Theoretical Population Biology 100 (2015) 26–38

Contents lists available at ScienceDirect

Theoretical Population Biology

journal homepage: www.elsevier.com/locate/tpb

Can one hear the shape of a population history?☆

Junhyong Kim [a], Elchanan Mossel [b], Miklós Z. Rácz [b], Nathan Ross [c]

[a] University of Pennsylvania, United States
[b] University of California, Berkeley, United States
[c] University of Melbourne, Australia

ARTICLE INFO

Article history:
Received 1 April 2014
Available online 11 December 2014

Keywords:
Population size
Estimation
Coalescent

ABSTRACT

Reconstructing past population size from present day genetic data is a major goal of population genetics. Recent empirical studies infer population size history using coalescent-based models applied to a small number of individuals. Here we provide tight bounds on the amount of exact coalescence time data needed to recover the population size history of a single, panmictic population at a certain level of accuracy. In practice, coalescence times are estimated from sequence data and so our lower bounds should be taken as rather conservative.

1. Introduction

Reconstructing the past size and structure of the population of a species is a major goal of population genetics with applications in, for example, ecology, epidemiology (Heled and Drummond, 2008), and paleoanthropology (Li and Durbin, 2011). It is also important for understanding relationships between different evolutionary parameters, e.g., the dynamics of different parts of the genome or how demography affects selection (Li et al., 2012).

Inference is based on sequence data from individuals sampled from the population under consideration. Under a given population history, the coalescent is a model that provides likelihoods of observed genetic data and is one of the main tools used to infer population history. But the space of population histories typically considered is huge and so maximum likelihood estimation requires approximation techniques (Bhaskar et al., 2014; Excoffier et al., 2013; Harris and Nielsen, 2013; Li and Durbin, 2011; Nielsen, 2000; Palamara et al., 2012; Sheehan et al., 2013) which lack theoretical guarantees; the same statement applies to Bayesian methods (Drummond et al., 2005; Heled and Drummond, 2008). (These methods are discussed in greater detail in Section 1.1.)

In this paper we provide provable information-theoretic lower bounds on the amount of coalescence data needed to estimate, up to some specified accuracy, events in a population's past history (see Theorem 1.1). Our bounds are asymptotically tight as shown by analysis of a simple inference algorithm which recovers the history given slightly more data than required by the lower bounds.

Before stating our results in more detail, we provide a brief introduction to inference using the coalescent, as well as a summary of existing literature in this area.

1.1. Inference using the coalescent

Let $N(t)$ be the size of a single panmictic haploid population at time t "generations" in the past[1] and call $N = \{N(t)\}_{t\geqslant 0}$ the *shape of the population size history*, or simply the *population shape*. Given N, Kingman's coalescent (see Kingman, 1982b, Kingman, 1982a and Tavaré, 2004 for background) is a random genealogy on n sampled individuals from the present day population. The basic description is that the rate of coalescence between any two individuals/lineages at time t in the past is $1/N(t)$ and so given k lineages at time t, the rate of coalescence is $\binom{k}{2}/N(t)$. We focus primarily on the case where data comes from pairs of individuals, i.e., $n = 2$, just as in, e.g., Li and Durbin (2011).

The population shape $N = \{N(t)\}_{t\geqslant 0}$ determines a distribution $\mathbb{P}_N$ over coalescent trees; and in particular, $\mathbb{P}_N$ determines the distribution of coalescent trees of any finite number of individuals, at any number of independent loci. The first step to infer N using the coalescent is to ensure that the distribution over coalescent trees uniquely determines the shape of a population history, i.e., that $N \neq N'$ implies that $\mathbb{P}_N \neq \mathbb{P}_{N'}$. This is indeed true: if we know

☆ Our title is inspired by the famous paper of Kac (1966); analogously, we study the theoretical limits to inferring a population size history.

E-mail addresses: junhyong@sas.upenn.edu (J. Kim), mossel@stat.berkeley.edu (E. Mossel), racz@stat.berkeley.edu (M.Z. Rácz), nathan.ross@unimelb.edu.au (N. Ross).

http://dx.doi.org/10.1016/j.tpb.2014.12.002
0040-5809/

[1] Throughout the paper the unit of time is generations.

$\mathbb{P}_N$, then we also know the rate of coalescence of two arbitrary individuals at any time t, which is just $1/N(t)$. Thus with an infinite amount of coalescence time data, the population shape can be reconstructed.

Considering sequence data, the model assumes that for n individuals in a population, each genomic site follows an n-coalescent tree. Two sites have the same coalescent tree if there is no recombination breakpoint between them. At each site, mutations occur on top of the trees according to a Poisson process with small mutation rate and so, in principle, likelihoods of statistics of sequence data can be derived. Unfortunately, recombination is a complicated process and even under simplifying assumptions, likelihood functions are typically intractable, both analytically and computationally. Thus inexact methods must be developed, which we now describe.

Given whole genome data, likelihoods of various population parameters can be estimated across the parameter space by MCMC (Nielsen, 2000; Excoffier et al., 2013). Using a simplified model of recombination (McVean and Cardin, 2005), simpler likelihood functions arise; however, these must still be analyzed using approximation schemes (Li and Durbin, 2011; Sheehan et al., 2013). Sequence data can also be used to infer the lengths of nonrecombinant blocks (Tataru et al., 2014), the distribution of which can be used to infer various aspects of the population history (Palamara et al., 2012).

The problem simplifies when it is assumed that all loci in a given sequence are linked, that is, not separated by recombination events or in parts of the genome where it may be assumed that no recombination occurs (such as mitochondrial DNA). In such cases, the coalescent trees at each of the sites are *identical*. For such data, given the coalescent tree, the number of segregating sites (where mutations have occurred) follows a Poisson distribution and analytic (though intractable) expressions for likelihoods can be derived.

If all sites are unlinked, one can use inference tools involving the population allele frequency spectrum (Bhaskar et al., 2014) or Bayesian approaches such as the "Bayesian Skyline" (Drummond et al., 2005; Heled and Drummond, 2008) for both single and multi-locus data. Outside of the coalescent framework, the allele frequency spectrum and its diffusion approximation (Gutenkunst et al., 2009; Lukić et al., 2011) (which is derived from the underlying Wright–Fisher dynamics that also drive the coalescent) can also be used, though the same computational caveats as above apply. The allele frequency spectrum suffers from identifiability issues in general (Myers et al., 2008), though not under biologically realistic assumptions (Bhaskar and Song, 2014).

1.2. Results overview and applications

We provide lower bounds on the amount of *exact* coalescence time data necessary to infer past population history events. The assumption that our data are exact coalescence times is unrealistic but idealized: for a single, panmictic population, the rate of coalescence t generations in the past determines the population size at that time and so the most direct route to estimating the population history is through the coalescence times. Since our lower bounds on the amount of samples are for idealized data, the bounds should also be taken to apply to methods which use sequence data (and should be considered as underestimates for such methods). In fact, all the previously mentioned coalescent-based methods used to infer population history based on sequence data also infer the coalescence times along the way (usually implicitly).

The following theorem provides bounds on the probability of correctly distinguishing between two population histories that differ only on an interval $(T, T+S)$ over which each is constant, given coalescence times between pairs of individuals at L independent loci. See Fig. 1 for an illustration of two such histories.

Theorem 1.1. *Let a, b, and S be positive constants and let $T \geqslant 0$. Consider the following hypothesis testing problem: under both hypotheses the population sizes are equal in the intervals $[0, T)$ and $[T+S, \infty)$, given by some function $N(\cdot)$, but under H_1 the population size is constant $aN(0) =: aN_0$ in the interval $[T, T+S)$ while under H_2 the population size during the interval $[T, T+S)$ is constant bN_0. If L independent coalescence times are observed from either H_1 or H_2, with prior probability $1/2$, then the Bayes error rate for any classifier is at least $(1-\mathcal{E})/2$, where*

$$\mathcal{E}^2 \leqslant 2L \exp\left(-\int_0^T 1/N(t)\,dt\right)\left(1-e^{-\frac{S}{2N_0}\frac{a+b}{ab}}\right) \times \frac{\left(\sqrt{a}-\sqrt{b}\right)^2}{a+b} \tag{1.1}$$

$$\leqslant 2L \exp\left(-\int_0^T 1/N(t)\,dt\right) \min\left\{\frac{S}{2N_0}, \frac{ab}{a+b}\right\} \times \frac{\left(\sqrt{a}-\sqrt{b}\right)^2}{ab}. \tag{1.2}$$

In other words, for any classification procedure, the chance of correctly determining whether the samples came from H_1 or H_2, is at most $(1+\mathcal{E})/2$.

The main features of the bound of the theorem above are that if $L \ll (S/N_0)^{-1}$, or if $L \ll 1/(\sqrt{a}-\sqrt{b})^2$, or if T is large enough, then the chance of distinguishing between the two histories will be near $1/2$. Consequently, given a, b, S, and T, the theorem provides a lower bound on the number L of independent coalescence times necessary in order to distinguish between the two histories with a given probability.

To understand the bound in more concrete settings and to compare Theorem 1.1 to previous work, consider Li and Durbin (2011), who apply the pairwise sequentially Markovian coalescent model (PSMC) to the complete diploid genome sequences of seven individuals in order to infer human population size history, with one of their main goals being to infer the timing of the out-of-Africa event which caused a bottleneck in East Asian and European populations. To validate their model, they apply PSMC to simulated data where the population histories consist of a sharp out-of-Africa bottleneck followed by a population expansion. They note that the simulations "reveal a limitation of PSMC in recovering sudden changes in effective population size". We use Theorem 1.1 to quantitatively show that *every method* must suffer from this to a certain extent.

Take the population history considered in Li and Durbin (2011, Fig. 2(a)), reproduced in the left panel of Fig. 2. Here the present day effective population size is $N_0 := N(0) = 2.732 \times 10^4$; the effective population size is N_0 in the time interval $[0, 2.732 \times 10^4)$ back in time (measured in years, assuming 25 years per generation), it is $0.05 \times N_0$ in the time interval $[2.732 \times 10^4, 1.0245 \times 10^5)$ back in time, it is $0.5 \times N_0$ in the time interval $[1.0245 \times 10^5, 3.415 \times 10^6)$ back in time, and it is N_0 in the time interval $[3.415 \times 10^6, \infty)$ back in time. We apply Theorem 1.1 to obtain bounds on the amount of data needed to estimate the timing of the bottleneck at approximately 100 kyr to a given accuracy. We have $a = 0.05$, $b = 0.5$, and $N_0 = 2.732 \times 10^4$. Assuming 25 years per generation, we have $T = 0.15N_0$ and $\int_0^T 1/N(t)\,dt = \int_0^{0.04N_0} 1/N_0 dt + \int_{0.04N_0}^{0.15N_0} 1/(0.05N_0)dt = 2.24$.

