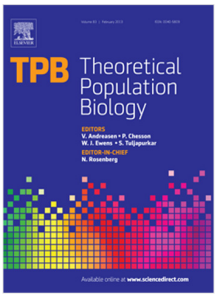
Journal Pre-proof



Rates of convergence in the two-island and isolation-with-migration models

Brandon Legried, Jonathan Terhorst

PII: S0040-5809(22)00051-X
DOI: https://doi.org/10.1016/j.tpb.2022.08.001
Reference: YTPBI 2854

To appear in: *Theoretical Population Biology*

Received date : 21 August 2021

Please cite this article as: B. Legried and J. Terhorst, Rates of convergence in the two-island and isolation-with-migration models. *Theoretical Population Biology* (2022), doi: https://doi.org/10.1016/j.tpb.2022.08.001.

Rates of convergence in the two-island and isolation-with-migration models

Brandon Legried and Jonathan Terhorst*
Department of Statistics, University of Michigan

August 10, 2022

Abstract

A number of powerful demographic inference methods have been developed in recent years, with the goal of fitting rich evolutionary models to genetic data obtained from many populations. In this paper we investigate the statistical performance of these methods in the specific case where there is continuous migration between populations. Compared with earlier work, migration significantly complicates the theoretical analysis and requires new techniques. We employ the theories of phase-type distributions and concentration of measure in order to study the two-island and isolation-with-migration models, resulting in both upper and lower bounds on rates of convergence for parametric estimators in migration models. For the upper bounds, we consider inferring rates of coalescent and migration on the basis of directly observing pairwise coalescent times, and, more realistically, when (conditionally) Poisson-distributed mutations dropped on latent trees are observed. We complement these upper bounds with information-theoretic lower bounds which establish a limit, in terms of sample size, below which inference is effectively impossible.

1 Introduction

Demographic inference–the estimation of past gene flow, migration, and size history events experienced by a population–is now a significant research area in evolutionary biology and mathematical genetics. Stimulated by an ongoing explosion in data

*Corresponding author: jonth@umich.edu

availability, a series of increasingly sophisticated statistical methods has been developed to infer rich, highly parameterized demographic models using patterns of population genetic variation. These methods have seen significant uptake in biology, with some (e.g., Gutenkunst et al., 2009; Li and Durbin, 2011) having been used in thousands of studies across a wide variety of species.

Formidable mathematical and computational hurdles must be overcome in order to estimate complex evolutionary models; often, even evaluating the likelihood function is nontrivial. As a result, research in this area has, to date, tended to focus on developing efficient inference methods. A much smaller number of authors have studied the question which interests us here: when is it theoretically (im)possible to estimate parameters of these models from data?

A starting point in the literature on the theoretical statistical aspects of demographic inference is Myers, Fefferman, and Patterson (2008), who proved the striking result that population size history is unidentifiable from the site frequency spectrum. That is, given an arbitrary size history function, there exists a smooth perturbation of it which produces exactly the same frequency spectrum in expectation. Subsequently, Bhaskar and Song (2014) showed that identifiability can be achieved by restricting the space of size history functions to be finite dimensional, for example piecewise constant or piecewise exponential. Terhorst and Song (2015) derived minimax lower bounds for demographic inference from the site frequency spectrum, and showed that there is a fundamental limit in our ability to infer size history for populations which have experienced a bottleneck. Baharian and Gravel (2018) showed that even in non-bottlenecked populations, there may be little to no statistical power to distinguish between different size history hypotheses on the basis of a finite amount of data. Working in a different setting, J. Kim, Mossel, Rácz, et al. (2015a) studied nonparametric estimation of the size history using samples of coalescent times from pairs of chromosomes, deriving both upper and lower bounds for hypothesis testing and estimation of the size history function. Johndrow and Palacios (2019) extended the analysis J. Kim, Mossel, Rácz, et al. (2015a) to coalescent trees on three samples, studied the benefit of incorporating ancient samples, and derived exact lower bounds on the Bayes error rate for distinguishing between population size histories.

All of the above papers consider the case of a panmictic population. Less attention still has been paid to inference in structured population models. Y. Kim et al. (2020) furthered the analysis of J. Kim, Mossel, Rácz, et al. (2015a) to the case where pairwise coalescent data is used to infer population structure, and showed in particular that the amount of data needed to accurately reconstruct the demography of a structured population may be exponential in the number of demes. Sousa, Grelaud, and Hey (2011) showed that the times of migration events in gene trees are

not identifiable under a standard coalescent model.

A related thread concerns reconstructing the phylogeny or "species tree" of a set of populations tree under a structured coalescent model. Up to this point, coalescent-based approaches have mostly considered complications arising only from incomplete lineage sorting, which causes gene trees to have a different topology than the background species tree (Rannala and Yang, 2003; Allman, Degnan, and Rhodes, 2011; Mirarab et al., 2014). Although there has been some recent progress on phylogenetic inference with migration (Hey, Chung, et al., 2018; Flouri et al., 2019), the focus of this line of work, species tree estimation, is ultimately different from that of demographic inference, where we seek to infer distributional parameters from a collection of latent genealogies.

Despite these many interesting and useful contributions, it is fair to say that our ability to estimate complex demographic models has far outpaced our theoretical understanding of those estimators. As noted above, only a handful of theoretical studies consider inference in the presence of complex population structure. Nevertheless, such models are now routinely fit in practice, often in consideration of numerous populations and different migration events (e.g., Gutenkunst et al., 2009; Jouganous et al., 2017; Kamm, Terhorst, and Song, 2017; Rodríguez et al., 2018; Kamm, Terhorst, Durbin, et al., 2020). Given that the theoretical results in the panmictic setting have so far been mainly negative, it seems important to extend the analysis to other types of structured population models that are becoming prevalent.

In this paper, we address this gap by theoretically analyzing some inference problems that arise in a structured coalescent model with continuous migration. Although some of our proof techniques are based on these earlier works (in particular, that of J. Kim, Mossel, Rácz, et al., 2015a), as we will see, migration introduces significant challenges into the analysis, requiring different approaches than have been used previously. Consequently, we restrict our focus to the simplest non-trivial structured population model of two islands with continuous migration between them, and a variant of it known as the isolation-with-migration model. In Section 2, we lay out our notation. Section 3 formalizes the model and introduces key definitions. Section 4 derives some eigenvalue bounds for migration transitions and tail bounds for estimation errors in the two-island model, some of which may be useful more generally. Section 5 studies moment-based estimation of the key parameters in the two-island model. Section 6 derives information-theoretic lower bounds on the ability to distinguish different island models from data. Section 7 concludes with some discussion.

2 Notation

Throughout the paper, n is used to denote sample size, and we suppress explicit dependence on it when there is no possibility of confusion. For any $r > 0$ and $\mathbf{x} \in \mathbb{R}^d$, let $B_r(\mathbf{x})$ be the ball of radius r around $\mathbf{x}$, i.e.

$$B_r(\mathbf{x}) = \{\mathbf{p} \in \mathbb{R}^d : \|\mathbf{x} - \mathbf{p}\| < r\}.$$

The constant

$$a_j := \binom{j}{2}$$

appears throughout the paper.

Matrices and vectors are denoted in boldface. The L^p norm of $\mathbf{x}$ is denoted

$$\|\mathbf{x}\|_p = (x_1^p + \cdots + x_d^p)^{1/p},$$

and the L^∞ norm is denoted $\|\mathbf{x}\|_\infty = \max_i |x_i|$. If p is not indicated, then $\|\mathbf{x}\| = \|\mathbf{x}\|_2 = (x_1^2 + \cdots + x_d^2)^{1/2}$ is taken to be the L^2 (Euclidean) norm. If $\mathbf{A} \in \mathbb{R}^{m \times n}$ is a matrix, then

$$\|\mathbf{A}\|_p = \sup_{\|\mathbf{x}\|_p = 1} \|\mathbf{A}\mathbf{x}\|_p$$

denotes the induced p-norm, with $\|\mathbf{A}\|$ denoting the operator norm. The Frobenius norm is given by

$$\|\mathbf{A}\|_F^2 = \sum_{i=1,j=1}^{m,n} a_{ij}^2,$$

where a_{ij} is the ijth entry of the matrix $\mathbf{A}$. If $m = n$ such that $\mathbf{A}$ is square, then the trace and determinant of $\mathbf{A}$ are denoted $\operatorname{tr} \mathbf{A}$ and $\det \mathbf{A}$, respectively. The identity matrix is denoted $\mathbf{I}$. The standard basis vectors are denoted

$$\mathbf{e}_1 = (1, 0, \ldots, 0), \mathbf{e}_2 = (0, 1, 0, \ldots, 0), \ldots, \mathbf{e}_d = (0, 0, \ldots, 1) \in \mathbb{R}^d,$$

and the vector of all ones is denoted $\mathbf{1} = (1, 1, \ldots, 1)^\intercal$. The dimensionality of $\mathbf{I}$, $\mathbf{e}_i$, and $\mathbf{1}$ may vary from usage to usage, but will be obvious from context.

In this paper, the dimension of all vector spaces is either 3 or 4, independent of any other problem-specific quantities. Hence, by equivalence of matrix norms, there exist universal constants $C_1, C_2 > 0$ such that

$$\|\mathbf{A}\|_p \leq C_1 \|\mathbf{A}\|_q \leq C_2 \|\mathbf{A}\|_p$$

for all $p, q > 0$, including $p = \infty$ or $p = F$ (Frobenius norm). In particular, for $\mathbf{A} \in \mathbb{R}^4$ we have

$$\begin{aligned} \|\mathbf{A}\|_F \leq \|\mathbf{A}\|_2 &\leq 2\,\|\mathbf{A}\|_F \\ \frac{1}{2}\|\mathbf{A}\|_\infty \leq \|\mathbf{A}\|_2 &\leq 2\,\|\mathbf{A}\|_\infty\,. \end{aligned}$$

3 The model

We consider a structured coalescent model with two demes and continuous migration between them, sometimes referred to as the "two-island" model (Takahata, 1988; Notohara, 1990). The model considers the probability distribution of a genealogy formed by sampling a pair of chromosomes. The time $t = 0$ corresponds to the present while positive t corresponds to t generations in the past. Let m_1 be the rate at which an individual migrates from island 2 to island 1, similarly for m_2. For any pair of individuals in the present, the time to their most recent common ancestor is called the coalescent time. Let c_1 be the corresponding rate of coalescence for the two individuals if they both live on island 1 and c_2 be the respective rate for island 2. It is not possible for coalescence to occur for pairs of individuals living on separate islands.

For any $t \geq 0$, the vector

$$\mathbf{p}_t = (p_{12}(t), p_{11}(t), p_{22}(t), p_{\text{coal}}(t))$$

gives a probability distribution on the finite sample space $\Omega = \{12, 11, 22, \text{coal}\}$. Here, $p_{12}(t)$ is the probability that a pair of individuals sampled in the present descend from a pair of individuals separated into the islands 1 and 2 at time t. Similarly, $p_{11}(t)$ is the probability that they descend from a pair of individuals both on island 1 at time t, with an analogous definition for $p_{22}(t)$. Lastly, $p_{\text{coal}}(t)$ is the probability that they descend from a common ancestor at time t.

The movement of a pair of individuals between these four states is modeled by a continuous-time Markov chain (CTMC) with state probabilities

$$\begin{aligned} \mathbf{p}_0 &= (p_{12}(0), p_{11}(0), p_{22}(0), p_{\text{coal}}(0)) \\ \frac{\mathrm{d}\mathbf{p}_t}{\mathrm{d}t} &= \mathbf{p}_t\mathbf{Q}, t > 0, \end{aligned} \tag{1}$$

where is $\mathbf{Q}$ is the transition rate matrix

$$\mathbf{Q} = Q(\mathbf{c}, \mathbf{m}) = \begin{pmatrix} -(m_1 + m_2) & m_1 & m_2 & 0 \\ 2m_2 & -(2m_2 + c_1) & 0 & c_1 \\ 2m_1 & 0 & -(2m_1 + c_2) & c_2 \\ 0 & 0 & 0 & 0 \end{pmatrix}. \tag{2}$$

Recall $\mathbf{p}_0$ is the distribution of the locations for a sampled pair of individuals in the present. In the present, we sample two individuals assuming they are not coalesced, so $p_{\text{coal}}(0) = 0$. To avoid degeneracies, we assume henceforth that both of the coalescent rates c_i and at least one of the migration rates m_i are strictly positive. The solution to (1) is $\mathbf{p}_t = \mathbf{p}_0 e^{\mathbf{Q}t}$.

