Statistical, algorithmic, and robustness aspects of population demographic inference from genomic variation data

by

Anand Bhaskar

A dissertation submitted in partial satisfaction of the
requirements for the degree of
Doctor of Philosophy

in

Computer Science

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Yun S. Song, Chair
Professor Steven N. Evans
Professor Montgomery Slatkin
Professor David Tse

Fall 2013

Statistical, algorithmic, and robustness aspects of population demographic inference from genomic variation data

Abstract

Statistical, algorithmic, and robustness aspects of population demographic inference from genomic variation data

by

Anand Bhaskar

Doctor of Philosophy in Computer Science

University of California, Berkeley

Professor Yun S. Song, Chair

The recent availability of large-sample high-throughput sequencing data has given us an unprecedented opportunity to very finely resolve the details of historical demographic processes that have shaped the genomes of modern human populations. Such understanding of population demography is important for several applications — to avoid false positives in genome-wide association studies; to calibrate null models of neutral genome evolution in order to find regions under selection; to study the impact of bottlenecks and small founder populations on genetic mutational load; to reconstruct large-scale historical human migration and admixture events; and so on.

In this dissertation, we consider some statistical, algorithmic and robustness aspects of demographic inference from genomic variation data. In particular, we study the problem of determining the historical effective size of a population from the sample frequency spectrum (SFS), which measures the distribution of allele frequencies in a sample of sequences drawn from the population.

From the statistical or information-theoretic perspective, it is known that this inverse problem does not have a unique solution in general, no matter how large the sample size. For any population allele frequency distribution, there exist infinitely many population size functions that are consistent with this distribution. While such a non-identifiability result might appear to pose a serious problem to statistical inference algorithms, we show that the situation is not so bad in practice. In particular, we prove that if the true population size function is *piecewise-defined* with each piece belonging to some family of biologically-motivated functions, then the SFS of a finite sample of sequences uniquely determines the underlying demography. We obtain a general bound on the sample size sufficient for identifiability; this bound depends on the number of pieces in the demographic model and on the family of functions for each piece. We also give concrete instantiations of this bound for piecewise-constant and piecewise-exponential models that are commonly used in demographic inference analyses.

From the algorithmic perspective, we build on analytic results for the expected SFS of a time-varying population size function and develop an efficient likelihood-based algorithm to infer piecewise-exponential population size histories from large sample allele frequency data. By considering very large samples, our method can resolve details of the population history from the very recent past that are not otherwise accessible using smaller samples.

The third aspect of this dissertation is concerned with understanding the robustness of widely used evolutionary models to violations of model assumptions. Continuous-time evolutionary models like Kingman's coalescent and its dual diffusion process are derived from discrete models of random mating by assuming that the sample size being analyzed is much smaller than the the population size. However, the very large sample datasets being produced due to advances in high-throughput sequencing technologies are approaching the limits of this assumption. To investigate this issue, we develop exact algorithms for computation under the discrete-time Wright-Fisher model and use these algorithms to study the distortions in several genealogical quantities arising due to the coalescent approximation. Our findings indicate that for several demographic models inferred from large-scale sequence data, there can be substantial genealogical deviations introduced by the coalescent approximation that might influence the results of inference studies.

To appa and amma, for instilling in me the importance of learning.

Contents

1 Introduction **1**
1.1 Motivation . . . 1
1.2 Demographic inference . . . 2
1.2.1 Importance of understanding demography . . . 2
1.2.2 Representations of genomic data . . . 2
1.3 Preliminaries . . . 4
1.3.1 Wright-Fisher model . . . 4
1.3.2 Kingman's coalescent . . . 5
1.3.3 Extension to variable population size . . . 6
1.3.4 Sample frequency spectrum . . . 7
1.4 Overview . . . 8
1.4.1 Statistical . . . 8
1.4.2 Algorithmic . . . 9
1.4.3 Robustness . . . 9

2 Identifiability of the population size function from SFS data **10**
2.1 Overview . . . 10
2.2 Notation . . . 12
2.3 Main results . . . 13
2.3.1 Determining expected times to the first coalescence event from the SFS 13
2.3.2 Piecewise population size models and sign change complexities . . . 14
2.3.3 Identifiability from the SFS . . . 18
2.3.4 Extension to the folded frequency spectrum . . . 22
2.4 The counterexample of Myers *et al.* . . . 24
2.5 Discussion . . . 25

3 Inference of the population size function from SFS data **27**
3.1 Overview . . . 27
3.2 Notation . . . 29
3.3 Method . . . 29
3.3.1 Poisson Random Field approximation . . . 30
3.3.2 Computing the expected SFS under a variable population size . . . 32

3.4 Results 34

4 Distortion of genealogical properties for very large sample sizes 41
4.1 Demographic models 41
4.2 Multiple and simultaneous mergers in the DTWF model 43
4.3 Number of lineages as a function of time 47
4.4 Expected sample frequency spectrum 47
4.5 A hybrid method for computing the frequency spectrum 49
4.6 Computational details 49
4.6.1 Notation 49
4.6.2 Expected NLFT under the DTWF model 50
4.6.3 Expected number of multiple- and simultaneous-mergers in the DTWF model 50
4.6.4 Expected SFS under the DTWF model 51
4.6.5 Expected NLFT under the coalescent 52
4.7 Dicussion 53

5 Future directions 69

Bibliography 71

A Some identities related to Kingman's coalescent 79
A.1 Analytic expressions for the distribution of the coalescent waiting times . . . 79
A.2 Expectation and standard deviation of the NLFT under the coalescent . . . 83

Acknowledgments

It is hard to overstate the feelings of gratitude and respect I hold for my advisor, Yun S. Song, who always provided me unrestrained freedom and encouragement to explore my interests (both academic, and non-academic) and find my own path. Graduate school has undoubtedly been one of the best experiences in my life so far, and I am certain that the rapport I enjoyed with Yun has been a big reason for this. Yun has been an incredibly inspiring role model and mentor who always amazes me with his ability to see the big picture even while tackling the devil in the technical details. Furthermore, I will consider it a great triumph if even a fraction of his work ethic has rubbed off on me. Many thanks are also due to Yun and his wonderful wife, Young, for introducing me to the delicacies of Korean cuisine, and for so warmly and graciously hosting us graduate students at their house on numerous occasions.

I am also grateful to the other members of my dissertation and qualifying examination committees — Steve Evans, Monty Slatkin, Bernd Sturmfels, and David Tse. I am grateful to Steve for patiently sharing his wealth of knowledge through several excellent courses; to Monty, for always offering an enthusiastic ear for my research ideas, and for providing sage advice over the years; to Bernd, for inspiring my incipient interest in algebraic statistics and geometry; and to David, with whom being a graduate student instructor was a fun experience.

Yun's research group has always provided a very welcoming and collegial environment to talk about research (and other unrelated topics). Ma'ayan Bresler, Andrew Chan, Kelley Harris, Paul Jenkins, Jack Kamm, Wei-Chun Kao, Joyce Liu, Jasmine Nirody, Joshua Paul, Sara Sheehan, Matthias Steinrücken, Jonathan Terhorst, Rachel Wang, and Junming Yin, it has been an honor learning from and working with you, and I really cherish the friendship and camaraderie we've built over the years. I am also very thankful to Andy Clark, Jack Kamm, Uri Keich, Matthias Steinrücken, and Rachel Wang for their collaboration on several research projects.

I would like to specially thank Uri Keich, whose kindness, patience and enthusiasm to work with me while I was still a clueless undergraduate student was extremely valuable for my development and self-confidence. Moreover, I am extremely grateful to Uri for always giving me candid advice, and for being instrumental in my decision to come to Berkeley. I am also very grateful to Jayavel Shanmugasundaram and John Hopcroft for giving me my first tastes of undergraduate research.

I have often heard it said that graduate school can be a lonely experience. However, an incredible group of friends never let me see any truth in that statement. My office mates of several years, Siu Man Chan, Siu On Chan, and Tom Watson have always regaled me with their stories and have been excellent buddies for jaunts around the Bay Area. My friends in the theory group, especially James Cook, Anindya De, and Piyush Srivastava, have always been a source of random interesting conversations. Longtime friends from high school in India and Singapore, and friends from Cornell have always kept me grounded and reminded me that there is a world outside academia. For this I would like to thank Devdeepta Bose,

Raquel Furtado, Nitesh Goel, Caitlin Heyden, Devashish Kumar, Aziz Lalljee, Maria Ligai, Nikhil Mehta, Rishi Ramchand, Chaitanya Reddy, Vikas Reddy, Corey Siegel, Byron Singh, Ted Suzman, Aditya Thatte, Tejas Viswanath, Ti Zhao, and several others. To my wonderful former housemates in Berkeley — Andrew Chan, Nicholas Kong, and Andrew Krioukov — thank you for being so easy-going and understanding even when I wasn't always doing my share of the chores.

Some of my most enjoyable experiences at Berkeley were the fantastic Russian language courses with Zachary Johnson, Mark Kaiser, and Lisa Little. Their passion and enthusiasm for teaching have been truly inspirational. Захар, Марк и Лиза, спасибо вам за всё и извините, что часто опаздывал на занятия. I am also very grateful to my Russian language classmate of several years, Jin Jin, who became one of my closest friends and introduced me to the amazing Berkeley Art Studio and Cal Performances.

I am grateful to the generous funding support provided by the Berkeley Fellowship during my initial years of graduate school. I am also thankful to La Shana Porlaris, Xuan Quach, and Shirley Salanio for helping me with various administrative matters, and to Brian McClendon for organizing many fantastic computational biology retreats.

Finally, and most importantly, I cannot thank my parents enough for giving me complete freedom to make my own decisions from a very early age. Their unconditional love and support mean the world to me and I am grateful to them each and every day for raising me the way they have. Amma and appa, this dissertation is dedicated to you.

Chapter 1

Introduction

1.1 Motivation

Human genetics has entered a new era with study sample sizes on the order of thousands to tens of thousands of individuals sequenced at hundreds of genomic regions [1, 12, 19, 62, 85]. A consistent finding arising from several large-sample studies is that human genomes harbor a substantial excess of rare variants compared to that predicted using previously applied demographic models. For example, in a large-scale exome-sequencing study of 14,000 individuals at over 200 genes, Nelson *et al.* [62] found that over 70% of the single nucleotide variants only appear in one or two haplotypes in the sample, which corresponds to a minor allele frequency on the order of 0.01%. There are several factors that may contribute to the discrepancy between observations in the data and theoretical predictions, including the following possible explanations:

1. Previously applied demographic models are wrong. In particular, the observed polymorphism patterns are indicative of a recent rapid growth in the effective population size, much more rapid than in previously applied demographic models. This conclusion would be consistent with historical records of census population size [39].

2. Population substructure [71,81] and natural selection have distorted the observed polymorphism patterns while previous demographic inference studies have failed to adequately account for these factors.

3. Theoretical predictions for a given demographic model are inaccurate when the sample size is very large compared to the effective population size. Kingman's coalescent process [44–46], which arises as a continuum limit of a large class of discrete-time random mating models [58, 59], provides an accurate approximation to these discrete-time models only if the sample size is sufficiently small compared to the effective population size. Violation of this assumption may distort the genealogical properties of a large sample in a way that may inflate rare variants relative to the predictions of coalescent theory.

The aim of this dissertation is to investigate the first and third possibilities in detail through the problem of demographic inference.

1.2 Demographic inference

1.2.1 Importance of understanding demography

Given a sample of genomic sequences from a large population, an important inference problem with a wide variety of applications is to determine the underlying demography of the population. Understanding the population demography is important for many reasons:

- Genome-wide association studies for inferring the genetic basis of complex heritable diseases can produce false positive associations if the demography of the population is not taken into account properly [7, 10, 53, 65, 70, 86].
- The population demography is needed for calibrating correct null models of neutral genome evolution so that one can conduct principled statistical tests to find regions of the genome that are subject to evolutionary pressure from natural selection [6, 49, 63, 73, 93].
- Genomic data can serve as an independent source of evidence for important anthropological events like population migrations, splits, admixture and introgression events [23, 41, 48, 51, 75, 80, 87] that have occurred in the history of our species.
- Demographic events such as bottlenecks and small founder populations can lead to a higher frequency of deleterious mutations [49]. Conversely, increased genetic drift due to bottlenecks might also serve as an explanation for speciation [2].
- Population demography is also important in forensic science applications, where a low probability of a random genetic profile from the population matching the crime scene sample is used as evidence for further investigation. This random match probability strongly depends on the substructure and the amount of relatedness between individuals of the population [72].

1.2.2 Representations of genomic data

Demographic inference from full sequence data is a very challenging problem, and most existing methods for inferring demography can be broadly categorized into one of three types depending on the representation of the sequence data that they operate on.

Full sequence data

The evolution of full genomic sequences with intragenic recombination can be modeled using rich generative processes like the coalescent with recombination [29, 37]. However, because

recombination induces correlations between distant genetic loci [94], it is computationally cumbersome to even simulate long sequences at many samples, let alone perform any kind of inference from data. Several approximations to the coalescent with recombination have been proposed which avoid these long-range correlations by assuming that the ancestral trees at each genomic locus form a Markov chain of small order [54, 57], thus enabling much more efficient sequence simulation [8,17]. However, it is still quite challenging to perform inference under these Markov approximations to the coalescent with recombination. Methods for inferring piecewise-constant historical population sizes have been developed that can utilize anywhere from a single diploid sequence (or a pair of haploid sequences) [48] to tens of sequences [79]. These methods are able to quite accurately recover population size changes in the ancient past, but have limited ability to infer very recent population expansion events because of the small sample size and hence limited number of coalescence events that they can access.

Tracts of sequence similarity

A lower-dimensional representation of full sequence data are the demography dependent distribution of contiguous tract lengths where a pair of haplotypes are identical by descent (IBD) or identical by state (IBS). Inference methods that use these data [34, 64] work by calculating the expected distribution of IBD or IBS tract lengths for a pair of randomly sampled haplotypes under a given demographic model. Using the observed distribution of pairwise IBS and IBD tract lengths in a sample of sequences, they compute a composite likelihood function of the demographic parameters which can then be optimized. The advantage of these methods is that they can capture linkage information between sites to some extent while still summarizing a set of sequences of length L using L numbers. Another advantage of working with tracts of sequence similarity is that one can easily visualize the fit of a demographic model by plotting the expected and observed distributions of tract lengths. One disadvantage of using IBD tracts is that it is not obvious from the data as to which segments of a pair of haplotypes are shared identically by descent. Hence, one typically has to preprocess the data using a method that can infer tracts shared by descent [30], and uncertainties in this inference can also impact the downstream demographic inference analyses.

Sample frequency spectrum

The sample frequency spectrum (SFS) is a succinct representation of a sample of homologous sequences which tracks the distribution of alleles frequencies without regard to where the segregating sites are located along the sequence. For a randomly mating population with low enough mutation rate so that each segregating site has exactly two distinct alleles, the SFS of a sample of size n is a vector of length $n-1$ that counts the proportion of segregating sites as a function of the frequency ($\frac{b}{n}$, where $1 \leq b \leq n-1$) of the mutant allele in the sample. The SFS is useful for several reasons. First, the SFS is a succinct summary of a large sample of genomic sequences, where the information in n sequences

of arbitrary length can be summarized by just $n-1$ numbers. This makes the SFS both mathematically and algorithmically tractable. In particular, since the SFS ignores linkage information between sites, one can avoid challenging mathematical and computational issues associated with rigorously modeling genetic recombination. Furthermore, the statistical properties of the SFS and their dependence on the population demographic history are well understood under the coalescent and the diffusion models of neutral evolution [9, 20, 26, 28, 42, 68, 95]. This dependence of the SFS on demography, along with the assumption of free recombination between sites, has been exploited in several methods for inferring historical population demography [16, 31, 52, 55]. Second, the SFS can effectively capture the impact of recent demography on genetic variation. Because the leading entries of the SFS count the rare variants in the sample, one might be able to use the large number of rare variants in recent large-sample studies [1, 12, 19, 62, 85] to infer demographic events in the recent past at a much finer resolution than possible using smaller samples. Third, similar to the distribution of IBD and IBS tract lengths, the SFS also provides a simple way of visualizing the goodness of fit of a demographic model to data, since one can easily compare the entries of the observed and fitted SFS. Because of these advantages of the SFS, we will work with this representation of the sequence data in the rest of this dissertation.

1.3 Preliminaries

Before proceeding any further, we first give a brief overview of the discrete-time Wright-Fisher model and its continuum limit, Kingman's coalescent. We also briefly describe how the coalescent can be extended to incorporate population demography in the form of a time-varying population size function.

1.3.1 Wright-Fisher model

The Wright-Fisher model is a widely used discrete-time random mating model for describing the evolution of a finite population. In its simplest form, each generation of the population is composed of N individuals, where an individual is a genomic region that is inherited as a unit (henceforth called a *locus*). A locus could be a single nucleotide, a copy of a gene, etc. In each generation, an offspring population of size N is produced, where each offspring picks a parent from the previous generation uniformly at random. The ancestry of a sample of sequences drawn from a given generation can be represented by a tree, where each internal node corresponds to the most recent common ancestor of its descendant leaves.

The population size N used in the Wright-Fisher model is called the *effective population size* to distinguish it from the more physical notion of the census population size. In real biological populations, the Wright-Fisher model assumptions of random mating and non-overlapping generations are obviously violated because of geographical and other demographic constraints. The notion of effective population size captures the degrees of freedom in a random mating model which generates the same values for the population genetic quan-

tities of interest as the actual biological population under study. In this work, the population genetic quantity we are interested in is the site frequency spectrum of a finite sample. Keeping the above points in mind, we use the terms population size and effective population size interchangeably in the rest of this dissertation, reserving the use of the term census population size to refer to the physical concept. A related discussion of the relation between effective and census population sizes is given in Section 4.7 of Chapter 4.

In the Wright-Fisher model, the probability that two randomly sampled individuals from a given generation have a common parent in the previous generation is $1/N$ by construction. If we consider a random sample of n individuals from a particular generation, where $n \ll N$, the probability that some two individuals in the sample will have a common parent in the previous generation is $\binom{n}{2}/N + O\left(1/N^2\right)$. The probability that no two individuals in the sample will have a common ancestor in k previous generations is then given by

$$\left(1 - \frac{\binom{n}{2}}{N} + O\left(\frac{1}{N^2}\right)\right)^k = \exp\left(-\binom{n}{2}\frac{k}{N}\right) + O\left(\frac{1}{N^2}\right)$$

If we set $t = k/N$ in the above expression, let N tend to infinity, and measure time in units of N generations, then the waiting time for a sample of size n to have $n-1$ ancestors is distributed as an exponential random variable with rate $\binom{n}{2}$. This scaling limit is called Kingman's coalescent.

1.3.2 Kingman's coalescent

Kingman's coalescent [44–46] — henceforth simply referred to as the coalescent — on a sample of size n is a partition-valued Markov process $(\Pi_n(t), t \geq 0)$ on $\mathcal{P}[n]$, the set of partitions of $[n]$. It is defined as:

- $\Pi_n(0) = \{\{1\}, \{2\}, \ldots, \{n\}\}$
- At any time $t > 0$, from any $\Pi_n(t) = \{A_1, \ldots, A_k\} \in \mathcal{P}[n]$, the only possible transitions are to Π_{ij} at rate 1, where $1 \leq i < j \leq k$ and $\Pi_{ij} = \Pi_n(t) - \{A_i, A_j\} \cup \{A_i \cup A_j\}$, i.e. Π_{ij} is $\Pi_n(t)$ with blocks A_i and A_j merged.

Each realization of the above partition-valued process has a natural genealogical tree associated with it. This tree can be sampled along with the partitions as follows.

- Let $\Pi_{n,n} = \{\{1\}, \{2\}, \ldots, \{n\}\}$, $\Sigma_{n+1} = 0$ and $k = n$
- Repeat until $k = 1$
  - Sample $T_k \sim \text{Exp}\left(\binom{k}{2}\right)$ and let $\Sigma_k = \Sigma_{k+1} + T_k$
  - Merge two blocks of $\Pi_{n,k}$ uniformly at random to obtain $\Pi_{n,k-1}$
  - Set k to $k-1$

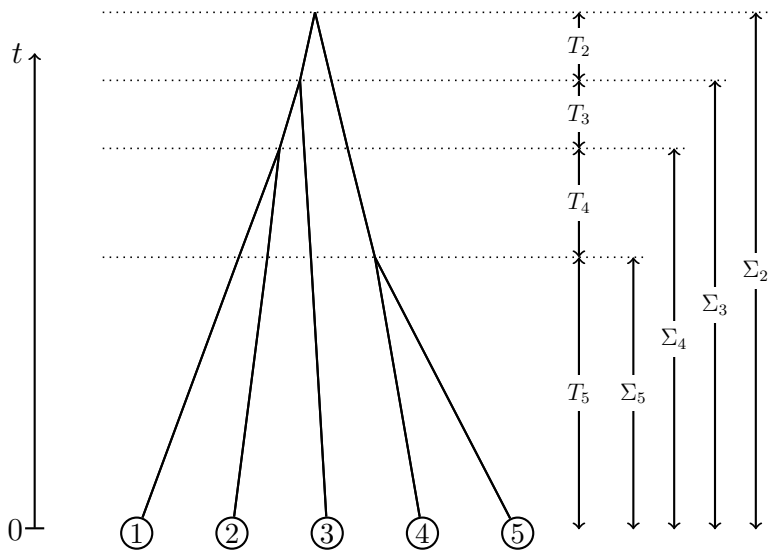


Figure 1.1: **A realization of a genealogical tree for a sample of size 5 from the coalescent.** The waiting time while there are exactly k ancestral lineages is denoted by T_k, and the total time spent in the process while there are at least k ancestral lineages is given by Σ_k.

Figure 1.1 shows a realization of a genealogical tree on 5 individuals sampled according to the above algorithm along with the waiting times T_k and cumulative waiting times Σ_k.

Since its introduction, the coalescent has found wide use in modern population genetics analyses, and has been extended to incorporate time-varying population sizes [27], recombination [25], natural selection [47], and population subdivision [35]. Although we motivated the coalescent through a scaling limit of the discrete-time Wright-Fisher model, it is the continuum limit of a large class of discrete-time models of random mating [46, 58, 59], and hence robust to the minutiae of these random mating schemes. Furthermore, generalizations of Kingman's coalescent have been developed which allow multiple lineages to merge in a single coalescence event [67, 74] and even multiple sets of lineages to merge simultaneously [58, 78].

1.3.3 Extension to variable population size

We can extend the previous description of the coalescent to a deterministically varying population size as follows. Consider a discrete-time Wright-Fisher model where the population size at generation t is given by $N(t)$. In generation $t + 1$, an offspring population of $N(t + 1)$ individuals is produced, where each individual chooses a parent uniformly at random. Let N_r be some reference population size and define the relative population size

$\eta_{N_r}(t)$ in population-scaled units of time as

$$\eta_{N_r}(t) = \frac{N(\lceil N_r t \rceil)}{N_r}. \tag{1.1}$$

If N_r and the population size function $N(t)$ tend to infinity while the following hold,

- $\lim_{N_r \to \infty} \eta_{N_r}(t) = \eta(t)$ exists,
- $\eta(t)$ is positive and bounded for all $t \geq 0$,

then it can be shown that when time is measured in units of N_r generations, the genealogy of a random sample of size n drawn at time $t = 0$ from the discrete-time Wright-Fisher model with time varying population size $N(t)$ approaches that given by the following continuous-time coalescent process:

- Let $\Pi_{n,n} = \{\{1\}, \{2\}, \ldots, \{n\}\}$, $\Sigma_{n,n+1} = 0$ and $k = n$
- Repeat until $k = 1$
  - Sample $T_{n,k}$ from an exponential distribution with time variable rate $\binom{k}{2}/\eta(t)$ started at time $\Sigma_{n,k+1}$
  - Let $\Sigma_{n,k} = \Sigma_{n,k+1} + T_{n,k}$
  - Merge two blocks of $\Pi_{n,k}$ uniformly at random to obtain $\Pi_{n,k-1}$
  - Set k to $k - 1$

We label the times $T_{n,k}$ and $\Sigma_{n,k}$ by two indices to indicate the fact that the distribution of $T_{n,k}$ and $\Sigma_{n,k}$ depend on both n and k. In the constant population size case described in Section 1.3.2, these distributions depend only on k and not on n. Moreover, in the constant population size case, the sequence of random variables $T_{n,k}$ are independent. For general variable population size functions, this is no longer true. However, for $k < j$, $T_{n,k}$ is conditionally independent of $T_{n,j}$ given $\Sigma_{n,k+1}$, and this fact will be important later for obtaining analytic expressions for the probability density function of $T_{n,k}$ and $\Sigma_{n,k}$.

1.3.4 Sample frequency spectrum

We can also introduce mutations in the discrete-time Wright-Fisher model described in Section 1.3.1, where each reproduction event is hit by a mutation with probability μ. If we let μ tend to 0 while the population size N tends to infinity such that the population-scaled mutation rate $2N\mu$ approaches some constant θ, then the mutation process on the genealogy of a sample of size n approaches that given by a Poisson process with rate $\theta/2$ on the genealogy generated by Kingman's coalescent on n leaves. The analogous statement also applies to the variable population size setting, with the constant population size N replaced by some reference population size N_r with respect to which the function $N(t)$ stays bounded for all $t \geq 0$ in the scaling limit.

In the rest of this dissertation, we will assume that the mutation rate is low enough that each mutation affects a new site at the locus under consideration. This is called the infinite-sites model of mutation [14]. For a random sample of size n drawn from the population at time 0, let $\boldsymbol{\gamma}_n = (\gamma_{n,1}, \ldots, \gamma_{n,n-1})$ be the vector where $\gamma_{n,b}$ is the expected number of sites having b copies of the mutant allele and $n - b$ copies of the ancestral allele in a sample of size n. Furthermore, let $\boldsymbol{\xi}_n$ be $\boldsymbol{\gamma}_n$ normalized to be a probability distribution. The vector $\boldsymbol{\xi}_n$ is known as the expected sample frequency spectrum (SFS) of a sample of size n. The expected SFS $\boldsymbol{\xi}_n$ depends on the population size function $\eta(t)$ and can be computed using the expectation of the waiting times $T_{n,k}$ and $\Sigma_{n,k}$. We describe these details in Section A.1 of Appendix A.

1.4 Overview

In this dissertation, we study the problem of demographic inference for a large randomly mating population, where we will be interested in recovering the effective population size as a function of time from the SFS representation of a sample of sequences generated according to the coalescent with a variable population size function and infinite-sites model of mutation as described in Section 1.3. The next three chapters in the dissertation deal with the following three aspects of demographic inference, and the last chapter concludes with a discussion of problems for future research.

1.4.1 Statistical

The most natural statistical or information-theoretic question to ask is whether the expected SFS of a random sample of sequences uniquely determines the historical population size function. If the true underlying population size function is allowed to be an arbitrary continuous function, then Myers *et al.* [61] have shown that it is information-theoretically not possible to recover the population size function from the SFS, no matter how large the sample size. In particular, they showed that even if one has perfect information about the SFS, i.e. even if the underlying population-wide probability density function of allele frequencies can be known exactly, there might be several population size functions that generate this same distribution. We can then sharpen the question and ask about what reasonable biologically motivated assumptions might we impose on the true population size function so that it is uniquely recoverable from the SFS of a finite sample? In Chapter 2, we use tools from real analysis to show that the space of piecewise-defined population size functions where each piece does not oscillate too much (in some technical sense) is identifiable from the SFS of finite samples. The contents of Chapter 2 are largely based on a preprint of Bhaskar and Song [5].

1.4.2 Algorithmic

If we are able to restore statistical identifiability by restricting the space of underlying population size functions, then the next question is an algorithmic one. Given SFS data from very large samples, how might we efficiently infer the population size function in the class of functions that are biologically realistic and identifiable from the data? In Chapter 3, we develop an efficient, exact and numerically stable likelihood-based algorithm for inferring piecewise-exponential population size functions from the SFS of very large samples of tens to hundreds of thousands of sequences. As a by-product of this inference procedure, we will also be able to estimate the mutation rate at each locus. Our algorithm is based on analytic, numerically stable expressions for several genealogical quantities that are intimately tied to the SFS [69]. Because our method will be based on an exactly computable likelihood function rather than Monte-Carlo simulations, we can also compute gradients with respect to the demographic parameters using automatic differentiation [24], thus enabling the use of gradient-based optimization methods that work well in practice.

1.4.3 Robustness

With modern datasets having sample sizes on the order of tens of thousands of individuals, a number commonly cited as the effective population size of humans [15, 32, 33, 82], we are rapidly approaching samples that form a non-negligible fraction of the entire population. Kingman's coalescent [44–46], henceforth simply referred to as the coalescent, has been a central model in modern population genetics for studying the ancestry of a sample of individuals taken from a large randomly mating population. The coalescent is a continuous-time Markov process that can be constructed as a scaling limit of a discrete-time Wright-Fisher (DTWF) model of random mating by taking the population size to infinity while suitably rescaling the unit of time. The coalescent is an excellent approximation to the original discrete-time model if, for all times, the population size is sufficiently large relative to the number of ancestral lineages of the sample. However, with modern datasets having tens of thousands of individuals, the difference in genealogical quantities between the coalescent and these discrete-time models of mating is not clear. In Chapter 4, we develop exact computation algorithms for the DTWF model to study the deviations in several genealogical quantities, such as the number of surviving lineages as a function of time, the number of multiple and simultaneous mergers, and the entries of the expected SFS. We find that for several recently inferred demographic models and large sample sizes, there can be substantial genealogical deviations introduced by the coalescent approximation that might influence the results of demographic inference analyses. The contents of Chapter 4 are based on a preprint of Bhaskar *et al.* [3].

Chapter 2

Identifiability of the population size function from SFS data

2.1 Overview

In Chapter 1, we saw that the SFS is a particularly succinct representation of a large sample of genome sequences. While the SFS has algorithmic advantages for demographic inference, it is believed to suffer from a statistical shortcoming. Specifically, Myers *et al.* [61] showed that even with perfect knowledge of the *population* frequency spectrum (i.e., the proportion of polymorphic sites with population-wide allele frequency in $(x, x + dx)$ for all $x \in (0, 1)$), the historical population size function $\eta(t)$ as a function of time is not identifiable. Using Müntz-Szász theory, they showed that for any population size function $\eta(t)$, one can construct arbitrarily many smooth functions $F(t)$ such that both $\eta(t)$ and $\eta(t) + \alpha F(t)$ generate the same population frequency spectrum for suitably chosen values of α. They also constructed explicit examples of such functions $\eta(t)$ and $F(t)$. While this non-identifiability could pose serious challenges to demographic inference from frequency spectrum data, the population size functions involved in their example are arguably unrealistic for biological populations. In particular, their explicit example involves a population size function which oscillates at an increasingly higher frequency as the time parameter approaches the present. Real biological population sizes can be expected to vary over time in a mathematically more well-behaved fashion. In particular, populations can be expected to evolve in discrete units of time, which, when approximated by a continuous-time model, restricts the frequency of oscillations in the population size function to be less than the number of generations of reproduction per unit time. Furthermore, since a population size model being inferred must have a finite representation for obvious algorithmic reasons, most previous demographic inference analyses have focused on inferring population size models that are piecewise-defined over a restricted class of functions, such as piecewise-constant and piecewise-exponential models [23, 40, 48, 51, 62, 77, 85].

In this chapter, we revisit the question of demographic model identifiability under the

assumption that the population size is a *piecewise-defined* function of time where each piece comes from a family of biologically-motivated functions, such as the family of constant or exponential functions. We also re-examine the assumption that one has access to the population-wide patterns of polymorphism. In real applications, we do not expect to know the allele frequency spectrum for an entire population but rather only the SFS for a randomly drawn finite sample of individuals. Here, we investigate whether one can learn piecewise population size functions given perfect knowledge of the expected SFS for a sufficiently large sample of size n. Unlike in the case of arbitrary continuous population size functions considered by Myers *et al.* [61], the answer to this question is affirmative. More precisely, we obtain bounds on the sample size n that are sufficient to distinguish population size functions among piecewise demographic models with K pieces, where each piece comes from some family of functions (see Theorems 5 and 9). Our bound on the sample size can be expressed as an affine function of the number K of pieces, where the slope of the function is a measure of the complexity of the family to which each piece belongs. In the cases of piecewise-constant and piecewise-exponential models, which are often assumed in population genetic analyses, the slope of this affine function can be calculated explicitly, as shown in Corollaries 6 and 7. We also obtain analogous results for the "folded" SFS (see Theorem 10), a variant of the SFS which circumvents the ambiguity in the identity of the ancestral allele type by grouping the polymorphic sites in a sample according to the sample minor allele frequency.

There are two main technical elements underlying our proofs of the identifiability results mentioned above. The first step is to show that the expected SFS of a sample of size n is in bijection with the Laplace transform of a time-rescaled version of the population size function evaluated at a particular sequence of $n-1$ points. This reduces the problem of identifiability from the SFS to that of identifiability from the values of the Laplace transform at a fixed set of points. The second step relies on a generalization of Descartes' rule of signs for polynomials to the Laplace transform of general piecewise-continuous functions. This technique yields an upper bound on the number of roots of the Laplace transform of a function by the number of sign changes of the function. We think that this proof technique based on sign changes might be of independent interest for proving statistical identifiability results in other settings. We also provide an alternate proof of identifiability for piecewise-constant population models, where the aforementioned second step is replaced by a linear algebraic argument that has a constructive flavor. We include this alternate proof in the hope that it could be used to develop an algebraic inference algorithm for piecewise-constant models.

