrix $\mathbf{Q}$ whose off-diagonal elements determine how often a chain in each state transitions to another state, and whose diagonal elements allow each row to sum to zero. The transition *probability* matrix $\mathbf{P}(t) = e^{\mathbf{Q}t}$, and its i, jth element represents the chance that the CTMC ends in state j at time t, having started at state i at time 0. These models are also memoryless, so $\mathbf{P}(t)$ also represents the probabilities of each transition ending at time $u+t$ when starting at time u. A note on terminology, due to the fact that computational evolutionary biology uses the term "transition rate" differently than in statistics, for clarity from here on we refer to $\mathbf{Q}$ as the *substitution rate matrix*. Populating the substitution rate matrix leads to several important models, discussed below.

Jukes-Cantor 69

Jukes et al. [1969] proposed the first CTMC-based model of DNA evolution, denoted JC69 in the literature. It assumes that the off-diagonal substitution rates are all identical and equal to λ, resulting in the substitution rate matrix (with states ordered alphabetically A,C,G,T),

$$\mathbf{Q}_{JC69} = \begin{bmatrix} * & \lambda & \lambda & \lambda \\ \lambda & * & \lambda & \lambda \\ \lambda & \lambda & * & \lambda \\ \lambda & \lambda & \lambda & * \end{bmatrix},$$

where $*$ represents the necessary terms to make the rows sum to zero. JC69 has the stationary distribution $\boldsymbol{\pi} = \{\pi_A, \pi_C, \pi_G, \pi_T\} = \{0.25, 0.25, 0.25, 0.25\}$, which means that as time goes on the probability of being in each state approaches 0.25, since the CTMC is irreducible and positive recurrent.

Kimura 80

Kimura [1980] extended JC69 to account for that fact that there is a biological and chemical difference in rate between *transitions* (substitutions within the sets $\{A, G\}$ and $\{C, T\}$, with

rate α) and *transversions* (substitutions between the sets $\{A, G\}$ and $\{C, T\}$, with rate β). Kimura's model (K80) has the substitution rate matrix,

$$\mathbf{Q}_{K80} = \begin{bmatrix} * & \beta & \alpha & \beta \\ \beta & * & \beta & \alpha \\ \alpha & \beta & * & \beta \\ \beta & \alpha & \beta & * \end{bmatrix},$$

with the same uniform stationary distribution $\boldsymbol{\pi} = \{\pi_A, \pi_C, \pi_G, \pi_T\} = \{0.25, 0.25, 0.25, 0.25\}$ as JC69.

Felsenstein 81

Felsenstein [1981] proposed an extension to JC69 called F81, which allows for differing stationary probabilities. F81 has the substitution rate matrix,

$$\mathbf{Q}_{F81} = \begin{bmatrix} * & \pi_C & \pi_G & \pi_T \\ \pi_A & * & \pi_G & \pi_T \\ \pi_A & \pi_C & * & \pi_T \\ \pi_A & \pi_C & \pi_G & * \end{bmatrix},$$

with a stationary distribution that matches the parameters in $\mathbf{Q}_{F81}$, $\boldsymbol{\pi} = \{\pi_A, \pi_C, \pi_G, \pi_T\}$.

Hasegawa, Kishino and Yano 85

Hasegawa et al. [1985] combined the approaches of K80 and F81 to allow for the distinction between transitions (with rate α) and transversions (with rate β), as well as arbitrary stationary probabilities. Their model, HKY85, has the substitution rate matrix,

$$\mathbf{Q}_{HKY85} = \begin{bmatrix} * & \beta\pi_C & \alpha\pi_G & \beta\pi_T \\ \beta\pi_A & * & \beta\pi_G & \alpha\pi_T \\ \alpha\pi_A & \beta\pi_C & * & \beta\pi_T \\ \beta\pi_A & \alpha\pi_C & \beta\pi_G & * \end{bmatrix},$$

with a stationary distribution that matches the π parameters, $\boldsymbol{\pi} = \{\pi_A, \pi_C, \pi_G, \pi_T\}$.

2.3.2 Sequence Likelihood

With a selection of CTMC-based models available to us, it becomes important to have a practical algorithm for calculating the likelihood relating a genetic sequence alignment to its genealogy. We assume that multiple sites are mutating independently, so we can break the sequence alignment likelihood into factors by site. For a single site, if we knew its state at every node and tip in the genealogy it would be a simple matter of, for every branch of the genealogy, multiplying the substitution probabilities $P_{ij}(t)$, where i is the state at the start of the branch, j is the state at the end of the branch, and t is the length of the branch.

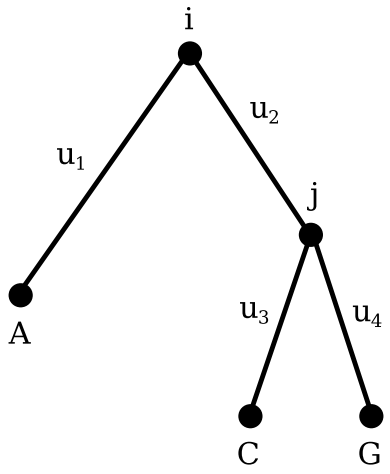

Figure 2.3: **Illustration of an example phylogenetic genealogy with $n = 3$ lineages.** Observed tip states A,C,G are marked below the genealogy, unobserved internal nodes i, j are marked above, and branch lengths $u_1, \ldots, u_4$ are next to their branch.

However, in practice we only observe sequence data at the tips. We illustrate an example phylogenetic genealogy in Figure 2.3. Given the observed tips, the substitution model $\mathbf{Q}$, the tree topology $\boldsymbol{\tau}$, and the branch lengths $\{u_i\}$, we can still calculate the likelihood using

the law of total probability, which for the tree and data in Figure 2.3 is,

$$\begin{aligned}\Pr(A, C, G|\boldsymbol{\tau}, \mathbf{u}, \mathbf{Q}) &= \sum_i \sum_j \Pr(A, C, G, i, j|\boldsymbol{\tau}, \mathbf{u}, \mathbf{Q}) \\ &= \sum_i \sum_j \pi_i P_{iA}(u_1) P_{ij}(u_2) P_{jC}(u_3) P_{jG}(u_4).\end{aligned}$$

This quickly becomes a intractably large summation as the number of tips grows. However, the distributive property allows for a more efficient computation. Felsenstein [1973] proposed the *tree-pruning algorithm* which gathers like terms to reduce the number of required operations. Applied to our example tree, the likelihood becomes,

$$\begin{aligned}\Pr(A, C, G|\boldsymbol{\tau}, \mathbf{u}, \mathbf{Q}) &= \sum_i \sum_j \pi_i P_{iA}(u_1) P_{ij}(u_2) P_{jC}(u_3) P_{jG}(u_4) \\ &= \sum_i \pi_i P_{iA}(u_1) \left[\sum_j P_{ij}(u_2) P_{jC}(u_3) P_{jG}(u_4) \right].\end{aligned}$$

The savings in operations comes from being able to cache the results of

$$\sum_j P_{ij}(u_2) P_{jC}(u_3) P_{jG}(u_4)$$

in our example, and similar results around each node in the general case. Even in our simple example we see a savings, going from 4^2 summation calculations down to $4 \cdot 2$.

2.3.3 Applications in Epidemiology

Coalescent-based inference has many useful epidemiological applications. The effective population size itself can be considered a proxy for the number of infections or number of infected individuals. However, effective population size is often more difficult to interpret, and care should be taken to preserve statistical and epidemiological integrity [Pybus et al., 2001, Frost and Volz, 2010]. Estimates of changes in the effective population size for an epidemic provide a possible estimate of the disease's basic reproduction number R_0, which would otherwise be challenging to obtain via traditional surveillance methods [Volz et al., 2009]. Similarly, estimates of within-host viral growth are achievable, allowing for better understanding of

infection mechanics [Gray et al., 2012]. Phylodynamic methods also allow for insight into the effectiveness of epidemiological control efforts [Van Ballegooijen et al., 2009]. Inferring effective population size changes can show how effective vaccines and isolation efforts (for example, school closings and quarantines) are at affecting an epidemic's effective R_0. Estimating changes in the mutation rate can reveal evolutionary pressure and the effectiveness of biological control efforts (for example, anti-retrovirals) [Drummond et al., 2001].

Heterochronous data allows for estimating the mutation rate in molecular clock models. This also allows coalescent events to be estimated in terms of calendar time, up to and including the time of the most recent common ancestor (TMRCA) for that sample of sequence data [Fraser et al., 2009]. Estimates of the TMRCA sampling infections from different hosts provide information about the disease's epidemiological origins, while estimates of the TMRCA for a sample from within a single host produce a lower bound for the individual's infection time [Lemey et al., 2006].

Chapter 3

QUANTIFYING AND MITIGATING THE EFFECT OF PREFERENTIAL SAMPLING ON PHYLODYNAMIC INFERENCE

3.1 Introduction

Phylodynamics — a set of techniques for estimating population dynamics from genetic data — has proven useful in ecology and epidemiology [Grenfell et al., 2004, Holmes and Grenfell, 2009]. Phylodynamics is especially useful in cases where ascertaining population sizes via traditional sampling methods is infeasible; e.g., in infectious disease epidemiology it is impossible to obtain the total number of infected individuals in a large population. Estimating population dynamics from a limited sample of genetic data is possible because changes in population size leave evidence in the molecular sequences of the population. Recently, techniques employing a nonparametric approach to inferring population trajectories have improved upon earlier models in terms of flexibility, accuracy, and precision by, e.g., employing Gaussian Markov random fields [Minin et al., 2008, Gill et al., 2013] and Gaussian processes [Palacios and Minin, 2013]. However, none of these state-of-the-art methods currently account for randomness in sampling time data, potentially introducing bias in studies where sampling times have a relationship to population dynamics. Through a simulation study we characterize this bias in a demographic scenario with seasonally varying population size. We also extend the state-of-the-art by incorporating a sampling time model into phylodynamic inference, mitigating the bias and improving precision.

Phylodynamic methods use Kingman's coalescent model that, given a particular effective population size trajectory, defines the density of a genealogy relating the sampled individuals [Kingman, 1982]. Effective population size measures genetic diversity present in the pop-

ulation and relates to census population size if certain assumptions are met [Wakeley and Sargsyan, 2009]. Many early coalescent-based phylodynamic methods required strict parametric assumptions about the effective population size trajectory, such as constant through time [Griffiths and Tavaré, 1994b] or exponential growth [Drummond et al., 2002, Kuhner et al., 1998]. A major alternative arose with the advent of nonparametric methods, one of the earliest and most influential being the piecewise constant classical skyline model [Pybus et al., 2000]. This approach greatly increases the number of estimated parameters, leading to noisy effective population size trajectories. A number of algorithms seeking compromise between the relative stability of parametric approaches and the flexibility of nonparametric approaches have been implemented [Drummond et al., 2005, Minin et al., 2008, Gill et al., 2013]. For a detailed comparison, see [Ho and Shapiro, 2011].

Many successful applications of phylodynamics methodology come from infectious disease epidemiology, where the effective population size is interpreted, albeit with caution, as the effective number of infections [Frost and Volz, 2010]. In these epidemiological applications, disease agent DNA or RNA sequences are collected at multiple times. When analyzing such heterochronous data, researchers implicitly assume that sampling times are either fixed or follow a distribution that is functionally independent of the effective population size trajectory. However, it is conceivable that the infectious disease agent DNA samples are collected more frequently when the number of infections is high and less frequently during time periods with few infections. Therefore, the implicit assumption of no relationship between sampling times and population dynamics, made by all state-of-the-art phylodynamic methods, is troublesome, since unrecognized preferential sampling leads to systematic estimation bias, as explored by Diggle et al. [2010] in the context of spatial statistics. Furthermore, preferential sampling could be present in the sequence databases, but it could also be introduced accidentally or intentionally by filtering during database queries or data mining.

To test the effect of preferential sampling on phylodynamic inference we first perform a simulation study. We simulate sampling times according to multiple distributions, contrasting distributions functionally dependent on effective population size with a functionally

independent distribution. We then simulate genealogies based on the sampling times and perform state-of-the-art phylodynamic analyses, and we find that ignoring preferential sampling can bias effective population size estimation and that the size of the bias depends on the local properties of the effective population size trajectory.

In order to account for preferential sampling, we formulate a new phylodynamic model in which sampling times are generated from an inhomogeneous Poisson process with intensity functionally dependent on effective population size. Our model is similar to the augmented coalescent model of Volz and Frost, who work with a specific parametric model [Volz and Frost, 2014]. In contrast, we work within a nonparametric framework by incorporating our Poisson preferential sampling model into a Gaussian process-based Bayesian phylodynamic method [Minin et al., 2008, Gill et al., 2013, Palacios and Minin, 2013]. Applying our new sampling-aware method to our simulations shows that modeling preferential sampling eliminates the aforementioned bias and can increase precision of the phylodynamic inference. In all of our developments, we assume that the genealogy of the sample is known without error. This assumption allows us to use an integrated nested Laplace approximation (INLA) to make our Bayesian inference computationally efficient [Rue et al., 2009, Palacios and Minin, 2012], which is important for executing our simulation studies.

Finally, we examine the performance of our algorithm on two real-world examples. Rambaut et al. [2008] explore the seasonal variation of genetic diversity in the genes that code for several of the most important proteins in the two most common influenza subtypes, H3N2 and H1N1. For the sake of brevity we only analyze the hemagglutinin gene in H3N2. We find evidence of preferential sampling in the dataset, and our sampling-aware method produces a large improvement in precision over the conditional (sampling un-aware) method. Zinder et al. [2014] specifically explore the patterns of seasonal migration of genetic diversity of H3N2 influenza across the regions of the world. We examine the regions separately and find differing strengths of preferential sampling, but in all regions our method performs better than the conditional model. In some regions, we see stronger relationships between sampling frequency and population size, most often in regions with the most seasonal variation in

incidence.

3.2 Methods

3.2.1 State-of-the-art phylodynamics

Consider a sample of individuals from a well-mixed population. Some individuals will share a common ancestor more recently than others. One pair of individuals in particular will have the pairwise most recent common ancestor. Moving backwards in time, we can consider those two individuals to have *coalesced*, treating the two individuals as one. We can then repeat this process of finding the pairwise most recent common ancestor and coalescing individuals until we reach the most recent common ancestor of the entire sample. If we keep track of the ancestral lineages and coalescences of the individuals, we see the data take the shape of a bifurcating tree, and we refer to this ancestry tree as a *genealogy* (illustrated in Figure 3.1).

We refer to the branching points of the genealogy tree as *coalescent events*. If the samples are all taken simultaneously, we refer to the genealogy as *isochronous*. Kingman's original coalescent provided a density for isochronous genealogies with a fixed effective population size [Kingman, 1982]. Later extensions to the coalescent allowed for parametric and non-parametric specifications of effective population size trajectories along with *heterochronous* sampling times. Heterochronous sampling times (also called sampling events) can occur at any time up to the present.

We consider first the case of a fixed, heterochronous genealogy [Felsenstein and Rodrigo, 1999]. The coalescent likelihood has sufficient statistics $\mathbf{g}$ and $\mathbf{s}$ with

$$\mathbf{g} = \{t_i\}_{i=1}^{n}, 0 = t_n < t_{n-1} < \ldots < t_1$$

representing the coalescent times and

$$\mathbf{s} = \{s_i, n_i\}_{i=1}^{m}, 0 = s_m < s_{m-1} < \ldots < s_1, \sum_{j=1}^{m} n_j = n$$

representing the sampling times along with the corresponding number of lineages sampled (see Figure 3.1). We define the number of *active lineages* at time t as the number of lineages

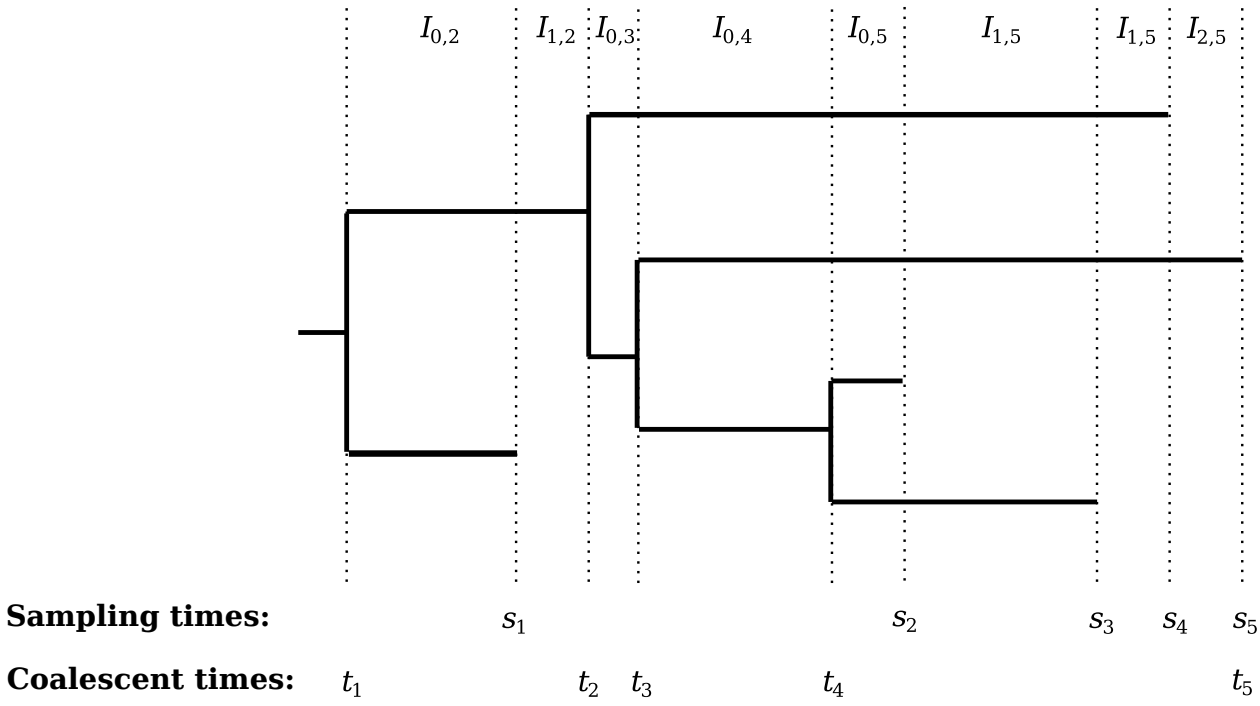

Figure 3.1: **Illustration of an example heterochronous genealogy with $n = 5$ lineages.** Sampling times $s_1, \ldots, s_5$ and coalescent times $t_1, \ldots, t_5$ are marked below the genealogy.

sampled between t and the present, minus the number of coalescent events between t and the present. In Figure 3.1, this appears as the number of horizontal lines that a vertical line at time t will cross.

We define a partition of $(0, t_1)$ with intervals $I_{i,k}$ for $k = 1, \ldots, n$. We let $I_{0,k}$ represent the intervals ending with a coalescent event and let $I_{i,k}$ for $i = 1, \ldots, m_k$ represent the m_k intervals ending in a sampling event between the $(k-1)$th and kth coalescent events (see Intervals in Figure 3.1). We let $C_{i,k} = \binom{n_{i,k}}{2}$, where $n_{i,k}$ is the number of active lineages in the interval $I_{i,k}$. Suppose $\mathbf{s}$ is fixed, then the coalescent likelihood is

$$\Pr[\mathbf{g}|N_e(t), \mathbf{s}] \propto \prod_{k=2}^{n} \frac{C_{0,k}}{N_e(t_{k-1})} \exp\left[-\sum_{i=0}^{m_k} \int_{I_{i,k}} \frac{C_{i,k}}{N_e(t)} dt\right].$$

In Bayesian phylodynamic inference, our aim is to explore the posterior distribution of the effective population size trajectory $N_e(t)$, so we employ a Gaussian process prior $\Pr[N_e(t) \mid \tau]$, where $N_e(t) = \exp[\gamma(t)]$, with $\gamma(t) \sim \mathcal{BM}(\tau)$ following a Brownian motion with precision parameter τ [Palacios and Minin, 2012]. We assign a Gamma(0.01, 0.01) hyperprior to τ. This results in the posterior $\Pr[N_e(t), \tau \mid \mathbf{g}] \propto \Pr[\mathbf{g} \mid N_e(t)] \Pr[N_e(t) \mid \tau] \Pr(\tau)$.

The continuous case as written above involves an infinite-dimensional object—the function $N_e(t)$—which makes the problem as stated intractable. However, we can approximate the continuous function with a piecewise constant function. We construct a fine, regular grid $\mathbf{x} = \{x_j\}_{j=1}^{B}$ with grid width w over the interval that supports the genealogy and let $\gamma_j = \log[N_e(x_j)]$. We construct a piecewise constant approximation $N^{\boldsymbol{\gamma}}(t) = \sum_{i=1}^{B} \exp(\gamma_i) 1_{t \in [x_i - w/2, x_i + w/2)}$. The discretized coalescent likelihood becomes

$$\Pr(\mathbf{g} \mid \boldsymbol{\gamma}) \propto \prod_{k=2}^{n} \frac{C_{0,k}}{N^{\boldsymbol{\gamma}}(t_{k-1})} \exp\left[-\sum_{i=0}^{m_k} \int_{I_{i,k}} \frac{C_{i,k}}{N^{\boldsymbol{\gamma}}(t)} dt\right], \tag{3.1}$$

where $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_B)$ and the integrals are simple to compute over the step function $N^{\boldsymbol{\gamma}}(t)$. We discretize the Brownian process prior with an intrinsic random walk prior,

$$\Pr(\boldsymbol{\gamma} \mid \tau) \propto \tau^{(n-1)/2} \exp\left[-\frac{\tau}{2} \sum_{k=1}^{B-1} (\gamma_{k+1} - \gamma_k)^2\right].$$

Finally, the discretized posterior becomes $\Pr(\boldsymbol{\gamma}, \tau \mid \mathbf{g}) \propto \Pr(\mathbf{g} \mid \boldsymbol{\gamma}) \Pr(\boldsymbol{\gamma}|\tau) \Pr(\tau)$.

With the posterior known (up to a proportionality constant), we can proceed with numerical integration techniques such as Markov chain Monte Carlo (MCMC) or INLA — a deterministic algorithm for approximating posterior distributions. We select INLA and name the implementation Bayesian nonparametric phylodynamic reconstruction (BNPR).

3.2.2 Phylodynamics with preferential sampling

In the previous section we made the assumption that we could safely ignore any potential dependence of sampling times $\mathbf{s}$ on effective population size $N^{\gamma}(t)$ in our calculations. In this section, we relax this assumption. We model sampling times according to an inhomogeneous Poisson process in a fixed sampling window $[0, s_0]$, with intensity $\lambda(t) = \exp(\beta_0)[N^{\gamma}(t)]^{\beta_1}$, i.e. proportional to a power of the effective population size, where β_0 and β_1 are unknown parameters. The sampling log-likelihood is

$$\log[\Pr(\mathbf{s} \mid \boldsymbol{\gamma}, \beta_0, \beta_1)] = C + n\beta_0 + \sum_{i=1}^{n} \beta_1 \log[N^{\gamma}(s_i)] - \int_{s_m}^{s_0} \exp(\beta_0)[N^{\gamma}(r)]^{\beta_1} dr.$$

To illustrate our parameterization, sampling with $\beta_1 = 1$ would result in collecting genetic sequences with intensity directly proportional to effective population size, while higher β_1 values result in more clustered samples. Conversely, $\beta_1 = 0$ produces a uniform distribution of sampling times, with a Poisson distribution on the number of individuals sampled.

In many datasets, the sampling time data will have low enough resolution (for instance, only recording the date but not time of sampling) that some sampling times will appear to be coincident. Our sampling model is compatible with simultaneous sampling times because the model naturally bins the samples along our earlier discretization. The likelihood is proportional to a product of Poisson mass functions centered at the grid points $\mathbf{x}$.

The genealogy depends on the sampling times, so we condition on $\mathbf{s}$ in the likelihood for $\mathbf{g}$. We are treating $\mathbf{s}$ as random, so we insert the likelihood term for it as well as independent Normal priors for parameters β_0 and β_1—specifically $\beta_i \sim N(\text{mean} = 0, \text{variance} = 1000)$ for $i = 0, 1$. We retain the same hyperprior for the precision parameter τ as above. This results

in the posterior that accounts for preferential sampling,

$$\Pr(\boldsymbol{\gamma}, \tau, \boldsymbol{\beta} \mid \mathbf{g}, \mathbf{s}) \propto \Pr(\mathbf{g} \mid \mathbf{s}, \boldsymbol{\gamma}) \Pr(\mathbf{s} \mid \boldsymbol{\gamma}, \boldsymbol{\beta}) \Pr(\boldsymbol{\gamma} \mid \tau) \Pr(\tau) \Pr(\boldsymbol{\beta}),$$

where $\Pr(\mathbf{g} \mid \mathbf{s}, \boldsymbol{\gamma})$ is defined by Equation (3.1), but now we add conditioning on $\mathbf{s}$ explicitly. In the case where the density of sampling times $\mathbf{s}$ is functionally independent of the vector of log effective population sizes $\boldsymbol{\gamma}$, the posterior for $\mathbf{g}$ simplifies to the form it had in the previous section, because the likelihood for $\mathbf{s}$ becomes a constant in $\boldsymbol{\gamma}$. We incorporate our sampling model into an INLA framework similar to BNPR and name the implementation Bayesian nonparametric phylodynamic reconstruction with preferential sampling (BNPR-PS).

3.2.3 INLA framework

Here we present a brief outline of the INLA methodology [Rue et al., 2009] in the context of our BNPR and BNPR-PS implementations. We first examine BNPR as the simpler model. In the end, we intend to estimate the marginal posteriors of the precision hyperparameter $\Pr(\tau \mid \mathbf{g})$ and the latent points $\Pr(\gamma_i \mid \mathbf{g}), i = 1, \ldots, B$, most often focusing on the posterior medians and the end points of the 95% Bayesian credible intervals. We approximate the marginal of τ with

$$\widehat{\Pr}(\tau \mid \mathbf{g}) \propto \left. \frac{\Pr(\boldsymbol{\gamma}, \tau, \mathbf{g})}{\widehat{\Pr}_{\mathbf{G}}(\boldsymbol{\gamma} \mid \tau, \mathbf{g})} \right|_{\boldsymbol{\gamma} = \boldsymbol{\gamma}^*(\tau)},$$

where $\widehat{\Pr}_{\mathbf{G}}(\boldsymbol{\gamma} \mid \tau, \mathbf{g})$ is the Gaussian approximation generated from a Taylor expansion around $\boldsymbol{\gamma}^*(\tau)$, the mode of $\Pr(\boldsymbol{\gamma} \mid \tau, \mathbf{g})$ for a given τ. We can find $\boldsymbol{\gamma}^*(\tau)$ using the Newton-Raphson method.

Next, we need to approximate the distribution of γ_i conditional on τ. The simplest method of using the Gaussian approximations above can produce errors [Rue et al., 2009], so we briefly describe the use of nested Laplace approximations. The full implementation details can be found in [Rue et al., 2009]. We define

$$\widehat{\Pr}_{LA}(\gamma_i \mid \tau, \mathbf{g}) \propto \left. \frac{\Pr(\boldsymbol{\gamma}, \tau, \mathbf{g})}{\widehat{\Pr}_{\mathbf{GG}}(\boldsymbol{\gamma}_{-i} \mid \gamma_i, \tau, \mathbf{g})} \right|_{\boldsymbol{\gamma}_{-i} = \boldsymbol{\gamma}^*_{-i}},$$

where $\widehat{\Pr}_{\mathbf{GG}}(\boldsymbol{\gamma}_{-i} \mid \gamma_i, \tau, \mathbf{g})$ is a Gaussian approximation of $\Pr(\boldsymbol{\gamma}_{-i} \mid \gamma_i, \tau, \mathbf{g})$ obtained by a Taylor expansion around $\boldsymbol{\gamma}^*_{-i} = \mathrm{E}_{\mathbf{G}}(\boldsymbol{\gamma}_{-i} \mid \gamma_i, \tau, \mathbf{g})$, which is computed using $\widehat{\Pr}_{\mathbf{G}}(\boldsymbol{\gamma} \mid \tau, \mathbf{g})$. Finally, we normalize and combine the two approximations, then use numerical integration to calculate

$$\widehat{\Pr}(\gamma_i \mid \mathbf{g}) = \int \widehat{\Pr}(\gamma_i \mid \tau, \mathbf{g}) \widehat{\Pr}(\tau \mid \mathbf{g}) d\tau.$$

The outline for BNPR-PS is very similar. The approximate marginal of the hyperparameters is

$$\widehat{\Pr}(\tau, \boldsymbol{\beta} \mid \mathbf{g}, \mathbf{s}) \propto \left. \frac{\Pr(\boldsymbol{\gamma}, \tau, \boldsymbol{\beta}, \mathbf{g}, \mathbf{s})}{\widehat{\Pr}_{\mathbf{G}}(\boldsymbol{\gamma} \mid \tau, \boldsymbol{\beta}, \mathbf{g}, \mathbf{s})} \right|_{\boldsymbol{\gamma} = \boldsymbol{\gamma}^*(\tau, \boldsymbol{\beta})},$$

for similarly defined factors. We take advantage of an INLA extension byMartins et al. [2013] that allows for multiple likelihoods. The approximate distribution of γ_i conditional on $\tau, \boldsymbol{\beta}$ becomes

$$\widehat{\Pr}_{LA}(\gamma_i \mid \tau, \boldsymbol{\beta}, \mathbf{g}, \mathbf{s}) \propto \left. \frac{\Pr(\boldsymbol{\gamma}, \tau, \boldsymbol{\beta}, \mathbf{g}, \mathbf{s})}{\widehat{\Pr}_{\mathbf{GG}}(\boldsymbol{\gamma}_{-i} \mid \gamma_i, \tau, \boldsymbol{\beta}, \mathbf{g}, \mathbf{s})} \right|_{\boldsymbol{\gamma}_{-i} = \boldsymbol{\gamma}^*_{-i}},$$

and the final numerical integration is analogously more complex but still tractable, since we integrate over both τ and $\boldsymbol{\beta}$.

We use the R-INLA package [Rue et al., 2009, Martins et al., 2013] to perform the above calculations. We make INLA approximations of BNPR and BNPR-PS posteriors available, along with other phylodynamic tools, in the R package `phylodyn` which can be found at https://github.com/mdkarcher/phylodyn.

3.3 Results

3.3.1 Simulation study

We investigate estimating effective population size in the presence of preferential sampling via simulated data. First, we seek to show where and how the model misspecification resulting from ignoring preferential sampling manifests itself in terms of posterior median and Bayesian credible interval width estimation. Our second goal is to show what we gain by properly modeling preferential sampling.