For $x \geq 0$ and $t > x$, the hazard rate of coalescence $h(t \mid x) = h^{\mathbf{c},\mathbf{m}}(t \mid x)$ is the rate of entry into the "coal" state, given that the process has not already done so up to time x. Viewing x as the present time, conditioning on noncoalescence implies $p_{\text{coal}}(x|x) = 0$. If we define $\mathbf{p}_{t|x}$ to be the density of the coalescence time in each state at time $t \geq x$ conditioned on noncoalescence up to time x, then

$$\begin{aligned} \mathbf{p}_{x|x} &= \frac{1}{1 - p_{\text{coal}}(x|0)} \left(p_{12}(x|0), p_{11}(x|0), p_{22}(x|0), 0\right) \\ \mathbf{p}_{t|x} &= \mathbf{p}_{x|x} e^{(t-x)\mathbf{Q}}. \end{aligned} \tag{3}$$

The (conditional) hazard rate of coalescence at time t is given by multiplying $\mathbf{p}_{t|x}$ with the fourth column of $\mathbf{Q}$:

$$h(t \mid x) = c_1 p_{11}(t \mid x) + c_2 p_{22}(t \mid x). \tag{4}$$

Suppose $n \geq 2$ individuals are sampled at time $t = 0$ and consider the sequence of coalescent times $0 = x_{n+1} < x_n < \cdots < x_2$ in the genealogy of the sample. For any $j \leq n$, recall that $a_j = j(j-1)/2$ is the number of ways to choose a particular pair of individuals from j samples. Any pair of non-coalesced lineages that exist in the same deme are as likely to coalesce as any other, with that deme's coalescence rate. Then, the conditional hazard rate for coalescence time x, given that the $(j+1)$th coalescence time is x_{j+1}, is obtained by averaging over the possibilities for sampling a pair of individuals at time x_{j+1}, i.e.

$$h_j(x \mid x_{j+1}) = a_j \left[c_1 p_{11}(x \mid x_{j+1}) + c_2 p_{22}(x \mid x_{j+1})\right]. \tag{5}$$

The conditional density of the j-th coalescent time given the $(j+1)$-th coalescent time is then

$$f_j(x \mid x_{j+1}) = \exp\left(-\int_{t=x_{j+1}}^{x} h_j(t \mid x_{j+1}) \, \mathrm{d}t\right) h_j(x \mid x_{j+1}).$$

From the Markov property, the joint density of the coalescent times $x_n < ... < x_2$ is then

$$\begin{aligned} f(x_2, ..., x_n) &= \prod_{j=2}^{n} f_j(x_j \mid x_{j+1}) \\ &= \prod_{j=2}^{n} \exp\left(- \int_{t=x_{j+1}}^{x_j} h_j(t \mid x_{j+1}) \, \mathrm{d}t \right) h_j(x_j \mid x_{j+1}). \end{aligned}$$

4 Tail bounds

We now collect some results about the two-island model with migration which will be used below. These results exploit the fact that coalescent times in this model follow a so-called phase-type distribution. The phase-type distribution of a time-homogeneous, continuous time Markov chain has a simple closed form expression for statistics of interest such as the mean, variance, and moment generating function. With these formulas, we derive quantitative bounds on extreme-event probabilities involving coalescence times. In particular, we show that estimators of the migration and coalescence rates are concentrated around the true parameters, meaning the estimators converge in probability exponentially fast in the number of samples. A useful reference on these distributions is Asmussen and Albrecher (2010). Hobolth, Siri-Jegousse, and Bladt (2019) have also recently studied phase-type distributions in a related setting.

Definition 1 (Phase-type distribution)**.** Let $\{X_t\}_{t \geq 0}$ be a homogeneous, continuous time Markov chain on a finite state space $\mathcal{S}$ with single absorbing state $\Delta \in \mathcal{S}$. The rate matrix of X_t may be written in block form as

$$\mathbf{Q} = \begin{pmatrix} \mathbf{Q}' & \mathbf{s} \\ \mathbf{0} & 0 \end{pmatrix} \in \mathbb{R}^{|\mathcal{S}| \times |\mathcal{S}|}, \tag{6}$$

where $\mathbf{Q}' \in \mathbb{R}^{(|\mathcal{S}|-1) \times (|\mathcal{S}|-1)}$ is the sub-intensity matrix giving the transition rates between the transient states in $\mathcal{S}$, and $\mathbf{s} \in \mathbb{R}^{|\mathcal{S}|-1}$ is a column vector giving the transition rates from each transient state to Δ. Also, let $\boldsymbol{\alpha} \in \mathbb{R}^{|\mathcal{S}|}$ be the distribution of X_0. The first hitting time $\zeta = \inf\{t > 0 : X_t = \Delta\}$ is said to be of *phase-type with representation* $(\mathbf{Q}, \boldsymbol{\alpha})$. We denote this as

$$\zeta \sim \mathrm{PH}(\mathbf{Q}, \boldsymbol{\alpha}), \tag{7}$$

and denote the distribution function of ζ by $\mathbb{P}_{\boldsymbol{\alpha}}(\zeta < t)$.

Below we will need to simultaneously consider phase-type distributions with multiple initial distributions $\boldsymbol{\alpha}_{ij}$, $i, j \in \{1, 2\}$. This indexing of states is chosen so that a pair of individuals is in state ij if one individual is on island i and the other is on island j. We slightly abuse notation and denote these by $\zeta_{\boldsymbol{\alpha}_{ij}}$.

An important quantity in the theory of phase-type distributions is the spectrum of $\mathbf{Q}$. In particular, the long-term behavior of the process is controlled by the size of the gap between the largest and second-largest eigenvalues (Asmussen and Albrecher, 2010). Our first result quantifies this gap for the chain defined in (2).

Proposition 2. *Let $\mathbf{Q} = Q(\mathbf{c}, \mathbf{m})$. Then the eigenvalues $\lambda_0, \ldots, \lambda_3$ of $\mathbf{Q}$ are non-positive and real: $\lambda_3 \leq \lambda_2 \leq \lambda_1 < \lambda_0 = 0$, with*

$$\|\mathbf{c}\|_1 + 3\|\mathbf{m}\|_1 \geq |\lambda_1| \geq \frac{c_1c_2(m_1 + m_2) + 2(c_1m_1^2 + c_2m_2^2)}{c_1c_2 + 3(m_1c_1 + m_2c_2) + m_1c_2 + m_2c_1 + 2\,(m_1 + m_2)^{\,2}}. \quad (8)$$

Since only the leading eigenvalue plays a role in the sequel, we define

$$\lambda := \min\{|\lambda_i| : \lambda_i < 0 \text{ is an eigenvalue of } \mathbf{Q}\}$$

for the rest of the paper. The right-hand side of (8) is an explicit bound on the leading eigenvalue, and this will be used to prove bounds on error in demographic inference.

The following quantity appears repeatedly in the results to come.

Definition 3. Given a rate matrix $\mathbf{Q}$ with leading eigenvalue λ, the *condition number* of $\mathbf{Q}$ is $\kappa := \|\mathbf{Q}\|/\lambda$.

Remark. κ differs slightly from the usual definition: it is the ratio of the largest singular value to smallest (absolute) eigenvalue of $\mathbf{Q}$.

The following Lemma and Corollary establish a Chernoff-type bound on the phase-type random variable. The proof of the Lemma is given in the Appendix.

Lemma 4. *Let $M_{\zeta_{\boldsymbol{\alpha}}}(r) := \mathbb{E}\exp(r\zeta_{\boldsymbol{\alpha}})$ be the moment generating function of ζ in (7). Then $M_{\zeta_{\boldsymbol{\alpha}}}(r)$ is defined for all $r < \lambda$. Furthermore, for any such r,*

$$M_{\zeta_{\boldsymbol{\alpha}}}(r) \leq \kappa.$$

Corollary 5. *For any $t > 0$,*

$$\mathbb{P}(\zeta_{\boldsymbol{\alpha}} > t) \leq \kappa e^{-\lambda t}. \quad (9)$$

Proof. Apply the previous lemma to the Chernoff-type bound

$$\mathbb{P}(\zeta_{\boldsymbol{\alpha}} > t) \leq \inf_{0<r<\lambda} e^{-rt} M_{\zeta_{\boldsymbol{\alpha}}}(r).$$

□

The next results pertain to two possibly different island models specified by rate matrices $\mathbf{Q}^{(i)} = Q(\mathbf{c}^{(i)}, \mathbf{m}^{(i)})$ (cf. equation 2) for $i = 1, 2$, with leading eigenvalues $\lambda^{(i)}$ and condition numbers $\kappa^{(i)}$. We let $\mathbf{p}^{(i)}_{t|x}$, $h^{(i)}_j(t \mid x)$, and $f^{(i)}$ refer to the transition probability, hazard rate, and joint density for the corresponding model.

Definition 6. Given a pair of two-island models $\mathbf{Q}^{(1)} = Q(\mathbf{c}^{(1)}, \mathbf{m}^{(1)})$ and $\mathbf{Q}^{(2)} = Q(\mathbf{c}^{(2)}, \mathbf{m}^{(2)})$, we say *model 2 is δ-close to model 1* if there exist diagonal matrices $\mathbf{D}_c, \mathbf{D}_m$ with $\max\{\|\mathbf{D}_c\|, \|\mathbf{D}_m\|\} < \delta$ such that

$$\begin{aligned} \mathbf{c}^{(2)} &= (\mathbf{I}+\mathbf{D}_c)\mathbf{c}^{(1)} \\ \mathbf{m}^{(2)} &= (\mathbf{I}+\mathbf{D}_m)\mathbf{m}^{(1)}. \end{aligned} \tag{10}$$

By abuse of notation, we also refer to $\mathbf{Q}^{(2)}$ being δ-close to $\mathbf{Q}^{(1)}$. It follows from the definition that if $\mathbf{Q}^{(2)}$ is δ-close to $\mathbf{Q}^{(1)}$ then

$$\left\|\mathbf{Q}^{(1)} - \mathbf{Q}^{(2)}\right\| \leq 2\left\|\mathbf{Q}^{(1)} - \mathbf{Q}^{(2)}\right\|_F \leq 2\delta\left\|\mathbf{Q}^{(1)}\right\|_F \leq 2\delta\left\|\mathbf{Q}^{(1)}\right\|, \tag{11}$$

whence

$$\|\mathbf{Q}^{(2)}\| \leq (1 + 2\delta)\left\|\mathbf{Q}^{(1)}\right\|. \tag{12}$$

The next bound establishes convergence to zero linearly in δ. The following result, due to Mitrophanov (2003, Corollary 2.1), is stated in an adapted form below. Its proof follows from technical convergence results stated and proved in the Appendix. The proof of this theorem is also given in the Appendix.

Theorem 7 (Mitrophanov 2003)**.** *Let $\mathbf{Q}^{(1)}$ and $\mathbf{Q}^{(2)}$ be as in Proposition 18, and further suppose that*

$$\mathbf{p}^{(1)}_{x|x} = \mathbf{p}^{(2)}_{x|x}. \tag{13}$$

Then

$$\left\|\mathbf{p}^{(2)}_{t|x} - \mathbf{p}^{(1)}_{t|x}\right\|_1 \leq 2\delta\kappa^{(1)} \log\left[64(1 + 2\delta)\right]. \tag{14}$$

5 Upper bounds

In this section, we derive an estimator for the parameters in the symmetric two-island model, and prove some results about its accuracy on finite samples. Before presenting our results, we first outline some of the challenges of inference in structured population models.

In the panmictic setting, J. Kim, Mossel, Rácz, et al. (2015a) derive a nonparametric estimator of the effective population size function $N(t)$ based on some ideas from survival analysis. They rely on an explicit expression for the hazard rate function $\mathbb{P}(T \in [x, x+dx) \mid T > x, N(t))$ of the coalescent time T (see Remark 2.1 in their paper). This function is then inverted, yielding a histogram-type estimator for $\hat{N}(t)$. The simple form of the estimator makes it possible to precisely analyze its performance on finite samples.

Unfortunately, it does not seem possible to extend their approach to the case of (even relatively simple) structured population models. The hazard rate function (4) depends on $\mathbf{m}$ and $\mathbf{c}$ in a complicated way, and cannot be analytically inverted. Thus, although classical (asymptotic) guarantees are obtainable, it is difficult to study the finite-sample behavior of likelihood-based estimators in structured population models. For simplicity, we restrict attention to the case where the number of leaves n is equal to 2.

To make progress, we turn to a moment-based estimator instead. Let E_{12} be the expected time to coalescence for a pair of individuals that live on different islands and E_{11}, E_{22} be the analogous expectations for pairs living on the same islands. With $\mathbf{Q}'$ defined as in Section 3, we have

$$\mathbf{E} := (E_{12}, E_{11}, E_{22})^\intercal = \mathbb{E}\left(\mathbf{e}_{12}, \mathbf{e}_{11}, \mathbf{e}_{22}\right) = -\left(\mathbf{Q}'\right)^{-1}\mathbf{1} \tag{15}$$

where the initial distribution $\mathbf{e}_{ij}$ places all mass on state ij, and the final equality is a property of phase-type distributions (Asmussen and Albrecher, 2010, Theorem IX.1.5).