Theorem 1.1 tells us that, given coalescence times from L independent loci, in order to distinguish between the two histories considered in Fig. 2 with probability at least 0.95, it is necessary that $\mathcal{E}^2 \geqslant 0.81$, so using (1.1) and plugging in the numbers above, it is necessary that

$$2Le^{-2.24}\left(1-e^{-\frac{S}{54640}\times\frac{0.55}{0.025}}\right)\frac{\left(\sqrt{0.5}-\sqrt{0.05}\right)^2}{0.55} \geqslant 0.81. \tag{1.3}$$

From this we immediately see that there is no solution for S when $L \leqslant 8$, i.e., when the number of independent loci is too small, the

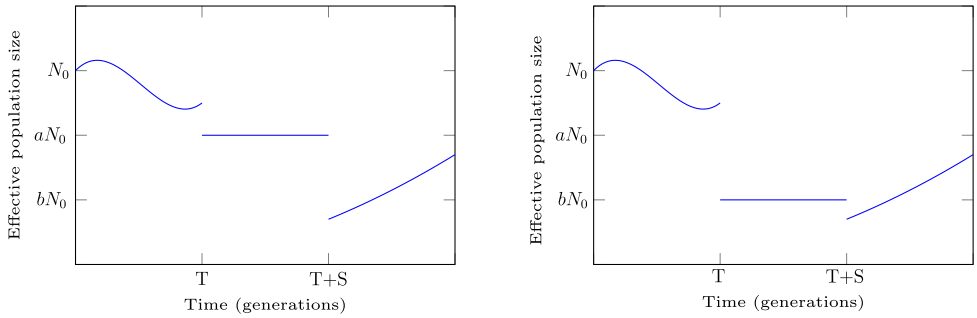


Fig. 1. An illustration of two population histories for which Theorem 1.1 provides lower bounds on the amount of coalescence time data needed to distinguish between them.

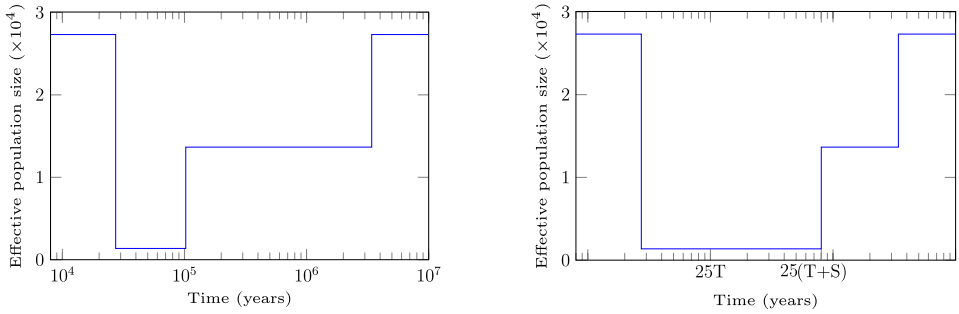


Fig. 2. The population histories compared in Table 1. The history on the left is Fig. 2(a) in Li and Durbin (2011) and the history on the right is a modified version. For each value of S, Theorem 1.1 provides lower bounds on the amount of coalescence time data needed to distinguish between the two histories.

Table 1
Lower bounds on the lengths of the 95% uncertainty intervals for determining the timing of a bottleneck given a sample of L independent loci in the scenario depicted in Fig. 2.

Number of loci	$\leqslant 8$	10	20	30	50
Interval length (yr)	∞	1.3×10^5	3.6×10^4	2.1×10^4	1.2×10^4

length of the 95% "uncertainty interval"[2] is infinite. When $L \geqslant 9$, (1.3) is equivalent to

$$S \geqslant \frac{27320}{11} \log\left(1 + \frac{\frac{891\times e^{2.24}}{4000\times\left(\sqrt{0.5}-\sqrt{0.05}\right)^2}}{L - \frac{891\times e^{2.24}}{4000\times\left(\sqrt{0.5}-\sqrt{0.05}\right)^2}}\right), \tag{1.4}$$

i.e., the length of the 95% "uncertainty interval" is inversely proportional to L for large L. Table 1 collects the numerical values of some of the lower bounds on the lengths of the 95% uncertainty intervals given by (1.3) and (1.4) (note that in (1.3) and (1.4) the unit of S is generations, while in Table 1 the unit of time is years, where we assume 25 years per generation). These estimates are in line with the simulation results of Li and Durbin (2011), where in the PSMC reconstruction of the population history the sudden drop in population is spread out over several tens of thousands of years.

Similarly, Theorem 1.1 also provides bounds on the amount of coalescence data needed to estimate the time of the final jump in the population to N_0 in this same scenario. We may consider two population histories, one the same as Fig. 2(a) in Li and Durbin (2011), and the other a modified version where the final jump in the population to N_0 occurs at some time in the interval $[1.0245 \times 10^5, 3.415 \times 10^6]$ years in the past; see Fig. 3.

[2] For a statistical hypothesis $H_1 : \theta = \theta_1$ on a real parameter θ and a given set of data, we define the 95% *uncertainty interval* to be the set of values θ_2 such that for any classification procedure, the chance of correctly determining whether the samples came from H_1 or $H_2 : \theta = \theta_2$ is no greater than 0.95.

Here we thus have $a = 0.5$, $b = 1$, $25(T+S) = 3.415 \times 10^6$, and $25T \in [1.0245 \times 10^5, 3.415 \times 10^6]$. For each such T we have

$$\begin{aligned}\int_0^T 1/N(t)dt &= \int_0^{.04N_0} 1/N_0 dt + \int_{.04N_0}^{.15N_0} 1/(.05N_0)dt \\ &\quad + \int_{.15N_0}^{T} 1/(.5N_0)dt \\ &= 1.94 + 2(T/N_0).\end{aligned}$$

Plugging these expressions into (1.2) of Theorem 1.1 and taking $ab/(a+b)$ in the minimum, we find that, in order to recover the true population history with probability at least 0.95 given L independent samples, we must have

$$T \leqslant 13660 \log(L/49.2). \tag{1.5}$$

Notice again that recovery with 95% chance is impossible when $L \leqslant 49$, so a considerable amount of coalescence time data is required for accurate inference. When $L \geqslant 50$, Table 2 summarizes lower bounds on the length of the 95% uncertainty intervals implied by (1.5) (again converted to years).

The PSMC reconstruction of the population history ends at approximately 5 Myr, so the lengths of their uncertainty intervals on the timing of the last population size change are unclear, but appear to be at least a few Myr. Our results are therefore in line with these simulation results, and show that no method can perform substantially better than PSMC.

1.3. Organization of the paper

The layout of the paper is as follows: in the next section we describe a procedure that infers a population history given slightly more data than required by the lower bounds implied by Theorem 1.1. We prove our results on lower bounds, including Theorem 1.1, in Section 3, and then prove results about the inference procedure in Section 4. We support our results by simulations presented in Section 5, and we end with a summarizing discussion section with some open problems.

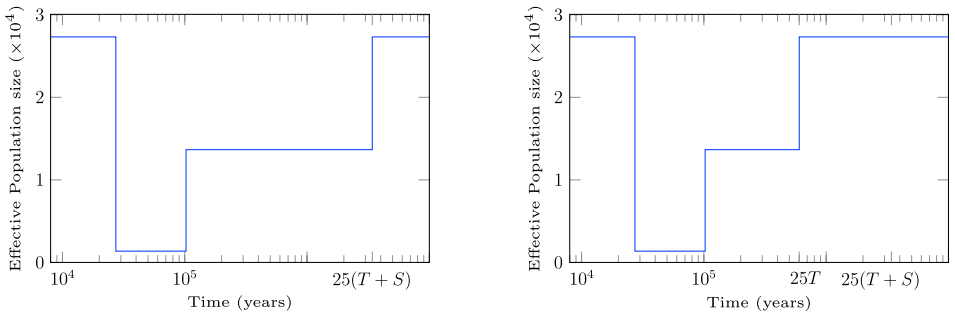


Fig. 3. The population histories compared in Table 2. The history on the left is Fig. 2(a) in Li and Durbin (2011) and the history on the right is a modified version. Here $25(T+S) = 3.415 \times 10^6$ and for each value of T, Theorem 1.1 provides lower bounds on the amount of coalescence data needed to distinguish between the two histories.

Table 2
Lower bounds on the lengths of the 95% uncertainty intervals for determining the timing of the last population size change in the scenario depicted in Fig. 3, given a sample of L independent loci.

Number of loci	100	200	500	10^3	10^4
Interval length (yr)	3.1×10^6	2.9×10^6	2.6×10^6	2.3×10^6	1.6×10^6

2. A simple reconstruction algorithm

To complement the results on lower bounds detailed above, we describe a simple estimation procedure (analyzed in Section 4) that takes coalescence time data and returns an estimate for the population shape. The analysis of this procedure shows that the amount of data it requires almost matches the lower bounds stated in our results above.

The procedure takes i.i.d. pairwise coalescence times $\mathbf{t}^L = \{t_1, \ldots, t_L\}$ and returns a piecewise constant estimate $\widehat{N} = \left\{\widehat{N}(t)\right\}_{t \geqslant 0}$ for the population shape. The procedure involves a single parameter, ε, which controls the length of the time intervals where our estimate is constant, and which then also affects the accuracy of our estimate in each time interval. Assume that there are N_0 individuals initially, at time 0.

1. Partition time backwards in time into intervals of length εN_0, i.e., let $I_1 = [0, \varepsilon N_0], I_2 = [\varepsilon N_0, 2\varepsilon N_0], \ldots, I_K = [(K-1)\varepsilon N_0, K\varepsilon N_0]$. ($K$ is the minimum integer such that the interval $[0, K\varepsilon N_0]$ covers the data and we do not provide estimates past time $K\varepsilon N_0$.)
2. For $k = 1, \ldots, K$, denote the fraction of data points lying in the time interval I_k by
$$\widehat{X}_k := \frac{1}{L} \#\{i : t_i \in I_k\},$$
and furthermore let $\widehat{S}_0 = 0$ and $\widehat{S}_k = \sum_{i=1}^{k} \widehat{X}_i$, the fraction of data points lying in the time interval $[0, k\varepsilon N_0]$.
3. Our estimate $\widehat{N}_k$ in the time interval I_k is
$$\widehat{N}_k := \frac{\varepsilon N_0}{-\log\left(1 - \frac{\widehat{X}_k}{1-\widehat{S}_{k-1}}\right)}, \tag{2.1}$$
provided that $\widehat{X}_k > 0$, i.e., we have at least one data point in the time interval I_k. If $\widehat{X}_k = 0$, then we do not give an estimate.