The remainder of this chapter is organized as follows. In Section 2.2, we introduce the model and notation. We describe and prove our main results in Section 2.3. We also discuss the counterexample of Myers *et al.* in light of our findings in Section 2.4.

2.2 Notation

We consider a population evolving according to Kingman's coalescent [44–46] with the infinite-sites model of mutation [43] and selective neutrality as described in Section 1.3 of Chapter 1. The population size is assumed to change deterministically with time and is described by a function $\eta : \mathbb{R}_{\geq 0} \to \mathbb{R}_+$, such that the instantaneous coalescence rate between any pair of lineages at time t is $1/\eta(t)$.

Let $T^{(\eta)}_{n,k}$ denote the time (in coalescent units) while there are k ancestral lineages for a sample of size n obtained at time 0. Defining $R_\eta(t)$ as

$$R_\eta(t) := \int_0^t \frac{1}{\eta(x)}\, dx, \tag{2.1}$$

the expected time $\mathbb{E}[T^{(\eta)}_{m,m}]$ to the first coalescence event for a sample of size m is given by

$$\mathbb{E}[T^{(\eta)}_{m,m}] = \int_0^\infty t\, \frac{\binom{m}{2}}{\eta(t)} \exp\left[-\binom{m}{2} R_\eta(t)\right] dt. \tag{2.2}$$

Following the notation of Myers *et al.* [61], define a time-rescaled version $\widetilde{\eta}$ of the population size function η as

$$\widetilde{\eta}(\tau) = \eta(R^{-1}_\eta(\tau)), \tag{2.3}$$

where $\tau \in \mathbb{R}_{\geq 0}$. The function $\widetilde{\eta}(\tau)$ reparameterizes the population size as a function of the cumulative rate of coalescence $\tau = R_\eta(t)$. For a given population size function $\widetilde{\eta}$ parameterized by the total coalescence rate τ, there corresponds a unique population size function η parameterized by time t. Specifically, $\eta(t) = \widetilde{\eta}(S^{-1}_{\widetilde{\eta}}(t))$, for all $t \in \mathbb{R}_{\geq 0}$, where $S_{\widetilde{\eta}}(t)$ is an invertible function given by

$$S_{\widetilde{\eta}}(t) = \int_0^t \widetilde{\eta}(x)\, dx. \tag{2.4}$$

Applying integration by parts to (2.2) and using the condition that $\mathbb{E}[T^{(\eta)}_{m,m}] < \infty$, we have

$$\mathbb{E}[T^{(\eta)}_{m,m}] = \int_0^\infty \exp\left[-\binom{m}{2} R_\eta(t)\right] dt. \tag{2.5}$$

Furthermore, since R_η is monotonically increasing and continuous from $\mathbb{R}_{\geq 0}$ to $\mathbb{R}_{\geq 0}$, it is a bijection over $\mathbb{R}_{\geq 0}$. For notational convenience, for any interval $I \subseteq \mathbb{R}_{\geq 0}$, we define $R_\eta(I)$ to be the interval

$$R_\eta(I) = \{R_\eta(x) \mid x \in I\}. \tag{2.6}$$

By making the substitution $\tau = R_\eta(t)$ in (2.5) and using (2.3), we have the following expression for $\mathbb{E}[T^{(\eta)}_{m,m}]$:

$$\mathbb{E}[T^{(\eta)}_{m,m}] = \int_0^\infty \widetilde{\eta}(\tau) \exp\left[-\binom{m}{2}\tau\right] d\tau. \tag{2.7}$$

Equation (2.7) states that the time to the first coalescence event for a sample of size m is given by the Laplace transform of the time-rescaled population size function $\widetilde{\eta}$ evaluated at the point $\binom{m}{2}$. For a sample of size n, let $\xi_{n,b}$ denote the probability that a polymorphic site has b mutant alleles and $n-b$ ancestral alleles. We refer to $(\xi_{n,1}, \ldots, \xi_{n,n-1})$ as the expected sample frequency spectrum (SFS).

2.3 Main results

2.3.1 Determining expected times to the first coalescence event from the SFS

The following lemma shows that the expected SFS for a sample of size n tightly constrains the expected time to the first coalescence event for all sample sizes $2, \ldots, n$:

Lemma 1. *Under an arbitrary variable population size model $\{\eta(t), t \geq 0\}$, suppose $\xi_{n,1}$, $\ldots, \xi_{n,n-1}$ are known and define $c_m := \mathbb{E}[T^{(\eta)}_{m,m}]$ for $2 \leq m \leq n$. Then, up to a common positive multiplicative constant, the quantities $c_2, \ldots, c_n$ can be determined uniquely from $\xi_{n,1}, \ldots, \xi_{n,n-1}$.*

Proof. In the coalescent for a sample of size n, let $\gamma_{n,b}$ denote the total expected branch length subtending b leaves, for $1 \leq b \leq n-1$. Then, $\xi_{n,b} = \gamma_{n,b} / \sum_{k=1}^{n-1} \gamma_{n,k}$, which implies that there exists a positive constant κ such that $\gamma_{n,b} = \kappa \xi_{n,b}$ for all $1 \leq b \leq n-1$. We now prove that $c_2, \ldots, c_n$ can be determined uniquely from $\gamma_{n,1}, \ldots, \gamma_{n,n-1}$.

Let $\phi_{n,k} = \mathbb{E}[T^{(\eta)}_{n,k}]$. Then, by a result of Griffiths and Tavaré [28],

$$\gamma_{n,b} = \sum_{k=2}^{n-b+1} k \frac{\binom{n-b-1}{k-2}}{\binom{n-1}{k-1}} \phi_{n,k}, \tag{2.8}$$

for $1 \leq b \leq n-1$. The system of equations (2.8) can be rewritten succinctly as a linear system

$$\boldsymbol{\gamma} = \boldsymbol{M}\boldsymbol{\phi}, \tag{2.9}$$

where $\boldsymbol{\gamma} = (\gamma_{n,1}, \ldots, \gamma_{n,n-1})$, $\boldsymbol{\phi} = (\phi_{n,2}, \ldots, \phi_{n,n})$, and $\boldsymbol{M} = (m_{bk})$ with $m_{bk} = k\binom{n-b-1}{k-2} / \binom{n-1}{k-1}$, for $1 \leq b \leq n-1$ and $2 \leq k \leq n$. The matrix $\boldsymbol{M}$ is upper-left triangular since $\binom{n-b-1}{k-2} = 0$

if $k > n - b + 1$, and the anti-diagonal entries are $\frac{k}{\binom{n-1}{k-1}} > 0$. Hence, $\det(\boldsymbol{M}) \neq 0$ and $\boldsymbol{M}$ is therefore invertible. Thus, given $\boldsymbol{\gamma}$, we can determine $\boldsymbol{\phi}$ uniquely as $\boldsymbol{M}^{-1}\boldsymbol{\gamma}$.

Let $\psi_{n,k} = \sum_{j=k}^{n} \mathbb{E}[T_{n,j}^{(\eta)}]$. Then, defining $\psi_{n,n+1} := 0$, observe that $\psi_{n,k} = \phi_{n,k} + \psi_{n,k+1}$ for $2 \leq k \leq n$. This implies that $\psi_{n,2}, \ldots, \psi_{n,n}$ can be determined uniquely from $\phi_{n,2}, \ldots, \phi_{n,n}$. Polanski *et al.* [68] showed that $\psi_{n,k}$ can be written as

$$\psi_{n,k} = \sum_{m=k}^{n} a_{n,k,m} c_m, \tag{2.10}$$

where $a_{n,k,m}$, for $k \leq m \leq n$, are given by

$$a_{n,k,m} = \frac{\prod_{l=k, l\neq m}^{n} \binom{l}{2}}{\prod_{l=k, l\neq m}^{n} \left[\binom{l}{2} - \binom{m}{2}\right]}, \tag{2.11}$$

and $c_m = \mathbb{E}[T_{m,m}^{(\eta)}]$, shown in (2.5). An alternate derivation of (2.10) and (2.11) can also be found in Section A.1 of Appendix A. Again, the system of equations (2.10) can be written as a triangular linear system

$$\boldsymbol{\psi} = \boldsymbol{A}\boldsymbol{c}, \tag{2.12}$$

where $\boldsymbol{\psi} = (\psi_{n,2}, \ldots, \psi_{n,n})$, $\boldsymbol{c} = (c_2, \ldots, c_n)$, and $\boldsymbol{A} = (a_{n,k,m})$, for $2 \leq k, m \leq n$. Note that $\boldsymbol{A}$ is an upper triangular matrix since $a_{n,k,m} := 0$ if $m < k$. Since $\boldsymbol{A}$ has non-zero entries on its diagonal, $\boldsymbol{A}^{-1}$ exists, and $\boldsymbol{c}$ can be determined uniquely as $\boldsymbol{A}^{-1}\boldsymbol{\psi}$. □

This lemma implies that the problem of identifying the population size function $\eta(t)$ from $\xi_{n,1}, \ldots, \xi_{n,n-1}$ can be reduced, up to a multiplicative constant, to the problem of identifying $\eta(t)$ from $c_2, \ldots, c_n$.

2.3.2 Piecewise population size models and sign change complexities

To state our main result in full generality, we first need a few definitions:

Definition 1 ($\mathcal{F}$, family of continuous population size functions)**.** *A family $\mathcal{F}$ of continuous population size functions is a set of positive continuous functions $f : \mathbb{R}_{\geq 0} \to \mathbb{R}_+$ of a particular type parameterized by a collection of variables.*

We use $\mathcal{F}_c$ to denote the family of constant population size functions; i.e., functions of the form $f(t) = \nu$ for all t, where $\nu \in \mathbb{R}_+$ is the only parameter of the family. Further, we use $\mathcal{F}_e$ to denote the family of exponential population size functions of the form $f(t) = \nu \exp(\beta t)$, where $\nu \in \mathbb{R}_+$ and $\beta \in \mathbb{R}$ are the parameters of the family.

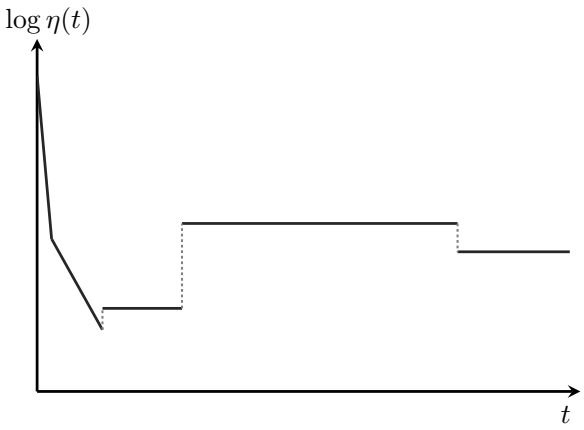


Figure 2.1: **A piecewise-exponential population size function $\eta \in \mathcal{M}_K(\mathcal{F}_e)$, where $K \geq 5$.** Note that the y-axis is in a log scale. This piecewise-exponential function depicts the historical population size changes of a European population that was estimated from the SFS of a sample of 1,351 (diploid) individuals of European ancestry [85].

Definition 2 ($\mathcal{M}_K(\mathcal{F})$, piecewise models over $\mathcal{F}$ with at most K pieces)**.** *Given a family $\mathcal{F}$ of continuous population size functions, a population size function $\eta(t)$ defined over $\mathbb{R}_{\geq 0}$ is said to be piecewise over $\mathcal{F}$ with at most K pieces if there exists an integer p, where $1 \leq p \leq K-1$, and a sequence of p time points $0 < t_1 < \cdots < t_p < \infty$ such that for each $1 \leq i \leq p+1$, there exists a positive continuous function $f_i \in \mathcal{F}$ such that $\eta(t) = f_i(t - t_{i-1})$ for all $t \in [t_{i-1}, t_i)$. For convention, we define $t_0 = 0$ and $t_{p+1} = \infty$. Note that η may not be continuous at the change points $t_1, \ldots, t_p$. We use $\mathcal{M}_K(\mathcal{F})$ to denote the space of such piecewise population size models with at most K pieces, each of which belongs to function family $\mathcal{F}$. Illustrated in Figure 2.1 is an example piecewise-exponential population size function $\eta \in \mathcal{M}_K(\mathcal{F})$ where $K \geq 5$ and $\mathcal{F} = \mathcal{F}_e$.*

Definition 3 ($\sigma(f)$, number of sign changes of a function)**.** *For a function g (not necessarily continuous) defined over some interval (a, b), we say that $t \in (a, b)$ is a sign change point of g if there exist some $\epsilon > 0$, $t' \geq t$, and an interval $(t', t' + \epsilon) \subseteq (a, b)$ such that*

1. *$(t - \epsilon, t) \subseteq (a, b)$,*
2. *$g(z) = 0$ for $z \in (t, t')$,*
3. *$g(x)g(y) < 0$ for all $x \in (t - \epsilon, t)$ and $y \in (t', t' + \epsilon)$.*

We define the number $\sigma(g)$ of sign changes of g as the number of such sign change points in its domain (a, b). See Figure 2.2 for an illustration.

Note that the above definition of the number of sign changes counts the number of times the function g changes value from positive to negative (and vice versa) while ignoring

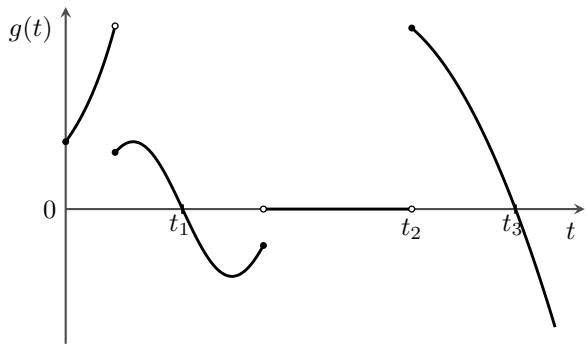


Figure 2.2: **Illustration of the sign changes of a function.** For the domain shown, $\sigma(g) = 3$ and the sign change points of g are denoted t_1, t_2, and t_3.

intervals where it is identically zero. While the above definition is not restricted to piecewise continuous functions, we will restrict our attention to such functions for the remainder of this chapter.

Definition 4 ($\mathscr{S}(\mathcal{F})$ and $\mathscr{S}(\mathcal{M}_K(\mathcal{F}))$, sign change complexities)**.** *For a family $\mathcal{F}$ of continuous population size functions, we define the sign change complexity $\mathscr{S}(\mathcal{F})$ as*

$$\mathscr{S}(\mathcal{F}) = \sup_{\substack{f_1,f_2\in\mathcal{F},\\ a_1,a_2\in\mathbb{R}_{\geq 0}}} \left\{ \sigma(g) \;\middle|\; \begin{array}{l} g(\tau) := \widetilde{f}_1(\tau - a_1) - \widetilde{f}_2(\tau - a_2) \text{ with domain} \\ \mathrm{Dom}(g) = \left\{ \tau \in \mathbb{R}_{\geq 0} \;\middle|\; \begin{array}{l} \tau - a_1 \in \mathrm{Dom}(\widetilde{f}_1), \\ \tau - a_2 \in \mathrm{Dom}(\widetilde{f}_2) \end{array} \right\} \end{array} \right\} \tag{2.13}$$

$$= \sup_{\substack{f_1,f_2\in\mathcal{F},\\ a\in\mathbb{R}_{\geq 0}}} \left\{ \sigma(g) \;\middle|\; \begin{array}{l} g(\tau) := \widetilde{f}_1(\tau) - \widetilde{f}_2(\tau - a) \text{ with domain} \\ \mathrm{Dom}(g) = \left\{ \tau \in \mathbb{R}_{\geq 0} \;\middle|\; \begin{array}{l} \tau \in \mathrm{Dom}(\widetilde{f}_1), \\ \tau - a \in \mathrm{Dom}(\widetilde{f}_2) \end{array} \right\} \end{array} \right\}, \tag{2.14}$$

where $\widetilde{f}_j$ are the time-rescaled versions of f_j as defined in (2.3)*, and $\mathrm{Dom}(\widetilde{f}_j) = R_{f_j}(\mathbb{R}_{\geq 0})$ is the domain of $\widetilde{f}_j$. Similarly, for the space $\mathcal{M}_K(\mathcal{F})$ of piecewise population size models with at most K pieces over some function family $\mathcal{F}$, we define the sign change complexity $\mathscr{S}(\mathcal{M}_K(\mathcal{F}))$ as*

$$\mathscr{S}(\mathcal{M}_K(\mathcal{F})) = \sup_{\eta_1,\eta_2\in\mathcal{M}_K(\mathcal{F})} \left\{ \sigma(\widetilde{\eta}_1 - \widetilde{\eta}_2) \right\}, \tag{2.15}$$

where, again, $\widetilde{\eta}_j$ are related to η_j as given in (2.3)*.*

The following lemma gives a bound on the sign change complexity of a model with at most K pieces in terms of the underlying family of population size functions for each piece.

Lemma 2. *The sign change complexity of the space $\mathcal{M}_K(\mathcal{F})$ of piecewise models with at most K pieces in a function family $\mathcal{F}$ is bounded by the sign change complexity of $\mathcal{F}$ as*

$$\mathscr{S}(\mathcal{M}_K(\mathcal{F})) \leq (2K-2) + (2K-1)\mathscr{S}(\mathcal{F}). \tag{2.16}$$

Proof. Given a pair of piecewise population size functions $\eta_1, \eta_2 \in \mathcal{M}_K(\mathcal{F})$, let $\widetilde{\eta}_1$ and $\widetilde{\eta}_2$ be their respective time-rescaled versions, defined by (2.3). Let $0 < t_1^{(1)} < \cdots < t_{p_1}^{(1)} < \infty$, where $0 \leq p_1 \leq K-1$, (respectively, $0 < t_1^{(2)} < \cdots < t_{p_2}^{(2)} < \infty$, where $0 \leq p_2 \leq K-1$) be the change points of the pieces of η_1 (respectively, η_2). We define $t_0^{(1)} = t_0^{(2)} = 0$ and $t_{p_1+1}^{(1)} = t_{p_2+1}^{(2)} = \infty$. The change points of $\widetilde{\eta}_1$ are given by $R_{\eta_1}(t_i^{(1)})$, where $1 \leq i \leq p_1$, while the change points of $\widetilde{\eta}_2$ are given by $R_{\eta_2}(t_i^{(2)})$, where $1 \leq i \leq p_2$. Let $0 < \tau_1 < \cdots < \tau_p < \infty$ be the union of the change points of $\widetilde{\eta}_1$ and $\widetilde{\eta}_2$, where $0 \leq p \leq p_1 + p_2$. For convention, let $\tau_0 = 0$ and $\tau_{p+1} = \infty$.

Consider the piece (τ_i, τ_{i+1}) for $0 \leq i \leq p$. Let $I_1 = (t_k^{(1)}, t_{k+1}^{(1)})$, where $0 \leq k \leq p_1$, and $I_2 = (t_l^{(2)}, t_{l+1}^{(2)})$, where $0 \leq l \leq p_2$, be the pieces of the original population size functions η_1 and η_2, respectively, such that $(\tau_i, \tau_{i+1}) \subseteq R_{\eta_1}(I_1)$ and $(\tau_i, \tau_{i+1}) \subseteq R_{\eta_2}(I_2)$. Since $\eta_1 \in \mathcal{M}_K(\mathcal{F})$, there exists a function $f_1 \in \mathcal{F}$ such that $\eta_1(t) = f_1(t - t_k^{(1)})$ for all $t \in I_1$. Then, for all $\tau \in R_{\eta_1}(I_1)$,

$$\widetilde{\eta}_1(\tau) = \eta_1\big(R_{\eta_1}^{-1}(\tau)\big) = f_1\big(R_{\eta_1}^{-1}(\tau) - t_k^{(1)}\big) \tag{2.17}$$

$$= \widetilde{f}_1\big(R_{f_1}\big(R_{\eta_1}^{-1}(\tau) - t_k^{(1)}\big)\big) \tag{2.18}$$

$$= \widetilde{f}_1\big(\tau - R_{\eta_1}(t_k^{(1)})\big). \tag{2.19}$$

Similarly, there exists some function $f_2 \in \mathcal{F}$ such that, for all $\tau \in R_{\eta_2}(I_2)$,

$$\widetilde{\eta}_2(\tau) = \widetilde{f}_2\big(\tau - R_{\eta_2}(t_l^{(2)})\big). \tag{2.20}$$

Using (2.19) and (2.20), we see that the number of sign change points of $\widetilde{\eta}_1 - \widetilde{\eta}_2$ in the piece (τ_i, τ_{i+1}) is at most the number of sign change points of $\widetilde{f}_1(\tau - R_{\eta_1}(t_k^{(1)})) - \widetilde{f}_2(\tau - R_{\eta_2}(t_l^{(2)}))$ for $\tau \in (\tau_i, \tau_{i+1})$. Hence, by (2.14), it follows that within each piece (τ_i, τ_{i+1}) for $0 \leq i \leq p$, $\widetilde{\eta}_1 - \widetilde{\eta}_2$ has at most $\mathscr{S}(\mathcal{F})$ sign change points. Also, the point τ_{i+1} itself could be a sign change point in the interval between the last sign change point in piece (τ_i, τ_{i+1}) and the first sign change point in piece (τ_{i+1}, τ_{i+2}) where $0 \leq i \leq p-1$. These are all the possible sign change points of $\widetilde{\eta}_1 - \widetilde{\eta}_2$. Hence,

$$\begin{aligned}\sigma(\widetilde{\eta}_1 - \widetilde{\eta}_2) &\leq p + (p+1)\mathscr{S}(\mathcal{F}) \\ &\leq (p_1 + p_2) + (p_1 + p_2 + 1)\mathscr{S}(\mathcal{F}) \\ &\leq (2K-2) + (2K-1)\mathscr{S}(\mathcal{F}). \end{aligned} \tag{2.21}$$

Since (2.21) holds for all $\eta_1, \eta_2 \in \mathcal{M}_K(\mathcal{F})$, the lemma follows. □

Note that the bound in Lemma 2 is tight for the family $\mathcal{F}_c$ of constant population sizes, for which $\mathscr{S}(\mathcal{F}_c) = 0$ and $\mathscr{S}(\mathcal{M}_K(\mathcal{F}_c)) = 2K-2$.

2.3.3 Identifiability from the SFS

Our main results on identifiability will be proved using a generalization of Descartes' rule of signs for polynomials:

Theorem 3 (Descartes' rule of signs for polynomials)**.** *Consider a degree-n polynomial $p(x) = a_0 + a_1 x + \cdots + a_n x^n$ with real-valued coefficients a_i. The number of positive real roots (counted with multiplicity) of p is at most the number of sign changes between consecutive non-zero terms in the sequence $a_0, a_1, \ldots, a_n$.*

The following theorem generalizes the above classic result to relate the number of sign changes of a piecewise-continuous function f to the number of roots of its Laplace transform:

Theorem 4 (Generalized Descartes' rule of signs)**.** *Let $f : \mathbb{R}_{\geq 0} \to \mathbb{R}$ be a piecewise-continuous function which is not identically zero and with a finite number $\sigma(f)$ of sign changes. Then, the function $G(x)$ defined by*

$$G(x) = \int_0^\infty f(t) e^{-tx} \, dt \tag{2.22}$$

has at most $\sigma(f)$ roots in $\mathbb{R}$ (counted with multiplicity).

Proof. The proof is by induction on the number of sign changes of f. If f has zero sign changes, then without loss of generality, $f(t) \geq 0$ for $t \in (0, \infty)$ and $f(t) > 0$ for some interval $(a, b) \subseteq (0, \infty)$. Hence, $G(x) > 0$ for all x, and the base case holds. Suppose f has $m+1$ sign changes for some $m \geq 0$. Let $\{t_0, \ldots, t_m\}$ be the set of sign change points of $G(x)$. Note that $G(x)$ and $F(x) = e^{t_0 x} G(x)$ have the same real-valued roots (with multiplicity) since $e^{t_0 x} > 0$ for all $x \in \mathbb{R}$. $F'(x)$ is given by

$$F'(x) = \frac{d}{dx} \left(\int_0^\infty f(t) e^{-(t-t_0)x} \, dt \right) = \int_0^\infty (t_0 - t) f(t) e^{-(t-t_0)x} \, dt, \tag{2.23}$$

where the interchange of the differential and integral operators in the second equality is justified by the Leibniz integral rule because f is piecewise continuous over $\mathbb{R}_{\geq 0}$, and both $f(t)e^{-(t-t_0)x}$ and $\frac{d}{dx}(f(t)e^{-(t-t_0)x})$ are jointly continuous over $(p_i, p_{i+1}) \times (-\infty, \infty)$ for each piece (p_i, p_{i+1}) over which f is continuous. Note that the set of sign change points of $(t_0 - t)f(t)$ is $\{t_1, \ldots, t_m\}$. Hence $(t_0 - t)f(t)$ has only m sign changes. By the induction hypothesis, F' has at most m real-valued roots. By Rolle's theorem, the number of real-valued roots of F is at most one more than the number of real-valued roots of F'. Hence, F has at most $m+1$ real-valued roots, implying that G has at most $m+1$ real-valued roots. □

The statement of Theorem 4 and its proof are adapted from Jameson [38, Lemma 4.5] for our setting. Using Theorem 4, we can prove the following main theorem on the identifiability of piecewise models when the pieces are from a family with finite sign change complexity:

Theorem 5. *For a sample of size n, let $\boldsymbol{c} = (c_2, \ldots, c_n)$, where $c_m = \mathbb{E}[T^{(\eta)}_{m,m}]$, for $2 \leq m \leq n$, defined in (2.5). If $\mathscr{S}(\mathcal{F}) < \infty$ and $n \geq 2K + (2K-1)\mathscr{S}(\mathcal{F})$, then no two distinct models $\eta_1, \eta_2 \in \mathcal{M}_K(\mathcal{F})$ can produce the same $(c_2, \ldots, c_n)$. In other words, for $n \geq 2K + (2K-1)\mathscr{S}(\mathcal{F})$, the map $\boldsymbol{c} : \mathcal{M}_K(\mathcal{F}) \to \mathbb{R}_+^{n-1}$ is injective.*

Proof. First, note that if $\mathscr{S}(\mathcal{F}) < \infty$, it follows from Lemma 2 that $\mathscr{S}(\mathcal{M}_K(\mathcal{F})) < \infty$. Suppose there exist two distinct models $\eta_1, \eta_2 \in \mathcal{M}_K(\mathcal{F})$ that produce exactly the same c_m for all $2 \leq m \leq n$. From (2.7), we have that

$$\int_0^\infty (\widetilde{\eta}_1(\tau) - \widetilde{\eta}_2(\tau)) e^{-\binom{m}{2}\tau} d\tau = 0 \tag{2.24}$$

for $2 \leq m \leq n$. If we define the function $G(x)$ as

$$G(x) = \int_0^\infty (\widetilde{\eta}_1(\tau) - \widetilde{\eta}_2(\tau)) e^{-x\tau} d\tau, \tag{2.25}$$

then from (2.24), we see that $\binom{m}{2}$ is a root of $G(x)$ for $2 \leq m \leq n$, and hence, G has at least $n-1$ roots. Applying Theorem 4 to the piecewise continuous function $\widetilde{\eta}_1 - \widetilde{\eta}_2$, we see that G can have at most $\sigma(\widetilde{\eta}_1 - \widetilde{\eta}_2)$ roots. Taking the supremum over all population size functions η_1 and η_2 in $\mathcal{M}_K(\mathcal{F})$, we see that G can have at most $\mathscr{S}(\mathcal{M}_K(\mathcal{F}))$ roots, and by Lemma 2, at most $(2K-2) + (2K-1)\mathscr{S}(\mathcal{F})$ roots. Hence, if $n - 1 > (2K-2) + (2K-1)\mathscr{S}(\mathcal{F})$, we get a contradiction. This implies that if $n \geq 2K + (2K-1)\mathscr{S}(\mathcal{F})$, no two distinct population size functions in $\mathcal{M}_K(\mathcal{F})$ can produce the same $(c_2, \ldots, c_n)$. □

Using Theorem 5, it is simple to derive identifiability results for piecewise-defined population size models over several function families $\mathcal{F}$ that are of biological interest. In particular, we have the following result for the case of piecewise-constant models:

Corollary 6 (Identifiability of piecewise-constant population size models)**.** *For the space $\mathcal{M}_K(\mathcal{F}_c)$ of piecewise-constant population size models, the map $\boldsymbol{c} : \mathcal{M}_K(\mathcal{F}_c) \to \mathbb{R}_+^{n-1}$ is injective if the sample size $n \geq 2K$.*

Proof. As remarked after Lemma 2, for the constant population size function family $\mathcal{F}_c$, $\mathscr{S}(\mathcal{F}_c) = 0$. Hence, by Theorem 5, if $n \geq 2K$, the map $\boldsymbol{c} : \mathcal{M}_K(\mathcal{F}_c) \to \mathbb{R}_+^{n-1}$ is injective. □

The bound in Corollary 6 on the sample size sufficient for identifying piecewise-constant population models is actually tight, since $\mathcal{M}_K(\mathcal{F}_c)$ has $2K-1$ parameters in $\mathbb{R}_+$ and there is no continuous injective function from $\mathbb{R}_+^{2K-1}$ to $\mathbb{R}_+^{n-1}$ if $n < 2K$. (This fact can be proved in multiple ways, such as by the Borsuk-Ulam theorem or the Constant Rank theorem.) We also provide an alternate proof of Corollary 6 that does not explicitly rely on Theorem 5. This alternate proof is based on an argument from linear algebra, and it might be possible to adapt this approach to develop an algebraic algorithm for inferring the parameters of a piecewise-constant population function from the set of expected first coalescence times c_m.

An alternate proof of Corollary 6 based on linear algebra. Let $n \geq 2K$, and suppose there exist two distinct models $\eta^{(1)}, \eta^{(2)} \in \mathcal{M}_K(\mathcal{F}_c)$ that produce exactly the same c_m for all $2 \leq m \leq n$. Let $\widetilde{\eta}^{(1)}$ and $\widetilde{\eta}^{(2)}$ denote the time-rescaled versions of $\eta^{(1)}$ and $\eta^{(2)}$, respectively, as in (2.3). Since $\eta^{(j)}$ is piecewise-constant with at most K pieces, $\widetilde{\eta}^{(j)}$ is also piecewise-constant with the same number of pieces as $\eta^{(j)}$, and $\eta^{(1)} \neq \eta^{(2)}$ implies $\widetilde{\eta}^{(1)} \neq \widetilde{\eta}^{(2)}$. Therefore, $\widetilde{\Delta} := \widetilde{\eta}^{(1)} - \widetilde{\eta}^{(2)}$ is a piecewise-constant function over $[0, \infty)$ with p pieces, where $1 \leq p \leq 2K-1$, and $\widetilde{\Delta}$ is not identically zero. Let $\tau_1 < \cdots < \tau_{p-1}$ denote the change points of $\widetilde{\Delta}$, and define $\tau_0 = 0$ and $\tau_p = \infty$. Suppose $\widetilde{\Delta}(\tau) = \delta_i \in \mathbb{R}$ for all $\tau \in [\tau_{i-1}, \tau_i)$, where $1 \leq i \leq p$. Since $\widetilde{\eta}^{(1)}$ and $\widetilde{\eta}^{(2)}$ produce the same c_m for all $2 \leq m \leq n$, we know that $\widetilde{\Delta}$ satisfies

$$\int_0^\infty \widetilde{\Delta}(\tau) e^{-\binom{m}{2}\tau} \, d\tau = 0, \tag{2.26}$$

for all $2 \leq m \leq n$. Substituting the definition of $\widetilde{\Delta}$ into (2.26) and multiplying by $\binom{m}{2}$, we obtain

$$\sum_{i=1}^{p} \delta_i \left[e^{-\binom{m}{2}\tau_{i-1}} - e^{-\binom{m}{2}\tau_i} \right] = 0, \tag{2.27}$$

for $2 \leq m \leq n$. This defines a linear system $\boldsymbol{A}\boldsymbol{\delta} = \boldsymbol{0}$, where $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_p)$ and $\boldsymbol{A} = (a_{mi})$ is an $(n-1) \times p$ matrix with $a_{mi} := e^{-\binom{m}{2}\tau_{i-1}} - e^{-\binom{m}{2}\tau_i}$ for $2 \leq m \leq n$ and $1 \leq i \leq p$.