Our primary set of simulation results use the family of seasonally-varying effective population size functions characterized by

$$N_{e,a,o}(t) = \begin{cases} 10 + 90/(1 + \exp\{a[3 - (t + o \pmod{12})]\}), & \text{if } t + o \pmod{12} \leq 6, \\ 10 + 90/(1 + \exp\{a[3 + (t + o \pmod{12}) - 12]\}), & \text{if } t + o \pmod{12} > 6. \end{cases} \tag{3.2}$$

For all of our experiments, the smoothness parameter $a = 2$ will be used. This family emulates a cyclical population time series with t in nominal months. The shape is loosely modeled after flu seasons, with o controlling which part of the year $t = 0$ represents ($o = 0, 3, 6$ emulates summer, spring, and winter, respectively). We simulate genealogies with varying tip sampling times using two sampling schedules. The uniform schedule distributes n sampling times uniformly throughout a given sampling interval. The proportional schedule distributes sampling times in the sampling interval according to an inhomogeneous Poisson process with intensity proportional to effective population size. The proportionality constant here is tuned to have an expected number of sampling times equal to n.

We explore the properties of our two methods using a Monte Carlo approach. To create a Monte Carlo iteration, we generate our sampling times according to their sampling schedules, then simulate our genealogies using coalescent theory via the rejection sampling method of [Palacios and Minin, 2013]. Given the genealogy and the samples, we infer the effective population time series, using BNPR and BNPR-PS to approximate grids of marginal posteriors. For each iteration, this gives us approximate estimates of the posterior median and quantiles at each point in the effective population size time series. In Figure 3.2, we see outputs from BNPR and BNPR-PS on the same example iteration.

Our first set of experiments is aimed at determining the extent of the bias introduced by unaccounted preferential sampling. With r Monte Carlo iterations, we take two approaches to locating model misspecification error—time interval analysis and pointwise analysis. For time interval analyses, we calculate summary statistics for a pre-specified time interval (a, b) and average them over the set of r simulation iterations. For pointwise analyses however,

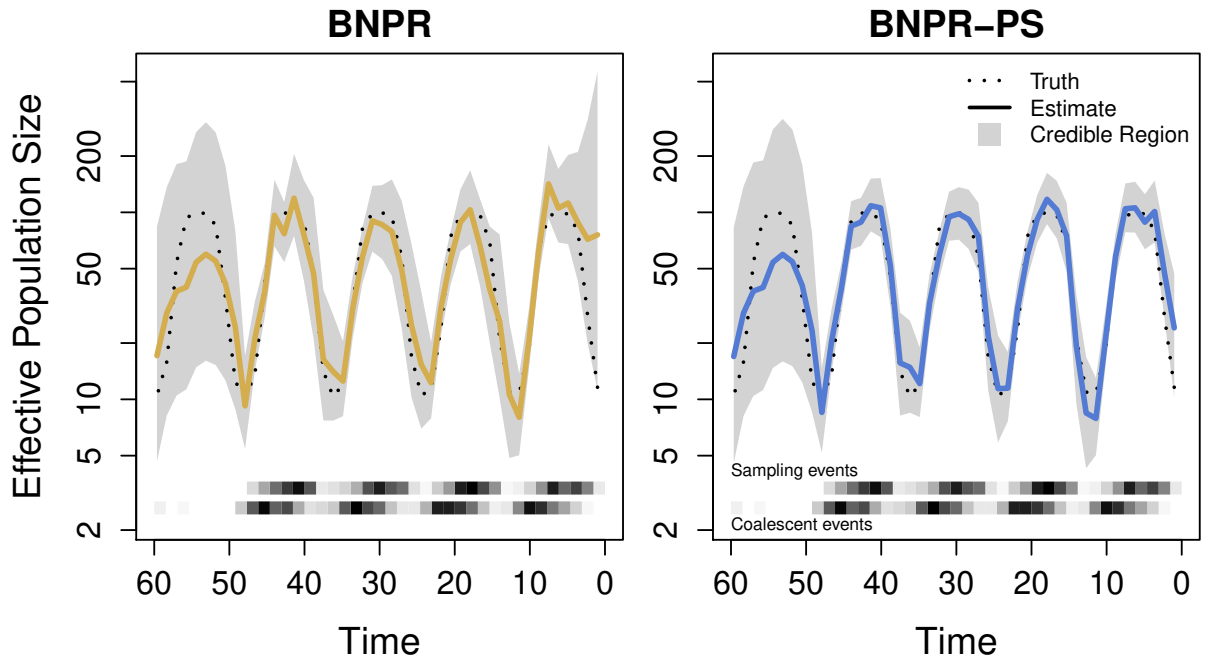

Figure 3.2: **Graphical representation of the output of a single genealogy simulation and integrated nested Laplace approximation (INLA) estimation.** The dotted black lines represent the true population trajectory. The solid colored lines represent the posterior median estimates, while the shaded regions represent the 95% credible regions. At bottom, the upper and lower heatmaps represent frequencies of sampling events and coalescent events, respectively. For this figure, we sampled individuals according to an inhomogeneous Poisson process with intensity proportional to effective population size $N_e(t)$. The plot on the left is generated by Bayesian nonparametric phylodynamic reconstruction (BNPR) and does not account for preferential sampling, while the plot on the right is generated by Bayesian nonparametric phylodynamic reconstruction with preferential sampling (BNPR-PS) and incorporates our sampling time model. Time is in months.

we consider the time series of point estimates from each iteration, and then on a pointwise basis we calculate aggregate point estimates and confidence intervals.

Our time interval summary statistics are *mean relative deviation*,

$$\text{MRD} = \frac{1}{r}\sum_{i=1}^{r}\left[\frac{1}{b-a}\int_a^b \frac{|\hat{N}_i^{\gamma}(t) - N^{\gamma}(t)|}{N^{\gamma}(t)}dt\right],$$

mean relative width of the 95% Bayesian credible intervals,

$$\text{MRW} = \frac{1}{r}\sum_{i=1}^{r}\left[\frac{1}{b-a}\int_a^b \frac{\hat{N}_{i,0.975}^{\gamma}(t) - \hat{N}_{i,0.025}^{\gamma}(t)}{N^{\gamma}(t)}dt\right],$$

where $N^{\gamma}(t)$ is the discretized true effective population size trajectory, $\hat{N}_i^{\gamma}(t)$ is the estimated posterior median of effective population sizes for iteration i, and $\hat{N}_{i,q}^{\gamma}(t)$ is the estimated qth posterior quantile for iteration i. We also look at *mean envelope*, ME, the proportion of grid points where the credible interval contains the true trajectory, averaged over all grid points contained in $[a, b]$ across all Monte Carlo iterations.

For a given grid of time points $\{t_j\}_{j=0}^{k}$, pointwise analysis computes the means of pointwise posterior medians,

$$\text{mpmedian}(t_j) = \frac{1}{r}\sum_{i=1}^{r}\hat{N}_{i,0.5}^{\gamma}(t_j), \text{ for } j = 0, \ldots, k,$$

pointwise mean relative errors,

$$\text{mre}(t_j) = \frac{1}{r}\sum_{i=1}^{r}\frac{\hat{N}_{i,0.5}^{\gamma}(t_j) - N^{\gamma}(t_j)}{N^{\gamma}(t_j)}, \text{ for } j = 0, \ldots, k,$$

and a sequence of mean relative widths of the pointwise Bayesian credible intervals,

$$\text{mrw}(t_j) = \frac{1}{r}\sum_{i=1}^{r}\frac{\hat{N}_{i,0.975}^{\gamma}(t_j) - \hat{N}_{i,0.025}^{\gamma}(t_j)}{N^{\gamma}(t_j)}, \text{ for } j = 0, \ldots, k.$$

We choose grid size $k = 100$, number of simulation iterations $r = 512$, and expected number of lineages per genealogy $n = 500$. We choose the sampling interval $[0, 48]$ for all simulations.

Ignoring preferential sampling

Table 3.1 shows the averaged time interval summary statistics for simulated genealogies under uniform and proportional schedules for the time intervals $(0, 6)$ and $(6, 48)$. Genealogies were simulated assuming effective population size function $N_{e,2,0}(t)$ defined in Equation 3.2. We show the time interval summary statistics for inferred effective population sizes both ignoring and considering preferential sampling. Ignoring preferential sampling (Table 1 under BNPR), we note a 17% increase in mean relative deviation from uniform to proportional schedules, as well as a 20% increase in mean relative width of Bayesian credible intervals for $(6, 48)$. For $(0, 6)$ the increase is more stark. We see a 407% increase in mean relative deviation from uniform to proportional, and a 799% increase in mean relative width of Bayesian credible intervals. Under proportional sampling, we see a notable increase in mean envelope, ME, on the $(0, 6)$ interval. All other cases show BNPR and BNPR-PS having ME within Monte Carlo error. These results confirm that ignoring preferential sampling affects both bias and variance of Bayesian nonparametric estimators of the effective populations size.

Figure 3.3 (solid lines) compares the average pointwise statistics for the uniform and proportional sampling schedules. Note the marked increase in mean relative error in several locations. We also see much larger mean relative widths in the same locations. Figure 3.4 compares the time interval statistics for the uniform and proportional sampling schedules, and we see increases in mean relative deviation and mean relative width. We conjecture that these features are representative of the model misspecification error that we would expect while sampling sequences/lineages preferentially in time but not accounting for it in the model.

Accounting for preferential sampling

Table 3.1 under BNPR-PS shows the time interval statistics for the sampling-aware model. For interval $(6, 48)$, mean relative deviation decreases by 23% versus BNPR under proportional sampling, while mean relative width of Bayesian credible intervals decreases by a

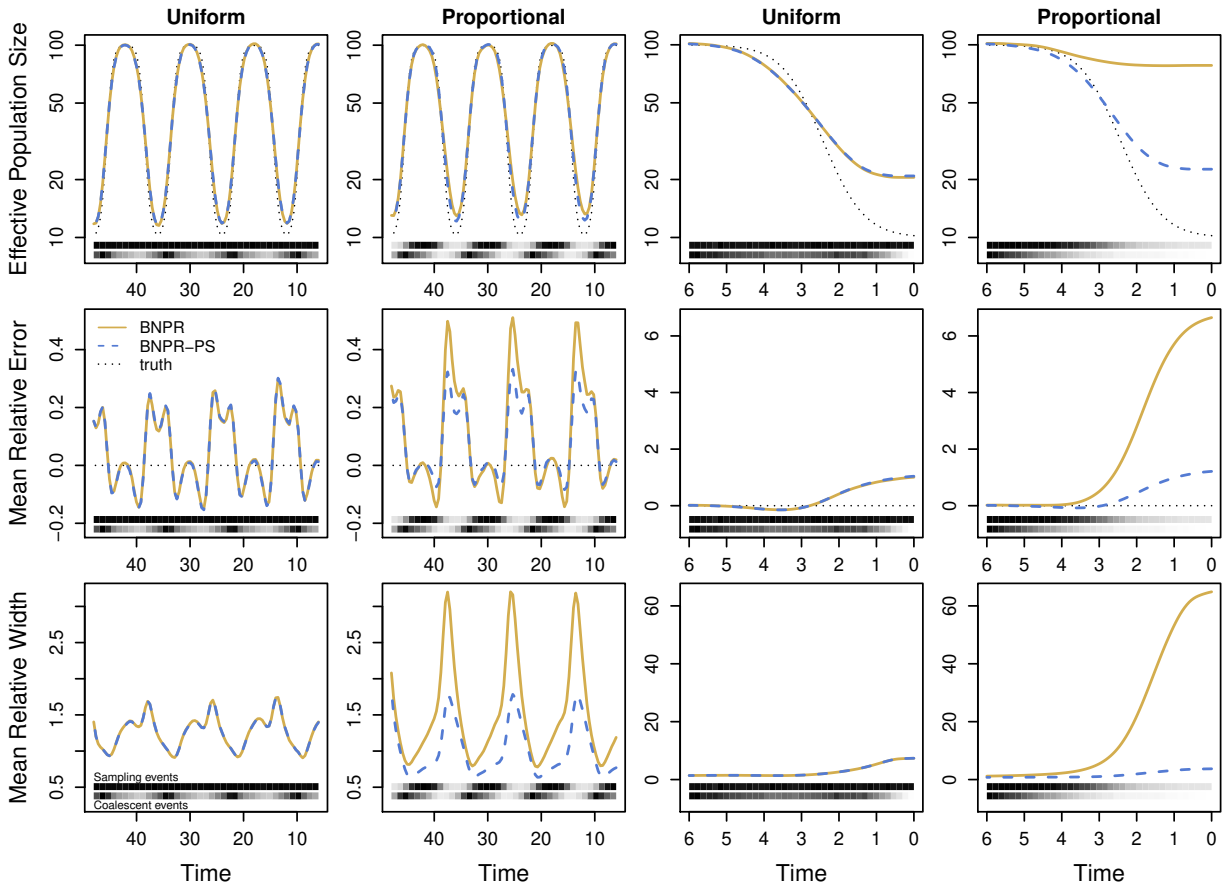

Figure 3.3: **Comparison of pointwise statistics.** Dotted black lines represent the truth, where applicable. Solid yellow lines represent the conditional method BNPR (ignoring preferential sampling), while dashed blue lines represent the sampling-aware method BNPR-PS (accounting for preferential sampling). The first row shows true and estimated effective population sizes, the second shows mean relative error, while the third shows mean relative width of the 95% Bayesian credible interval. The left two columns show the interval $(6, 48)$ where both models perform at their best. The right two columns show $(0, 6)$, where BNPR-PS performs significantly better. At the bottom of each plot, the distribution of sampling events (above) and coalescent events (below) are shown as heat maps. Time is in months.

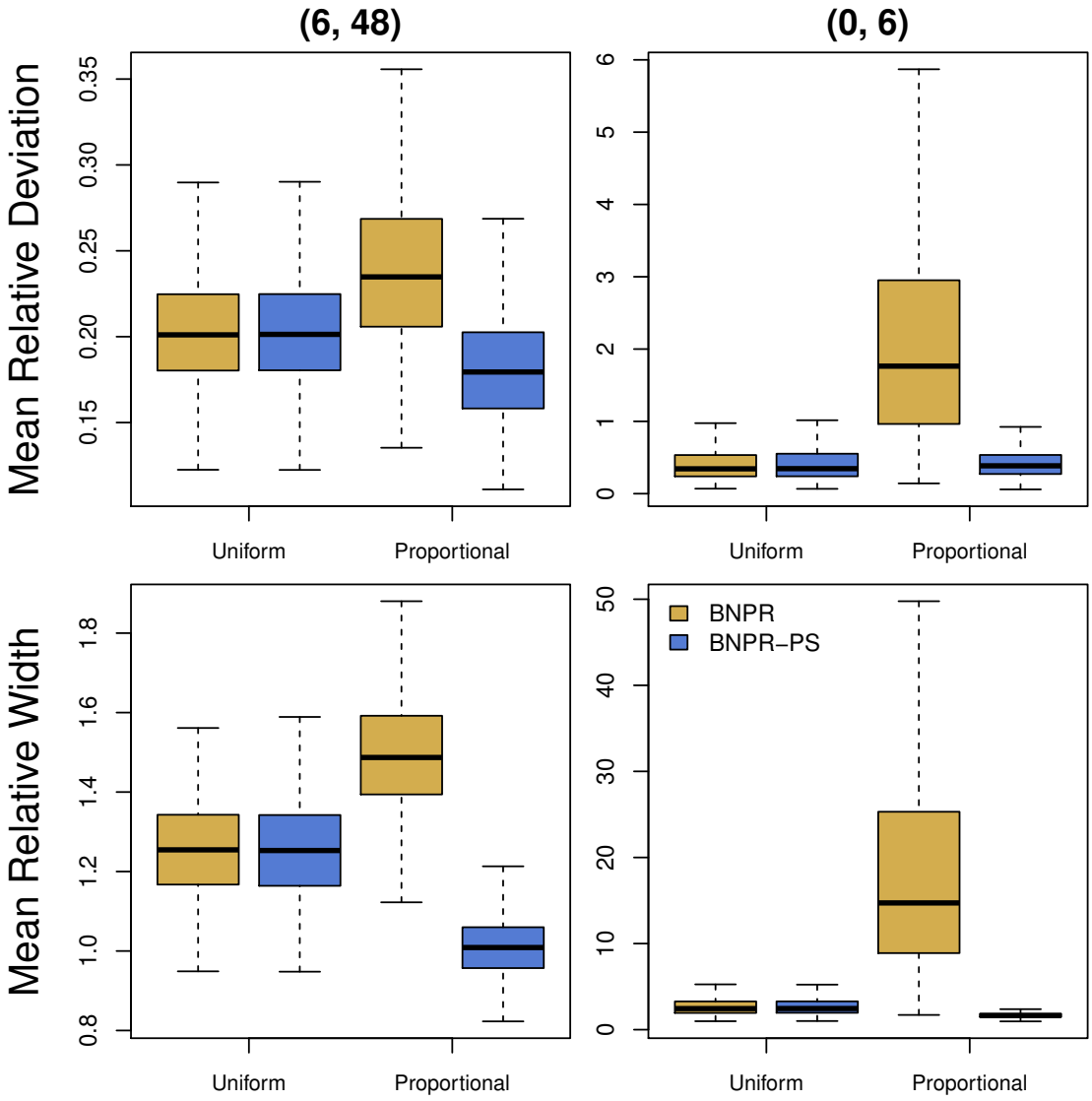

Figure 3.4: **Comparison of time interval statistics.** Within each plot, we apply BNPR and BNPR-PS to sampling times generated according to a Uniform distribution on the left and proportionally to effective population size on the right. In the left column of plots, we examine the interval $(6, 48)$ where the performances of both models are comparable. In the right column, we show $(0, 6)$, and note that BNPR-PS performs well, while BNPR performs considerably worse.

Table 3.1: **Averaged time interval summary statistics for BNPR and BNPR-PS.**

	Uniform—$(6, 48)$		Proportional—$(6, 48)$		Uniform—$(0, 6)$		Proportional—$(0, 6)$	
	BNPR	BNPR-PS	BNPR	BNPR-PS	BNPR	BNPR-PS	BNPR	BNPR-PS
MRD	0.205	0.205	0.239	0.183	0.430	0.436	2.181	0.432
MRW	1.255	1.255	1.500	1.008	2.816	2.816	19.681	1.682
ME	0.965	0.964	0.962	0.957	0.950	0.948	0.833	0.898

We compare the performance of the models under two different sampling distributions. Uniform distributes sampling times according to a uniform distribution on the interval $(0, 48)$, while proportional distributes sampling times according to a inhomogeneous Poisson process with intensity proportional to effective population size $N_e(t)$ on the same interval. We examine the statistics mean relative deviation (MRD), mean relative width of the 95% Bayesian credible interval (MRW), and mean envelope (ME). We average over statistics over the interval $(6, 48)$ where both methods perform well and over the most recent interval $(0, 6)$ where BNPR-PS performs considerably better.

larger margin of 33%. For interval $(0, 6)$ mean relative deviation and mean relative width decrease by 80% and 91%, respectively. Under uniform sampling, BNPR-PS performs almost identically to BNPR for both intervals.

Figure 3.3 (dashed lines) compares the average pointwise statistics for the uniform and proportional sampling schedules under BNPR-PS. We see that BNPR-PS does not experience the increase in relative error that BNPR experiences under preferential sampling. The plots also show an improvement in mean relative width of Bayesian credible intervals under preferential sampling due to the additional information available. Figure 3.4 compares the time interval statistics for the uniform and proportional sampling schedules under BNPR-PS, and shows improvements in mean relative deviation and mean relative width.

Negative control simulations

In the previous sections, we find a pattern of increased mean relative deviation and mean relative width while using a conditional model in a scenario involving preferential sampling. However, it is possible that this behavior of the conditional, state-of-the-art coalescent model can be seen under other simulation scenarios that cluster sampling times, even when such clustering has no relationship to the effective population size fluctuations. To test this assertion, we design a pair of negative control simulation studies to have random clusters of sampling times, but no preferential sampling.

First, we apply BNPR to genealogies generated from randomly constructed piecewise constant sampling intensity functions, independent of effective population size; see Appendix. We see some examples of increased mean relative error, but nothing as consistent nor prevalent as in the preferential sampling case above (see Figure 3.2 and 3.3). Similarly, we see increased mean relative width in several locations, but decreased widths in others. Second, we apply BNPR to genealogies generated from Gaussian process evaluations (subsampled for relatively similar shapes and number of peaks and troughs to the true population trajectory); see Appendix. This model has shape characteristics closer to the population trajectory since we are sampling from a Gaussian process. Despite the similar shapes, we see fewer increases in mean relative width and smaller increases in mean relative width. We conclude that unaccounted preferential sampling produces markedly more error more consistently than the negative control cases.

We also apply BNPR-PS to the same scenarios as above. BNPR-PS's performance suffers significantly due to both scenarios violating its fundamental assumption of a fixed relationship between effective population size and sampling intensity. We see BNPR-PS performs worst locally when there is a nearly fixed relationship which is suddenly reversed in a small time interval.

Parametric simulations

Finally, we also explored model misspecification in a correctly-specified parametric context. We simulated 100,000 genealogies from the coalescent with an exponential effective population size trajectory $N_e(t) = \exp(a + bt)$, under uniform and proportional sampling schedules. We applied an exponential growth/decline parametric maximum likelihood method and summarized the results in the Appendix. In both uniform and preferential sampling we see small, but comparable biases in estimates of parameter a. However, estimates of the exponential growth rate parameter b have no detectable bias under uniform sampling, but have small but significant bias under preferential sampling. This verifies that ignoring preferential sampling causes systematic bias, perhaps of small magnitude, in maximum likelihood phylodynamic estimation even under a simple low-dimensional parametric model.

3.3.2 Case studies

New York influenza

We base our first case study on a subset of the data from [Rambaut et al., 2008], also analyzed byPalacios and Minin [2013]. We focus on the 709 hemagglutinin gene sequences of H3N2 human influenza type A obtained from the National Center for Biotechnology Information (NCBI) Influenza Virus Sequence Database for years 1992 through 2005 from New York State. We align the sequences using the software MUSCLE [Edgar, 2004], and infer a maximum clade credibility genealogy using the software BEAST [Drummond et al., 2012]. We infer the genealogy branch lengths in units of years using a strict molecular clock, a constant effective population size prior, and an HKY substitution model with the first two nucleotides of a codon sharing the same estimated transition matrix, while the third nucleotide's transition matrix is estimated separately. We then apply our two algorithms to the estimated genealogy.

We find that BNPR produces results in line with previous analyses of this dataset, showing a characteristic uncertainty around the flu seasons of 2000-2001 and 2002-2003 (see

Figure 3.5). In contrast, BNPR-PS shows a marked improvement in the regularity of the reconstructed flu seasons, as well as thinner Bayesian credible intervals across the the whole observation interval. Estimations also improved during the unusual flu seasons of 2000-2001 and 2002-2003, consistent with these seasons being H1N1 dominant seasons instead of H3N2 dominant [Goldstein et al., 2011].

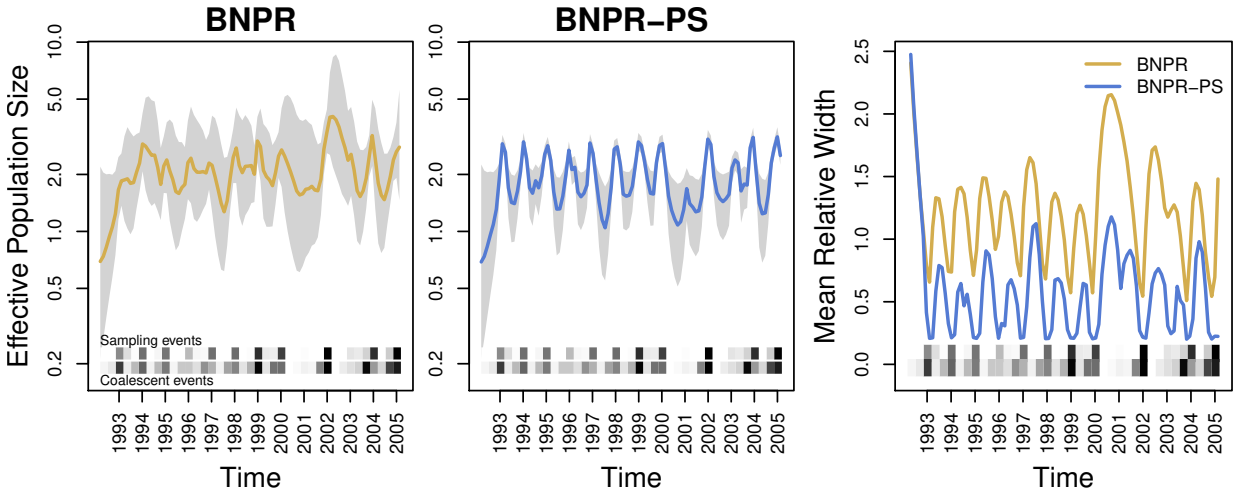

Figure 3.5: **BNPR and BNPR-PS models applied to the genealogy inferred from the New York influenza data [Rambaut et al., 2008].** Years mark January of the corresponding year. Note the correlation of higher effective population size $N^{\gamma}(t)$ with more intense sampling frequencies (darker regions in the Sampling events heatmap), suggesting preferential sampling. We see a marked improvement in discerning the seasonal influenza patterns and significantly thinner credible regions under BNPR-PS.

To compare performance of the BNPR and BNPR-PS models, we introduce an empirical measure of performance because we cannot know the true population size trajectory. We calculate the time interval and pointwise *empirical mean relative width* (EMRW) of the 95% Bayesian credible intervals,

$$\text{EMRW} = \frac{1}{r}\sum_{i=1}^{r}\left[\frac{1}{b-a}\int_a^b \frac{\hat{N}^{\gamma}_{i,0.975}(t) - \hat{N}^{\gamma}_{i,0.025}(t)}{\hat{N}^{\gamma}_i(t)}dt\right], \text{ for } j = 0, \ldots, k,$$

and

$$emrw(t_j) = \frac{1}{r}\sum_{i=1}^{r}\frac{\hat{N}^{\gamma}_{i,0.975}(t_j) - \hat{N}^{\gamma}_{i,0.025}(t_j)}{\hat{N}^{\gamma}_{i}(t_j)}, \text{ for } j = 0, \ldots, k.$$

Table 3.2 shows a very high value of β_1 for this dataset, suggesting a strong pattern of preferential sampling, and accordingly we see a marked improvement of BNPR-PS model over its BNPR counterpart in estimation precision as measured by the Bayesian credible interval widths (EMRW).

Table 3.2: **Case studies' empirical mean relative widths and Bayesian credible intervals of β_0 and β_1.**

		EMRW		95% credible		95% credible	
	n	BNPR	BNPR-PS	interval of β_0		interval of β_1	
New York influenza	709	1.23	0.58	(-47.4,	-30.3)	(5.88,	10.23)
Regional influenza							
USA & Canada	520	1.83	1.11	(-3.02,	-0.79)	(2.52,	4.05)
South America	191	0.86	0.91	(-4.21,	-0.42)	(3.27,	7.52)
Europe	361	1.73	0.96	(-6.61,	-2.44)	(3.68,	6.88)
India	233	1.79	1.30	(-2.18,	0.50)	(2.34,	4.78)
Japan & Korea	444	1.82	1.09	(-2.23,	-0.25)	(2.35,	3.76)
North China	384	1.80	1.09	(-2.63,	-0.27)	(2.22,	3.89)
South China	528	1.27	0.78	(-1.05,	1.00)	(1.68,	3.23)
Southeast Asia	494	0.99	0.54	(-7.93,	-2.55)	(4.39,	8.86)
Oceania	461	1.53	0.88	(-1.51,	0.43)	(2.71,	4.52)

The regional influenza dataset is broken down into world regions. In all but one region, we see improvements, or at worst near-equality, in empirical mean relative width (EMRW) using BNPR-PS over BNPR.

Regional influenza

Zinder et al. examine world-wide seasonal patterns of migration of H3N2 influenza across the regions of the world [Zinder et al., 2014]. They also examine different seasonal incidence patterns, with tropical regions having a relatively flat incident rate throughout the year, while temperate regions show larger seasonal variation with higher incidence in winter months. In order to explore the effects of seasonality on preferential sampling, we examine the regions separately. We align the sequences using the software MUSCLE [Edgar, 2004], and infer a maximum clade credibility genealogy using the software BEAST [Drummond et al., 2012]. We infer the genealogy branch lengths in units of years using a strict molecular clock, a constant effective population size prior, and an HKY substitution model with the first two nucleotides of a codon sharing the same estimated transition matrix, while the third nucleotide's transition matrix is estimated separately. We then apply our two algorithms to the estimated genealogy.

We find that none of the regions contain 0 in their β_1 Bayesian credible interval (see Table 3.2), suggesting a relationship between effective population size and sampling frequency. Across all regions except South America, we see improvements of the BNPR-PS model over the BNPR model in estimation precision (EMRW). We examine three of the regions more closely in Figure 3.6 and 3.7 and the remaining six regions in the appendix. We see noticeable improvements in the relative widths of the Bayesian credible intervals. We also see more pronounced seasonality in the estimated effective population size trajectories produced by BNPR-PS. The USA/Canada region shows the expected seasonal peak in January-February, while the Oceania region shows the same in July-September. South China shows less seasonality overall, but BNPR-PS shows a more pronounced August peak despite the region being in the northern hemisphere. This is, however, in line with previous findings, most likely due to southern China's more tropical climate [Shu et al., 2010].