Remark 2.1. The estimate (2.1) is motivated by the fact that
$$\frac{\mathbb{P}(t_1 \in I_k)}{\mathbb{P}(t_1 \notin [0, (k-1)\varepsilon N_0])} = 1 - \exp\left(-\int_{(k-1)\varepsilon N_0}^{k\varepsilon N_0} \frac{1}{N(t)} dt\right).$$

Remark 2.2. In Step 1 above, we partition time into intervals of equal length. This is done solely to make the subsequent analysis and discussion as simple as possible. Depending on the specific application, it might be of interest to consider other choices of partitions, for instance, choosing intervals whose lengths grow exponentially backwards in time. Our estimation procedure (and also the subsequent analysis) works in an analogous way: $\widehat{X}_k$ and $\widehat{S}_k$ can be defined in the same way in Step 2, and the only change in the estimate (2.1) is to replace εN_0 in the numerator of the fraction with the length of the appropriate interval, $|I_k|$.

In order to state the properties of this procedure, define for $k \geqslant 1$ the "effective constant population size in the time interval I_k" by
$$\widetilde{N}_k := \frac{\varepsilon N_0}{\int_{(k-1)\varepsilon N_0}^{k\varepsilon N_0} \frac{1}{N(t)} dt};$$
the $\widetilde{N}_k$ give a natural piecewise constant approximation of the population shape $N(t)$ that is directly comparable to the piecewise estimate $\widehat{N}$. Let
$$E_k := \sup_{t \in I_k} \left|\log N(t) - \log \widehat{N}_k\right|$$
be the absolute error of our estimate $\widehat{N}$ on a logarithmic scale for each time interval. When estimating the error E_k, there are two types of errors to consider. One is the inherent error coming from the fact that we are approximating the shape with a piecewise constant function; the other error comes from the finite sample size L. By the triangle inequality we can bound the error E_k by the sum of these two errors:
$$E_k \leqslant E_{k,1} + E_{k,2},$$
where
$$E_{k,1} := \sup_{t \in I_k} \left|\log N(t) - \log \widetilde{N}_k\right|$$
is the error coming from approximating the shape in the time interval I_k with a constant, and
$$E_{k,2} := \left|\log \widetilde{N}_k - \log \widehat{N}_k\right|$$
is the error coming from the finite sample size. Ignoring the error $E_{k,1}$ for now, we can use concentration inequalities to derive the following finite sample estimate for the accuracy of our estimator:

Proposition 2.3. *Given that ℓ samples "survived" the first $k-1$ intervals, the probability that $\log \widetilde{N}_k$, is in the (random) interval*

$$\left[\log(\varepsilon N_0) - \log\left(-\log\left(\left(1 - \frac{L}{\ell}\widehat{X}_k - c\right) \vee 0\right)\right),\right.$$
$$\left.\log(\varepsilon N_0) - \log\left(-\log\left(\left(1 - \frac{L}{\ell}\widehat{X}_k + c\right) \wedge 1\right)\right)\right]$$

is at least $1 - 2\exp\left(-2c^2\ell\right)$ *for all* $c \geqslant 0$.[3]

Note that the interval in the proposition contains the estimate $\log \widehat{N}_k$.

To understand in what sense Proposition 2.3 and Theorem 1.1 are matching bounds, first consider the following easy corollary of Theorem 1.1 that better matches the setting of Proposition 2.3.

Theorem 2.4. *Let* a, b, *and* S *be positive constants and* $T \geqslant 0$. *Consider the following hypothesis testing problem: Under* H_1 *the population size is constant* $aN(0) =: aN_0$ *in the interval* $[T, T + S)$ *while under* H_2 *the population size during the interval* $[T, T + S)$ *is constant* bN_0. *Assume the data are independent coalescence times and let* ℓ *be the number of pairs that have not coalesced by time* T. *If the true history is given by either* H_1 *or* H_2, *each with prior probability* $1/2$, *then the Bayes error rate for any classifier is at least* $(1 - \Delta)/2$, *where* Δ *satisfies:*

$$\Delta^2 \leqslant 2\ell \frac{\left(\sqrt{b} - \sqrt{a}\right)^2}{a + b}.$$

In other words, for any classification procedure, the chance of correctly determining whether the samples came from H_1 *or* H_2, *is at most* $(1 + \Delta)/2$.

Writing $b = a(1 + \eta)$ for $\eta > 0$, the bound of the theorem becomes $2\ell(1 - 2\sqrt{1 + \eta}/(2 + \eta)) \leqslant \ell\eta^2/4$ and so we see that if $\eta \ll \ell^{-1/2}$, then *no procedure* will distinguish between the two histories given by H_1 and H_2 with good probability. On the other hand, Proposition 2.3 implies that for a fixed confidence α,

$$1 - 2\exp\left(-2c^2\ell\right) = \alpha,$$

and the constant c is of order $\ell^{-1/2}$ as ℓ becomes large, and thus the width of the interval in Proposition 2.3 is of order $\log(1 + C\ell^{-1/2})$ where C is some constant. To summarize, if $\eta \gg \ell^{-1/2}$, then our method will distinguish between the histories with high probability; but if $\eta \ll \ell^{-1/2}$, Theorem 2.4 shows that *no procedure* will distinguish between the two histories with good probability. On a conceptual level, this last statement is the main purpose of the paper: a significant amount of data is needed to infer past population size, especially in deep history where there is likely to be little coalescence information.

For illustration, we implement our estimation procedure on simulated data in Section 5, where we find a good general performance, matching our theoretical results.

2.1. Related theoretical work

Our reconstruction algorithm is a special case of the following problem: given n i.i.d. copies of the *first point* of a Poisson point process on $[0, \infty)$ with intensity $\varphi(t)$, what is a good estimate of $\varphi(t)^{-1}$? Poisson process intensity estimation has a large literature, see for example Birgé (2007), Reynaud-Bouret (2003), Willett and Nowak (2007) and references therein, but the (natural) data assumed in this area is one realization of the point process, or the point process observed up to some fixed time, or i.i.d. copies of such data, which does not fit our framework.

For another perspective to this question, define the hazard rate for a positive random variable X with density f and distribution function F to be

$$-\frac{d}{dt}\log(1 - F(t)) = f(t)/(1 - F(t)). \tag{2.2}$$

A simple calculation shows that the time of the first point of a Poisson process with intensity $\varphi(t)$ has the same distribution as a positive random variable with hazard rate $\varphi(t)$. Due largely to their importance in applications in, e.g., insurance, medicine, and reliability theory (Lawless, 2003, Section 1.1), hazard rate estimation is well studied; some seminal papers are Rice and Rosenblatt (1976), Sethuraman and Singpurwalla (1981), Yandell (1983) and see the recent Cheng et al. (2006) and references therein. Without embellishments specific to lifetime data (such as censoring where some lifetimes are only known to be at least some value), the main technique to estimating (2.2) (which also applies to its inverse) is to adapt estimators of f and F.

Indeed, our reconstruction algorithm is essentially an adaptation of the histogram estimate of the density and distribution function to our setting. Other popular density estimation techniques such as those in the introduction of Silverman (1986) can be adapted to our setting through the use of (2.2); for example see Wang (2005) for a survey of kernel smoothing methods for hazard function estimation. Our particular estimation procedure was chosen due to its simplicity and explicitness; in particular, we mention two points. The first is that we desire results like Proposition 2.3 with explicit non-asymptotic confidence intervals. Asymptotic confidence intervals can be obtained and used as estimates for smoothed density estimators, but with error depending on unknown quantities related to the underlying density which can lead to poor coverage accuracy (Hall, 1992). Secondly, smoothed density estimators have improved performance only when the underlying density is itself smooth (expressed as differentiability and continuity conditions). A major purpose of estimating past population size is to discover drastic changes in population size such as bottlenecks (Harris and Nielsen, 2013; Li and Durbin, 2011; Palamara et al., 2012; Sheehan et al., 2013), when it is not clear such smoothness assumptions are appropriate.

3. Proof of lower bounds

In this section we prove Theorem 1.1, as well as derive some other lower bounds for the amount of data needed for a given accuracy of estimating the population shape. This is done by formulating hypothesis tests deciding between two population shapes, and proving upper bounds on the probability of correctly inferring the population shape.

3.1. Background on probability metrics

We first recall a few metrics between probability distributions (see Gibbs and Su, 2002 for a survey). Let P and Q be two probability measures that are absolutely continuous with respect to a third probability measure λ. Write $f_P = \frac{dP}{d\lambda}$ and $f_Q = \frac{dQ}{d\lambda}$ for the respective Radon–Nikodym derivatives. The square of the Hellinger distance between P and Q is then defined as

$$d_H^2(P, Q) := \frac{1}{2}\int\left(\sqrt{f_P} - \sqrt{f_Q}\right)^2 d\lambda.$$

The definition does not depend on the choice of λ. A nice property of the Hellinger distance is that for product measures $P = P_1 \times P_2$, $Q = Q_1 \times Q_2$, we have that

$$1 - d_H^2(P, Q) = \left(1 - d_H^2(P_1, Q_1)\right)\left(1 - d_H^2(P_2, Q_2)\right),$$

which immediately implies that

$$d_H^2(P, Q) \leqslant d_H^2(P_1, Q_1) + d_H^2(P_2, Q_2).$$

Another commonly used metric is the total variation distance:

$$d_{TV}(P, Q) := \sup_{A \in \mathcal{F}} |P(A) - Q(A)|,$$

[3] Here and in the following we use the notation $a \vee b = \max\{a, b\}$ and $a \wedge b = \min\{a, b\}$.

or, equivalently:

$$d_{TV}(P,Q)=\frac{1}{2}\int\left|f_P-f_Q\right|d\lambda.$$

We use the following well-known fact:

Lemma 3.1. *With the notation above we have* $d_{TV}\leqslant\sqrt{2}d_H$.

Proof. This follows from the identity

$$f_P-f_Q=\left(\sqrt{f_P}-\sqrt{f_Q}\right)\left(\sqrt{f_P}+\sqrt{f_Q}\right),$$

the Cauchy–Schwarz inequality, and the inequality

$$\left(\sqrt{f_P}+\sqrt{f_Q}\right)^2\leqslant 2\left(f_P+f_Q\right). \quad \square$$

3.2. A lower bound on the amount of data needed to recover a constant history

We start in a simpler setting than Theorem 1.1 where we are trying to differentiate with good probability between two populations of *constant* size. For this simple setup we assume our data are L i.i.d. copies of coalescent trees on n individuals from a constant population, and we want to estimate the size of the population. We derive lower bounds on the amount of data needed for recovery.

