Identifiability and inference of phylogenetic birth-death models

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

August 26, 2022

Abstract

Recent theoretical work on phylogenetic birth-death models offers differing viewpoints on whether they can be estimated using lineage-through-time data. Louca and Pennell (2020) showed that the class of models with continuously differentiable rate functions is nonidentifiable: any such model is consistent with an infinite collection of alternative models, which are statistically indistinguishable regardless of how much data are collected. Legried and Terhorst (2021a) qualified this grave result by showing that identifiability is restored if only piecewise constant rate functions are considered.

Here, we contribute new theoretical results to this discussion, in both the positive and negative directions. Our main result is to prove that models based on piecewise polynomial rate functions of any order and with any (finite) number of pieces are statistically identifiable. In particular, this implies that spline-based models with an arbitrary number of knots are identifiable. The proof is simple and self-contained, relying only on basic algebra. We complement this positive result with a negative one, which shows that even when identifiability holds, rate function estimation is still a difficult problem. To illustrate this, we prove some rates-of-convergence results for hypothesis testing procedures on birth-death histories. These results are information-theoretic lower bounds, which apply to all potential estimators of birth-death models.

1 Introduction

The linear birth-death (BD) process (Feller, 1939; Kendall, 1948) has been used to study population growth in a variety of settings. Recently, in phylogenetics, it has also served as a model of tree formation, by viewing the surviving lineages of a tree as members of population, which randomly give birth to other lineages, or go extinct. Often, the rates at which these “births” and “deaths” occur in a phylogeny touch on important evolutionary questions. For example Nee et al. (1994); Quental and Marshall (2010); Morlon et al. (2011) used phylogenetic BD models to study extinction and speciation dynamics; Gernhard (2008); Heath et al. (2014) used them to calibrate divergence times; and Stadler (2009, 2010); Stadler et al. (2013) investigated the dynamics of pathogens in an infection tree. Estimating the rate at which lineages are born and die in an observed phylogeny is not trivial: deaths are not recorded at all, and the apparent rate of births will also be biased downwards, because some species went extinct before the present, or were simply not sampled. The birth-death model of tree formation provides a principled way to correct these biases.

Despite its widespread use, serious questions have recently been raised about whether it is even possible to estimate this model from phylogenetic data. Stadler (2009) showed that even when the birth and death rates are assumed constant over time, the birth-death model with present-day sampling fails to be identifiable if the sampling probability is not known in advance. That is, multiple distinct models produce exactly the same distribution over the observable data. Other studies have incorporated time-varying per capita birth and death rates in attempting to be more biologically realistic (Stadler et al., 2013). However, Louca and Pennell (2020) have recently shown that birth-death models with smoothly varying rate functions are unidentifiable from extant timetrees, meaning they cannot be reliably estimated using any amount of data.

In Legried and Terhorst (2021a), this grave finding was partly qualified by showing that a smaller class of candidate models, consisting of birth and death rates which are piecewise constant, is identifiable given a sufficiently large timetree. This sample size is small enough for many applications, such as the phylodynamic analysis of pathogens. A potential objection is that the class of piecewise-constant models considered by Legried and Terhorst (2021a) might not be sufficiently large if one believes that the birth and death rates could be continuous or satisfy other smoothness criteria. A partial response to this issue is given in their Theorem 5. Broadly speaking, that result states that any unidentifiable, but reasonably smooth, model can be approximated by an identifiable one. However, there is a sampling size requirement to have provable identifiability that diverges as the approximation error goes to zero. In their Conjecture 6, they suggest that their results extend to models with piecewise-polynomial (or spline) birth and death rates, with a similar sampling requirement to the piecewise-constant case. Resolution of this conjecture would give a sampling requirement for polynomial birth and death rates to be identifiable, but mathematical difficulty left the question open.

In this paper, we prove the following stronger version of their Conjecture 6: piecewise-polynomial models defined on an arbitrary (but finite) number of pieces are identifiable from extant timetree data. Further, the sample size requirement is removed, as we now possess a refined view of what constitutes a sample in this model. Besides removing the sampling requirement, our results demonstrate that it is *not* the presence of jump discontinuities that leads to identifiability. Indeed, our main result (Theorem 1) establishes the existence of identifiable model classes containing smooth (C^∞) birth and death rate functions. The proof relies on the fact that polynomials, uniquely and by definition, have finite power series. In contrast, the proof technique of Legried and Terhorst (2021a) depends on solving a certain differential equation satisfied by the rate functions, which is difficult to extend to functions that are not constant.

Identifiability is a minimal regularity criterion which must be possessed by any useful statistical model. Even when it holds, estimating rate functions from an extant time-tree is often difficult. Using methods from survival analysis (and following techniques developed by Kim et al., 2015; Legried and Terhorst, 2021b) we go on to show that hypothesis testing for constant rates requires substantial data, given quantitatively in the upcoming Theorems. This is an information-theoretic result, which extends to all potential estimators of the model. Moreover, as these results assume perfect knowledge of the underlying time-tree, the problem could in fact be much harder given limited or error-prone data.

2 Background

Let an extant timetree with $N \geq 2$ leaves be given. Under this assumption, there are $N-1$ branching times. The timetree is *extant* in that it traces out the observed ancestry of the sample. The branching times are denoted $\tau_1 > \ldots > \tau_{N-1} > 0$ where τ_i is the time of the ith branching point. These times are expressed looking backwards, with time 0 being the present and any positive time being the amount of time in the past. All N leaves are taken to be sampled in the present; that is, at time 0. There is also a source node called the *origin*, whose age is τ_o. In practice, extant timetrees are observed conditionally on non-extinction of the birth-death process up to the present. Taking τ_o as either random or deterministic, we condition on survival of the population for a period of length τ_o.

Extant timetrees are modeled using a linear birth-death (BD) process. Three parameters determine the dynamics. Two of them are the per-capita birth and death rate functions $\lambda : \mathbb{R}_{\geq 0} \to \mathbb{R}_{>0}$ and $\mu : \mathbb{R}_{\geq 0} \to \mathbb{R}_{>0}$. At each time looking in the past, λ and μ give the instantaneous rate that a given lineage gives birth to a new lineage or dies. The third parameter $\rho \in (0, 1]$ is the sampling probability. Each lineage surviving to the present is considered extant with probability ρ independently of all other surviving lineages. The triple (λ, μ, ρ) determines a BD process.

Next, we describe the function space where the birth and death rate functions come from. In this work, we study the case where they are given by piecewise-polynomial functions over the interval $[0, \tau_o]$. Let $\mathcal{P}$ denote the set of all polynomials with real coefficients, i.e.

$$f \in \mathcal{P} \iff \exists n \in \mathbb{Z}_{\geq 0}, \mathbf{a} \in \mathbb{R}^{n+1} \text{ such that } f(x) = a_0 + a_1 x + \cdots a_n x^n.$$

Definition 1. Let $\mathcal{P}^{(\oplus K)}[0, \tau_o]$ be the collection of piecewise polynomials with K internal breakpoints defined over $[0, \tau_o]$:

$$\mathcal{P}^{(\oplus K)}[0, \tau_o] = \left\{ \sum_{k=1}^{K} p_k(t) \mathbf{1}_{[t_{k-1}, t_k)}(t) : p_k \in \mathcal{P}, 0 = t_0 < t_1 < \ldots < t_K = \tau_o \right\}.$$

(The breakpoints $t_0, \ldots, t_K$ may vary between members of $\mathcal{P}^{(\oplus K)}[0, \tau_o]$.) Let $\mathcal{P}^{(\oplus)}[0, \tau_o] = \bigcup_{K=1}^{\infty} \mathcal{P}^{(\oplus K)}[0, \tau_o]$ be the set of piecewise polynomials with any finite number of pieces. Similarly, we define the positive subset of these polynomials, $\mathcal{P}_+^{(\oplus)}[0, \tau_o] \subset \mathcal{P}^{(\oplus)}[0, \tau_o]$, by

$$\mathcal{P}_+^{(\oplus)}[0, \tau_o] = \left\{ p \in \mathcal{P}^{(\oplus)}[0, \tau_o] : \inf_{t \in [0, \tau_o]} p(t) > 0 \right\}.$$

We now define the corresponding collection of models consisting of piecewise-polynomial birth and death rates. In the results that follow, we assume that ρ is known. If ρ is left to be estimated, then the triple (λ, μ, ρ) fails to be identifiable even in the case where λ and μ are assumed constant (Stadler, 2009; Stadler and Steel, 2019). Thus, the model class of interest supposes the sampling probability ρ is known in advance. There are several equivalent ways to define the phylogenetic BD model. In this paper, we find it convenient to work in the following parameterization.