Three parameters may be estimated from the first moments. We thus restrict our attention to the symmetric two-island model where $m_1 = m_2 = m$, wherefore

$$E_{12}(m, c_1, c_2) = \frac{8m^2 + 3m(c_1 + c_2) + c_1 c_2}{2m\left[c_1 c_2 + m(c_1 + c_2)\right]} \tag{16}$$

$$E_{11}(m, c_1, c_2) = \frac{2(c_2 + 2m)}{m(c_1 + c_2) + c_1 c_2} \tag{17}$$

$$E_{22}(m, c_1, c_2) = E_{11}(m, c_2, c_1). \tag{18}$$

Remark. Note that the expected time to coalescence for two lineages sampled from the same deme is invariant to m when $c_1 = c_2$, a special case of Strobeck's theorem (Strobeck, 1987; Durrett, 2008).

For each $ij \in \{12, 11, 22\}$, suppose we sample L i.i.d. coalescent trees $\{T_\ell^{(ij)}\}_{\ell=1}^L$ and form the sample version $\hat{\mathbf{E}}$ of $\mathbf{E}$ by averaging:

$$\hat{E}_{ij} = \frac{1}{L}\sum_{\ell=1}^{L} T_\ell^{(ij)}.$$

By the law of large numbers, $\hat{\mathbf{E}}$ is a consistent estimator of $\mathbf{E}$. Given $\hat{\mathbf{E}}$, we solve (16)–(18) for m, c_1, c_2 to obtain the following consistent estimators of the model parameters:

$$\begin{aligned}
\hat{m}(\hat{\mathbf{E}}) &= \frac{1}{2\hat{E}_{12} - \hat{E}_{11} - \hat{E}_{22}} \\
\hat{c_1}(\hat{\mathbf{E}}) &= \frac{4\hat{E}_{12} - 3\hat{E}_{11} - \hat{E}_{22}}{\hat{E}_{11}} \times \hat{m}(\hat{\mathbf{E}}) \\
\hat{c_2}(\hat{\mathbf{E}}) &= \frac{4\hat{E}_{12} - 3\hat{E}_{22} - \hat{E}_{11}}{\hat{E}_{22}} \times \hat{m}(\hat{\mathbf{E}}).
\end{aligned} \tag{19}$$

There is also a convenient formula for the variance of phase-type distributed random variables. Writing V_{ij} for the variance of $T_\ell^{(ij)}$, we have

$$V_{12}(m, c_1, c_2) = \mathrm{Var}(\mathbf{e}_{12}) = 2\mathbf{e}_1 \left(\mathbf{Q}'\right)^{-2} \mathbf{1} - \left[\mathbf{e}_1 \left(\mathbf{Q}'\right)^{-1} \mathbf{1}\right]^2$$

where $\mathbf{e}_1 = (1, 0, 0)$. The variances $V_{11}(m, c_1, c_2) = \mathrm{Var}(\mathbf{e}_{11})$ and $V_{22}(m, c_1, c_2) = \mathrm{Var}(\mathbf{e}_{12})$ are computed using the standard basis vectors $\mathbf{e}_2 = (0, 1, 0)$ and $\mathbf{e}_3$, respectively. Explicitly, these formulas are

$$\begin{aligned}
V_{12}(m, c_1, c_2) &= \frac{64m^4 + m^2(11c_1^2 + 14c_1c_2 + 11c_2^2) + m(4c_1^2c_2 + 4c_1c_2^2) + c_1^2c_2^2}{4m^2\left[c_1c_2 + m(c_1 + c_2)\right]^2} \\
V_{11}(m, c_1, c_2) &= \frac{c_1\left(c_2^2 + 5mc_2 + 10m^2\right) + m\left(3c_2^2 + 10mc^2 + 16m^2\right)}{m\left[c_1c_2 + m(c_1 + c_2)\right]^2} \\
V_{22}(m, c_1, c_2) &= V_{11}(m, c_2, c_1).
\end{aligned}$$

Using these expressions and the multivariate delta method (e.g., Casella and Berger, 2001), the asymptotic covariance matrix of $\hat{m}(\hat{\mathbf{E}})$ and $\hat{c_i}(\hat{\mathbf{E}})$ may also be obtained.

5.1 Error analysis

Now we derive bounds on the estimation error of the migration and coalescence parameters as a function of the number of samples L. Noting that the numerator and denominator of $\hat{m}$ and $\hat{c}_1$ (we omit discussion of $\hat{c}_2$ since it is symmetric to $\hat{c}_1$) are both homogeneous polynomials in $\hat{\mathbf{E}}$, this is most easily accomplished by considering the relative error.

Proposition 8. *Suppose that* $|E_{ij} - \hat{E}_{ij}| \leq \delta E_{ij}$ *for* $i, j \in \{1, 2\}$. *Then*

$$|\hat{m}/m - 1| \leq 3m\delta\|\mathbf{E}\| + \mathcal{O}(\delta^2) \tag{20}$$
$$|\hat{c}_1/c_1 - 1| \leq \delta\,(1 + 9m\|\mathbf{E}\|) + \mathcal{O}(\delta^2). \tag{21}$$

Proof. The supposition is equivalent to $\hat{\mathbf{E}} = (\mathbf{I} + \mathbf{D})\mathbf{E}$, where $\mathbf{D}$ is a diagonal matrix with $\|\mathbf{D}\| \leq \delta$. Thus, with $\mathbf{a}_m = (2, -1, -1)^\intercal$, we get

$$\begin{aligned}\left|\frac{m}{\hat{m}} - 1\right| &= \left|\frac{\left\langle \mathbf{a}_m, \hat{\mathbf{E}}\right\rangle}{\langle \mathbf{a}_m, \mathbf{E}\rangle} - 1\right| = \left|\frac{\langle \mathbf{a}_m, (\mathbf{I}+\mathbf{D})\mathbf{E}\rangle}{\langle \mathbf{a}_m, \mathbf{E}\rangle} - 1\right| \\ &= \left|\frac{\langle \mathbf{D}\mathbf{a}_m, \mathbf{E}\rangle}{\langle \mathbf{a}_m, \mathbf{E}\rangle}\right| \leq \|\mathbf{a}_m\| m \frac{\|\mathbf{D}\mathbf{a}_m\|}{\|\mathbf{a}_m\|}\|\mathbf{E}\| \leq 3m\delta\|\mathbf{E}\|. \end{aligned} \tag{22}$$

Equation (20) follows since $|m - \hat{m}|/\hat{m} < \delta \implies |m - \hat{m}|/m < \delta + \mathcal{O}(\delta^2)$.

Similarly, for $\hat{c}_1$ and $\mathbf{a}_c = (4, -3, -1)^\intercal$,

$$\left|\frac{\hat{c}_1}{c_1} - 1\right| = \left|\frac{\langle \mathbf{a}_c, (\mathbf{I}+\mathbf{D})\mathbf{E}\rangle\, \hat{m}(\hat{\mathbf{E}})/\hat{E}_{11}}{\langle \mathbf{a}_c, \mathbf{E}\rangle\, m/E_{11}} - 1\right|.$$

By our assumptions and the relative error bound (22), there exist ϵ_1, ϵ_2 such that

$$|\epsilon_1| < \delta \qquad \hat{E}_{11} = (1 + \epsilon_1)E_{11}$$
$$|\epsilon_2| < 3\delta\|\mathbf{E}\|m \qquad \hat{m}(\hat{\mathbf{E}}) = (1 + \epsilon_2)m.$$

Then

$$\begin{aligned}\left|\frac{\hat{c}_1}{c_1}-1\right| &= \left|\left(\frac{1+\epsilon_2}{1+\epsilon_1}\right)\frac{\langle \mathbf{a}_c, (\mathbf{I}+\mathbf{D})\mathbf{E}\rangle\, m/E_{11}}{\langle \mathbf{a}_c, \mathbf{E}\rangle\, m/E_{11}}-1\right| \\ &= \left|\left(\frac{1+\epsilon_2}{1+\epsilon_1}\right)\left(1+\frac{\langle \mathbf{a}_c, \mathbf{D}\mathbf{E}\rangle}{\langle \mathbf{a}_c, \mathbf{E}\rangle}\right)-1\right| \\ &= \left|\epsilon_2-\epsilon_1+\frac{\langle \mathbf{a}_c, \mathbf{D}\mathbf{E}\rangle}{\langle \mathbf{a}_c, \mathbf{E}\rangle}+\mathcal{O}(\delta^2)\right| \\ &\le \epsilon_1+\epsilon_2+\frac{\delta\|\mathbf{a}_c\|\|\mathbf{E}\|}{\langle \mathbf{a}_c, \mathbf{E}\rangle}+\mathcal{O}(\delta^2) \end{aligned} \tag{23}$$

Now since $\|\mathbf{a}_c\| < 6$ and

$$\langle \mathbf{a}_c, \mathbf{E}\rangle = 1/m + 2(E_{12}-E_{11}) > 1/m,$$

we have

$$\frac{\delta\|\mathbf{a}_c\|\|\mathbf{E}\|}{\langle \mathbf{a}_c, \mathbf{E}\rangle} \le 6\delta\|\mathbf{E}\|m. \tag{24}$$

Inserting (24) into (23) and simplifying yields (21). □

Next, we show that $\hat{\mathbf{E}}$ is concentrated around its expectation $\mathbf{E}$. This essentially follows from the fact that $\hat{\mathbf{E}}$ is the sample average of phase-type distributions (see equation 15), and the tail bounds we derived in Section 3.

Proposition 9. *Let $T_1^{(ij)}, T_2^{(ij)}, \ldots, T_L^{(ij)}$ be i.i.d. with distribution $\mathrm{PH}(\mathbf{Q}, \boldsymbol{\alpha}_{ij})$. Then*

$$\mathbb{P}\Big(\Big|\sum_{\ell=1}^{L} T_\ell^{(ij)} - E_{ij}\Big| > Lt\Big) \le 2\exp\left\{-c\min\left(\frac{Lt^2}{\gamma^2}, \frac{Lt}{\gamma}\right)\right\}, \tag{25}$$

where

$$\gamma \le (2/\lambda)\max\{1, \log_2\kappa\}.$$

and $c > 0$ is a universal constant.

Proof. By Jensen's inequality and Lemma 4, for sufficiently small r and any $K \ge 1$,

$$M_{\zeta_{\boldsymbol{\alpha}_{ij}}}(r/K) = \mathbb{E}\left[\left(e^{r\zeta_{\boldsymbol{\alpha}_{ij}}}\right)^{1/K}\right] \le M_{\zeta_{\boldsymbol{\alpha}_{ij}}}(r)^{1/K} \le \kappa^{1/K}.$$

Let

$$K = \max\{1, \log_2\kappa\}. \tag{26}$$

Then $M_{\zeta_{\boldsymbol{\alpha}_{ij}}}(r/K) \leq 2$. This implies that $\zeta_{\boldsymbol{\alpha}_{ij}}$ has a *sub-exponential* distribution, in the sense of Vershynin (2018, Proposition 2.7.1), with (see Vershynin, 2018, Definition 2.7.5)

$$\kappa := \|\zeta_{\boldsymbol{\alpha}_{ij}}\|_{\psi_1} \leq K/r = 2K/\lambda, \tag{27}$$

where $\|X\|_{\psi_1}$ denotes the Orlicz 1-norm of the random variable X, and we chose $r = \lambda/2$ (say). The bound (25) then follows from Bernstein's inequality (Vershynin, 2018, Corollary 2.8.3). □

As we have seen, the leading eigenvalue λ factors integrally into our convergence rates. To gain intuition, consider the completely symmetric case where $m_1 = m_2 = m$ and $c_1 = c_2 = c$. Then by (8),

$$\lambda \geq \frac{2cm(c+2m)}{(c+2m)^2 + 4cm},$$

and we can distinguish a few cases:

- If $m \ll c$ then λ is roughly lower-bounded by $2m$. This occurs when there is a low rate of migration between two islands with small effective population sizes. We then have $\log(\|\mathbf{c}\|/\lambda) \approx \log[\|\mathbf{c}\|/(2m)] \gg 1$. The bound (27) degenerates, such that we cannot rule out $\kappa \gg 1$. In turn, the concentration inequality (25) degrades and we longer have good control on $\|\mathbf{E} - \hat{\mathbf{E}}\|$. This result quantifies the intuitive statement that inference (in particular, estimation of m) is difficult when the rate of migration is small.

- If $m \gg c$, then $\lambda \approx c/2$. This occurs when there is migration between two islands with large effective population sizes. Then $\log(\|\mathbf{c}\|/\lambda) \approx \log(2\sqrt{2})$, so $\kappa \in \mathcal{O}(c)$ in equation (27). The right-hand side of (25) is essentially $\exp(-Lt/c)$, and the rate of convergence is dominated by the overall rate of coalescence.