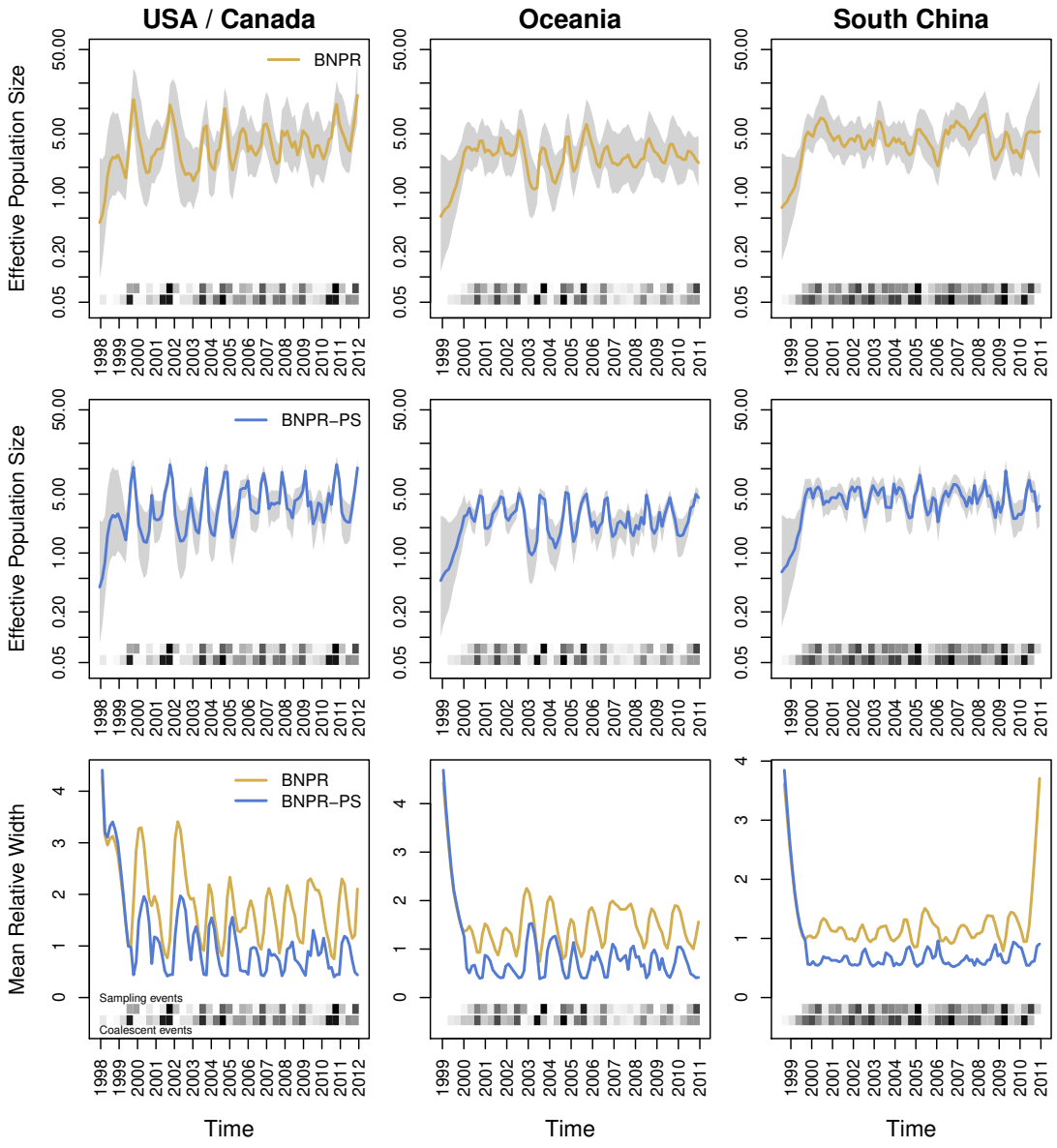

Figure 3.6: **BNPR and BNPR-PS models applied to the genealogies inferred from the regional influenza example [Zinder et al., 2014].** We see moderate correlation between effective population size $N^{\gamma}(t)$ and sampling frequencies in the data (Table 3.2). We see improvements in Bayesian credible interval widths, and BNPR-PS performs as well or better than BNPR everywhere in these examples.

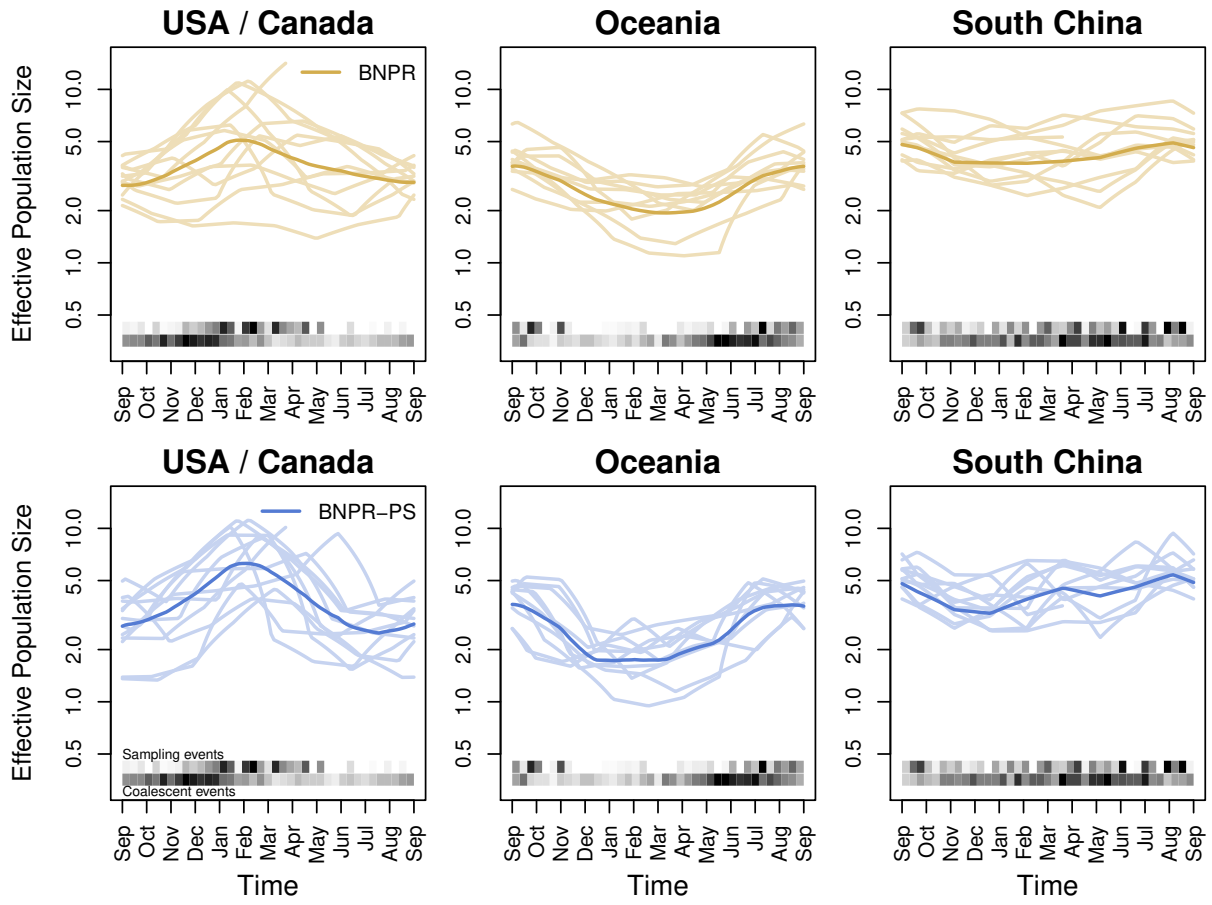

Figure 3.7: **Seasonality in regional influenza.** BNPR and BNPR-PS models applied to the genealogies inferred from the regional influenza example with years overlaid. We see more pronounced seasonality in the BNPR-PS plots.

3.4 Discussion

Researchers who study measurably evolving populations [Drummond et al., 2003], such as viruses, can inadvertently or purposefully preferentially select sequences in accordance to the changes in size of the population of interest. Failing to account for such an ascertainment bias can compromise the statistical properties of phylodynamic inference. Our simulation study shows that the effect of preferential sampling is particularly severe when the effective population size is decreasing. We propose an extension to the state-of-the-art in Gaussian process-based Bayesian phylodynamic methods, in which we assume that sampling times *a priori* follow an inhomogeneous Poisson process with intensity proportional to a power of the effective population size. This model extension eliminates the systematic estimation bias resulting from having unrecognized preferential sampling, and also gives us better population size estimates by incorporating sampling times as an additional source of information.

Applied to the real-world examples, our method produces improvements over the state-of-the-art. We see significantly improved precision, as well as more realistic estimation of seasonal variation of influenza diversity. In the presence of weaker preferential sampling, as in some of the regional influenza examples, we note that our method still performs better than the current state-of-the-art, with no loss of performance aside from a slightly longer computation time. In addition, by estimating β_1, the effect of population size on the log-intensity of sampling times, we gain the ability to quantify the strength of the preferential sampling relationship in the different regions. Such quantification is scientifically useful in infectious disease phylodynamics, because researchers may want to know whether frequency of sampling times can be used as a proxy for incidence.

One avenue of future exploration is to intentionally guarantee preferential sampling during the sequence data collection phase. For example, if an epidemiological study contains noisy incidence data, we can subsample sequences with intensity proportional to incidence and apply our sampling-aware BNPR-PS model to the resulting sequence data. Such a procedure will indirectly combine sequence and incidence data to estimate the effective number of

infections —a nontrivial task for the current methods [Rasmussen et al., 2011]. We contrast this to the approach of [Stack et al., 2010], which examined the effect of sampling infectious disease agent sequences in batches at different points in an epidemic's life-cycle compared to uniform and preferential sampling. They found that during epidemic declines their estimates had the largest mean squared error and benefited most in terms of this metric when samples were collected more frequently during the declines. This is consistent with our results, as we see the most error and widest credible intervals during effective population size declines. However, they did not consider the effect of the relationship between their proposed sampling intensity and population size trajectories on estimation of population dynamics—the primary goal of our work.

Our current implementation of the BNPR-PS model assumes a fixed, known genealogy. However, in practice, genealogies are inferred with inherent uncertainty from sequence data. We have found that point estimates produced by our method on the Regional influenza data are robust to genealogical uncertainty (see Regional influenza section in the Appendix), but a method that jointly estimates both genealogy and effective population is still necessary to properly assign uncertainty to population size estimates. One limitation of our method is that the INLA framework cannot be extended to include inference of genealogies. However, it should be straightforward to incorporate the core of our approach —the sampling times model— into an MCMC sampler that targets the joint posterior distribution of population size trajectory, genealogy of sampled sequences, and other parameters. We intend to implement such an MCMC approach in the software BEAST [Drummond et al., 2012].

The main goal of this manuscript is to point out the danger of ignoring preferential sampling in phylodynamics. Providing a solution to this problem, in the form of BNPR-PS model, remains our secondary goal, but we emphasize that much work is still needed to refine our proposed approach. The main weakness of our new model lies in its rigid parametric form of dependence between effective population size $N_e(t)$ and sequence sampling intensity $\lambda(t)$. In our negative control simulations we see that BNPR-PS performance suffers, possibly greatly, when this assumption of a fixed relationship between effective population size $N_e(t)$

and sampling intensity $\lambda(t)$ is violated. Similar results under model misspecification are observed by Volz and Frost in the context of birth-death-sampling models for phylodynamic inference [Stadler, 2010, Volz and Frost, 2014].

Sampling times model misspecification is most likely to occur if other variables besides effective population size $N_e(t)$ effect changes in the sampling intensity $\lambda(t)$. For instance, not accounting for a lag between $N_e(t)$ and $\lambda(t)$ may cause a severe model misspecification. Similarly, not accounting for increases in sampling intensity on longer time scales due to decreases in the cost of sequencing will bias our BNPR-PS estimation. We plan to address these issues by modeling our sampling intensity $\lambda(t)$ as a log-linear combination of effective sample size and other covariates:

$$\log[\lambda(t)] = \boldsymbol{\beta}^T \mathbf{c}(t),$$

where $\mathbf{c}(t)^T = (1, N_e(t), c_1(t), \ldots, c_p(t))$ and $c_i(t)$, $i = 1, \ldots, p$ are covariates of interest. For example, the cost of genome sequencing over time and lagged population size $N_e(t-l)$ are among prime candidates for covariates to be included into our BNPR-PS model. Another example of a promising time-varying sampling covariate is an indicator of 'outbreak' status, allowing for changes in sampling intensity during times of increased epidemiological oversight. We hope to explore these model extensions in our future research.

Chapter 4

EXTENDING THE MODEL: SEQUENCE DATA AND COVARIATES

4.1 Introduction

Phylodynamic inference—the study and estimation of population dynamics from genetic sequences—relies upon data sampled in a timeframe compatible with the evolutionary dynamics under question [Drummond et al., 2003]. One important class of phylodynamic methods seeks to estimate magnitudes and changes in a measure of genetic diversity called the *effective population size*, often considered proportional to the census population size [Wakeley and Sargsyan, 2009] or number of infections in epidemiological contexts [Frost and Volz, 2010]. One subtle and often ignored complication of phylodynamic inference occurs when there is a functional dependence between the effective population trajectory and the temporal frequency of collecting data samples, such as in case of sampling infectious disease agent genetic sequences with increasing urgency and intensity during a rising epidemic. This issue of *preferential sampling* is studied in depth by Karcher et al. [2016a] in the limited context of a known, fixed genealogy reconstructed from the genetic data, revealing that sampling protocols that (implicitly) depend on effective population size cause model misspecification bias in models that do not account for the possibility of preferential sampling. Here, we extend the work of Karcher et al. [2016a] and develop a Bayesian framework for accounting for preferential sampling during effective population size estimation directly from sequence data rather than from a fixed genealogy. We also propose a more flexible model for sequence sampling times that allows for inclusion of arbitrary time-dependent covariates and their interactions with the effective population size.

Methods for estimating effective population size from genealogical data and genetic se-

quence data have evolved from the earliest low dimensional parametric methods, such as constant population size [Griffiths and Tavaré, 1994b] and exponential growth models [Griffiths and Tavaré, 1994b, Drummond et al., 2002], to more flexible, nonparametric or highly parametric methods based on change-point models and Gaussian process smoothing [Drummond et al., 2005, Heled and Drummond, 2008, Minin et al., 2008, Palacios and Minin, 2013, 2012, Gill et al., 2013, 2016]. Most coalescent-based methods condition on the times of sequence sampling, rather than include these times into the model, leaving open the possibility of model misspecification if preferential sampling over time is in play. Volz and Frost [2014] and Karcher et al. [2016a] introduced coalescent models that include sampling times as random variables, whose distribution is allowed to depend on the effective population size. In particular, Karcher et al. [2016a] propose a method that models sampling times as an inhomogeneous Poisson process with log-intensity equal to an affine transformation of the log-transformed effective population size. In the presence of preferential sampling, this sampling-aware model demonstrates improved accuracy and precision compared to standard coalescent models due to eliminating an element of model misspecification and incorporating an additional source of information to estimate the effective population trajectory.

The main limitations of the approach of Karcher et al. [2016a] are a reliance on a fixed, known genealogy and lack of flexibility in the preferential sampling time model that currently does not allow the relationship between effective population size and sampling intensity to change over time. We address the issue of fixed-tree inference by implementing a preferential sampling time model in the popular phylodynamic Markov chain Monte Carlo (MCMC) software package BEAST [Drummond et al., 2012]. This allows us to perform inference directly from genetic sequence data, appropriately accounting for genealogical uncertainty, using a wide selection of molecular sequence evolution models and well tested phylogenetic MCMC transition kernels. Additionally, we implement a tuning parameter free elliptical slice sampling transition kernel [Murray et al., 2010] for high dimensional effective population size trajectory parameters, which allows us to update these parameters efficiently.

We also address the issue of an inflexible preferential sampling time model by incorpo-

rating time-varying covariates into the model. We model the sampling times as an inhomogeneous Poisson process with log-intensity equal to a linear combination of the log-effective population size and any number of functions of time. These functions can include products of covariates and the log-effective population size, referred to as *interaction covariates*. The addition of covariates into the sampling time model allows for incorporating additional sources of information into the relationship between effective population size and sampling intensity. One example of time-varying covariates includes an exponential growth function to account for a continuous decrease in sequencing costs that results in increased intensity of genetic data collection over time. In the context of endemic infectious disease surveillance, it is likely important to account for seasonality when modeling changes in genetic data sampling intensity, motivating inclusion of periodic functions as time varying covariates in the preferential sampling model.

We validate our methods first by simulating genealogies and sequence data and confirming that our methods successfully reconstruct the true effective population trajectories and true model parameters. We briefly simulate data in a fixed-tree context to demonstrate the fundamentals of incorporating covariates into the sampling time model and what bias is introduced by model misspecifications. We proceed to simulate genetic sequence data and demonstrate that our model successfully functions when we estimate effective population size trajectory and other parameters directly from sequence data. We also use simulations to test a combination of the two extensions of the preferential sampling model and work with covariates while sampling over genealogies during the MCMC. Finally, we use our method to analyze two real-world epidemiological datasets. We analyze a USA/Canada regional influenza dataset [Zinder et al., 2014] to determine if exponential growth of genetic sequencing or seasonal changes in sampling intensity are important to adjust for during effective population size reconstruction. We also analyze data from the recent Ebola outbreak in Western Africa to determine if preferential sampling has taken place and whether time-varying covariates or interaction covariates improve the phylodynamic inference.

4.2 Methods

4.2.1 Sequence Data and Substitution Model

Consider an *alignment* $\mathbf{y} = \{y_{ij}\}$, $i = 1, \ldots, n$, $j = 1, \ldots, l$, of n genetic sequences across l sites, collected from a well-mixed population at *sampling times*

$$\mathbf{s} = \{s_i\}_{i=1}^{n},\ s_1 \geq \ldots \geq s_n = 0.$$

The following example, shows an alignment of $n = 5$ samples across $l = 10$ sites, sampled at distinct times between time 7 and time 0—with time understood to be time *before* the latest sample:

$$\mathbf{y}_1 = ACATGAGCTT,\ s_1 = 7$$
$$\mathbf{y}_2 = ACTTGACCTG,\ s_2 = 4$$
$$\mathbf{y}_3 = TCTTGACCTT,\ s_3 = 2$$
$$\mathbf{y}_4 = AAATCTGCGT,\ s_4 = 1$$
$$\mathbf{y}_5 = AGATGTGCAT,\ s_5 = 0.$$

All of the individual sequences share a common ancestry, which can be represented by a bifurcating tree called a *genealogy*—illustrated in Figure 4.1.

We assume that sequence data $\mathbf{y}$ are generated by a continuous time Markov chain (CTMC) *substitution model* that models the evolution of the genetic sequence along the genealogy $\mathbf{g}$. According to this model, alignment sites are independent and identically distributed, with a transition matrix $\boldsymbol{\theta}$ controlling the CTMC substitution rates between the different nucleotide bases. Some relaxation of these assumptions is possible [Shapiro et al., 2005]. Different substitution models are then defined by different parameterizations of $\boldsymbol{\theta}$ [Hein et al., 2004]. It is simple to simulate from these models, and we can efficiently compute the probability of the observed sequence data $\mathbf{y}$

$$\Pr(\mathbf{y} \mid \mathbf{g}, \boldsymbol{\theta})$$

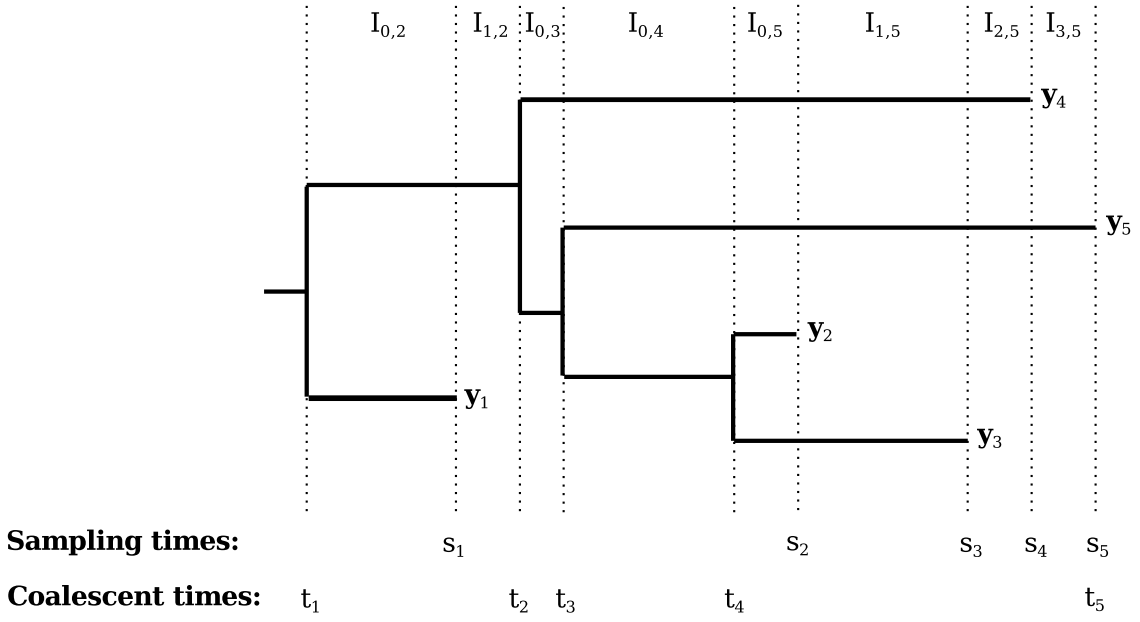

Figure 4.1: **Illustration of an example heterochronous genealogy with $n = 5$ lineages.** Sampling times $s_1, \ldots, s_5$ and coalescent times $t_1, \ldots, t_5$ are marked below the genealogy, and sequence data $\mathbf{y}_1, \ldots, \mathbf{y}_5$ are marked at their corresponding tips.

using Felsenstein's pruning algorithm [Felsenstein, 1973, 1981].

4.2.2 The Coalescent

Recall that we assume that the n sampled sequences share a common ancestry, which can be represented by a bifurcating tree called a *genealogy*—illustrated in Figure 4.1. The branching events of the tree $\mathbf{g} = \{t_i\}_{i=1}^{n-1}, t_1 > \ldots > t_{n-1}$ (with t greater the farther *back* in time an event occurs) are called *coalescent events*. The times associated with the tips of the tree $\mathbf{s} = \{s_i\}_{i=1}^{n}$, $s_1 \geq \ldots \geq s_n$ are called *sampling times* or *sampling events*. If all of the sampling events are simultaneous, the sampling is called *isochronous*. Assuming that the population evolves according to the Wright-Fisher model of genetic drift and that the size of the population is not changing, Kingman [1982] derived a probability density for an isochronous genealogy, where the population size plays the role of a parameter of this density.

Since the Wright-Fisher model is a simplified representation of the evolutionary process, the above parameter is called the *effective population size*, N_e. Later extensions to the coalescent model incorporated variable effective population size $N_e(t)$ [Griffiths and Tavaré, 1994b] and the ability to evaluate densities of genealogies with *heterochronously* sampled tips—genealogies with non-simultaneous sampling times [Felsenstein and Rodrigo, 1999].

Given sampling times $\mathbf{s}$ and effective population size trajectory $N_e(t)$, we would like to define the probability density for a particular genealogy $\mathbf{g}$. We use the term *active lineages*, $n(t)$, to refer to the difference between the number of samples taken and the number of coalescent events occurred between times 0 and t. To illustrate, in Figure 4.1, $n(t)$ can be seen as the number of horizontal lines that a vertical line at time t will cross. Suppose we partition the interval (s_n, t_1), from the most recent sampling event to the *time to most recent common ancestor* (TMRCA), into intervals $I_{i,k}$ with constant numbers of active lineages. Let $\lambda_c(t) = \binom{n(t)}{2}/N_e(t)$. Then the coalescent density evaluated at genealogy $\mathbf{g}$ is

$$\Pr(\mathbf{g} \mid N_e(t), \mathbf{s}) \propto \prod_{k=2}^{n} \left[\lambda_c(t_{k-1}) \exp\left(- \int_{I_{i,k}} \lambda_c(t) dt \right) \right]. \tag{4.1}$$

4.2.3 Population Size Prior

Note that without further assumptions the effective population size trajectory function $N_e(t)$ is infinite-dimensional, so inference about $N_e(t)$ without some manner of constraint is intractable. A number of approaches, reviewed in the Introduction, have been suggested to address this fact. Here, we take a regular grid approach that was used before in multiple studies [Palacios and Minin, 2012, Gill et al., 2013, 2016, Karcher et al., 2016a]. To review, we approximate $N_e(t)$ with a piecewise constant function, $N_{\boldsymbol{\gamma}}(t) = \exp[\boldsymbol{\gamma}(t)]$, where $\boldsymbol{\gamma}(t) = \sum_{i=1}^{p} \gamma_i 1_{\{t \in J_i\}}$ and $J_1, \ldots, J_p$ are consecutive time intervals of equal length. In contexts where the genealogy is known, we choose intervals that perfectly cover the interval between the TMRCA and the latest sample. However, in contexts where the genealogy is estimated from sequence data, the TMRCA is not necessarily fixed. To address this, we choose equal intervals that extend to a fixed point in time and append an additional interval

that extends from that point infinitely back in time. This allows us to estimate the effective population trajectory with user-defined resolution over a window that extends back in time as far as the user chooses. The choice of the end point of the grid is up to the user, but it is advisable to choose a point that is farther back in time than an *a priori* estimate of the TMRCA in order to extend the high-resolution grid to cover the entire true genealogy.

The population size trajectory $N_{\boldsymbol{\gamma}}(t)$ is parameterized by a potentially high dimensional vector $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_p)$. We assume that *a priori* $\boldsymbol{\gamma}$ follows a first order Gaussian random walk prior with precision hyperparameter κ: $\gamma_i \mid \gamma_{i-1} \sim \mathcal{N}(\gamma_{i-1}, 1/\kappa)$ or equivalently $\gamma_i - \gamma_{i-1} \sim \mathcal{N}(0, 1/\kappa)$, for $i = 2, \ldots, p$. We use a Gaussian prior on the first element: $\gamma_1 \sim \mathcal{N}(0, \sigma_p^2)$. Finally, we assign a Gamma(α, β) hyperprior to κ.

4.2.4 Preferential Sampling Model with Covariates

Karcher et al. [2016a] model times at which sequences are collected as a Poisson point process with intensity $\lambda_s(t)$ equal to a log-linear function of the effective population size. Although it is realistic to assume that the larger the population, the more members of the population gets sequenced, other factors may influence the distribution of sequence sampling times. For instance, decreasing sequencing costs may result in increasing sequence sampling intensity even if the population size remains constant. We propose an extension to the sampling model that allows for the incorporation of time-varying covariates as additional sources of information. Suppose we have one or more real-valued functions, $\mathcal{F} = \{f_2(t), \ldots, f_m(t)\}$. We let

$$\log \lambda_s(t; \mathcal{F}) = \beta_0 + \beta_1 \gamma(t) + \beta_2 f_2(t) + \ldots + \beta_m f_m(t) + [\delta_2 f_2(t) + \ldots + \delta_m f_m(t)]\, \gamma(t), \quad (4.2)$$

where we may set any or all of the $\beta_2, \ldots, \beta_m$ or $\delta_2, \ldots, \delta_m$ to zero if we want to avoid modeling effects of certain covariates or their interactions with the log-population size. Notice that we reserve $f_1(t)$ for $\gamma(t) = \log[N_e(t)]$, which is the covariate that is always present in our model. We also point out that even though Equation (4.2) is written in continuous time, in practice we assume that both the sampling intensity $\lambda_s(t)$ and our time varying covariates

are piecewise constant, with changes occurring at the grid points specified in Subsection 4.2.3. We assign independent $\mathcal{N}(0, \sigma_s^2)$ priors for all components of the preferential sampling model parameter vector $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_m, \delta_2, \ldots, \delta_m)$.

4.2.5 Putting It All Together: Posterior Approximation with MCMC

Having specified all parts of our data generating model, we are now ready to define the posterior distribution of all unknown variables of interest:

$$\begin{aligned}\Pr(\mathbf{g}, \boldsymbol{\gamma}, \kappa, \boldsymbol{\beta}, \boldsymbol{\theta} \mid \mathbf{y}, \mathbf{s}, \mathcal{F}) \propto &\Pr(\mathbf{y} \mid \mathbf{g}, \boldsymbol{\theta}) \Pr(\mathbf{g} \mid \boldsymbol{\gamma}, \mathbf{s}) \Pr(\mathbf{s} \mid \boldsymbol{\gamma}, \boldsymbol{\beta}, \mathcal{F}) \Pr(\boldsymbol{\gamma} \mid \kappa) \\ &\times \Pr(\kappa) \Pr(\boldsymbol{\beta}) \Pr(\boldsymbol{\theta}),\end{aligned} \tag{4.3}$$

where all probabilities and probability densities on the righthand side of equation (4.3) are defined in the previous subsections. Figure 4.2 illustrates conditional dependencies of model parameters and data in a graph form.

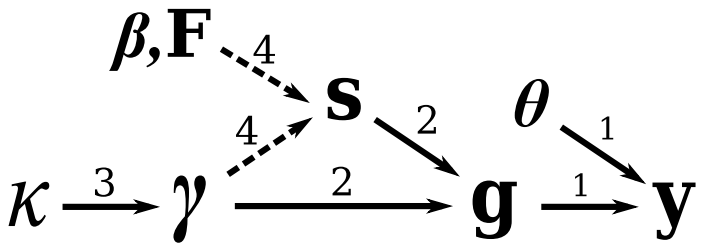

Figure 4.2: **Dependency graph for the phylodynamic model parameters and data.** Dependencies labeled 1 are explored in section 4.2.1, those labeled 2 are explored in section 4.2.2, those labeled 3 are explored in section 4.2.3, and those labeled 4 are explored in section 4.2.4. The dashed lines between $\boldsymbol{\gamma}, \boldsymbol{\beta}, \mathcal{F}$ and $\mathbf{s}$ represent preferential sampling.