Let $\boldsymbol{B} = (b_{mi})$ be the $(n-1) \times p$ matrix formed from $\boldsymbol{A}$ such that the i^{th} column of $\boldsymbol{B}$ is the sum of columns $i, i+1, \ldots, p$ of $\boldsymbol{A}$. Defining $\alpha_i = e^{-\tau_{i-1}}$, note that $b_{mi} = \alpha_i^{\binom{m}{2}}$ for $2 \leq m \leq n$ and $1 \leq i \leq p$. Now, consider the $p \times p$ submatrix $\boldsymbol{C}$ of $\boldsymbol{B}$ consisting of the first p rows of $\boldsymbol{B}$. Since $\alpha_1 > \alpha_2 > \cdots > \alpha_p > 0$, note that $\boldsymbol{C}$ is a generalized Vandermonde matrix, which implies $\det(\boldsymbol{C}) \neq 0$ [22, Ch. XIII, §8]. Hence, $\operatorname{rank}(\boldsymbol{B}) = p$. The rank of $\boldsymbol{A}$ is invariant under elementary column operations, and therefore $\operatorname{rank}(\boldsymbol{A}) = \operatorname{rank}(\boldsymbol{B}) = p$. Therefore, the kernel of $\boldsymbol{A}$ is trivial, and the only solution to (2.27) is $\delta_1 = \delta_2 = \cdots = \delta_p = 0$, which contradicts our assumption that $\widetilde{\Delta} = \widetilde{\eta}^{(1)} - \widetilde{\eta}^{(2)} \not\equiv 0$. □

Another class of models often assumed in population genetic analyses are piecewise-exponential functions, for which we have the following result:

Corollary 7 (Identifiability of piecewise-exponential population size models)**.** *For the space $\mathcal{M}_K(\mathcal{F}_e)$ of piecewise-exponential population size models, the map $\boldsymbol{c} : \mathcal{M}_K(\mathcal{F}_e) \to \mathbb{R}_+^{n-1}$ is injective if the sample size $n \geq 4K - 1$.*

Proof. Let $f_1, f_2 \in \mathcal{F}_e$ be given by

$$f_1(t) = \nu_1 \exp(\beta_1 t), \tag{2.28}$$
$$f_2(t) = \nu_2 \exp(\beta_2 t), \tag{2.29}$$

where $t \in \mathbb{R}_{\geq 0}$, $\nu_1, \nu_2 \in \mathbb{R}_+$ and $\beta_1, \beta_2 \in \mathbb{R}$. Then, for $i = 1, 2$, the time-rescaled function $\widetilde{f}_i$ is given by

$$\widetilde{f}_i(\tau) = \frac{\nu_i}{1 - \nu_i \beta_i \tau}, \tag{2.30}$$

for $\tau \in \mathrm{Dom}(\widetilde{f}_i) = R_{f_i}(\mathbb{R}_{\geq 0}) = [0, \frac{1}{\nu_i \beta_i})$. From (2.30), it can be seen that $\widetilde{f}_1$ and $\widetilde{f}_2$ are continuous in their domains. Furthermore, for any given $a \in \mathbb{R}_{\geq 0}$, there is at most one τ, where $\tau \in \mathrm{Dom}(\widetilde{f}_1)$ and $\tau - a \in \mathrm{Dom}(\widetilde{f}_2)$, such that $g(\tau) := \widetilde{f}_1(\tau) - \widetilde{f}_2(\tau - a) = 0$, implying $\sigma(g) \leq 1$. By the definition of sign change complexity in (2.14), it then follows that $\mathscr{S}(\mathcal{F}_e) \leq 1$ for the exponential population family $\mathcal{F}_e$. Hence, applying Theorem 5, we conclude that $n \geq 4K - 1$ suffices for the map $\boldsymbol{c} : \mathcal{M}_K(\mathcal{F}_e) \to \mathbb{R}_+^{n-1}$ to be injective. □

For the identifiability of piecewise population size models from the SFS data, we first note the following lemma:

Lemma 8. *Consider a piecewise population size function $\eta \in \mathcal{M}_K(\mathcal{F})$. Consider a sample of size $n \geq 2K + (2K-1)\mathscr{S}(\mathcal{F})$ and suppose the function η produces $\mathbb{E}[T^{(\eta)}_{m,m}] = c_m$ for $2 \leq m \leq n$. Then, for every fixed $\kappa \in \mathbb{R}_+$, there exists a unique piecewise population size function $\zeta \in \mathcal{M}_K(\mathcal{F})$ with $\mathbb{E}[T^{(\zeta)}_{m,m}] = \kappa c_m$ for $2 \leq m \leq n$. Furthermore, this population size function ζ is given by $\zeta(t) = \kappa\eta(t/\kappa)$.*

Proof. For the population size function $\zeta(t)$ defined by $\zeta(t) = \kappa\eta(t/\kappa)$, note that $R_\zeta(t)$ is given by

$$R_\zeta(t) = \int_0^t \frac{1}{\zeta(x)} dx = \int_0^t \frac{1}{\kappa\eta(x/\kappa)} dx = \int_0^{t/\kappa} \frac{1}{\eta(x)} dx = R_\eta(t/\kappa).$$

$\mathbb{E}[T^{(\zeta)}_{m,m}]$ is then given by

$$\mathbb{E}[T^{(\zeta)}_{m,m}] = \int_0^\infty \exp\left[-\binom{m}{2} R_\eta\left(\frac{t}{\kappa}\right)\right] dt \tag{2.31}$$

$$= \kappa \int_0^\infty \exp\left[-\binom{m}{2} R_\eta(t)\right] dt \tag{2.32}$$

$$= \kappa\, \mathbb{E}[T^{(\eta)}_{m,m}]. \tag{2.33}$$

Since $n \geq 2K + (2K-1)\mathscr{S}(\mathcal{F})$, by Theorem 5, ζ is the unique population size function in $\mathcal{M}_K(\mathcal{F})$ with $\mathbb{E}[T^{(\zeta)}_{m,m}] = \kappa c_m$ for $2 \leq m \leq n$. □

Given two models $\eta, \zeta \in \mathcal{M}_K$, we say that η and ζ are equivalent, and write $\eta \sim \zeta$, if they are related by a rescaling of change points and population sizes as described in Lemma 8. Then, combining Lemma 1, Theorem 5, and Lemma 8, we obtain the following theorem:

Theorem 9. *If $\mathscr{S}(\mathcal{F}) < \infty$ and $n \geq 2K + (2K-1)\mathscr{S}(\mathcal{F})$, then, for each expected SFS $(\xi_{n,1}, \ldots, \xi_{n,n-1})$, there exists a unique equivalence class $[\eta]$ of models in $\mathcal{M}_K(\mathcal{F})/\!\sim$ consistent with $(\xi_{n,1}, \ldots, \xi_{n,n-1})$.*

2.3.4 Extension to the folded frequency spectrum

To generate the SFS from genomic sequence data, one needs to know the identities of the ancestral and mutant alleles at each site. To avoid this problem, a commonly employed strategy in population genetic inference involves "folding" the SFS. More precisely, for a sample of size n, the i-th entry of the folded SFS $\boldsymbol{\chi} = (\chi_{n,1}, \ldots, \chi_{\lfloor n/2 \rfloor})$ is defined by

$$\chi_{n,i} = \frac{\xi_{n,i} + \xi_{n,n-i}}{1 + \delta_{i,n-i}}, \tag{2.34}$$

where $1 \leq i \leq \lfloor n/2 \rfloor$. In particular, $\chi_{n,i}$ is the proportion of polymorphic sites that have i copies of the minor allele. For any sample size n, since $\boldsymbol{\chi}$ is a vector of approximately half the dimension as $\boldsymbol{\xi}$, we might expect to require roughly twice as many samples to recover the demographic model from $\boldsymbol{\chi}$ compared to $\boldsymbol{\xi}$. This is indeed the case. Given the folded SFS $\boldsymbol{\chi}$, the following theorem establishes a sufficiency condition on the sample size for identifying demographic models in $\mathcal{M}_K(\mathcal{F})$:

Theorem 10. *If $\mathscr{S}(\mathcal{F}) < \infty$ and $n \geq 2(2K-1)(1+\mathscr{S}(\mathcal{F}))$, then, for each expected folded SFS $\boldsymbol{\chi} = (\chi_{n,1}, \ldots, \chi_{n,\lfloor n/2 \rfloor})$, there exists a unique equivalence class $[\eta]$ of models in $\mathcal{M}_K(\mathcal{F})/\!\sim$ consistent with $\boldsymbol{\chi}$.*

To prove Theorem 10, we first need a lemma that characterizes a certain symmetry property of the invertible matrix that relates the genealogical quantities $\boldsymbol{\gamma}$ and $\boldsymbol{c}$ introduced in the proof of Lemma 1.

Lemma 11. *For a sample of size n, let W be the $(n-1) \times (n-1)$ invertible matrix such that $\gamma_{n,b} = \sum_{m=2}^{n} W_{b,m} c_m$, where $\gamma_{n,b}$ is the total expected branch length subtending b leaves and $c_m = \mathbb{E}[T^{(\eta)}_{m,m}]$. Then, for every b and m, where $1 \leq b \leq n-1$ and $2 \leq m \leq n$, we have the following identities:*

$$W_{b,m} + W_{n-b,m} = 0, \quad \text{if } m \text{ is odd}, \tag{2.35}$$

$$W_{b,m} - W_{n-b,m} = 0, \quad \text{if } m \text{ is even}. \tag{2.36}$$

Proof. From the proof of Lemma 1, it can be seen that the matrix $\boldsymbol{W}$ is the product of 3 matrices whose entries are explicitly given combinatorial expressions. However, using Zeilberger's algorithm [66], Polanksi and Kimmel [69, Equations 13–15] also derived the following recurrence relation for the entries of $\boldsymbol{W}$:

$$\begin{aligned} W_{b,2} &= \frac{6}{(n+1)}, \\ W_{b,3} &= 30\frac{(n-2b)}{(n+1)(n+2)}, \\ W_{b,m+2} &= f(n,m)W_{b,m} + g(n,m)(n-2b)W_{b,m+1}, \end{aligned} \tag{2.37}$$

where $f(n,m)$ and $g(n,m)$ are rational functions of n and m given by

$$f(n,m) = -\frac{(1+m)(3+2m)(n-m)}{m(2m-1)(n+m+1)},$$
$$g(n,m) = \frac{(3+2m)}{m(n+m+1)}.$$

It will be easy to prove our lemma by induction on m using (2.37). The base cases are easy to check:

$$W_{b,2} - W_{n-b,2} = 0,$$
$$W_{b,3} + W_{n-b,3} = 30\frac{(n-2b)+(n-2(n-b))}{(n+1)(n+2)} = 0.$$

Using (2.37), we see that if m is odd,

$$\begin{aligned}
&W_{b,m+2} + W_{n-b,m+2} \\
&= f(n,m)(W_{b,m} + W_{n-b,m}) \\
&\quad + g(n,m)\{(n-2b)W_{b,m+1} + [n-2(n-b)]W_{n-b,m+1}\} \\
&= f(n,m)(W_{b,m} + W_{n-b,m}) + g(n,m)(n-2b)(W_{b,m+1} - W_{n-b,m+1}) \\
&= 0,
\end{aligned}$$

where the last equality follows from the induction hypothesis which implies $W_{b,m}+W_{n-b,m} = 0$ and $W_{b,m+1} - W_{n-b,m+1} = 0$. Similarly, if m is even,

$$\begin{aligned}
&W_{b,m+2} - W_{n-b,m+2} \\
&= f(n,m)(W_{b,m} - W_{n-b,m}) \\
&\quad + g(n,m)\{(n-2b)W_{b,m+1} - [n-2(n-b)]W_{n-b,m+1}\} \\
&= f(n,m)(W_{b,m} - W_{n-b,m}) + g(n,m)(n-2b)(W_{b,m+1} + W_{n-b,m+1}) \\
&= 0,
\end{aligned}$$

where again the last equality follows from the induction hypothesis. □

Proof of Theorem 10. For a sample of size n in the coalescent, let $\gamma_{n,b}$ be the total expected branch length subtending b leaves, for $1 \le b \le n-1$. Then, there exists a positive constant κ such that

$$\frac{\gamma_{n,d} + \gamma_{n,n-d}}{1 + \delta_{d,n-d}} = \kappa\chi_{n,d}, \tag{2.38}$$

for all $1 \le d \le \lfloor n/2 \rfloor$. Let $f_{n,d} = \frac{\gamma_{n,d}+\gamma_{n,n-d}}{1+\delta_{d,n-d}}$. The relationship between $\boldsymbol{f} = (f_{n,1}, \ldots, f_{n,\lfloor n/2 \rfloor})$ and $\boldsymbol{\gamma} = (\gamma_{n,1}, \ldots, \gamma_{n,n-1})$ can be described by the linear equation

$$\boldsymbol{f} = \boldsymbol{Z}\boldsymbol{\gamma}, \tag{2.39}$$

where $\boldsymbol{Z}$ is an $\lfloor n/2 \rfloor \times (n-1)$ matrix with entries given by

$$Z_{dj} = \begin{cases} 1, & \text{if } j = d \text{ or } j = n-d, \\ 0, & \text{otherwise,} \end{cases} \tag{2.40}$$

for $1 \leq d \leq \lfloor n/2 \rfloor$ and $1 \leq j \leq n-1$. Hence, $\dim(\ker(\boldsymbol{Z})) = \lfloor (n-1)/2 \rfloor$.

From Lemma 1, we know that $\boldsymbol{\gamma}$ and $\boldsymbol{c} = (c_2, \ldots, c_n)$ are related as $\boldsymbol{\gamma} = \boldsymbol{W}\boldsymbol{c}$, where $\boldsymbol{W} = (W_{b,m})$ is an $(n-1) \times (n-1)$ invertible matrix, where $1 \leq b \leq n-1$ and $2 \leq m \leq n$. Hence,

$$\boldsymbol{f} = \boldsymbol{Y}\boldsymbol{c}, \tag{2.41}$$

where $\boldsymbol{Y} := \boldsymbol{Z}\boldsymbol{W}$. Since $Y_{b,m} = W_{b,m} + W_{n-b,m}$, we know from Lemma 11 that $Y_{b,m} = 0$ for all odd values of m. Therefore, every other column of the matrix $\boldsymbol{Y}$ is zero. This implies that $\text{span}(\{\boldsymbol{e}_3, \boldsymbol{e}_5, \ldots, \boldsymbol{e}_{n-\nvDash\{n \text{ even}\}}\}) \subseteq \ker(\boldsymbol{Y})$, where $\boldsymbol{e}_i$ is an $(n-1)$-dimensional unit vector defined as $\boldsymbol{e}_i = (e_{i,2}, \ldots, e_{i,n})$, with $e_{i,i} = 1$ and $e_{i,j} = 0$ for $i \neq j$. Note that $n - \nvDash\{n \text{ even}\} = 2\lfloor (n-1)/2 \rfloor + 1$ and $\dim(\text{span}(\{\boldsymbol{e}_3, \boldsymbol{e}_5, \ldots, \boldsymbol{e}_{2\lfloor (n-1)/2 \rfloor + 1}\})) = \lfloor (n-1)/2 \rfloor$. Now, since $\boldsymbol{W}$ is invertible, $\dim(\ker(\boldsymbol{Y})) = \dim(\ker(\boldsymbol{Z}\boldsymbol{W})) = \dim(\ker(\boldsymbol{Z})) = \lfloor (n-1)/2 \rfloor$. Therefore,

$$\ker(\boldsymbol{Y}) = \text{span}\left(\{\boldsymbol{e}_3, \boldsymbol{e}_5, \ldots, \boldsymbol{e}_{2\lfloor (n-1)/2 \rfloor + 1}\}\right). \tag{2.42}$$

Suppose there exist two distinct models $\eta_1, \eta_2 \in \mathcal{M}_K(\mathcal{F})$ that produce the same folded SFS $\boldsymbol{f}$. Let $\boldsymbol{c}^{(1)}$ and $\boldsymbol{c}^{(2)}$ be the vector of genealogical quantities for models η_1 and η_2 respectively, where $c_m^{(1)} = \mathbb{E}[T_{m,m}^{(\eta_1)}]$ and $c_m^{(2)} = \mathbb{E}[T_{m,m}^{(\eta_2)}]$, $2 \leq m \leq n$. From (2.41), we know that $\boldsymbol{c}^{(1)} - \boldsymbol{c}^{(2)} \in \ker(\boldsymbol{Y})$. Using (2.42), $c_m^{(1)} - c_m^{(2)}$ can be written as

$$c_m^{(1)} - c_m^{(2)} = \sum_{l=1}^{\lfloor (n-1)/2 \rfloor} \alpha_l e_{2l+1,m}, \tag{2.43}$$

for some $\alpha_l \in \mathbb{R}$. Since $e_{ij} = 0$ for $i \neq j$, (2.43) implies that $c_m^{(1)} - c_m^{(2)} = 0$ for all even values of m, where $2 \leq m \leq n$. Now applying a similar argument as in the proof of Theorem 5 to $c_m^{(1)} - c_m^{(2)}$ for even values of m, we conclude that if $\lceil (n-1)/2 \rceil > (2K-2) + (2K-1)\mathscr{S}(\mathcal{F})$, then no two distinct models $\eta_1, \eta_2 \in \mathcal{M}_K(\mathcal{F})$ can produce the same $\boldsymbol{f}$. This implies that a sample size $n \geq 2(2K-1)(1+\mathscr{S}(\mathcal{F}))$ suffices for identifying the population size function in $\mathcal{M}_K(\mathcal{F})$ from the folded SFS $\boldsymbol{f}$, and the conclusion of the theorem follows from (2.38) and Lemma 8. □

2.4 The counterexample of Myers *et al.*

Myers *et al.* [61] provided an explicit counterexample to the identifiability of population size models from the allelic frequency spectrum. In our notation, they provided two time-rescaled population size functions $\widetilde{\eta}_1$ and $\widetilde{\eta}_2$ given by

$$\widetilde{\eta}_1(\tau) = N, \tag{2.44}$$

$$\widetilde{\eta}_2(\tau) = N(1 - 9F(\tau)), \tag{2.45}$$

where N is an arbitrary positive constant, and the function F is given by the convolution

$$F(\tau) = \int_0^\tau f_0(\tau - u) f_1(u) du, \tag{2.46}$$

where f_0 and f_1 are given by

$$f_0(\tau) = \exp(-1/\tau^2), \tag{2.47}$$

$$f_1(\tau) = \frac{\cos(\pi^2/\tau)\exp(-\tau/8)}{\sqrt{\tau}}. \tag{2.48}$$

Both functions f_1 and F have increasingly frequent oscillations as $\tau \downarrow 0$ so that $\sigma(\widetilde{\eta}_1 - \widetilde{\eta}_2) = \sigma(F) = \infty$. This is why Theorem 5 does not apply to this example. Indeed, by an argument using the Laplace transforms of f_1 and F, Myers *et al.* showed that the function $G(x)$ defined in (2.22) in terms of F has roots at $-\binom{m}{2}$ for each $m \geq 2$.

2.5 Discussion

In human genetics, several large-sample datasets have recently become available, with sample sizes on the order of several thousands to tens of thousands of individuals [1, 12, 19, 62, 85]. The patterns of polymorphism observed in these datasets deviate significantly from that expected under a constant population size, and there has been much interest to infer recent and ancient human demographic changes that might explain these deviations [23, 48, 51]. Clearly, model identifiability is an important prerequisite for such statistical inference problems. In this chapter, we have obtained mathematically rigorous identifiability results for demographic inference by showing that piecewise-defined population size functions over a wide class of function families are completely determined by the SFS, provided that the sample is sufficiently large. Furthermore, we have provided explicit bounds on the sample sizes that are sufficient for identifying such piecewise population size functions. These bounds depend on the number of pieces and the functional type of each piece. For piecewise-constant population size models, which have been extensively applied in demographic inference studies, our bounds are tight. We have also given analogous results for identifiability from the folded SFS, a variant of the SFS that is oblivious to the identities of the ancestral and mutant alleles.

Our work suggests several interesting avenues for future research. In our work, we examined the identifiability of demography from the expected SFS data. However, if one were to use the complete sequence data or other summary statistics such as the length distribution of shared haplotype tracts, it might be possible to uniquely identify the demography using even smaller sample sizes than that needed when using only the SFS. Indeed, several demographic inference methods have been developed to infer historical population size changes

from such data using anywhere from a pair of genomic sequences [34, 48, 64] to tens of such sequences [79], and it is important to theoretically characterize the power and limitations of both the data and the inference methods.

Another interesting direction for exploration is to understand the extent to which ancient demographic events can be inferred from the SFS in practice. The population size changes sufficiently far back in the past are likely to have only a marginal effect on the SFS since the individuals in the sample are highly likely to have found a common ancestor by such ancient times. Our identifiability results apply in the limit that the genome length is infinite, which allows one to estimate the entries of the expected SFS exactly. However, a finite length genome does not permit exact estimation of the SFS, and it might be difficult to resolve very ancient demographic events. Another question related to this issue is understanding the sensitivity of the SFS to perturbations in the demographic parameters. This is important for quantifying the extent to which errors in estimating the expected SFS from data affect the parameter estimates in inferred demographic models.

It would also be interesting to consider the possibility of developing an algebraic algorithm for demographic inference that closely mimics the linear algebraic proof of Corollary 6 provided in Section 2.3.3. For example, using a sample of size $K + 1$, one could consider inferring a piecewise-constant model with K pieces, with one piece for each of the most recent $K - 1$ generations and another piece for the population size further back in time. (Here we are considering a restricted class of piecewise-constant population size functions with fixed change points, so the minimum sample size needed for distinguishing such models using the SFS is $K + 1$ rather than $2K$.) Such an algebraic algorithm could provide a more principled way of inferring demographic parameters, compared to existing inference methods that rely on optimization procedures which lack theoretical guarantees for functions with multiple local optima.

Chapter 3

Inference of the population size function from SFS data

3.1 Overview

Motivated by recent large-sample exome-sequencing studies [62,85] which have greater power to capture rare variants than previous analyses involving small sample sizes [23, 31], in this chapter, we study the problem of efficiently inferring population size changes and mutation rates in a large randomly mating population using the SFS of a sample of individuals at multiple loci.

There have been several previous approaches to this problem. Coventry *et al.* [12] developed a method based on coalescent tree simulations to infer population size changes and per-locus mutation rates, and applied this to exome-sequencing data from approximately 10,000 individuals at 2 genes. Using coalescent simulations to empirically estimate the expected SFS under a given demographic model, Coventry *et al.* compute a likelihood function for the demographic model by comparing the expected and observed SFS. Nelson *et al.* [62] have also applied this method to a dataset of 11,000 individuals of European ancestry (CEU) sequenced at 188 genes to infer a recent epoch of exponential population growth. Excoffier *et al.* [16] have developed a software package that uses coalescent tree simulations to estimate the expected joint SFS of multiple subpopulations for inferring potentially very complex demographic scenarios. This problem has also been approached from the diffusion perspective. Gutenkunst *et al.* [31] used numerical schemes to solve the partial differential equation for the density of segregating sites at a given derived allele frequency, while Lukić *et al.* [52] approximated the solutions to these equations using spectral methods.

Similar to the inference methods of Coventry *et al.* and Excoffier *et al.*, we also work in the coalescent framework rather than in the diffusion setting. However, our method differs from existing methods in several ways. First, our method is based on exact computation and does not require any Monte-Carlo coalescent simulations. We use an efficient algorithmic adaption of the analytic theory of the SFS for deterministically varying population size

models that was developed by Polanski *et al.* [68] and Polanski and Kimmel [69]. As a result, our method is extremely efficient compared to simulation-based approaches, allowing us to more thoroughly search the space of demographic models of interest. Second, our method uses the Poisson Random Field (PRF) approximation proposed by Sawyer and Hartl [76]. Under this approximation, all the segregating sites in a single locus are assumed to be far enough apart to be completely unlinked. This is also the same approximation made by the numerical and spectral methods for demographic inference under the Wright-Fisher diffusion process [31,52]. At the other extreme, the method of Coventry *et al.* for population size inference assumes that all the segregating sites in a locus are completely linked. Both these model simplifications, which assume that a locus is a collection of completely linked or completely unlinked sites, are biologically unrealistic. However, as we demonstrate in our results on simulated data, our method can recover demographic parameters accurately even when the data is generated under realistic recombination rates. Third, working in the PRF model also confers our method a significant computational benefit. Under the PRF model, we can write down efficiently computable expressions for the the maximum likelihood estimate of the mutation rates at each locus. This contrasts with the method of Coventry *et al.* where a line search has to be performed over the mutation rates due to the Monte-Carlo integration involved. Fourth, under the PRF model, the maximum likelihood estimate for the demographic model has a nice mathematical interpretation in that it is the demographic model which minimizes the KL divergence of the expected SFS from the observed SFS. Our method also has advantages over the diffusion-based methods of Gutenkunst *et al.* [31] and Lukić *et al.* [52]. Since we are working in the coalescent framework, the running time of our method is independent of the population size and our computations are exact up to numerical precision. On the other hand, the method of Gutenkunst *et al.* discretizes the allele frequency space, while the spectral method of Lukić *et al.* [52] has to choose the number of basis elements for the representation of the Wright-Fisher transition density function. Finally, since our method is based on a likelihood function that is computed exactly, we can take advantage of automatic differentiation [24] to compute numerically exact gradients of the likelihood function, and thus take advantage of gradient-based algorithms for optimization over the space of demographic parameters.

As we have seen in previous chapters, a result of Myers *et al.* [61] shows that it is information-theoretically not possible to uniquely recover general population size functions from the SFS. However, as we proved in Chapter 2, if we assume that the true population size function is piecewise-constant or piecewise-exponential, then the expected SFS of a large enough sample uniquely determines the piecewise population size function. Taking advantage of this identifiability result, we perform inference under the assumption that the true population size function is piecewise-exponential (which also allows for constant pieces).

The rest of this chapter is organized as follows. We introduce the relevant notation in Section 3.2, and describe some of the details of our method in Section 3.3. The details of the theory and computation in our method are given in 3.3.2. In Section 3.4, we demonstrate our method on simulated data using biologically reasonable parameters.

3.2 Notation

The data we wish to analyze, denoted by $\mathcal{D}$, consists of the unnormalized SFS for n haploid (or $n/2$ diploid) individuals at each of G loci located sufficiently far apart in the genome. The unnormalized SFS at locus g is a vector $\boldsymbol{o}^{(g)} = (o_1^{(g)}, \ldots, o_{n-1}^{(g)})$, where $o_i^{(g)}$ is the number (rather than the probability) of segregating sites at locus g which have i copies of the mutant allele among the n alleles at that site. For notational convenience, let $o^{(g)} = \sum_{i=1}^{n} o_i^{(g)}$ be the total number of segregating sites in the sample at locus g. We are also given the length $m^{(g)}$ of each locus g. Given the data $\mathcal{D}$, our goal is to infer the haploid piecewise-exponential effective population size function $N(t)$ and the per-base locus specific mutation rates $\mu^{(g)}$. Note that in this chapter, we work with the population size functions $N(t)$ in physical units rather than coalescent rate functions $\eta(t)$ used in Chapter 2. The conversion from $\eta(t)$ to $N(t)$ can be done if we have knowledge of, for example, the ancestral population size or the mutation rate at some locus in physical units (i.e. per generation). We will use Φ to denote the parameterization of the population size function in the family of piecewise-exponential models we consider in this work. While we state our method assuming we have access to the SFS, we can just as easily work with the folded SFS if the identities of the ancestral and mutant alleles are uncertain.

3.3 Method

Let us first restrict attention to a single locus g. For a locus with length m bases and per-base per-generation mutation rate μ, let θ denote the population-scaled mutation rate for the whole locus. Specifically, in the infinite-sites model, θ is given by,

$$\theta = 4N_r m\mu,$$

where N_r denotes a reference population size which is used as a scaling parameter. We wish to compute the probability of the observed frequency spectrum $\boldsymbol{o}$ at locus g under the infinite-sites model. If all the sites in the locus were completely linked and the n individuals in the sample are related according to the coalescent tree T, then the probability of observing the frequency spectrum $\boldsymbol{o}$ is given by,

$$\mathbb{P}(\boldsymbol{o} \mid T, \Phi, \theta) = \prod_{i=1}^{n-1} \exp\left(-\frac{\theta}{2}\tau_{n,i}(T)\right) \frac{\left(\frac{\theta}{2}\tau_{n,i}(T)\right)^{o_i}}{o_i!}, \tag{3.1}$$

where $\tau_{n,i}(T)$ is the sum of the lengths of branches in the coalescent tree T which subtend $i$descendant leaves. The explanation for (3.1) is as follows. In the infinite-sites model, mutations are dropped on the coalescent tree according to a Poisson process with rate $\theta/2$, and every mutation generates a new segregating site. A mutation creates a segregating site with i mutant alleles if it occurs on a branch that subtends i descendants. To avoid unwieldy notation, we drop the dependence on the tree T for the branch lengths $\tau_{n,i}(T)$. To compute

the probability of the observed frequency spectrum $\boldsymbol{o}$, we have to integrate (3.1) over the distribution $f(T \mid \Phi)$ of coalescent trees T in the space of coalescent trees $\mathcal{T}_n$ over n leaves. Abusing notation, this can be written as

$$\begin{aligned}
\mathbb{P}(\boldsymbol{o} \mid \Phi, \theta) &= \int_{\mathcal{T}_n} \mathbb{P}(\boldsymbol{o} \mid T, \Phi, \theta) f(T \mid \Phi)\ dT \\
&= \int_{\mathcal{T}_n} \left[\prod_{i=1}^{n-1} \exp\left(-\frac{\theta}{2}\tau_{n,i}\right) \frac{\left(\frac{\theta}{2}\tau_{n,i}\right)^{o_i}}{o_i!} \right] f(T \mid \Phi)\ dT \\
&= \int_{\mathcal{T}_n} \left[\prod_{i=1}^{n-1} \frac{\left(\frac{\theta}{2}\tau_{n,i}\right)^{o_i}}{o_i!} \right] \exp\left(-\frac{\theta}{2}L_n\right) f(T \mid \Phi)\ dT \\
\mathbb{P}(\boldsymbol{o} \mid \Phi, \theta) &= \int_{\mathcal{T}_n} \binom{o}{o_1, \ldots, o_n} \left[\prod_{i=1}^{n-1} \left(\frac{\tau_{n,i}}{L_n}\right)^{o_i} \right] \exp\left(-\frac{\theta}{2}L_n\right) \frac{\left(\frac{\theta}{2}L_n\right)^{o}}{o!} f(T \mid \Phi)\ dT. \qquad (3.2)
\end{aligned}$$

In (3.2), L_n is the total branch length of the tree T on n haploid individuals. It is not known how to efficiently and exactly compute (3.2), even when Φ represents the constant population size model. Coventry *et al.* [12] and Nelson *et al.* [62] approximate the integral in (3.2) by drawing random coalescent trees under the demographic model Φ. In order to find the MLE for θ, they repeat this Monte-Carlo integration for each value of θ in some grid.