Definition 2. Let

$$\mathcal{I}_\rho = \left\{(\lambda, r, \rho) : \lambda \in \mathcal{P}_+^{(\oplus)}[0, \tau_o], r \in \mathcal{P}^{(\oplus)}[0, \tau_o], \lambda - r \geq 0\right\}$$

be the space of all piecewise-polynomial BD parameterizations with birth rates $\lambda \in \mathcal{P}_+^{(\oplus)}[0, \tau_o]$, net diversification rates $r \in \mathcal{P}^{(\oplus)}[0, \tau_o]$, and fixed sampling fraction $\rho \in (0, 1]$.

Note that elements of the model class $\mathcal{I}_\rho$ are in 1–1 correspondence with models defined by the more common birth/death rate parameterization, via the identity $r = \lambda - \mu$. The function r is commonly referred to as the *net diversification rate* (Rabosky, 2010).

Louca and Pennell (2020) consider a new quantity called the *pulled speciation rate* λ_p, defined as

$$\lambda_p(t) = \lambda(t) \left[1 - E(t)\right], \tag{1}$$

where E is the probability that a lineage that exists at time t goes extinct by the present. The extinction probability E satisfies the differential equation

$$\frac{dE}{dt} = \mu - (\lambda + \mu)E + \lambda E^2,$$

with initial condition $E(0) = 1 - \rho$ (Morlon et al., 2011). This Bernoulli-type equation has the solution

$$E(t) = 1 - \frac{e^{R(t)}}{\rho^{-1} + \int_{u=0}^{t} \lambda(u) e^{R(u)}\, du}, \tag{2}$$

where

$$R(t) = \int_{u=0}^{t} r(u)\, du \tag{3}$$

is the cumulative net speciation rate. Thus,

$$\lambda_p(t) = \frac{\lambda(t) e^{R(t)}}{\rho^{-1} + \int_{u=0}^{t} \lambda(u) e^{R(u)}\ du}. \tag{4}$$

The cumulative integral Λ_p is given by

$$\Lambda_p(t) = \int_{u=0}^{t} \lambda_p(u)\ du.$$

Finally, we write the likelihood of a timetree. For the phylogenetic BD process, the likelihood depends only on the number and timing of branching events, and is independent of the tree topology (Nee et al., 1994; Morlon et al., 2011). Given that the number of tips is N and the process survives to the present over a period of length τ_o, (Louca and Pennell, 2020, supp. eqn. 34) show that the likelihood of a tree with bifurcation times $\tau_1, ..., \tau_{N-1}$ is

$$L^{(\lambda_p)}(\tau_1, ..., \tau_{N-1} | \tau_o) \propto \prod_{i=1}^{N-1} \lambda_p(\tau_i) e^{-\Lambda_p(\tau_i)}, \quad \tau_o > \tau_1 > \cdots > \tau_{N-1}. \tag{5}$$

From the preceding display, it is clear that a sample of merger times $(\tau_1, \ldots, \tau_{N-1})$ from an extant timetree can be equivalently viewed as the order statistics of $N-1$ i.i.d. draws from a distribution with density

$$f(x) \propto \lambda_p(x) e^{-\Lambda_p(x)} \tag{6}$$

supported on $[0, \tau_o]$, and that λ_p is the hazard rate function of this distribution. (See Section 5.2.1 for additional discussion.) Thus, λ_p completely characterizes the distribution of merger times in a timetree, and different BD models have different likelihoods if and only if their respective pulled rate functions are not equal almost everywhere on $[0, \tau_o]$. Moreover, subject to standard regularity conditions on hazard rate functions, λ_p (at least) can be consistently estimated from a timetree as the number of leaves N tends to infinity.

3 Results

We first show that piecewise polynomial phylogenetic birth-death models are identifiable from time-tree data, meaning that that different models in $\mathcal{I}_\rho$ have different likelihood functions. As noted in the preceding section, it suffices to prove that different models possess different pulled rate functions.

Theorem 1 (Identifiability of piecewise polynomial BD models)**.** Let $\mathcal{M}_1 = (\lambda^{(1)}, r^{(1)}, \rho)$ and $\mathcal{M}_2 = (\lambda^{(2)}, r^{(2)}, \rho)$ be two models in $\mathcal{I}_\rho$. Then $\lambda_p^{(1)} = \lambda_p^{(2)}$ if and only if $\mathcal{M}_1 = \mathcal{M}_2$.[1]

Theorem 1 is the main result of this paper, and should be contrasted with that of Louca and Pennell (2020). There, it is shown that, within the model class consisting of continuously differentiable birth and death rate functions, there are *infinitely* (in fact, uncountably) many pairs of such functions which all map to the same pulled rate function. Hence, statistical estimates of λ_p, even if they were somehow uncontaminated by error, do not automatically translate into accurate estimates of the underlying rate functions. The best one can hope for is to estimate an equivalence or "congruence" class of rate functions, whose members can be qualitatively quite different, as Louca and Pennell exhibit. Theorem 1 asserts that this unfortunate situation cannot occur if one places additional and somewhat mild assumptions on the model class: different models in $\mathcal{I}_\rho$ have different pulled rate functions, and conversely. This is a fairly strong result, for it asserts that *every* unique piecewise-polynomial rate parameterization corresponds to a different pulled rate function. Hence, the number and placement of break points, as well as the underlying rates themselves, can, at least in principle, all be estimated given a sufficiently large time tree.

However, this seemingly positive outlook does not paint the full picture. In Section 4, we demonstrate that phylogenetic BD models can fail to be "practically" identifiable, in the sense that it may be impossible to ascertain the correct model given a realistic amount of data. To formalize this, we consider a testing problem where the task is to choose between two competing models which are hypothesized to have generated the data. One hypothesis is $H_1 : (\lambda, r, \rho)$, where now the rate functions λ and r can be arbitrary, and are not restricted in form as in Theorem 1. The second hypothesis differs from the first by only a multiplicative

[1] Throughout the paper, we take $f = g$ to mean that the rate functions f and g are equal almost everywhere with respect to Lebesgue measure.

perturbation of the birth-rate function, $H_2 : ((1+\eta)\lambda, r, \rho)$, where $\eta > 0$ is a constant. Intuitively, if η is too small relative to the amount of data that has been collected, then it is difficult to distinguish between H_1 and H_2. Our next result quantifies this intuition.

Theorem 2. Consider the hypothesis testing problem where H_1 states that the birth-death model over $[0,\infty)$ is (λ, r, ρ) while H_2 states that the birth-death model is $((1+\eta)\lambda, r, \rho)$, where $\eta > 0$ is a constant. If an extant timetree on N leaves is observed, then for sufficiently small η, the Bayes error rate for distinguishing between H_1 and H_2 is at least $(1-\Upsilon)/2$, where

$$\Upsilon \leq \frac{\sqrt{N}\eta}{2}. \tag{7}$$

We prove a complementary result for when the net diversification rate r is scaled. However, for technical reasons, we are forced to make the strong assumption that λ and r are constant when proving this theorem.

Theorem 3. Consider the hypothesis testing problem where H_1 states that the birth-death model over $[0,\infty)$ is (λ, r, ρ) while H_2 states that the birth-death model is $(\lambda, (1+\eta)r, \rho)$, where both λ and r are constant over time, and $\eta > 0$ is a constant. If an extant time tree on N leaves is observed, the for sufficiently small η, the Bayes error rate for any classifier is at least $(1-\Upsilon)/2$, where

$$\Upsilon \leq \frac{\sqrt{N}\eta}{2}.$$

Since the Bayes error rate lower bounds classification accuracy, the theorems imply that the probability of correctly determining whether the timetree was generated by H_1 or H_2 is at most $(1+\Upsilon)/2$ using *any* method. The key feature of the bounds is that they degrade in only the *square root* of the size of the time-tree N. Hence, unless $N \gg 1/\eta^2$, no procedure can distinguish between H_1 and H_2 with high probability.