When the distribution of sampling times does not depend on the effective population size trajectory (in our model, this happens when $\beta_1 = 0$ and $\delta_2 = \cdots = \delta_m = 0$), the posterior

takes the following form:

$$\Pr(\mathbf{g}, \boldsymbol{\gamma}, \kappa, \boldsymbol{\theta}, \boldsymbol{\beta} \mid \mathbf{y}, \mathbf{s}, \mathcal{F}) \propto \underbrace{\Pr(\mathbf{y} \mid \mathbf{g}, \boldsymbol{\theta}) \Pr(\mathbf{g} \mid \boldsymbol{\gamma}, \mathbf{s}) \Pr(\boldsymbol{\gamma}|\kappa) \Pr(\kappa) \Pr(\boldsymbol{\theta})}_{\propto \Pr(\mathbf{g}, \boldsymbol{\gamma}, \kappa, \boldsymbol{\theta} | \mathbf{y}, \mathbf{s})}$$
$$\times \underbrace{\Pr(\mathbf{s} \mid \boldsymbol{\beta}, \mathcal{F}) \Pr(\boldsymbol{\beta})}_{\propto \Pr(\boldsymbol{\beta} | \mathbf{s}, \mathcal{F})}.$$

The factorization above demonstrates that when $\boldsymbol{\gamma}$ is absent from the $\Pr(\mathbf{s} \mid \cdot)$ term, joint and separate estimations of effective population size parameters $\boldsymbol{\gamma}$ and preferential sampling model parameters $\boldsymbol{\theta}$ will yield identical results. Moreover, in this case estimation of sampling model parameters can be dropped from the analysis entirely, since typically these parameters would be considered nuisance. If we drop preferential sampling, our model specifications reduces to the Bayesian skygrid model of Gill et al. [2013], with the corresponding posterior:

$$\Pr(\mathbf{g}, \boldsymbol{\gamma}, \kappa, \boldsymbol{\theta} \mid \mathbf{y}, \mathbf{s}) \propto \Pr(\mathbf{y} \mid \mathbf{g}, \boldsymbol{\theta}) \Pr(\mathbf{g} \mid \boldsymbol{\gamma}, \mathbf{s}) \Pr(\boldsymbol{\gamma}|\kappa) \Pr(\kappa) \Pr(\boldsymbol{\theta}). \tag{4.4}$$

We approximate posteriors (4.3) and (4.4) by devising MCMC algorithms, implemented in the software package BEAST [Drummond et al., 2012], that target these distributions. We update model parameters in blocks — 1) genealogy $\mathbf{g}$, 2) substitution parameters $\boldsymbol{\theta}$, 3) population size parameters $\boldsymbol{\gamma}$, 4) random walk prior precision κ, 5) preferential sampling model parameters $\boldsymbol{\beta}$ — keeping parameters outside of the block fixed. We update the genealogy and substitution model parameters via the default BEAST Markov kernels. We update the log effective population latent field $\boldsymbol{\gamma}$ via an elliptical slice sampler (ESS) operator [Murray et al., 2010, Lan et al., 2015], which takes advantage of the Gaussian prior distribution of the latent field to perform efficient updates. Informally, it does this by sampling a set of parameter values from the prior and iteratively moving the values closer to the current values via elliptical interpolation if the coalescent likelihood falls below a random, but small, neighborhood of the current likelihood. Because the stepwise differences of the log effective population size trajectory, $\Delta\boldsymbol{\gamma}$, are modeled as independent Gaussians with precision κ, and because we give κ a Gamma(α, β) hyperprior, we update κ using a Normal-Gamma Gibbs

update kernel with full conditional

$$\kappa \mid \Delta\boldsymbol{\gamma} \sim \text{Gamma}\left[\alpha + \frac{p}{2}, \beta + \frac{1}{2}\sum_{i=2}^{p}\left(\gamma_i - \gamma_{i-1}\right)^2\right],$$

where p is the number of parameters in the latent field. For our sampling conditional model with posterior (4.4), we finish here and refer to the method as *ESS/BEAST*, abbreviated when appropriate as *ESS*. For our sampling-aware model with the posterior (4.3), we update components of the preferential sampling model parameter vector $\boldsymbol{\beta}$ with univariate Gaussian random walk Metropolis-Hastings kernels. We refer to the method as *SampESS/BEAST*, abbreviated when appropriate as *SampESS*.

4.3 Implementation

We implemented INLA-based, fixed-genealogy BNPR-PS method with simple covariates in R package `phylodyn` (`https://github.com/mdkarcher/phylodyn`). The package has also MCMC functionality that can handle inference from a fixed genealogy with simple and interaction sampling model covariates. See `phylodyn` vignettes for more details. MCMC for direct inference from sequence data is available in the development branch of software package `BEAST` (`https://github.com/beast-dev/beast-mcmc`). We provide examples of how to specify our preferential sampling models in `BEAST` xml files at `https://github.com/mdkarcher/phylodyn`.

4.4 Results

4.4.1 Simulation Study

Inference Assuming Fixed Genealogy

In Section 4.2.4, we proposed an extended sampling time model that incorporated time-varying covariates. We perform a simulation study to confirm the ability of our method to recover the true effective population trajectory and model coefficients with covariates

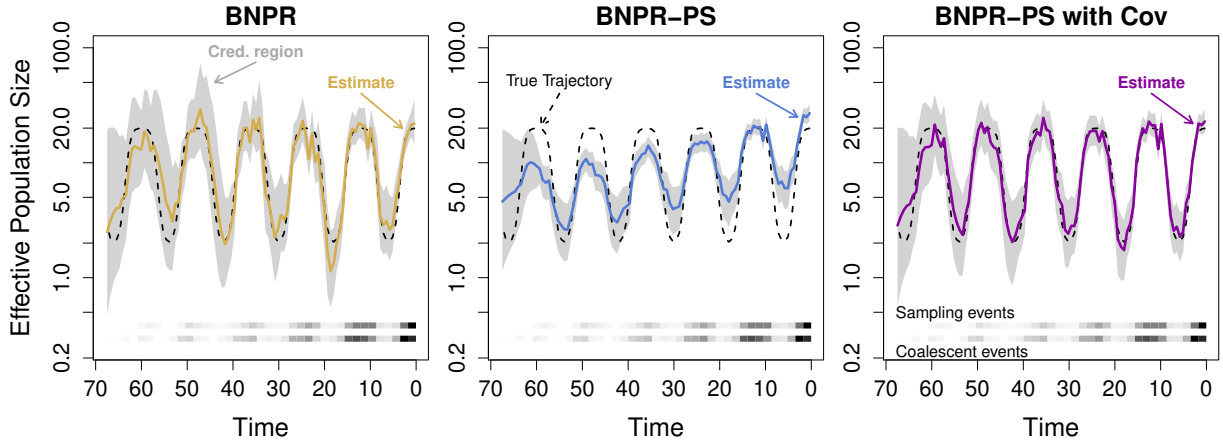

Figure 4.3: **Effective population size reconstruction for BNPR, BNPR-PS, and BNPR-PS with simple covariates.** The dotted black line represents the true effective population trajectory. The solid colored line represents the marginal posterior median effective population trajectory inferred by BNPR (yellow), BNPR-PS (blue), and BNPR-PS with simple covariates (purple), and the gray region represents the corresponding pointwise 95% credible intervals for the effective population trajectory. The log sampling intensity was $1.557 + \gamma(t) - 0.025t$.

affecting the sampling intensity. We begin here with fixed genealogies and move on to direct inference from sequence data in the next section.

We start with the inhomogeneous Poisson process sampling model with log-intensity as in Equation 4.2. If we restrict all βs and δs to be zero aside from β_0, the model collapses to homogenous Poisson process sampling (equivalently, uniformly sampling a Poisson number of points across the sampling interval). If we allow β_1 to be nonzero, the model becomes the sampling-aware model of Karcher et al. [2016a]. If we allow additional βs, each corresponding to a fixed function of time, to be nonzero (but not δs) we say that the model includes *simple* or *ordinary covariates*.

For computational efficiency in this simulation study, we build upon the methods of

Karcher et al. [2016a], including Bayesian Nonparametric Population Reconstruction (BNPR) which uses integrated nested Laplace approximation (INLA) to efficiently approximate the marginal posterior for fixed-genealogy data, and Bayesian Nonparametric Population Reconstruction with Preferential Sampling (BNPR-PS) which does the same but includes our sampling time model (without covariates). We incorporate our extended sampling time model into BNPR-PS, but due to constraints in the INLA R package, upon which BNPR-PS relies, we can only include simple covariates.

Because our sampling time model is an inhomogeneous Poisson process, it is straightforward to simulate sampling times. We use a *time-transformation* method [Çinlar, 1975, pages 98–99], which, informally, treats the waiting times between events as transformations of exponential waiting times based on the intensity function following the previous event. Because the coalescent likelihood is sufficiently similar to an inhomogeneous Poisson process, we can use a similar time-transformation technique to generate the coalescent events of simulated genealogies [Slatkin and Hudson, 1991]. We implement these methods for simulating sampling times and coalescent times in R package `phylodyn` [Karcher et al., 2017].

In Figure 4.3, we illustrate BNPR, BNPR-PS, and BNPR-PS with simple covariates applied to a single simulated genealogy with sampling events distributed according to log-intensity $1.56 + \gamma(t) - 0.05t$, resulting in 1013 tips, where $\gamma(t) = \log[N_{e,2,6}(t)]$ and $N_{e,a,o}(t)$ is a family of functions that approximate seasonal changes in effective population size, defined as follows:

$$N_{e,a,o}(t) = \begin{cases} 2 + 18/(1 + \exp\{a[3 - (t + o \pmod{12})]\}), & \text{if } t + o \pmod{12} \leq 6, \\ 2 + 18/(1 + \exp\{a[3 + (t + o \pmod{12}) - 12]\}), & \text{if } t + o \pmod{12} > 6. \end{cases} \tag{4.5}$$

We see that BNPR (the sampling conditional model) suffers from the kind of model misspecification induced bias illustrated in [Karcher et al., 2016a]. BNPR-PS with no additional covariates beyond $\gamma(t) = \log[N_e(t)]$, in contrast, suffers even more strongly from a misspecified sampling model. Table 4.1 shows that the model fails to correctly infer the coefficient of $\gamma(t)$. This illustrates the care one must take in choosing parameterizations of

Model	Coef	Q0.025	Median	Q0.975	Truth
$\{\gamma(t)\}$	$\gamma(t)$	1.67	1.99	2.34	1.0
$\{\gamma(t), -t\}$	$\gamma(t)$	0.86	1.01	1.16	1.0
	$-t$	0.040	0.047	0.053	0.050

Table 4.1: **Summary of simulated fixed-tree data inference.** Posterior distribution quantile summaries for BNPR-PS with no covariates (model: $\{\gamma(t)\}$) and BNPR-PS with an ordinary covariate (model: $\{\gamma(t), -t\}$).

the sampling model. BNPR-PS with simple covariates, $\gamma(t)$ and $-t$, the correctly-specified model, produces a reconstruction of the effective population trajectory that is very close to the true trajectory used to simulate the data. Table 4.1 shows that the true values of the sampling model coefficients are within 95% Bayesian credible intervals produced by our inference method with the correctly specified model.

Direct Inference from Sequence Data

We simulate several genealogies and DNA sequences from different sampling scenarios in order to evaluate how well our population reconstruction and parameter inference performs. Given a sampling model and, optionally, an effective population size trajectory, we generate sampling times within a *sampling window*. We generate sampling and coalescent times for a genealogy using the same time-transformation methods as for our fixed-tree simulations. We simulate the topology of the genealogy by proceeding backward in time, adding an active lineage at each sampling time and joining a pair of active lineages uniformly at random at each coalescent event. We provide an implementation of this tree-topology simulation method in `phylodyn`. We generate simulated sequence alignments using the software SeqGen [Rambaut and Grass, 1997], using the Jukes-Cantor 1969 (JC69) [Jukes et al., 1969] substitution model. We set the substitution rate to produce an expected 0.9 mutations per site, in order to produce a sequence alignment with many sites having one mutation, and some sites having

zero or multiple mutations. For all of our simulations, we use the same seasonal effective population trajectory, $N_{e,2,6}(t)$, as for our fixed-tree simulations.

First, we simulate a genealogy with 200 tips and sequence data with 1500 sites and uniform sampling times and apply both of our sampling-conditional methods. We apply the INLA-based fixed-tree BNPR from [Karcher et al., 2016a] to the true genealogy, and we apply the MCMC-based tree-sampling ESS/BEAST (specified above) to the sequence data. In Figure 4.4 (upper left), we compare the truth with the resulting pointwise posterior medians and credible intervals. The two methods' results are mutually consistent, with additional uncertainty in the tree-sampling method (visible in the wider credible intervals) due to having to estimate the genealogy jointly with other model parameters. We see similar results comparing BNPR-PS with SampESS/BEAST in Figure 4.4 (upper right), where we sample sequences (1500 sites) with sampling times generated from an inhomogeneous Poisson process with intensity proportional to effective population size (log-intensity $2.9 + \gamma(t)$) resulting in 170 samples and infer using a sampling model with log-intensity $\beta_0 + \beta_1\gamma(t)$. We also see similar results in Figure 4.4 (lower left), where we add time as an additional covariate and sample sequences (1500 sites) with log-intensity $3.35 + \gamma(t) - 0.5t$, resulting in 199 samples, and infer using a sampling model with log-intensity $\beta_0 + \beta_1\gamma(t) + \beta_2 \cdot (-t)$. Table 4.2 shows that SampESS does a reasonable job at reconstructing the true model coefficients, though the credible interval for $-t$ includes 0.

We also simulate a genealogy and sequence data (1500 sites) with log-intensity $1.89 + \gamma(t) + \gamma(t) \cdot 1_{t\in[0.5,1]}$, resulting in 210 samples. This produces an interval we refer to as a *sampling spike* which requires the use of an interaction covariate. Because of design limitations of the R implementation of INLA, we are limited in how we may implement interaction covariates in BNPR-PS. Therefore, in Figure 4.4 (lower right) we plot SampESS/BEAST with the correct interaction covariate (and a corresponding ordinary covariate) against BNPR-PS with no covariates. We see SampESS (with covariates) perform better than BNPR-PS (without covariates) at reconstructing the correct trajectory. We also see that our method, using the full covariate model, with log-intensity $\beta_0 + \beta_1\gamma(t) + \beta_2 \cdot 1_{t\in[0.5,1]} + \delta_2 \cdot \gamma(t) \cdot 1_{t\in[0.5,1]}$, produces

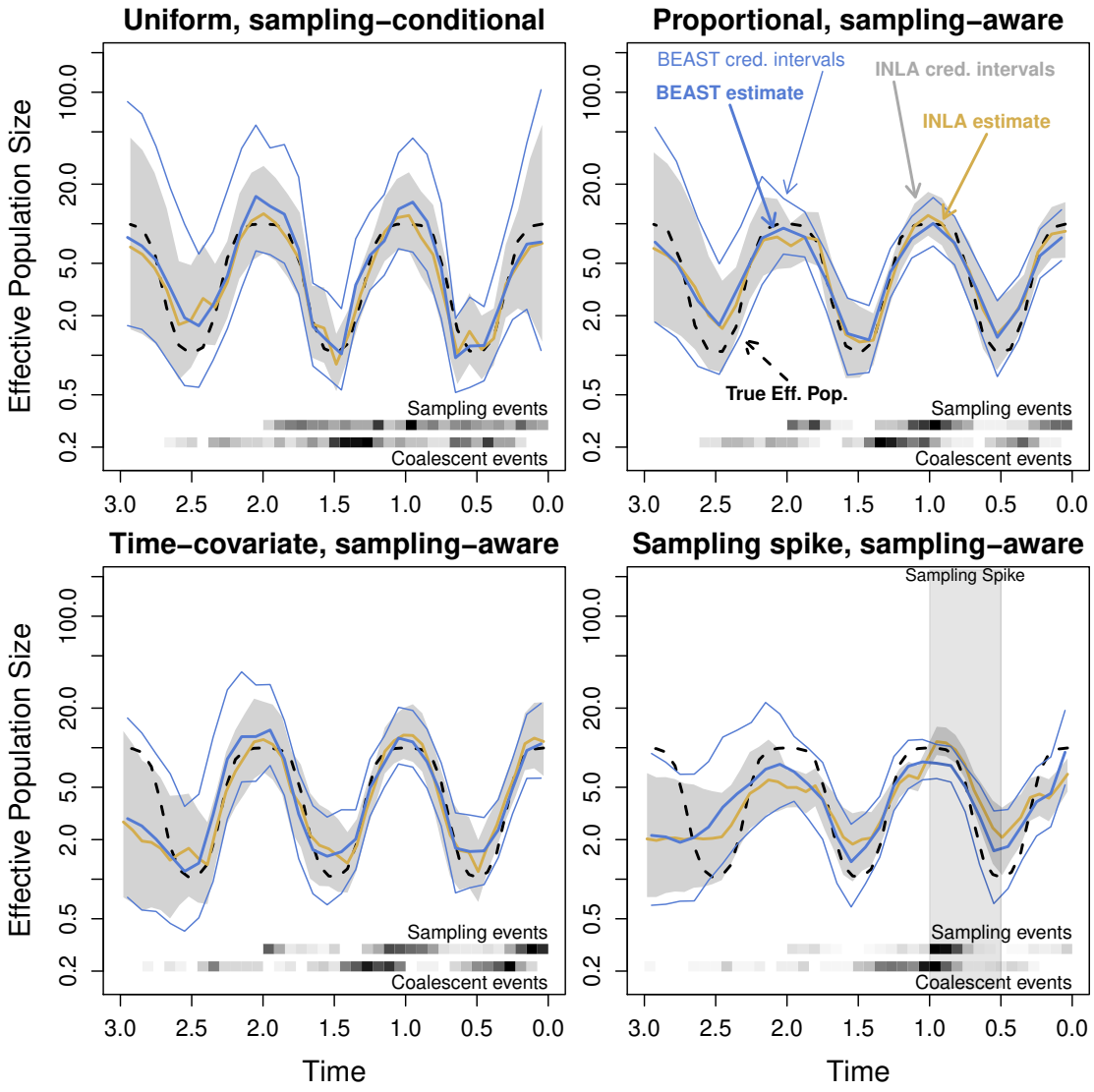

Figure 4.4: **Effective population size reconstructions for four sequence data simulations, all based on the same seasonal effective population size trajectory.** *Upper left*: Uniform sampling times, sampling-conditional posterior. *Upper right*: Sampling frequency proportional to effective population size, sampling-aware posterior. *Lower left*: Sampling frequency proportional to effective population times a time-covariate ($\exp(t)$), sampling- and covariate-aware posterior. *Lower right*: Sampling frequency proportional to effective population size with a sampling spike, sampling- and covariate-aware posterior.

a 95% Bayesian credible interval for the coefficient of the ordinary covariate that contains the true value ($\beta_2 = 0$), while the true value of the interaction covariate coefficient ($\delta_2 = 1$) is correctly inside the 95% Bayesian credible interval produced by SampESS/BEAST.

Model	**Coef**	**Q0.025**	**Median**	**Q0.975**	**Truth**
$\{\gamma(t)\}$	$\gamma(t)$	0.98	1.42	2.16	1.0
$\{\gamma(t), -t\}$	$\gamma(t)$	0.75	1.06	1.55	1.0
	$-t$	-0.06	0.44	0.94	0.5
$\{\gamma(t), 1_{t\in[0.5,1]}, 1_{t\in[0.5,1]} \cdot \gamma(t)\}$	$\gamma(t)$	0.72	1.26	2.14	1.0
	$1_{t\in[0.5,1]}$	-9.01	-1.50	1.64	0.0
	$1_{t\in[0.5,1]} \cdot \gamma(t)$	0.13	1.75	5.75	1.0

Table 4.2: **Summary of simulated sequence data inference** Posterior distribution quantile summaries for SampESS with no covariates (model: $\{\gamma(t)\}$), SampESS with an ordinary covariate (model: $\{\gamma(t), -t\}$), and SampESS with both an ordinary and interaction covariate (model: $\{\gamma(t), 1_{t\in[0.5,1]}, 1_{t\in[0.5,1]} \cdot \gamma(t)\}$).

4.4.2 Seasonal Influenza Example

We reanalyze the H3N2 regional influenza data for the USA/Canada region as analyzed with fixed-tree methods in [Karcher et al., 2016a]. The data contains 520 sequences aligned to form a multiple sequence alignment with 1698 sites of the hemagglutinin gene. This dataset is a subset of the dataset of influenza sequences from around the world analyzed in [Zinder et al., 2014]. We use ESS/BEAST with our tree-sampling MCMC targeting posterior (4.4) to analyze these data and mark the pointwise posterior median and 95% credible region in black, summarized in Figure 4.5 (upper row). We observe a seasonal pattern consistent with flu seasons observed in the temperate northern hemisphere [Zinder et al., 2014]. Our results are also consistent with previous fixed-tree method results but with larger credible interval widths due to correctly accounting for genealogical uncertainty in our analysis.

We apply our sampling-aware model SampESS/BEAST to the USA/Canada influenza data, following the posterior from Equation (4.3). We used several different log-sampling-intensity models. The simplest one has log-intensity $\beta_0 + \beta_1\gamma(t)$ (abbreviated $\{\gamma(t)\}$) and is summarized in Figure 4.5 (upper left). We include a time term in one model, with log-intensity $\beta_0 + \beta_1\gamma(t) + \beta_2 \cdot (-t)$ (abbreviated $\{\gamma(t), -t\}$) summarized in Figure 4.5 (upper center). We use seasonal indicator functions in the final model, defined as,

$$I_{\text{winter}}(t) = I_{(t \bmod 1.0)\in[0,0.25)},$$
$$I_{\text{autumn}}(t) = I_{(t \bmod 1.0)\in[0.25,0.5)},$$
$$I_{\text{summer}}(t) = I_{(t \bmod 1.0)\in[0.5,075)},$$

with t measured in decimal calendar years (going forward in time). This results in the log-intensity $\beta_0 + \beta_1\gamma(t) + \beta_2 I_{\text{winter}}(t) + \beta_3 I_{\text{autumn}}(t) + \beta_4 I_{\text{summer}}(t)$ (abbreviated $\{\gamma(t), I_{\text{winter}}$, $I_{\text{autumn}}, I_{\text{summer}}\}$), summarized in Figure 4.5 (upper right).

We summarize the sampling model coefficient results for each model in Table 4.3. The $\{\gamma(t)\}$ model corresponds to the preferential sampling model of Karcher et al. [2016a], but has noticeably different estimates. We attribute this to the differences between the fixed-tree (with a tree inferred using a constant effective population size BEAST model), INLA-based approach of Karcher et al. [2016a], and the tree-sampling MCMC-based approach of this paper. We also note that the $\{\gamma(t), -t\}$ model does not perform better (or even noticably differently) than the $\{\gamma(t)\}$ model. The coefficient summary for $\{\gamma(t), -t\}$ bears this out, because the 95% Bayesian credible interval for the coefficient for $-t$ contains 0. We do observe differences in the $\{\gamma(t), I_{\text{winter}},\ I_{\text{autumn}},\ I_{\text{summer}}\}$ model. The coefficient of $\gamma(t)$ is close to 1.0, which is the easiest value to interpret under preferential sampling, suggesting a baseline sampling rate proportional to effective population size. The coefficients for the indicators suggest increased sampling in the flu season intervals, as compared to the summer intervals and especially the spring intervals—with spring treated as a baseline rate without an indicator for the sake of identifiability.

We observe the seasonality of our estimates of the effective population size trajectory.

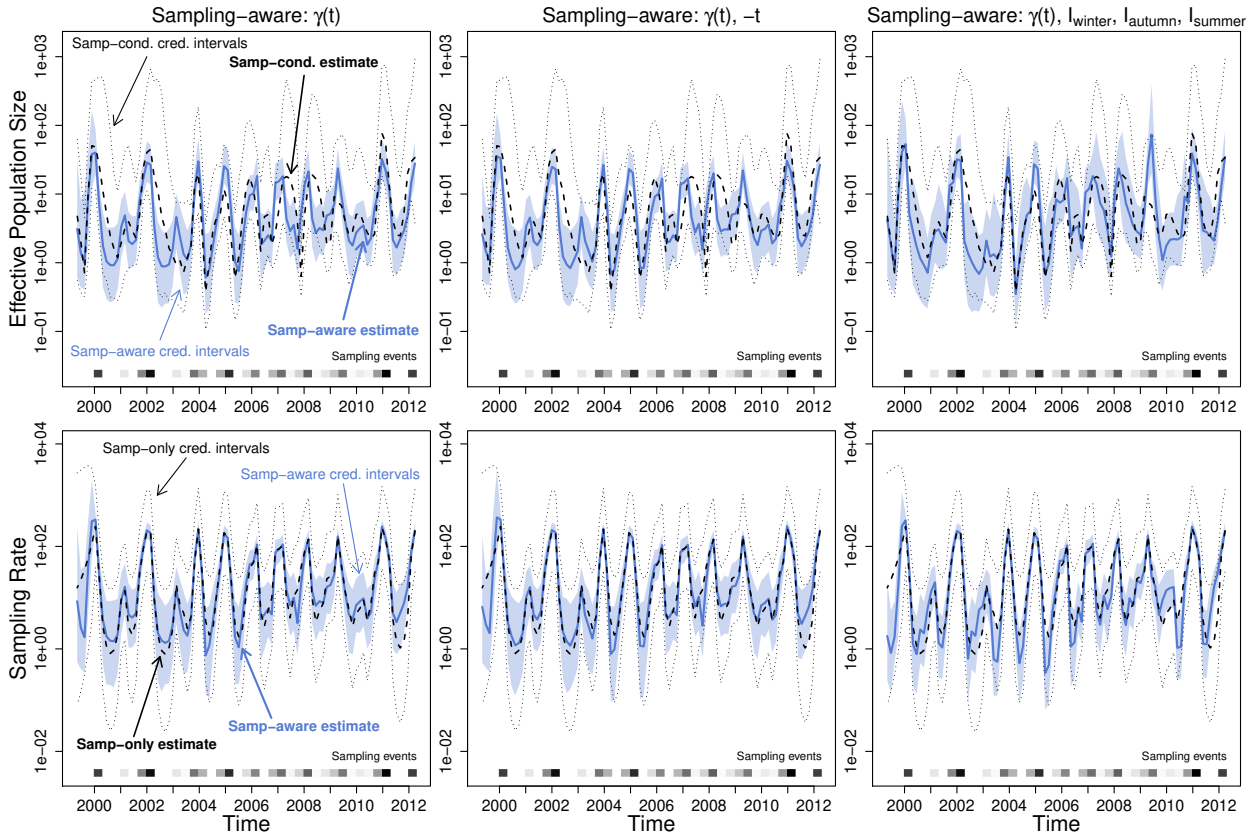

Figure 4.5: **Effective population size and sampling rate reconstructions for the USA and Canada influenza dataset.** *Upper row*: Dashed lines and dotted black lines are the pointwise posterior effective population size estimates and credible intervals of the sampling-conditional model. The blue lines and the light blue regions are the pointwise posterior effective population size estimates and credible intervals of that column's sampling-aware model. *Lower row*: Dashed lines and dotted black lines are the pointwise posterior sampling rate estimates and credible intervals of a nonparametric sampling-time-only model. The blue lines and the light blue regions are the pointwise posterior sampling rate estimates and credible intervals of that column's sampling-aware model.

Model	Coef	Q0.025	Median	Q0.975
$\{\gamma(t)\}$	$\gamma(t)$	1.11	1.45	2.01
$\{\gamma(t), -t\}$	$\gamma(t)$	1.21	1.52	2.00
	$-t$	-0.10	-0.02	0.07
$\{\gamma(t), I_{\text{winter}}, I_{\text{autumn}}, I_{\text{summer}}\}$	$\gamma(t)$	0.72	0.92	1.21
	I_{winter}	1.91	2.79	3.83
	I_{autumn}	1.88	2.85	3.85
	I_{summer}	0.44	1.52	2.58

Table 4.3: **Summary of USA/Canada influenza data inference.** Posterior distribution quantile summaries for SampESS with no covariates (model: $\{\gamma(t)\}$), SampESS with an ordinary covariate (model: $\{\gamma(t), -t\}$), and SampESS with seasonal indicator covariates (model: $\{\gamma(t), I_{\text{winter}}, I_{\text{autumn}}, I_{\text{summer}}\}$).

In Figure 4.6, we superimpose the twelve years of estimates per model, and plot the posterior median annual estimate. We note that the sampling aware models all show increased seasonality compared to the sampling conditional model. We also note that the 2008-2009 flu season stands out on the seasonality plot for having a peak in the summer months, particularly in the $\{\gamma(t), I_{\text{winter}}, I_{\text{autumn}}, I_{\text{summer}}\}$ model. We note that the competing H1N1 strain of influenza began an outbreak in spring 2009, likely outcompeting the H3N2 strain and resulting in the low effective population size that we observe in 2009 and 2010 in the sampling-conditional model. However, one possibility is that the CDC collected additional influenza sequence data around that time, resulting in the increased effective population size estimates we see in the sampling-aware models over the summer of 2009.

Finally, we compare the sampling rates we derive from our BEAST runs to a nonparametric INLA-based estimate of the sampling rate (using a method similar to BNPR-PS without the coalescent likelihood or covariates). Figure 4.5 (lower row) shows the comparison. The methods produce very similar estimates, with the BEAST/MCMC methods having thinner

credible intervals due to incorporating additional information from the coalescent likelihood.

4.4.3 Ebola Oubreak

Finally, we analyze a subset of the sequence data collected from the recent African Ebola outbreak [Dudas et al., 2017]. The data consists of 1610 samples collected from mid-2014 to mid-2015, and aligned across the entire genome (18992 sites). The dataset represents over 5% of known cases of Ebola during that outbreak, and presents an unprecedented insight into the epidemiological dynamics of an Ebola outbreak. We consider two subsets of the data, corresponding to the samples from Sierra Leone and Liberia. For Sierra Leone, we subsampled 200 sequences (out of 1010) for computational tractability, chosen randomly from the larger dataset. For Liberia, we use the entire collection of 205 sequences.

We begin with the Sierra Leone dataset. We apply ESS/BEAST with our tree-sampling posterior from Equation (4.4) and mark the pointwise posterior median and 95% credible region in black, summarized the results in Figure 4.7 (upper row). We observe an effective population size trajectory visually resembles a typical time trajectory of prevalence or incidence that peaks in Autumn of 2014. We apply our sampling-aware model SampESS/BEAST to the Ebola data, following the posterior from Equation (4.3). We used several different log-sampling-intensity models. The simplest has log-intensity $\beta_0 + \beta_1\gamma(t)$ (abbreviated $\{\gamma(t)\}$) and is summarized in Figure 4.7 (upper left). We include a t term in one model, with log-intensity $\beta_0 + \beta_1\gamma(t) + \beta_2 \cdot (-t)$ (abbreviated $\{\gamma(t), -t\}$) summarized in Figure 4.7 (upper center). We make $-t$ an interaction covariate as well in the final model, resulting in the log-intensity $\beta_0 + \beta_1\gamma(t) + \beta_2 \cdot (-t) + \delta_2\gamma(t) \cdot (-t)$ (abbreviated $\{\gamma(t), -t, -t \cdot \gamma(t)\}$), summarized in Figure 4.7 (upper right). We summarize the coefficient results for each model in Table 4.4. We note that the $\{\gamma(t), -t\}$ model performs slightly poorer than the $\{\gamma(t)\}$ model. The coefficient summary for $\{\gamma(t), -t\}$ bears this out, because the 95% Bayesian credible interval for the coefficient for $-t$ contains 0. We do observe improvement in the $\{\gamma(t), -t, -t \cdot \gamma(t)\}$ model. In this more parameter rich model, the posterior distribution of the sampling model parameters suggests that time by itself is not a good predictor of changes

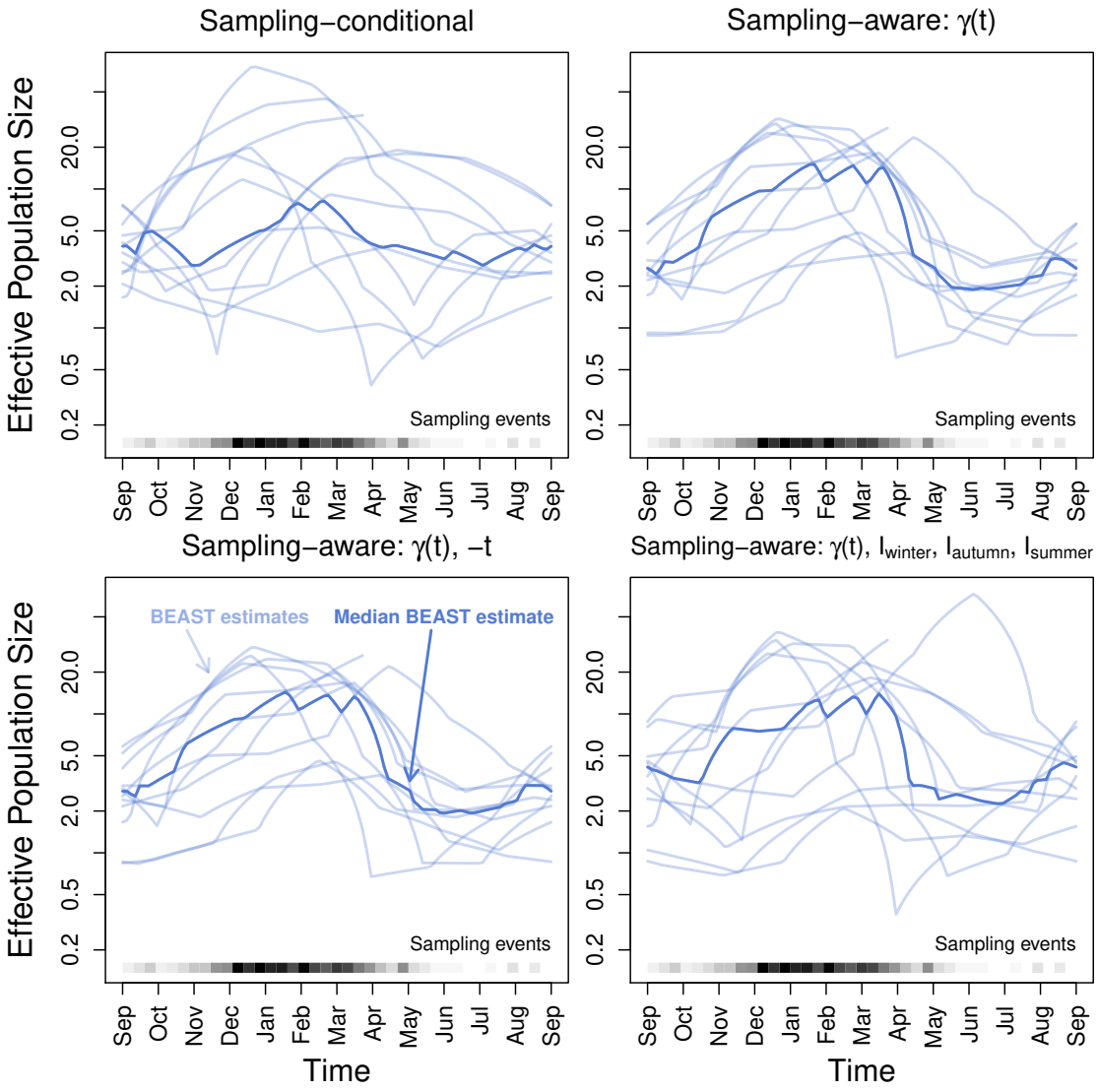

Figure 4.6: **Effective population size seasonal overlay for the USA and Canada influenza dataset.** The light blue lines are the pointwise posterior estimates for each year, and the dark blue line is the median annual estimate. *Upper left*: Sampling-conditional posterior. *Upper right*: Sampling-aware posterior with only log-effective population size $\gamma(t)$ informing the sampling time model. *Lower left*: Sampling- and covariate-aware posterior, with $\gamma(t)$ and $-t$. *Lower right*: Sampling- and covariate-aware posterior, with $\gamma(t)$ and seasonal indicators $I_{\text{winter}}, I_{\text{autumn}}, I_{\text{summer}}$.

in sampling intensity (the 95% Bayesian credible interval for β_2 contains both positive and negative values). However, the 95% credible interval for β_3 contains only positive values, suggesting that the preferential sampling intensity strength increased during the course of Sierra Leone's Ebola outbreak. More specifically, taking posterior medians of all parameters, our most complex preferential sampling model says that sampling intensity started with $141.2N_e^{1.76-1.37\cdot 1.65}(t) = 141.2N_e^{-0.5}(t)$ during the early stages of the outbreak and increased to $141.2N_e^{1.76}(t)$ by the Autumn of 2015. This form of the sampling intensity points at a possible increase over time of the proportion of all cases, whose Ebola genome ended up sequenced — a plausible outcome of an aggressive disease surveillance during an ongoing outbreak.

Model	**Coef**	**Q0.025**	**Median**	**Q0.975**
$\{\gamma(t)\}$	$\gamma(t)$	0.28	0.46	0.71
$\{\gamma(t), -t\}$	$\gamma(t)$	0.30	0.49	0.83
	$-t$	-0.58	0.27	1.46
$\{\gamma(t), -t, -t\cdot\gamma(t)\}$	$\gamma(t)$	1.01	1.76	3.32
	$-t$	-0.88	0.18	1.02
	$-t\cdot\gamma(t)$	0.89	1.65	3.11

Table 4.4: **Summary of Sierra Leone Ebola sequence data inference.** Posterior distribution quantile summaries for SampESS with no covariates (model: $\{\gamma(t)\}$), SampESS with an ordinary covariate (model: $\{\gamma(t), -t\}$), and SampESS with both an ordinary and interaction covariate (model: $\{\gamma(t), -t, -t\cdot\gamma(t)\}$).

We apply the same models to the Liberia Ebola dataset, summarized across the upper row of Figure 4.8 and in Table 4.5. We note that the $\{\gamma(t)\}$ and $\{\gamma(t), -t\}$ models perform very similarly, but the $\{\gamma(t), -t\}$ model has slightly wider pointwise credible intervals in places. This is consistent with the coefficients, as the credible interval for the $-t$ term contains 0. The $\{\gamma(t), -t, -t\cdot\gamma(t)\}$ model has even wider pointwise credible intervals, and

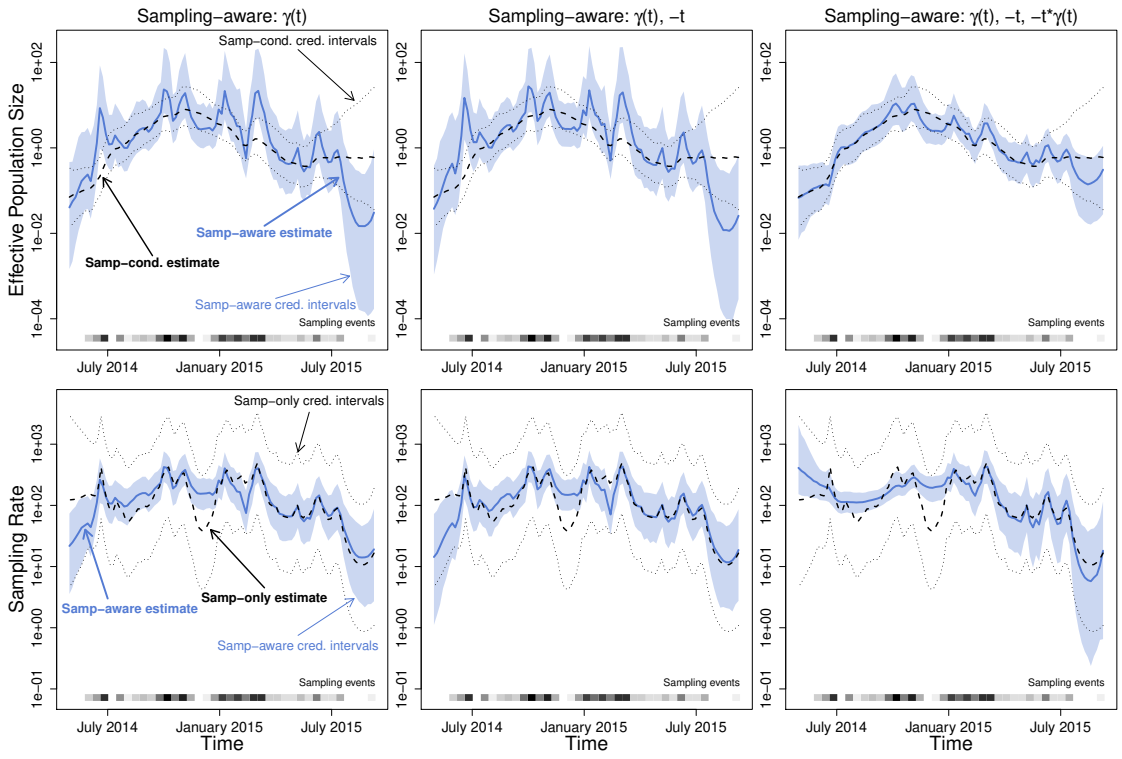

Figure 4.7: **Effective population size and sampling rate reconstructions for the Sierra Leone Ebola dataset.** *Upper row*: Dashed lines and dotted black lines are the pointwise posterior effective population size estimates and credible intervals of the sampling-conditional model. The blue lines and the light blue regions are the pointwise posterior effective population size estimates and credible intervals of that column's sampling-aware model. *Lower row*: Dashed lines and dotted black lines are the pointwise posterior sampling rate estimates and credible intervals of a nonparametric sampling-time-only model. The blue lines and the light blue regions are the pointwise posterior sampling rate estimates and credible intervals of that column's sampling-aware model.

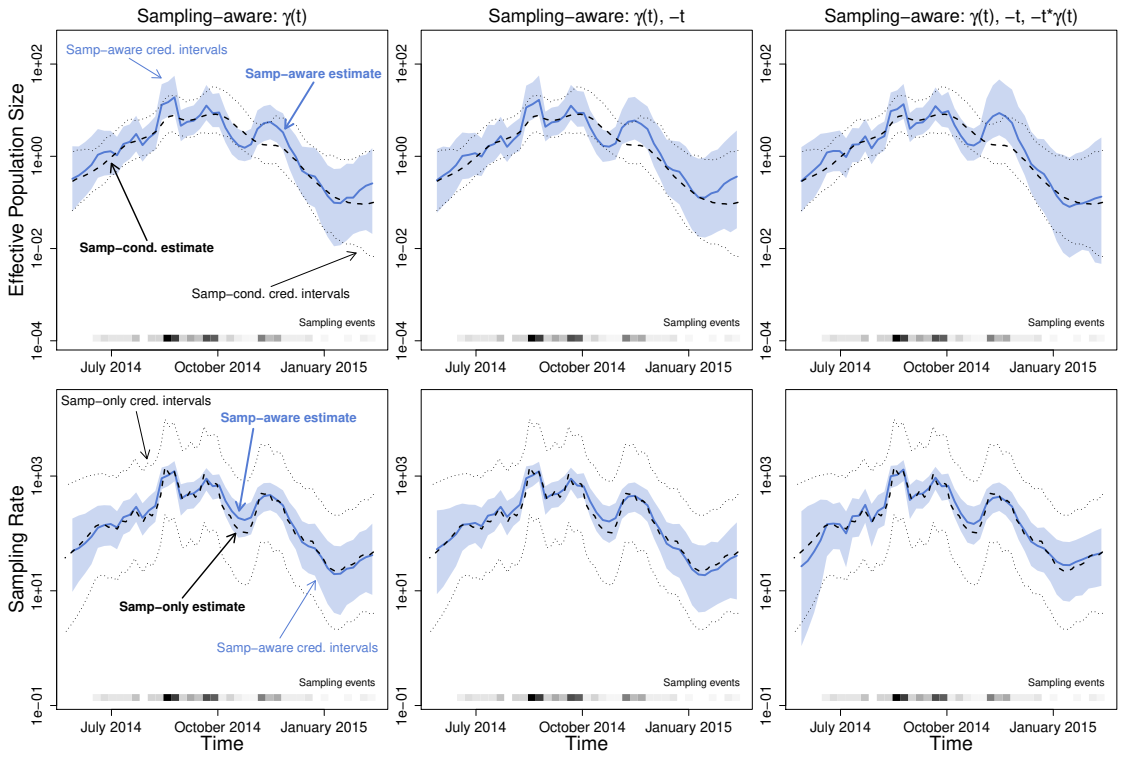

Figure 4.8: **Effective population size and sampling rate reconstructions for the Liberia Ebola dataset.** *Upper row*: Dashed lines and dotted black lines are the pointwise posterior effective population size estimates and credible intervals of the sampling-conditional model. The blue lines and the light blue regions are the pointwise posterior effective population size estimates and credible intervals of that column's sampling-aware model. *Lower row*: Dashed lines and dotted black lines are the pointwise posterior sampling rate estimates and credible intervals of a nonparametric sampling-time-only model. The blue lines and the light blue regions are the pointwise posterior sampling rate estimates and credible intervals of that column's sampling-aware model.

Model	Coef	Q0.025	Median	Q0.975
$\{\gamma(t)\}$	$\gamma(t)$	0.53	0.78	1.20
$\{\gamma(t), -t\}$	$\gamma(t)$	0.53	0.81	1.23
	$-t$	-3.39	-0.74	1.84
$\{\gamma(t), -t, -t \cdot \gamma(t)\}$	$\gamma(t)$	-0.07	0.40	1.51
	$-t$	-3.26	-0.31	2.67
	$-t \cdot \gamma(t)$	-2.98	-1.21	1.20

Table 4.5: **Summary of Liberia Ebola sequence data inference.** Posterior distribution quantile summaries for SampESS with no covariates (model: $\{\gamma(t)\}$), SampESS with an ordinary covariate (model: $\{\gamma(t), -t\}$), and SampESS with both an ordinary and interaction covariate (model: $\{\gamma(t), -t, -t \cdot \gamma(t)\}$).

the credible intervals for the coefficients all contain 0. This suggests that in Liberia, of the three sampling-aware models, the sampling model most consistent with the data is simple preferential sampling. We also note that in the $\{\gamma(t)\}$ model, the median estimate for the coefficient for $\gamma(t)$ is close to 1.0, suggesting direct proportional sampling.

As in the previous section, we compare the sampling rates we derive from our BEAST runs to a nonparametric INLA-based estimate of the sampling rate. Figures 4.7 (lower row) and 4.8 (lower row) show the comparisons. The two methods produce very similar estimates, and again the sampling-aware methods have thinner credible intervals due to incorporating additional information from the coalescent likelihood.

4.5 Discussion

Currently, few phylodynamic methods incorporate sampling time models in order to address model misspecification and take advantage of the additional information contained in sampling times in preferential sampling contexts. Even fewer methods implement sampling time models by appropriately integrating over genealogies relating the sampled genetic sequences

and performing inference directly from these sequence data. Furthermore, we extend previous sampling models to incorporate time-varying covariates in order to allow the sampling time model to be more flexible under different scientific circumstances.

However, this additional flexibility comes with additional uncertainty around which set of covariates is the best one for a given scientific context. For instance, regarding the USA/Canada influenza data, it is unclear which set of covariates most accurately reconstructs the effective population size trajectory. The model including seasonal indicators has credibly nonzero coefficients for each of the covariates, but the pointwise credible intervals are not better, and possibly worse, than the sampling time model only including effective population size. Similarly, the Sierra Leone Ebola data seems to support a sampling time model including an interaction covariate $-t \cdot \gamma(t)$, at least in terms of the credibility of the model coefficients and the widths of the pointwise credible intervals. However, additional precision does not guarantee, nor even imply, that the model is correct. We require a method to judge how well the data (sampling times, coalescent times, and/or sequence data) fit the model and covariate set we choose. One potential method for model evaluation relies on the concept of a posterior predictive check [Gelman et al., 1996]—taking a model with a posterior sample of parameters estimated from data, using the same model and estimated parameters to simulate new data, and comparing the observed data and the simulated data. This approach gives one way of judging how two or more models compare at explaining the real data.

Another approach to extending and increasing the flexibility of the sampling model is to decouple the fixed temporal relationship between effective population size and sampling intensity. Introducing an estimated lag parameter to the sampling time model would allow for cause-and-effect phenomena and delays to be accounted for within the model. Incorporating an estimated lag parameter would also allow for an additional avenue of model verification. Under most imaginable circumstances, if there is a relationship between the effective population size and sampling frequency, changes to the population size would effect sampling frequency with zero or positive delay. Estimating a credibly negative lag would be a possible

indicator that some element of the model or data is worth re-examining.

In terms of flexibility, the ideal sampling time model would be a separate Gaussian latent field distinct from the (log) effective population size. However, methods for primarily phylodynamic inference with this feature would suffer from severe identifiability problems. One approach that would retain most of the flexibility of the separate Gaussian field while also retaining the identifiability of the original model would be to model the (log) effective population size and sampling intensities as *correlated* Gaussian fields. Estimating the correlation parameter between the fields would illuminate the different between tightly and loosely bound population and sampling trajectories.

Chapter 5

MODEL CHECKS AND MODEL SELECTION

5.1 Methods

5.1.1 Transformed Exponentials

Suppose random variable $X \sim \text{Exp}(1)$, and thus its PDF is $f_X(x) = \exp(-x)$. Define $g_\lambda(u) = \int_0^u \lambda(t)dt$ for nonnegative $\lambda(\cdot)$ integrable on $[0, \infty)$. Then $g_\lambda(u)$ is monotonic non-decreasing, so $g_\lambda^{-1}(\cdot)$ is well-defined almost everywhere. If we let $U = g_\lambda^{-1}(X)$, then the PDF of U is $f_U(u) = \lambda(u)\exp(-\int_0^u \lambda(t)dt)$.

We then have two useful results. If we wish to sample U, we may do so by sampling an Exp(1) random variable X, then apply the transformations $U = g_\lambda^{-1}(X)$, which will result in the desired distribution. There generally is not an explicit, closed-form solution for $g^{-1}(\cdot)$, but it can be implicitly solved using root-finding methods and, if necessary, numerical integration. Conversely, if we wish to recover the original Exp(1) random variable X from U, we can apply the transformations $X = g_\lambda(U)$.

5.1.2 Heterochronous Coalescent Time Transformation

Consider the heterochronous coalescent model, as presented in Chapter 4. The coalescent provides the probability density of a genealogy (illustrated in Figure 5.1), given an effective population size trajectory $N_e(t)$ and a collection of n *sampling times* $\mathbf{s} = \{s_1 \geq \ldots \geq s_n = 0\}$. We typically refer to the bifurcation events of the genealogy tree as *coalescent events* and label them $\mathbf{g} = \{t_1 \geq \ldots \geq t_n = 0\}$. If $s_i = 0$ for all i, then we call the genealogy *isochronous*. Griffiths and Tavaré [1994a] prove that for isochronous data, the sequence of coalescent events of a genealogy (and allowing variable effective population size) is a continuous time

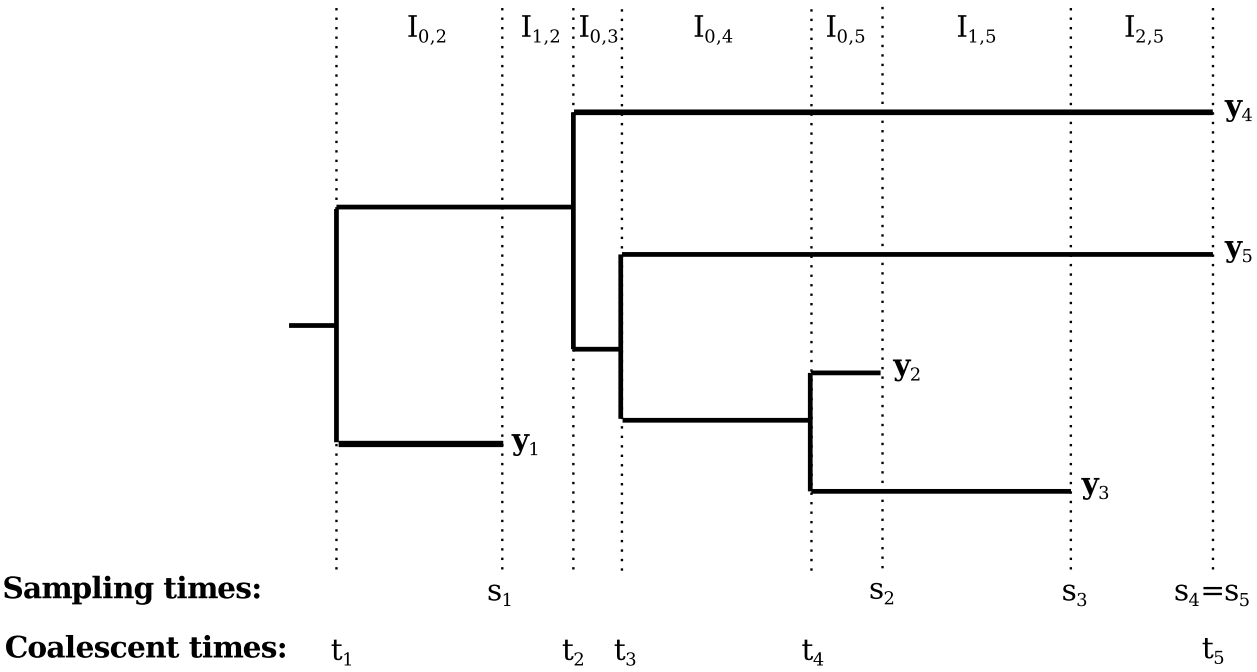

Figure 5.1: **Illustration of an example heterochronous genealogy with $n = 5$ lineages.** Sampling times $s_1, \ldots, s_5$ and coalescent times $t_1, \ldots, t_4$ are marked below the genealogy, and sequence data $\mathbf{y}_1, \ldots, \mathbf{y}_5$ are marked at their corresponding tips.

Markov chain and that the function $A_n(t)$, representing the number of distinct ancestors at time t and called the *ancestral process*, is a pure death process starting at value n at time 0 and decreasing by one at every coalescent event proceeding into the past.

We seek to extend this framework to allow heterochronous genealogies as well. Consider a Wright-Fisher population with population $N(i)$, i generations in the past. We assume that sampled individuals cannot be ancestors to future sampled individuals, so if we sample an individual at generation i, we segregate that individual from the other $N(i)$ individuals in the population until the sampled individual "selects" an ancestor in generation $i+1$, at which point the usual Wright-Fisher process proceeds until another individual is sampled farther in the past. Suppose we have a fixed schedule of n individuals sampled at generations $g_1 \leq g_2 \leq \ldots \leq g_n$, and we consider any particular generation i, having counted k coalescent events between generation 0 and generation i. Let $b_i = \sum_{i=1}^{n} 1_{[g_i > i]}$ represent the number of individuals that are sampled farther into the past than generation i. In an isochronous scenario, b_i would be 0 for all i, and the number of distinct lineages at generation i would be $n-k$. However, here we suppose that $b_i > 0$. We see that if there are no individuals sampled at generations i or $i+1$, then this iteration of the Wright-Fisher process is identical to an iteration of an isochronous Wright-Fisher process with the same population and $n-k-b_i$ distinct lineages. If there is an individual sampled at generation $i+1$, the outcome is the same since we can safely ignore the (segregated) sampled individual until iterating from generation $i+1$ to $i+2$. If there is an individual sampled at generation i, then we consider the (segregated) sampled individual to be an additional distinct lineage, but we see the iteration still behaves as if it were an iteration of an isochronous Wright-Fisher process with $n-k-b_i$ distinct lineages.

We now switch to continuous time, applying our heterochronous distinct lineage counts into the results from [Griffiths and Tavaré, 1994a]. Let $b(t) = \sum_{i=1}^{n} 1_{[s_i > t]}$ be the count of samples that occur farther into the past than time t. Let $B_n(t) = n - k(t)$, where $k(t)$ is the number of coalescent events between time 0 and time t. Under isochronous sampling, $B_n(t) = A_n(t)$ is the ancestral process. Under heterochronous sampling, $B_n(t)$ is merely

the pure death process that is directly analogous to $A_n(t)$. Substituting our results from the heterochronous Wright-Fisher process into the key results reveals the transition rates for $B_n(t)$,

$$\Pr(B_n(t+h) = j \mid B_n(t) = i) = \begin{cases} \binom{i-b(t)}{2}\frac{1}{N_e(t)}h + o(h), & j = i-1 \\ 1 - \binom{i-b(t)}{2}\frac{1}{N_e(t)}h + o(h), & j = i \\ 0 & \text{otherwise,} \end{cases}$$

and the joint density for the Markov chain of coalescent events,

$$\Pr(\mathbf{g} \mid N_e(t), \mathbf{s}) = \prod_{k=2}^{n} \left[\lambda_k(t_{k-1}) \exp\left(- \int_{t_k}^{t_{k-1}} \lambda_k(t) dt \right) \right],$$

where $\lambda_k(t) = \binom{k-b(t)}{2}\frac{1}{N_e(t)}$.

Following the results from [Griffiths and Tavaré, 1994a], we note that the terms in the product are in the form of transformed exponentials, and can be sampled by transforming $n-1$ independent, identically distributed (i.i.d.) Exp(1) random variables. Finally, we note that we can recover these exact $n-1$ i.i.d. Exp(1) random variables by applying the inverse transformation.

5.1.3 Coalescent Posterior Predictive Check

We consider the Bayesian approach for phylodynamic analysis laid out in Chapter 4, with a posterior with the dependency graph we see in 5.2. We observe molecular sequence data $\mathbf{y}$, with a sequence data likelihood (labeled 1, hyperparameters $\boldsymbol{\theta}$) relating to the genealogy $\mathbf{g}$. The coalescent (labeled 2), relates the $\mathbf{g}$ to the sampling times $\mathbf{s}$ and a piecewise constant parameterization of the log-effective population size $\boldsymbol{\gamma}$. We apply a Gaussian random-walk prior (labeled 3, precision hyperparameter κ) to $\boldsymbol{\gamma}$, and optionally apply a sampling time model (labeled 4, hyperparameters $\boldsymbol{\beta}$, and optional covariates $\mathcal{F}$) to $\mathbf{s}$. The posterior takes the form,

$$\begin{aligned} \Pr(\mathbf{g}, \boldsymbol{\gamma}, \kappa, \boldsymbol{\beta}, \boldsymbol{\theta} \mid \mathbf{y}, \mathbf{s}, \mathcal{F}) \propto & \Pr(\mathbf{y} \mid \mathbf{g}, \boldsymbol{\theta}) \Pr(\mathbf{g} \mid \boldsymbol{\gamma}, \mathbf{s}) \Pr(\mathbf{s} \mid \boldsymbol{\gamma}, \boldsymbol{\beta}, \mathcal{F}) \Pr(\boldsymbol{\gamma} \mid \kappa) \\ & \times \Pr(\kappa) \Pr(\boldsymbol{\beta}) \Pr(\boldsymbol{\theta}). \end{aligned} \tag{5.1}$$

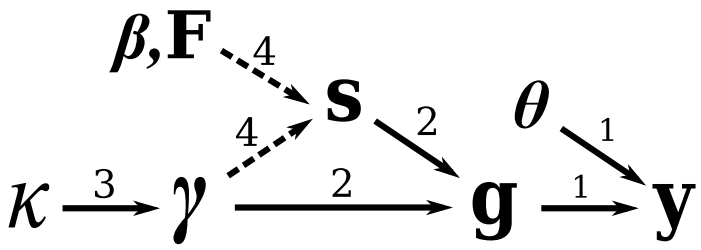

Figure 5.2: **Dependency graph for the phylodynamic model parameters and data.** Dependencies labeled 1 represent the sequence data likelihood, those labeled 2 represent the coalescent model, those labeled 3 represent the effective population latent field model, and those labeled 4 represent the sampling time model. The dashed lines between $\boldsymbol{\gamma}, \boldsymbol{\beta}, \mathcal{F}$ and $\mathbf{s}$ represent preferential sampling.

If we do not supply a sampling time model, our model reduces to the corresponding posterior:

$$\Pr(\mathbf{g}, \boldsymbol{\gamma}, \kappa, \boldsymbol{\theta} \mid \mathbf{y}, \mathbf{s}) \propto \Pr(\mathbf{y} \mid \mathbf{g}, \boldsymbol{\theta}) \Pr(\mathbf{g} \mid \boldsymbol{\gamma}, \mathbf{s}) \Pr(\boldsymbol{\gamma}|\kappa) \Pr(\kappa) \Pr(\boldsymbol{\theta}). \tag{5.2}$$

Similar to Gelman et al. [1996]'s mixed predictive distribution approach, we simulate data and certain latent variables from our models, informed by our posterior sample, in order to judge how well those models adhere to observed and inferred realities. In the context of our posterior with no sampling time model, we replicate $\{\mathbf{y}_i^{\text{rep}}\}_{i=1}^N$ and $\{\mathbf{g}_i^{\text{rep}}\}_{i=1}^N$ according to this joint posterior,

$$\Pr(\mathbf{y}^{\text{rep}}, \mathbf{g}^{\text{rep}}, \boldsymbol{\gamma}, \kappa, \boldsymbol{\theta} \mid \mathbf{y}, \mathbf{s}) \propto \Pr(\mathbf{y}^{\text{rep}} \mid \mathbf{g}^{\text{rep}}, \boldsymbol{\theta}) \Pr(\mathbf{g}^{\text{rep}} \mid \boldsymbol{\gamma}, \mathbf{s}) \Pr(\boldsymbol{\gamma}, \kappa, \boldsymbol{\theta} \mid \mathbf{y}, \mathbf{s}), \tag{5.3}$$

simulating from the coalescent $\Pr(\mathbf{g}^{\text{rep}} \mid \boldsymbol{\gamma}, \mathbf{s})$ and (if necessary, see below) the substitution model $\Pr(\mathbf{y}^{\text{rep}} \mid \mathbf{g}^{\text{rep}})$. We sample the final term on the right side via MCMC.

With posterior-sampled replicates available, we construct a discrepancy D_c [Gelman et al., 1996, Sinharay and Stern, 2003] on the observables and the inferred latent variables. Let $G(\mathbf{g}, \boldsymbol{\gamma})$ be the transformation (explored in the previous section) that, given the correct effective population trajectory, and valid assumptions for the coalescent model, will produce

a sample of $n-1$ i.i.d. Exp(1)-distributed random variables. Let K be the Kolmogorov-Smirnov statistic [Massey Jr, 1951],

$$K_{\text{Exp}(1)}(\mathbf{e}) = \sup_{x \in \mathbb{R}} |F_{\mathbf{e}}(x) - F_{\text{Exp}(1)}(x)|, \tag{5.4}$$

where $F_{\mathbf{e}}(x)$ is the empirical cumulative distribution function (ECDF) of $\mathbf{e}$, and $F_{\text{Exp}(1)}(x)$ is the true cumulative distribution function (CDF) of the Exp(1) distribution. We define

$$D_c(\mathbf{y}, \mathbf{g}, \mathbf{s}, \boldsymbol{\gamma}, \kappa) = K_{\text{Exp}(1)}(G(\mathbf{g}, \boldsymbol{\gamma})).$$

Then when we run MCMC, we then compare the *observed discrepancies*,

$$\{D_c(\mathbf{y}, \mathbf{g}_i, \mathbf{s}, \boldsymbol{\gamma}_i, \kappa_i)\}_{i=1}^N,$$

to the *replicate discrepancies*,

$$\{D_c(\mathbf{y}_i^{\text{rep}}, \mathbf{g}_i^{\text{rep}}, \mathbf{s}, \boldsymbol{\gamma}_i, \kappa_i)\}_{i=1}^N.$$

Note that the D_c we constructed does not depend on $\mathbf{y}^{\text{rep}}$, so we can save computation time by not simulating $\mathbf{y}^{\text{rep}} \mid \mathbf{g}^{\text{rep}}$. If we wish to check the sampling-aware posterior with the sampling time model, the replicate posterior remains mostly the same as in Equation 5.3, but the final term becomes $\Pr(\boldsymbol{\gamma}, \kappa, \boldsymbol{\beta}, \boldsymbol{\theta} \mid \mathbf{y}, \mathbf{s}, \mathcal{F})$ to match the sampling-aware posterior.

One method we have to compare the observed and replicate discrepancies is the posterior predictive p-value [Gelman et al., 1996]. We calculate the posterior predictive p-value by finding the proportion of MCMC iterations where the replicated discrepancy values are larger than its corresponding observed discrepancy value. The smaller the posterior predictive p-value, the more unusual the observed data is in the context of the chosen model. Note that this posterior predictive p-value does not have the usual frequentist p-value properties such as uniformity under a null model. However, values close to 50% suggest that the current model is adequate, and for discrepancies that become larger as the observed data becomes less likely given a set of parameters, the posterior predictive p-value tends to be smaller, to some degree, under under inadequate models [Gelman et al., 1996].

5.1.4 Sampling Posterior Predictive Checks

Similarly to the previous section, we replicate $\{\mathbf{y}_i^{\text{rep}}\}_{i=1}^N$, $\{\mathbf{g}_i^{\text{rep}}\}_{i=1}^N$, and $\{\mathbf{s}_i^{\text{rep}}\}_{i=1}^N$ according to this joint posterior,

$$\begin{aligned}\Pr(\mathbf{y}^{\text{rep}}, \mathbf{g}^{\text{rep}}, \mathbf{s}^{\text{rep}}, \boldsymbol{\gamma}, \kappa, \boldsymbol{\beta}, \boldsymbol{\theta} \mid \mathbf{y}, \mathbf{s}) \propto \Pr(\mathbf{y}^{\text{rep}} \mid \mathbf{g}^{\text{rep}}) \Pr(\mathbf{g}^{\text{rep}} \mid \boldsymbol{\gamma}, \mathbf{s}^{\text{rep}}) \Pr(\mathbf{s}^{\text{rep}} \mid \boldsymbol{\gamma}, \boldsymbol{\beta}) \\ \times \Pr(\boldsymbol{\gamma}, \kappa, \boldsymbol{\beta}, \boldsymbol{\theta} \mid \mathbf{y}, \mathbf{s}),\end{aligned} \tag{5.5}$$

with $\Pr(\boldsymbol{\gamma}, \kappa, \boldsymbol{\beta}, \boldsymbol{\theta} \mid \mathbf{y}, \mathbf{s})$ sampled via MCMC. We simulate from the sampling model $\Pr(\mathbf{s}^{\text{rep}} \mid \boldsymbol{\gamma}, \beta)$, and, if necessary, the coalescent $\Pr(\mathbf{g}^{\text{rep}} \mid \boldsymbol{\gamma}, \mathbf{s})$, and the substitution model $\Pr(\mathbf{y}^{\text{rep}} \mid \mathbf{g}^{\text{rep}})$.

Suppose we divide the sampling interval into a grid $K_1, \ldots, K_l$, potentially the same grid as used by grid-based priors for the effective population trajectory. The sampling model is inhomogeneous Poisson, so we can bin the numbers of sampling times within each interval $m_1, \ldots, m_l$, each with expected values $E_i = \int_{K_i} \lambda_s(t)dt$. A common approach to problems with independent Poisson bins is a Chi-squared test with statistic $\chi_s^2 = \sum_{i=1}^l \frac{(m_i - E_i)^2}{E_i}$ [Pearson, 1900]. We can then define a discrepancy

$$D_{\chi^2}(\mathbf{y}, \mathbf{g}, \mathbf{s}, \boldsymbol{\gamma}, \kappa) = \sum_{i=1}^{l} \frac{(m_i - E_i)^2}{E_i}, \tag{5.6}$$

for m_i and E_i derived from $\mathbf{s}$ as above.

5.2 Results

5.2.1 Simulation Study

Fixed Genealogy Inference

We perform a simulation study in order to explore the capabilities of the posterior predictive checks proposed above in Sections 5.1.3 and 5.1.4. We begin with a simplified version of the phylodynamic data-to-inference methodology. Here we take genealogies to be our observed data (and move on to inference based on observed sequence data in the next section). We simulate sampling times according to inhomogeneous Poisson processes with different intensity trajectories via a time-transformation method [Çinlar, 1975] as we implemented in our

Scenario	Sampling Model	Post. Pred. p-val	
		Coalescent	Sampling
Uniform	Conditional	0.58	—
Proportional	Aware: $\boldsymbol{\gamma}(t)$	0.59	0.72
Unrelated	Conditional	0.46	—
Unrelated	Aware: $\boldsymbol{\gamma}(t)$	0.00	0.15

Table 5.1: **Posterior predictive p-values for simulated fixed-tree data.**

R package `phylodyn` [Karcher et al., 2016b]. Give sampling time data, we simulate from the coalescent model using a similar time-tranformation method for the coalescent [Slatkin and Hudson, 1991], again as implemented in `phylodyn`. For all of our fixed-tree simulations, we use an effective population size trajectory designed to mimic the seasonal effective population size changes of a seasonal disease such as influenza in North America [Zinder et al., 2014], defined as follows:

$$N_{e,l,u,p,o}(t) = \begin{cases} l + \frac{(u-l)}{1+\exp\{2[3-(\frac{t+o}{p} \pmod{12})]\}}, & \text{if } \frac{t+o}{p} \pmod{12} \leq 6, \\ l + \frac{(u-l)}{1+\exp\{2[3+(\frac{t+o}{p} \pmod{12}-12]\}}, & \text{if } \frac{t+o}{p} \pmod{12} > 6. \end{cases} \tag{5.7}$$

Specifically, we use $N_{e,10,100,12,0}(t)$ which is most comparable to an influenza effective population size trajectory as measured in units of weeks, with $t = 0$ representing the summer effective population size minimum. We compare the results of our posterior predictive checks across different sampling scenario and choice-of-posterior combinations.

In our first scenario, we simulate 500 sampling times, distributed according to a uniform distribution between $t = 0$ and $t = 24$ (weeks), and simulate a genealogy with effective population size $N_{e,10,100,12,0}(t)$. We infer the underlying effective population size trajectory with a sampling-conditional posterior using a Markov chain Monte Carlo (MCMC) method with an elliptical slice sampling transition kernel (ESS) [Murray et al., 2010] as implemented in `phylodyn` (illustrated in the first row, first column of Figure 5.3) We use the MCMC

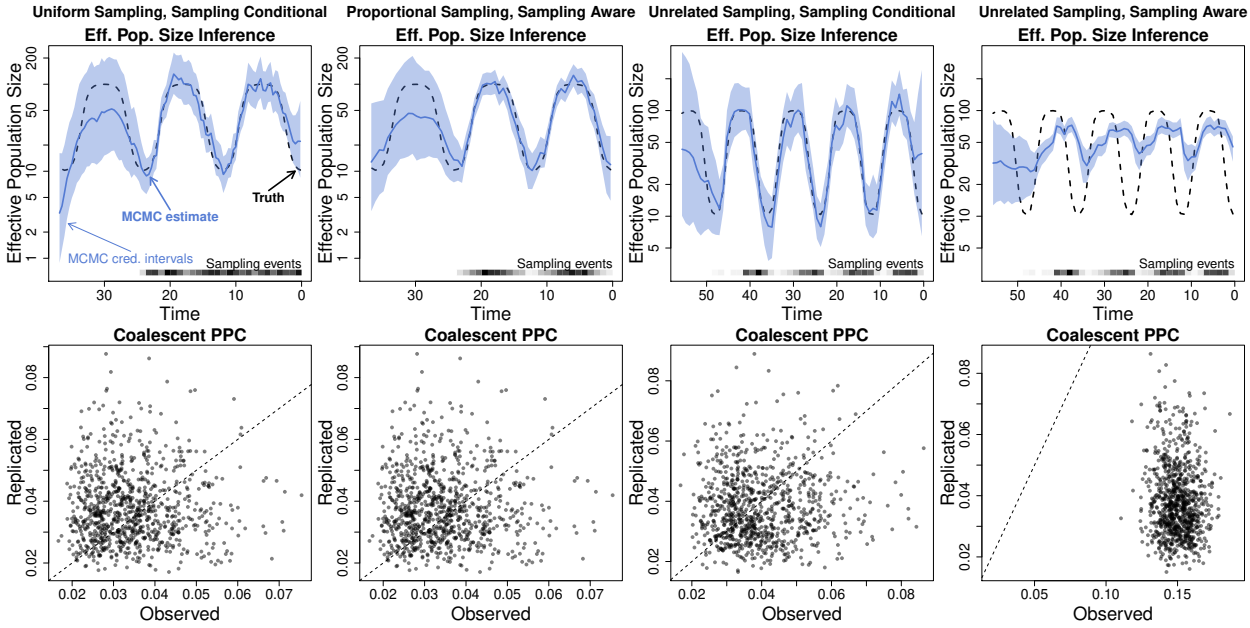

Figure 5.3: **Effective population size inference and coalescent posterior predictive check for fixed-tree simulations.** The dashed black line represents the true effective population trajectory. The solid blue line represents the posterior median effective population trajectory inferred by fixed-tree MCMC and the light blue region represents the corresponding pointwise 95% credible intervals for the effective population trajectory.

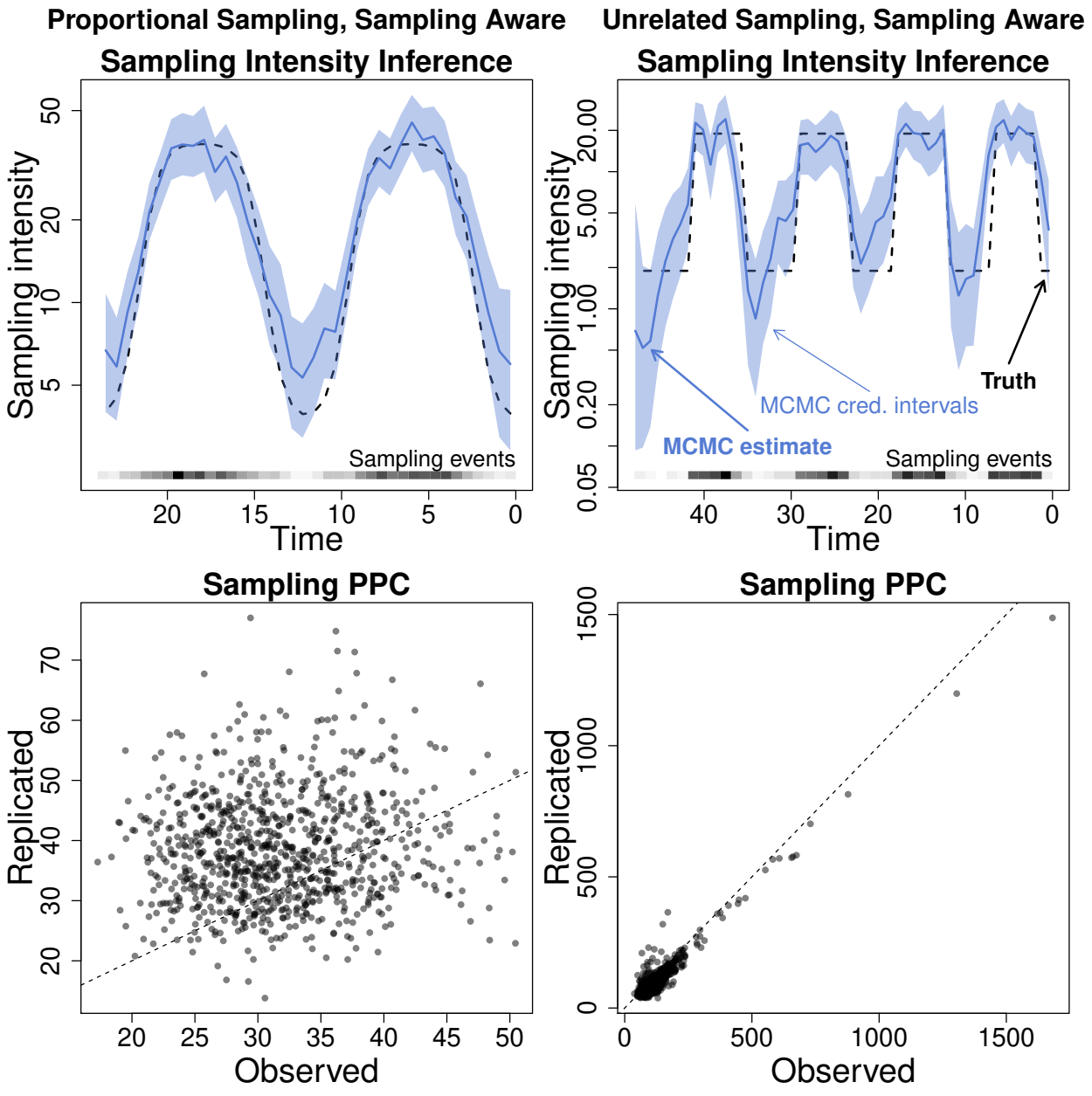

Figure 5.4: **Sampling intensity inference and sampling time posterior predictive check for fixed-tree simulations.** The dashed black line represents the true sampling intensity. The solid blue line represents the posterior median sampling intensity inferred by fixed-tree MCMC, and the light blue region represents the corresponding pointwise 95% credible intervals for the sampling intensity.

output to generate replicate coalescent data as laid out in Section 5.1.3 and calculate our coalescent discrepancy D_c for the observed MCMC results as well as for the replicated results. We plot the discrepancy comparison in the second row, first column of Figure 5.3, and note that the posterior predictive p-value is 0.58, which is close to 0.5, correctly suggesting that the model is adequate.

We proceed with several additional scenarios. We simulate 514 sampling times between $t = 0$ and $t = 24$ (weeks), distributed proportionally to the effective population size, with sampling log-intensity $\log[\lambda_c(t)] = -0.97 + N_{e,10,100,12,0}(t)$. We infer the underlying effective population size trajectory with a sampling-aware posterior (illustrated in the second column of Figure 5.3), with sampling time model $\log[\lambda_s(t)] = \beta_0 + \beta_1 \cdot \boldsymbol{\gamma}(t)$. We calculate the posterior predictive p-value as 0.59, again correctly suggesting adequacy. We also simulate 509 sampling times between $t = 0$ and $t = 48$ (weeks), distributed proportionally to a piecewise constant function $P(t)$ (illustrated in the second column of Figure 5.4) unrelated to the effective population size, with log-sampling intensity $\log[\lambda_c(t)] = -1.67 + P(t)$. We infer the underlying effective population size trajectory using two different methods. We use the sampling-conditional method (illustrated in the third column of Figure 5.3) and the sampling-aware method (illustrated in the fourth column of Figure 5.3) with sampling log-intensity $\log[\lambda_s(t)] = \beta_0 + \beta_1 \cdot \boldsymbol{\gamma}(t)$. The sampling-conditional posterior predictive p-value becomes 0.46, suggesting that this method (which only considers the coalescent model) does produce an adequate estimate of the effective population size trajectory. The sampling-aware posterior predictive p-value becomes zero, suggesting that this method produced a very poor estimate of the effective population size trajectory (very visible in Figure 5.3). This is likely due to the sampling time model mistaking fluctuations in sampling intensity for information about the effective population size trajectory, illustrating the importance of model checking when the true sampling model is uncertain.

For our sampling-aware scenarios, we apply our sampling time posterior predictive check as well. Our chi-squared sampling discrepancy D_{χ^2} generates a posterior predictive p-value of 0.72, correctly suggesting a good fit. The unrelated sampling scenario also produces a

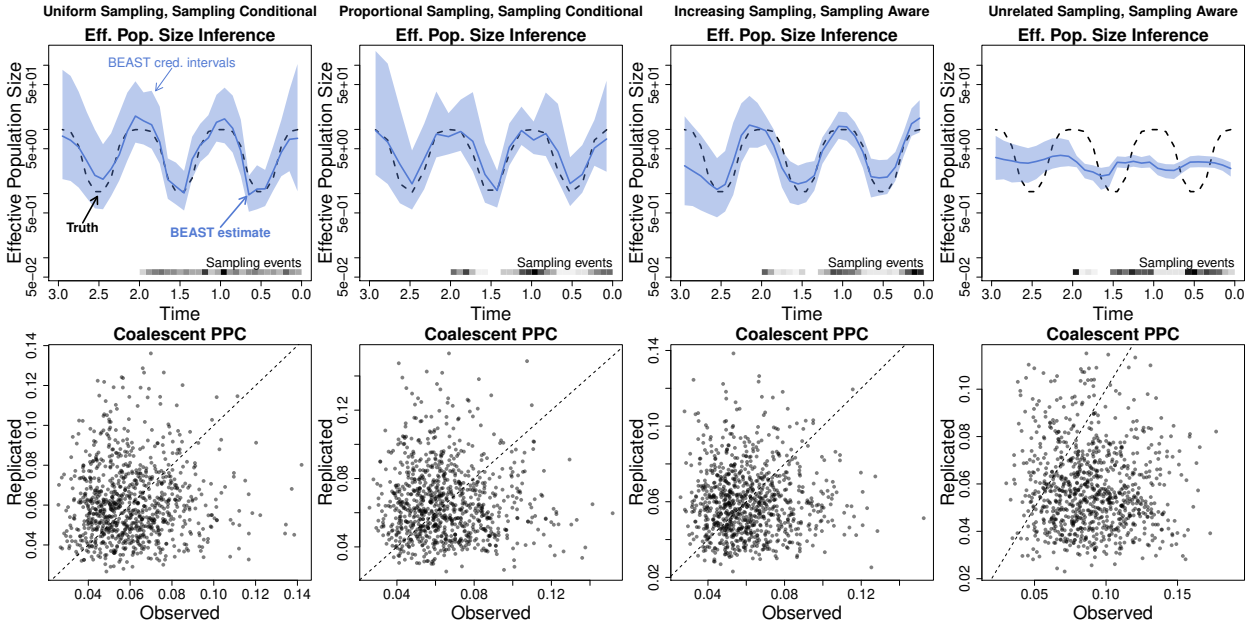

Figure 5.5: **Effective population size inference and coalescent posterior predictive check for sequence data simulations.** The dashed black line represents the true effective population trajectory. The solid blue line represents the posterior median effective population trajectory inferred by BEAST, and the light blue region represents the corresponding pointwise 95% credible intervals for the effective population trajectory.

sampling posterior predictive p-value. We see a relatively low posterior predictive p-value of 0.15, reacting to differences between the true and inferred sampling intensity trajectories.

Sequence Data Inference

Now, we expand the scope of our simulation study to be based on simulated sequence alignment data instead of a known genealogy. In this section, all of our examples will be based on an effective population size trajectory of $N_{e,1,10,1,0.5}(t)$, mimicking the trajectory of a seasonal disease as measured in units of years. Similar to the previous section, we generate sampling times and genealogies according to different sampling scenarios and the coalescent,

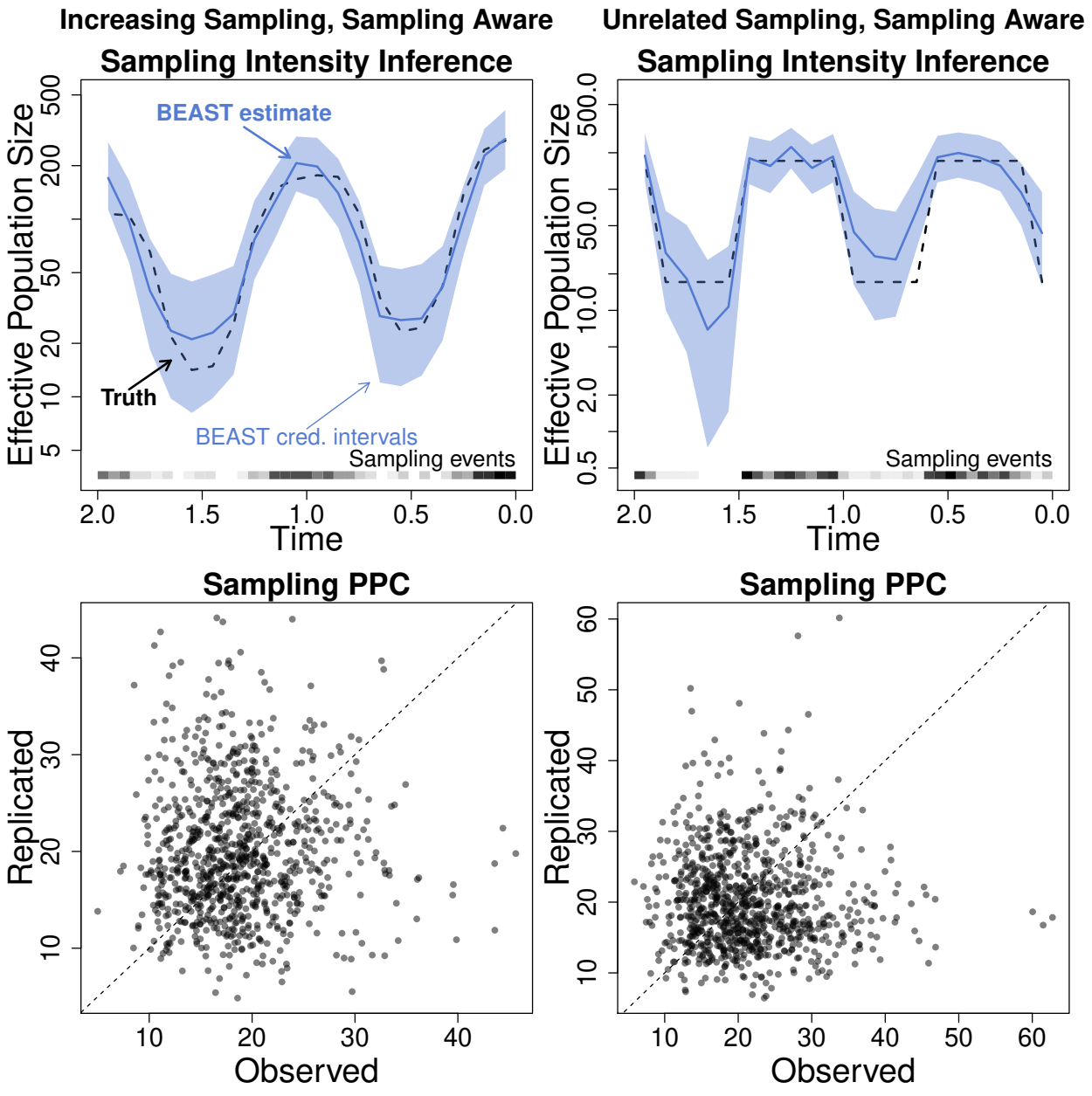

Figure 5.6: **Sampling intensity inference and sampling time posterior predictive check for sequence data simulations.** The dashed black line represents the true sampling intensity. The solid blue line represents the posterior median sampling intensity inferred by BEAST, and the light blue region represents the corresponding pointwise 95% credible intervals for the sampling intensity.

Scenario	Sampling Model	Post. Pred. p-val	
		Coalescent	Sampling
Uniform	Conditional	0.51	—
Proportional	Conditional	0.50	—
Increasing	Aware: $\boldsymbol{\gamma}(t)$	0.47	0.56
Unrelated	Aware: $\boldsymbol{\gamma}(t)$	0.17	0.46

Table 5.2: **Posterior predictive p-values for simulated sequence data.**

respectively. Given a genealogy, we simulate sequence data using the software `SeqGen` [Rambaut and Grass, 1997] using the Jukes-Cantor 1969 [Jukes et al., 1969] substitution model to generate 1500 sites. We set the substitution rate to produce an expected 0.9 mutations per site, in order to produce a sequence alignment with many sites having one mutation and some sites having zero or multiple mutations.

For our first simulation, we distribute 200 sampling times uniformly between $t = 0$ and $t = 2$ (years). We infer the underlying genealogy and effective population size trajectory using the software `BEAST` [Drummond et al., 2012] with an elliptical slice sampling transition kernel (ESS) [Murray et al., 2010] as implemented in Chapter 4, with a sampling-conditional posterior. Finally, we generate replicate genealogies as in the previous section, and we calculate our coalescent discrepancy D_c for the observed BEAST results as well as the replicates. In Figure 5.5 (first column), we see that the effective population estimate is close to the true trajectory, and when we compare the observed and replicate discrepancies, we calculate a posterior predictive p-value of 0.51, corroborating the model's adequacy. Next, we distribute 170 sampling times between $t = 0$ and $t = 2$ (years) with sampling log-intensity $\log[\lambda_c(t)] = 2.90 + N_{e,1,10,1,0.5}(t)$. We infer the underlying genealogy and effective population size trajectory using the sampling-conditional model and calculate discrepancies as above. Note this is a model misspecification applying a sampling-conditional model to a preferential sampling sampling scenario in the style of [Karcher et al., 2016a]. Unfortunately, the poste-

rior predictive p-value (0.50) does not detect this mismatch, as the bias effective population size estimate is hard to visually detect in Figure 5.5.

In our third scenario, we distribute 199 sampling times between $t = 0$ and $t = 2$ (years) with increasing sampling log-intensity $\log[\lambda_c(t)] = 3.35 - 0.5t + N_{e,1,10,1,0.5}(t)$. We infer as above, but targeting the sampling-aware posterior with sampling log-intensity $\log[\lambda_s(t)] = \beta_0 + \beta_1 \cdot \boldsymbol{\gamma}(t)$. This is again a misspecification, as the model cannot recover the $-0.5t$ term. However, the posterior predictive check does not clearly detect the mismatch, with a posterior predictive p-value of 0.47. Our sampling posterior predictive check does not detect the misspecification either, with a posterior predictive p-value of 0.56. In our final scenario, we distribute 222 sampling times between $t = 0$ and $t = 2$ (years) with a sampling log-intensity $\log[\lambda_c(t)] = 2.84 + P'(t)$ ($P'(t)$ illustrated in Figure 5.6, second column) unrelated to the effective population size. We target the sampling-aware posterior, with sampling log-intensity $\log[\lambda_s(t)] = \beta_0 + \beta_1 \cdot \boldsymbol{\gamma}(t)$. The model reconstructs the effective population size trajectory poorly, and this is successfully reflected in the posterior predictive p-value of 0.17. However, our sampling posterior predictive check does not detect the misspecification, with a posterior predictive p-value of 0.46.

5.2.2 Seasonal Influenza

We apply our posterior predictive check methods to the real world epidemiological datasets explored in the previous chapter. Here we analyze a North American subset of global H3N2 influenza [Zinder et al., 2014]. The data contains 520 sequences aligned to form a multiple sequence alignment with 1698 sites of the hemagglutinin gene. We use the same sequence data `BEAST` framework as the previous section, choosing four different specific sampling time models. We use a sampling-conditional model with no sampling time model and three sampling-aware models with log-intensities $\log[\lambda_s(t)] = \beta_0 + \beta_1 \cdot \boldsymbol{\gamma}(t)$, $\log[\lambda_s(t)] = \beta_0 + \beta_1 \cdot \boldsymbol{\gamma}(t) + \beta_2 \cdot (-t)$, and $\log[\lambda_s(t)] = \beta_0 + \beta_1 \cdot \boldsymbol{\gamma}(t) + \beta_2 \cdot I_w(t) + \beta_3 \cdot I_a(t) + \beta_4 \cdot I_s(t)$, where $I_w(t) = I_{(t \bmod 1) \in [0,0.25)}$ is an indicator function for winter, $I_a(t) = I_{(t \bmod 1) \in [0.25,0.5)}$ is an indicator function for autumn, and $I_s(t) = I_{(t \bmod 1) \in [0.5,075)}$ is an indicator function for

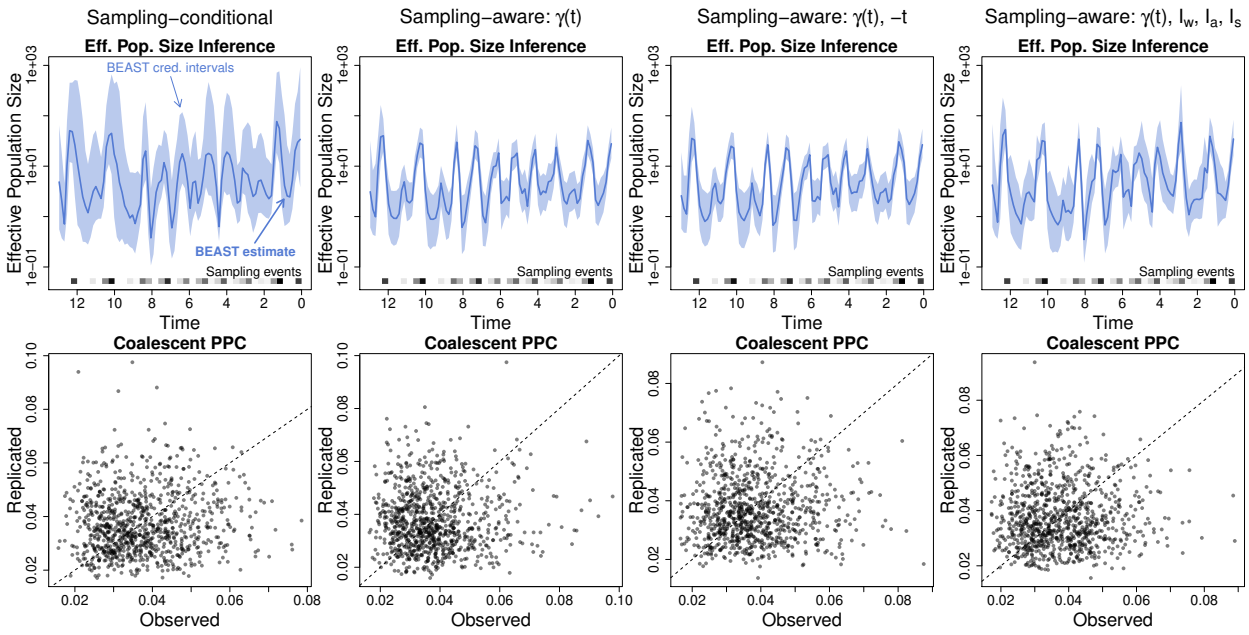

Figure 5.7: **Effective population size inference and coalescent posterior predictive check for seasonal influenza data.** The solid blue line represents the posterior median effective population trajectory inferred by BEAST, and the light blue region represents the corresponding pointwise 95% credible intervals for the effective population trajectory.