Theorem 3.2. *Consider the following hypothesis testing problem: H_1 states that the population size during the interval $[0,\infty)$ is constant N, while H_2 states that the population size during the interval $[0,\infty)$ is the constant $(1+\eta)N$, where $\eta>0$ is fixed. If L i.i.d. coalescent trees on n individuals are observed from either H_1 or H_2, each with prior probability $1/2$, then the Bayes error rate for any classifier is at least $(1-\Upsilon)/2$, where Υ satisfies:*

$$\Upsilon^2\leqslant 2L\left(1-\left(\frac{2\sqrt{1+\eta}}{2+\eta}\right)^{n-1}\right)\leqslant\frac{L(n-1)\eta^2}{4}.$$

In other words, for any classification procedure, the chance of correctly determining whether the samples come from H_1 or H_2, is at most $(1+\Upsilon)/2$.

The interpretation of the theorem is that if $\eta\ll(nL)^{-1/2}$, then *no procedure* will distinguish between the two histories given by H_1 and H_2 with good probability. In other words, we need $L=\Omega\left(1/\left(n\eta^2\right)\right)$ samples to differentiate between the two histories N_1 and N_2.[4] We reiterate that these bounds hold knowing exact rather than estimated coalescence times, and so should be considered as underestimates in more realistic data settings.

We set up for the proof of Theorem 3.2; the same paradigm will be used to prove Theorem 1.1. Consider the following hypothesis testing problem. Let $\eta>0$, and let $N_1(\cdot)\equiv N$ and $N_2(\cdot)\equiv(1+\eta)N$ be two population size histories. Let κ be uniform in $\{1,2\}$, and, given κ, let $\mathbf{R}^{\kappa,L}=\left\{R_1^\kappa,\ldots,R_L^\kappa\right\}$ be a collection of L i.i.d. coalescence trees on n individuals drawn from the distribution induced by the population size history N_κ. The problem is to infer κ from $\mathbf{R}^{\kappa,L}$.

The probability of correctly inferring κ using the optimal reconstruction strategy is clearly at least $1/2$; denote this probability by $(1+\Upsilon)/2$ (here $\Upsilon=\Upsilon(L,n,N,\eta)$). The reconstruction method which gives the largest probability of correctly inferring κ is maximum likelihood: let $\widehat{\kappa}=1$ if $\mathbb{P}\left(\kappa=1\,\middle|\,\mathbf{R}^{\kappa,L}\right)\geqslant\mathbb{P}\left(\kappa=2\,\middle|\,\mathbf{R}^{\kappa,L}\right)$ and $\widehat{\kappa}=2$ otherwise. Then we have

$$\Upsilon=\mathbb{P}(\widehat{\kappa}=\kappa)-\mathbb{P}(\widehat{\kappa}\neq\kappa)=d_{TV}\left(\mathbf{R}^{1,L},\mathbf{R}^{2,L}\right).$$

Proof of Theorem 3.2. By the facts in Section 3.1 we have

$$\begin{aligned}\Upsilon(L,n,N,\eta)^2&=d_{TV}^2\left(\mathbf{R}^{1,L},\mathbf{R}^{2,L}\right)\leqslant 2d_H^2\left(\mathbf{R}^{1,L},\mathbf{R}^{2,L}\right)\\&\leqslant 2Ld_H^2\left(R_1^1,R_1^2\right).\end{aligned}\tag{3.1}$$

Since for $i\in\{1,2\}$ the increasing sequence of times of coalescence of the trees R_1^i, denoted by $\mathbf{s}^i=(s_1^i,\ldots,s_{n-1}^i)$, are sufficient statistics for R_1^i, we have $d_H^2\left(R_1^1,R_1^2\right)=d_H^2(\mathbf{s}^1,\mathbf{s}^2)$. We can directly compute the density $f_i(\mathbf{x})$ of $\mathbf{s}^i$ as

$$\begin{aligned}f_i(\mathbf{x})=&\prod_{j=1}^{n-1}\exp\left\{-\binom{n-j+1}{2}(x_j-x_{j-1})/N_i\right\}\\&\times\frac{\binom{n-j+1}{2}}{N_i}\mathbf{1}_{\{0<x_1<\cdots<x_{n-1}\}},\end{aligned}$$

where we have set $N_1:=N$, $N_2:=(1+\eta)N$, and $x_0=0$. Using these densities in the definition of the Hellinger distance and noting especially that since f_i is a density, we have for any $\alpha>0$ that

$$\begin{aligned}&\int_{0<x_1<\cdots<x_{n-1}}\prod_{j=1}^{n-1}\exp\left\{-\alpha\binom{n-j+1}{2}(x_j-x_{j-1})\right\}\\&\times\binom{n-j+1}{2}d\mathbf{x}=\frac{1}{\alpha^{n-1}},\end{aligned}$$

a calculation shows that

$$d_H^2\left(s_1^1,s_1^2\right)=1-\left(\frac{2\sqrt{1+\eta}}{2+\eta}\right)^{n-1}.\tag{3.2}$$

Plugging this into (3.1) yields the first bound of the result. When $\eta>0$, we can upper bound the right hand side of (3.2) by $(n-1)\eta^2/8$ to get the simpler bound $\Upsilon^2\leqslant L(n-1)\eta^2/4$. $\square$

3.3. Proof of Theorem 1.1

We prove Theorem 1.1 using the same strategy as that of Section 3.2. For $i=1,2$, let $N_i(\cdot)$ be the history corresponding to hypothesis H_i. Let κ be uniform in $\{1,2\}$, and, given κ, let $\mathbf{t}^{\kappa,L}=\left\{t_1^\kappa,\ldots,t_L^\kappa\right\}$ be a collection of L i.i.d. coalescence times of pairs of individuals drawn from the distribution induced by the population size history N_κ. The problem is to infer κ from $\mathbf{t}^{\kappa,L}$.

Proof of Theorem 1.1. As above, the chance that we infer κ correctly from $\mathbf{t}^{\kappa,L}$ is bounded above by $(1+\mathcal{E}(L,a,b,T,S))/2$ where

$$\mathcal{E}(L,a,b,T,S)^2=d_{TV}^2\left(\mathbf{t}^{1,L},\mathbf{t}^{2,L}\right)\leqslant 2Ld_H^2\left(t_1^1,t_1^2\right).\tag{3.3}$$

Writing $a_1:=a$ and $a_2:=b$ to shorten formulas, a straightforward calculation shows that the density of t_1^i is

$$f_i(x)=\begin{cases}\exp\left(-\int_0^x\frac{1}{N(s)}ds\right)\frac{1}{N(x)}, & x<T,\\ \exp\left(-\int_0^T\frac{1}{N(s)}ds\right)\exp\left(-\frac{x-T}{a_iN_0}\right)\frac{1}{a_iN_0}, & T\leqslant x<T+S,\\ \exp\left(-\int_0^T\frac{1}{N(s)}ds\right)\exp\left(-\frac{S}{a_iN_0}\right)\\ \quad\times\exp\left(-\int_{T+S}^x\frac{1}{N(s)}ds\right)\frac{1}{N(x)}, & T+S\leqslant x.\end{cases}$$

[4] We use the standard asymptotic notation Ω, which means "at least on the order of". Formally, if a_n and b_n are two sequences such that there exists a positive constant c and an integer n_0 such that for every $n\geqslant n_0$, $a_n\geqslant c\times b_n$, then $a_n=\Omega(b_n)$ as $n\to\infty$. Similarly, if f and g are two functions such that there exist positive constants c and x_0 such that for every $x\in(0,x_0)$, $f(x)\geqslant c\times g(x)$, then $f(x)=\Omega(g(x))$ as $x\searrow 0$. Equivalently, $f(x)=\Omega(g(x))$ if and only if $g(x)=O(f(x))$.

Using these densities in the definition of the Hellinger distance, we find after some simple calculations that

$$d_H^2\left(t_1^1, t_1^2\right) = \exp\left(-\int_0^T 1/N\left(t\right) dt\right)\left(1 - e^{-\frac{S}{2N_0}\frac{a+b}{ab}}\right) \times \frac{\left(\sqrt{a}-\sqrt{b}\right)^2}{a+b}.$$

Finally, using the inequality $1 - e^{-x} \leqslant \min\{x, 1\}$, simplifying, and plugging the result into (3.3) implies the bound of the theorem. □

4. Estimating the population shape

Recall our setting of the estimation procedure for the population shape $N = \{N(t)\}_{t\geqslant 0}$ from the i.i.d. coalescence times $\mathbf{t}^L = \{t_1, \ldots, t_L\}$. In this section we analyze our piecewise constant population shape estimator $\widehat{N} = \left\{\widehat{N}(t)\right\}_{t\geqslant 0}$ introduced in Section 2.

We consider the absolute error of our estimate $\widehat{N}$ on a logarithmic scale for each time interval, i.e., for $k \geqslant 1$ we consider

$$E_k := \sup_{t\in I_k}\left|\log N(t) - \log \widehat{N}_k\right|.$$

Recall also that $E_k \leqslant E_{k,1} + E_{k,2}$, where $E_{k,1}$ and $E_{k,2}$ are also both defined in Section 2.

To bound the error $E_{k,1}$ it is necessary to make an additional assumption on the population shape. We introduce an additional parameter, δ, which controls how much the population size can vary within a time interval, and we make the following assumption.

Assumption 1. We assume that in each time interval the population size can increase by a factor of at most $e^{\delta\varepsilon}$, and can decrease by a factor of at most $e^{-\delta\varepsilon}$.

Using this assumption, it is simple to bound the first type of error.

Lemma 4.1. *Given Assumption 1, we have that $E_{k,1} \leqslant 2\delta\varepsilon$.*

Proof. Assumption 1 implies that

$$\log N\left((k-1)\varepsilon N_0\right) - \delta\varepsilon \leqslant \log N\left(t\right) \leqslant \log N\left((k-1)\varepsilon N_0\right) + \delta\varepsilon$$

for all $t \in I_k$, and consequently also that

$$\log N\left((k-1)\varepsilon N_0\right) - \delta\varepsilon \leqslant \log \widetilde{N}_k \leqslant \log N\left((k-1)\varepsilon N_0\right) + \delta\varepsilon.$$

These inequalities then imply that $E_{k,1} \leqslant 2\delta\varepsilon$. □

To estimate the second type of error, $E_{k,2}$, it is not necessary to make any assumptions. We use concentration results for sums of i.i.d. random variables, and, in particular, we use the following simple corollary of the Chernoff bound.

Theorem 4.2. *Let $Y_1, \ldots, Y_n$ be i.i.d. Bernoulli (p) random variables, and let $Y = \sum_{i=1}^n Y_i$. Then for any $\lambda > 0$ we have*

$$\left.\begin{matrix}\mathbb{P}\left(Y \leqslant np - \lambda\right)\\ \mathbb{P}\left(Y \geqslant np + \lambda\right)\end{matrix}\right\} \leqslant \exp\left(-\frac{2\lambda^2}{n}\right). \tag{4.1}$$

The bounds in Theorem 4.2 imply the following concentration bound.