3.3.1 Poisson Random Field approximation

In our method, we use the Poisson Random Field (PRF) approximation of Sawyer and Hartl [76] which assumes that all the sites in a given locus are completely unlinked, and hence the underlying coalescent tree at each site is independent. Under this assumption, the probability of the frequency spectrum $\boldsymbol{o}$ is given by,

$$\begin{aligned}
\mathbb{P}(\boldsymbol{o} \mid \Phi, \theta) &= \prod_{i=1}^{n-1} \frac{\left(\frac{\theta}{2}\mathbb{E}[\tau_{n,i}]\right)^{o_i}}{o_i!} \exp\left(-\frac{\theta}{2}\mathbb{E}[L_n]\right) \\
&= C \prod_{i=1}^{n-1} \left(\frac{\theta}{2}\mathbb{E}[\tau_{n,i}]\right)^{o_i} \exp\left(-\frac{\theta}{2}\mathbb{E}[L_n]\right) \qquad (3.3)
\end{aligned}$$

where the expectation in (3.3) is taken over the distribution on coalescent trees T over n leaves, and $C = \prod_{i=1}^{n-1} \frac{1}{o_i!}$ is a data-dependent constant that can be ignored for maximum likelihood estimation. Hence, under the PRF approximation, the problem of computing the likelihood in (3.3) reduces to that of computing the quantities $\mathbb{E}[\tau_{n,i}]$ and $\mathbb{E}[L_n]$ under the given demographic model Φ. Using the theory of the SFS for variable population sizes developed by Polanski *et al.* [68] and Polanski and Kimmel [69], we can develop an efficient algorithm to *numerically stably* and *exactly* compute $\mathbb{E}[\tau_{n,i}]$ and $\mathbb{E}[L_n]$ for a wide class of population size functions $N(t)$. In our work, we consider the family of piecewise-exponential

functions. The details of computing $\mathbb{E}[\tau_{n,i}]$ and $\mathbb{E}[L_n]$ for such a class of population size functions is given in Section 3.3.2. Taking logarithms on both sides in (3.3), we get the following log-likelihood for the demographic model Φ and mutation rate θ at this locus,

$$\mathcal{L}(\Phi, \theta) = \log \mathbb{P}(\boldsymbol{o} \mid \Phi, \theta) = \sum_{i=1}^{n-1} o_i \left(\log \mathbb{E}[\tau_{n,i}] + \log \theta\right) - \frac{\theta}{2}\mathbb{E}[L_n] + \text{constant}(\boldsymbol{o}). \tag{3.4}$$

Assuming the loci are all completely unlinked, the log-likelihood for one locus given in (3.4) can be summed across all loci g to get a log-likelihood for the entire dataset $\mathcal{D}$,

$$\begin{aligned}\mathcal{L}(\Phi, \{\theta^{(g)}\}_{g=1}^{G}) &= \log \mathbb{P}(\mathcal{D} \mid \Phi, \{\theta^{(g)}\}_{g=1}^{G}) \\ &= \sum_{g=1}^{G} \left[\sum_{i=1}^{n-1} o_i^{(g)} \left(\log \mathbb{E}[\tau_{n,i}] + \log \theta^{(g)}\right) - \frac{\theta^{(g)}}{2}\mathbb{E}[L_n]\right] + \text{constant}(\boldsymbol{o}). \end{aligned} \tag{3.5}$$

It is easy to see that $\mathcal{L}$ is a concave function of the mutation rates $\theta^{(g)}$, since the Hessian H of $\mathcal{L}$ with respect to $\theta^{(g)}$ is given by,

$$H_{g,h} = \frac{\partial^2 \mathcal{L}}{\partial \theta^{(g)} \partial \theta^{(h)}} = -\delta_{g,h} \frac{1}{{\theta^{(g)}}^2} \sum_{i=1}^{n-1} o_i^{(g)}, \tag{3.6}$$

showing that H is negative definite. Hence, the mutation rates of the loci that maximize $\mathcal{L}$ are the solutions of,

$$0 = \frac{\partial \mathcal{L}}{\partial \theta^{(g)}} = \frac{1}{\theta^{(g)}} \sum_{i=1}^{n-1} o_i^{(g)} - \frac{1}{2}\mathbb{E}[L_n], \tag{3.7}$$

yielding the maximum likelihood estimate for the mutation rate $\theta^{(g)}$ at locus g given the demographic model,

$$\widehat{\theta}^{(g)} = \frac{2\sum_{i=1}^{n-1} o_i^{(g)}}{\mathbb{E}[L_n]}. \tag{3.8}$$

Note that for a constant population size, (3.8) is the same as Watterson's estimator $\widehat{\theta}_W^{(g)}$ for the mutation rate [92],

$$\widehat{\theta}_W^{(g)} = \frac{\sum_{i=1}^{n-1} o_i^{(g)}}{\sum_{i=1}^{n-1} \frac{1}{i}},$$

since for a constant population size, $\mathbb{E}[L_n] = \sum_{i=1}^{n-1} \frac{1}{i}$. Substituting the MLE for $\theta^{(g)}$ in (3.8) into (3.5), we have the log-likelihood with the optimal mutation rates,

$$\mathcal{L}(\Phi) = \sum_{i=1}^{n-1} \left[\left(\sum_{g=1}^{G} o_i^{(g)}\right) \log\left(\frac{\mathbb{E}[\tau_{n,i}]}{\mathbb{E}[L_n]}\right)\right] + \text{constant}(\boldsymbol{o}). \tag{3.9}$$

If we define the n-dimensional discrete probability distributions $\widetilde{o}$ and $\widetilde{\tau}_n$ by

$$\widetilde{o}_k = \frac{\sum_{g=1}^{G} o_k^{(g)}}{\sum_{i=1}^{n-1} \sum_{g=1}^{G} o_i^{(g)}},$$

and

$$\widetilde{\tau}_{n,k} = \frac{\mathbb{E}[\tau_{n,k}]}{\mathbb{E}[L_n]},$$

where $1 \leq k \leq n-1$, then we see that the demographic model Φ^* that is the MLE of the likelihood function $\mathcal{L}(\Phi)$ in (3.9) is given by,

$$\begin{aligned} \Phi^* &= \arg\max_{\Phi} \mathcal{L}(\Phi) \\ &= \arg\min_{\Phi} \mathrm{KL}(\widetilde{o} \,||\, \widetilde{\tau}_{2n}), \end{aligned} \tag{3.10}$$

where $\mathrm{KL}(P \,||\, Q)$ denotes the KL divergence of distribution Q from P. Hence, given the demographic model, we can efficiently infer the optimal mutation rate at each locus independently according to (3.8) and compute the log-likelihood using (3.9). We can also compute the gradient of $\mathcal{L}(\Phi)$ with respect to Φ using automatic differentiation [24], and search over the space of demographic models using a gradient-based optimization method.

3.3.2 Computing the expected SFS under a variable population size

In this section, we describe the details of computing the quantities $\mathbb{E}[\tau_{n,i}]$ and $\mathbb{E}[L_n]$, the expected branch length subtending i leaves and the expected total branch length respectively, when a coalescent tree is drawn over n individuals according to Kingman's coalescent with demographic model Φ.

Polanski and Kimmel [69] showed that $\tau_{n,i}$ and $\mathbb{E}[L_n]$ can be computed efficiently and numerically stably using the relations,

$$\mathbb{E}[\tau_{n,i}] = \sum_{m=2}^{n} W_{n,i,m} c_m, \tag{3.11}$$

$$\mathbb{E}[L_n] = \sum_{m=2}^{n} V_{n,m} c_m, \tag{3.12}$$

where the coefficients $W_{n,i,m}$ and $V_{n,m}$ are efficiently computable by dynamic programming [69, Equations 12–15], the coefficients c_m are given by the integral,

$$c_m = \int_0^\infty \exp\left[-\binom{m}{2} R(t)\right] \, dt, \tag{3.13}$$

and $R(t)$ is a time-rescaling function for the coalescent process, given by the expression,

$$R(t) = \int_0^t \frac{N_r}{N(\tau)}\, d\tau. \tag{3.14}$$

In this work, we consider the class of piecewise-exponential population size functions. Any population size function in this family of functions can be described by $M+1$ time points, $0 = t_0 < t_1 < \cdots < t_M = \infty$, where the effective population size in time interval $[t_i, t_{i+1})$, $0 \le i \le M-1$, is given by $N(t) = N(t_i)\exp(-\beta_{i+1}(t-t_i))$. The times t_i are in units of N_r generations. In the piece corresponding to time interval $[t_{i-1}, t_i)$, exponential population growth (decline) is encoded by $\beta_i > 0$ ($\beta_i < 0$), while $\beta_i = 0$ represents a constant population. For this family of population models, we can compute the integrals in (3.13) as follows. For $t \in [t_i, t_{i+1})$, $0 \le i \le M-1$,

$$R(t) - R(t_i) = \mathbf{1}[\beta_{i+1} = 0]\frac{N_r}{N(t_i)}(t - t_i) + \mathbf{1}[\beta_{i+1} \neq 0]\frac{1}{\beta_{i+1}}\frac{N_r}{N(t_i)}(\exp(\beta_{i+1}(t-t_i)) - 1) \tag{3.15}$$

$$\begin{aligned}
c_m &= \sum_{i=1}^{M} \int_{t_{i-1}}^{t_i} \exp\left(-\binom{m}{2} R(t)\right)\, dt \\
&= \sum_{i=1}^{M} \exp\left(-\binom{m}{2} R(t_{i-1})\right) \left\{ \int_{t_{i-1}}^{t_i} \exp\left(-\binom{m}{2}(R(t) - R(t_{i-1}))\right)\, dt \right\} \\
&= \sum_{i=1}^{M} \exp\left(-\binom{m}{2} R(t_{i-1})\right) \left\{ \mathbf{1}[\beta_i = 0] \int_{t_{i-1}}^{t_i} \exp\left(-\binom{m}{2}\frac{N_r}{N(t_{i-1})}(t - t_{i-1})\right)\, dt + \right. \\
&\qquad \left. \mathbf{1}[\beta_i \neq 0] \int_{t_{i-1}}^{t_i} \exp\left(-\binom{m}{2}\frac{1}{\beta_i}\frac{N_r}{N(t_{i-1})}(\exp(\beta_i(t - t_{i-1})) - 1)\right) \right\} \\
c_m &= \sum_{i=1}^{M} \mathbf{1}[\beta_i = 0]\frac{1}{\binom{m}{2}}\frac{N(t_{i-1})}{N_r}\left\{ \exp\left(-\binom{m}{2} R(t_{i-1})\right) - \exp\left(-\binom{m}{2} R(t_i)\right) \right\} \\
&\quad + \sum_{i=1}^{M} \mathbf{1}[\beta_i \neq 0]\frac{1}{\beta_i}\exp\left(-\binom{m}{2} R(t_{i-1})\right) \exp\left(\frac{1}{\beta_i}\binom{m}{2}\frac{N_r}{N(t_{i-1})}\right) \times \\
&\qquad \left\{ \mathrm{Ei}\left(-\binom{m}{2}\frac{N_r}{N(t_{i-1})}\frac{\exp(\beta_i(t_i - t_{i-1}))}{\beta_i}\right) - \mathrm{Ei}\left(-\binom{m}{2}\frac{N_r}{N(t_{i-1})}\frac{1}{\beta_i}\right) \right\},
\end{aligned} \tag{3.16}$$

where $\mathrm{Ei}(x)$ is the exponential integral special function, given by,

$$\mathrm{Ei}(x) = -\int_{-x}^{\infty} \frac{e^{-t}}{t}\, dt. \tag{3.17}$$

Equation (3.16) can be further simplified to get

$$\begin{aligned} c_m = &\sum_{i=1}^{M} \mathbf{1}[\beta_i = 0] \frac{1}{\binom{m}{2}} \frac{N(t_{i-1})}{N_r} \left\{ \exp\left(-\binom{m}{2} R(t_{i-1})\right) - \exp\left(-\binom{m}{2} R(t_i)\right) \right\} + \\ &\sum_{i=1}^{M} \mathbf{1}[\beta_i \neq 0] \frac{1}{\beta_i} \left\{ \exp\left(-\binom{m}{2} R(t_i)\right) \exp\left(\binom{m}{2} \frac{N_r}{N(t_{i-1})} \frac{\exp(\beta_i(t_i - t_{i-1}))}{\beta_i}\right) \times \right. \\ &\qquad \mathrm{Ei}\left(-\binom{m}{2} \frac{N_r}{N(t_{i-1})} \frac{\exp(\beta_i(t_i - t_{i-1}))}{\beta_i}\right) \\ &\qquad \left. - \exp\left(-\binom{m}{2} R(t_{i-1})\right) \exp\left(\binom{m}{2} \frac{N_r}{N(t_{i-1})} \frac{1}{\beta_i}\right) \mathrm{Ei}\left(-\binom{m}{2} \frac{N_r}{N(t_{i-1})} \frac{1}{\beta_i}\right) \right\}, \end{aligned} \tag{3.18}$$

We evaluate terms of the form $\exp(x)\mathrm{Ei}(-x)$ that appear in (3.18) for large values of x using the following asymptotic expansion,

$$\exp(x)\mathrm{Ei}(-x) = -\frac{1}{x} \sum_{k=0}^{\infty} (-1)^k \frac{k!}{x^k}. \tag{3.19}$$

For the results in Section 3.4, we truncated the divergent expansion (3.19) after 10 terms for $x \geq 45$.

3.4 Results

To test the performance of our method, we simulated data under the following demographic scenarios:

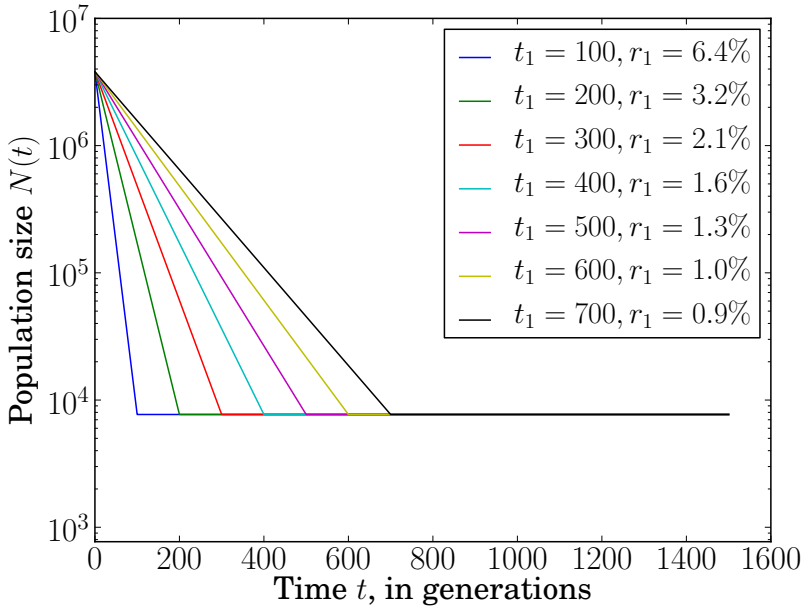
- Scenario 1 – One epoch of exponential growth, with several different parameter values for the per-generation growth rate r_1 and the duration of growth t_1. For all of these choices, the population expansion factor, $(1 + r_1)^{t_1}$, was fixed at approximately 512, which was also roughly equal to the MLE for the population expansion factor in the CEU subpopulation inferred by Nelson *et al.* [62]. The ancestral population size prior to the onset of exponential growth was fixed at 7700 diploids following the estimate of Schaffner *et al.* [77] for the CEU subpopulation.

- Scenario 2 – Two epochs of exponential growth, the more ancient of which lasts for $t_2 = 300$ generations with a growth rate of $r_2 = 1\%$ per generation, and a more recent epoch of rapid growth lasting $t_1 = 100$ generations with a growth rate of $r_1 = 4\%$ per generation. This model also has two population bottlenecks inferred by Keinan *et al.* [40]. The ancestral population size was fixed at 10^4 diploids.

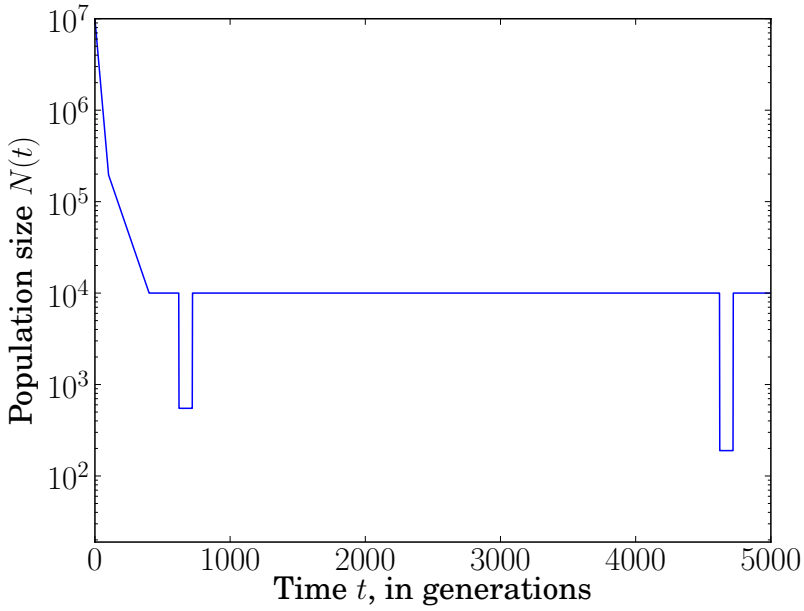
The population size functions, $N(t)$, for both these scenarios are shown in Figure 3.1. For each scenario, we generated 100 simulated datasets using Hudson's `ms` program [36] with a sample size of 10,000 diploid individuals. For Scenario 1, we simulated 200 loci of length 10kb, with the mutation rates at each locus randomly chosen from the range 1.1×10^{-8} to 3.8×10^{-8} per bp per generation, and then held fixed across all the simulated datasets. These are also the range of mutation rates estimated from family trios by Conrad *et al.* [11]. For Scenario 2, we simulated 1000 loci of length 1kb, each with mutation rate 2.5×10^{-8} per bp per generation. For both scenarios, we used a realistic recombination rate of 10^{-8} per bp per generation within each locus.

Figure 3.2 shows the inferred values of the duration and rate of exponential growth for the various values of r_1 and t_1 in Scenario 1, while Figure 3.3 shows the inferred values of the duration and growth rates for each of the two epochs in Scenario 2. The inferred mutation rates for one setting of parameter values for Scenario 1, and for Scenario 2 are shown in Figures 3.4 and 3.5 respectively. As can be seen from Figures 3.2 and 3.4, our method is able to accurately infer the growth parameters and per-locus mutation rate at each of the loci in Scenario 1. While there is more uncertainty in the inference of the growth parameters and mutation rates in Scenario 2, the estimates appear unbiased. With larger sample sizes or more segregating sites, this uncertainty in the parameter estimates is expected to decrease.

For all these simulation results, we combined the counts of those segregating sites which fell in the top 90% of segregating sites ordered by minor allele frequency. To search over the space of demographic models, we used the optimization program IPOPT [88] with gradients computed using the automatic differentiation library ADOL [91]. Using the binning scheme mentioned above, the optimization over the two parameters in Scenario 1 takes about 10 CPU seconds, and that over the four parameters in Scenario 2 takes about a CPU minute.

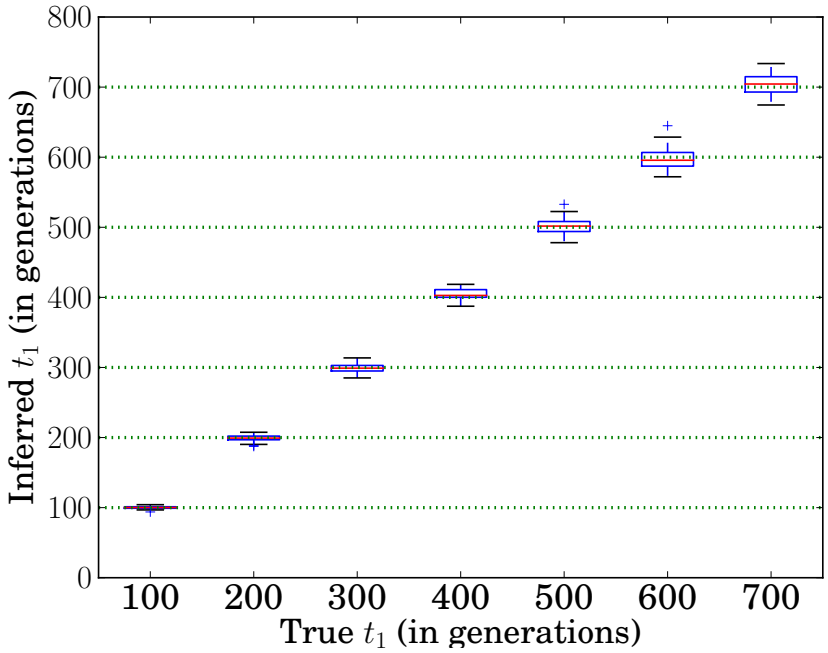


(a)

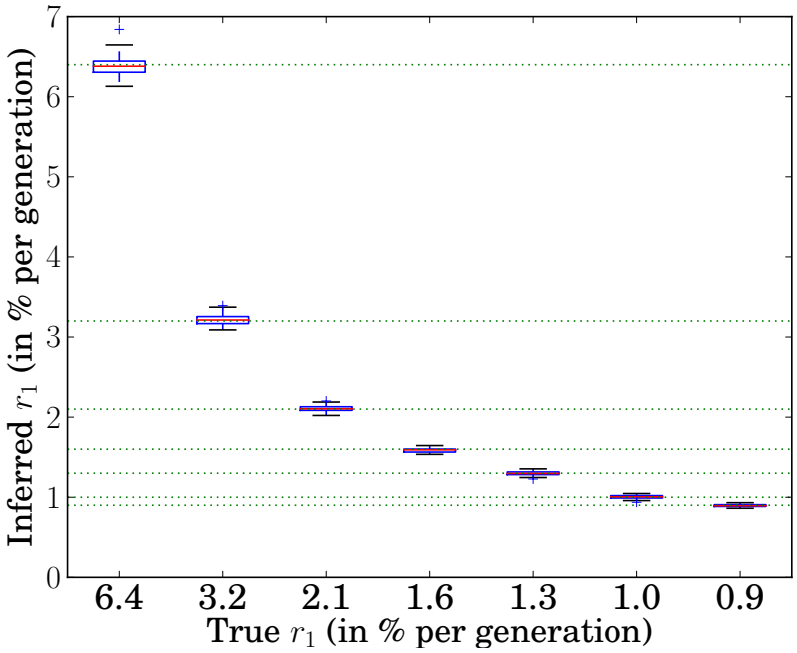


(b)

Figure 3.1: **Population size $N(t)$ as a function of time (number of generations ago) for (a) several choices of t_1 and r_1 in Scenario 1 and (b) Scenario 2.** The present time corresponds to $t = 0$.

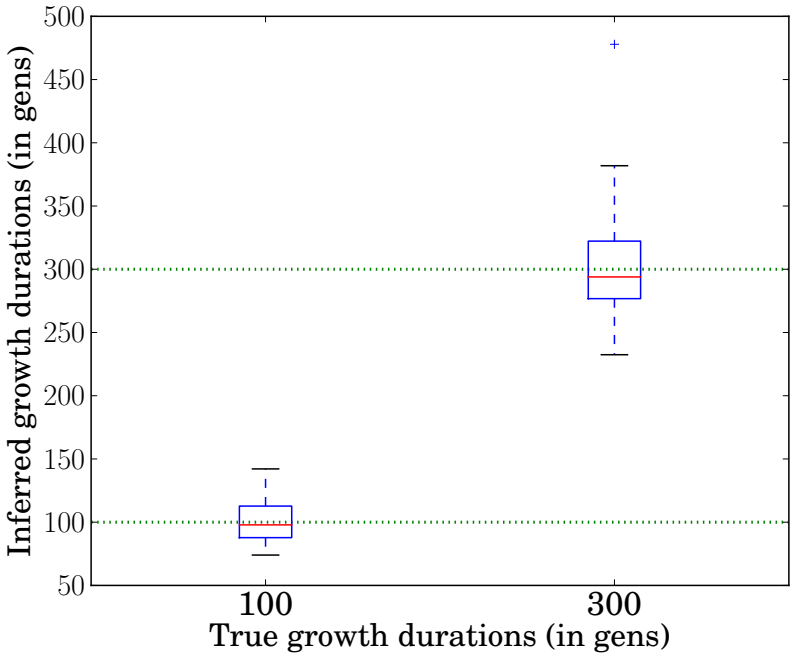


(a)

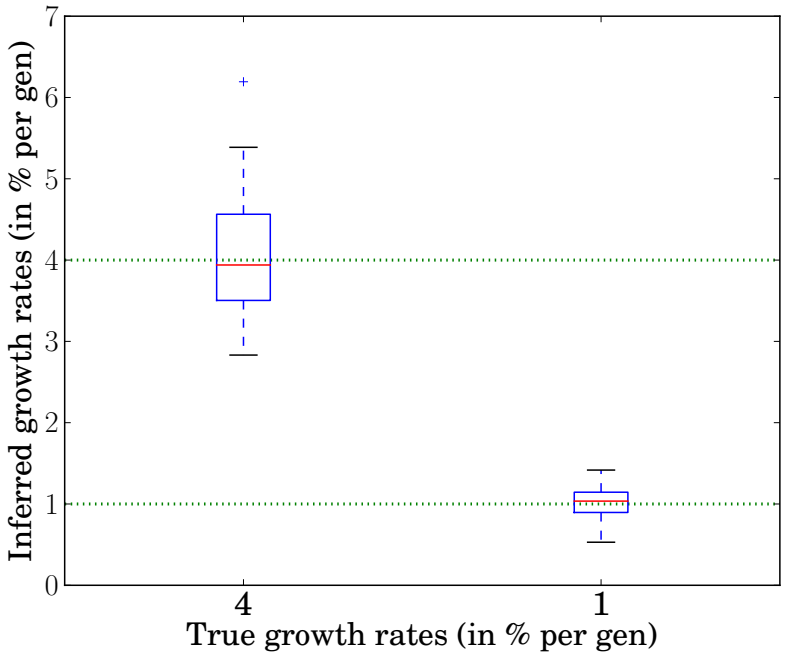


(b)

Figure 3.2: **Box plots of the (a) duration and (b) rate of exponential growth in the population size for different values of t_1 and r_1 in Scenario 1 over 100 simulated datasets.**



(a)



(b)

Figure 3.3: **Box plots of the (a) durations and (b) rates of exponential growth for each of the two epochs in Scenario 2, over 100 simulated datasets.**

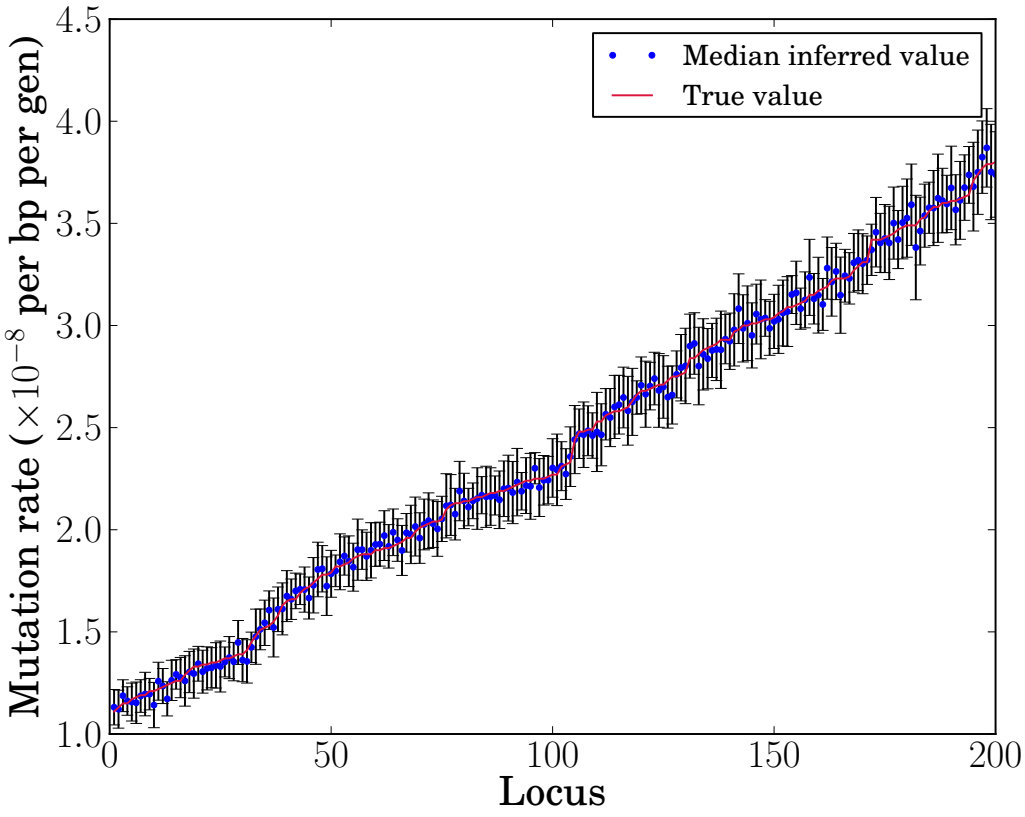


Figure 3.4: **Box plots of the inferred mutation rates at each of 200 loci for Scenario 1 with $t_1 = 100$, and $r_1 = 6.4\%$.** The loci are sorted in ascending order of the true mutation rates. The red line indicates the true mutation rate used in the simulations, while the blue circles and error bars denote the median and 1 standard deviation of the inferred mutation rates over 100 simulated datasets.

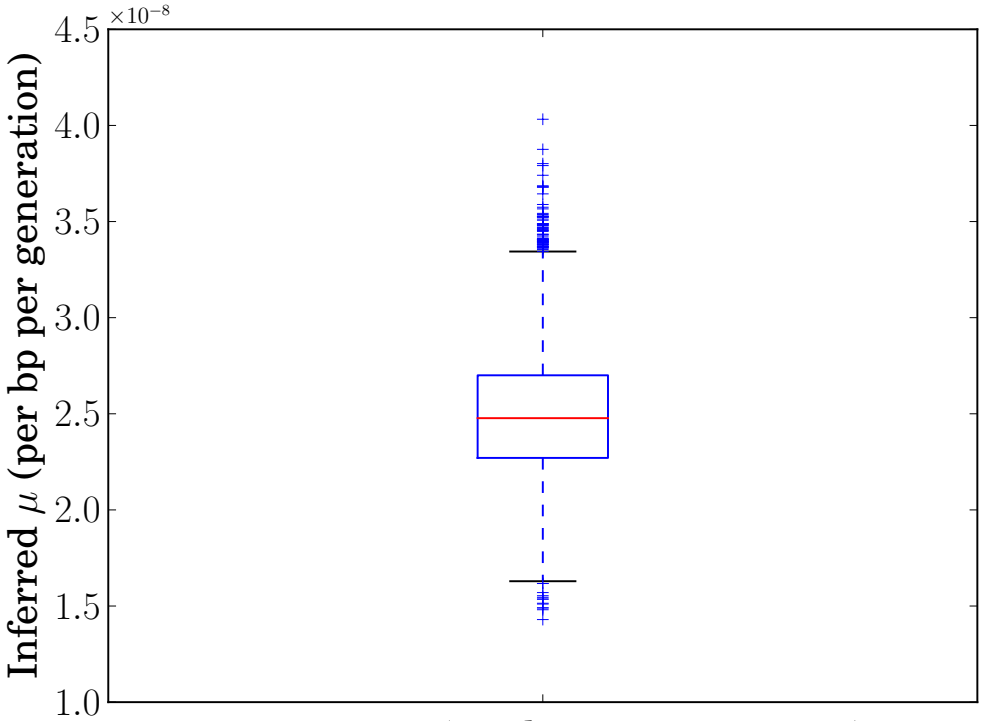


Figure 3.5: **Box plot of the inferred common mutation rate for Scenario 2 with 2 epochs of recent exponential growth.** The true mutation rate used in the simulations was $\mu = 2.5 \times 10^{-8}$ per bp per generation at all loci.

Chapter 4

Distortion of genealogical properties for very large sample sizes

In this chapter, we investigate whether the coalescent continues to be a good approximation to the DTWF model in the case where the sample size increases to the point where the coalescent assumptions may be violated. We compare the two models under certain demographic scenarios previously considered in the literature, including the case of recent rapid population growth for humans [23, 85]. We examine several key genealogical statistics of interest such as the number of multiple and simultaneous mergers in the DTWF model, the number of lineages as a function of time (NLFT), and the sample frequency spectrum. A key feature of our work is that all our results, under both the coalescent and the DTWF model, are based on *exact* deterministic computations rather than Monte-Carlo simulations.

To perform exact computation in the DTWF model, we exploit the Markov property of the model and devise dynamic programming algorithms to compute various genealogical quantities of interest exactly. These algorithms are computationally expensive, so we also consider a hybrid method that uses the DTWF model for the recent past and the coalescent for the more distant past. We demonstrate that this hybrid approach produces substantially more accurate predictions than does the coalescent, while being more efficient than performing computation in the full DTWF model.

4.1 Demographic models

In addition to the case of a constant population size, we consider three models of variable population size. The details of the demographic models we consider are provided below and illustrated in Figure 4.1:

- Model 1: A constant population size of 10,000 diploid individuals [32, 33].
- Model 2: Proposed by Keinan *et al.* [40], this model has two population bottlenecks, the most recent of which lasted for 100 generations starting from 620 generations in

the past, and a more ancient bottleneck lasting 100 generations, starting from 4,620 generations in the past. Further back in time, the population size is fixed at 10,085 diploids.

- MODEL 3: This demographic model was inferred by Gravel *et al.* [23] for the CEU subpopulation from the 1000 Genomes [1] exon pilot data. In this model, a population expansion in the last 920 generations occurs at a rate of 0.38% per generation.

- MODEL 4: This demographic model was inferred by Tennessen *et al.* [85, Figure 2B] for the CEU subpopulation from exome-sequencing of 2440 individuals. The ancient demography is similar to that in MODEL 3. However, following the most recent bottleneck, there are two epochs of exponential expansion in the most recent 920 generations – a slower expansion phase for 716 generations at 0.307% per generation, followed by a rapid expansion rate of 1.95% per generation for 204 generations.

While other models exhibiting recent rapid population expansion have been inferred [19, 62], we focus on MODEL 3 and MODEL 4 in this chapter because we want to consider sample sizes that are on the order of the current effective population size while also considering demographic models that incorporate realistic changes in the population size. Due to computational limitations, the largest sample size for which we can perform exact computation in the DTWF model is on the order of 10^5 haploids, which is of the same order of magnitude as the effective population size in MODEL 3, and about 10% of the current effective population size of MODEL 4.

Using the above four demographic models, we examine deviations in the following quantities between the coalescent and the DTWF model:

(a) Multiple and simultaneous mergers in the DTWF model.

(b) Number of lineages as a function of time (NLFT).

(c) Expected sample frequency spectrum.

The computation of the various genealogical quantities in the DTWF model, such as the number of simultaneous- and multiple-mergers, the NLFT, and the expected frequency spectrum, rely on the Markov property of the DTWF model. By considering the types and counts of the mergers occurring in the previous generation, one can write down one-step recurrence equations relating these genealogical quantities over time and solve these recurrences by dynamic programming. The details of these recurrence equations for the various genealogical quantities considered in this work are provided in Section 4.6, and software programs implementing them can be downloaded from `http://www.eecs.berkeley.edu/~yss/software`.