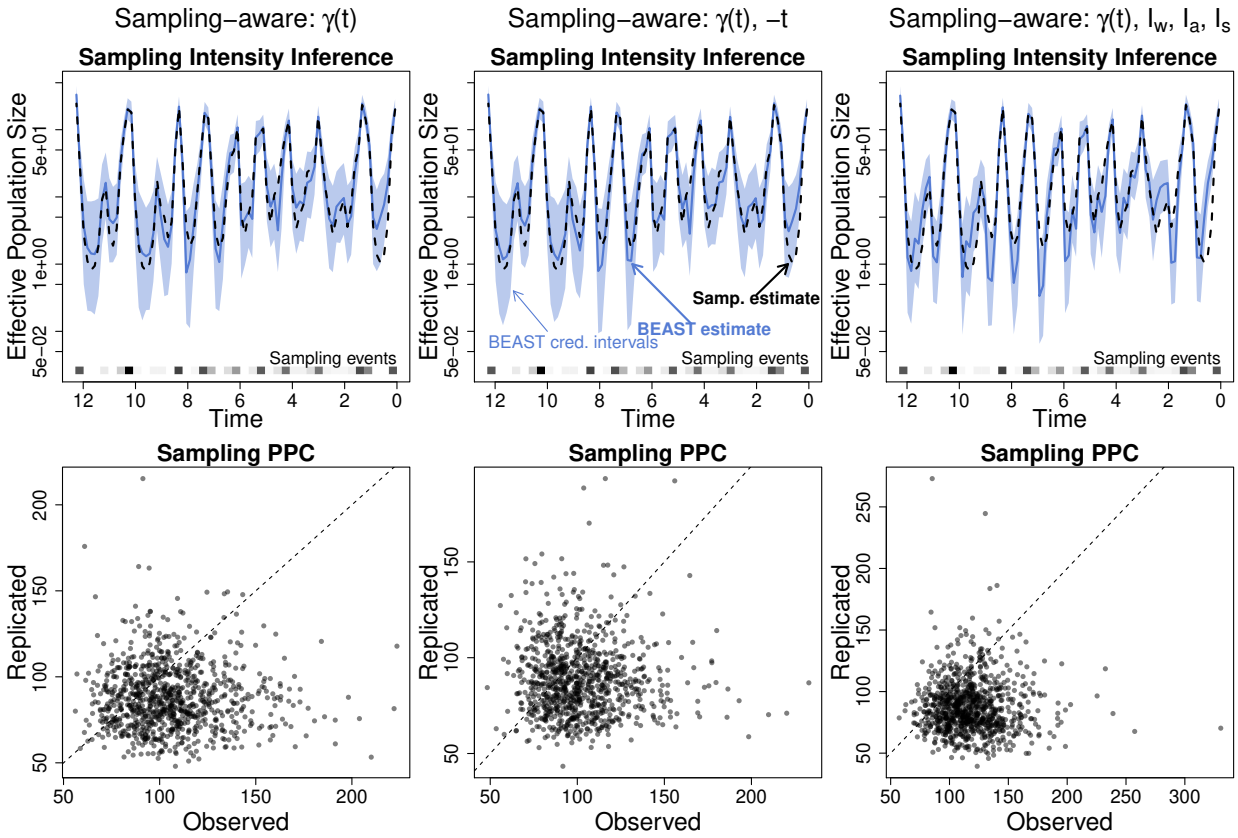

Figure 5.8: **Sampling intensity inference and sampling time posterior predictive check for seasonal influenza data.** The dashed black line represents the true sampling intensity. The solid blue line represents the posterior median sampling intensity inferred by BEAST, and the light blue region represents the corresponding pointwise 95% credible intervals for the sampling intensity.

Sampling Model	Post. Pred. p-val	
	Coalescent	Sampling
Conditional	0.47	—
Aware: $\boldsymbol{\gamma}(t)$	0.48	0.29
Aware: $\boldsymbol{\gamma}(t), -t$	0.47	0.32
Aware: $\boldsymbol{\gamma}(t), I_w, I_a, I_s$	0.49	0.16

Table 5.3: **Posterior predictive p-values for seasonal influenza data.**

summer.

Figure 5.7 shows the inferred effective population size trajectories and coalescent posterior predictive checks for the four models. All four estimated trajectories follow a similar seasonal trajectory, and the discrepancy comparison suggests that the estimated trajectory produces reasonable results with large posterior predictive p-values (Table 5.3). Figure 5.8 shows the inferred sampling intensities compared against a nonparametric sampling time-only estimate of the sampling intensity, as well as sampling posterior predictive checks for the four models. The sampling posterior predictive check produces moderate-to-low posterior predictive p-values, suggesting some model inadequacy manifesting in the sampling intensity estimates.

5.2.3 Ebola Outbreak

Finally, we analyze a subset of sequence data from the recent African Ebola outbreak [Dudas et al., 2017]. The data consists of 1610 samples collected from mid-2014 to mid-2015 aligned across the entire genome (18992 sites). We use the same sequence data `BEAST` framework as the previous section, choosing four different specific sampling time models. We use a sampling-conditional model with no sampling time model and three sampling-aware models with log-intensities $\log[\lambda_s(t)] = \beta_0 + \beta_1 \cdot \boldsymbol{\gamma}(t)$, $\log[\lambda_s(t)] = \beta_0 + \beta_1 \cdot \boldsymbol{\gamma}(t) + \beta_2 \cdot (-t)$, and $\log[\lambda_s(t)] = \beta_0 + \beta_1 \cdot \boldsymbol{\gamma}(t) + \beta_2 \cdot (-t) + \beta_3(-t) \cdot \boldsymbol{\gamma}(t)$.

Figure 5.9 shows the inferred effective population size trajectories and coalescent poste-

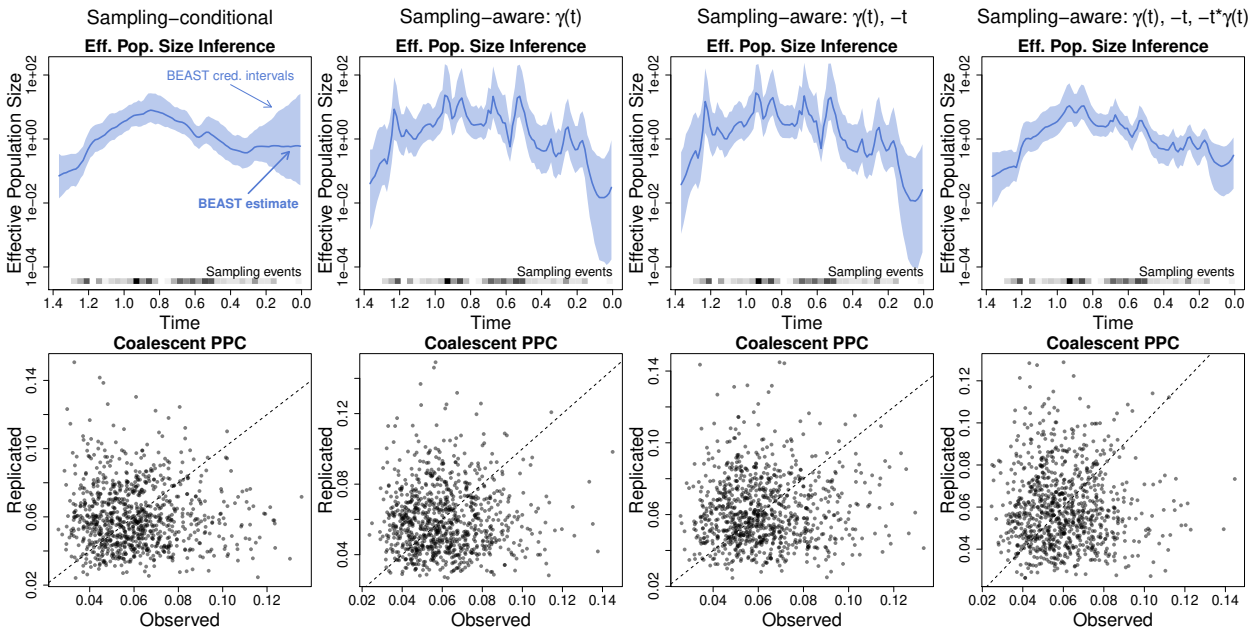

Figure 5.9: **Effective population size inference and coalescent posterior predictive check for Ebola data.** The solid blue line represents the posterior median effective population trajectory inferred by BEAST, and the light blue region represents the corresponding pointwise 95% credible intervals for the effective population trajectory.

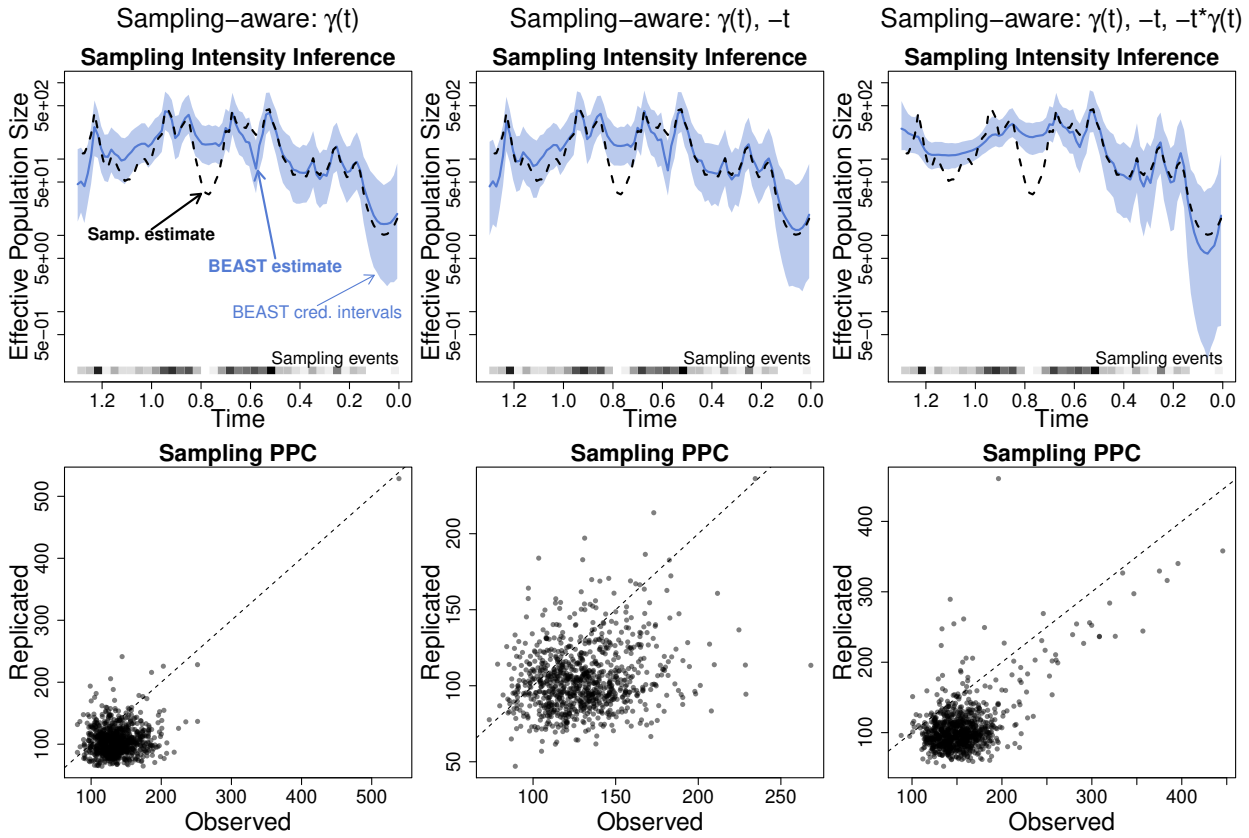

Figure 5.10: **Sampling intensity inference and sampling time posterior predictive check for Ebola data.** The dashed black line represents the true sampling intensity. The solid blue line represents the posterior median sampling intensity inferred by BEAST, and the light blue region represents the corresponding pointwise 95% credible intervals for the sampling intensity.

Sampling Model	Post. Pred. p-val	
	Coalescent	Sampling
Conditional	0.48	—
Aware: $\boldsymbol{\gamma}(t)$	0.47	0.15
Aware: $\boldsymbol{\gamma}(t), -t$	0.50	0.18
Aware: $\boldsymbol{\gamma}(t), -t, -t \cdot \boldsymbol{\gamma}(t)$	0.50	0.06

Table 5.4: **Posterior predictive p-values for Ebola data.**

rior predictive checks for the four models. All four estimated trajectories follow a similar effective population size trajectory that visually resembles a typical time trajectory of prevalence or incidence that peaks in Autumn of 2014. The discrepancy comparison suggests that the estimated trajectory produces reasonable results with large posterior predictive p-values (Table 5.4). Figure 5.10 shows the inferred sampling intensities compared against a nonparametric sampling time-only estimate of the sampling intensity, as well as sampling posterior predictive checks for the four models. The sampling posterior predictive check produces small posterior predictive p-values (Table 5.4), suggesting notable model inadequacy manifesting in the sampling intensity estimates.

BIBLIOGRAPHY

E. Çinlar. Introduction to Stochastic Processes. Prentice-Hall, Inc., Englewood Cliffs, New Jersey, 1975.

P. J. Diggle, R. Menezes, and T. Su. Geostatistical inference under preferential sampling. Journal of the Royal Statistical Society: Series C (Applied Statistics), 59(2):191–232, 2010.

A. Drummond, R. Forsberg, and A. G. Rodrigo. The inference of stepwise changes in substitution rates using serial sequence samples. Molecular Biology and Evolution, 18 (7):1365–1371, 2001.

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

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

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

A.J. Drummond, O.G. Pybus, A. Rambaut, R. Forsberg, and A.G. Rodrigo. Measurably evolving populations. Trends in Ecology & Evolution, 18(9):481–488, 2003.

G. Dudas, L. M. Carvalho, T. Bedford, A. J. Tatem, G. Baele, N. R. Faria, D. J. Park, J. T. Ladner, A. Arias, D. Asogun, et al. Virus genomes reveal factors that spread and sustained the ebola epidemic. Nature, 544(7650):309–315, 2017.

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

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

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

J. Felsenstein and A. G. Rodrigo. Coalescent Approaches to HIV Population Genetics. In The Evolution of HIV, pages 233–272. Johns Hopkins University Press, 1999. ISBN 9780801861512.

C. Fraser, C. A. Donnelly, S. Cauchemez, W. P. Hanage, M. D. Van Kerkhove, T. D. Hollingsworth, J. Griffin, R. F. Baggaley, H. E. Jenkins, E. J. Lyons, et al. Pandemic potential of a strain of influenza a (h1n1): early findings. science, 2009.

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

Andrew Gelman, Xiao-Li Meng, and Hal Stern. Posterior predictive assessment of model fitness via realized discrepancies. Statistica sinica, pages 733–760, 1996.

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

M.S Gill, P. Lemey, S.N. Bennett, R. Biek, and M.A. Suchard. Understanding past population dynamics: Bayesian coalescent-based modeling with covariates. Systematic Biology, 65(6):1041–1056, 2016.

E. Goldstein, S. Cobey, S. Takahashi, J. C. Miller, and M. Lipsitch. Predicting the epidemic sizes of influenza A/H1N1, A/H3N2, and B: a statistical method. PLoS medicine, 8(7): 952, 2011.

R. R. Gray, M. Salemi, P. Klenerman, and O. G. Pybus. A new evolutionary model for hepatitis c virus chronic infection. PLoS pathogens, 8(5):e1002656, 2012.

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

R. C Griffiths and S. Tavaré. Ancestral inference in population genetics. Statistical science, pages 307–319, 1994a.

R. C. Griffiths and S. Tavaré. 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, 1994b.

M. Hasegawa, H. Kishino, and T. Yano. Dating of the human-ape splitting by a molecular clock of mitochondrial dna. Journal of Molecular Evolution, 22(2):160–174, 1985.

J. Hein, M. Schierup, and C. Wiuf. Gene genealogies, variation and evolution: a primer in coalescent theory. Oxford University Press, USA, 2004.

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

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

E. C. Holmes and B. T. Grenfell. Discovering the phylodynamics of RNA viruses. PLoS Computational Biology, 5(10):e1000505, 2009.

T. H. Jukes, C. R. Cantor, H. N. Munro, et al. Evolution of protein molecules. Mammalian protein metabolism, 3(21):132, 1969.

M. D. Karcher, J. A. Palacios, T. Bedford, M. A. Suchard, and V. N. Minin. Quantifying and mitigating the effect of preferential sampling on phylodynamic inference. PLoS Computational Biology, 12:e1004789, 2016a.

M. D. Karcher, J. A Palacios, S. Lan, and V. N. Minin. phylodyn: an R package for phylodynamic simulation and inference. Molecular ecology resources, 2016b.

M.D. Karcher, J.A. Palacios, S. Lan, and V.N. Minin. phylodyn: an R package for phylodynamic simulation and inference. Molecular Ecology Resources, 17(1):96–100, 2017.

M. Kimura. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. Journal of molecular evolution, 16(2):111–120, 1980.

J. F. C. Kingman. The coalescent. Stochastic Processes and Their Applications, 13(3): 235–248, 1982.

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

S. Lan, J. A. Palacios, M. D. Karcher, V. N. Minin, and B. Shahbaba. An efficient Bayesian inference framework for coalescent-based nonparametric phylodynamics. Bioinformatics, 31:3282–3289, 2015.

P. Lemey, A. Rambaut, and O. G. Pybus. Hiv evolutionary dynamics within and among hosts. AIDS Rev, 8(3):125–140, 2006.

T. G. Martins, D. Simpson, F. Lindgren, and H. Rue. Bayesian computing with INLA: new features. Computational Statistics & Data Analysis, 67:68–83, 2013.

Frank J Massey Jr. The kolmogorov-smirnov test for goodness of fit. Journal of the American statistical Association, 46(253):68–78, 1951.

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

I. Murray, R.P. Adams, and D. Mackay. Elliptical slice sampling. In International Conference on Artificial Intelligence and Statistics, pages 541–548, 2010.

J. A. Palacios and V. N. Minin. Integrated nested Laplace approximation for Bayesian non-parametric phylodynamics. In Proceedings of the Twenty-Eighth International Conference on Uncertainty in Artificial Intelligence, pages 726–735, 2012.

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

K. Pearson. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science, 50(302):157–175, 1900.

O. G. Pybus, A. Rambaut, and P. H. Harvey. An integrated framework for the inference of viral population history from reconstructed genealogies. Genetics, 155(3):1429–1437, 2000.

O. G. Pybus, M. A. Charleston, S. Gupta, A. Rambaut, E. C. Holmes, and P. H. Harvey. The epidemic behavior of the hepatitis c virus. Science, 292(5525):2323–2325, 2001.

A. Rambaut and N. C. Grass. Seq-gen: an application for the monte carlo simulation of dna sequence evolution along phylogenetic trees. Bioinformatics, 13(3):235–238, 1997.

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

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

H. Rue, S. Martino, and N. Chopin. 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, 2009.

B. Shapiro, A. Rambaut, and A.J. Drummond. Choosing appropriate substitution models for the phylogenetic analysis of protein-coding sequences. Molecular Biology and Evolution, 23(1):7–9, 2005.

Yue-Long Shu, Li-Qun Fang, Sake J de Vlas, Yan Gao, Jan Hendrik Richardus, and Wu-Chun Cao. Dual seasonal patterns for influenza, china. Emerging Infectious Diseases, 16 (4):725, 2010.

S. Sinharay and H. S. Stern. Posterior predictive model checking in hierarchical models. Journal of Statistical Planning and Inference, 111(1):209–221, 2003.

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

J. C. Stack, J. D. Welch, M. J. Ferrari, B. U. Shapiro, and B. T. Grenfell. Protocols for sampling viral sequences to study epidemic dynamics. Journal of The Royal Society Interface, 7(48):1119–1127, 2010.

Tanja Stadler. Sampling-through-time in birth–death trees. Journal of Theoretical Biology, 267(3):396–404, 2010.

W. M. Van Ballegooijen, S. M. Van Houdt, R.and Bruisten, H. J. Boot, R. A. Coutinho, and J. Wallinga. Molecular sequence data of hepatitis b virus and genetic diversity after vaccination. American journal of epidemiology, 170(12):1455–1463, 2009.

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

E. M. Volz, S. L. K. Pond, M. J. Ward, A. J. L. Brown, and S. D. W. Frost. Phylodynamics of infectious disease epidemics. Genetics, 2009.

J. Wakeley and O. Sargsyan. Extensions of the coalescent effective population size. Genetics, 181(1):341–345, 2009.

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

D. Zinder, T. Bedford, E. B. Baskerville, R. J. Woods, M. Roy, and M. Pascual. Seasonality in the migration and establishment of H3N2 influenza lineages with epidemic growth and decline. BMC Evolutionary Biology, 14(1):272, 2014.

Appendix A

ADDITIONAL FIXED-TREE RESULTS

A.1 Hyperproportional simulations

To explore preferential sampling relationships with greater clustering than direct proportionality, or *hyperproportional* preferential sampling, we also perform a simulation study with $\beta_1 = 2, 3$. Figure A-1 shows the pointwise statistics, while Table A-1 lists the time interval statistics. The results are largely consistent with $\beta_1 = 1$, but with slightly more bias under the BNPR.

	$\beta_1 = 2.0$ $(6, 48)$		$\beta_1 = 3.0$ $(6, 48)$		$\beta_1 = 2.0$ $(0, 6)$		$\beta_1 = 3.0$ $(0, 6)$	
	BNPR	BNPR-PS	BNPR	BNPR-PS	BNPR	BNPR-PS	BNPR	BNPR-PS
MRD	0.240	0.146	0.242	0.147	3.784	0.237	3.839	0.311
MRW	1.865	0.979	1.919	0.938	51.507	1.413	57.701	1.444
ME	0.989	0.987	0.990	0.978	0.852	0.988	0.891	0.972

Table A-1: **Averaged time interval summary statistics of the hyperproportional simulations.** Over the interval $(6, 48)$ where both methods perform well, and the most recent interval $(0, 6)$ where BNPR-PS performs considerably better.

A.2 Negative control simulations

In the negative control simulations section above, we seek to differentiate between misspecification error due to preferential sampling and error due to few observations (long periods without coalescent events). In Figure A-2 we simulated sampling times from random piecewise constant sampling intensities independent from effective population size. We generated

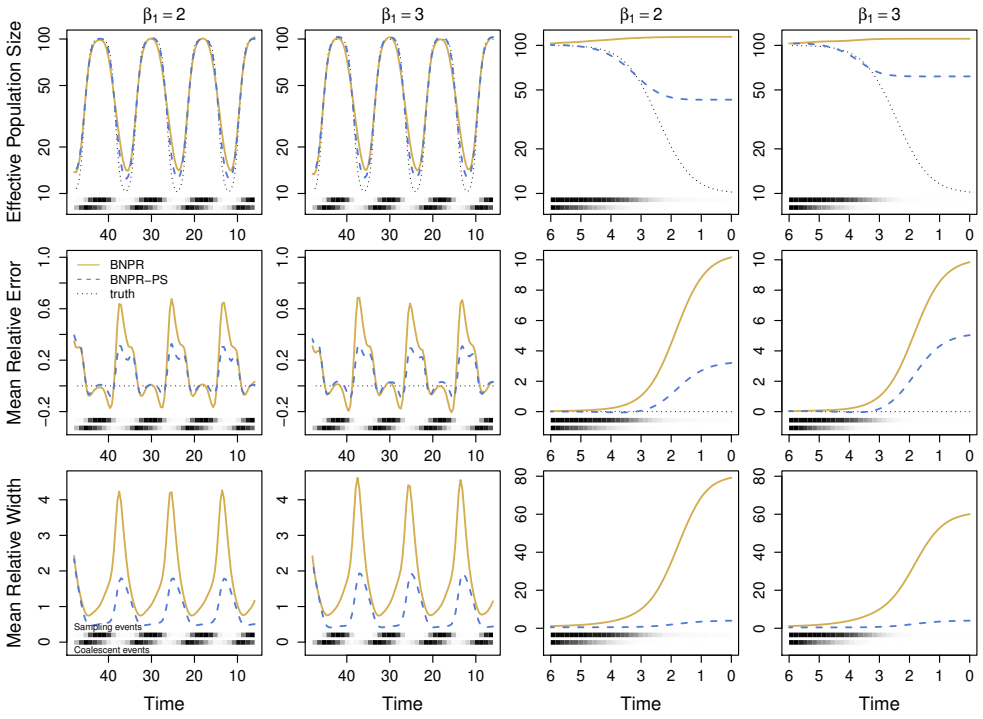

Figure A-1: **Comparison of pointwise statistics with hyperproportional preferential sampling.** Dotted lines represent the truth, where applicable. Solid yellow lines represent the conditional method BNPR (ignoring preferential sampling), while dashed blue lines represent the sampling-aware method BNPR-PS (accounting for preferential sampling). The first row shows true and estimated effective population sizes, the second shows mean relative error, while the third shows mean relative width of the 95% Bayesian credible interval. The left two columns show the interval $(6, 48)$ where both models perform at their best. The right two columns show $(0, 6)$, where BNPR-PS performs significantly better. At the bottom of each plot, the distribution of sampling events (above) and coalescent events (below) are shown. Time is in months.

the sampling intensity trajectories by selecting change points such that the trajectory has a similar number of low-to-high and high-to-low transitions to the seasonal trajectory used in our simulation study, as well as similar amounts of time with high and low intensity. In Figure A-3 we simulated sampling times from a Gaussian process sampling intensity, also independent from effective population size. We generated the sampling intensity trajectories by selecting Gaussian process realizations that have similar ranges of values and number of transitions to the seasonal trajectory used in our simulation study. Under BNPR, we see relative errors and relative widths less severe than in the case of preferential sampling as we see in Figure 3.2, 3.3, and A-1. However, the performance of BNPR-PS suffers due to the wildly changing ratios of sampling intensity and effective population size.

A.3 Parametric simulations

For completeness, we also explored model misspecification in a correctly-specified parametric context. We use an exponential effective population size trajectory $N_e(t) = \exp(a + bt)$, seeking to simulate a growth scenario and a decline scenario. Working backwards in time, perpetual exponential decline is impossible and results in potentially unbounded times to most recent common ancestor (TMRCA). In order to maintain the correctness of our model, we need a values sufficiently small as to result in reasonable TMRCA in our 100,000 genealogies. We also choose b values so that across the sampling window $t \in [0, 10]$ the effective population size will change by a realistic one order of magnitude. Finally, we choose a values that make ranges of the effective population size trajectories in the sampling interval comparable under both growth and decline scenarios. We select $(a, b) = (-\log(10), 0.2)$ and $(-\log(100), -0.2)$. For both uniform and proportional sampling schedules, we simulated 100,000 collections of sampling times between $t = 0$ and $t = 10$, expecting 500 sampling times each collection. We then simulated genealogies for each sampling time collection from the coalescent as in our other simulation studies. We applied an exponential growth/decline parametric maximum likelihood method and summarized the results in Table A-2. In both uniform and preferential sampling we see small, but comparable biases in estimates of parameter a. However,

Growth	$a = -\log(10) = -2.30$			$b = 0.2$		
	Bias	(±2SE)		Bias	(±2SE)	
Unif	-0.00216	(-0.00273,	-0.00158)	-0.00003	(-0.00012,	0.00008)
Pref	0.00128	(0.00069,	0.00187)	0.00066	(0.00054,	0.00078)

Decline	$a = -\log(100) = -4.61$			$b = -0.2$		
	Bias	(±2SE)		Bias	(±2SE)	
Unif	-0.00165	(-0.00222,	-0.00108)	0.00003	(-0.00006,	0.00013)
Pref	0.00189	(0.00088,	0.00291)	0.00042	(0.00028,	0.00055)

Table A-2: **Estimates and confidence intervals for the bias of estimating the parameters of a correctly specified exponential growth/decline model with preferential sampling.**

estimates of the exponential growth rate parameter b have no detectable bias under uniform sampling, but have small but significant bias under preferential sampling. This verifies that ignoring preferential sampling causes systematic bias, perhaps of small magnitude, in maximum likelihood phylodynamic estimation even under a simple low-dimensional parametric model.

A.4 Regional influenza

We examine three of the remaining regions more closely in Figure A-4 and A-5 and the final three regions in Figure A-6 and A-7. We see much more pronounced seasonality in the estimated effective population size trajectories produced by BNPR-PS, as well as noticeable improvements in the relative widths of the Bayesian credible intervals. We see three regions with unusual results. Whereas most of the regions show some seasonality under BNPR which becomes more visible under BNPR-PS, India and Southeast Asia both have little seasonality

under both methods. Furthermore, South America has little seasonality under BNPR but some seasonality appears under BNPR-PS. We suspect that these results may be due to inclusion of countries with different flu seasonal patterns into this region.

As a final experiment, we explored the variability in our results when we allow for uncertainty in our genealogical inference. Figure A-8 shows the results of running BNPR and BNPR-PS on eight randomly chosen trees from our BEAST run. We see that the inferred effective population size trajectories are quite similar, suggesting the potential for some robustness to uncertainty associated with genealogical inference.