4 Discussion

In this paper, we have derived new theoretical results concerning estimation of phylogenetic birth-death models using time-tree data. Our main result, Theorem 1, establishes identifiability of phylogenetic BD models parameterized by piecewise polynomial rate functions. One implication that may be surprising is that there exist identifiable classes of birth-death models with smoothly varying birth and death rates that are not purely constant. In particular, spline rate functions are identifiable. Splines are piecewise polynomials with additional smoothness constraints, and are widely used to model natural systems.

Theorem 1 improves on an earlier result of Legried and Terhorst (2021a), who established identifiability of piecewise constant BD models; we obtain their result as a special case. However, the sampling models underlying the two results are slightly different. Legried and Terhorst (2021a) assume access to a finite collection of *moments* of merger times $(\tau_1, \ldots, \tau_N)$, the number of which increases with the number of pieces K of the rate functions. Conceivably, these could be estimated using a large collection of independent time trees all of size N. Here, we assume access to the complete distribution of $(\tau_1, \ldots, \tau_N)$, which, as noted above,

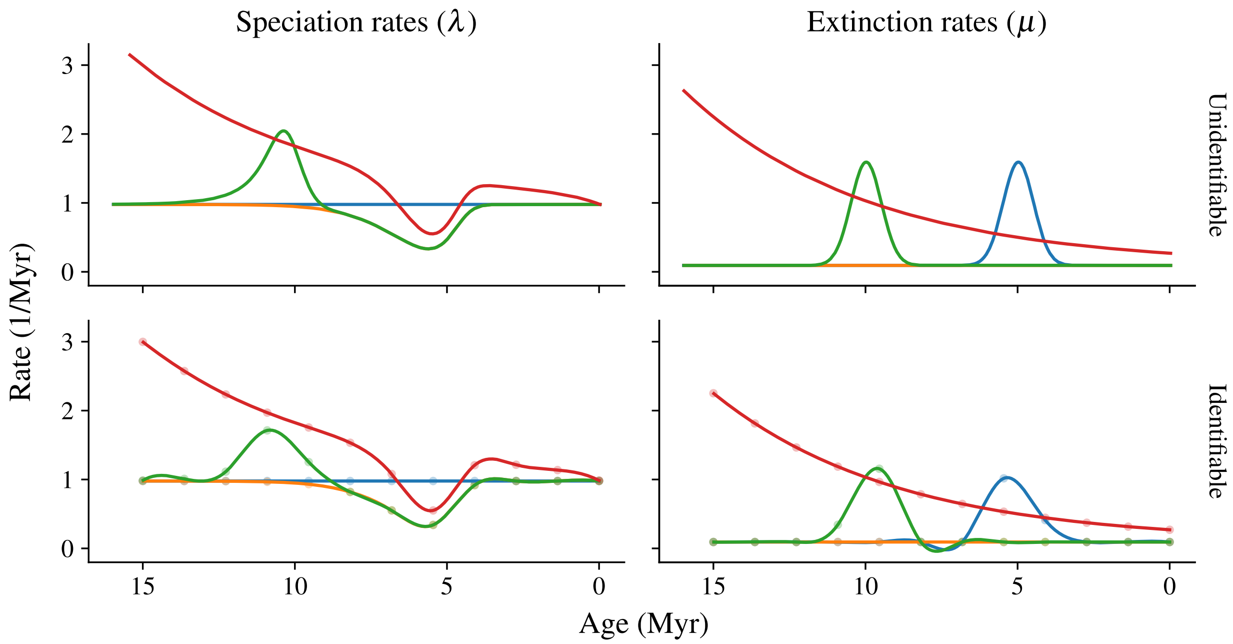

Figure 1: Identifiable versus unidentifiable models. The top row contains four unidentifiable models shown in Figure 1 of Louca and Pennell (2020). Each color-coded (λ, μ) pair in the top maps to the same pulled rate function; hence they are indistinguishable. In the bottom row, we approximated these functions using interpolating cubic splines, with knots at $t = 15, 14, \dots, 0$. According to Theorem 1, these models have different pulled rate functions, and may be distinguished in the infinite-data limit.

is given essentially by the density $f(x)$ in equation (6). This scheme is more in keeping with the model studied by Louca and Pennell (2020), where there is a single time-tree tending in size to infinity.

To better understand the relationship between Theorem 1 and the non-identifiability result of Louca and Pennell, consider Figure 1. The top row is adapted from Figure 1 of their paper, and depicts four pairs of (λ, μ) rate functions which all have the pulled sampling rate λ_p. In the bottom row, we approximated each of these functions using a cubic spline, with sixteen knots placed at $t = 15, \dots, 0$. The models shown in the bottom row are identifiable; the ones in the top are not. Note that there are some visual differences between the two panels; for example the speciation rates of the rate and green models intersect in the top row, whereas the spline smoothness constraints prevent them from doing so in the bottom. (For illustrative purposes, we used a very simple interpolating spline with equispaced knots; a closer approximation could be obtained using a more complex fitting procedure.) Practitioners must decide if the spline model class (or a related class of identifiable models) can faithfully model population dynamics in their application.

The results we present here restore to some extent the mathematical footing beneath the many published studies that have utilized the linear birth-death process to describe evolution. However, identifiability is a minimal regularity condition one can impose on a statistical model, and estimation remains challenging. To see this, consider now Figure 2,

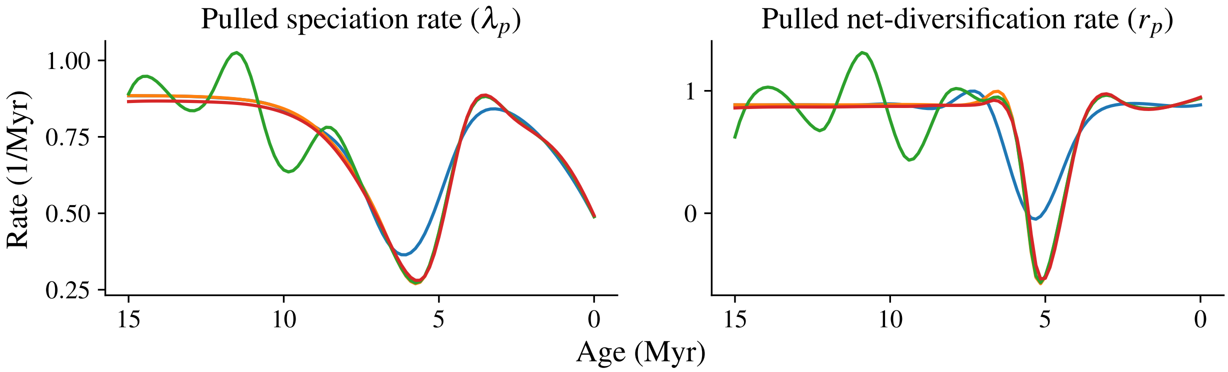

Figure 2: The pulled rates of speciation and net diversification for each of the spline models shown in the bottom row of Figure 1. The sampling fraction was $\rho = 0.5$ in each model.

which plots the pulled speciation and net diversification rates for each of the spline-based models from Figure 1. By Theorem 1, these functions are necessarily different—but in terms of estimation, the important question is *how* different they are. It is evident from Figure 2 that distinguishing the green model from the other three is probably feasible. In contrast, the red and orange models appear almost identical in terms of λ_p, and identifying which of them generated a particular data set is likely to be difficult given a realistic amount of data.

This leads to our second set of results, Theorems 2 and 3, which study the hardness of distinguishing between competing BD hypothesis using only a finite amount of data. We prove that even answering the relatively simple question of whether the data are generated by a particular model, or one in which its rate function(s) are a scalar multiple of it, scales poorly in the time-tree size N. (This theoretical limitation is also found in coalescent models, see Kim et al., 2015.) As Figure 2 already suggests, the question of when practical estimation is possible is likely to be quite subtle even in identifiable model classes. Much more deserves to be said, and this is an important area for future research.