5.2 Poisson hierarchical model

The previous section derives error bounds under the assumption that we could directly sample pairwise coalescent times within and between demes. In this section, we relax this unrealistically favorable assumption, and instead consider a model where the data consist of counts of the number of pairwise mismatches between randomly sampled genes. Specifically, we suppose

$$D_{ij} \mid T_{ij} \sim \text{Poisson}(\theta T_{ij}), \tag{28}$$
$$T_{ij} \sim \text{PH}(\mathbf{Q}, \boldsymbol{\alpha}_{ij}) \tag{29}$$

where the initial distribution $\boldsymbol{\alpha}_{ij}$ places all mass on deme $ij \in \{11, 22, 12\}$, so that T_{ij} are (i.i.d.) pairwise coalescent times between two genes sampled from demes i and j, which may be equal. Here $\theta/2$ is the rate of mutation per unit of coalescent time, assumed known. This type of model is known in the literature as the Poisson random field (Sawyer and Hartl, 1992), and is realistic if there is no recombination within genes; there is free recombination between genes; and θ is low such that there is no recurrent mutation (the so-called *infinite sites* assumption). It forms the basis of methods designed to estimate population history from the site frequency spectrum (e.g., Gutenkunst et al., 2009; Excoffier et al., 2013; Bhaskar and Song, 2014; Kamm, Terhorst, and Song, 2017; Jouganous et al., 2017; Kamm, Terhorst, Durbin, et al., 2020), and has had significant impact in applications (e.g., Yi et al., 2010; Gravel et al., 2011; Tennessen et al., 2012; Gazave et al., 2013).

For each $ij \in \{11, 22, 12\}$, we sample L i.i.d. mutation counts $\{D_\ell^{(ij)}\}_{\ell=1}^L$ and estimate the E_k using sample averages:

$$\hat{E}_{ij} = \frac{1}{\theta L} \sum_{\ell=1}^{L} D_\ell^{(ij)}.$$

Then we use the estimators developed in the previous section.

To get a rate of convergence, we need to extend Proposition 9 to the marginal distribution of D_{ij} in (28).

Lemma 10. *Let D_{ij} be distributed according to (28)–(29), and let $M_{D_{ij}/\theta}(s)$ denote the moment generating function of D_{ij}/θ. Then for all $s \le \theta \log(1 + \lambda/\theta)$,*

$$M_{D_{ij}/\theta}(s) \le \kappa.$$

Proof. We have

$$\begin{aligned}
\mathbb{E} \exp(sD_{ij}/\theta) &= \mathbb{E}\left[\mathbb{E}(\exp(sD_{ij}/\theta) \mid T_{ij})\right] \\
&= \mathbb{E} \exp\left[\theta T_{ij}(e^{s/\theta} - 1)\right] \\
&= M_{\zeta_{\boldsymbol{\alpha}_{ij}}}\left[\theta(e^{s/\theta} - 1)\right],
\end{aligned}$$

so the claim follows from Lemma 4. □

Remark. Lemma 10 also follows from general results on subexponential mixtures of Poisson random variables (Schmidli, 1999), but our earlier results enable a direct and more quantitative proof.

Proposition 11. *Let* $D_1^{(ij)}, D_2^{(ij)}, \ldots, D_L^{(ij)}$ *be i.i.d. with distribution* D_{ij} *in* (28). *Then*

$$\mathbb{P}\Big(\Big|\sum_{i=1}^{L} D_\ell^{(ij)}/\theta - E_{ij}\Big| > Lt\Big) \le 2\exp\left\{-\min\left(\frac{Lt^2}{\gamma^2}, \frac{Lt}{\gamma}\right)\right\}, \tag{30}$$

where

$$\gamma \le \max\{1, \log_2 \kappa\} \Big/ \Big(\theta \log\sqrt{1+\lambda/\theta}\Big)$$

and $c > 0$ *is a universal constant.*

Proof. As in the proof of Proposition 9, we find that $M_{D_{ij}/\theta}(r/K) \le 2$ for $r < \theta\log(1+\lambda/\theta)$ and the same constant K. Taking $r = \theta\log\sqrt{1+\lambda/\theta}$, we get

$$\gamma := \|D_{ij}/\theta\|_{\psi_1} \le K/r.$$

□

5.3 Simulations

Using Propositions 8–10, we can bound the accuracy of migration and coalescent rate estimates in the two-island model from finite amounts of data. For example, setting $t = \delta E_{ij}$ in (30) and finding L such that the right-hand side is less than or equal to ϵ, we get a bound on the relative error $|\hat{E}_{ij} - E_{ij}|/E_{ij} < \delta$ that holds with probability at least $1-\epsilon$.

In Figure 1 we consider using sample averages of T_{ij} and D_{ij} to estimate $\hat{E}_{12}$, the average coalescent time for lineages originating in different demes. We set $\delta = \epsilon = 0.1$, i.e. $< 10\%$ relative error with $> 90\%$ probability, and for simplicity we took $c_1 = c_2 = 1$. In the simulations of D_{ij}, the mutation rate was set to $\theta = 0.1$. The area between the shaded blue region contains the .05–.95 quantiles of the sampling distribution of $\hat{E}_{12}$, obtained over 1,000 independent trials. The red lines are $(1 \pm 0.1)E_{12}$. Based on our theoretical calculations, we found the value of L^* needed to ensure that the blue region was contained between the red lines. Thus, the sharpness of our bounds is reflected in the gap between the red lines (bounds) and blue region, with a larger gap indicating that our bounds predicted more samples were required than were actually necessary.

We see that the bounds are fairly accurate, particularly for using T_{12} (direct sampling of coalescent times) in order to estimate the population means. The actual number of samples L^* is plotted in Figure 2 (left panel). As expected, estimating E_{12} with Poisson noise is more difficult, requiring 1–2 order of magnitude more data

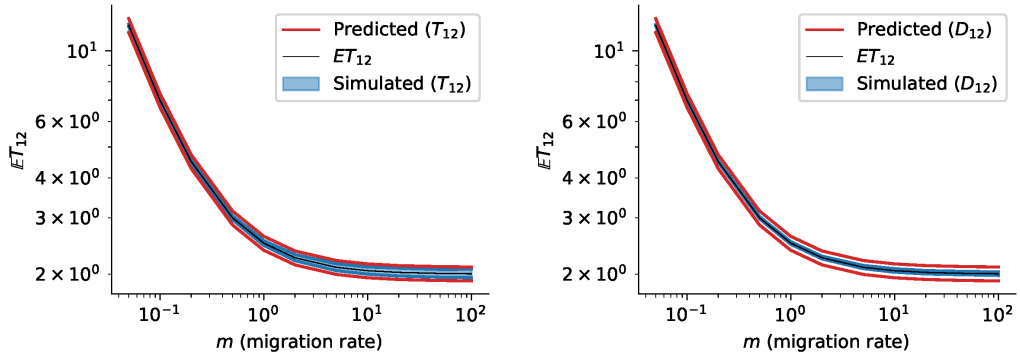


Figure 1: Observed and predicted 90% confidence intervals for the 5^{th} and 95^{th} percentiles of the sample mean. Left panel: directly sampling (T_{ij}). Right panel: Poisson noise (D_{ij}).

to obtain the same level of accuracy, and estimation requires more data when the migration rate is low.

Next, we used simulations of D_{ij} to estimate the migration rate m using (19). We studied the relative error $|m/\hat{m} - 1|$ and compared it to the bound (20), where, as noted above, $\delta = 0.1$ and m varied across a range of values. (Of course, (22) depends on the true parameters m and $\|\mathbf{E}\|$, so the bound may have limited practical use, but we can use it to get intuition for how the methods perform on real data.) The results are shown in the right panel of Figure 2, where we plugged $\delta = 0.1$ and relevant values c_1, c_2, m and $\mathbf{E}$ into (22) to obtain the upper bound. We can see that the bound is loose by a (large) constant, but has the correct functional dependence in m. Since Figure 1 showed that the concentration bounds on D_{ij} are accurate, this imprecision is probably due to the fairly rudimentary bounds employed in the proof of Proposition 8.

6 Lower bounds

In this section, we prove several lower bounds on parameter estimation in the two-island migration models. The starting point of our work is the following result of J. Kim, Mossel, Rácz, et al. (2015a) on distinguishing between different single-population coalescent models.

Theorem (J. Kim, Mossel, Rácz, et al. 2015a, Theorem 3.2)**.** Consider the following

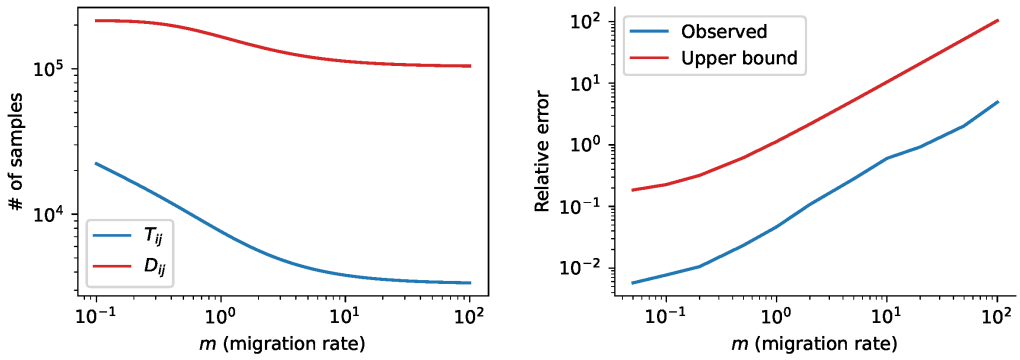


Figure 2: (Left panel) Number of samples calculated to obtain concentration bounds in Figure 1. (Right panel) Relative error in estimating $\hat{m}$ from Poisson-distributed mutation data.

hypothesis testing problem: H_1 states that the effective population size during the interval $[0, \infty)$ is constant N, while H_2 states that the population size during the same interval is the constant $(1+\eta)N$ for a fixed $\eta > 0$. 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

$$\Upsilon^2 \leq 2L \left(1 - \left(\frac{2\sqrt{1+\eta}}{2+\eta}\right)^{n-1}\right) \leq \frac{L(n-1)\eta^2}{4}. \tag{31}$$

In this section, we prove analogous results for population sizes and migration parameters in the two-island model and the isolation-with-migration model. There is a limitation in distinguishing between two hypotheses on population history for an arbitrary period of time with any estimation method, even though simple moment-based estimators converge quickly to their respective model parameters. Our bounds are given in terms of the rates of coalescence and migration in the two-island model model. The form of these results differs slightly from those of J. Kim, Mossel, Rácz, et al., which are stated in terms of (perturbations of) the effective population size N. In the panmictic setting of their paper, the effective population size and coalescence rate are inversely related, but such a simple relationship no longer holds here: there are multiple effective population sizes, and the rate of coalescence has a complicated expression that depends on all of the model parameters (see Section 3). Thus, it seems more natural to work with the rates of coalescence and migration directly.

6.1 Probability metrics and Bayes error rate

The section expands on the discussion in Section 3 of J. Kim, Mossel, Rácz, et al. (2015a). Let $(\Omega, \mathcal{F}, P)$ and $(\Omega, \mathcal{F}, Q)$ be two measures defined on a common probability space, with corresponding probability density functions f_P and f_Q. The *total variation distance* between P and Q is defined to be

$$d_{TV}(P, Q) := \sup_{A \in \mathcal{F}} |P(A) - Q(A)| = \frac{1}{2} \int |f_P - f_Q|.$$

By abuse of notation, we may sometimes write $d_{TV}(f_P, f_Q)$ to mean the same thing.

Suppose we are given a datum D that has been generated under either P or Q, and are asked to decide which measure was used assuming both choices are equally likely. The total variation distance between P and Q bounds the ability of any classifier to do so. Indeed, let $\chi \in \{P, Q\}$ denote the true data generating distribution, and $\hat{\chi}(D) \in \{P, Q\}$ be a classifier. The probability that $\hat{\chi}$ correctly classifies D can be written

$$\mathbb{P}(\hat{\chi} = \chi) = \frac{1 + \Upsilon}{2}, \tag{32}$$

where $\Upsilon > 0$ since the error of any binary classifier can be made less than $1/2$. Note that (32) rearranges to

$$\Upsilon = \mathbb{P}(\hat{\chi} = \chi) - \mathbb{P}(\hat{\chi} \neq \chi).$$

It can be shown (e.g., Devroye, Györfi, and Lugosi, 2013) that the best possible classification rule is the likelihood ratio: is $\hat{\chi} = P$ iff $P(D) > Q(D)$, in which case

$$\mathbb{P}(\hat{\chi} = \chi) - \mathbb{P}(\hat{\chi} \neq \chi) = \frac{1}{2} \left[\int_{f_P > f_Q} f_P + \int_{f_Q > f_P} f_Q - \int_{f_Q > f_P} f_P - \int_{f_P > f_Q} f_Q \right] = d_{TV}(P, Q).$$

This classification rule is said to achieve the minimal or "Bayes" error rate. If multiple samples are given, say $D_1, \ldots, D_L$, then $\Upsilon = d_{TV}(P^{\otimes L}, Q^{\otimes L})$, where $P^{\otimes L}$ denotes product measure.