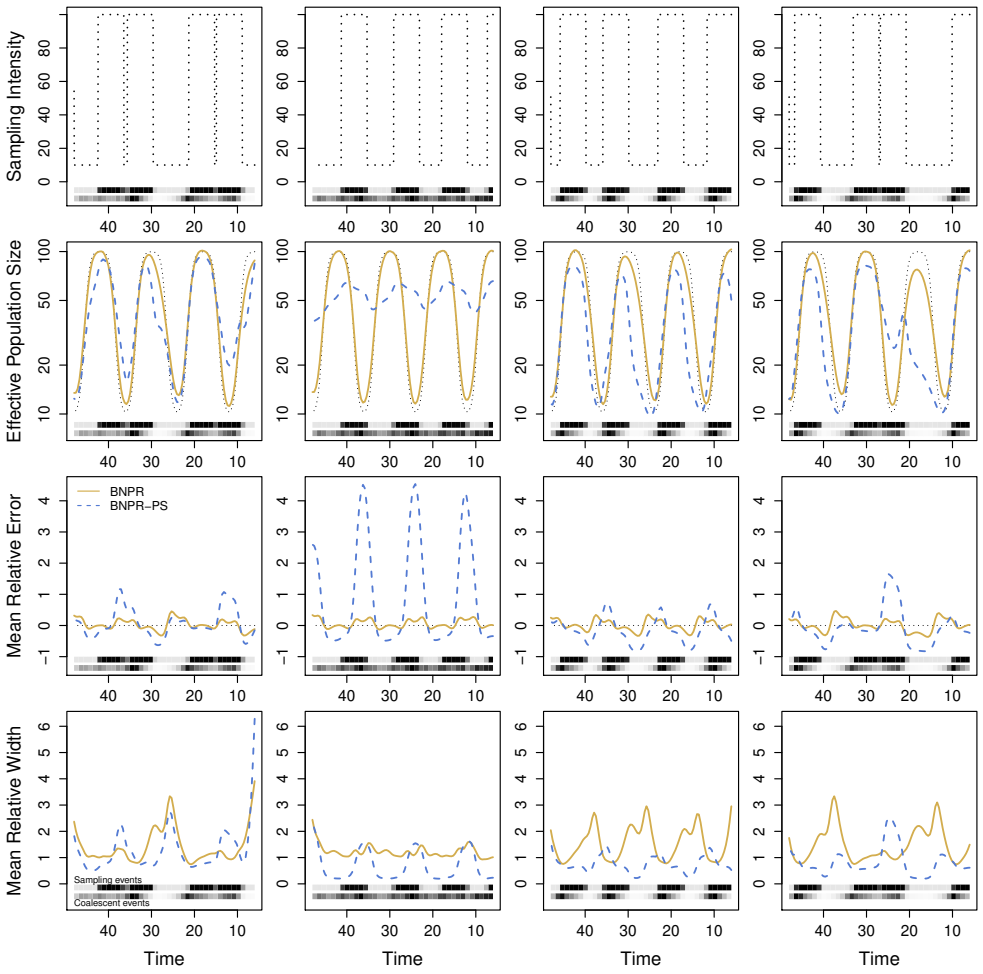

Figure A-2: **Comparison of pointwise statistics for a randomly generated piecewise constant sampling intensity trajectory independent of effective population size.** Dotted lines represent the sampling intensity trajectory and true effective population size trajectory. Solid yellow lines represent the conditional method BNPR, while the dashed blue lines represent the sampling-aware model BNPR-PS. The first row shows the sampling intensity, the second shows true and estimated effective population sizes, the third shows mean relative error, while the fourth shows mean relative width of the 95% Bayesian credible interval. The columns represent four realizations of the random sampling intensity trajectory. At the bottom of each plot, the distribution of sampling events (above) and coalescent events (below) are shown.

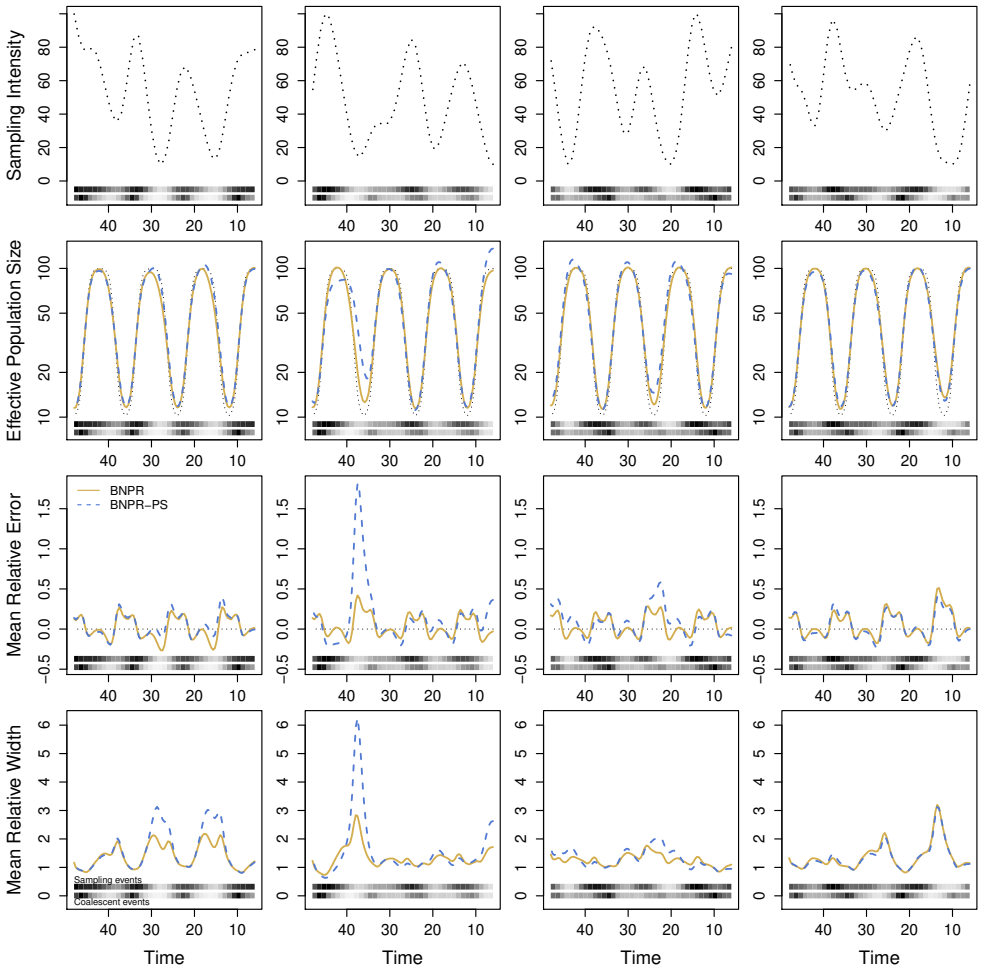

Figure A-3: **Comparison of pointwise statistics for a randomly generated Gaussian process sampling intensity independent of effective population size.** Visuals as in Figure A-2.

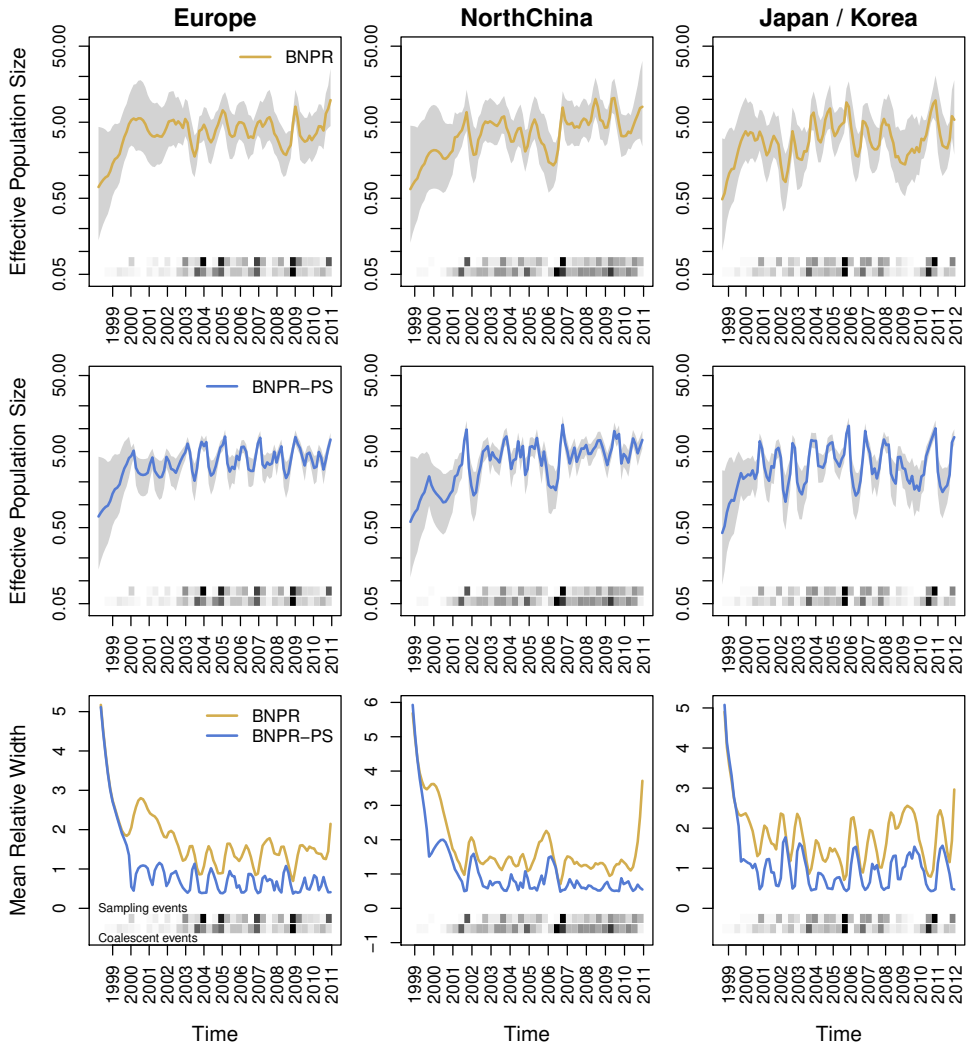

Figure A-4: **BNPR and BNPR-PS models applied to the genealogies inferred from the regional influenza example.** We see moderate correlation between effective population size $N^{\gamma}(t)$ and sampling frequencies in the data (Table 3.2). We see improvements in Bayesian credible interval widths, and BNPR-PS performs as well or better than BNPR everywhere in these examples.

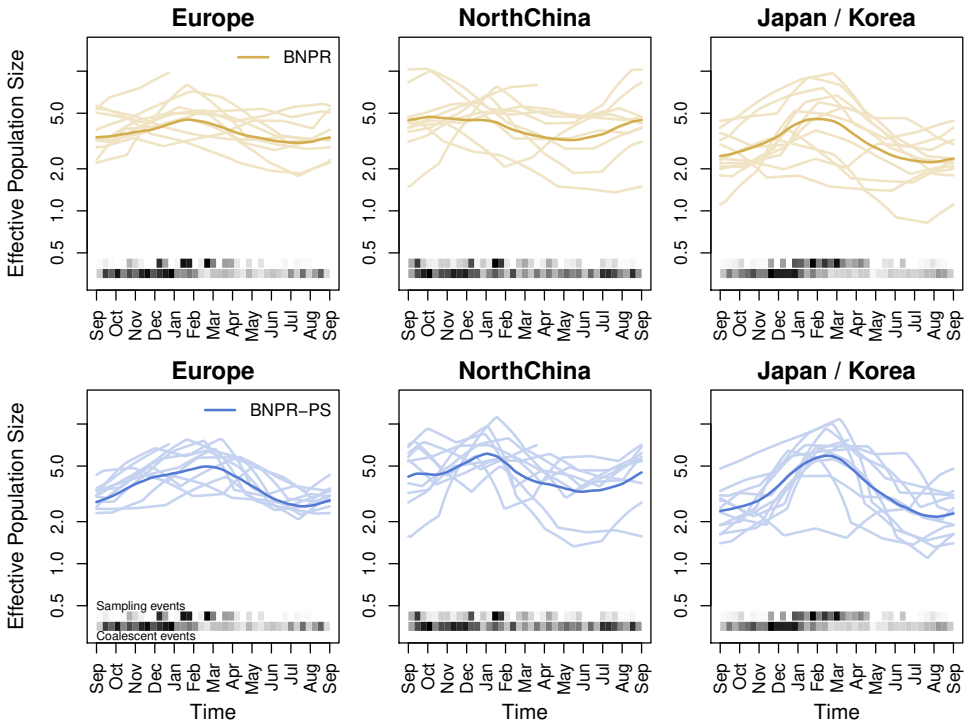

Figure A-5: **BNPR and BNPR-PS models applied to the genealogies inferred from the regional influenza example with years overlaid.** We see more pronounced seasonality in the BNPR-PS plots.

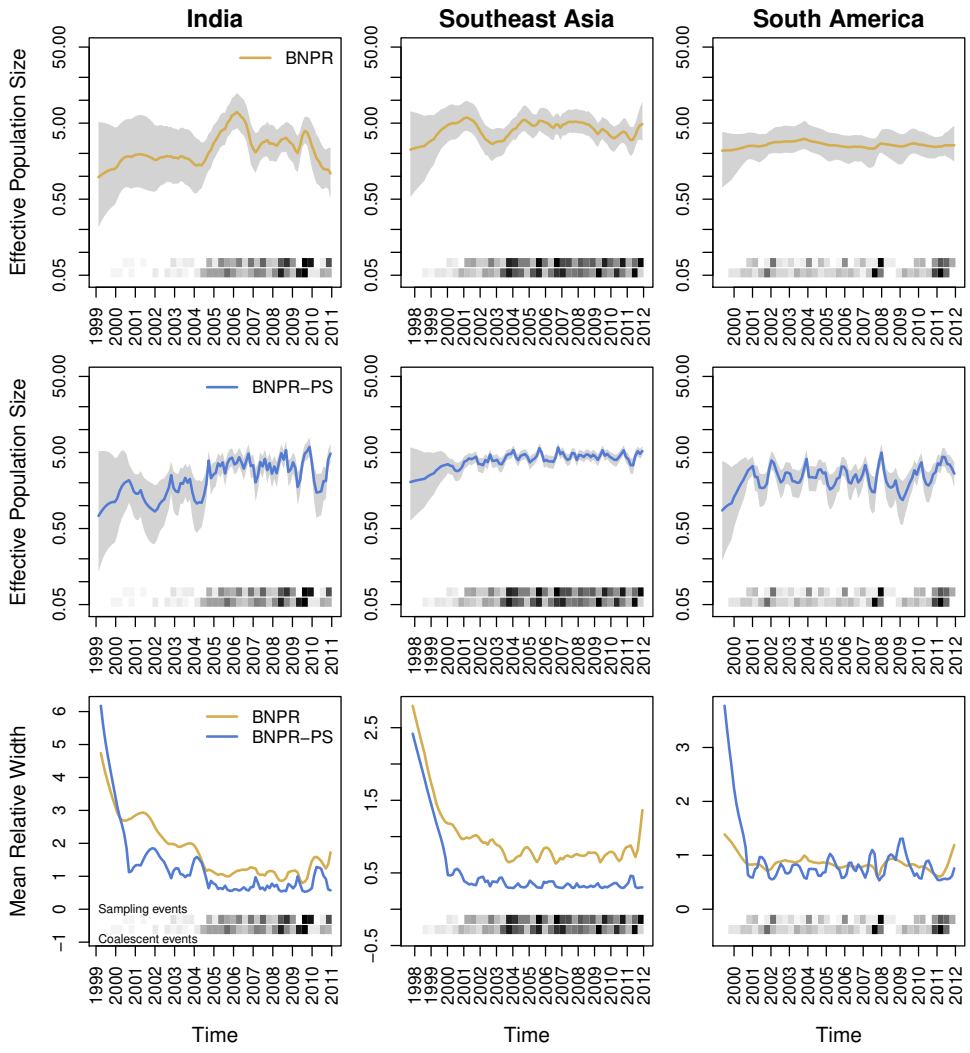

Figure A-6: **BNPR and BNPR-PS models applied to the genealogies inferred from the regional influenza example.** Visuals as in Figure A-4. In South America, we see moderate correlation between effective population size $N^{\gamma}(t)$ and sampling frequencies in the data (Table 3.2). We see improvements in Bayesian credible interval widths, and BNPR-PS performs as well or better than BNPR everywhere in these examples.

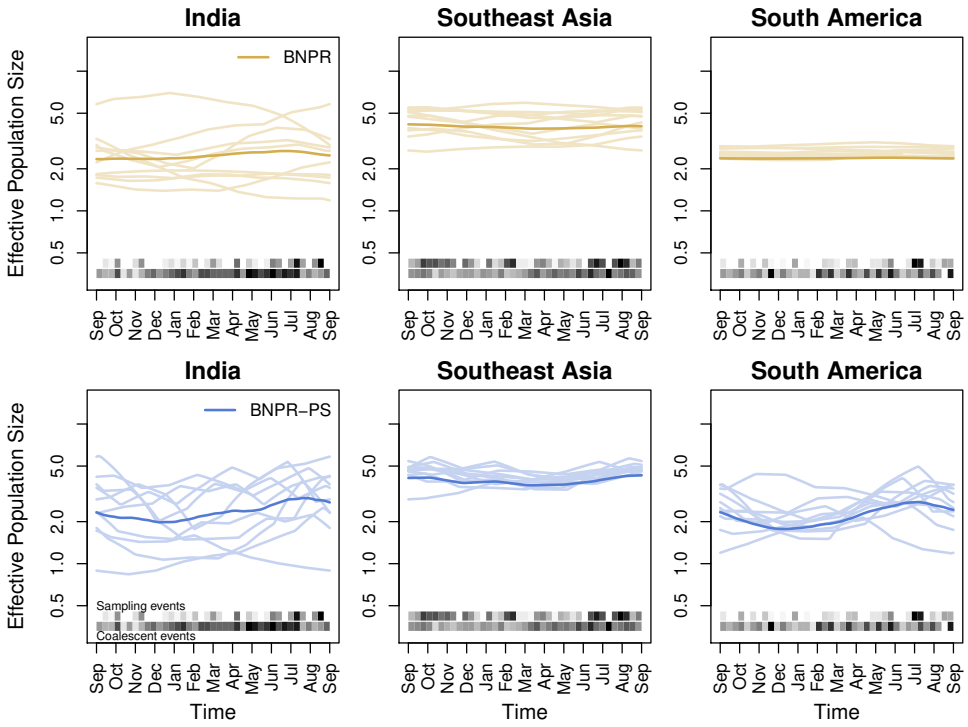

Figure A-7: **BNPR and BNPR-PS models applied to the genealogies inferred from the regional influenza example with years overlaid.** We see more pronounced seasonality in the BNPR-PS plots.

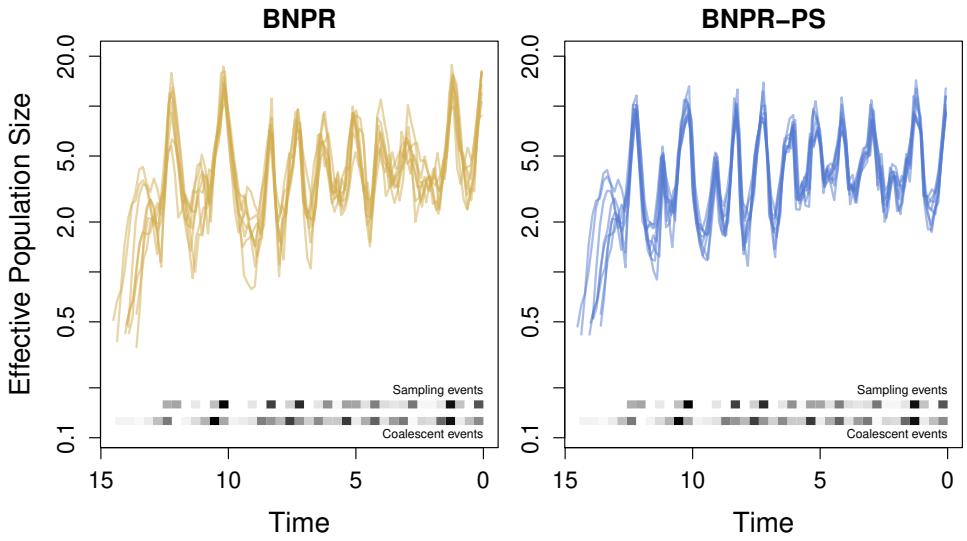

Figure A-8: **BNPR and BNPR-PS models applied to eight randomly selected genealogies from our BEAST inference.**

Appendix B

ADDITIONAL SEQUENCE DATA RESULTS

B.1 Seasonal Influenza

We consider one additional model for the USA/Canada influenza data with log-intensity,

$$\begin{aligned}\beta_0 + \beta_1\gamma(t) + \beta_2 I_{\text{winter}}(t) + \beta_3 I_{\text{autumn}}(t) + \beta_4 I_{\text{summer}}(t) \\ + \delta_2 I_{\text{winter}}(t)\cdot\gamma(t) + \delta_3 I_{\text{autumn}}(t)\cdot\gamma(t) + \delta_4 I_{\text{summer}}(t)\cdot\gamma(t),\end{aligned}$$

abbreviated $\{\gamma(t),\ I_{\text{winter}},\ I_{\text{autumn}},\ I_{\text{summer}},\ I_{\text{winter}}\cdot\gamma(t),\ I_{\text{autumn}}\cdot\gamma(t),\ I_{\text{summer}}\cdot\gamma(t)\}$, or more succinctly as $\{\gamma(t),\ I_{\text{winter}},\ I_{\text{autumn}},\ I_{\text{summer}},\ \text{interactions}\}$. The results are summarized in Figure B-1 and Table B-1. We see that only the coefficients for $\gamma(t)$, I_{winter}, and I_{autumn} have credible intervals that do not contain zero, suggesting that additional terms are not necessary.

B.2 Ebola Outbreak

We consider three additional models for our subsample of 200 sequences from the Sierra Leone Ebola outbreak data with log-intensities,

$$\begin{aligned}&\beta_0 + \beta_1\gamma(t) + \beta_2\cdot(-t) + \beta_3\cdot(-t^2),\\ &\beta_0 + \beta_1\gamma(t) + \beta_2\cdot(-t) + \beta_3\cdot(-t^2)\\ &\qquad\qquad + \delta_2\gamma(t)\cdot(-t) + \delta_3\gamma(t)\cdot(-t^2),\ \text{and}\\ &\beta_0 + \beta_1\gamma(t) + \delta_2\gamma(t)\cdot(-t) + \delta_3\gamma(t)\cdot(-t^2),\end{aligned}$$

abbreviated as $\{\gamma(t),\ -t, -t^2\}$, $\{\gamma(t), -t, -t^2,\ -t\cdot\gamma(t),\ -t^2\cdot\gamma(t)\}$, and $\{\gamma(t),\ -t\cdot\gamma(t),\ -t^2\cdot\gamma(t)\}$, respectively. The results are summarized in Figure B-2 and Table B-2. We see that the coefficients for $\gamma(t)$, $-t$, and $-t^2$ tend to have credible intervals that do not contain

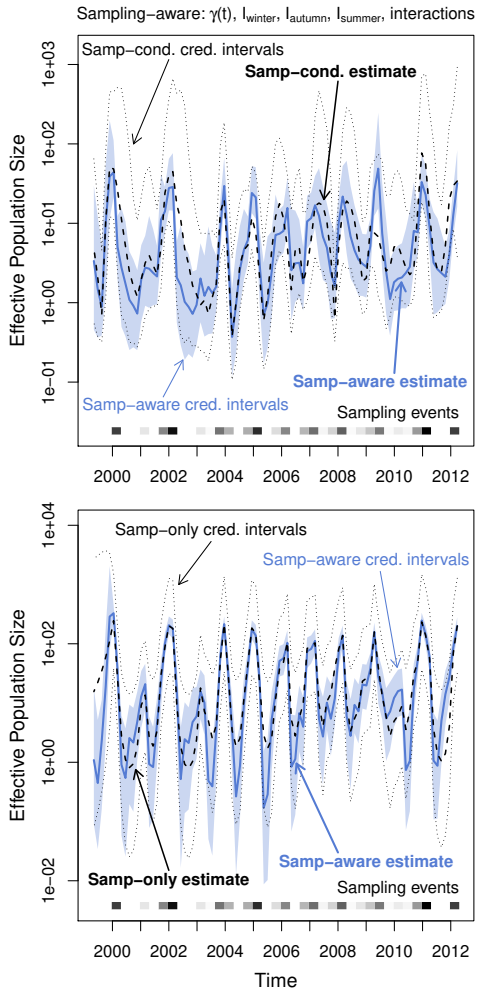

Figure B-1: **Effective population size and sampling rate reconstructions for the USA and Canada influenza dataset.** *Upper row*: Dashed lines and dotted black lines are the pointwise posterior effective population size estimates and credible intervals of the sampling-conditional model. The blue line and the light blue region are the pointwise posterior effective population size estimates and credible intervals of that column's sampling-aware model. *Lower row*: Dashed lines and dotted black lines are the pointwise posterior sampling rate estimates and credible intervals of a nonparametric sampling-time-only model. The blue line and the light blue region are the pointwise posterior sampling rate estimates and credible intervals of that column's sampling-aware model.

Model	Coef	Q0.025	Median	Q0.975
$\{\gamma(t), I_{\text{winter}}, I_{\text{autumn}}, I_{\text{summer}}, \text{interactions}\}$	$\gamma(t)$	0.64	1.20	1.97
	I_{winter}	1.83	3.48	5.58
	I_{autumn}	1.31	3.16	5.28
	I_{summer}	-0.15	2.08	4.52
	$I_{\text{winter}} \cdot \gamma(t)$	-1.08	-0.29	0.34
	$I_{\text{autumn}} \cdot \gamma(t)$	-1.00	-0.14	0.53
	$I_{\text{summer}} \cdot \gamma(t)$	-1.24	-0.24	0.92

Table B-1: **Summary of USA/Canada influenza data inference.** Posterior distribution quantile summaries for SampESS with seasonal indicator and interaction covariates (model: $\{\gamma(t), I_{\text{winter}}, I_{\text{autumn}}, I_{\text{summer}}$,interactions$\}$).

zero (except for the interaction-only model $\{\gamma(t), -t \cdot \gamma(t), -t^2 \cdot \gamma(t)\}$), but the other terms do not, suggesting that the additional terms are not necessary.

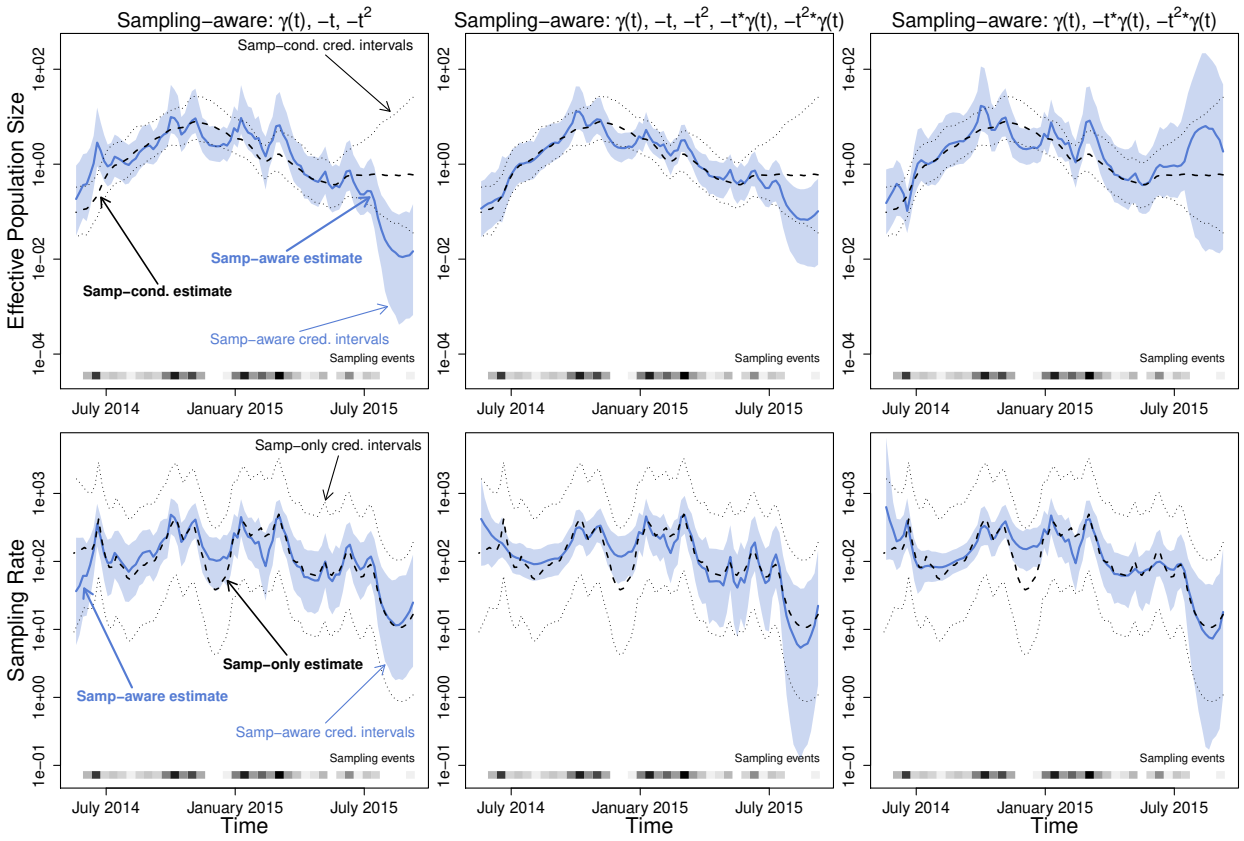

Figure B-2: **Effective population size and sampling rate reconstructions for the Sierra Leone Ebola dataset.** *Upper row*: Dashed lines and dotted black lines are the pointwise posterior effective population size estimates and credible intervals of the sampling-conditional model. The blue line and the light blue region are the pointwise posterior effective population size estimates and credible intervals of that column's sampling-aware model. *Lower row*: Dashed lines and dotted black lines are the pointwise posterior sampling rate estimates and credible intervals of a nonparametric sampling-time-only model. The blue line and the light blue region are the pointwise posterior sampling rate estimates and credible intervals of that column's sampling-aware model.

Model	**Coef**	**Q0.025**	**Median**	**Q0.975**
$\{\gamma(t), -t, -t^2\}$	$\gamma(t)$	0.47	1.00	1.80
	$-t$	2.02	9.05	20.63
	$-t^2$	-13.08	-5.58	-1.09
$\{\gamma(t), -t, -t^2, -t \cdot \gamma(t), -t^2 \cdot \gamma(t)\}$	$\gamma(t)$	0.71	2.20	4.69
	$-t$	1.21	9.75	20.29
	$-t^2$	-12.67	-6.00	-0.79
	$-t \cdot \gamma(t)$	-2.72	0.95	6.49
	$-t^2 \cdot \gamma(t)$	-2.64	0.72	3.26
$\{\gamma(t), -t \cdot \gamma(t), -t^2 \cdot \gamma(t)\}$	$\gamma(t)$	-3.00	-1.39	2.08
	$-t \cdot \gamma(t)$	-11.30	-6.59	2.00
	$-t^2 \cdot \gamma(t)$	-0.16	4.90	8.16

Table B-2: **Summary of Sierra Leone Ebola sequence data inference.** Posterior distribution quantile summaries for SampESS with models: $\{\gamma(t), -t, -t^2\})$, $\{\gamma(t), -t, -t^2, -t \cdot \gamma(t), -t^2 \cdot \gamma(t)\}$, and $\{\gamma(t), -t \cdot \gamma(t), -t^2 \cdot \gamma(t)\}$.