Corollary 4.3. *For any $k \geqslant 1$ and $\lambda > 0$ we have*

$$\mathbb{P}\left(\left|\widehat{X}_k - \mathbb{E}\left(\widehat{X}_k\right)\right| \geqslant \lambda\right) \leqslant 2\exp\left(-2\lambda^2 L\right). \tag{4.2}$$

In the following we present two bounds on the error $E_{k,2}$. We first present a bound for the first interval (i.e., when $k = 1$), which then also implies conditional bounds for general intervals, by conditioning on the number of data points that have not coalesced by a given time.

4.1. Bounds for the first interval

Proposition 4.4. *For any $c \geqslant 0$, with probability at least $1 - 2\exp\left(-2c^2L\right)$, the logarithm of the effective constant population size in I_1, $\log \widetilde{N}_1$, is in the interval*

$$\left[\log\left(\varepsilon N_0\right) - \log\left(-\log\left(\left(1 - \widehat{X}_1 - c\right)\vee 0\right)\right), \log\left(\varepsilon N_0\right) - \log\left(-\log\left(\left(1 - \widehat{X}_1 + c\right)\wedge 1\right)\right)\right]. \tag{4.3}$$

Note that the interval in (4.3) is an interval around our estimate $\log \widehat{N}_1$.

Proof. The inequality (4.2) for $k = 1$ can be rephrased as

$$\mathbb{P}\left(\mathbb{E}\widehat{X}_1 \in \left[\left(\widehat{X}_1 - c\right)\vee 0, \left(\widehat{X}_1 + c\right)\wedge 1\right]\right) \geqslant 1 - 2\exp\left(-2c^2L\right).$$

By algebraic manipulation, $\mathbb{E}\widehat{X}_1 \in \left[\left(\widehat{X}_1 - c\right)\vee 0, \left(\widehat{X}_1 + c\right)\wedge 1\right]$ is equivalent to $\log \widetilde{N}_1$ being contained in the interval in (4.3). □

This bound is useful because we can immediately determine a confidence interval for our estimate. To achieve a confidence level of $1 - \alpha$, we can choose $c = c(\alpha, L)$ to satisfy $2\exp\left(-2c^2L\right) = \alpha$, i.e., choose

$$c = \sqrt{\frac{\log\left(2/\alpha\right)}{2L}}. \tag{4.4}$$

Then the interval in (4.3) with c given by (4.4) has a confidence level of $1 - \alpha$.

4.2. Conditional bounds

Next, we present conditional bounds: given the number of samples that did not coalesce in the time interval $[0, (k-1)\varepsilon N_0]$, what is the error we make when estimating the population size in the time interval I_k? The following result is the same as Proposition 2.3 but worded more precisely.

Proposition 4.5. *For any $c \geqslant 0$, the probability conditioned on*

$$L\left(1 - \widehat{S}_{k-1}\right) = \ell$$

(i.e., that ℓ samples "survived" the first $k-1$ intervals) that the logarithm of the effective constant population size, $\log \widetilde{N}_k$, is in the interval

$$\left[\log\left(\varepsilon N_0\right) - \log\left(-\log\left(\left(1 - \frac{L}{\ell}\widehat{X}_k - c\right)\vee 0\right)\right), \log\left(\varepsilon N_0\right) - \log\left(-\log\left(\left(1 - \frac{L}{\ell}\widehat{X}_k + c\right)\wedge 1\right)\right)\right] \tag{4.5}$$

is at least $1 - 2\exp\left(-2c^2\ell\right)$.

Note that the interval in (4.5) is an interval around our estimate $\log \widehat{N}_k$, given $L\left(1 - \widehat{S}_{k-1}\right) = \ell$.

Proof. Let $\widehat{Y}_k := \frac{L}{\ell}\widehat{X}_k$. Given $L\left(1 - \widehat{S}_{k-1}\right) = \ell$, $\widehat{Y}_k$ is the average of ℓ i.i.d. indicator variables. Therefore Chernoff's bound gives that

$$\mathbb{P}\left(\left|\widehat{Y}_k - \mathbb{E}\widehat{Y}_k\right| \geqslant c \,\middle|\, L\left(1 - \widehat{S}_{k-1}\right) = \ell\right) \leqslant 2\exp\left(-2c^2\ell\right).$$

In other words,

$$\mathbb{P}\left(\mathbb{E}\widehat{Y}_k \in \left[\left(\widehat{Y}_k - c\right)\vee 0, \left(\widehat{Y}_k + c\right)\wedge 1\right] \,\middle|\, L\left(1 - \widehat{S}_{k-1}\right) = \ell\right) \geqslant 1 - 2\exp\left(-2c^2\ell\right).$$

Just as in the proof of Proposition 4.4, by algebraic manipulation, given $L\left(1 - \widehat{S}_{k-1}\right) = \ell$, $\mathbb{E}\widehat{Y}_k \in \left[\left(\widehat{Y}_k - c\right)\vee 0, \left(\widehat{Y}_k + c\right)\wedge 1\right]$ is equivalent to $\log \widetilde{N}_k$ being contained in the interval in (4.5). □

Again, this bound is useful because we can immediately determine a confidence interval for our estimate. To achieve a con-

fidence level of $1-\alpha$, we can choose $c = c(\alpha, \ell)$ to satisfy $2\exp\left(-2c^2\ell\right) = \alpha$, i.e., choose

$$c = \sqrt{\frac{\log(2/\alpha)}{2\ell}}. \tag{4.6}$$

Then the interval in (4.5) with c given by (4.6) has a confidence level of $1-\alpha$.

5. Simulations

We illustrate our estimation procedure on simulated data for the following settings: (1) constant size population, (2) piecewise constant size population, and (3) a population experiencing recent exponential growth; the last setting being germane to recent human population history (see, e.g., Tennessen et al., 2012 for a study on how recent accelerated population growth, together with weak purifying selection, can lead to an excess of rare functional variants). In each case, we simulate L independent pairwise coalescence times and apply our estimation procedure described in Section 4 with a given ε to the data; the outcome is summarized in Figs. 4–7. Each figure plots $\log(N(t)/N(0))$ versus $t/N(0)$, i.e., we scale time according to the coalescent timescale, and we plot the population size on a logarithmic scale.

The true history is the blue line, the estimates over each interval are the red lines, and the confidence intervals at the 95% level are given in pink. Recall that the logarithm of our estimate (2.1) is

$$\log\widehat{N}_k = \log(\varepsilon N_0) - \log\left(-\log\left(1-\frac{\widehat{X}_k}{1-\widehat{S}_{k-1}}\right)\right),$$

and our confidence interval for confidence level $1-\alpha$ is given by (4.5):

$$\left[\log(\varepsilon N_0) - \log\left(-\log\left(\left(1-\frac{\widehat{X}_k}{1-\widehat{S}_{k-1}} - c\right)\vee 0\right)\right),\right.$$
$$\left.\log(\varepsilon N_0) - \log\left(-\log\left(\left(1-\frac{\widehat{X}_k}{1-\widehat{S}_{k-1}} + c\right)\wedge 1\right)\right)\right],$$

where

$$c = \sqrt{\frac{\log(2/\alpha)}{2L\left(1-\widehat{S}_{k-1}\right)}}.$$

In the case that $\widehat{X}_k = 0$ we do not give an estimate but can sometimes still obtain a lower bound on the confidence interval. The intervals where the pink extends to the upper or lower margin of the graphing area represent a confidence bound that is infinite or zero, i.e., where the minimum or maximum are taken to be one or zero in the expressions for the confidence bounds above.

The error $E_{k,1}$ is not represented in the plots, but would add δ to each side of each confidence interval. Alternatively, we can view our statistic as an estimate of the effective constant population size over the given interval.

The plots in Fig. 7 compare our lower bounds to the confidence intervals of our estimation procedure. The black lines are the 95% uncertainty intervals: in a given time interval, populations within the interval given by the black lines cannot be distinguished from the true blue line population with the amount of data in hand with probability 0.95. Thus if the red line is within the interval given by the black lines, then our estimate is in some sense the best that can be achieved. Note that when an interval has no upper black line, then there is not enough data to distinguish between a history of the blue line size in the interval and *any* larger size population with 95% certainty.

Even though our assumed data is idealized and unrealistic (exact coalescent data at thousands of independent sites), these simulations allow us to make some general qualitative observations. The major determining factor of the performance of our procedure in a time interval is the number of coalescence times that have survived to that interval. So having more data (i.e., larger L) leads to more accurate estimates, and the estimates lose accuracy moving back in time as the number of data points decreases. Moreover, there is a rare event effect when there are few coalescences in an interval – having no coalescences in an interval is not very informative – and this leads to the consistent underestimates in the deepest part of the histories. Possibly this effect would be lessened by lengthening the widths of the intervals as they go back in time, although this would smooth out big features in the history. Also note that in the presence of a bottleneck, i.e., a time period where the population becomes small, there are many coalescences due to the increased rate. In turn this decreases the number of available data deeper in history. For example, compare the accuracy of the estimates at time $t/N(0) = 4$ in the constant population of Fig. 4 to that of the piecewise constant population of Fig. 5 where there is a bottleneck starting around time $t/N(0) = 1$. Finally, note that in Fig. 7 the confidence intervals of our procedure and the lower bound black lines are rather tight in the presence of a significant amount of data, but loosen as the number of data points decreases.

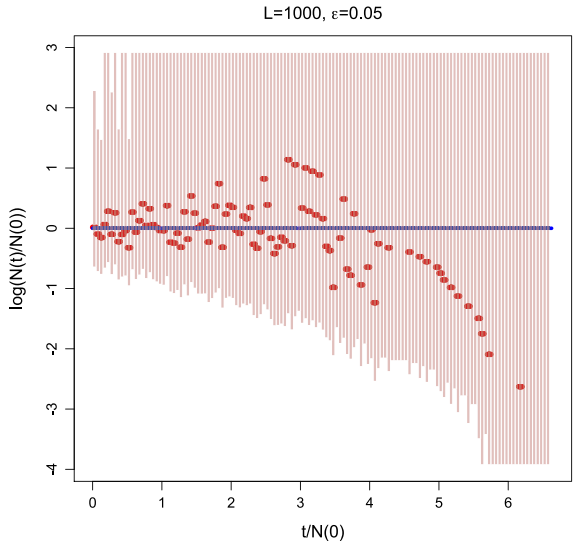
6. Discussion

An assortment of methods has been developed to infer a population's history from (an ever increasing amount of) genetic data (Bhaskar et al., 2014; Drummond et al., 2005; Excoffier et al., 2013; Harris and Nielsen, 2013; Heled and Drummond, 2008; Li and Durbin, 2011; Nielsen, 2000; Palamara et al., 2012; Sheehan et al., 2013). These methods are necessarily computational and approximate and so the quality of the outputs of these methods cannot be rigorously justified. However, understanding the theoretical limitations of inferring past population history (Bhaskar and Song, 2014; Myers et al., 2008) is of the utmost importance, since it is becoming increasingly common for such analyses to be used as the main tool for inference, with less emphasis on external verification (e.g., fossil or paleontological record) (Bos et al., 2014).

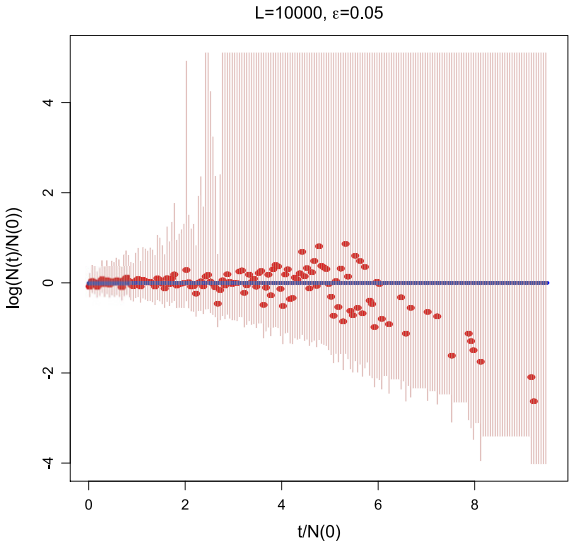
Here we have provided lower bounds on the amount of idealized data needed to infer a population history to a given accuracy. Our bounds should be considered as underestimates of the amount of data necessary for inference in methods which use sequence data, so they can be used as a guide when performing such analyses. We end with some further avenues of study and open problems.

6.1. Open problems

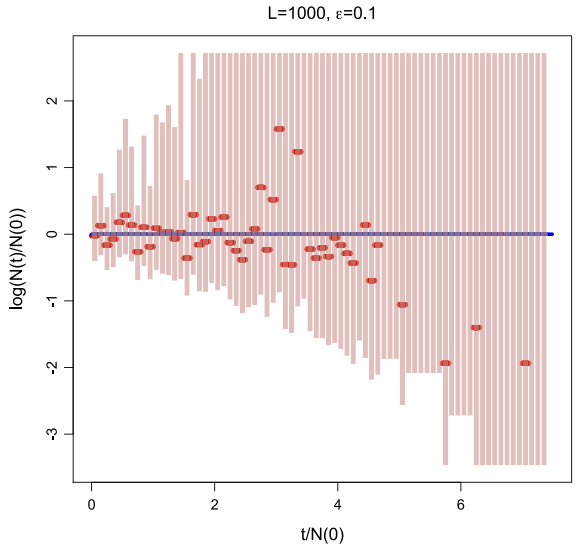
***n*-coalescence trees**. With the exception of Theorem 3.2, we assume that our data are L i.i.d. coalescence times between pairs of individuals. If instead our data are L i.i.d. coalescence trees among n individuals then how does this affect the bounds? In the setting of Theorem 3.2 when comparing two constant populations, increasing the number of individuals n in the coalescent tree is as good as increasing the number of independent loci L. This should not hold true in general since adding individuals does not greatly increase the depth of the tree. For moderate values of n, we expect that estimates of the deep history will not be greatly affected since the time of coalescence for the final two lineages is roughly half the length of the coalescent tree started from *infinitely* many individuals. On the other hand, explosive growth in the near history should be estimated better using more individuals since the amount of coalescing in the near past will increase. It would be interesting to better understand how increasing the number of individuals in the tree affects the lower bounds of Section 1.2 and the upper bounds provided by some generalization of our inference algorithm of Section 2. For some discussion on the affect of increasing the size of the tree versus increasing independent loci, see Heled and Drummond (2008).



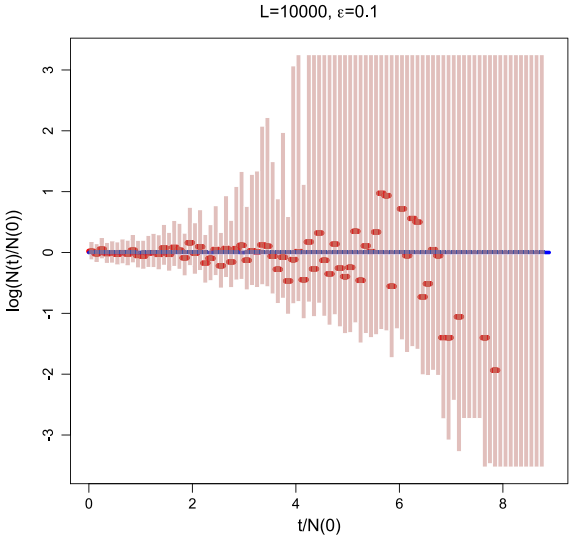
(a) $L = 10^3$, $\varepsilon = 0.05$.



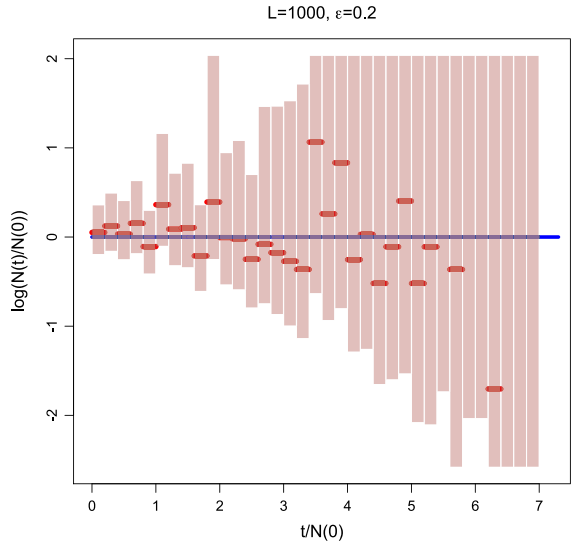
(b) $L = 10^4$, $\varepsilon = 0.05$.



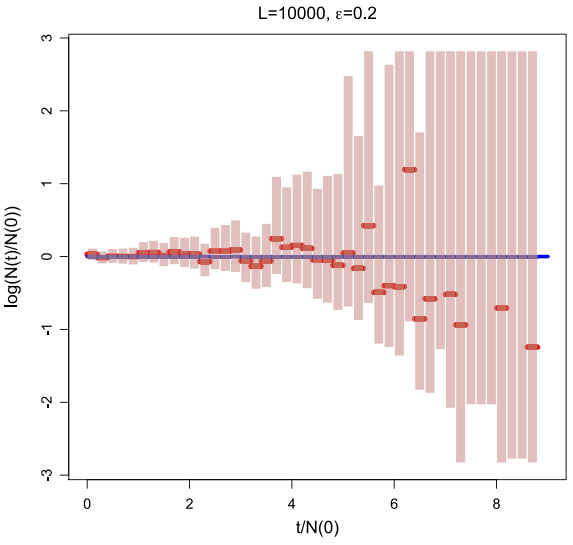
(c) $L = 10^3$, $\varepsilon = 0.1$.



(d) $L = 10^4$, $\varepsilon = 0.1$.

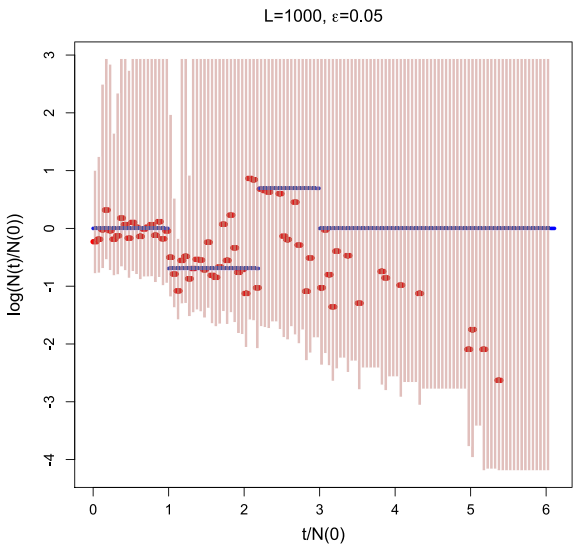


(e) $L = 10^3$, $\varepsilon = 0.2$.

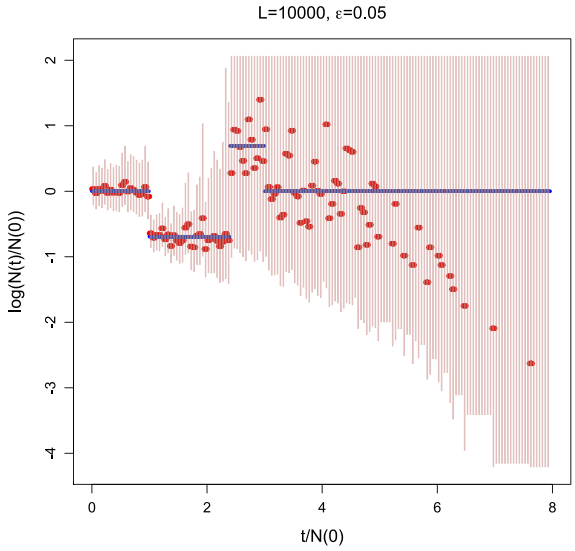


(f) $L = 10^4$, $\varepsilon = 0.2$.

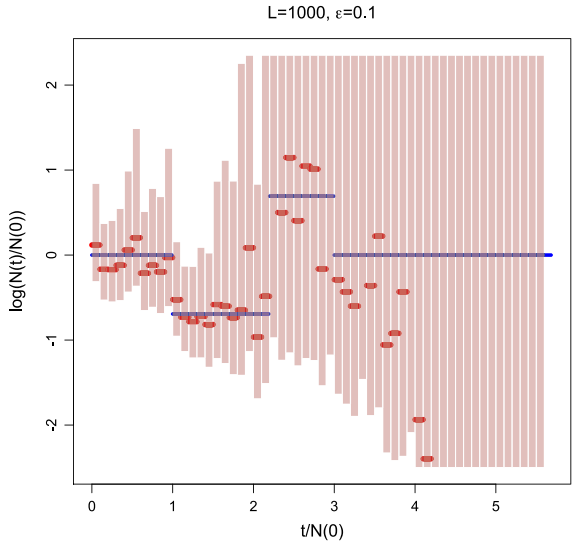
Fig. 4. Estimating a constant population size.



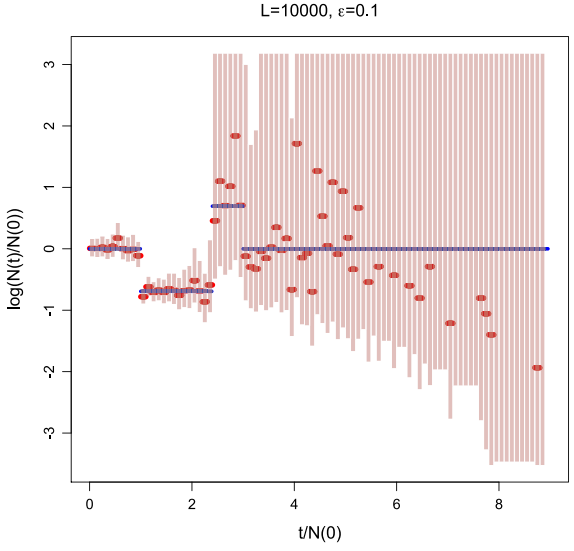
(a) $L = 10^3$, $\varepsilon = 0.05$.



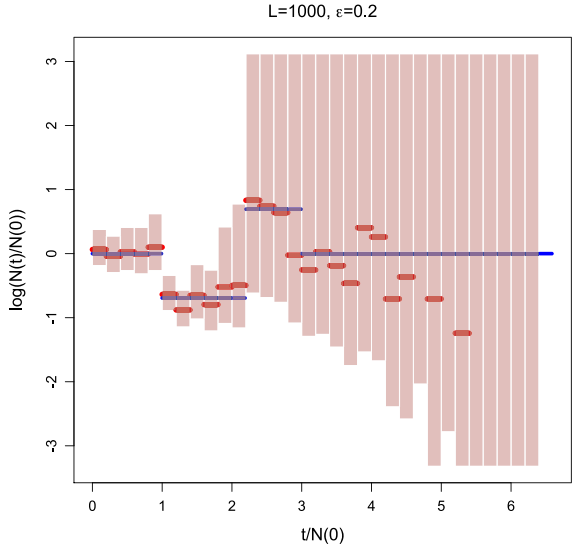
(b) $L = 10^4$, $\varepsilon = 0.05$.



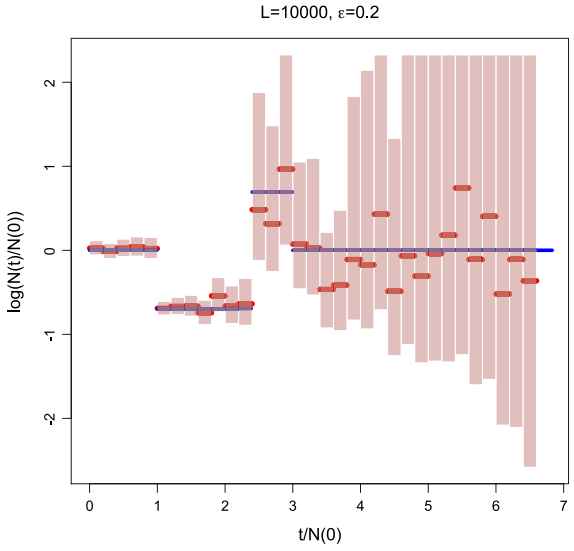
(c) $L = 10^3$, $\varepsilon = 0.1$.



(d) $L = 10^4$, $\varepsilon = 0.1$.

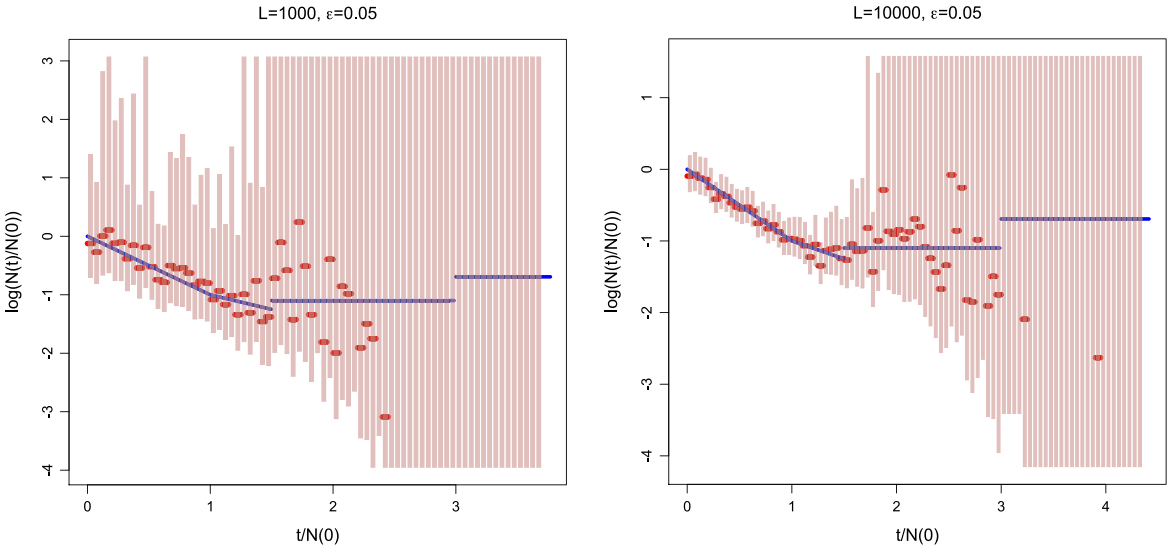


(e) $L = 10^3$, $\varepsilon = 0.2$.



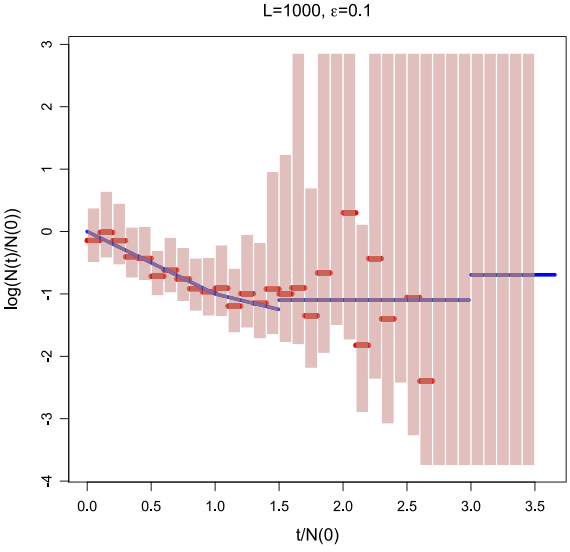
(f) $L = 10^4$, $\varepsilon = 0.2$.

Fig. 5. Estimating a piecewise constant population size.

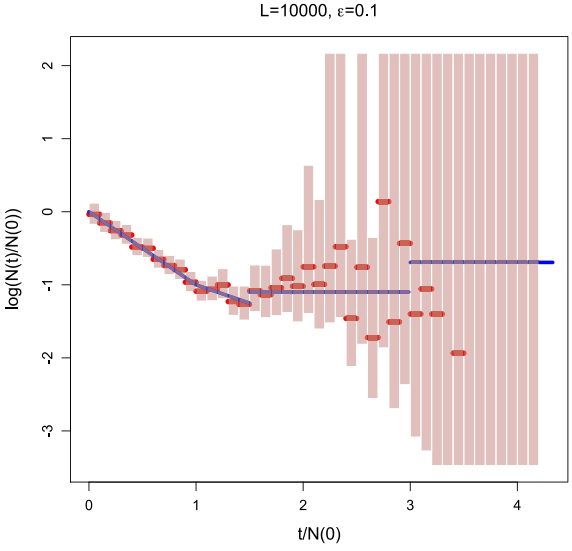


(a) $L = 10^3$, $\varepsilon = 0.05$.

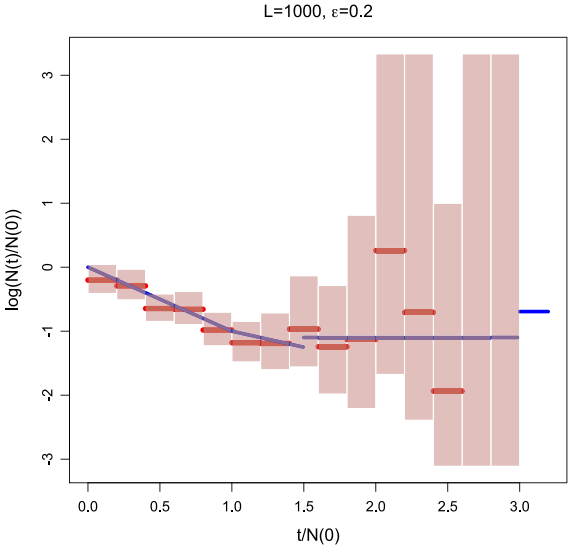
(b) $L = 10^4$, $\varepsilon = 0.05$.



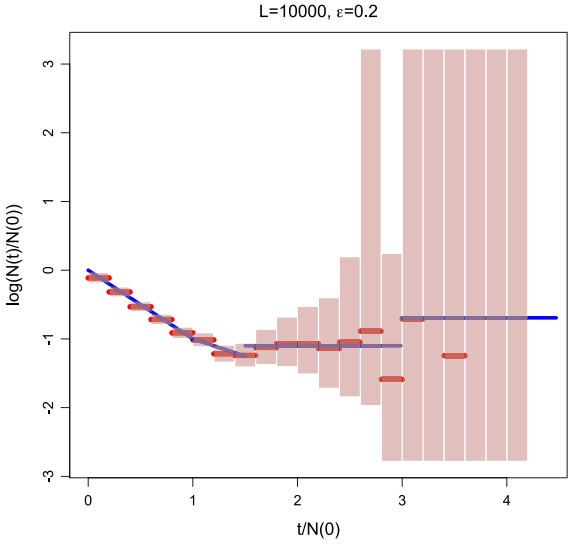
(c) $L = 10^3$, $\varepsilon = 0.1$.



(d) $L = 10^4$, $\varepsilon = 0.1$.

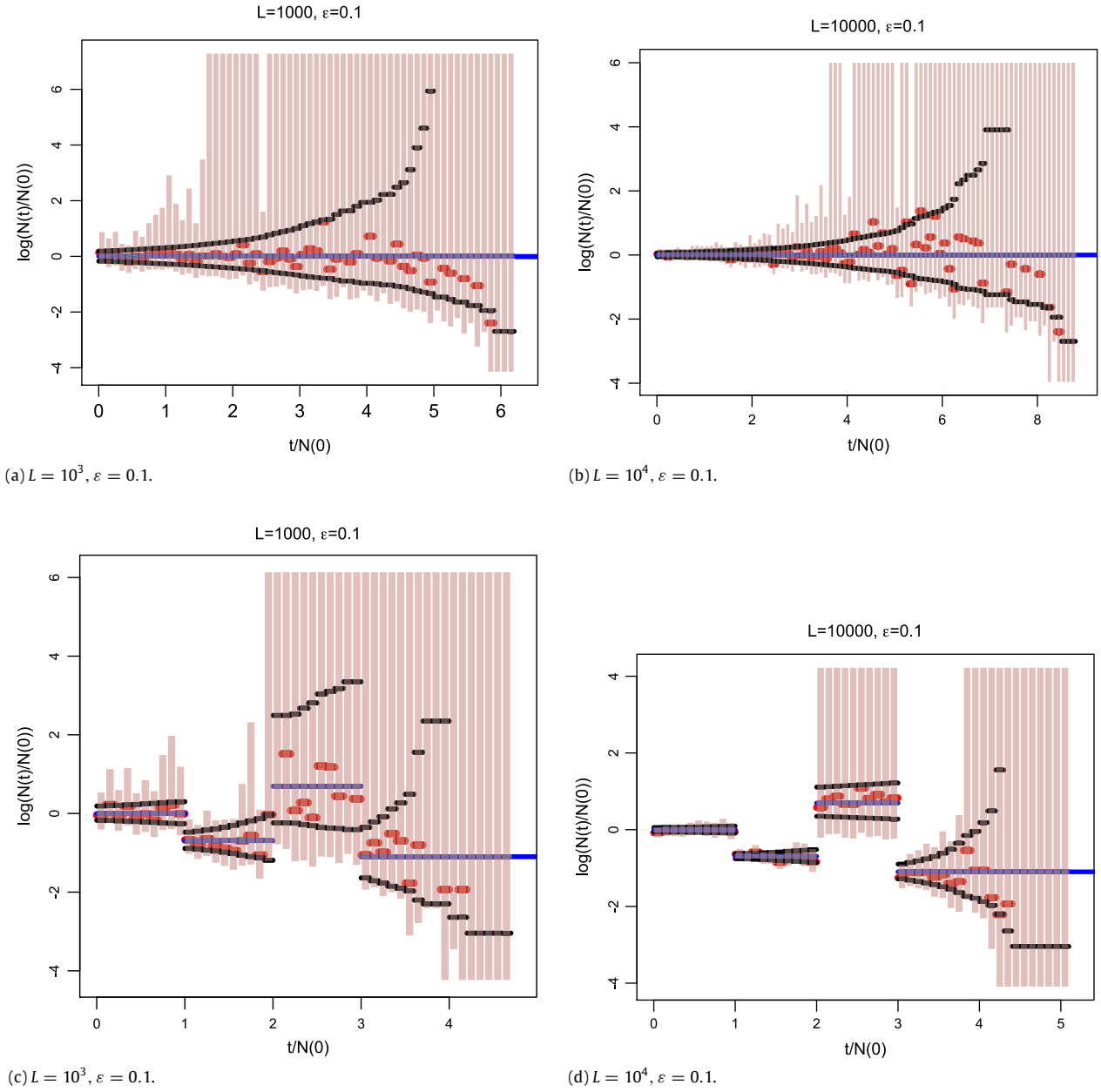


(e) $L = 10^3$, $\varepsilon = 0.2$.



(f) $L = 10^4$, $\varepsilon = 0.2$.

Fig. 6. Estimating a population history with piecewise exponential change.



(a) $L = 10^3$, $\varepsilon = 0.1$.

(b) $L = 10^4$, $\varepsilon = 0.1$.

(c) $L = 10^3$, $\varepsilon = 0.1$.

(d) $L = 10^4$, $\varepsilon = 0.1$.

Fig. 7. Constant and piecewise constant population histories with uncertainty intervals.

Estimation from sequence data. The assumption that we know exact coalescence times is unrealistic. These times need to be estimated from sequence data at independent loci with good accuracy. How do we estimate coalescence times from sequence data with quantitative upper and lower bounds analogous to those here?

Population substructure. Assume we want to estimate a population that is not only changing over time, but also has subpopulations that merge and split, and which may have migration rates between them. Are there analogs of our results in this setting? Note that identifiability can be an issue here since, for example, a constant population that splits at a given point in the past has the same distribution of coalescence times among two individuals as a single population that grows exponentially at a specific rate (backward in time) starting at the time of the split.

Acknowledgments

We thank Anand Bhaskar, Luke Gandolfo, Jasmine Nirody, Sara Sheehan, and Yun Song for helpful discussions and relevant references. We also thank anonymous reviewers for their careful reading of the manuscript which lead to several improvements. EM was supported by NSF grants DMS 1106999 and CCF 1320105 and by DOD ONR grant N000141110140. MZR was supported by NSF grant DMS 1106999 and by DOD ONR grant N000141110140. A portion of the work for this project was completed when NR was at University of California, Berkeley with support from NSF grants DMS-0704159, DMS-0806118, DMS-1106999 and ONR grant N00014-11-1-0140.

References

Bhaskar, A., Song, Y.S., 2014. Descartes' rule of signs and the identifiability of population demographic models from genomic variation data. Ann. Statist. 42 (6), 2469–2493.

Bhaskar, A., Wang, Y.X.R., Song, Y.S., Efficient inference of population size histories and locus-specific mutation rates from large-sample genomic variation data, 2014. Preprint http://biorxiv.org/content/early/2014/06/28/006742.1.

Birgé, L., 2007. Model selection for Poisson processes. In: Asymptotics: Particles, Processes and Inverse Problems. In: IMS Lecture Notes Monogr. Ser., vol. 55. Inst. Math. Statist., Beachwood, OH, pp. 32–64.

Bos, K.I., Harkins, K.M., Herbig, A., Coscolla, M., Weber, N., Comas, I., Forrest, S.A., Bryant, J.M., Harris, S.R., Schuenemann, V.J., Campbell, T.J., Majander, K., Wilbur, A.K., Guichon, R.A., Wolfe Steadman, D.L., Cook, D.C., Niemann, S., Behr, M.A., Zumarraga, M., Bastida, R., Huson, D., Nieselt, K., Young, D., Parkhill, J., Buikstra, J.E., Gagneux, S., Stone, A.C., Krause, J., 2014. Pre-Columbian mycobacterial genomes reveal seals as a source of New World human tuberculosis. Nature 514, 494–497.

Cheng, M.-Y., Hall, P., Tu, D., 2006. Confidence bands for hazard rates under random censorship. Biometrika 93 (2), 357–366.

Drummond, A., Rambaut, A., Shapiro, B., Pybus, O., 2005. Bayesian coalescent inference of past population dynamics from molecular sequences. Mol. Biol. Evol. 22 (5), 1185–1192.

Excoffier, L., Dupanloup, I., Huerta-Sánchez, E., Sousa, V.C., Foll, M., 2013. Robust demographic inference from genomic and SNP data. PLoS Genet. 9 (10), e1003905.

Gibbs, A.L., Su, F.E., 2002. On choosing and bounding probability metrics. Internat. Statist. Rev. 70 (3), 419–435.

Gutenkunst, R.N., Hernandez, R.D., Williamson, S.H., Bustamante, C.D., 2009. Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. PLoS Genet. 5 (10), e1000695.

Hall, P., 1992. Effect of bias estimation on coverage accuracy of bootstrap confidence intervals for a probability density. Ann. Statist. 20 (2), 675–694.

Harris, K., Nielsen, R., 2013. Inferring demographic history from a spectrum of shared haplotype lengths. PLoS Genet. 9 (6), e1003521.

Heled, J., Drummond, A., 2008. Bayesian inference of population size history from multiple loci. BMC Evol. Biol. 8 (1), 289.

Kac, M., 1966. Can one hear the shape of a drum? Amer. Math. Monthly 73 (4), 1–23.

Kingman, J.F.C., 1982a. On the genealogy of large populations. J. Appl. Probab. 27–43.

Kingman, J.F.C., 1982b. The coalescent. Stochastic Process. Appl. 13 (3), 235–248.

Lawless, J.F., 2003. Statistical Models and Methods for Lifetime Data, second ed. In: Wiley Series in Probability and Statistics, Wiley-Interscience [John Wiley & Sons], Hoboken, NJ.

Li, H., Durbin, R., 2011. Inference of human population history from individual whole-genome sequences. Nature 475, 493–496.

Li, J., Li, H., Jakobsson, M., Li, S., Sjödin, P., Lascoux, M., 2012. Joint analysis of demography and selection in population genetics: where do we stand and where could we go? Mol. Ecol. 21 (1), 28–44.

Lukić, S., Hey, J., Chen, K., 2011. Non-equilibrium allele frequency spectra via spectral methods. Theor. Popul. Biol. 79 (4), 203–219.

McVean, G.A., Cardin, N.J., 2005. Approximating the coalescent with recombination. Philos. Trans. R. Soc. B 360 (1459), 1387–1393.

Myers, S., Fefferman, C., Patterson, N., 2008. Can one learn history from the allelic spectrum? Theor. Popul. Biol. 73 (3), 342–348.

Nielsen, R., 2000. Estimation of population parameters and recombination rates from single nucleotide polymorphisms. Genetics 154 (2), 931–942.

Palamara, P.F., Lencz, T., Darvasi, A., Pe'er, I., 2012. Length distributions of identity by descent reveal fine-scale demographic history. Am. J. Hum. Genet. 91 (5), 809–822.

Reynaud-Bouret, P., 2003. Adaptive estimation of the intensity of inhomogeneous Poisson processes via concentration inequalities. Probab. Theory Related Fields 126 (1), 103–153.

Rice, J., Rosenblatt, M., 1976. Estimation of the log survivor function and hazard function. Sankhyā Ser. A 38 (1), 60–78.

Sethuraman, J., Singpurwalla, N.D., 1981. Large sample estimates and uniform confidence bounds for the failure rate function based on a naive estimator. Ann. Statist. 9 (3), 628–632.

Sheehan, S., Harris, K., Song, Y.S., 2013. Estimating variable effective population sizes from multiple genomes: a sequentially Markov conditional sampling distribution approach. Genetics 194, 647–662.

Silverman, B.W., 1986. Density Estimation for Statistics and Data Analysis. In: Monographs on Statistics and Applied Probability, Chapman & Hall, London.

Tataru, P., Nirody, J.A., Song, Y.S., 2014. diCal-IBD: demography-aware inference of identity-by-descent tracts in unrelated individuals. Bioinformatics 30 (23), 3430–3431.

Tavaré, S., 2004. Ancestral inference in population genetics. In: Lectures on Probability Theory and Statistics. In: Lecture Notes in Math., vol. 1837. Springer, Berlin, pp. 1–188.

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., Kang, H.M., Jordan, D., Leal, S.M., Gabriel, S., Rieder, M.J., Abecasis, G., Altshuler, D., Nickerson, D.A., Boerwinkle, E., Sunyaev, S., Bustamante, C.D., Bamshad, M.J., Akey, J.M., 2012. Evolution and functional impact of rare coding variation from deep sequencing of human exomes. Science 337 (6090), 64–69.

Wang, J.-L., 2005. Smoothing hazard rates. Encyclopedia Biostat. 7, 4986–4997.

Willett, R.M., Nowak, R.D., 2007. Multiscale Poisson intensity and density estimation. IEEE Trans. Inform. Theory 53 (9), 3171–3187.

Yandell, B.S., 1983. Nonparametric inference for rates with censored survival data. Ann. Statist. 11 (4), 1119–1135.