In our setting, it is easier to work with a related quantity known as the Hellinger distance:

$$\begin{aligned} d_H^2(P, Q) &:= \frac{1}{2} \int \left(\sqrt{f_P} - \sqrt{f_Q} \right)^2 \\ &= 1 - \int \sqrt{f_P f_Q}. \end{aligned} \tag{33}$$

It is easily shown that $d_{TV}^2 \leq 2d_H^2$. Furthermore, the Hellinger distance distributes over product measures: if $P = P_1 \times P_2$ and $Q = Q_1 \times Q_2$ represent product measures, then

$$d_H^2(P, Q) \leq d_H^2(P_1, Q_1) + d_H^2(P_2, Q_2). \tag{34}$$

Hence, given L i.i.d. samples hypothesized to have been generated under either P or Q, it follows that

$$\Upsilon \leq d_{TV}(P^{\otimes L}, Q^{\otimes L}) \leq 2d_H^2(P^{\otimes L}, Q^{\otimes L}) \leq 2Ld_H^2(P, Q). \tag{35}$$

Going forward, we may abuse notation by identifying $\mathbb{Q}_i$ with its corresponding probability density function $f^{(i)}$ and compute $d_H^2(f^{(1)}, f^{(2)})$ and $d_{TV}(f^{(1)}, f^{(2)})$.

6.2 Two-island models

In this subsection, we study the ability to statistically distinguish between different two-island models as a function of how close they are to each other. For $i = 1, 2$, we suppose that under H_i the coalescent times are generated under a two-island model with rate matrix $\mathbf{Q}^{(i)} = Q(\mathbf{c}^{(i)}, \mathbf{m}^{(i)})$, and that $\mathbf{Q}^{(2)}$ is close to $\mathbf{Q}^{(1)}$ in a sense that is made precise below.

To improve readability, for the remainder of the section we suppress dependence of the density and hazard functions on x and y when there is no risk of confusion. Let $\bar{h} = [h_j^{(1)} + h_j^{(2)}]/2$, $\bar{H} = \int^x \bar{h}$, and $R = \bar{h} - \sqrt{h_j^{(1)} h_j^{(2)}} \geq 0$. Then

$$\int \sqrt{f_j^{(1)} f_j^{(2)}} = \int e^{-\bar{H}}(\bar{h} - R) = 1 - \int e^{-\bar{H}} R, \tag{36}$$

all integrals being over the positive reals. So by (33),

$$d_H^2(f_j^{(1)}, f_j^{(2)}) \leq \int e^{-\bar{H}} R.$$

If

$$h_j^{(2)} = (1 + \delta) h_j^{(1)}, \tag{37}$$

then $R = \left[(2 + \delta)/2 - \sqrt{1 + \delta}\right] h_j^{(1)}$, whence

$$\int e^{-\bar{H}} R = 1 - \frac{2\sqrt{1 + \delta}}{2 + \delta} \leq \frac{\delta^2}{8}. \tag{38}$$

This is essentially the bound obtained by J. Kim, Mossel, Rácz, et al. (2015a, Theorem 3.2).

Below we extend this result to the two-island model. Theorem 12 covers the case when the two model hypothesis are δ-close in the sense of Definition 6, without placing any additional assumptions on the relationship between the hypotheses. This result is general, but as can be seen from equation (39), the bound is on the order $\mathcal{O}(\delta)$, so it is asymptotically looser than the $\mathcal{O}(\delta^2)$ bound indicated by (38). Getting the $\mathcal{O}(\delta^2)$ rate turns out to depend rather delicately on a cancellation of the first-order coefficients in the Taylor expansion of $2\sqrt{1+\delta}/(2+\delta)$. This, in turn, only seems to happen if the hazard rate function $h_j^{(2)}$ is an exact scalar multiple of $h_j^{(1)}$, as in (37). When there is more than one population, such an equality no longer holds *even when* the parameters of the underlying model differ only by a multiplicative factor. Currently, we do not know if the difference in rates is an artifact of our proof technique, or if having data from multiple populations in fact renders the inference problem quantitatively easier.

Theorem 12. *Let H_1 and H_2 be hypotheses with corresponding rate matrices $\mathbf{Q}^{(1)}$ and $\mathbf{Q}^{(2)}$ such that H_2 is δ-close to H_1. If L i.i.d. pairwise coalescent trees are sampled under H_1 or H_2, each with probability $1/2$, then for sufficiently small $\delta > 0$, the Bayes error rate for any classifier is at least $(1-\Upsilon)/2$, where*

$$\Upsilon^2 \leq 2L\left[C\delta + \mathcal{O}(\delta^2)\right], \tag{39}$$

where

$$C = 3 + \frac{20\kappa^{(1)}\log(64)}{\left[\lambda^{(1)}\right]^2}.$$

The proof actually derives nonasymptotic (in δ) bounds on Υ, but to simplify the exposition we choose to present the result in the form (39). To prove the theorem, we establish some lemmas that enable upper-bounding the distance between the probability measures corresponding to H_1 and H_2. For $i = 1, 2$, we suppose that under H_i the coalescent times are generated under a two-island model with rate matrix $\mathbf{Q}^{(i)} = Q(\mathbf{c}^{(i)}, \mathbf{m}^{(i)})$. We assume that $\mathbf{Q}^{(2)}$ is δ-close to $\mathbf{Q}^{(1)}$. By rescaling coalescent time, we may assume $\left\|\mathbf{Q}^{(1)}\right\| = 1$, so that

$$\begin{aligned} \|\mathbf{Q}^{(1)} - \mathbf{Q}^{(2)}\| &\leq 2\delta \\ \|\mathbf{Q}^{(2)}\| &\leq 1 + 2\delta. \end{aligned} \tag{40}$$

Lemma 13. *Let $h_j^{(1)}$ and $h_j^{(2)}$ be the hazard rate functions corresponding to H_1 and*

H_2 when there are j lineages remaining.

$$\left|h_j^{(1)} - h_j^{(2)}\right| \leq 2\delta h_j^{(1)} + 2(1+2\delta)a_j \left\|\mathbf{p}_{y|x}^{(1)} - \mathbf{p}_{y|x}^{(2)}\right\|_1 \tag{41}$$

$$\frac{h_j^{(1)} + h_j^{(2)}}{2} \leq (1+\delta)h_j^{(1)} + (1+2\delta)a_j \left\|\mathbf{p}_{y|x}^{(1)} - \mathbf{p}_{y|x}^{(2)}\right\|_1. \tag{42}$$

Lemma 14. *Let $h_j^{(1)}$ and $h_j^{(2)}$ as above. Then*

$$\sqrt{h_j^{(1)} h_j^{(2)}} \geq (1-2\delta)h_j^{(1)} - 2(1+2\delta)a_j \left\|\mathbf{p}_{y|x}^{(1)} - \mathbf{p}_{y|x}^{(2)}\right\|_1.$$

We prove the Theorem in the case where the number of sampled leaves n equals 2, so the previous Lemmas are applied only in the case where $n = j = 2$. It follows from Theorem 7 that the error terms in the above expressions converge to 0 linearly in δ.

Proof of Theorem 12. The squared Hellinger distance between the two hypotheses is

$$d_H^2(f^{(1)}, f^{(2)}) = 1 - \int_{y=0}^{\infty} \sqrt{f_2^{(1)}(y \mid 0) f_2^{(2)}(y \mid 0)}\, \mathrm{d}y.$$

Recall that the density can be expressed using the hazard rate as

$$f_2^{(i)}(y|0) = h_2^{(i)}(y|0) e^{-\int_{t=0}^{y} h_2^{(i)}(t|0)\mathrm{d}t}, i \in \{1,2\}.$$

Substituting these expressions and applying Lemmas 13 and 14 implies

$$\begin{aligned}
&\int_{y=0}^{\infty} \sqrt{f_2^{(1)}(y \mid 0) f_2^{(2)}(y \mid 0)}\, \mathrm{d}y \\
&= \int_{y=0}^{\infty} \exp\left(-\int_{t=0}^{y} \frac{h_2^{(1)}(t \mid 0) + h_2^{(2)}(t \mid 0)}{2}\, \mathrm{d}t\right) \sqrt{h_2^{(1)}(y \mid 0) h_2^{(2)}(y \mid 0)}\, \mathrm{d}y \\
&\geq \int_{y=0}^{\infty} \exp\left(-(1+\delta)\int_{t=0}^{y} h_2^{(1)}(t \mid 0)\, \mathrm{d}t\right) && (43) \\
&\quad \times \exp\left(-\frac{1+2\delta}{2}\int_{t=0}^{y} \left\|\mathbf{p}_{t|0}^{(1)} - \mathbf{p}_{t|0}^{(2)}\right\|_1 \mathrm{d}t\right) && (44) \\
&\quad \times \left[(1-2\delta)h_2^{(1)}(y \mid 0) - 2(1+2\delta)\left\|\mathbf{p}_{y|0}^{(1)} - \mathbf{p}_{y|0}^{(2)}\right\|_1\right] \mathrm{d}y. && (45)
\end{aligned}$$

(Note that in the preceding display, we have $a_j = a_2 = 1$ since we assume $j = 2$.) We combine lines (43) and (44) in the above display to form

$$k_2(y \mid 0) := (1+\delta)h_2^{(1)}(y \mid 0) + \frac{1+2\delta}{2}\left\|\mathbf{p}_{y|0}^{(1)} - \mathbf{p}_{y|0}^{(2)}\right\|_1,$$

and then subtract $(1+\delta)/(1-2\delta)$ times line (45) from k_2 to form

$$\begin{aligned} v(\delta) &= \frac{(1+2\delta)(5+2\delta)}{2(1-2\delta)} \\ R_2(y \mid x) &= v(\delta)\left\|\mathbf{p}_{y|0}^{(1)} - \mathbf{p}_{y|0}^{(1)}\right\|_1. \end{aligned} \tag{46}$$

This gives us

$$\begin{aligned} &\int_{y=0}^{\infty} \sqrt{f_2^{(1)}(y \mid 0) f_2^{(2)}(y \mid 0)}\,\mathrm{d}y \\ &\qquad \geq \frac{1-2\delta}{1+\delta}\int_{y=0}^{\infty} \exp\left(-\int_{t=0}^{y} k_2(t \mid 0)\mathrm{d}t\right) \times \left[k_2(y \mid 0) - R_2(y \mid 0)\right]\mathrm{d}y. \end{aligned}$$

Splitting the integral into two pieces, we first have

$$\int_{y=0}^{\infty} \exp\left\{-\int_{t=0}^{y} k_2(t \mid 0)\right\} k_2(y \mid 0)\mathrm{d}y = 1.$$

For the other piece, the Chernoff bound (9) gives

$$\begin{aligned} \int_{y=0}^{\infty} \exp\left(-\int_{t=0}^{y} h_2^{(1)}(t \mid x)\mathrm{d}t\right)\mathrm{d}y &\leq \int_{y=0}^{\infty} \exp\left(-\int_{t=0}^{y} h_2^{(1)}(t \mid x)\mathrm{d}t\right)\mathrm{d}y \\ &\leq \sup_{\boldsymbol{\alpha}} \int_{t=0}^{\infty} \mathbb{P}(\zeta_{\boldsymbol{\alpha}} > t)\,\mathrm{d}t \leq \frac{\kappa^{(1)}}{\lambda^{(1)}} = \frac{1}{[\lambda^{(1)}]^2}. \end{aligned}$$

By Theorem 7,

$$\left\|\mathbf{p}_{y|0}^{(1)} - \mathbf{p}_{y|0}^{(1)}\right\|_1 \leq 2\delta\kappa^{(1)} \log\left[8(1+2\delta)\right] =: w(\delta). \tag{47}$$

The preceding display and (46) imply

$$-\int_{y=0}^{\infty} \exp\left(-\int_{t=0}^{y} k_2(t \mid x)\mathrm{d}t\right) R_2(y \mid x)\,\mathrm{d}y \geq -\frac{v(\delta)w(\delta)}{\left[\lambda^{(1)}\right]^2}.$$

Putting it together, we have

$$\int_{y=0}^{\infty} \sqrt{f_2^{(1)}(y \mid 0) f_2^{(2)}(y \mid 0)}\, \mathrm{d}y \geq \frac{1-2\delta}{1+\delta}\left(1 - \frac{v(\delta)w(\delta)}{[\lambda^{(1)}]^2}\right) =: z.$$

Expanding z in powers of δ, we find that

$$z = 1 - \left(3 + \frac{20\kappa^{(1)}\log(64)}{[\lambda^{(1)}]^2}\right)\delta + \mathcal{O}(\delta^2).$$

The Hellinger distance then satisfies

$$d_H^2(f^{(1)}, f^{(2)}) \leq 1 - z.$$

As δ tends to 0, we eventually have $0 < z < 1$. Finally, we have $1 - z \leq \mathcal{O}(\delta)$, so we obtain (39). □

Remark. In the proof above we used Theorem 7 to bound $\|\mathbf{p}_{y|0}^{(1)} - \mathbf{p}_{y|0}^{(2)}\|$ in equation (47). This makes use of the assumption that $\mathbf{p}_{0|0}^{(1)} = \mathbf{p}_{0|0}^{(2)}$, i.e. the starting distributions under the two hypotheses are the same. If we were to consider sample sizes larger than two, we would have to control the difference between the conditional distributions (see equation 3) $\mathbf{p}_{y|x_j}^{(1)}$ and $\mathbf{p}_{y|x_j}^{(2)}$, where x_j is a (random) time at which the j-th coalescent event takes place. This turns out to be difficult without placing additional and somewhat unnatural assumptions on $\mathbf{Q}^{(1)}$ and $\mathbf{Q}^{(2)}$, so the result is limited in its current form to the case $n = 2$. Note that this difficulty is specific to multi-population models and does not arise in the single-population analysis.