Finally, an obvious caveat to the results we have presented is that it may not be possible to estimate λ_p in the first place. As noted in Section 1, inferring λ_p is tantamount to estimating the hazard rate function of the distribution shown in (5). If it were possible to directly sample from this distribution, estimation of λ_p would be routine, however this is not possible in practice. Rather, one must first estimate the time-tree itself, and then treat the estimated merger times as though they were samples from (5). Tree inference is itself a difficult problem, and often these estimates contain considerable error, so it is not obvious that such a procedure leads to accurate downstream estimates of the underlying BD model, even in the infinite-data limit.

The assumption that λ_p is known colors our results as follows. First, it implies that the lower bounds derived in Theorems 2 and 3 are likely not sharp, even for identifiable model classes, and that inferring phylogenetic BD models can be even harder than is indicated by theorems. Conversely, it qualifies Theorem 1 somewhat, since the theorem does not resolve the question of whether phylogenetic BD models are identifiable *on the basis of the observed data*. In this sense, unidentifiability results like those derived in Stadler (2009); Louca and Pennell (2020) are stronger, since they imply that the model cannot be estimated *even if* we somehow had direct access to the underlying time-tree. We note in closing that there is a growing literature on phylogenetic identifiability (e.g., Rhodes and Sullivant, 2012; Mossel and Roch, 2013), which derives conditions under which it is possible to consistently estimate an underlying phylogeny given character data evolving along its branches. If it can be shown that an asymptotically expanding time tree of the variety considered here is estimable, it would, together with our results, automatically imply identifiability of polynomial rate functions from character data.

5 Proofs

This section contains the proofs of Theorems 1, 2 and 3.

5.1 Polynomial identifiability

In this subsection, we prove Theorem 1. We start with the case of polynomial models.

Lemma 1. Consider two models $(\lambda^{(1)}, r^{(1)}, \rho)$ and $(\lambda^{(2)}, r^{(2)}, \rho)$ where $\lambda^{(i)}, r^{(i)} \in \mathcal{P}$ are polynomials with $\inf_{t\in[0,\tau_o]} \lambda^{(i)}(t) > 0$, and define

$$p = \lambda^{(2)} \left[\lambda^{(1)}\right]' - \lambda^{(1)} \left[\lambda^{(2)}\right]' + \lambda^{(1)}\lambda^{(2)}(r^{(1)} - r^{(2)}) \in \mathcal{P}. \tag{8}$$

If u is a limit point of the set $\{t : \lambda_p^{(1)}(t) = \lambda_p^{(2)}\}$, then $p(u) = 0$. In particular, if there are infinitely many such points, then $p \equiv 0$.

Proof. For any such u, we have by continuity

$$0 = \lambda_p^{(1)}(u) - \lambda_p^{(2)}(u) = \frac{d}{dt} \left[\log \lambda_p^{(1)}(t) - \log \lambda_p^{(2)}(t)\right]\Big|_{t=u}.$$

By (4),

$$\begin{aligned} \frac{d}{dt} \log \lambda_p(t) &= \frac{\lambda'(t)}{\lambda(t)} + r(t) - \frac{d}{dt} \log \left(\rho^{-1} + \int_{u=0}^{t} \lambda(u) e^{R(u)}\, du\right) \\ &= \frac{\lambda'(t)}{\lambda(t)} + r(t) - \lambda_p(t). \end{aligned}$$

Thus

$$\frac{\left[\lambda^{(1)}\right]'(u)}{\lambda^{(1)}(u)} + r^{(1)}(u) - \frac{\left[\lambda^{(2)}\right]'(u)}{\lambda^{(2)}(u)} - r^{(2)}(u) = 0.$$

Clearing denominators gives $p(u) = 0$. If there are infinitely many limit points, then p has infinitely many zeros and is hence identically zero. □

Our main new technical result is the following sufficient condition for equality of two polynomial models, which works by an algebraic argument.

Proposition 1. Let $(\lambda^{(1)}, r^{(1)}, \rho)$, $(\lambda^{(2)}, r^{(2)}, \rho)$, and p be as defined in Lemma 1. If $p \equiv 0$, then $(\lambda^{(1)}, r^{(1)}) = (\lambda^{(2)}, r^{(2)})$.

Proof. By assumption, for $i \in \{1, 2\}$ there exist non-negative degrees $\deg \lambda^{(i)} = m_i$ and $\deg r^{(i)} = n_i$ and real coefficients $\lambda_j^{(i)}$ and $r_j^{(i)}$, with $\lambda_{m_i}^{(i)} \neq 0$ such that

$$\lambda^{(i)}(t) = \sum_{j=0}^{m_i} \lambda_j^{(i)} t^j$$
$$r^{(i)}(t) = \sum_{j=0}^{n_i} r_j^{(i)} t^j.$$

We first establish $r^{(1)} = r^{(2)}$. Define $n = \max(n_1, n_2)$. If $n = -\infty$ then $r^{(1)} = r^{(2)} \equiv 0$ and there is nothing left to show. Otherwise, let

$$q(t) = r^{(1)} - r^{(2)} = \sum_{j=0}^{n} \left(r_j^{(1)} - r_j^{(2)} \right) t^j =: \sum_{j=0}^{n} q_j t^j, \tag{9}$$

where we set $r_j^{(i)} = 0$ for all $j > n_i$. In order to prove $r^{(1)} = r^{(2)}$, it suffices to show that the coefficients $q_j = 0$ for all $j \in \{0, \ldots, n\}$.

Next, define the polynomials

$$\begin{aligned} p_1 &= \lambda^{(2)} \left[\lambda^{(1)} \right]' \\ p_2 &= -\lambda^{(1)} \left[\lambda^{(2)} \right]' \\ p_3 &= q \lambda^{(1)} \lambda^{(2)}, \end{aligned} \tag{10}$$

so that

$$p_1 + p_2 + p_3 = p \equiv 0. \tag{11}$$

We observe that $\deg(p_1 + p_2) \leq m_1 + m_2 - 1$, and that p_3 is a polynomial of degree at most $m_1 + m_2 + n$:

$$p_3(t) = \sum_{j=0}^{m_1+m_2+n} \gamma_j t^j,$$

for some $\gamma_0, \ldots, \gamma_{m_1+m_2+n} \in \mathbb{R}$. Hence, by (11),

$$\gamma_{m_1+m_2+n} = \gamma_{m_1+m_2+n-1} = \cdots = \gamma_{m_1+m_2} = 0. \tag{12}$$

We proceed to establish that $q_{n-j} = 0$ for $0 \leq j \leq n$ by induction on j. For the base case $j = 0$, we have by (10) that the leading coefficient of p_3 equals the product of the leading coefficients of $\lambda^{(1)}$, $\lambda^{(2)}$, and q:

$$\gamma_{m_1+m_2+n} = q_n \lambda_{m_1}^{(1)} \lambda_{m_2}^{(2)}. \tag{13}$$

Since $\lambda_{m_i}^{(i)} \neq 0$, it must be that $q_n = 0$.

Now suppose the claim is true for all $0 \leq j < k$. By the convolution rule for polynomial multiplication, we have

$$\gamma_{m_1+m_2+n-k} = \sum_{(u,v,w) \in \mathcal{S}} q_u \lambda_v^{(1)} \lambda_w^{(2)},$$

where the summation is over the index set

$$\mathcal{S} = \left\{(u,v,w) \in \mathbb{Z}^3 : u+v+w = m_1+m_2+n-k, u \in [0,n], v \in [0,m_1], w \in [0,m_2]\right\}.$$

We partition $\mathcal{S}$ into disjoint subsets

$$\begin{aligned}\mathcal{S}' &= \{(u,v,w) \in \mathcal{S} : u > n-k\}\\ \mathcal{S}'' &= \{(u,v,w) \in \mathcal{S} : u = n-k\} = \{(n-k, m_1, m_2)\}.\end{aligned}$$

By the inductive hypothesis,

$$\sum_{(u,v,w)\in\mathcal{S}'} q_u \lambda_v^{(1)} \lambda_w^{(2)} = 0,$$

so that

$$\gamma_{m_1+m_2+n-k} = \sum_{(u,v,w)\in\mathcal{S}''} q_u \lambda_v^{(1)} \lambda_w^{(2)} = q_{n-k} \lambda_{m_1}^{(1)} \lambda_{m_2}^{(2)}.$$

This together with (12) implies that $q_{n-k} = 0$. Hence $q = 0$, so $r^{(1)} = r^{(2)}$.