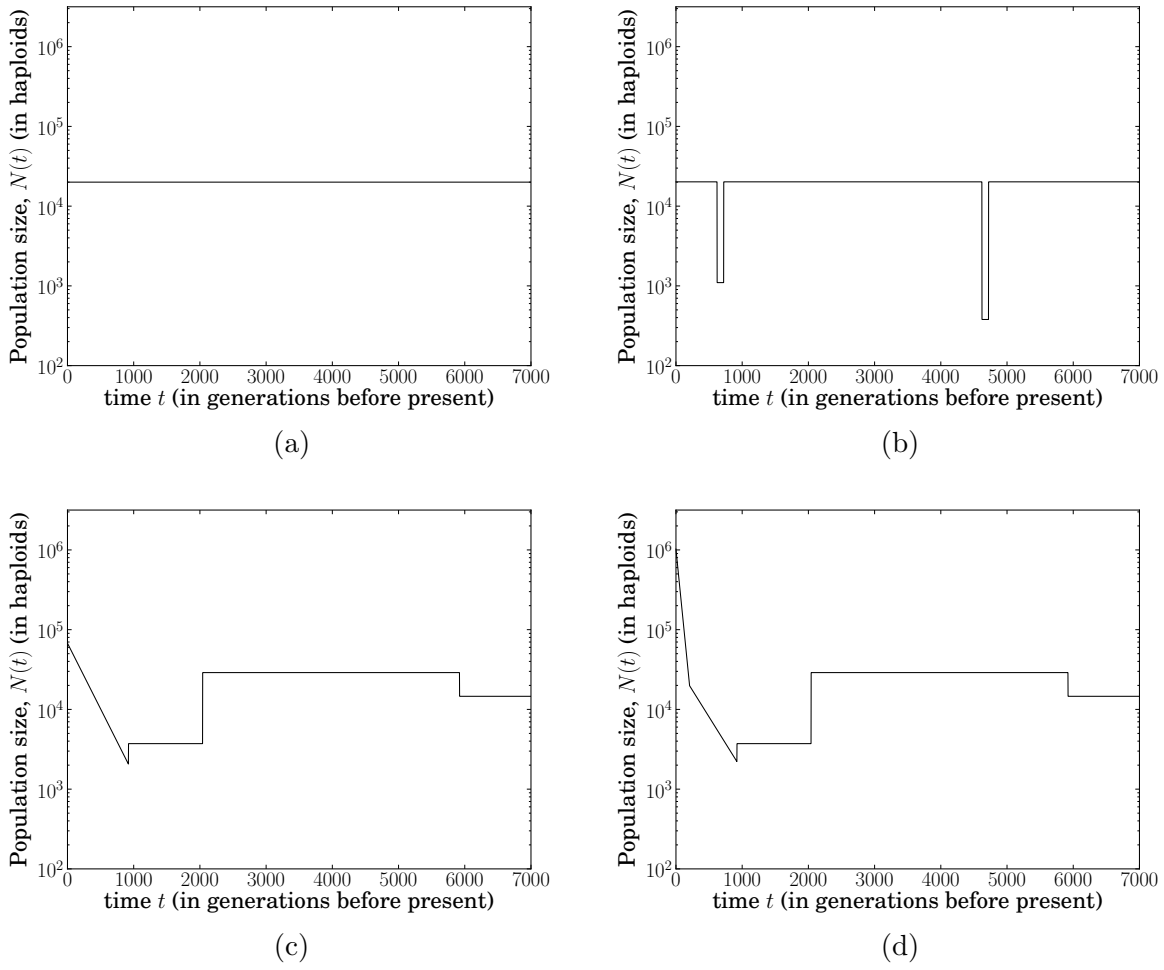


Figure 4.1: **Demographic models examined in this work.** Each graph shows the effective population size $N(t)$ (in haploids) as a function of time (number of generations ago) in a Wright-Fisher model of random mating. The present time corresponds to $t = 0$. The demographic models are labeled as (a) MODEL 1, (b) MODEL 2, (c) MODEL 3, and (d) MODEL 4.

4.2 Multiple and simultaneous mergers in the DTWF model

For a given demographic model in the DTWF framework, it becomes more likely that multiple lineages may be lost in a single generation as the sample size n increases. The first-order approximations used in the derivation of the coalescent from the DTWF model assume that the sample size n is on the order of $\sqrt{N}$, with N being the population size. For example,

Table 4.1: **Expected percentage of lineages (relative to $n-1$, where n is the sample size) lost due to either multiple or simultaneous mergers for Models 1–4.** For large sample sizes ($n = 2 \times 10^3$ or 2×10^4), in all demographic models, most of the lineages are lost in generations where multiple lineages are lost.

n	Model 1	Model 2	Model 3	Model 4
20	0.28%	2.16%	0.84%	0.86%
2×10^2	25.13%	30.86%	19.44%	24.79%
2×10^3	91.16%	92.18%	90.81%	92.53%
2×10^4	99.12%	99.22%	99.14%	99.31%

consider a sample of size $n = 250$ with an effective population size of $N = 20{,}000$ haploids. Figure 4.2 shows the probability distribution of the number of parents of the sample in the previous generation. There is a high probability that the sample will have less than $n - 1$ parents in the previous generation, an event which is ignored in the asymptotic calculation used in the coalescent derivation from the DTWF model. Figure 4.3 shows the expected fraction of lineages (relative to $n - 1$) that are lost due to either multiple or simultaneous mergers, from the present up to time t in the past. Table 4.1 shows the numerical values of the expected fraction as $t \to \infty$. The sharp jump in the plot for Model 2 (Figure 4.3b) corresponds to the beginning (backwards in time) of population bottlenecks when the population size declines substantially, thus instantaneously increasing the rate at which lineages find common ancestors and are lost. For small sample sizes relative to the population size, it is unlikely for more than one lineage to be lost in a single generation, as can be seen in the plots for $n = 20$ and $n = 200$. In contrast, for large sample sizes ($n = 2 \times 10^4$), almost all the lineages are lost in generations when more than one lineage is lost.

When multiple lineages are lost in a single generation of the DTWF model, there are several ways this could happen. For example, suppose 2 out of m lineages are lost in one generation. This could be the result of 3 lineages finding the same parent in the previous generation, or two pairs of lineages each finding a common parent, with the two parents being different. In general, there are $S(m, j)$ different ways that m labeled lineages can have j distinct parents in the previous generation, where $S(m, j)$ is the Stirling number of the second kind, counting the number of ways of partitioning a set of m labeled objects into j non-empty subsets. A particular pattern of mergers of m lineages which leads to j distinct parents, where $\lceil \frac{m}{2} \rceil \leq j \leq m$, is illustrated in Figure 4.4. Here, $m - j$ pairs of lineages each find a common parent distinct from all other parents, and the remaining $2j - m$ lineages do not merge with any other lineages. There are j ancestral lineages left after this type of merger. We call this an $(m - j)$-*pairwise-simultaneous merger*. For $k \geq 2$, we use the term k-*merger* to denote an event where exactly k lineages find the same common parent in the previous generation. It is possible to have several multiple merger events in a single generation. For example, a j-pairwise-simultaneous merger is equivalent to there

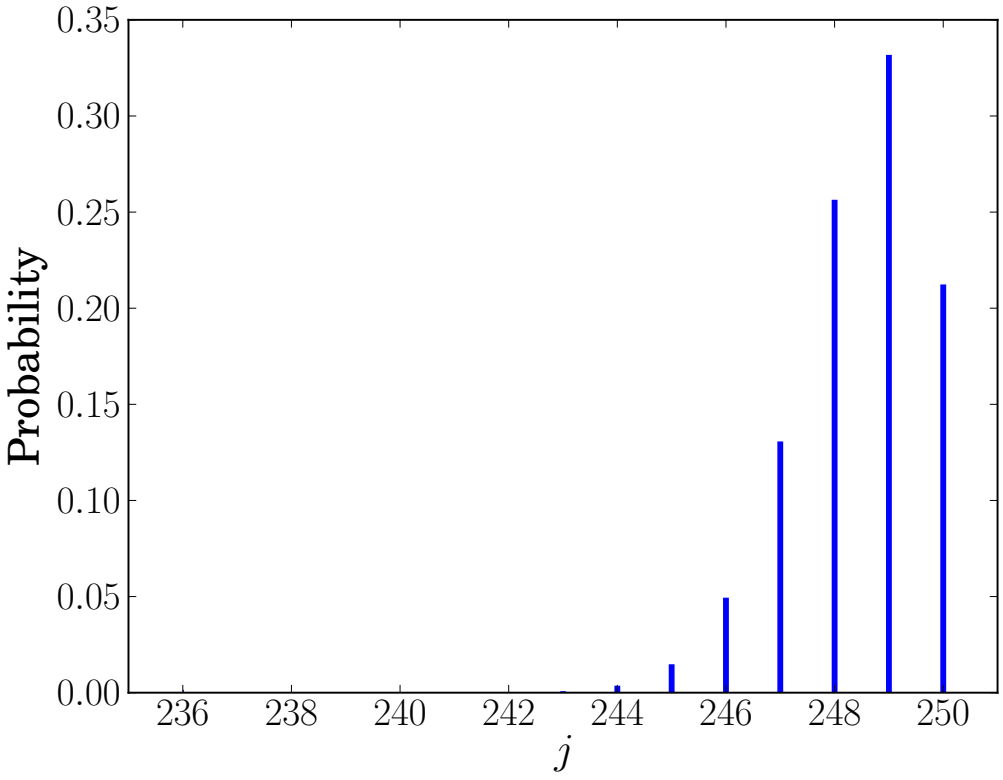


Figure 4.2: **Probability distribution of the number of parents of a sample of size $n = 250$ and population size $N = 20{,}000$ in the Wright-Fisher model.** For each value of j on the x-axis, the y-axis is the probability that the sample of size n has j parents in the previous generation. There is a substantial probability of the sample having less than $n - 1$ parents in the previous generation, which corresponds to multiple or simultaneous mergers in the ancestral process.

being exactly j 2-merger events and no other merger events in a single generation.

In the coalescent, since at most 2 lineages find a common ancestor in any given time, the only kind of possible merger is a single 2-merger (or equivalently, a 1-pairwise-simultaneous merger). On the other hand, in a DTWF model with m lineages at a given time, there are $\frac{1}{2}\binom{m}{2}\binom{m-2}{2}$ possible 2-pairwise-simultaneous mergers, and $\binom{m}{3}$ possible 3-mergers, yielding the following expression for the total number of different ways for m lineages to find $m - 2$ distinct parents in the previous generation:

$$S(m, m-2) = \binom{m}{3} + \frac{1}{2}\binom{m}{2}\binom{m-2}{2}. \tag{4.1}$$

Since the second term is $O(m^4)$ while the first term is $O(m^3)$, for large m we expect 2-pairwise-simultaneous mergers to be the dominant reason for losing 2 lineages in a single generation.

Table 4.2: **Expected percentage of lineages (relative to $n-1$, where n is the sample size) lost due to k-mergers in Models 1–4.**

	Model 1		Model 2	
k	$n = 2 \times 10^3$	$n = 2 \times 10^4$	$n = 2 \times 10^3$	$n = 2 \times 10^4$
2	96.68%	68.70%	96.66%	68.93%
3	3.24%	23.03%	3.26%	22.93%
4	0.08%	6.44%	0.08%	6.36%

	Model 3		Model 4	
k	$n = 2 \times 10^3$	$n = 2 \times 10^4$	$n = 2 \times 10^3$	$n = 2 \times 10^4$
2	98.77%	89.99%	98.96%	98.11%
3	1.22%	9.25%	1.03%	1.87%
4	0.01%	0.72%	0.01%	0.02%

In Model 1 and Model 2 for $n = 2 \times 10^4$, a substantial number of lineages are involved in 3-mergers, and more than 6% of the lineages are involved in 4-mergers, because the sample size is of the same order as the current population size. Even in Model 3 and Model 4 for $n = 2 \times 10^4$, around 9% and 2% of the lineages participate in 3-mergers, respectively.

Figure 4.5 illustrates the ratio of the sum of the expected number of lineages lost due to k-pairwise-simultaneous mergers, for $k \geq 2$, to the results shown in Figure 4.3, the expected number of lineages lost due to multiple or simultaneous mergers, from the present up to time t in the past. As Figure 4.5 shows, a substantial fraction of the lineages that are lost in generations with multiple lost lineages (i.e. in generations with mergers forbidden in the coalescent) are due to pairwise simultaneous mergers. Incidentally, that the curves for $n = 20$ starts out near 0.93 can be attributed to the fact that the ratio of the second term in the right hand side of (4.1) to $S(m, m-2)$ is $\frac{51}{55}$ for $m = 20$.

The expected fraction (relative to $n-1$) of lineages lost due to k-mergers is shown in Table 4.2. A substantial number of lineages are lost to 3-mergers in Model 1, Model 2 and Model 3 for $n = 2 \times 10^4$ because the sample size is of the same order as the population size at time 0. Even in Model 4, about 1.9% of lineages participate in 3-mergers. Figure 4.6 shows the fraction of 3-mergers up to time t relative to the total expected number of 3-mergers as $t \to \infty$. As expected, in Model 1, Model 2, and Model 3, due to the large sample size relative to the population size at time 0, a substantial portion of the 3-mergers take place very early when the number of surviving lineages drops quickly. It is rather surprising that in Model 4, where there is a rapid exponential population growth, a large fraction of the 3-mergers in fact take place during this period of growth. In particular, more than 25% of the expected 3-mergers for $n = 2 \times 10^4$ occur in the most recent 32 generations when the effective population size is at least 5.5×10^5.

Based on the results described above, one would expect that the number of ancestral

lineages remaining at a given time decreases more rapidly under the DTWF model than under the coalescent, and we investigate this quantity next.

4.3 Number of lineages as a function of time

Here, we compare the expected number of lineages as a function of time (NLFT) in the coalescent and in the DTWF model. In what follows, we let $A_n^C(t)$ and $A_n^D(t)$ denote the random variables for the number of lineages at generation t in the coalescent and the DTWF model, respectively, starting with a sample of size n at time 0. Under the coalescent, the expectation and standard deviation of the NLFT, $\mathbb{E}[A_n^C(t)]$ and $\sigma(A_n^C(t))$, can be computed exactly in a numerically-stable fashion for an arbitrary variable population size model as described in Section 4.6.5. An algorithm to compute $\mathbb{E}[A_n^D(t)]$ and $\sigma(A_n^D(t))$ under the DTWF model is also described there.

For the four demographic models considered, Figure 4.7 shows the expectation and standard deviation of the NLFT under the DTWF model, while Figures 4.8 and 4.9 show the relative differences in the expectation and standard deviation, respectively, of the NLFT in the coalescent with respect to the NLFT in the DTWF model. For large sample sizes under Model 1, Model 2, and Model 3, it can be seen that the lineages are lost at a faster rate in the DTWF model than in the coalescent. This pattern is consistent with the fact that these demographic models exhibit a substantial number of 3-mergers in the DTWF model for large sample sizes (see Table 4.2), though the deviation in the expected NLFT is still substantially less than the expected number of 3-mergers. The deviation disappears after about 1000 generations when enough time has passed for the number of ancestral lineages to become sufficiently small that the coalescent approximation holds.

For Model 4, the expected NLFT in the coalescent provides a fairly good approximation to that in the DTWF model for all times and for all sample sizes considered. This is because the population size remains much larger than the number of ancestral lineages at all times.

4.4 Expected sample frequency spectrum

Given a sample of n haploid (or $n/2$ diploid) individuals, a common summary of the sample used in various population genetic analyses is the sample frequency spectrum, $\hat{\boldsymbol{\tau}}_n = (\hat{\tau}_{n,1}, \ldots, \hat{\tau}_{n,n-1})$. Under the infinite-sites model of mutation, the kth entry $\hat{\tau}_{n,k}$ corresponds to the number of polymorphic sites in the sample that have k derived alleles and $n-k$ ancestral alleles, where $1 \leq k \leq n-1$. For a sample of n haploids randomly drawn from the population, we denote the expected value of $\hat{\tau}_{n,k}$ in the coalescent and the DTWF models by $\tau_{n,k}^C$ and $\tau_{n,k}^D$, respectively. In the case of a constant population size, $\tau_{n,k}^C$ under the infinite-sites model of mutation is given exactly by the expression

$$\tau_{n,k}^C = \frac{\theta}{k}, \tag{4.2}$$

where θ denotes a population-scaled mutation rate. (Mutations arise according to a Poisson process with intensity $\theta/2$ in each lineage, independently of all other lineages.) For variable population size models, the results of Polanski and Kimmel [69] can be used to compute the expected sample frequency spectrum numerically-stably under the coalescent. In Section 4.6.4, we develop an algorithm to compute the expected sample frequency spectrum under the DTWF model, denoted by $\boldsymbol{\tau}_n^D = (\tau_{n,1}^D, \ldots, \tau_{n,n-1}^D)$.

Figure 4.10 illustrates the relative difference between the coalescent and the DTWF model in the number of singletons ($\tau_{n,1}$) and doubletons ($\tau_{n,2}$) as a function of the sample size (n). As the figure shows, the number of singletons predicted by the DTWF model is *larger* than the coalescent prediction by as much as 11% when the sample size is comparable to the current population size. It is interesting to note that even though there are a substantial number of 3-mergers and 4-mergers in MODEL 1 and MODEL 2, the deviations in the frequency spectrum are not nearly as large as one might have expected. This is probably because even though the coalescent forbids multiple mergers by construction, successive two mergers can be separated by arbitrarily small amounts of time (as opposed to being separated by at least 1 generation in a discrete model). This allows the coalescent to simulate the effect of multiple mergers without explicitly allowing them, leading to fairly similar frequency spectra as a DTWF model.

The deviations in the singletons and doubletons for MODEL 1 match those computed by Fisher [18] (and tabulated in [90, Table 1]) when the sample size equals the population size and in the limit that the population size tends to infinity. For MODEL 4 (Figure 4.10d), we could not consider sample sizes $> 10^5$ because of computational burden, but the results for MODELS 1–3 suggest that we should expect to observe $\geq 10\%$ deviation when the sample size n is increased to 10^6, the current population size in MODEL 4. The deviation in the number of doubletons is also significant when the sample size is comparable to the current population size; the DTWF prediction for doubletons is *smaller* than the coalescent prediction by about 4.8%.

The findings described above are especially important given that rare variants comprise a large fraction of segregating sites when the sample size is large. In Figure 4.11, we plot the cumulative distribution of the frequency spectrum in the DTWF model for MODELS 1–4. The number of singletons in MODELS 3 and 4 is higher than in MODELS 1 and 2 due to exponential population growth. The rapid population expansion in MODEL 4 results in about 51% of the segregating sites being singletons and over 80% of the segregating sites having less than 5 copies of the derived allele in a sample of size $n = 2 \times 10^4$. Figure 4.12 shows the expected proportion of rare variants (derived allele frequency $\leq 0.01\%$) as a function of the sample size n for MODELS 3 and 4 under the coalescent. It can be seen that as n approaches the current population size, the proportion of rare variants increases substantially. Figure 4.13 shows the expected proportion of segregating sites that are singletons as a function of n for MODELS 1–4 under the coalescent. For small sample sizes (say, $n < 100$), the proportion of singletons in MODEL 3 and 4 (which incorporate rapid recent population expansion) is not much larger than that in MODELS 1 and 2. However, the difference increases considerably as the sample size goes beyond a few hundred individuals, illustrating

the need for large sample sizes to infer recent population expansion from frequency spectrum data.

4.5 A hybrid method for computing the frequency spectrum

As detailed in Section 4.6, computation in the DTWF model is substantially more involved than in the coalescent. In particular, while the runtime of the frequency spectrum computation in the coalescent depends only on the number of piecewise-exponential epochs and not the duration of each epoch, the runtime of our dynamic programming algorithm for the DTWF model actually depends on the number of generations over which the algorithm is run. Since noticeable deviation between the DTWF model and the coalescent arises when the number of ancestral lineages is not negligible compared to the population size, a reasonable tradeoff between accuracy and runtime would be to use the DTWF model for the recent past and the coalescent for the more distant past (when the number of ancestral lineages has decreased sufficiently).

To explore this idea, we implemented a hybrid method for computing the frequency spectrum which, for a specified switching generation t_s, uses the full DTWF model for generations $0 < t \leq t_s$, followed by the coalescent for generations $t > t_s$. In particular, for $t_s = 0$, this algorithm computes the frequency spectrum under the coalescent, while for $t_s = \infty$, it computes the frequency spectrum under the full DTWF model.

As Figure 4.14 illustrates for MODEL 3 in the case the sample size n is equal to the current effective population size N_0, the difference in the frequency spectrum between the full DTWF model and the hybrid algorithm decreases rapidly as t_s increases. With $t_s = 5$ generations, the largest deviation in the number of singletons is less than 1%, which is a substantial reduction from 11% for $t_s = 0$ (Figure 4.10c). Figure 4.15 shows these data in a different way, where the deviations in the frequency spectrum for MODEL 3 between the full DTWF model and the hybrid algorithm are shown as a function of sample size for several values of t_s.

4.6 Computational details

In this section, we present the details of the algorithms used to compute the various quantities such as the expected NLFT, the expected number of k-mergers and the sample frequency spectrum under the DTWF model, and the NLFT under the coalescent.

4.6.1 Notation

Let $N(t)$ be the number of (haploid) individuals at generation t in a DTWF model, where $t = 0$ corresponds to the present and t is increasing going back in the past. Let $p_{n,m}^{(t)}$ denote

the probability that a sample of n labeled individuals at generation t has m distinct ancestors at generation $t+1$. The probabilities $p_{n,m}^{(t)}$ can be computed using the following recursion for $1 \leq m \leq n$,

$$p_{n,m}^{(t)} = \frac{N(t+1)-m+1}{N(t+1)} p_{n-1,m-1}^{(t)} + \frac{m}{N(t+1)} p_{n-1,m}^{(t)}, \tag{4.3}$$

with the convention that $p_{0,0}^{(t)} := 1$, and $p_{n,m}^{(t)} = 0$ for all other values of m and n. This recursion can be derived by treating a sample of size n as a sample of size $n-1$ and an additional individual. The first term of (4.3) is the probability that the parent of this individual is distinct from the $m-1$ parents of the $n-1$ individuals in the sample, and the second term of (4.3) is the probability that the parent of this individual is one of the m parents of the $n-1$ individuals in the sample.

For an algorithmic reason that will become clear presently, we assume that there is a critical generation t_c such that $N(t) = N$ (some constant) for all $t > t_c$. This assumption is not so restrictive since for sufficiently large t, there will be only 1 lineage left with high probability, and the genealogical properties we study will not be affected. For $t > t_c$, we drop the dependence on t in the probabilities $p_{n,m}^{(t)}$, and simply write them as $p_{n,m}$.

4.6.2 Expected number of lineages as a function of time under the DTWF model

By conditioning on the number of ancestral lineages in the previous generation, it is easy to write a recurrence relation for the distribution of the number of lineages at generation t given that there are n lineages at time 0,

$$\mathbb{P}[A_n^D(t) = m] = \begin{cases} \sum_{k=m}^{n} p_{k,m}^{(t-1)} \mathbb{P}[A_n^D(t-1) = k], & \text{if } t > 0, \\ \delta_{n,m}, & \text{if } t = 0, \end{cases} \tag{4.4}$$

where $\delta_{n,m} = 1$ if $n = m$, and $\delta_{n,m} = 0$ otherwise. One can then compute the expectation $\mathbb{E}[A_n^D(t)]$ using (4.4).

4.6.3 Expected number of multiple- and simultaneous-mergers in the DTWF model

Let $M_{n,k}^{(t)}$ be the random variable denoting the number of k-mergers that occur in a genealogical tree starting with a sample of size n observed at generation t. The expected number of k-mergers in a sample of size n observed at present, $\mathbb{E}[M_{n,k}^{(0)}]$, can be computed by conditioning on the mergers that occur between generation t and $t+1$. We then have the following recurrence relations for $\mathbb{E}[M_{n,k}^{(t)}]$,

$$\mathbb{E}[M_{n,k}^{(t)}] = \begin{cases} \binom{n}{k} \sum_{m=k+1}^{n} p_{k,1}^{(t)} p_{n-k,m-k}^{(t)} \frac{N(t+1)-m+k}{N(t+1)} + \sum_{m=1}^{n} p_{n,m}^{(t)} \mathbb{E}[M_{m,k}^{(t+1)}], & \text{if } k < n, \\ p_{k,1}^{(t)} + p_{k,k}^{(t)} \mathbb{E}[M_{k,k}^{(t+1)}], & \text{if } k = n. \end{cases} \tag{4.5}$$

If the population size is constant, or for $t > t_c$ when the population size remains fixed in the past, we can drop the dependence on t in the notation $\mathbb{E}[M_{n,k}^{(t)}]$, and derive the following recurrence relations and boundary conditions for $\mathbb{E}[M_{n,k}]$:

$$\mathbb{E}[M_{n,k}] = \begin{cases} \binom{n}{k} \sum_{m=k+1}^{n} \frac{p_{k,1} p_{n-k,m-k}}{1-p_{n,n}} \frac{N-m+k}{N} + \sum_{m=1}^{n-1} \frac{p_{n,m}}{1-p_{n,n}} \mathbb{E}[M_{m,k}], & \text{if } k < n, \\ \frac{p_{k,1}}{1-p_{k,k}}, & \text{if } k = n. \end{cases} \tag{4.6}$$

One can write similar recurrence relations for the expected number of k-simultaneous mergers by conditioning on the mergers that occur during each generation of reproduction.

4.6.4 Expected sample frequency spectrum under the DTWF model

The expected frequency spectrum of a randomly drawn sample of n individuals is denoted $\boldsymbol{\tau}_n = (\tau_{n,1}, \ldots, \tau_{n,n-1})$, where $\tau_{n,k}$ corresponds to the number of polymorphic sites in the sample having k derived alleles and $n-k$ ancestral alleles for $1 \leq k < n$ under an infinite-sites model of mutation. For a given sample of individuals observed at present, the ancestral process in the DTWF model generates a genealogical tree, where the root of the tree is the most recent common ancestor (MRCA) of the sample, with the individuals in the sample forming the leaves of the tree. For the DTWF model, we can use dynamic programming to efficiently calculate $\tau_{n,k}$ as follows. Let $\gamma_{a,b}^{(t)}$ be a random variable denoting the total branch length (in number of generations) of a subtree that subtends a particular set of a labeled individuals in a larger set of $a+b$ individuals observed at time t. Then by the exchangeability of the individuals in the sample, the definition of γ, and linearity of expectation, we have,

$$\tau_{n,k} = \frac{\theta}{2} \binom{n}{k} \mathbb{E}[\gamma_{k,n-k}^{(0)}], \tag{4.7}$$

since there are $\binom{n}{k}$ subsamples of k individuals out of the n individuals in the original sample. By conditioning on the mergers between lineages that take place between generations t and $t+1$, we get the following recurrence relations for $\mathbb{E}[\gamma_{a,b}^{(t)}]$,

$$\mathbb{E}[\gamma_{a,b}^{(t)}] = \begin{cases} \sum_{j=1}^{a} \sum_{k=1}^{b} p_{a,j}^{(t)} p_{b,k}^{(t)} \frac{(N(t+1)-j)_{k\downarrow}}{(N(t+1))_{k\downarrow}} \mathbb{E}[\gamma_{j,k}^{(t+1)}] & \text{if } a > 1, \\ 1 + \sum_{m=1}^{b} \frac{N(t+1)-m}{N(t+1)} p_{b,m}^{(t)} \mathbb{E}[\gamma_{1,m}^{(t+1)}] & \text{if } a = 1 \text{ and } b > 1, \\ 1 + p_{2,2}^{(t)} \mathbb{E}[\gamma_{1,1}^{(t+1)}] & \text{if } a = b = 1. \end{cases} \tag{4.8}$$

If the population size is constant, or for $t > t_c$ when the population size remains fixed in the past, we can drop the dependence on t in the notation $\gamma_{a,b}^{(t)}$, and by conditioning on the previous genealogical event, we can derive the following recurrence relations and boundary conditions for $\mathbb{E}[\gamma_{a,b}]$,

$$\mathbb{E}[\gamma_{a,b}] = \begin{cases} \sum_{j=1}^{a} \sum_{k=1}^{b} (1 - \delta_{j,a}\delta_{k,b}) \frac{p_{a,j} p_{b,k}}{1-p_{a+b,a+b}} \frac{(N-j)_{k\downarrow}}{(N)_{k\downarrow}} \mathbb{E}[\gamma_{j,k}] & \text{if } a > 1, \\ \frac{1}{1-p_{b+1,b+1}} + \sum_{m=1}^{b-1} \frac{N-m}{N} \frac{p_{b,m}}{1-p_{b+1,b+1}} \mathbb{E}[\gamma_{1,m}] & \text{if } a = 1 \text{ and } b > 1, \\ N & \text{if } a = b = 1. \end{cases} \tag{4.9}$$

From recurrence relations (4.8) and (4.9), the expected frequency spectrum for a sample of size n can be computed in $O(n^4)$ and $O(n^4 t_c)$ time for the constant and variable population cases respectively. However, if one truncates the summation range for the indices j and k in (4.8) and (4.9) to only those j, k values where $p^{(t)}_{a,j}$ and $p^{(t)}_{b,k}$ (respectively, $p_{a,j}$ and $p_{b,k}$) are greater than some small tolerance parameter $\varepsilon > 0$, the time complexity of the above dynamic programs can be improved to $\widetilde{O}(n^2)$ and $\widetilde{O}(n^2 t_c)$, where the $\widetilde{O}$ notation signifies the dependence of the quantities on the truncation parameter ε.

We used (4.7) along with a truncation parameter of $\varepsilon = 10^{-120}$ to compute the expected frequency spectra values presented in the Results. Upon decreasing this threshold further, we did not observe any change to the frequency spectra, suggesting that the computed answers are accurate.

4.6.5 Expected number of lineages as a function of time under the coalescent

Suppose we have a panmictic population with size $N(t)$ at time t, evolving according to Kingman's coalescent. If we sample n lineages at time 0 and let $A^C_n(t)$ denote the number of ancestral lineages of this sample surviving at time t, then we have the following expression for the probability distribution function of $A^C_n(t)$ [83, 84],

$$\mathbb{P}[A^C_n(t) = m] = \sum_{i=m}^{n} e^{-\binom{i}{2}R(t)}(-1)^{i-m}\frac{(2i-1)(m)_{(i-1)\uparrow}(n)_{i\downarrow}}{m!(i-m)!(n)_{i\uparrow}}, \tag{4.10}$$

where

$$R(t) = \int_0^t \frac{N(0)}{N(\tau)}d\tau. \tag{4.11}$$

The summation in (4.10) has terms with alternating signs, and this leads to a loss of numerical precision due to catastrophic cancellation for even moderate sample sizes [56]. Hence, computing the expectation of the NLFT, $\mathbb{E}[A_n(t)]$, by naively using (4.10) will not produce reliable answers. However, using a formula of Tavaré [84, equation 5.11], one gets the following closed-form expressions for the expectation and variance of $A_n(t)$ that are numerically stable to evaluate,

$$\mathbb{E}[A^C_n(t)] = \sum_{i=1}^{n} e^{-\binom{i}{2}R(t)}(2i-1)\frac{(n)_{i\downarrow}}{(n)_{i\uparrow}}, \tag{4.12}$$

$$\mathrm{Var}\left(A^C_n(t)\right) = \sum_{i=1}^{n} e^{-\binom{i}{2}R(t)}(2i-1)(i^2-i+1)\frac{(n)_{i\downarrow}}{(n)_{i\uparrow}} - \left[\sum_{i=1}^{n} e^{-\binom{i}{2}R(t)}(2i-1)\frac{(n)_{i\downarrow}}{(n)_{i\uparrow}}\right]^2. \tag{4.13}$$

Each term in the summations in (4.12) and (4.13) is positive, and hence poses no numerical problems for evaluation. Furthermore, the terms in the sum decay rapidly for large i due to the exponential functions involved. We give elementary combinatorial proofs of (4.12) and (4.13) in Section A.2 of Appendix A.

4.7 Dicussion

Several analyses of genomic sequence variation in large samples of humans [19,39,62,85] have found a substantial excess of rare variation compared to those predicted using previously applied demographic models. The inference in these studies is that these results are consistent with a rapid growth of the effective population size in the recent past (much more rapid than in previously applied demographic models), a conclusion consistent with historical records of census population size [39]. These studies also employed sample sizes that would appear to be large enough to violate assumptions of the coalescent, potentially distorting genealogical properties in a way that may inflate rare variation relative to the predictions of coalescent theory. In this chapter, we have investigated this issue by developing a method for performing exact computation in the discrete-time Wright-Fisher model of random mating. We have studied the deviation between the coalescent and the Wright-Fisher model for several key genealogical quantities that are used for population genomic inference.

For several recently inferred demographic scenarios for humans, our results show that there are a significant number of multiple- and simultaneous-merger events under the Wright-Fisher model that are ignored by construction of the coalescent. Furthermore, there are noticeable differences in the expected number of rare variants between the coalescent and the DTWF model, especially in the regime where the sample size is on the order of the current effective population size. Even if the demographic models considered here might underestimate the true current effective population size of humans, sample sizes in population genetic studies are rapidly increasing and might grow to be large enough to cause the differences between the DTWF and the coalescent to become amplified.

A number of demographic inference methods are based on fitting the expected frequency spectra under the coalescent [12, 62, 85] or the diffusion process [23, 31, 51, 52] to observed data. For instance, the exponential growth parameters in MODELS 3 [23] and 4 [85] were inferred using a likelihood method based on the diffusion process approximation to the DTWF model, by fitting the predicted frequency spectrum to polymorphism patterns observed in a sample size of 876 individuals and 2,440 individuals, respectively. Since the diffusion process approximation to the DTWF model is equivalent to the coalescent approximation, the differences in the frequency spectrum (see Figure 4.10) between the coalescent and the DTWF model indicate that we might infer different demographies if the analysis were done using the DTWF model. In particular, for a sample of size n analyzed under the DTWF model, any inferred demography will have a current effective population size of at least n. However, the coalescent imposes no such restriction on the inferred current effective population size. In fact, under the coalescent, it is even possible to estimate a current effective population size N_0 that is smaller than the sample size n. This is because one can only infer a scaling function of time in the coalescent, which is the ratio of the variable effective population size to a fixed reference population size. The inferred scaling function can then be transformed into an effective population size function by using the empirically estimated per-generation mutation rate [23], or by setting the reference population size to a specific value [12,62] (e.g., using an ancestral population size inferred by previous studies [77]).

To balance the tradeoff between accuracy and computational efficiency, we have proposed a hybrid algorithm that uses the DTWF model for the recent past and the coalescent for the more distant past. This hybrid algorithm provides a way to obtain more accurate predictions of the frequency spectrum than in the coalescent, while being computationally more efficient than considering the full DTWF model. We leave the exploration of this method for demographic inference as future research.

Wakeley and Takahashi [90] have provided asymptotically accurate expressions (as the effective population size $N \to \infty$) for the number of singletons and the number of segregating sites under a variant of the DTWF model which allows for a larger number of offspring than the effective population size, assuming that the effective population size stays constant over time. Fu [21] has also examined the accuracy of the standard coalescent model and proposed an alternative continuous-time "exact" coalescent model applicable in the regime when $N(N-1)\cdots(N-n+1)\times N^{-n} \gg 0$, where N denotes the effective population size and n the sample size. That work was restricted to the case of a constant population size, while in this chapter we have considered several demographic scenarios inferred from recent large-scale population genomic studies. Moreover, for some of the demographic scenarios and sample sizes considered here, the assumption in Fu's work [21] that $N(N-1)\cdots(N-n+1)\times N^{-n} \gg 0$ is violated. Wakeley *et al.* [89] have shown that it is difficult to reject the coalescent even for data generated using *fixed* pedigrees with random genetic assortment. Our work is complementary to that study and compares the coalescent to the DTWF random mating model.

In this chapter, we have focused on the DTWF model for simplicity. However, it is known that under some weak conditions on the limiting probabilities of a 2-merger and a 3-merger, a large family of exchangeable random mating models converge to the same coalescent limit as the unit of time is rescaled appropriately and the population size gets large [58, 59]. The rate of convergence to the coalescent differs between different random mating models [4], and hence the accuracy of coalescent predictions for large sample sizes depends on the random mating model being considered. The deviation from the coalescent could be amplified for other random-mating models. It would be interesting to consider the accuracy of the coalescent for other random and realistic models of relevance to human genetics; e.g., models in which generations overlap and the distribution of offspring number more closely reflects the observed pattern for human populations (for example, the Swedish family data of Low and Clarke [50] or the Saguenay-Lac-Saint-Jean population considered by Moreau *et al.* [60]). Despite having access to large samples, recent studies [19,62,85] have inferred much smaller current effective population sizes (on the order of millions) than the current census size (on the order of billions) of the population from which the samples were drawn. It is possible that demographic inference methods that explicitly model realistic human mating patterns might be able to infer census population size histories more accurately than does the coalescent, which assumes random mating and can only infer effective population sizes that do not have a direct census interpretation.

Furthermore, it would be interesting to compare discrete-time random models and the coalescent with respect to haplotype sharing (identity-by-descent and identity-by-state), link-

age disequilibrium, and natural selection when the sample size is very large. For example, Davies *et al.* [13] employed simulations to demonstrate that for a constant population size model, recombination and gene conversion can increase the number of ancestral lineages of a sample of chromosomes to the extent that multiple and simultaneous mergers in the DTWF model can lead to substantial differences from the coalescent model in the rates of coalescence and in the number of sequences carrying ancestral material. It would be interesting to perform such comparisons for more realistic demographic models for humans.

We will soon enter an era where it will become routine to analyze samples with hundreds of thousands if not millions of individuals. For these large sample sizes, the standard coalescent will no longer serve as an adequate model for evolution. The DTWF model is mathematically cumbersome to work with, which was one of the original motivations for adopting the coalescent for modern population genetics analyses. However, for these very large sample sizes, we will need to develop new mathematically and computationally tractable stochastic processes that better approximate realistic models of human population evolution, and under which we can efficiently compute genealogical quantities like we have been able to under the coalescent.

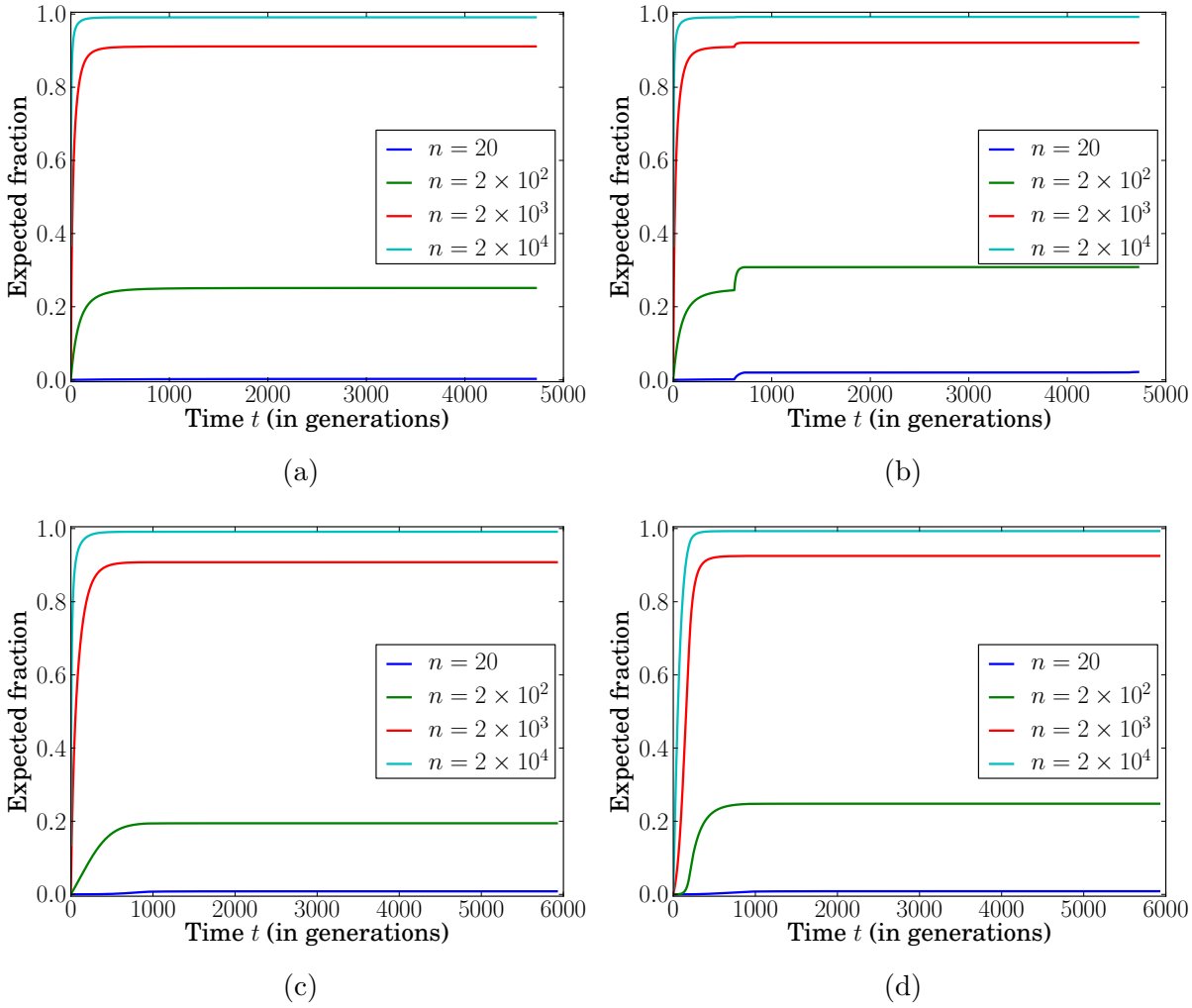


Figure 4.3: **Expected proportion of lineages (relative to $n-1$, where n is the sample size at present) that are lost in generations when more than one lineage is lost, from the present up to time t in the past.** For each generation t on the x-axis, the y-axis is the expected number of lineages that have been lost due to coalescence events in those generations from 0 to t where more than one lineage is lost, normalized by the quantity $n-1$, which is the total number of lineages that are lost over all generations. The plots correspond to (a) Model 1, (b) Model 2, (c) Model 3, and (d) Model 4. The sharp jumps in the plot for Model 2((b)) corresponds to the beginning (backwards in time) of population bottlenecks when the population size declines substantially, thus instantaneously increasing the rate at which lineages find common ancestors and are lost. For small sample sizes relative to the population size ($n = 20$ and $n = 200$), it is unlikely for more than one lineage to be lost in a single generation. In contrast, for large sample sizes ($n = 2 \times 10^4$), almost all the lineages are lost in generations when more than one lineage is lost.

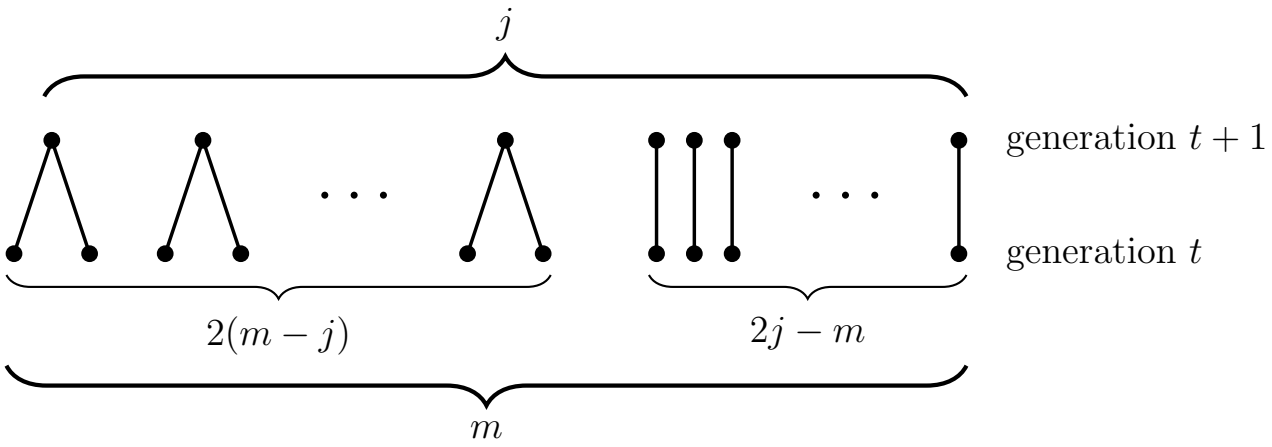


Figure 4.4: **An example generation of an $(m-j)$-pairwise-simultaneous merger during which m lineages in generation t find j parents in generation $t+1$ (backwards in time).** Each of $m-j$ pairs of lineages in generation t finds a different common parent in generation $t+1$, while the remaining $2j-m$ lineages of generation t each have a different parent in generation $t+1$.

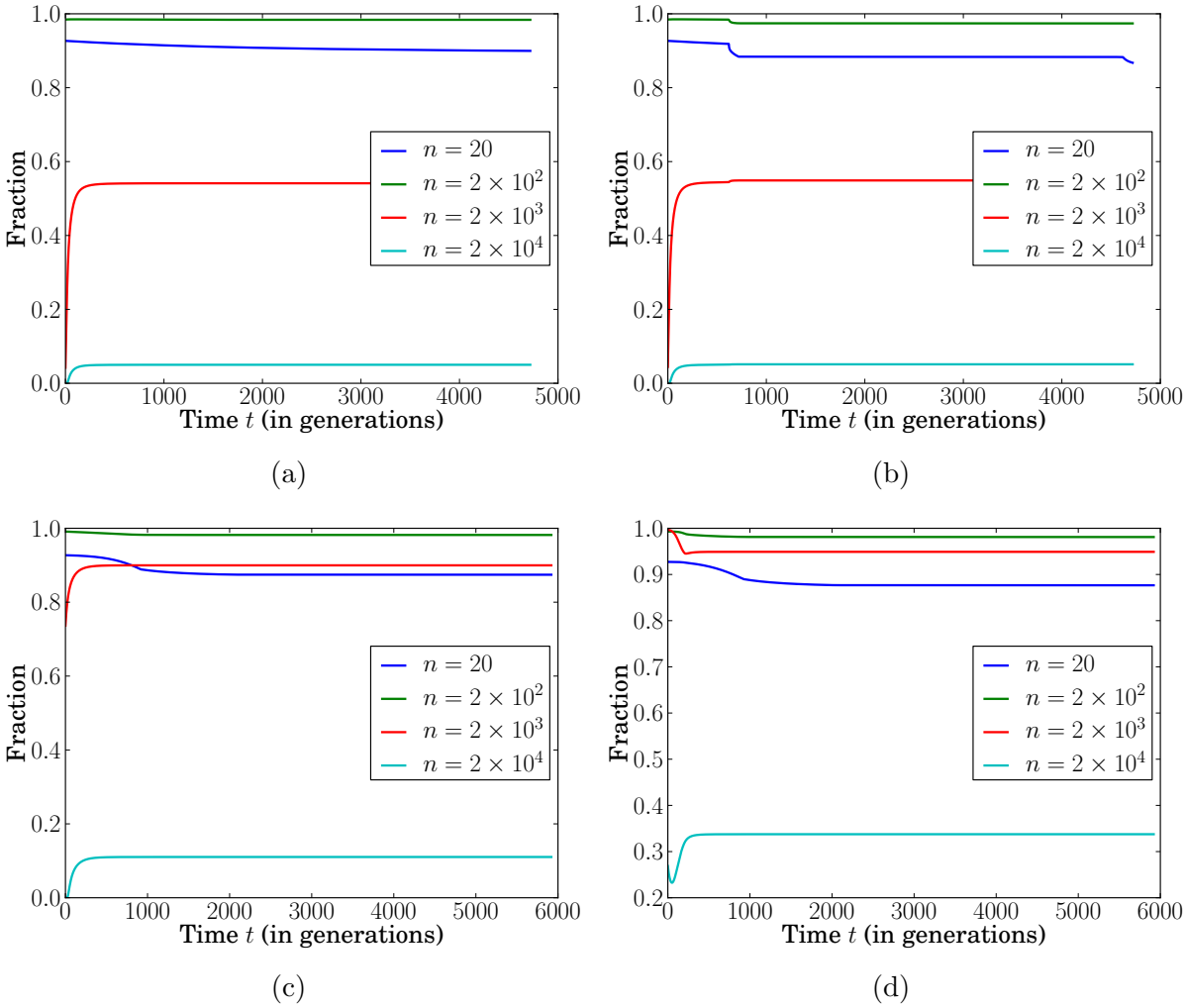


Figure 4.5: **Ratio of the sum of the expected number of lineages lost due to k-pairwise-simultaneous mergers up to time t, for $k \geq 2$, to the respective results shown in Figure 4.3.** This quantity measures the proportion of lineages lost in generations where more than one lineage is lost due to pairwise simultaneous merger events (and no multiple merger events). The plots correspond to (a) Model 1, (b) Model 2, (c) Model 3, and (d) Model 4. A substantial fraction of the lineages that are lost in generations with multiple lost lineages are due to pairwise simultaneous mergers.

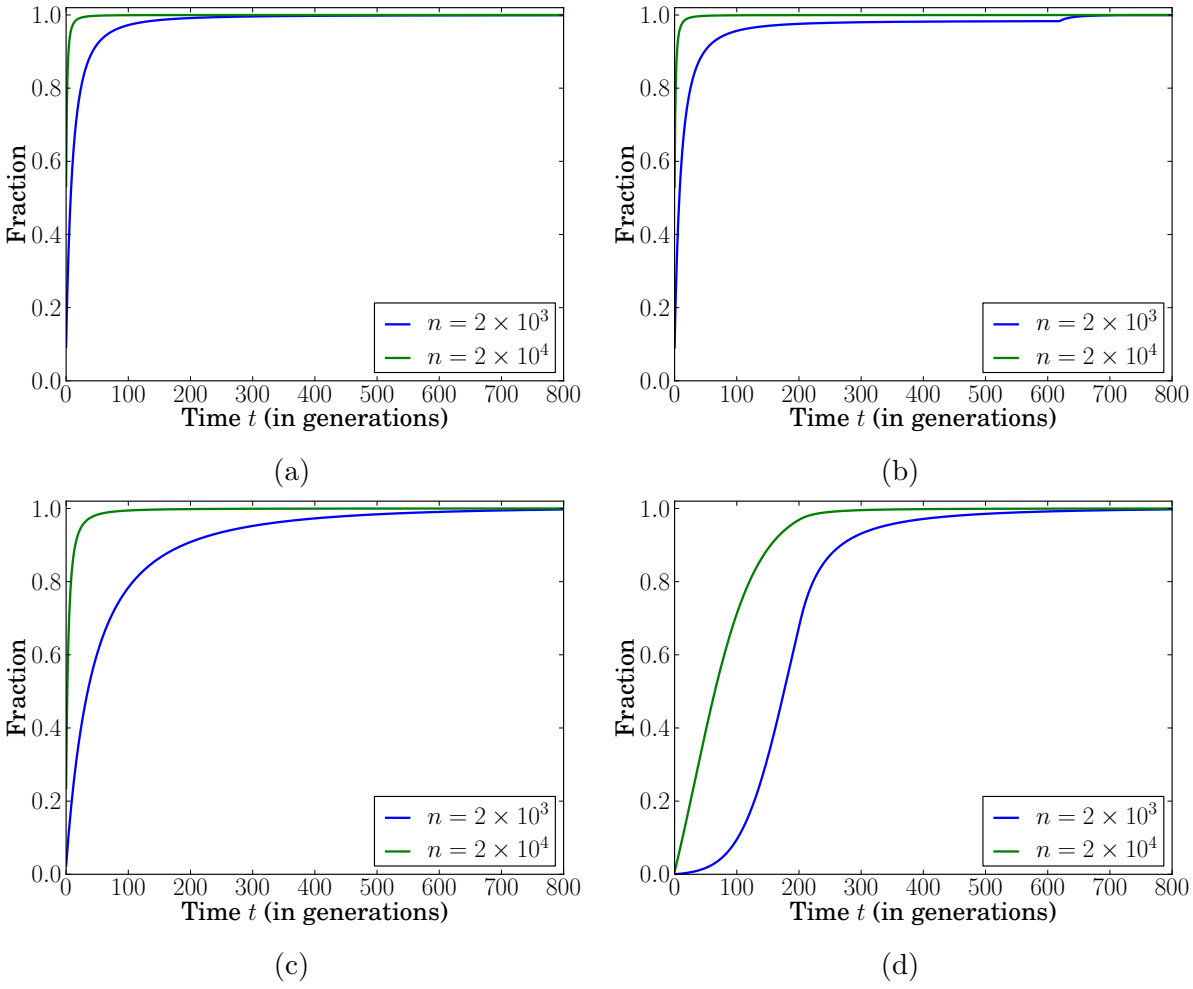


Figure 4.6: **Ratio of the expected number of 3-mergers until time t to the total expected number of 3-mergers.** The plots correspond to (a) Model 1, (b) Model 2, (c) Model 3, and (d) Model 4. In Model 1 and Model 2, due to the large sample size relative to the population size at time 0, a substantial portion of the 3-mergers take place very early when the number of surviving lineages drops quickly. Even in Model 4 where there is a rapid exponential population growth in the most recent 204 generations, more than 25% of the expected 3-mergers for $n = 2 \times 10^4$ occur in the most recent 32 generations when the effective population size is at least 5.5×10^5 haploids.

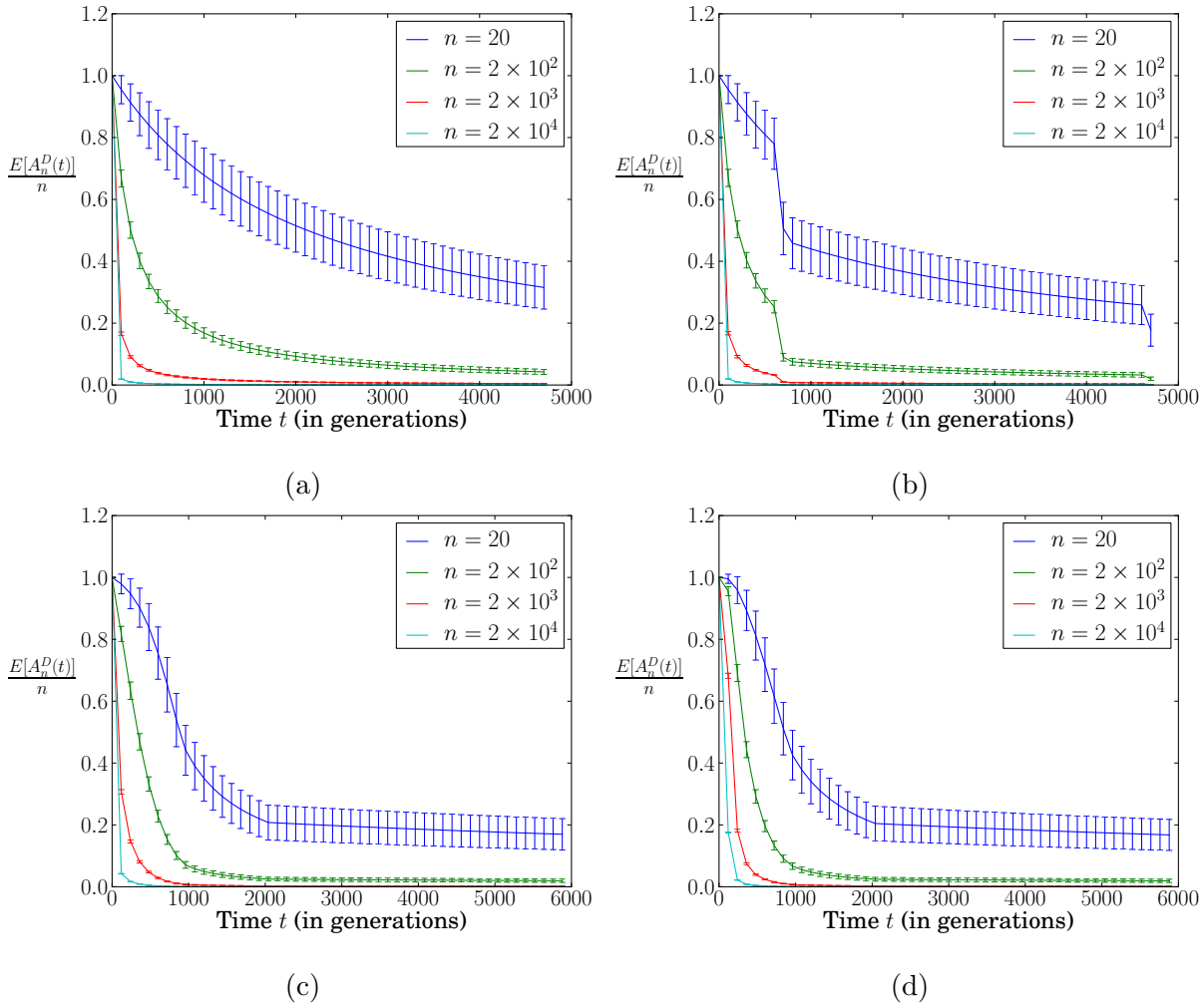


Figure 4.7: **The expectation (solid line) and standard deviation (vertical whiskers) of the NLFT under the DTWF model.** (a) Model 1. (b) Model 2. (c) Model 3. (d) Model 4.

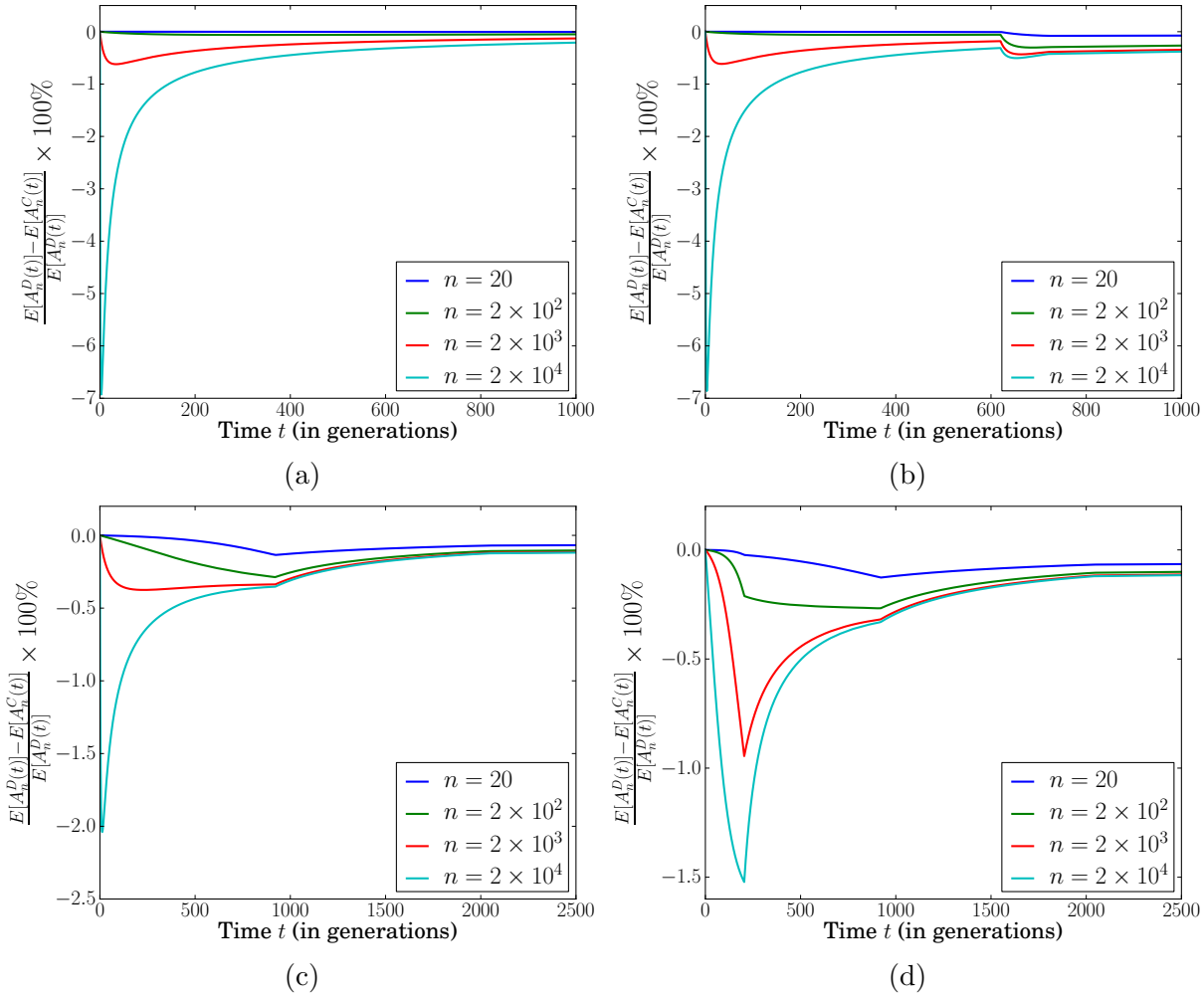


Figure 4.8: **The percentage difference in the expected NLFT between the coalescent and DTWF models, for a haploid sample of size $n \in \{20, 2\times10^2, 2\times10^3, 2\times10^4\}$.** The plots correspond to (a) MODEL 1, (b) MODEL 2, (c) MODEL 3, and (d) MODEL 4. For all demographic models, lineages are lost at a faster rate in the DTWF model than in the coalescent, consistent with the fact that there are a substantial number of 3-mergers in the DTWF model for large sample sizes. This deviation is more pronounced for larger sample sizes and for MODELS 1–3 where the sample size is comparable to the current population size.

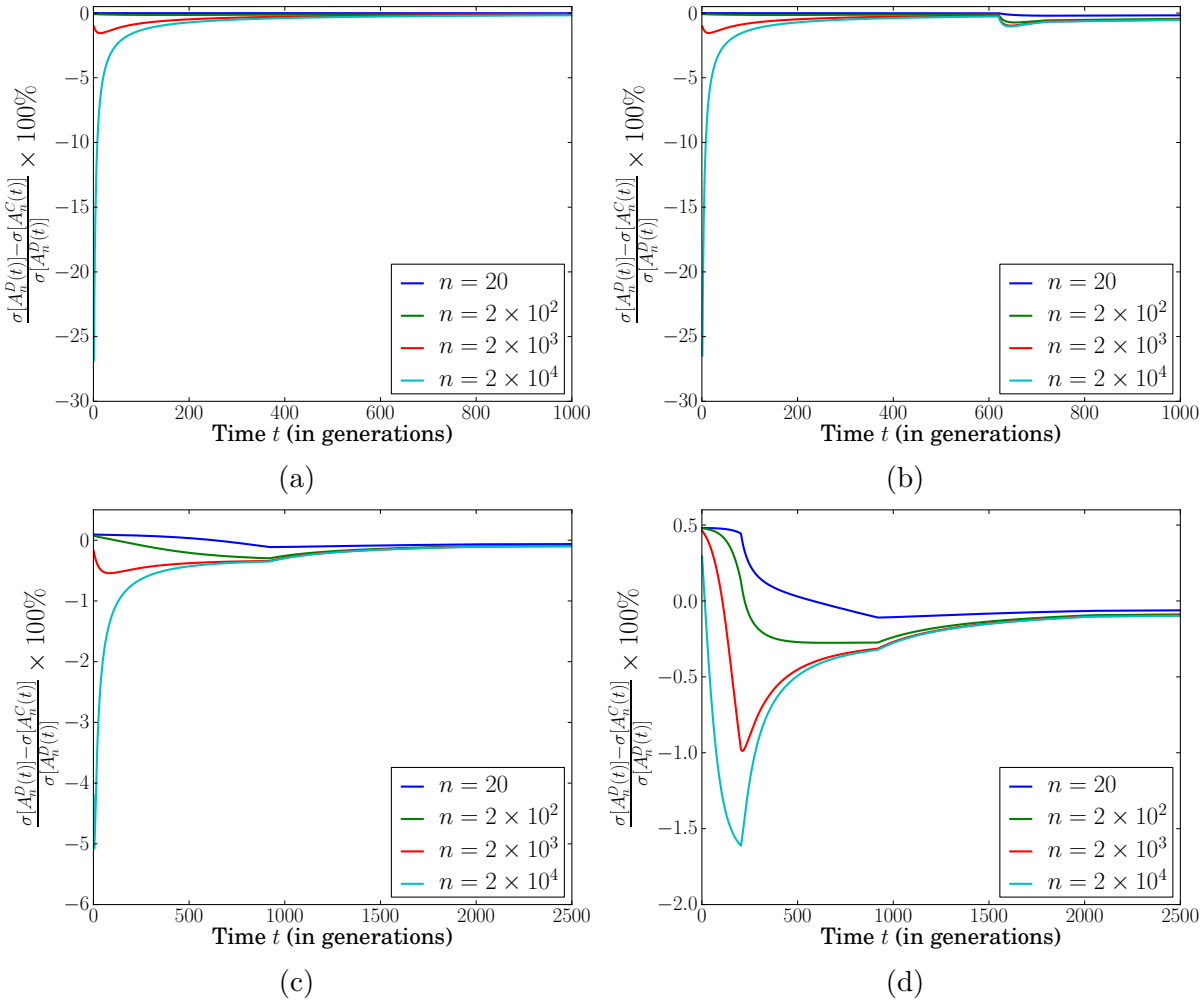


Figure 4.9: **The percentage difference in the standard deviation of the NLFT between the coalescent and DTWF models, for a haploid sample of size $n \in \{20, 2 \times 10^2, 2 \times 10^3, 2 \times 10^4\}$.** The plots correspond to (a) Model 1, (b) Model 2, (c) Model 3, and (d) Model 4.

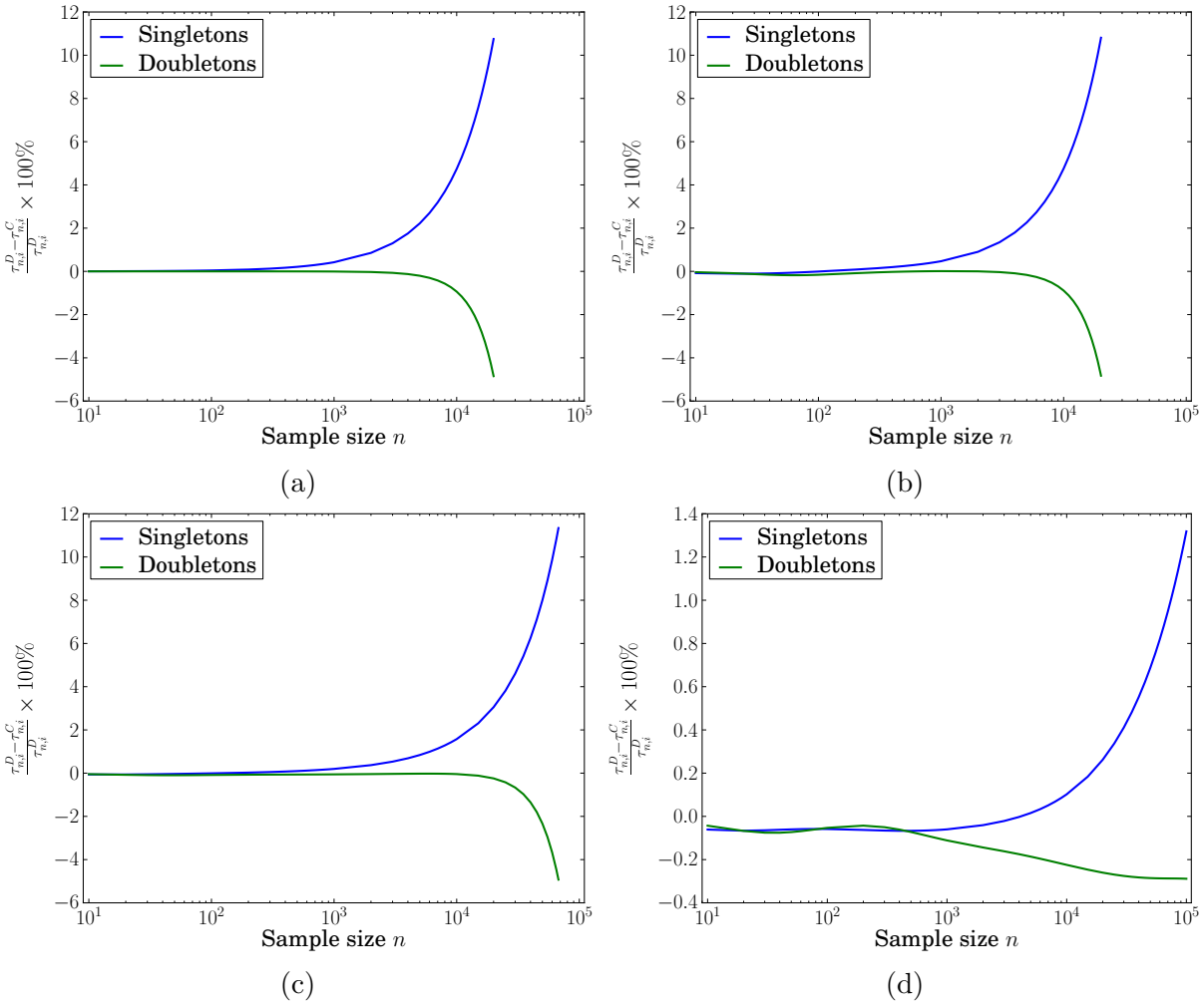


Figure 4.10: **The percentage relative error in the number of singletons and doubletons between the coalescent and DTWF models, as a function of the sample size n.** When the sample size is comparable to the current population size, the number of singletons predicted by the DTWF model is *larger* than the coalescent prediction by as much as 11%, while the number of doubletons predicted by the DTWF model is *smaller* than the coalescent prediction by about 4.8%. In MODEL 4, we could not consider a sample size comparable to the population size (10^6) because of computational burden, but we expect a similar extent of deviation as in MODELS 1–3 as n increases. Note that the y-axis scale for MODEL 4 is different from that for MODELS 1–3. (a) MODEL 1. (b) MODEL 2. (c) MODEL 3. (d) MODEL 4.

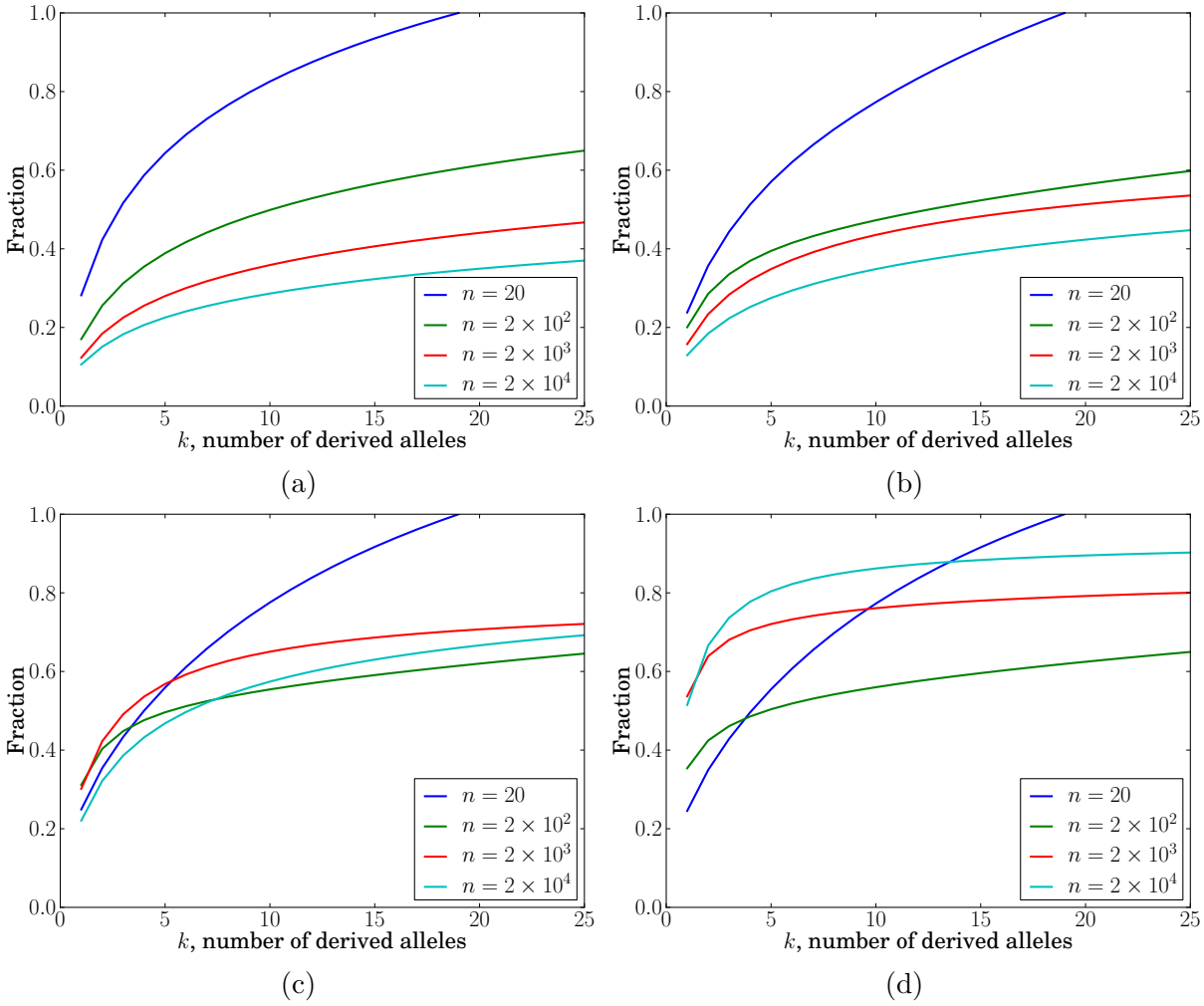


Figure 4.11: **The cumulative distribution function of the frequency spectrum, as a function of the number of derived alleles, k, in the DTWF model with sample sizes $n \in \{20, 2\times10^2, 2\times10^3, 2\times10^4\}$.** For each value of k on the x-axis, the y-axis is the proportion of segregating sites with at most k copies of the derived allele in the DTWF model, $\frac{\sum_{j=1}^{k}\tau^D_{n,j}}{\sum_{i=1}^{n-1}\tau^D_{n,i}}$. The plots correspond to (a) Model 1, (b) Model 2, (c) Model 3, and (d) Model 4. In Model 4 where the population grows rapidly in the recent past, about 51% of the segregating sites are singletons in a sample of size $n = 2\times10^4$. In Model 3, the fraction of singleton sites is lower than in Model 4 because the population growth rate is lower.

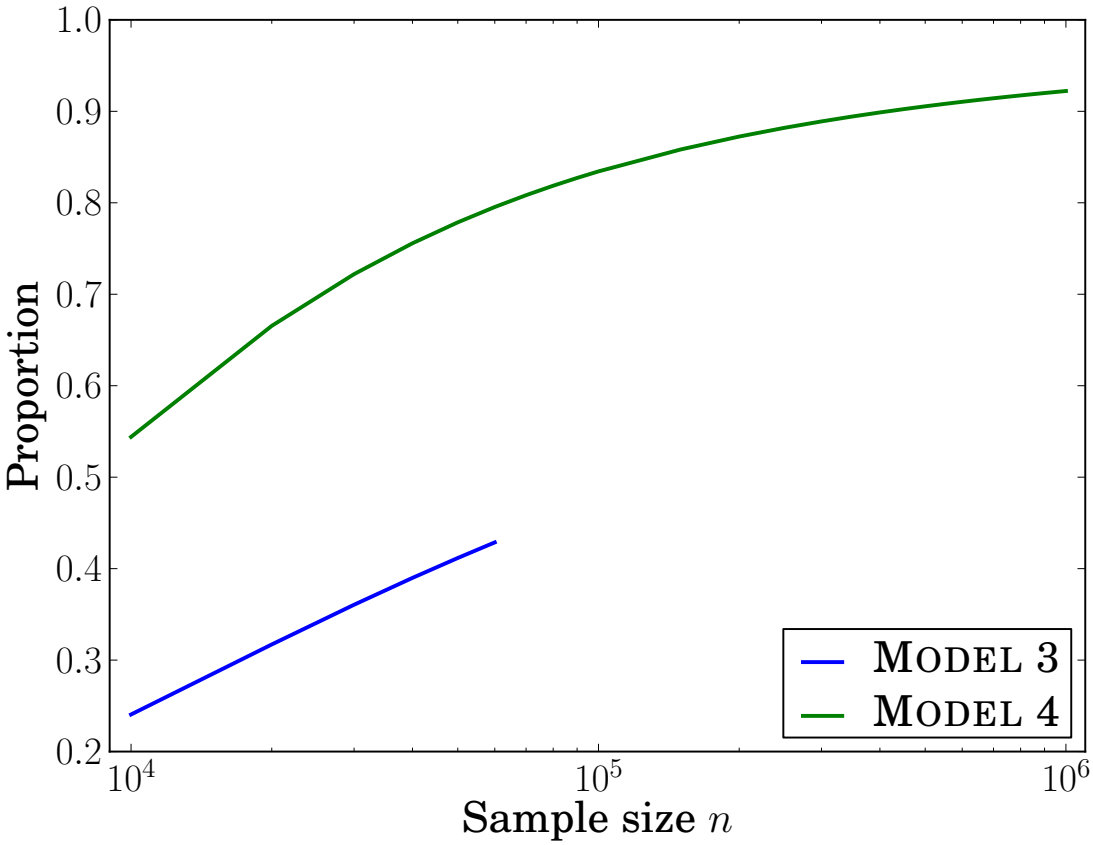


Figure 4.12: **Expected proportion of segregating sites with derived allele frequency $\leq 0.01\%$, as a function of the sample size n in the coalescent for** MODELS 3 **and** 4. The frequency of such rare variants is increasing in the sample size and in the case of MODEL 4, rare variants practically account for all the variants for very large sample sizes.

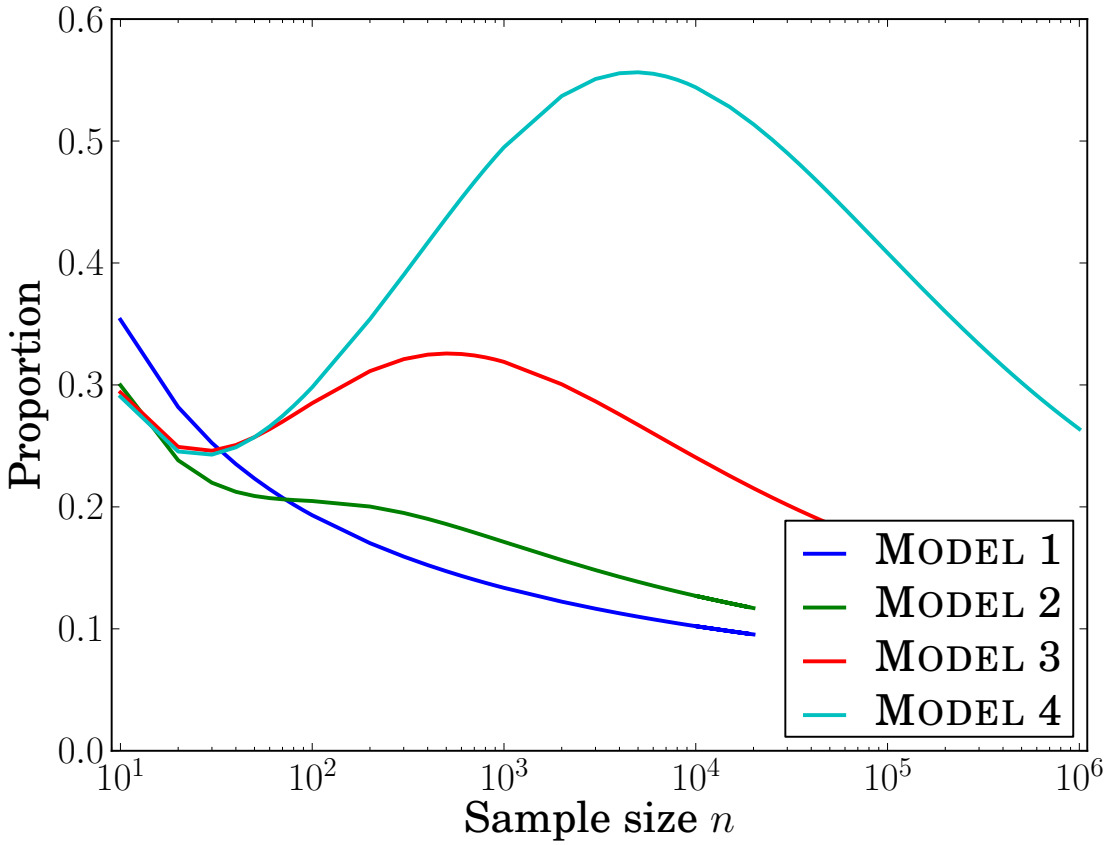


Figure 4.13: **Expected proportion of segregating sites that have only one copy of the derived allele (i.e., singletons), as a function of the sample size n in the coalescent for** MODELS 1–4. The difference in the number of singletons in MODEL 4 (which incorporates recent population expansion) and the number of singletons in MODELS 1 and 2 (without a recent expansion) rapidly increases for sample sizes beyond a few hundred individuals, indicating that large sample sizes are needed to infer demographic models of recent population growth from frequency spectrum data.

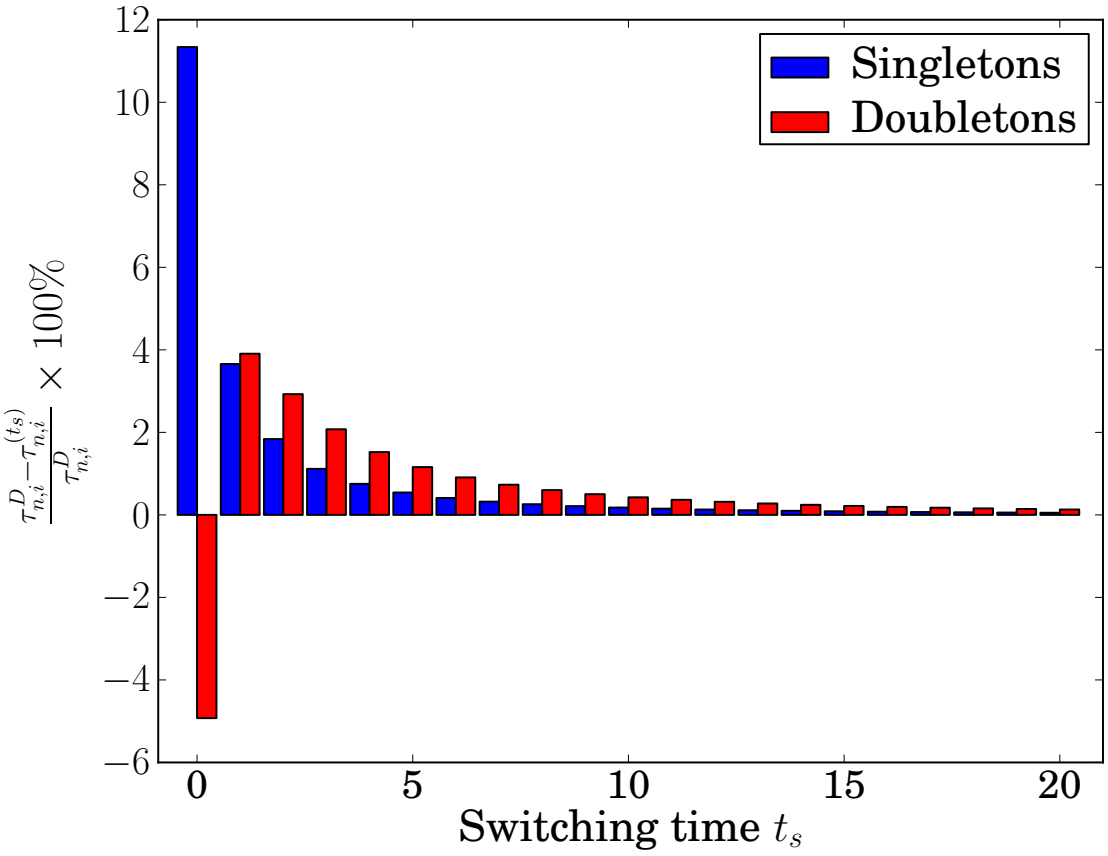


Figure 4.14: **The percentage relative error, with respect to the full DTWF model, in the number of singletons and doubletons in a hybrid algorithm with switching time t_s.** The hybrid method uses the DTWF model for generations $\leq t_s$ and the coalescent model in generations $> t_s$. The results are for MODEL 3 in the case the sample size n is equal to the current effective population size $N_0 = 67{,}627$. The case of $t_s = 0$ corresponds to using the coalescent model only. This plot shows that the difference in the frequency spectrum between the full DTWF model and the hybrid algorithm decreases very rapidly as the switching time t_s increases.

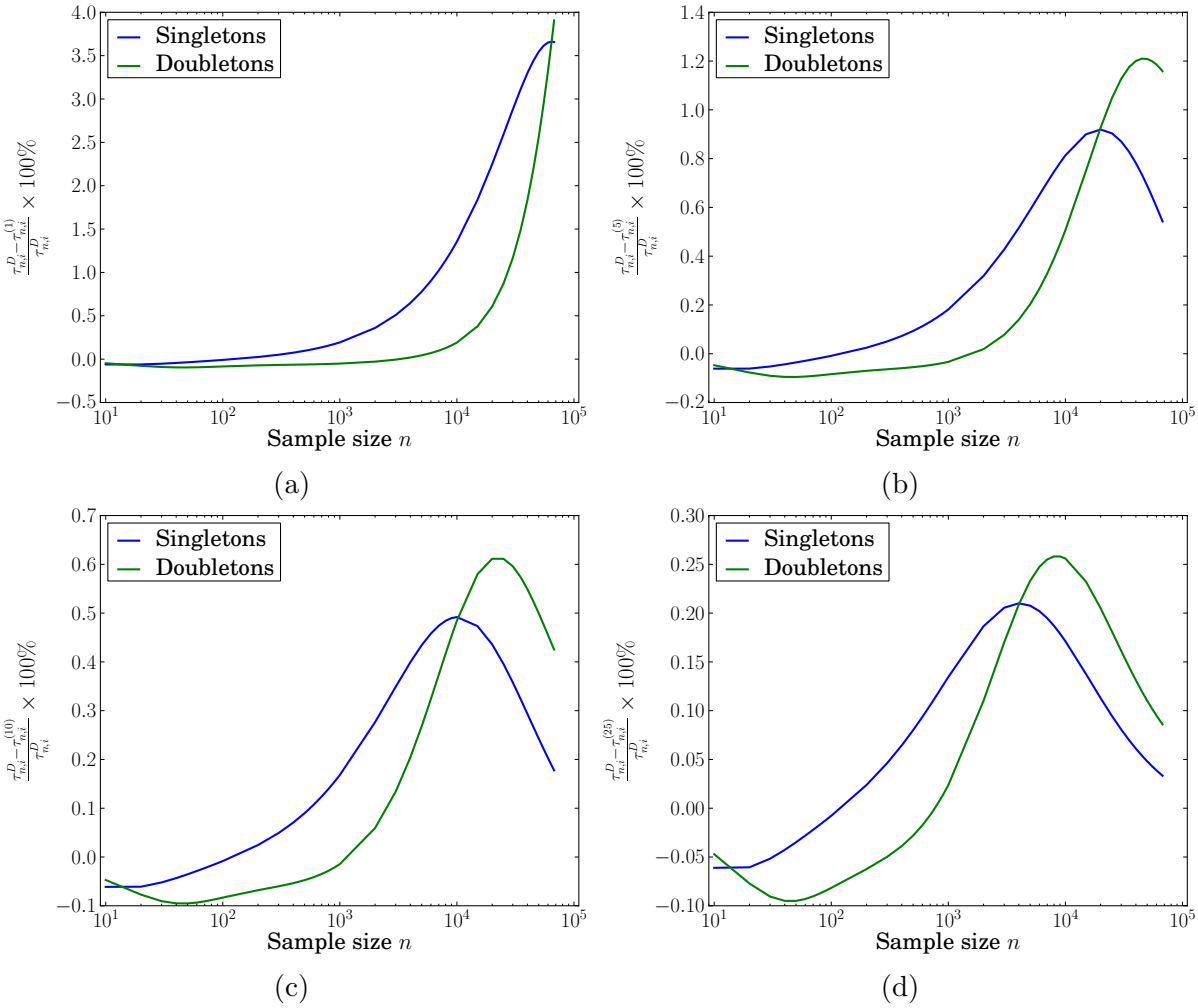


Figure 4.15: **The percentage relative error in the number of singletons and doubletons between the computations under the full DTWF model and a hybrid algorithm which uses the coalescent model in generations $t > t_s$ for** MODEL 3. The case of $t_s = 0$ corresponds to using the coalescent model only, illustrated in Figure 4.10c. We use $\boldsymbol{\tau}^{(t_s)}$ to denote the frequency spectrum using the hybrid algorithm with a switching time of t_s. As t_s increases, we see that the deviation between the computation under the DTWF model and the hybrid algorithm decreases monotonically, suggesting that one could use such an algorithm to efficiently approximate the frequency spectrum under the DTWF model. The values of t_s for the plots are (a) $t_s = 1$, (b) $t_s = 5$, (c) $t_s = 10$, (d) $t_s = 25$.

Chapter 5

Future directions

This dissertation examined several aspects of demographic inference for a randomly mating population using the sample frequency spectrum of a set of sequences. Beyond the open problems mentioned in the discussions in each chapter, there are several interesting questions for future research related to these aspects:

- What can we say about model identifiability from other representations of genomic data, such as full haplotype data or the distribution of IBD and IBS tract lengths? Several recent population size inference methods [48, 79] based on the sequentially Markov coalescent approximation [57] implicitly assume that piecewise-constant population sizes are identifiable from just two haplotypes. If this is indeed the case, what is the minimum sequence length required for identifiability under the full coalescent with recombination or approximations like the sequentially Markov coalescent? The same questions can also be asked about the evolutionary models that are used by demographic inference methods [34,64] that utilize the distribution of IBS and IBD tract lengths for a pair of haplotypes.

- What can we say about the number of samples sufficient for demographic identifiability if we have access to only a few entries of the expected SFS? More generally, how many samples are sufficient for identifiability given access to a linear transformation of the expected SFS? Such a result would generalize the identifiability theorem for the folded frequency spectrum given in Section 2.3.4 of Chapter 2.

- As the number of segregating sites increases, the empirical SFS converges in probability to the expected SFS, thus allowing the demographic model to be reliably inferred from sample sizes such as those given in Chapter 2. However, if we have access to far more samples than the minimum needed for identifiability, then simulations indicate that we might require fewer segregating sites for reliable inference than if we used a smaller sample size. This suggests the following interesting question: for a given total number of bases of sequencing capacity, what is the trade-off between the sample size and sequence length that maximizes the accuracy of our inference method?

- Can we develop demographic inference methods based on exact computation that use the joint SFS from samples from multiple populations while taking into account complex demographic events like admixture and continuous migration? Chen [9] has derived analytic expressions for the entries of the expected joint SFS of multiple populations that are related by split and point admixture events. It would be interesting to extend those results to the case of multiple populations with continuous migration between them.

- What are the extremal demographic models for the deviations of the various genealogical quantities between the discrete-time Wright-Fisher and coalescent models? What can we say about this question for other discrete-time models of mating that also converge to the coalescent, such as the discrete-time Moran model or the more general family of Cannings exchangeable models?

Bibliography

[1] 1000 Genomes Project Consortium (2010). A map of human genome variation from population-scale sequencing. *Nature* **467,** 1061–1073.

[2] Barton, N. H. and Charlesworth, B. (1984). Genetic revolutions, founder effects, and speciation. *Annual Review of Ecology and Systematics* **15,** 133–164.

[3] Bhaskar, A., Clark, A. G. and Song, Y. S. (2013). Distortion of genealogical properties when the sample is very large. arXiv:1308.0091 [q-bio].

[4] Bhaskar, A. and Song, Y. S. (2009). Multi-locus match probability in a finite population: a fundamental difference between the Moran and Wright–Fisher models. *Bioinformatics* **25,** i187–i195.

[5] Bhaskar, A. and Song, Y. S. (2013). The identifiability of piecewise demographic models from the sample frequency spectrum. arXiv:1309.5056 [q-bio].

[6] Boyko, A. R., Williamson, S. H., Indap, A. R., Degenhardt, J. D., Hernandez, R. D., Lohmueller, K. E., Adams, M. D., Schmidt, S., Sninsky, J. J., Sunyaev, S. R. et al. (2008). Assessing the evolutionary impact of amino acid mutations in the human genome. *PLoS Genetics* **4,** e1000083.

[7] Campbell, C. D., Ogburn, E. L., Lunetta, K. L., Lyon, H. N., Freedman, M. L., Groop, L. C., Altshuler, D., Ardlie, K. G. and Hirschhorn, J. N. (2005). Demonstrating stratification in a European American population. *Nature Genetics* **37,** 868–872.

[8] Chen, G. K., Marjoram, P. and Wall, J. D. (2009). Fast and flexible simulation of dna sequence data. *Genome Research* **19,** 136–142.

[9] Chen, H. (2012). The joint allele frequency spectrum of multiple populations: A coalescent theory approach. *Theoretical Population Biology* **81,** 179–195.

[10] Clayton, D. G., Walker, N. M., Smyth, D. J., Pask, R., Cooper, J. D., Maier, L. M., Smink, L. J., Lam, A. C., Ovington, N. R., Stevens, H. E. et al. (2005). Population structure, differential bias and genomic control in a large-scale, case-control association study. *Nature Genetics* **37,** 1243–1246.

[11] Conrad, D., Keebler, J., DePristo, M., Lindsay, S., Zhang, Y., Casals, F., Idaghdour, Y., Hartl, C., Torroja, C., Garimella, K. et al. (2011). Variation in genome-wide mutation rates within and between human families. *Nature* **201,** 1.

[12] Coventry, A., Bull-Otterson, L. M., Liu, X., Clark, A. G., Maxwell, T. J., Crosby, J., Hixson, J. E., Rea, T. J., Muzny, D. M., Lewis, L. R. et al. (2010). Deep resequencing reveals excess rare recent variants consistent with explosive population growth. *Nature Communications* **1,** 131.

[13] Davies, J. L., Simančík, F., Lyngsø, R., Mailund, T. and Hein, J. (2007). On recombination-induced multiple and simultaneous coalescent events. *Genetics* **177,** 2151–2160.

[14] Durrett, R. (2008). *Probability models for DNA sequence evolution* 2nd ed. Springer, New York.

[15] Erlich, H. A., Bergström, T. F., Stoneking, M. and Gyllensten, U. (1996). HLA sequence polymorphism and the origin of humans. *Science* **274,** 1552–1554.

[16] Excoffier, L., Dupanloup, I., Huerta-Sánchez, E., Sousa, V. C. and Foll, M. (2013). Robust demographic inference from genomic and SNP data. *PLoS Genetics* **9,** e1003905.

[17] Excoffier, L. and Foll, M. (2011). Fastsimcoal: a continuous-time coalescent simulator of genomic diversity under arbitrarily complex evolutionary scenarios. *Bioinformatics* **27,** 1332–1334.

[18] Fisher, R. A. et al. (1930). The distribution of gene ratios for rare mutations. *Proceedings of the Royal Society of Edinburgh* **50,** 205–220.

[19] Fu, W., O'Connor, T. D., Jun, G., Kang, H. M., Abecasis, G., Leal, S. M., Gabriel, S., Altshuler, D., Shendure, J., Nickerson, D. A. et al. (2012). Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. *Nature* **493,** 216–220.

[20] Fu, Y.-X. (1995). Statistical properties of segregating sites. *Theoretical Population Biology* **48,** 172–197.

[21] Fu, Y.-X. (2006). Exact coalescent for the Wright-Fisher model. *Theoretical Population Biology* **69,** 385–394.

[22] Gantmacher, F. R. (2000). *The theory of matrices* vol. 2. Chelsea Publishing Company, New York.

[23] Gravel, S., Henn, B. M., Gutenkunst, R. N., Indap, A. R., Marth, G. T., Clark, A. G., Yu, F., Gibbs, R. A., Bustamante, C. D., Altshuler, D. L. et al. (2011). Demographic history and rare allele sharing among human populations. *Proceedings of the National Academy of Sciences* **108,** 11983–11988.

[24] Griewank, A. and Corliss, G. F. (1991). *Automatic differentiation of algorithms: theory, implementation, and application*. Society for industrial and Applied Mathematics Philadelphia, PA.

[25] Griffiths, R. (1991). The two-locus ancestral graph. *Selected Proceedings of the Sheffield Symposium on Applied Probability. IMS Lecture Notes–Monograph Series* **18,** 100–117.

[26] Griffiths, R. (2003). The frequency spectrum of a mutation, and its age, in a general diffusion model. *Theoretical Population Biology* **64,** 241–251.

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

[28] Griffiths, R. and Tavaré, S. (1998). The age of a mutation in a general coalescent tree. *Communications in Statistics. Stochastic Models* **14,** 273–295.

[29] Griffiths, R. C. and Marjoram, P. (1997). An ancestral recombination graph. *Institute for Mathematics and its Applications* **87,** 257.

[30] Gusev, A., Lowe, J. K., Stoffel, M., Daly, M. J., Altshuler, D., Breslow, J. L., Friedman, J. M. and Pe'er, I. (2009). Whole population, genome-wide mapping of hidden relatedness. *Genome Research* **19,** 318–326.

[31] Gutenkunst, R. N., Hernandez, R. D., Williamson, S. H. and Bustamante, C. D. (2009). Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. *PLoS Genetics* **5,** e1000695.

[32] Harding, R. M., Fullerton, S., Griffiths, R. et al. (1997). Archaic African and Asian lineages in the genetic ancestry of modern humans. *American Journal of Human Genetics* **60,** 772.

[33] Harpending, H. C., Batzer, M. A., Gurven, M. et al. (1998). Genetic traces of ancient demography. *Proceedings of the National Academy of Sciences* **95,** 1961–1967.

[34] Harris, K. and Nielsen, R. (2013). Inferring demographic history from a spectrum of shared haplotype lengths. *PLoS Genetics* **9,** e1003521.

[35] Herbots, H. (1997). The structured coalescent. *Progress in Population Genetics and Human Evolution* **87,** 231–255.

[36] Hudson, R. (2002). Generating samples under a wright–fisher neutral model of genetic variation. *Bioinformatics* **18,** 337–338.

[37] Hudson, R. R. (1983). Properties of a neutral allele model with intragenic recombination. *Theoretical Population Biology* **23,** 183–201.

[38] Jameson, G. (2006). Counting zeros of generalised polynomials: Descartes' rule of signs and Laguerre's extensions. *The Mathematical Gazette* **90,** 223–234.

[39] Keinan, A. and Clark, A. G. (2012). Recent explosive human population growth has resulted in an excess of rare genetic variants. *Science* **336,** 740–743.