6.3 Isolation-with-migration

In this section, we consider a two-island problem where the two populations were part of a panmictic ancestral population until time $\tau > 0$ in the past, sometimes referred to as the isolation-with-migration (IwM) model (Hey and Nielsen, 2007).

Let $\mathbf{c} = (c_0, c_1, c_2)$ be the vector of coalescent rates where $\overline{\mathbf{c}} = (c_1, c_2)$ are for the islands under the two-island portion of the model and c_0 is for the ancestral population. We will need

$$c_{\max} = \max(c_0, c_1, c_2) \qquad\qquad c_{\min} = \min(c_0, c_1, c_2)$$

as well. As in the previous section, we consider the ability to distinguish between two hypothesized models, so $\mathbf{c}^{(i)}$, $\overline{\mathbf{c}}^{(i)}$, $c_{\max}^{(i)}$, etc. are defined for $i = 1, 2$.

The hazard rate function now depends on t and τ:

$$h_j(t \mid x) = \begin{cases} a_j \left[c_1 p_{11}^{\bar{\mathbf{c}},m}(t|x) + c_2 p_{22}^{\bar{\mathbf{c}},m}(t|x) \right] & t \le \tau \\ a_j c_0, & t > \tau. \end{cases}$$

The final result is an analog of Theorem 12 for the case where two IwM models are compared. The proof is similar to the previous theorem, so it is given in the Appendix.

Theorem 15. *Let H_1 and H_2 be hypotheses with the same ancestral coalescent rate c_0 and two-island rate matrix $\mathbf{Q}$ but different divergence times $\tau^{(1)}$ and $\tau^{(2)} = (1+\delta)\tau^{(1)}$. Suppose L i.i.d. coalescent trees on n individuals are sampled under H_1 or H_2, each with probability $1/2$, then for sufficiently small $\delta > 0$, the Bayes error rate for any classifier is at least $(1 - \Upsilon)/2$, where*

$$\Upsilon^2 \le 2L \left\{ 1 - (1 - 7\delta)^{n-1} \right\} \le 14L(n-1)\delta.$$

7 Discussion

In this paper, we studied upper and lower bounds for parameter estimation in the two-island model with migration. In Section 5 we derived some upper bounds on estimation error of the migration and coalescence parameters in the symmetric two-island model, and confirmed by simulations that our theoretical results are accurate (up to constant factors). In Section 6, we obtained lower bounds on the Bayes error rate for distinguishing between different two-island and isolation-with-migration models. Our results have basically the same consistent message: if the "sample size" Ln is much smaller than $1/\delta$, where δ is some measure of relative closeness between the hypotheses, then no procedure is able to reliably distinguish between them on the basis of sampled coalescent trees.

It is instructive to compare our results to those of J. Kim, Mossel, Rácz, et al., which inspired the present work and whose proof techniques we have adapted. Our results differ by leading order in δ (η, in their notation): J. Kim, Mossel, Rácz, et al. obtain $\Upsilon^2 \le \mathcal{O}(Ln\delta^2)$ whereas the bounds in this paper are merely (at worst) $\mathcal{O}(Ln\delta)$. The bounds have different leading orders, and for small δ theirs is tighter. In the J. Kim, Mossel, Rácz, et al. paper, the simplicity of their model implies a direct relationship between the hazard rates of the original and perturbed models (see equation 37, above), which leads to a sharp rate via the argument summarized in Section 6.2. The situation is not as simple when there are multiple populations,

and we have to settle for bounds (Lemmas 13 and 14) instead of equality to quantify the relationship between h_1 and h_2. These culminate in the somewhat looser result of Theorem 12. At present, we do not know whether the difference is due to our proof method, or whether estimating the coalescent and migration rates may be easier under a multi-population model. Note that their setting is not technically a special case of the one we consider here since we need to assume that $\mathbf{m} \neq 0$ in the definition of the model (2); if we do not assume this, the condition number $\kappa \to \infty$ and many of the results in Section 4 become vacuous.

The models we have analyzed here are very basic, consisting of only a few parameters and at most two populations. Even if this restricted setting, the theoretical analysis is already cumbersome. Nowadays, significantly larger and more complicated models involving many populations and migration events between them are routinely estimated from large genetic datasets; there is a large gap between theory and practice. We have attempted to fill that gap, but there are many possible extensions and avenues for future work. In particular, we are not able to say anything about likelihood-based estimation in multi-population models, despite it being by far the dominant mode of method of estimation in applications. A useful, though seemingly difficult, future direction would be to study the likelihood function of genetic data under multi-population models with migration.

Acknowledgments

This research was supported by the National Science Foundation (grant numbers DMS-1646108 and DMS-2052653).

References

Allman, Elizabeth S, James H Degnan, and John A Rhodes (2011). "Identifying the rooted species tree from the distribution of unrooted gene trees under the coalescent". In: *Journal of Mathematical Biology* 62.6, pp. 833–862. DOI: 10.1007/s00285-010-0355-7.

Asmussen, Soren and Hansjorg Albrecher (2010). *Ruin Probabilities (2nd Edition)*. Vol. 2nd ed. Advanced Series on Statistical Science and Applied Probability. World Scientific.

Baharian, Soheil and Simon Gravel (Mar. 2018). "On the decidability of population size histories from finite allele frequency spectra". en. In: *Theor. Popul. Biol.* 120, pp. 42–51.

Bhaskar, Anand and Yun S Song (2014). *Descartes' rule of signs and the identifiability of population demographic models from genomic variation data.*

Casella, George and Roger Berger (2001). *Statistical Inference*. 2nd. Cengage Learning.

Devroye, Luc, Laszlo Györfi, and Gabor Lugosi (Nov. 2013). *A Probabilistic Theory of Pattern Recognition*. en. Springer Science & Business Media.

Durrett, R. (2008). *Probability Models for DNA Sequence Evolution*. 2nd. Springer, New York.

Excoffier, Laurent et al. (2013). "Robust Demographic Inference from Genomic and SNP Data". In: *PLoS Genetics* 9.10, e1003905.

Flouri, Tomáš et al. (Dec. 2019). "A Bayesian Implementation of the Multispecies Coalescent Model with Introgression for Phylogenomic Analysis". In: *Molecular Biology and Evolution* 37.4, pp. 1211–1223. DOI: 10.1093/molbev/msz296.

Gazave, Elodie et al. (2013). "Population growth inflates the per-individual number of deleterious mutations and reduces their mean effect". In: *Genetics* 195.3, pp. 969–978.

Gravel, Simon et al. (2011). "Demographic history and rare allele sharing among human populations". In: *Proceedings of the National Academy of Sciences* 108.29, pp. 11983–11988.

Gutenkunst, Ryan N et al. (Oct. 2009). "Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data". en. In: *PLoS Genet.* 5.10, e1000695.

Hey, Jody, Yujin Chung, et al. (Aug. 2018). "Phylogeny Estimation by Integration over Isolation with Migration Models". In: *Molecular Biology and Evolution* 35.11, pp. 2805–2818. DOI: 10.1093/molbev/msy162.

Hey, Jody and Rasmus Nielsen (Feb. 2007). "Integration within the Felsenstein equation for improved Markov chain Monte Carlo methods in population genetics". en. In: *Proc. Natl. Acad. Sci. U. S. A.* 104.8, pp. 2785–2790. ISSN: 0027-8424. DOI: 10.1073/pnas.0611164104.

Hobolth, Asger, Arno Siri-Jegousse, and Mogens Bladt (2019). "Phase-type distributions in population genetics". In: *Theoretical population biology* 127, pp. 16–32.

Horn, Roger A and Charles R Johnson (2012). *Matrix analysis*. Cambridge university press.

Johndrow, James E and Julia A Palacios (Feb. 2019). "Exact limits of inference in coalescent models". en. In: *Theor. Popul. Biol.* 125, pp. 75–93.

Jouganous, Julien et al. (July 2017). "Inferring the Joint Demographic History of Multiple Populations: Beyond the Diffusion Approximation". en. In: *Genetics*

206.3, pp. 1549–1567. ISSN: 0016-6731, 1943-2631. DOI: 10.1534/genetics.117.200493.

Kamm, Jack, Jonathan Terhorst, Richard Durbin, et al. (2020). "Efficiently inferring the demographic history of many populations with allele count data". en. In: *J. Am. Stat. Assoc.* 115.531, pp. 1472–1487. ISSN: 0162-1459. DOI: 10.1080/01621459.2019.1635482.

Kamm, Jack, Jonathan Terhorst, and Yun S Song (Feb. 2017). "Efficient computation of the joint sample frequency spectra for multiple populations". en. In: *J. Comput. Graph. Stat.* 26.1, pp. 182–194. ISSN: 1061-8600. DOI: 10.1080/10618600.2016.1159212.

Kim, Junhyong, Elchanan Mossel, Miklós Z. Rácz, et al. (2015a). "Can one hear the shape of a population history?" In: *Theoretical Population Biology* 100, pp. 26–38. ISSN: 0040-5809. DOI: 10.1016/j.tpb.2014.12.002.

— (2015b). "Can one hear the shape of a population history?" In: *Theoretical population biology* 100, pp. 26–38.

Kim, Younhun et al. (Apr. 2020). "How Many Subpopulations Is Too Many? Exponential Lower Bounds for Inferring Population Histories". en. In: *J. Comput. Biol.* 27.4, pp. 613–625.

Li, Heng and Richard Durbin (July 2011). "Inference of human population history from individual whole-genome sequences". en. In: *Nature* 475.7357, pp. 493–496.

Mirarab, S. et al. (2014). "ASTRAL: Genome-scale coalescent-based species tree estimation". In: *Bioinformatics* 30.17, pp. i541–i548. DOI: 10.1093/bioinformatics/btu462.

Mitrophanov, A. Yu. (2003). "Stability and exponential convergence of continuous-time Markov chains". In: *Journal of Applied Probability* 40.4, pp. 970–979. DOI: 10.1239/jap/1067436094.

Myers, Simon, Charles Fefferman, and Nick Patterson (May 2008). "Can one learn history from the allelic spectrum?" en. In: *Theor. Popul. Biol.* 73.3, pp. 342–348.

Notohara, M (1990). "The coalescent and the genealogical process in geographically structured population". en. In: *J. Math. Biol.* 29.1, pp. 59–75. ISSN: 0303-6812. DOI: 10.1007/BF00173909.

Rannala, Bruce and Ziheng Yang (2003). "Bayes Estimation of Species Divergence Times and Ancestral Population Sizes Using DNA Sequences From Multiple Loci". In: *Genetics* 164.4, pp. 1645–1656.

Rodríguez, Willy et al. (Dec. 2018). "The IICR and the non-stationary structured coalescent: towards demographic inference with arbitrary changes in population structure". en. In: *Heredity* 121.6, pp. 663–678. ISSN: 0018-067X, 1365-2540. DOI: 10.1038/s41437-018-0148-0.

Sawyer, Stanley A. and Daniel L. Hartl (1992). "Population genetics of polymorphism and divergence." In: *Genetics* 132.4, pp. 1161–1176.

Schmidli, Hanspeter (1999). "Compound sums and subexponentiality". In: *Bernoulli* 5.6, pp. 999–1012.

Sousa, Vitor C, Aude Grelaud, and Jody Hey (Oct. 2011). "On the nonidentifiability of migration time estimates in isolation with migration models". en. In: *Mol. Ecol.* 20.19, pp. 3956–3962.