We conclude by showing that $\lambda^{(1)} = \lambda^{(2)}$. Since $q = 0$, we have $p_1 = -p_2$ by (11). Therefore,

$$\frac{d}{dt}\log\lambda^{(1)} = \frac{\left[\lambda^{(1)}\right]'}{\lambda^{(1)}} = \frac{\left[\lambda^{(2)}\right]'}{\lambda^{(2)}} = \frac{d}{dt}\log\lambda^{(2)} \implies \lambda^{(1)} = C\lambda^{(2)}$$

for some constant $C > 0$. Define

$$e(t) = \int_0^t \lambda^{(2)}(s) e^{R^{(2)}(s)}\, ds,$$

and let $z \in [0, \tau_o]$ be a zero of $\lambda_p^{(1)} - \lambda_p^{(2)}$. Noting that $e'(z) > 0$, we have that

$$\frac{Ce'(z)}{\rho^{-1} + Ce(z)} = \lambda_p^{(1)}(z) = \lambda_p^{(2)}(z) = \frac{e'(z)}{\rho^{-1} + e(z)}$$

implies $C = 1$.

□

The preceding results immediately yield the following corollary on identifiability for polynomial models.

Proposition 2. Let $(\lambda^{(1)}, r^{(1)}, \rho)$ and $(\lambda^{(2)}, r^{(2)}, \rho)$ be as defined in Lemma 1. If $(\lambda^{(1)}, r^{(1)}) \neq (\lambda^{(2)}, r^{(2)})$, then $\lambda_p^{(1)}(t) \neq \lambda_p^{(2)}(t)$ almost everywhere on $[0, \tau_o]$.

Proof. By Lemma 1 and Proposition 1, the set $\{t : \lambda_p^{(1)}(t) = \lambda_p^{(2)}(t)\}$ contains at most a finite number of limit points, so it is null. Therefore, its relative complement in $[0, \tau_o]$ has full measure. □

Using Proposition 2, to we can prove Theorem 1 for piecewise-polynomial birth and death rates on an arbitrary number of pieces. The proof is an extension of Proposition A.5 in Legried and Terhorst (2021a).

Proof of Theorem 1. The "if" direction is immediate. To establish the "only if" direction, we show its contrapositive: $\left(\lambda^{(1)}, r^{(1)}\right) \neq \left(\lambda^{(2)}, r^{(2)}\right)$ implies $\lambda_p^{(1)} \neq \lambda_p^{(2)}$ on a set of positive measure. We may assume that the two models are defined on the same set of breakpoints, $0 = t_0 < \cdots < t_K = \tau_o$, since this can always be achieved by increasing K. Then there exists a non-empty interval $[u, v) \subset [0, \tau_o]$ such that

1. $\left(\lambda^{(1)}(s), r^{(1)}(s)\right) = \left(\lambda^{(2)}(s), r^{(2)}(s)\right)$ for all $0 < s < u$;
2. $\left(\lambda^{(1)}(s), r^{(1)}(s)\right) \neq \left(\lambda^{(2)}(s), r^{(2)}(s)\right)$ for all $s \in [u, v)$; and
3. $[u, v) \subset [t_k, t_{k+1})$ for some k.

(We could have $u = 0$, rendering the first condition vacuous.) Let the birth and net diversification rates for the two models over $[u, v)$ be denoted $\lambda_k^{(i)}$ and $r_k^{(i)}$, respectively.

Recall the function $E(t)$ defined in equation (2). Because E is continuous and $E^{(1)}(0) = E^{(2)}(0) = 1 - \rho$, condition (1) above implies that $E^{(1)}(u) = E^{(2)}(u) = 1 - \epsilon \in (0, 1)$ for some ϵ. Define the shifted models $(\pi^{(i)}, q^{(i)}, \epsilon)$ for $i \in \{1, 2\}$, where

$$\begin{aligned} \pi^{(i)}(t) &= \lambda_k^{(i)}(t + u) \\ q^{(i)}(t) &= r_k^{(i)}(t + u). \end{aligned}$$

Denote the corresponding pulled rate functions $\pi_p^{(i)}$, so that $\pi_p^{(i)}(t) = \lambda_p^{(i)}(t + u)$ for $t \in [0, v - u)$. By Proposition 2, $\pi_p^{(1)} \neq \pi_p^{(2)}$ almost everywhere on $[0, v - u)$. The same is therefore true of $\lambda_p^{(1)}$ and $\lambda_p^{(2)}$ on $[u, v)$. □

5.2 Lower bounds

In this section, we prove Theorems 2 and 3. Subsections 5.2.1 and 5.2.2 contain some necessary background and definitions, and can be omitted if the reader is already familiar with them. The results of this section use techniques developed by Kim et al. (2015) and, later, Legried and Terhorst (2021b).

5.2.1 The pulled speciation rate as a hazard rate

Let the origin time τ_o have an (improper) uniform prior distribution on $(0, \infty)$. Then the likelihood of $\tau_o > \tau_1 > ... > \tau_{N-1}$ is given by

$$L^{(\lambda_p)}(\tau_o, \tau_1, ..., \tau_{N-1}) = L^{(\lambda_p)}(\tau_1, ..., \tau_{N-1} | \tau_o).$$

The likelihood formula holds, in particular, when we take $\tau_o \to \infty$. Then for any sequence $\tau_1 > \tau_2 > ... > \tau_{N-1} > \tau_N = 0$, the likelihood is

$$L^{(\lambda_p)}(\tau_1, ..., \tau_{N-1}) = (N-1)! \prod_{i=1}^{N-1} \lambda_p(\tau_i) e^{-\Lambda_p(\tau_i)} = (N-1)! \prod_{i=1}^{N-1} \lambda_p(\tau_i) e^{-i[\Lambda_p(\tau_i) - \Lambda_p(\tau_{i+1})]}.$$

This shows that λ_p is a hazard rate function. To ensure that the integral of the likelihood function converges, we assume that

$$\lim_{t\to\infty} \Lambda_p(t) = \int_{u=0}^{\infty} \lambda_p(u)\ du = \infty,$$

so this requirement is satisfied. Since λ_p is a hazard rate, it corresponds to a probability density function $f_p(t) = \lambda_p(t)e^{-\Lambda_p(t)}$.

5.2.2 Bounding the Hellinger distance

Here, we recall the total variation distance and Hellinger distance between measures, borrowing from the explanation in Legried and Terhorst (2021b). Consider a measurable space $(\Omega, \mathcal{F})$ with two possible probability measures P and Q on it. The two probability spaces $(\Omega, \mathcal{F}, P)$ and $(\Omega, \mathcal{F}, Q)$ then have corresponding probability density functions f_P and f_Q. The *total variation distance* between P and Q is defined as

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

which equals $\frac{1}{2}\int |f_P - f_Q|$. As this integral is itself a metric, one might abuse notation and write $d_{TV}(f_P, f_Q)$ to mean the same thing.

Suppose a datum D has been generated under P or Q and we want 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 decide the measure correctly. Let χ be the true data generating distribution (from the set $\{P, Q\}$), and let $\hat{\chi}(D) \in \{P, Q\}$ be a putative classifier. The probability that $\hat{\chi}$ correctly classifies D, i.e. $\hat{\chi} = \chi$, is

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

where Υ is a positive number, because the error of any binary classifier can be made less than $1/2$. Another way to write Υ is

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

It can be shown that the best possible classification rule is the likelihood ratio: set $\hat{\chi} = P$ if and only if $P(D) > Q(D)$. In that case, we have

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

The likelihood ratio is said to achieve the *minimal error rate* or *Bayes error rate*. If multiple data samples $D_1, ..., D_L$ are given, then similarly Υ equals $d_{TV}(P^{\otimes L}, Q^{\otimes L})$, where $P^{\otimes L}$ denotes the product measure over the L samples.

In our problem setting, we find it easier to work with the Hellinger distance, which is given defined as

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

The Hellinger distance is related to total variation via

$$d_{TV}^2(P,Q) \leq 2d_H^2(P,Q) = 2\left(1 - \int \sqrt{f_P f_Q}\right), \tag{14}$$

and it also possess the following subadditivity property: if $P = \otimes_{i=1}^N P_i$ and $Q = \otimes_{i=1}^N Q_i$ are product measures, then

$$d_H^2(P,Q) \leq \sum_{i=1}^N d_H^2(P_i, Q_i).$$

By bounding the right-hand side above by a quantity converging to 0 quickly, it follows that the total variation distance also converges to 0 quickly.

We will prove Theorems 2 and 3 by computing the Hellinger distance of competing joint distributions of $(\tau_1, ..., \tau_{N-1})$. For this, we need to establish precise two-sided bounds of $\lambda_p^{(2)}$ in terms of $\lambda_p^{(1)}$. We start in the setting of Theorem 2, where λ is the birth rate corresponding to $\lambda_p^{(1)}$ and $(1+\eta)\lambda$ is the birth rate corresponding to $\lambda_p^{(2)}$. The two-sided bound of $\lambda_p^{(2)}(t)$ is expressed in the following Lemma.

Lemma 2. Let $t > 0$ and $\eta > 0$ be arbitrary. Then

$$\lambda_p^{(1)}(t)\left[1 + a(t)\eta - \frac{1}{2}b(t)\eta^2\right] \leq \lambda_p^{(2)}(t) \leq \lambda_p^{(1)}(t)\left[1 + a(t)\eta + \frac{1}{2}b(t)\eta^2\right],$$

where

$$\begin{aligned} a(t) &= \frac{\rho^{-1}}{\rho^{-1} + \int_{u=0}^t \lambda(u)e^{R(u)}\ du} \\ b(t) &= -\frac{2\rho^{-1}\int_{u=0}^t \lambda(u)e^{R(u)}\ du}{\left[\rho^{-1} + \int_{u=0}^t \lambda(u)e^{R(u)}\ du\right]^2}. \end{aligned}$$

Proof. The quotient is

$$\begin{aligned} \frac{\lambda_p^{(2)}(t)}{\lambda_p^{(1)}(t)} &= \frac{(1+\eta)\lambda(t)e^{R(t)}}{\rho^{-1} + (1+\eta)\int_{u=0}^t \lambda(u)e^{R(u)}\ du} \frac{\rho^{-1} + \int_{u=0}^t \lambda(u)e^{R(u)}\ du}{\lambda(t)e^{R(t)}} \\ &= (1+\eta)\frac{\rho^{-1} + \int_{u=0}^t \lambda(u)e^{R(u)}\ du}{\rho^{-1} + (1+\eta)\int_{u=0}^t \lambda(u)e^{R(u)}\ du} \\ &= \frac{\rho^{-1} + \int_{u=0}^t \lambda(u)e^{R(u)}\ du}{\rho^{-1}(1+\eta)^{-1} + \int_{u=0}^t \lambda(u)e^{R(u)}\ du}. \end{aligned}$$

The coefficients $a(t)$ and $b(t)$ are obtained by differentiating the above expression with respect to η and evaluating at 0. The first derivative is

$$\frac{\rho^{-1}(\rho^{-1} + \int_{u=0}^t \lambda(u)e^{R(u)}\ du)}{(1+\eta)^2\left[\rho^{-1}(1+\eta)^{-1} + \int_{u=0}^t \lambda(u)e^{R(u)}\ du\right]^2}$$

which evaluates at 0 to

$$a(t) = 1 - \frac{\int_{u=0}^{t} \lambda(u) e^{R(u)}\ du}{\rho^{-1} + \int_{u=0}^{t} \lambda(u) e^{R(u)}\ du} = \frac{\rho^{-1}}{\rho^{-1} + \int_{u=0}^{t} \lambda(u) e^{R(u)}\ du}.$$

Note that the right-hand side is a positive number between 0 and 1. The second derivative is

$$\begin{aligned} b(\eta, t) &= -\frac{2\rho^{-1} \int_{u=0}^{t} \lambda(u) e^{R(u)}\ du \left(\rho^{-1} + \int_{u=0}^{t} \lambda(u) e^{R(u)}\ du\right)}{(1+\eta)^3 \left[\rho^{-1}(1+\eta)^{-1} + \int_{u=0}^{t} \lambda(u) e^{R(u)}\ du\right]^3} \\ &= -\frac{2\rho^{-1} \int_{u=0}^{t} \lambda(u) e^{R(u)}\ du \left(\rho^{-1} + \int_{u=0}^{t} \lambda(u) e^{R(u)}\ du\right)}{\left[\rho^{-1} + (1+\eta) \int_{u=0}^{t} \lambda(u) e^{R(u)}\ du\right]^3}, \end{aligned}$$

which is always negative. We now apply Taylor's theorem, observing that

$$\frac{\lambda_p^{(2)}(t)}{\lambda_p^{(1)}(t)} = 1 + a(t)\eta + R_2(\eta; 0),$$

where the remainder term $R_2(\eta; 0)$ is

$$R_2(\eta; 0) = \frac{b(c,t)}{2} \eta^2$$

for some $c \in (0, \eta)$. Since $\partial |b(c,t)| / \partial c < 0$, we have $|R_2(\eta; 0)| < b(t)\eta^2/2$, where

$$b(t) := |b(0,t)| = \frac{2\rho^{-1} \int_{u=0}^{t} \lambda(u) e^{R(u)}\ du}{\left[\rho^{-1} + \int_{u=0}^{t} \lambda(u) e^{R(u)}\ du\right]^2}.$$

Then

$$\left| \frac{\lambda_p^{(2)}(t)}{\lambda_p^{(1)}(t)} - 1 - a(t)\eta \right| \le \frac{1}{2} b(t) \eta^2,$$

which implies the result. □

To use the Lemma to prove Theorem 2, we show that $a(t)$ and $b(t)$ are each bounded by constants over $t \ge 0$.

Proof of Theorem 2. From (14), the squared error rate satisfies

$$\Upsilon^2 \le 2 d_H^2(f_1, f_2).$$

Starting with the case of $N = 2$, we have $d_H^2(f_1, f_2) = 1 - I(f_1, f_2)$ where $I(f_1, f_2) = \int_{t=0}^{\infty} \sqrt{f_1(t) f_2(t)}\ dt$. Defining the probability density function

$$g(t) = \exp\left[-\int_{y=0}^{t} \frac{h_1(y) + h_2(y)}{2}\ dy \right] \frac{h_1(t) + h_2(t)}{2},$$

we have

$$I(f_1, f_2) = \int_{t=0}^{\infty} g(t) \frac{2\sqrt{h_1(t)h_2(t)}}{h_1(t) + h_2(t)}\, dt.$$

Now we use an upper bound on $h_1 + h_2$ and a lower bound on $h_1 h_2$ from Lemma 2 to obtain a lower bound for $2\sqrt{h_1 h_2}/(h_1 + h_2)$. We have

$$\frac{2\sqrt{h_1(t)h_2(t)}}{h_1(t) + h_2(t)} \geq \frac{2\lambda_p^{(1)}(t)\sqrt{1 + a(t)\eta - \frac{1}{2}b(t)\eta^2}}{\lambda_p^{(1)}(t)\left[2 + a(t)\eta + \frac{1}{2}b(t)\eta^2\right]}.$$

The power series expansion of the right-hand side is

$$1 - \frac{1}{8}\left[a(t)\right]^2 \eta^2 + R_3(\eta; 0)$$

with the remainder term satisfying

$$|R_3(\eta; 0)| \leq \left|\frac{1}{8}\left[a(t)\right]^2 (a(t) - b(t))\eta^3\right|$$

for some $t \in [0, \eta)$. Both functions $a(t)$ and $b(t)$ are bounded in t. So for η close enough to 0, independently of t, we have

$$1 - \frac{1}{8}\left[a(t)\right]^2 \eta^2 + R_3(\eta; 0) \geq 1 - \frac{1}{8}\eta^2 + D\eta^3$$

where D is a non-zero constant independent of t. So since $g(t)$ is a density,

$$I(f_1, f_2) \geq \left(1 - \frac{1}{8}\eta^2 + D\eta^3\right) \int_{t=0}^{\infty} g(t)\, dt = 1 - \frac{1}{8}\eta^2 + D\eta^3$$

for η sufficiently close to 0. Putting it together, for such η the Hellinger distance is bounded as

$$d_H^2(f_1, f_2) \leq \frac{1}{8}\eta^2,$$

and the Theorem follows in the $N = 2$ case.

Now we derive the bound for an arbitrary number N of extant individuals. There are $N - 1$ independent coalescent times with identically distributed arrival times. As f_1 and f_2 are product measures of $N - 1$ independent random variables, the subadditivity property implies

$$d_H^2(f_1, f_2) \leq (N-1)\frac{\eta^2}{8} \leq N\frac{\eta^2}{8}.$$

Then Theorem 2 follows from (14).

□

For Theorem 3, we compare hypothetical models with the same birth rate and sampling probability, but $R^{(1)}(t) = R(t)$ and $R^{(2)}(t) = (1 + \eta)R^{(1)}(t)$ for all t. Although we prove Theorem 3 in the case where $R^{(1)}$ is constant over time, the analogue to Lemma 2 can be proved for general $\lambda^{(1)}$ and $R^{(1)}$.

Lemma 3. Let t be fixed and η be a positive number close to 0. Then

$$\frac{\lambda_p^{(2)}(t)}{\lambda_p^{(1)}(t)} = 1 + a(t)\eta + \frac{1}{2}b(t)\eta^2 + O(\eta^3)$$

as $\eta \to 0$, where

$$\begin{aligned}
a(t) &= R(t) - \frac{\int_{u=0}^t \lambda(u)R(u)e^{R(u)}\ du}{\rho^{-1} + \int_{u=0}^t \lambda(u)e^{R(u)}\ du} \\
b(t) &= \Bigg\{ - \frac{R(t)\int_{u=0}^t \lambda(u)R(u)e^{R(u)}\ du}{\rho^{-1} + \int_{u=0}^t \lambda(u)e^{R(u)}\ du} - \frac{\int_{u=0}^t \lambda(u)[R(u)]^2 e^{R(u)}\ du}{\rho^{-1} + \int_{u=0}^t \lambda(u)e^{R(u)}\ du} \\
&\qquad + \frac{2\left[\int_{u=0}^t \lambda(u)R(u)e^{R(u)}\ du\right]^2}{\left[\rho^{-1} + \int_{u=0}^t \lambda(u)e^{R(u)}\ du\right]^2} \Bigg\} + a(t)R(t)
\end{aligned}$$

Proof. The quotient is

$$\frac{\lambda_p^{(2)}(t)}{\lambda_p^{(1)}(t)} = \frac{\lambda(t)e^{(1+\eta)R(t)}}{\rho^{-1} + \int_{u=0}^t \lambda(u)e^{(1+\eta)R(u)}\ du} \frac{\rho^{-1} + \int_{u=0}^t \lambda(u)e^{R(u)}\ du}{\lambda(t)e^{R(t)}}$$

As before, the coefficients $a(t)$ and $b(t)$ are obtained by differentiating the above expression with respect to η and evaluating at 0. The first derivative is

$$\left\{ \frac{R(t)}{\rho^{-1} + \int_{u=0}^t \lambda(u)e^{(1+\eta)R(u)}\ du} - \frac{\int_{u=0}^t \lambda(u)R(u)e^{(1+\eta)R(u)}\ du}{\left[\rho^{-1} + \int_{u=0}^t \lambda(u)e^{(1+\eta)R(u)}\ du\right]^2} \right\} \times e^{\eta R(t)} \left(\rho^{-1} + \int_{u=0}^t \lambda(u)e^{R(u)}\ du \right),$$

which evaluates at $\eta = 0$ to

$$a(t) = R(t) - \frac{\int_{u=0}^t \lambda(u)R(u)e^{R(u)}\ du}{\rho^{-1} + \int_{u=0}^t \lambda(u)e^{R(u)}\ du}.$$

The second derivative is

$$\begin{aligned}
&\Bigg\{ - \frac{R(t)\int_{u=0}^t \lambda(u)R(u)e^{(1+\eta)R(u)}\ du}{\left[\rho^{-1} + \int_{u=0}^t \lambda(u)e^{(1+\eta)R(u)}\ du\right]^2} - \frac{\int_{u=0}^t \lambda(u)[R(u)]^2 e^{(1+\eta)R(u)}\ du}{\left[\rho^{-1} + \int_{u=0}^t \lambda(u)e^{(1+\eta)R(u)}\ du\right]^2} \\
&+ \frac{2\left[\int_{u=0}^t \lambda(u)R(u)e^{(1+\eta)R(u)}\ du\right]^2}{\left[\rho^{-1} + \int_{u=0}^t \lambda(u)e^{(1+\eta)R(u)}\ du\right]^3} \Bigg\} e^{\eta R(t)} \left(\rho^{-1} + \int_{u=0}^t \lambda(u)e^{R(u)}\ du \right) \\
&+ \left\{ \frac{R(t)}{\rho^{-1} + \int_{u=0}^t \lambda(u)e^{(1+\eta)R(u)}\ du} - \frac{\int_{u=0}^t \lambda(u)R(u)e^{(1+\eta)R(u)}\ du}{\left[\rho^{-1} + \int_{u=0}^t \lambda(u)e^{(1+\eta)R(u)}\ du\right]^2} \right\} \\
&\times R(t)e^{\eta R(t)} \left(\rho^{-1} + \int_{u=0}^t \lambda(u)e^{R(u)}\ du \right),
\end{aligned}$$

which evaluates at $\eta = 0$ to

$$\begin{aligned} &\Bigg\{ - \frac{R(t) \int_{u=0}^{t} \lambda(u) R(u) e^{R(u)}\ du}{\rho^{-1} + \int_{u=0}^{t} \lambda(u) e^{R(u)}\ du} - \frac{\int_{u=0}^{t} \lambda(u) [R(u)]^2 e^{R(u)}\ du}{\rho^{-1} + \int_{u=0}^{t} \lambda(u) e^{R(u)}\ du} \\ &+ \frac{2 \left[\int_{u=0}^{t} \lambda(u) R(u) e^{R(u)}\ du \right]^2}{\left[\rho^{-1} + \int_{u=0}^{t} \lambda(u) e^{R(u)}\ du \right]^2} \Bigg\} \\ &+ a(t) R(t). \end{aligned}$$

□

The proof of Theorem 3 is similar. However, because the $R(t)$ function appears inside exponents, bounding $a(t)$ and $b(t)$ is more complicated. This is why we assume that $R(t) = Rt$ is linear in the theorem.

Proof of Theorem 3. Going forward, we assume that the rates are constant, implying there exists a real number R such that $R(t) = Rt$ for all t. Then we use Taylor's theorem:

$$\frac{\lambda_p^{(2)}(t)}{\lambda_p^{(1)}(t)} = 1 + a(t)\eta + R_2(\eta; 0),$$

where now

$$a(t) = \frac{\lambda\rho(e^{Rt} - 1) + R^2 t - \lambda\rho R t}{\lambda\rho(e^{Rt} - 1) + R}$$

by Lemma 3. If $R > 0$, then $|a(t)| \leq 1$ for all $t \geq 0$. If $R < 0$, then the limit of $a(t)$ as $t \to +\infty$ is $-\infty$, which is not usable. So we proceed with $R > 0$. The remainder term $R_2(\eta; 0)$ has the same structure as before, but

$$\begin{aligned} b(c,t) = &\frac{e^{cRt} R(R + \lambda\rho(e^{Rt} - 1))}{[(1+c)R + \lambda\rho(e^{(1+c)Rt} - 1)]^3} \Bigg[(1+c)^3 R^3 t^2 \\ &+ \lambda^2 \rho^2 t \left(2 + (1+c)Rt + e^{(1+c)Rt}((1+c)Rt - 2)\right) \\ &- \lambda\rho \left\{ e^{(1+c)Rt}(2 - 4(1+c)Rt + (1+c)^2 R^2 t^2) + 2(-1 + (1+c)Rt + (1+c)^2 R^2 t^2) \right\} \Bigg]. \end{aligned}$$