[40] Keinan, A., Mullikin, J. C., Patterson, N. and Reich, D. (2007). Measurement of the human allele frequency spectrum demonstrates greater genetic drift in East Asians than in Europeans. *Nature Genetics* **39,** 1251–1255.

[41] Kidd, J. M., Gravel, S., Byrnes, J., Moreno-Estrada, A., Musharoff, S., Bryc, K., Degenhardt, J. D., Brisbin, A., Sheth, V., Chen, R. et al. (2012). Population genetic inference from personal genome data: impact of ancestry and admixture on human genomic variation. *The American Journal of Human Genetics* **91,** 660–671.

[42] Kimura, M. (1955). Solution of a process of random genetic drift with a continuous model. *Proceedings of the National Academy of Sciences* **41,** 144–150.

[43] Kimura, M. (1969). The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutations. *Genetics* **61,** 893.

[44] Kingman, J. F. C. (1982). The coalescent. *Stochastic Processes and Their Applications* **13,** 235–248.

[45] Kingman, J. F. C. (1982). Exchangeability and the evolution of large populations. In *Exchangeability in probability and statistics*. ed. G. Koch and F. Spizzichino. North-Holland Publishing Company pp. 97–112.

[46] Kingman, J. F. C. (1982). On the genealogy of large populations. *Journal of Applied Probability* **19,** 27–43.

[47] Krone, S. M. and Neuhauser, C. (1997). Ancestral processes with selection. *Theoretical Population Biology* **51,** 210–237.

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

[49] Lohmueller, K. E., Indap, A. R., Schmidt, S., Boyko, A. R., Hernandez, R. D., Hubisz, M. J., Sninsky, J. J., White, T. J., Sunyaev, S. R., Nielsen, R. et al. (2008). Proportionally more deleterious genetic variation in European than in African populations. *Nature* **451,** 994–997.

[50] Low, B. and Clarke, A. (1991). Family patterns in nineteenth-century sweden: impact of occupational status and landownership. *Journal of Family History* **16,** 117–138.

[51] Lukić, S. and Hey, J. (2012). Demographic inference using spectral methods on SNP data, with an analysis of the human out-of-Africa expansion. *Genetics* **192,** 619–639.

[52] Lukić, S., Hey, J. and Chen, K. (2011). Non-equilibrium allele frequency spectra via spectral methods. *Theoretical Population Biology* **79,** 203–219.

[53] Marchini, J., Cardon, L. R., Phillips, M. S. and Donnelly, P. (2004). The effects of human population structure on large genetic association studies. *Nature Genetics* **36,** 512–517.

[54] Marjoram, P. and Wall, J. D. (2006). Fast "coalescent" simulation. *BMC Genetics* **7,** 16.

[55] Marth, G., Czabarka, E., Murvai, J. and Sherry, S. (2004). The allele frequency spectrum in genome-wide human variation data reveals signals of differential demographic history in three large world populations. *Genetics* **166,** 351–372.

[56] Maruvka, Y. E., Shnerb, N. M., Bar-Yam, Y. and Wakeley, J. (2011). Recovering population parameters from a single gene genealogy: An unbiased estimator of the growth rate. *Molecular Biology and Evolution* **28,** 1617–1631.

[57] McVean, G. A. and Cardin, N. J. (2005). Approximating the coalescent with recombination. *Philosophical Transactions of the Royal Society B: Biological Sciences* **360,** 1387–1393.

[58] Möhle, M. and Sagitov, S. (2001). A classification of coalescent processes for haploid exchangeable population models. *The Annals of Probability* **29,** 1547–1562.

[59] Möhle, M. and Sagitov, S. (2003). Coalescent patterns in diploid exchangeable population models. *Journal of Mathematical Biology* **47,** 337–352.

[60] Moreau, C., Bhérer, C., Vézina, H., Jomphe, M., Labuda, D. and Excoffier, L. (2011). Deep human genealogies reveal a selective advantage to be on an expanding wave front. *Science* **334,** 1148–1150.

[61] Myers, S., Fefferman, C. and Patterson, N. (2008). Can one learn history from the allelic spectrum? *Theoretical Population Biology* **73,** 342–348.

[62] Nelson, M. R., Wegmann, D., Ehm, M. G., Kessner, D., Jean, P. S., Verzilli, C., Shen, J., Tang, Z., Bacanu, S.-A., Fraser, D. et al. (2012). An abundance of rare functional variants in 202 drug target genes sequenced in 14,002 people. *Science* **337,** 100–104.

[63] Nielsen, R. (2005). Molecular signatures of natural selection. *Annual Review of Genetics* **39,** 197–218.

[64] Palamara, P. F., Lencz, T., Darvasi, A. and Pe'er, I. (2012). Length distributions of identity by descent reveal fine-scale demographic history. *American Journal of Human Genetics* **91,** 809–822.

[65] Pasaniuc, B., Zaitlen, N., Lettre, G., Chen, G. K., Tandon, A., Kao, W. L., Ruczinski, I., Fornage, M., Siscovick, D. S., Zhu, X. et al. (2011). Enhanced statistical tests for GWAS in admixed populations: assessment using African Americans from CARe and a breast cancer consortium. *PLoS Genetics* **7,** e1001371.

[66] Petkovšek, M., Wilf, H. and Zeilberger, D. (1996). *A=B*. AK Peters Ltd, Wellesley, MA.

[67] Pitman, J. (1999). Coalescents with multiple collisions. *Annals of Probability* **27,** 1870–1902.

[68] Polanski, A., Bobrowski, A. and Kimmel, M. (2003). A note on distributions of times to coalescence, under time-dependent population size. *Theoretical Population Biology* **63,** 33–40.

[69] Polanski, A. and Kimmel, M. (2003). New explicit expressions for relative frequencies of single-nucleotide polymorphisms with application to statistical inference on population growth. *Genetics* **165,** 427–436.

[70] Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A. and Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics* **38,** 904–909.

[71] Ptak, S. E. and Przeworski, M. (2002). Evidence for population growth in humans is confounded by fine-scale population structure. *Trends in Genetics* **18,** 559–563.

[72] Rohlfs, R. V., Murphy, E., Song, Y. S. and Slatkin, M. (2013). The influence of relatives on the efficiency and error rate of familial searching. *PLoS ONE* **8,** e70495+.

[73] Roux, C., Pauwels, M., Ruggiero, M.-V., Charlesworth, D., Castric, V. and Vekemans, X. (2013). Recent and ancient signature of balancing selection around the S-Locus in *Arabidopsis halleri* and *A. lyrata*. *Molecular Biology and Evolution* **30,** 435–447.

[74] Sagitov, S. (1999). The general coalescent with asynchronous mergers of ancestral lines. *Journal of Applied Probability* **36,** 1116–1125.

[75] Sankararaman, S., Patterson, N., Li, H., Pääbo, S. and Reich, D. (2012). The date of interbreeding between Neandertals and modern humans. *PLoS Genetics* **8,** e1002947.

[76] Sawyer, S. A. and Hartl, D. L. (1992). Population genetics of polymorphism and divergence. *Genetics* **132,** 1161–76.

[77] Schaffner, S., Foo, C., Gabriel, S., Reich, D., Daly, M. and Altshuler, D. (2005). Calibrating a coalescent simulation of human genome sequence variation. *Genome Research* **15,** 1576–1583.

[78] Schweinsberg, J. (2000). Coalescents with simultaneous multiple collisions. *Electronic Journal of Probability* **5,** 1–50.

[79] Sheehan, S., Harris, K. and Song, Y. S. (2013). Estimating variable effective population sizes from multiple genomes: A sequentially Markov conditional sampling distribution approach. *Genetics* **194,** 647–662.

[80] Skoglund, P. and Jakobsson, M. (2011). Archaic human ancestry in East Asia. *Proceedings of the National Academy of Sciences* **108,** 18301–18306.

[81] Städler, T., Haubold, B., Merino, C., Stephan, W. and Pfaffelhuber, P. (2009). The impact of sampling schemes on the site frequency spectrum in nonequilibrium subdivided populations. *Genetics* **182,** 205–216.

[82] Takahata, N. (1993). Allelic genealogy and human evolution. *Molecular Biology and Evolution* **10,** 2–22.

[83] Takahata, N. and Nei, M. (1985). Gene genealogy and variance of interpopulational nucleotide differences. *Genetics* **110,** 325–344.

[84] Tavaré, S. (1984). Line-of-descent and genealogical processes, and their applications in population genetics models. *Theoretical Population Biology* **26,** 119–164.

[85] Tennessen, J. A., Bigham, A. W., O'Connor, T. D., Fu, W., Kenny, E. E., Gravel, S., McGee, S., Do, R., Liu, X., Jun, G. et al. (2012). Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science* **337,** 64–69.

[86] Tian, C., Gregersen, P. K. and Seldin, M. F. (2008). Accounting for ancestry: population substructure and genome-wide association studies. *Human molecular genetics* **17,** R143–R150.

[87] Voight, B. F., Adams, A. M., Frisse, L. A., Qian, Y., Hudson, R. R. and Di Rienzo, A. (2005). Interrogating multiple aspects of variation in a full resequencing data set to infer human population size changes. *Proceedings of the National Academy of Sciences* **102,** 18508–18513.

[88] Wächter, A. and Biegler, L. T. (2006). On the implementation of an interior-point filter line-search algorithm for large-scale nonlinear programming. *Mathematical Programming* **106,** 25–57.

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

[90] Wakeley, J. and Takahashi, T. (2003). Gene genealogies when the sample size exceeds the effective size of the population. *Molecular Biology and Evolution* **20,** 208–213.

[91] Walther, A. and Griewank, A. (2012). Getting started with ADOL-C. In *Combinatorial Scientific Computing*. ed. U. N. und O. Schenk. Chapman-Hall CRC Computational Science pp. 181–202.

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

[93] Williamson, S. H., Hernandez, R., Fledel-Alon, A., Zhu, L., Nielsen, R. and Bustamante, C. D. (2005). Simultaneous inference of selection and population growth from patterns of variation in the human genome. *Proceedings of the National Academy of Sciences* **102,** 7882–7887.

[94] Wiuf, C. and Hein, J. (1999). Recombination as a point process along sequences. *Theoretical Population Biology* **55,** 248–259.

[95] Živković, D. and Stephan, W. (2011). Analytical results on the neutral non-equilibrium allele frequency spectrum based on diffusion theory. *Theoretical Population Biology* **79,** 184–191.

[96] Živković, D. and Wiehe, T. (2008). Second-order moments of segregating sites under variable population size. *Genetics* **180,** 341–357.

Appendix A

Some identities related to Kingman's coalescent

In this appendix, we derive analytic expressions for several genealogical quantities under the coalescent with a time-varying population size function. While these results are known in previous works, we give independent elementary combinatorial proofs here.

A.1 Analytic expressions for the distribution of the coalescent waiting times

In this section, we derive analytic expressions for the entries of the expected SFS $\boldsymbol{\xi}_n$ of a sample of size n under a time-varying population size and the infinite-sites model of mutation. These expressions also appear in the works of Polanski *et al.* [68] and Živković and Wiehe [96]. Consider Kingman's coalescent with a time-varying population size function $\eta(t)$, where $t = 0$ corresponds to the present, and t increases going back into the past. Similar to the notation in Chapter 2, we let $T^{(\eta)}_{n,k}$ denote the random variable representing the time (in coalescent units) spent while there are k lineages in the coalescent process, given that we start with a sample of size n at $t = 0$. We also let $\Sigma^{(\eta)}_{n,k}$ be the random variable denoting the total time spent in the process while there are at least k surviving lineages, given that we start with a sample of size n at $t = 0$. For notational brevity, we will drop the dependence of these random variables on the population size function η where it is clear from context. The random variables $\Sigma_{n,k}$ and $T_{n,k}$, for $2 \leq k \leq n$, are related by

$$\Sigma_{n,k} = \sum_{j=k}^{n} T_{n,j},$$
$$T_{n,k} = \Sigma_{n,k} - \Sigma_{n,k+1},$$

where for consistency, we define $\Sigma_{n,n+1} = 0$.

For a coalescent-rescaled population size function $\eta(t)$, the total intensity of any pairwise coalescence up to time t, $R_\eta(t)$, is given by

$$R_\eta(t) := \int_0^t \frac{1}{\eta(x)}\, dx. \tag{A.1}$$

From the Markov property of the ancestral process mentioned in Section 1.3.4 of Chapter 1, the sequence of random variables $\Sigma_{n,n}, \ldots, \Sigma_{n,2}$ form a first-order Markov chain with the pdf of $\Sigma_{n,k}$ conditional on $\Sigma_{n,k+1} = \tau$ given by

$$f_{\Sigma_{n,k}|\Sigma_{n,k+1}=\tau}(t) = \begin{cases} \binom{k}{2}\frac{1}{\eta(t)} \exp\left(-\binom{k}{2}(R_\eta(t) - R_\eta(\tau))\right) & \text{if } t \geq \tau \\ 0 & \text{if } t < \tau. \end{cases} \tag{A.2}$$

Theorem 12. *The pdf $f_{\Sigma_{n,k}}$ of the total time $\Sigma_{n,k}$ while there are at least k ancestors for a sample of size n at time 0 is given by,*

$$f_{\Sigma_{n,k}}(t) = \sum_{j=k}^{n} a_{n,k,j} \binom{j}{2} \frac{1}{\eta(t)} \exp\left(-\binom{j}{2} R_\eta(t)\right), \tag{A.3}$$

where the $a_{n,k,j}$ are given by

$$a_{n,k,j} = \frac{\prod_{l=k, l\neq j}^{n} \binom{l}{2}}{\prod_{l=k, l\neq j}^{n} \left[\binom{l}{2} - \binom{j}{2}\right]}. \tag{A.4}$$

Proof. Letting $\lambda_k = \binom{k}{2}$, for any sample size n, the pdf $f_{\Sigma_{n,n}}$ of the waiting time to the first coalescence event $\Sigma_{n,n}$ is given by

$$f_{\Sigma_{n,n}}(t) = \lambda_n \frac{1}{\eta(t)} \exp\left(-\lambda_n R_\eta(t)\right). \tag{A.5}$$

Now consider the pdf of the total waiting time to the second coalescence event $\Sigma_{n,n-1}$. Using (A.5) and (A.2), we have

$$\begin{aligned}
f_{\Sigma_{n,n-1}}(t) &= \int_0^t f_{\Sigma_{n,n}}(\tau) f_{\Sigma_{n,n-1}|\Sigma_{n,n}=\tau}(t) d\tau \\
&= \int_0^t \lambda_n \frac{1}{\eta(\tau)} \exp\left(-\lambda_n R_\eta(\tau)\right) \lambda_{n-1} \frac{1}{\eta(t)} \exp\left(-\lambda_{n-1}(R_\eta(t) - R_\eta(\tau))\right) d\tau \\
&= \lambda_n \lambda_{n-1} \frac{1}{\eta(t)} \exp\left(-\lambda_{n-1} R_\eta(t)\right) \int_0^t \frac{1}{\eta(\tau)} \exp\left[-\left(\lambda_n - \lambda_{n-1}\right) R_\eta(\tau)\right] d\tau \\
&= \lambda_n \lambda_{n-1} \frac{1}{\eta(t)} \exp\left(-\lambda_{n-1} R_\eta(t)\right) \frac{1 - \exp\left[-\left(\lambda_n - \lambda_{n-1}\right) R_\eta(t)\right]}{\lambda_n - \lambda_{n-1}} \\
f_{\Sigma_{n,n-1}}(t) &= \frac{\lambda_n \lambda_{n-1}}{\lambda_n - \lambda_{n-1}} \frac{1}{\eta(t)} \exp\left(-\lambda_{n-1} R_\eta(t)\right) - \frac{\lambda_n \lambda_{n-1}}{\lambda_n - \lambda_{n-1}} \frac{1}{\eta(t)} \exp\left(-\lambda_n R_\eta(t)\right).
\end{aligned} \tag{A.6}$$

Equations (A.5) and (A.6) suggest the following ansatz for the pdf $f_{\Sigma_{n,k}}$ of $\Sigma_{n,k}$:

$$f_{\Sigma_{n,k}}(t) = \sum_{j=k}^{n} b_{n,k,j} \frac{1}{\eta(t)} \exp\left(-\lambda_j R_\eta(t)\right), \tag{A.7}$$

where the coefficients $b_{n,k,j}$ are real-valued constants that only depend on n, k and j (and not on $\eta(t)$). We can get a recurrence relation between the coefficients $b_{n,k,j}$ by substituting the ansatz (A.7) into (A.2) as follows,

$$\begin{aligned}
f_{\Sigma_{n,k}}(t) &= \int_0^t f_{\Sigma_{n,k+1}}(\tau) f_{\Sigma_{n,k}|\Sigma_{n,k+1}=\tau}(t) d\tau \\
&= \int_0^t \left\{ \sum_{j=k+1}^{n} b_{n,k,j} \frac{1}{\eta(\tau)} \exp\left(-\lambda_j R_\eta(\tau)\right) \right\} \lambda_k \frac{1}{\eta(t)} \exp\left(-\lambda_k (R_\eta(t) - R_\eta(\tau))\right) d\tau \\
f_{\Sigma_{n,k}}(t) &= \sum_{j=k+1}^{n} b_{n,k+1,j} \frac{\lambda_k}{\lambda_j - \lambda_k} \frac{1}{\eta(t)} \exp\left(-\lambda_k R_\eta(t)\right) - \sum_{j=k+1}^{n} b_{n,k+1,j} \frac{\lambda_k}{\lambda_j - \lambda_k} \frac{1}{\eta(t)} \exp\left(-\lambda_j R_\eta(t)\right).
\end{aligned} \tag{A.8}$$

Equating the coefficients of (A.7) with the coefficients in (A.8), we get the following recurrence relations,

$$b_{n,k,k} = \sum_{j=k+1}^{n} b_{n,k+1,j} \frac{\lambda_k}{\lambda_j - \lambda_k}, \qquad 2 \le k \le n, \tag{A.9}$$

$$b_{n,k,j} = b_{n,k+1,j} \frac{\lambda_k}{\lambda_k - \lambda_j}, \qquad 2 \le k \le n \text{ and } k+1 \le j \le n, \tag{A.10}$$

with boundary conditions,

$$b_{n,n,n} = \lambda_n. \tag{A.11}$$

It is clear that (A.9)–(A.11) has a unique solution for the $b_{n,k,j}$. Using (A.9) and (A.10), we can explicitly compute the first few values of $b_{n,k,j}$,

$$\begin{aligned}
b_{n,n-1,n-1} &= \frac{\lambda_{n-1}\lambda_n}{\lambda_n - \lambda_{n-1}}, & b_{n,n-1,n} &= \frac{\lambda_{n-1}\lambda_n}{\lambda_{n-1} - \lambda_n}, \\
b_{n,n-2,n-2} &= \frac{\lambda_{n-2}\lambda_{n-1}\lambda_n}{(\lambda_{n-1} - \lambda_{n-2})(\lambda_n - \lambda_{n-2})}, & b_{n,n-2,n-1} &= \frac{\lambda_{n-2}\lambda_{n-1}\lambda_n}{(\lambda_{n-2} - \lambda_{n-1})(\lambda_n - \lambda_{n-1})}, \\
b_{n,n-2,n} &= \frac{\lambda_{n-2}\lambda_{n-1}\lambda_n}{(\lambda_{n-2} - \lambda_n)(\lambda_{n-1} - \lambda_n)}.
\end{aligned}$$

From these equations, we can conjecture that $b_{n,k,j}$, $2 \le k \le j \le n$, is given by

$$b_{n,k,j} = \frac{\prod_{l=k}^{n} \lambda_l}{\prod_{l=k, l\neq j}^{n} \lambda_l - \lambda_j}. \tag{A.12}$$

It is easy to see that the conjectured expression (A.12) for $b_{n,k,j}$ satisfies (A.10) and (A.11). To see that it satisfies (A.9), consider the partial fraction expansion of $\prod_{l=k+1}^{n} \frac{1}{\lambda_l - x}$ in the variable x,

$$\prod_{l=k+1}^{n} \frac{1}{\lambda_l - x} = \sum_{j=k+1}^{n} \frac{B_j}{\lambda_j - x}, \tag{A.13}$$

where the coefficients B_j can be found by the Heaviside cover-up method to be

$$B_j = \prod_{\substack{l=k+1 \\ l \neq j}}^{n} \frac{1}{\lambda_l - \lambda_j}. \tag{A.14}$$

Substituting $x = \lambda_k$ into (A.13), using (A.14), and multiplying both sides of (A.13) by $\prod_{l=k}^{n} \lambda_l$, we have,

$$\frac{\prod_{l=k}^{n} \lambda_l}{\prod_{l=k+1}^{n} \lambda_l - \lambda_k} = \sum_{j=k+1}^{n} \left(\frac{\prod_{l=k+1}^{n} \lambda_l}{\prod_{\substack{l=k+1 \\ l \neq j}}^{n} \lambda_l - \lambda_j} \right) \frac{\lambda_k}{\lambda_l - \lambda_k}. \tag{A.15}$$

Note that the term on the left hand side matches the conjectured expression (A.12) for $b_{n,k,k}$, and the term in the parentheses on the right hand side is the conjectured expression for $b_{n,k+1,j}$. Hence, (A.12) is the solution to (A.9)–(A.11). Using (A.12) and noting that $b_{n,k,j} = a_{n,k,j}\lambda_j$, we get (A.4). □

For the variable population size function $\eta(t)$, a result of Griffiths and Tavaré [28] states that the entries of the expected SFS $\boldsymbol{\xi}_n$ for a sample of size n are proportional to linear combinations of $\mathbb{E}[T_{n,k}^{(\eta)}]$,

$$\xi_{n,b} = \frac{\sum_{k=2}^{n-b+1} k \frac{\binom{n-b-1}{k-2}}{\binom{n-1}{k-1}} \mathbb{E}[T_{n,k}]}{\sum_{k=2}^{n} k \mathbb{E}[T_{n,k}]}. \tag{A.16}$$

Note that $\mathbb{E}[T_{n,k}] = \mathbb{E}[\Sigma_{n,k}] - \mathbb{E}[\Sigma_{n,k+1}]$, and $\mathbb{E}[\Sigma_{n,k}]$ can be calculated using the pdf of $\Sigma_{n,k}$ given in (A.3),

$$\begin{aligned} \mathbb{E}[\Sigma_{n,k}] &= \sum_{j=k}^{n} a_{n,k,j} \int_0^\infty t \binom{j}{2} \frac{1}{\eta(t)} \exp\left(-\binom{j}{2} R_\eta(t) \right) dt \\ &= \sum_{j=k}^{n} a_{n,k,j} c_j, \end{aligned} \tag{A.17}$$

where c_j is the expected time to the first coalescence event for a sample of size j,

$$c_j = \int_0^\infty t \binom{j}{2} \frac{1}{\eta(t)} \exp\left(-\binom{j}{2} R_\eta(t) \right) dt.$$

However, this method of computing the expected SFS $\boldsymbol{\xi}_n$ for a given population size function η is not numerically stable for large sample sizes n [68]. This is because for fixed values of n and k, the terms $a_{n,k,j}$ have alternating signs in j, which causes numerical difficulties in the computation of $\mathbb{E}[\Sigma_{n,k}]$ using (A.17) for large n.

A.2 Expectation and standard deviation of the NLFT under the coalescent

We first establish some combinatorial identities, which will be used in the proofs of the expressions for the mean and variance of the NLFT under the coalescent stated in Section 4.6.5 of Chapter 4.

Claim 13. *For all $t, n \in \mathbb{Z}_+$,*

$$\sum_{k=0}^{n}(-1)^{n-k}\binom{n}{k}\binom{t+k-1}{t} = \binom{t-1}{n-1}.$$

Proof. Let $\Omega = \{(x_1, \cdots, x_n) : x_i \in \mathbb{Z}_{\geq 0}, \sum_{i=1}^{n} x_i = t,\}$, and define $A_i = \Omega \bigcap \{(x_1, \cdots, x_n) : x_i = 0\}$. The number of positive integer solutions to $x_1 + \ldots + x_n = t$ is equal to $\binom{t-1}{n-1}$. Hence,

$$\left|\bigcap_i A_i^c\right| = \binom{t-1}{n-1}.$$

By the inclusion-exclusion principle,

$$\begin{aligned}
\left|\bigcap_i A_i^c\right| &= |\Omega| + \sum_{\varnothing \subsetneq S \subseteq [n]} (-1)^{|S|} \left|\bigcap_{i\in S} A_i\right| \\
&= \sum_{j=0}^{n} \sum_{\substack{\varnothing \subseteq S \subseteq [n], \\ |S|=j}} (-1)^j \binom{t+(n-j)-1}{t} \\
&= \sum_{j=0}^{n} (-1)^j \binom{n}{j}\binom{t+(n-j)-1}{t} \\
\left|\bigcap_i A_i^c\right| &= \sum_{k=0}^{n} (-1)^{n-k}\binom{n}{k}\binom{t+k-1}{t},
\end{aligned}$$

where in the second equality, we used the fact that the number of non-negative integer solutions to $y_1 + \ldots + y_l = m$ is equal to $\binom{m+l-1}{m}$. □

Claim 14. *For $n \in \mathbb{Z}_+$,*

$$\sum_{k=1}^{n}(-1)^{n-k}k\frac{(k)_{(n-1)\uparrow}}{k!(n-k)!} = 1$$

Proof.

$$\begin{aligned}\sum_{k=1}^{n}(-1)^{n-k}k\frac{(k)_{(n-1)\uparrow}}{k!(n-k)!} &= \sum_{k=1}^{n}(-1)^{n-k}\left((n+k-1)-(n-1)\right)\frac{(k)_{(n-1)\uparrow}}{k!(n-k)!} \\ &= \sum_{k=1}^{n}(-1)^{n-k}\frac{(k)_{n\uparrow}}{k!(n-k)!} - (n-1)\sum_{k=1}^{n}(-1)^{n-k}\frac{(k)_{(n-1)\uparrow}}{k!(n-k)!}. \quad \text{(A.18)}\end{aligned}$$

Considering the first term of (A.18),

$$\begin{aligned}\sum_{k=1}^{n}(-1)^{n-k}\frac{(k)_{n\uparrow}}{k!(n-k)!} &= \sum_{k=0}^{n}(-1)^{n-k}\binom{n}{k}\binom{n+k-1}{n} \\ &= 1, \quad \text{(A.19)}\end{aligned}$$

where in the second equality, we have used Claim 13 with $t = n$. Considering the second term of (A.18),

$$\begin{aligned}(n-1)\sum_{k=1}^{n}(-1)^{n-k}\frac{(k)_{(n-1)\uparrow}}{k!(n-k)!} &= \frac{(n-1)}{n}\sum_{k=0}^{n}(-1)^{n-k}\binom{n}{k}\binom{n-1+k-1}{n-1} \\ &= 0, \quad \text{(A.20)}\end{aligned}$$

where in the second equality, we have used Claim 13 with $t = n-1$. □

Claim 15. *For $n \in \mathbb{Z}_+$,*

$$\sum_{k=1}^{n}(-1)^{n-k}k^2\frac{(k)_{(n-1)\uparrow}}{k!(n-k)!} = n^2 - n + 1$$

Proof.

$$\begin{aligned}&\sum_{k=1}^{n}(-1)^{n-k}k^2\frac{(k)_{(n-1)\uparrow}}{k!(n-k)!} \quad \text{(A.21)} \\ &= \sum_{k=1}^{n}(-1)^{n-k}\left((k+n)(k+n-1)+(1-2n)(k+n-1)+(n-1)^2\right)\frac{(k)_{(n-1)\uparrow}}{k!(n-k)!} \\ &= (1-2n)\sum_{k=1}^{n}(-1)^{n-k}\frac{(k)_{n\uparrow}}{k!(n-k)!} + (n-1)^2\sum_{k=1}^{n}(-1)^{n-k}\frac{(k)_{(n-1)\uparrow}}{k!(n-k)!} +\end{aligned}$$

$$+\sum_{k=1}^{n}(-1)^{n-k}\frac{(k)_{(n+1)\uparrow}}{k!(n-k)!} \tag{A.22}$$

The first term of (A.22) is equal to $(1-2n)$ by (A.19), and the second term of (A.22) is equal to 0 by (A.20). Considering the third term of (A.22),

$$\begin{aligned}\sum_{k=1}^{n}(-1)^{n-k}\frac{(k)_{(n+1)\uparrow}}{k!(n-k)!} &= (n+1)\sum_{k=0}^{n}(-1)^{n-k}\binom{n}{k}\binom{n+1+k-1}{n+1}\\ &= n(n+1),\end{aligned}$$

where in the second equality, we used Claim 13 with $t = n+1$. □

Theorem 16. *The expected number of surviving lineages at time t in the coalescent, given that there were n lineages at time 0, is given by,*

$$\mathbb{E}\left[A_n^C(t)\right] = \sum_{i=1}^{n} e^{-\binom{i}{2}R(t)}(2i-1)\frac{(n)_{i\downarrow}}{(n)_{i\uparrow}}. \tag{A.23}$$

Proof. We have the following expression for the probability distribution of $A_n^C(t)$ [83, 84],

$$\mathbb{P}[A_n^C(t)=m] = \sum_{i=m}^{n} e^{-\binom{i}{2}R(t)}(-1)^{i-m}\frac{(2i-1)(m)_{(i-1)\uparrow}(n)_{i\downarrow}}{m!(i-m)!(n)_{i\uparrow}}. \tag{A.24}$$

Computing $\mathbb{P}[A_n^C(t)=m]$ for large sample sizes using (A.24) will produce inaccurate values due to catastrophic cancellations. Instead of computing $\mathbb{E}\left[A_n^C(t)\right]$ by summing (A.24) over all values of m, we can compute $\mathbb{E}\left[A_n^C(t)\right]$ in a more numerically stable fashion by an interchange of summations as follows,

$$\begin{aligned}\mathbb{E}\left[A_n^C(t)\right] &= \sum_{m=1}^{n} m\ \mathbb{P}[A_n^C(t)=m] \qquad (A.25)\\ &= \sum_{m=1}^{n} m\left(\sum_{i=m}^{n} e^{-\binom{i}{2}R(t)}(-1)^{i-m}\frac{(2i-1)(m)_{(i-1)\uparrow}(n)_{i\downarrow}}{m!(i-m)!(n)_{i\uparrow}}\right)\\ &= \sum_{i=1}^{n} e^{-\binom{i}{2}R(t)}(2i-1)\frac{(n)_{i\downarrow}}{(n)_{i\uparrow}}\left(\sum_{m=1}^{i}(-1)^{i-m}m\frac{(m)_{(i-1)\uparrow}}{m!(i-m)!}\right)\\ \mathbb{E}\left[A_n^C(t)\right] &= \sum_{i=1}^{n} e^{-\binom{i}{2}R(t)}(2i-1)\frac{(n)_{i\downarrow}}{(n)_{i\uparrow}},\end{aligned}$$

where in the last equality, we have used the combinatorial identity in Claim 14. Since each term in the summation in (A.23) is positive, this expression poses no numerical problems. □

Theorem 17. *The variance of the number of surviving lineages at time t in the coalescent, given that there were n lineages at time 0, is given by,*

$$Var\left(A_n^C(t)\right) = \sum_{i=1}^{n} e^{-\binom{i}{2}R(t)}(2i-1)(i^2-i+1)\frac{(n)_{i\downarrow}}{(n)_{i\uparrow}} - \left(\sum_{i=1}^{n} e^{-\binom{i}{2}R(t)}(2i-1)\frac{(n)_{i\downarrow}}{(n)_{i\uparrow}}\right)^2. \tag{A.26}$$

Proof. By a similar interchange of summations as in Theorem 16, we can also calculate the variance of $A_n^C(t)$ as,

$$\begin{aligned}
\mathrm{Var}\left(A_n^C(t)\right) &= \mathbb{E}\left[A_n^C(t)^2\right] - \mathbb{E}\left[A_n^C(t)\right]^2, && \text{(A.27)}\\
\mathbb{E}\left[A_n^C(t)^2\right] &= \sum_{m=1}^{n} m^2 \left(\sum_{i=m}^{n} e^{-\binom{i}{2}R(t)}(-1)^{i-m}\frac{(2i-1)(m)_{(i-1)\uparrow}(n)_{i\downarrow}}{m!(i-m)!(n)_{i\uparrow}}\right)\\
&= \sum_{i=1}^{n} e^{-\binom{i}{2}R(t)}(2i-1)\frac{(n)_{i\downarrow}}{(n)_{i\uparrow}}\left(\sum_{m=1}^{i}(-1)^{i-m}m^2\frac{(m)_{(i-1)\uparrow}}{m!(i-m)!}\right)\\
\mathbb{E}\left[A_n^C(t)^2\right] &= \sum_{i=1}^{n} e^{-\binom{i}{2}R(t)}(2i-1)(i^2-i+1)\frac{(n)_{i\downarrow}}{(n)_{i\uparrow}}, && \text{(A.28)}
\end{aligned}$$

where we have used the combinatorial identity in Claim 15 in the last equality. Combining (A.28) with Claim 16 gives the stated expression for $\mathrm{Var}\left(A_n^C(t)\right)$. □