Strobeck, Curtis (1987). "Average number of nucleotide differences in a sample from a single subpopulation: a test for population subdivision". In: *Genetics* 117.1, pp. 149–153.

Takahata, N (Dec. 1988). "The coalescent in two partially isolated diffusion populations". en. In: *Genet. Res.* 52.3, pp. 213–222. ISSN: 0016-6723. DOI: 10.1017/s0016672300027683.

Tennessen, Jacob A et al. (2012). "Evolution and Functional Impact of Rare Coding Variation from Deep Sequencing of Human Exomes". In: *Science* 337.6090, pp. 64–69.

Terhorst, Jonathan and Yun S Song (June 2015). "Fundamental limits on the accuracy of demographic inference based on the sample frequency spectrum". en. In: *Proc. Natl. Acad. Sci. U. S. A.* 112.25, pp. 7677–7682.

Vershynin, Roman (2018). *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press. DOI: 10.1017/9781108231596.

Yi, Xin et al. (2010). "Sequencing of 50 human exomes reveals adaptation to high altitude". In: *science* 329.5987, pp. 75–78.

8 Appendix

We first provide proofs of the main propositions of Section 4.

Proof of Proposition 2. By the Gershgorin circle theorem, the eigenvalues of $\mathbf{Q}$ have nonpositive real part, and $\lambda_0 = 0$ since $\mathbf{Q}$ is a rate matrix. Let $\mathbf{Q}'$ be the 3×3 leading principal minor of $\mathbf{Q}$. Then $v = (v_1, v_2, v_3, v_4)^\intercal$ is an eigenvector of $\mathbf{Q}$ with nonzero eigenvalue if and only if $v_4 = 0$ and $(v_1, v_2, v_3)^\intercal$ is an eigenvector of $\mathbf{Q}'$ with the same eigenvalue.

Defining $\mathbf{D} := \mathrm{diag}(1, \sqrt{m_1/(2m_2)}, \sqrt{m_2/(2m_1)})$, we have

$$\mathbf{D}\mathbf{Q}'\mathbf{D}^{-1} = \begin{pmatrix} -(m_1+m_2) & \sqrt{2m_1m_2} & \sqrt{2m_1m_2} \\ \sqrt{2m_1m_2} & -(c_1+2m_2) & 0 \\ \sqrt{2m_1m_2} & 0 & -(c_2+2m_1) \end{pmatrix} =: \mathbf{A}. \tag{48}$$

Thus, $\mathbf{Q}'$ is similar to the Hermitian matrix $\mathbf{A}$, so $\{\lambda_1, \lambda_2, \lambda_3\} \subset \mathbb{R}_{\leq 0}$. In fact, since

$$\det \mathbf{A} = -\left[2\left(c_1 m_1^2 + 2c_2 m_2^2\right) + c_1 c_2 \left(m_1 + m_2\right)\right] < 0,$$

$\mathbf{A}$ has no zero eigenvalues: $\lambda_3 \leq \lambda_2 \leq \lambda_1 < 0$. Finally,

$$-\frac{1}{\lambda_1} \leq -\sum_{i=1}^{3} \frac{1}{\lambda_i} = -\operatorname{tr}[(\mathbf{Q}')^{-1}] = -\operatorname{tr}[\mathbf{A}^{-1}],$$

which yields the lower bound in (8) by direct computation of $\mathbf{A}^{-1}$. Finally, for the upper bound we have

$$\lambda_1 \geq \lambda_1 + \lambda_2 + \lambda_3 = \operatorname{tr} \mathbf{A} = -\left(\|\mathbf{c}\|_1 + 3\|\mathbf{m}\|_1\right).$$

□

Proof of Lemma 4. Let $\lambda_3 \leq \lambda_2 \leq \lambda_1 < 0$ be defined as in Proposition 2. By Proposition IX.1.8 of Asmussen and Albrecher (2010), $\mathbb{P}(\zeta_{\boldsymbol{\alpha}} > t) \asymp e^{\lambda_1 t}$ as $t \to \infty$, which implies that

$$\int_0^\infty e^{r\zeta_{\boldsymbol{\alpha}}} \, \mathrm{d}\mathbb{P}(\zeta_{\boldsymbol{\alpha}}) < \infty$$

for $r < |\lambda_1|$. Next, by Proposition IX.1.7 of Asmussen and Albrecher,

$$M_{\zeta_{\boldsymbol{\alpha}}}(r) = \boldsymbol{\alpha}(-r\mathbf{I} - \mathbf{Q}')^{-1}\mathbf{c}. \tag{49}$$

With $\boldsymbol{\Lambda} = \operatorname{diag}(\lambda_1, \lambda_2, \lambda_3)$ and $0 < r < |\lambda_1|$, we get

$$\begin{aligned}
\left\|(-r\mathbf{I} - \mathbf{Q}')^{-1}\right\| &= \left\|(r\mathbf{I} + \boldsymbol{\Lambda})^{-1}\right\| \\
&= \max_i |r + \lambda_i|^{-1} \\
&= \left(\min_i |r + \lambda_i|\right)^{-1} \\
&= \frac{1}{r - \lambda_1} \leq \frac{1}{|\lambda_1|}.
\end{aligned}$$

The multiplication of $\mathbf{Q}$ with the fourth standard basis vector $\mathbf{e}_4 = (1, 0, 0, 0)^T$ equals the fourth column of $\mathbf{Q}$, i.e. $\mathbf{c}$. The result follows from (49) and the facts that

$$\begin{aligned}
\|\mathbf{c}\| = \|\mathbf{Q}\mathbf{e}_4\| \leq \|\mathbf{Q}\| \\
\|\boldsymbol{\alpha}\| \leq \|\boldsymbol{\alpha}\|_1 = 1.
\end{aligned} \tag{50}$$

□

We now prove a sequence of results that enter into the proof of Theorem 7.

Proposition 16. *For any two initial distributions* $\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2$*, we have*

$$\left\|\boldsymbol{\alpha}_1 e^{t\mathbf{Q}} - \boldsymbol{\alpha}_2 e^{t\mathbf{Q}}\right\|_1 \leq 16\kappa e^{-\lambda t}.$$

Proof. Let $\mathbb{P}_{\boldsymbol{\alpha}}(X(t) = s)$ be the sth component of $\boldsymbol{\alpha} e^{t\mathbf{Q}}$. For $i = 1, 2$ define

$$u_i = \mathbb{P}_{\boldsymbol{\alpha}_i}(X(t) = \text{coal}) = 1 - \mathbb{P}(\zeta_{\boldsymbol{\alpha}_i} > t)$$

for $\zeta_{\boldsymbol{\alpha}_i}$ defined by (7). Then for $s \neq \text{coal}$,

$$\mathbb{P}_{\boldsymbol{\alpha}_i}(X(t) = s) \leq 1 - u_i.$$

By repeated applications of the triangle inequality,

$$\begin{aligned}
\left\|\boldsymbol{\alpha}_1 e^{t\mathbf{Q}} - \boldsymbol{\alpha}_2 e^{t\mathbf{Q}}\right\|_1 &= \sum_{s\in\mathcal{S}} \left|\mathbb{P}_{\boldsymbol{\alpha}_1}(X(t) = s) - \mathbb{P}_{\boldsymbol{\alpha}_2}(X(t) = s)\right| \\
&= \left|\mathbb{P}_{\boldsymbol{\alpha}_1}(X(t) = \text{coal}) - \mathbb{P}_{\boldsymbol{\alpha}_2}(X(t) = \text{coal})\right| \\
&\quad + \sum_{s\neq\text{coal}} \left|\mathbb{P}_{\boldsymbol{\alpha}_1}(X(t) = s) - \mathbb{P}_{\boldsymbol{\alpha}_2}(X(t) = s)\right| \\
&= \left|\mathbb{P}(\zeta_{\boldsymbol{\alpha}_1} > t) - \mathbb{P}(\zeta_{\boldsymbol{\alpha}_2} > t)\right| + 6\max\{1 - u_1, 1 - u_2\} \\
&\leq 8\max\{1 - u_1, 1 - u_2\}.
\end{aligned}$$

Finally, by Corollary 5,

$$\max\{1 - u_1, 1 - u_2\} \leq (\|\boldsymbol{\alpha}_1\| + \|\boldsymbol{\alpha}_2\|)\, \|\mathbf{Q}\|\, \lambda^{-1} e^{-\lambda t}.$$

□

Lemma 17. *If* $\mathbf{Q}^{(2)}$ *is* δ*-close to* $\mathbf{Q}^{(1)}$ *then*

$$|\lambda^{(1)} - \lambda^{(2)}| \leq 2\delta \left\|\mathbf{Q}^{(1)}\right\|. \tag{51}$$

Proof. As in the proof of Proposition 2, let $\mathbf{A}^{(1)}$ and $\mathbf{A}^{(2)}$ be the Hermitian matrices to which (the upper-left submatrices of) $\mathbf{Q}^{(1)}$ and $\mathbf{Q}^{(2)}$ are similar, cf. equation (48). By Weyl's eigenvalue perturbation theorem (e.g., Horn and Johnson, 2012, Theorem 4.3.1)

$$|\lambda_1^{(1)} - \lambda_1^{(2)}| \leq \left\|\mathbf{A}^{(1)} - \mathbf{A}^{(2)}\right\|.$$

By the stated assumptions, we have for the entries of $\mathbf{A}^{(1)}$ and $\mathbf{A}^{(2)}$,

$$|a_{ij}^{(1)} - a_{ij}^{(2)}| < a_{ij}^{(1)}\delta$$

which implies that

$$\left\|\mathbf{A}^{(1)} - \mathbf{A}^{(2)}\right\|_F \leq \delta \left\|\mathbf{A}^{(1)}\right\|_F.$$

Hence,

$$\|\mathbf{A}^{(1)} - \mathbf{A}^{(2)}\| \leq \|\mathbf{A}^{(1)} - \mathbf{A}^{(2)}\|_F \leq \delta\|\mathbf{A}^{(1)}\|_F \leq 2\delta\|\mathbf{A}^{(1)}\| \leq 2\delta\|\mathbf{Q}^{(1)}\|,$$

where the final inequality is because

$$\left\|\mathbf{A}^{(1)}\right\| = \sup_{\substack{\mathbf{v}\in\mathbb{R}^3 \\ \|\mathbf{v}\|=1}} \left\|\mathbf{A}^{(1)}\mathbf{v}\right\| \leq \sup_{\substack{\mathbf{v}\in\mathbb{R}^4 \\ \|\mathbf{v}\|=1}} \left\|\mathbf{Q}^{(1)}\mathbf{v}\right\| = \left\|\mathbf{Q}^{(1)}\right\|.$$

□

Proposition 18. *Suppose* $\mathbf{Q}^{(1)}$ *and* $\mathbf{Q}^{(2)}$ *are* δ*-close for some* $\delta < 1/(4\kappa^{(1)})$*. Then for all* $t \geq 0$*,*

$$\left\|\mathbf{p}_{t|x}^{(2)} - \mathbf{p}_{t|x}^{(1)}\right\|_1 \leq 8(1+2\delta)\kappa^{(1)}e^{-\lambda^{(1)}t/2},$$

Proof. By Lemma 17 and the assumptions,

$$\min\{\lambda^{(1)}, \lambda^{(2)}\} \geq \lambda^{(1)} - 2\delta\left\|\mathbf{Q}^{(1)}\right\| > \lambda^{(1)}/2. \tag{52}$$

Observing that $\mathbf{e}_4 = (0,0,0,1)^\intercal$ is a left eigenvector of $e^{t\mathbf{Q}}$ for any t and $\mathbf{Q}$ defined by (2), we have

$$\begin{aligned}
\left\|\mathbf{p}_{t|x}^{(2)} - \mathbf{p}_{t|x}^{(1)}\right\|_1 &\leq \left\|\mathbf{p}_{t|x}^{(2)} - \mathbf{e}_4\right\|_1 + \left\|\mathbf{e}_4 - \mathbf{p}_{t|x}^{(1)}\right\|_1 \\
&= \left\|\mathbf{p}_{t|x}^{(2)} - \mathbf{e}_4 e^{t\mathbf{Q}^{(2)}}\right\|_1 + \left\|\mathbf{e}_4 e^{t\mathbf{Q}^{(1)}} - \mathbf{p}_{t|x}^{(1)}\right\|_1, \\
&\leq 2\left(\|\mathbf{c}^{(1)}\|e^{-\lambda^{(1)}t}/\lambda^{(1)} + \|\mathbf{c}^{(2)}\|e^{-\lambda^{(2)}t}/\lambda^{(2)}\right) \\
&\leq 4(1+2\delta)\|\mathbf{Q}^{(1)}\|\left(e^{-\lambda^{(1)}t}/\lambda^{(1)} + e^{-\lambda^{(2)}t}/\lambda^{(2)}\right) \qquad (53) \\
&\leq 8(1+2\delta)\|\mathbf{Q}^{(1)}\|e^{-\lambda^{(1)}t/2}/\lambda^{(1)}, \qquad (54)
\end{aligned}$$

where inequality (53) follows from equations (50) and (12), and inequality (54) follows from (52) and the fact that $x \mapsto e^{-xt}/x$ is decreasing. □

Proof of Theorem 7. In the notation of Mitrophanov (2003), Proposition 16 implies $b = \lambda^{(1)}/2$, $c = 64(1+2\delta) > 2$ in his equation (2.1), and (13) implies $\mathbf{z}(t) = 0$. Plugging these constants into Mitrophanov's equation (2.9) and using (11), we obtain

$$\left\|\mathbf{p}_{t|x}^{(2)} - \mathbf{p}_{t|x}^{(1)}\right\|_1 \leq \frac{2\delta \left\|\mathbf{Q}^{(1)}\right\|}{\lambda^{(1)}} \log\left[64(1+2\delta)\right].$$

□

We now give proofs of the two technical lemmas in Section 6.2.