(As $R > 0$, there is no possibility of division by 0.) Now we derive c-independent bounds for the bracketed expression. The negative terms are dropped and the remaining polynomial terms are replaced with exponentials:

$$\begin{aligned} &(1+c)^3 R^3 t^2 + \lambda^2 \rho^2 t \left(2 + (1+c)Rt + e^{(1+c)Rt}((1+c)Rt - 2)\right) \\ &- \lambda\rho \left(e^{(1+c)Rt}(2 - 4(1+c)Rt + (1+c)^2 R^2 t^2) + 2(-1 + (1+c)Rt + (1+c)^2 R^2 t^2) \right) \\ &\leq e^{2(1+c)Rt} \left[(1+c)^3 R^3 + \lambda^2 \rho^2 (2 + 2(1+c)R) + \lambda\rho(4(1+c) + 2) \right] \leq M e^{2(1+c)Rt}. \end{aligned}$$

On the right-hand side, the factor M can be taken to depend only on λ as $\rho \leq 1$, $R \leq \lambda$, and $c \in (0, \eta] \subset (0, 1)$. The companion lower bound for this polynomial is

$$e^{2(1+c)Rt} \left[-2\lambda^2 \rho^2 - 2\lambda\rho - (1+c)^2 R^2 - 2(1+c)R - 2(1+c)^2 R^2 \right] \geq -M' e^{2(1+c)Rt}$$

for a factor M' depending only on λ. Going back to $b(c,t)$, we drop the negative term in the numerator to get

$$\begin{aligned}|b(c,t)| &\leq \frac{e^{cRt}R(R+\lambda\rho(e^{Rt}-1))}{[(1+c)R+\lambda\rho(e^{(1+c)Rt}-1)]^3}\max\{M,M'\}e^{2(1+c)Rt} \\ &\leq \frac{\max\{M,M'\}(R^2+R\lambda\rho)e^{3(1+c)Rt}}{[(1+c)R+\lambda\rho(e^{(1+c)Rt}-1)]^3}. \end{aligned} \tag{15}$$

The proof is completed by using Lemma 4, below, to bound $|b(c,t)|$. By the lemma, we have $\max_i \sup_t |M_t^{(i)}| < \infty$ for the functions $M_t^{(i)}$ defined in the statement below. Therefore, $|b(c,t)|$ is bounded above by some constant independent of c and t. So $b(t) = b(0,t)$ is bounded in t. Since $a(t)$ and $b(t)$ are bounded in t, we can prove the analogous result for scaling R. This is sufficient to ensure $I(f_1, f_2) \geq 1 - \frac{1}{8}\eta^2$ as in Theorem 2. □

Lemma 4. Letting $M'' = \max\{M, M'\}$, we have $|b(c,t)| \leq \max\{M_t^{(1)}, M_t^{(2)}, M_t^{(3)}\}$ where

$$\begin{aligned} M_t^{(1)} &= \frac{e^{3Rt}M''R(R+\lambda\rho)}{(R+\lambda\rho(e^{Rt}-1))^3} \\ M_t^{(2)} &= \frac{e^{3+3\lambda\rho t}M''R(R+\lambda\rho)t^3}{(1+\lambda\rho e^{1+\lambda\rho t}t)^3} \\ M_t^{(3)} &= \frac{M''R(R+\lambda\rho)}{\lambda^3\rho^3}. \end{aligned}$$

Proof. In the two-sided bound (15), we optimize over $c \in (0,\infty)$ for a fixed t. The partial derivative in c is

$$\frac{3e^{3(1+c)Rt}M''R^2(R+\lambda\rho)(-1+(1+c)Rt-\lambda\rho t)}{[(1+c)R+\lambda\rho(e^{(1+c)Rt}-1)]^4},$$

yielding the critical point

$$c^* = \frac{1-Rt+\lambda\rho t}{Rt}.$$

Since the function is bounded, the global maximum is obtained either at a critical point or an end point of the interval $c \in [0,\infty)$. Evaluating the original bound at $c=0$ and c^* and taking the limit $c \to +\infty$ yields $M_t^{(1)}$, $M_t^{(2)}$, and $M_t^{(3)}$, respectively. □

Acknowledgments This research was supported by NSF grants DMS-2052653 and DMS-1646108.

References

Willy Feller. Die grundlagen der volterraschen theorie des kampfes ums dasein in wahrscheinlichkeitstheoretischer behandlung. *Acta Biotheoretica*, 5(1), 1939.

Tanja Gernhard. The conditioned reconstructed process. *Journal of theoretical biology*, 253 (4):769–778, 2008.

Tracy A Heath, John P Huelsenbeck, and Tanja Stadler. The fossilized birth–death process for coherent calibration of divergence-time estimates. *Proceedings of the National Academy of Sciences*, 111(29):E2957–E2966, 2014.

David G. Kendall. On the Generalized "Birth-and-Death" Process. *The Annals of Mathematical Statistics*, 19(1):1 – 15, 1948. doi: 10.1214/aoms/1177730285. URL https://doi.org/10.1214/aoms/1177730285.

Junhyong Kim, Elchanan Mossel, Miklós Z Rácz, and Nathan Ross. Can one hear the shape of a population history? *Theoretical population biology*, 100:26–38, 2015.

Brandon Legried and Jonathan Terhorst. A class of identifiable phylogenetic birth-death models. *bioRxiv*, 2021a. doi: 10.1101/2021.10.04.463015. URL https://www.biorxiv.org/content/early/2021/10/18/2021.10.04.463015.

Brandon Legried and Jonathan Terhorst. Rates of convergence in the two-island and isolation-with-migration models. *bioRxiv*, 2021b.

Stilianos Louca and Matthew W Pennell. Extant timetrees are consistent with a myriad of diversification histories. *Nature*, 580(7804):502–505, April 2020. ISSN 0028-0836, 1476-4687. doi: 10.1038/s41586-020-2176-1.

Hélène Morlon, Todd L Parsons, and Joshua B Plotkin. Reconciling molecular phylogenies with the fossil record. *Proc. Natl. Acad. Sci. U. S. A.*, 108(39):16327–16332, September 2011. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.1102543108.

Elchanan Mossel and Sebastien Roch. Identifiability and inference of non-parametric rates-across-sites models on large-scale phylogenies. *Journal of mathematical biology*, 67(4): 767–797, 2013.

Sean Nee, Robert M. May, and Paul H. Harvey. The reconstructed evolutionary process. *Philosophical Transactions: Biological Sciences*, 344(1309):305–311, 1994.

Tiago B. Quental and Charles R. Marshall. Diversity dynamics: molecular phylogenies need the fossil record. *Trends in ecology and evolution*, 25:434–441, 2010.

Daniel L. Rabosky. Extinction rates should not be estimated from molecular phylogenies. *Evolution*, 64(6):1816–1824, 2010. doi: https://doi.org/10.1111/j.1558-5646.2009.00926.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1558-5646.2009.00926.x.

John A Rhodes and Seth Sullivant. Identifiability of large phylogenetic mixture models. *Bulletin of mathematical biology*, 74(1):212–231, 2012.

Tanja Stadler. On incomplete sampling under birth-death models and connections to the sampling-based coalescent. *J. Theor. Biol.*, 261(1):58–66, November 2009. ISSN 0022-5193, 1095-8541. doi: 10.1016/j.jtbi.2009.07.018.

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

Tanja Stadler and Mike Steel. Swapping Birth and Death: Symmetries and Transformations in Phylodynamic Models. *Systematic Biology*, 68(5):852–858, 05 2019. ISSN 1063-5157. doi: 10.1093/sysbio/syz039. URL https://doi.org/10.1093/sysbio/syz039.

Tanja Stadler, Denise Kühnert, Sebastian Bonhoeffer, and Alexei J Drummond. Birth–death skyline plot reveals temporal changes of epidemic spread in hiv and hepatitis c virus (hcv). *Proceedings of the National Academy of Sciences*, 110(1):228–233, 2013.