Proof of Lemma 13. Letting $\mathbf{P}^{(i)} = \operatorname{diag}(0, c_1^{(i)}, c_2^{(i)}, 0)$ and recalling equation (10), we have

$$\begin{aligned}
&a_j^{-1}\left|h_j^{(1)} - h_j^{(2)}\right| \\
&= \left|\left(\mathbf{p}_{y|x}^{(1)}\mathbf{P}^{(1)} - \mathbf{p}_{y|x}^{(2)}\mathbf{P}^{(2)}\right)\mathbf{1}\right| \\
&= \left|\left[\mathbf{p}_{y|x}^{(1)} - \mathbf{p}_{y|x}^{(2)}\left(\mathbf{I} + \mathbf{D}_c\right)\right]\mathbf{P}^{(1)}\mathbf{1}\right| \\
&= \left|\left[\mathbf{p}_{y|x}^{(1)}\mathbf{D}_c + \left(\mathbf{p}_{y|x}^{(1)} - \mathbf{p}_{y|x}^{(2)}\right)\left(\mathbf{I} + \mathbf{D}_c\right)\right]\mathbf{P}^{(1)}\mathbf{1}\right| \\
&\leq \|\mathbf{D}_c\|_1 \left\|\mathbf{p}_{y|x}^{(1)}\mathbf{P}^{(1)}\right\|_1 + \|\mathbf{I} + \mathbf{D}_c\|_1 \left\|\left(\mathbf{p}_{y|x}^{(1)} - \mathbf{p}_{y|x}^{(2)}\right)\mathbf{P}^{(1)}\right\|_1 \\
&\leq \delta h_j^{(1)}(y \mid x) + \left\|\mathbf{p}_{y|x}^{(1)} - \mathbf{p}_{y|x}^{(2)}\right\|_1 \left\|\mathbf{P}^{(2)}\right\|_1.
\end{aligned} \tag{55}$$

By equations (50) and (40),

$$\left\|\mathbf{P}^{(2)}\mathbf{1}\right\|_1 \leq \left\|\mathbf{P}^{(2)}\mathbf{1}\right\|_\infty = \left\|\mathbf{c}^{(2)}\right\|_\infty \leq 2\left\|\mathbf{Q}^{(2)}\right\| \leq 2(1+2\delta),$$

establishing (41). Inequality (42) follows by writing

$$\frac{h_j^{(1)} + h_j^{(2)}}{2} = h_j^{(1)} + \frac{1}{2}\left(h_j^{(2)} - h_j^{(1)}\right) \tag{56}$$

and using (41). □

Proof of Lemma 14. Using the identity

$$\sqrt{ab} = a + \frac{\sqrt{a}}{\sqrt{a} + \sqrt{b}}(b - a), \quad a \wedge b > 0,$$

we have

$$\begin{aligned}
&a_j^{-1}\sqrt{h_j^{(1)}h_j^{(2)}} \\
&\geq \left\|\mathbf{p}_{y|x}^{(1)}\mathbf{P}^{(1)}\right\|_1^{1/2}\left\|\mathbf{p}_{y|x}^{(2)}\mathbf{P}^{(1)}\right\|_1^{1/2} \\
&= \left\|\mathbf{p}_{y|x}^{(1)}\mathbf{P}^{(1)}\right\|_1 + \frac{\left\|\mathbf{p}_{y|x}^{(1)}\mathbf{P}^{(1)}\right\|_1^{1/2}}{\left\|\mathbf{p}_{y|x}^{(1)}\mathbf{P}^{(1)}\right\|_1^{1/2} + \left\|\mathbf{p}_{y|x}^{(2)}\mathbf{P}^{(1)}\right\|_1^{1/2}}\left(\left\|\mathbf{p}_{y|x}^{(2)}\mathbf{P}^{(1)}\right\|_1 - \left\|\mathbf{p}_{y|x}^{(1)}\mathbf{P}^{(1)}\right\|_1\right) \\
&\geq \left\|\mathbf{p}_{y|x}^{(1)}\mathbf{P}^{(1)}\right\|_1 - \left\|\mathbf{P}^{(1)}\right\|_1\left\|\mathbf{p}_{y|x}^{(1)} - \mathbf{p}_{y|x}^{(2)}\right\|_1.
\end{aligned}$$

Multiplying both sides by a_j implies the result. □

We conclude with the proof of the main result.

Proof of Theorem 15. The proof is along the same lines as that of Theorem 12. Let $h_j^{(1)}$ correspond to $\tau^{(1)}$ and $h_j^{(2)}$ to $\tau^{(2)}$. We have

$$h_j^{(1)}(t \mid x_{j-1}) = h_j^{(2)}(t \mid x_{j-1}) \quad \forall t \notin [\tau^{(1)}, \tau^{(2)}].$$

Otherwise,

$$h_j^{(1)}(t \mid x) - h_j^{(2)}(t \mid x) = (c_0 - \|\mathbf{p}_{t|x}^{(2)}\mathbf{P}_{\bar{c}}\|_1)a_j.$$

We quantify the difference in hazard rates by adding and subtracting terms. Let $\mathbf{1}_{\{\tau^{(1)} \leq t \leq \tau^{(2)}\}}$ be the indicator function that equals 1 whenever $\tau^{(1)} \leq t \leq \tau^{(2)}$ and equals 0 otherwise. Then

$$\begin{aligned}
&\int_{t=x}^{y} \frac{h_j^{(1)}(t \mid x) + h_j^{(2)}(t \mid x)}{2}\,\mathrm{d}t \\
&= \int_{t=x}^{y} h_j^{(1)}(t \mid x)\,\mathrm{d}t + \int_{t=x}^{y} \frac{h_j^{(2)}(t \mid x) - h_j^{(1)}(t \mid x)}{2}\,\mathrm{d}t \\
&\leq \int_{t=x}^{y} h_j^{(1)}(t \mid x)\,\mathrm{d}t + \frac{1}{2}\int_{t=x}^{y} c_{\max}a_j\left(1 - \frac{1}{c_{\max}}\|\mathbf{p}_{t|x}^{(2)}\mathbf{P}_{\bar{c}}\|_1\right)\mathbf{1}_{\{\tau^{(1)} \leq t \leq \tau^{(2)}\}}\,\mathrm{d}t \\
&\leq \int_{t=x}^{y} h_j^{(1)}(t \mid x)\,\mathrm{d}t + \frac{1}{2}\int_{t=x}^{y} c_{\max}a_j\mathbf{1}_{\{\tau^{(1)} \leq t \leq \tau^{(2)}\}}\,\mathrm{d}t.
\end{aligned}$$

Now for some $\alpha \in [0,1]$ we have

$$\begin{aligned}
&\sqrt{h_j^{(1)}(y \mid x) h_j^{(2)}(y \mid x)} \geq h_j^{(1)}(y \mid x) - \alpha \left| h_j^{(1)}(y \mid x) - h_j^{(2)}(y \mid x) \right| \\
&\geq h_j^{(1)}(y \mid x) - \left| c_0 - \|\mathbf{p}_{t|x}^{(2)} \mathbf{P}_{\bar{c}}\|_1 \right| a_j \mathbf{1}_{\{\tau^{(1)} \leq y \leq \tau^{(2)}\}} \\
&\geq h_j^{(1)}(y \mid x) - 3 c_{\max} a_j \mathbf{1}_{\{\tau^{(1)} \leq y \leq \tau^{(2)}\}}.
\end{aligned}$$

Combining the preceding displays, we get

$$\begin{aligned}
&\int_{y=x}^{\infty} \sqrt{f_j^{(1)}(y|x) f_j^{(2)}(y|x)}\, \mathrm{d}y \\
&\geq \int_{y=x}^{\infty} \exp\left(- \int_{t=x}^{y} \left[h_j^{(1)}(t \mid x) + \frac{1}{2} c_{\max} a_j \mathbf{1}_{\{\tau^{(1)} \leq t \leq \tau^{(2)}\}} \right] \mathrm{d}t \right) \\
&\times \left[h_j^{(1)}(y \mid x) + \frac{1}{2} c_{\max} a_j \mathbf{1}_{\{\tau^{(1)} \leq y \leq \tau^{(2)}\}} \right. \\
&\left. - \frac{1}{2} c_{\max} a_j \mathbf{1}_{\{\tau^{(1)} \leq y \leq \tau^{(2)}\}} - 3 c_{\max} a_j \mathbf{1}_{\{\tau^{(1)} \leq y \leq \tau^{(2)}\}} \right] \mathrm{d}y.
\end{aligned}$$

The first part integrates to one:

$$\begin{aligned}
&\int_{y=x}^{\infty} \exp\left(- \int_{t=x}^{y} \left[h_j^{(1)}(t \mid x) + \frac{1}{2} c_{\max} a_j \mathbf{1}_{\{\tau^{(1)} \leq t \leq \tau^{(2)}\}} \right] \mathrm{d}t \right) \\
&\times \left[h_j^{(1)}(y \mid x) + \frac{1}{2} c_{\max} a_j \mathbf{1}_{\{\tau^{(1)} \leq y \leq \tau^{(2)}\}} \right] \mathrm{d}y = 1.
\end{aligned}$$

For the other part,

$$\begin{aligned}
&\int_{y=x}^{\infty} \exp\left(- \int_{t=x}^{y} \left[h_j^{(1)}(t \mid x) + \frac{1}{2} c_{\max} a_j \mathbf{1}_{\{\tau^{(1)} \leq t \leq \tau^{(2)}\}} \right] \mathrm{d}t \right) \mathrm{d}y \\
&\leq \int_{y=x}^{\infty} e^{-c_{\max} a_j (y-x)/2} \mathrm{d}y = \frac{2}{c_{\max} a_j}.
\end{aligned}$$

This implies

$$
\begin{aligned}
&- \int_{y=x}^{\infty} \exp\left(- \int_{t=x}^{y} \left[h_j^{(1)}(t \mid x) + \frac{1}{2} c_{\max} a_j \mathbf{1}_{\{\tau^{(1)} \le t \le \tau^{(2)}\}} \right] \mathrm{d}t \right) \\
&\times \left[\frac{1}{2} c_{\max} a_j \mathbf{1}_{\{\tau^{(1)} \le y \le \tau^{(2)}\}} + 3 c_{\max} a_j \mathbf{1}_{\{\tau^{(1)} \le y \le \tau^{(2)}\}} \right] \mathrm{d}y \\
&\ge - \frac{2}{c_{\max} a_j} \left[\frac{1}{2} c_{\max} a_j + 3 c_{\max} a_j \right] \int_{y=x}^{\infty} \mathbf{1}_{\{\tau^{(1)} \le y \le \tau^{(2)}\}} \mathrm{d}y \\
&\ge -7\delta.
\end{aligned}
$$

Putting it together, we have

$$
\int_{y=x}^{\infty} \sqrt{f_j^{(1)}(y \mid x) f_j^{(2)}(y \mid x)} \mathrm{d}y \ge 1 - 7\delta
$$

Hence,

$$
d_H^2(f^{(1)}, f^{(2)}) \le 1 - (1 - 7\delta)^{n-1},
$$

giving the first inequality. With $z := 1 - 7\delta$, we use the bound $1 - z^{n-1} \le (n-1)(1-z)$ to obtain the second inequality. □

Author Agreement Statement

We the undersigned declare that this manuscript is original, has not been published before and is not currently being considered for publication elsewhere.

We confirm that the manuscript has been read and approved by all named authors and that there are no other persons who satisfied the criteria for authorship but are not listed. We further confirm that the order of authors listed in the manuscript has been approved by all of us.

We understand that the Corresponding Author is the sole contact for the Editorial process. He/she is responsible for communicating with the other authors about progress, submissions of revisions and final approval of proofs

Signed by all authors as follows: