A class of identifiable phylogenetic birth-death models

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

October 18, 2021

Abstract

In a striking result, Louca and Pennell (2020) recently proved that a large class of phylogenetic birth-death models are statistically unidentifiable from lineage-through-time (LTT) data: any pair of sufficiently smooth birth and death rate functions is "congruent" to an infinite collection of other rate functions, all of which have the same likelihood for any LTT vector of any dimension. As Louca and Pennell argue, this fact has distressing implications for the thousands of studies that have utilized birth-death models to study evolution.

In this paper, we qualify their finding by proving that an alternative and widely used class of birth-death models is indeed identifiable. Specifically, we show that piecewise constant birth-death models can, in principle, be consistently estimated and distinguished from one another, given a sufficiently large extant time tree and some knowledge of the present-day population. Subject to mild regularity conditions, we further show that any unidentifiable birth-death model class can be arbitrarily closely approximated by a class of identifiable models. The sampling requirements needed for our results to hold are explicit, and are expected to be satisfied in many contexts such as the phylodynamic analysis of a global pandemic.

*Corresponding author: jonth@umich.edu

1 Introduction

The birth-death process (Feller, 1939; Kendall, 1948) is a classic model of population growth. Recently, it has also been used to study the processes of speciation and extinction (Nee et al., 1994; Morlon et al., 2010, 2011; Morlon, 2014), and also the evolution of pathogens (Stadler et al., 2013). Data-driven inquiry in these fields is inherently challenging because the majority of species and pathogens that have ever lived have left no record of their existence. Thus, we can only make inferences about evolution on the basis of a biased sample of the species or lineages that happened to survive to the present day (Quental and Marshall, 2010; Morlon, 2014). Interest in the birth-death process arises in part from the fact that it provides a principled way of correcting this bias (Höhna et al., 2011; Cusimano et al., 2012).

Realizations of the birth-death process can be viewed from a phylogenetic perspective as rooted trees, where each leaf node represents a species that survived until the present, internal nodes are unobserved species, and edges represent lines of descent. The shape of the tree is governed by two non-negative functions which describe, at any given time t before the present, the per-capita rates of birth and death. As noted above, a distinguishing feature of this model is that lineages that died out before the present are not reflected in the resulting tree. Given birth and death rates, as well as a third parameter known as the sampling rate, we refer to the resulting distribution over random trees as a phylogenetic birth-death (henceforth, BD) model. (A precise definition is given in the next section.) The BD model implies a distribution over observed evolutionary data, and given such data, we can use statistical estimation to make inferences about the model parameters.

BD models have been utilized in thousands of published studies (McPeek and Brown, 2007; Condamine et al., 2019; Henao Diaz et al., 2019), despite possessing known and somewhat troubling limitations. Stadler (2009) showed there exist different birth-death models that have the same likelihood in terms of observable data. In statistical terms, this implies that the BD model is *unidentifiable* without further assumptions. The models considered by Stadler are highly parsimonious, consisting of constant birth and death rates that do not change over time. The problem is made even more challenging if the rates are time-varying (Gavryushkina et al., 2014).

Very recently, Louca and Pennell (2020, cited hereafter as LP) proved that the situation is much worse than was previously realized: for *any* reasonably smooth birth and death rate functions, there are *infinitely* many

other such functions that result in the same likelihood over phylogenetic trees. Although each of these functions represents a qualitatively different evolutionary scenario, LP's result shows that it is impossible to tell which of them produced a given data set, even if the data were infinite. In light of the huge number of times that this model has appeared in the literature, this finding is highly worrisome. There they constructed a pulled (birth) rate function that is identifiable from LTT data.

Consistent estimation is impossible in an unidentifiable statistical model, so when faced with one, there are two ways forward: a) use a different model, or b) impose additional regularity conditions on the parameter space to restore identifiability. For the BD model, option a) may be warranted in some settings, but such a debate is beyond our scope. In this paper, we focus on option b). Our main result is to prove that there exists a class of BD models which are identifiable based on LTT data from an extant time tree. By identifiable, we mean that, within the space of rate functions we consider, each distinct BD model corresponds to one and only one likelihood function, and conversely. In fact, this space consists simply of piecewise constant functions, which are already widely used to fit BD models in practice.

Our results show that this class is identifiable once there are enough leaves in the extant tree, and we derive explicit lower bounds on the requisite number of samples. These bounds depend on a measure of parsimony of the underlying model class: they require that identifiable classes of birth-death rate functions do not oscillate unnaturally, in a sense which is made precise below. The same phenomenon has previously been noticed in population genetics (Myers et al., 2008; Bhaskar and Song, 2014), and our proofs are based in part on these earlier works.

2 Preliminaries

In this section, we define the BD model and introduce some key definitions.

Throughout the paper, n is used to denote sample size. We assume $n \geq 2$ and suppress explicit dependence on it when there is no risk of confusion. Given n sampled taxa, an *extant timetree* is a bifurcating tree that traces out the ancestry of the sample. Therefore, the extant timetree has $n-1$ internal nodes which denote the times at which various taxa diverged from common ancestors. These are denoted $0 \leq \tau_{n-1} < \cdots < \tau_1$, where time runs backwards from the present. As in LP, we assume that all n samples are

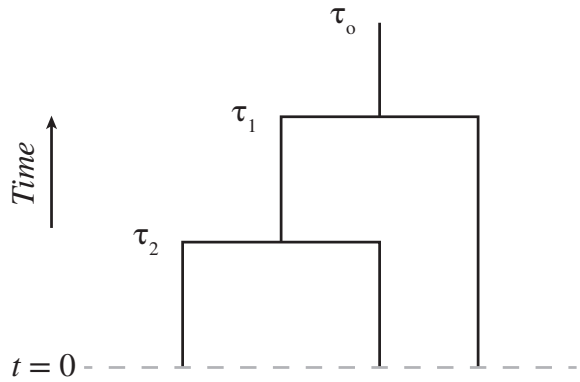

Figure 1: An extant timetree on $n + 1 = 3$ leaves.

collected at time $t = 0$. There is also a root node referred to as the *origin* which occurs at height $\tau_o < \infty$, when the process is assumed to have started. Note that the height of the origin node is not resolvable from character data evolving along the tree since it is ancestral to the entire sample. Its value is therefore conditioned on using prior information. An example of an extant time tree with sample size three is shown in Figure 1.

Extant timetrees are assumed to be stochastically generated by a BD process (Morlon et al., 2010). This process has three parameters: two positive rate functions $\lambda : \mathbb{R}_{\geq 0} \to \mathbb{R}_{>0}$ and $\mu : \mathbb{R}_{\geq 0} \to \mathbb{R}_{>0}$, and an initial sampling fraction $\rho \in (0, 1]$. Here λ and μ represent the instantaneous rate per capita at which lineages are born and die going forwards in time, and ρ is the fraction of extant lineages that are sampled at the present time $t = 0$. In the sequel, we refer to different BD models by their corresponding parameter triples (λ, μ, ρ). Under the BD model with parameters (λ, μ, ρ), the density of an extant time tree is denoted $L^{(\lambda,\mu,\rho)}(\tau_1, \ldots, \tau_n)$. The precise form of $L^{(\lambda,\mu,\rho)}$ is not important for what follows, but can be found in e.g. Morlon et al. (2011, equation 1). Note that the topology of the time tree is uninformative in this model; the likelihood depends only on the merger times τ_i.

Abstracting away from the BD model, we now turn to the concept of identifiability. Let Θ be an arbitrary parameter space, and let L_θ denote a likelihood function parameterized by $\theta \in \Theta$. The *statistical model* $\mathcal{L}_\Theta = \{L_\theta : \theta \in \Theta\}$ is the image of Θ under L_θ; that is, the set of all possible likelihood functions that can be obtained from the parameter space Θ. (If Θ

is a set of BD parameters, we use the notation

$$\mathcal{B}_\Theta = \left\{ L^{(\lambda,\mu,\rho)} : (\lambda, \mu, \rho) \in \Theta \right\} \tag{1}$$

to emphasize that we are focusing specifically on the BD model.)

Definition 1 (Identifiability)**.** The statistical model $\mathcal{L}_\Theta = \{L_\theta : \theta \in \Theta\}$ is *identifiable* if L_θ is injective; that is, for all $\theta_1, \theta_2 \in \Theta$, we have $L_{\theta_1} = L_{\theta_2} \implies \theta_1 = \theta_2$.

(In the BD model (1), the statement "$(\lambda_1, \mu_1, \rho_1) = (\lambda_2, \mu_2, \rho_2)$" is understood to mean that $\rho_1 = \rho_2$, and that the corresponding rate functions are equal almost everywhere.) If different parameters yield the same likelihood function, they cannot be distinguished using any amount of observable data. Identifiability is therefore the most minimal regularity condition one can place on a statistical model.

3 Results

In this section, we summarize LP's results, prove that piecewise constant BD models are identifiable, and explore some additional corollaries and conjectures.

3.1 The result of Louca and Pennell

The key new quantity that underlies LP's result is the so-called *pulled (birth) rate function*, which is defined as

$$\lambda_p = \lambda(1 - E), \tag{2}$$

where $E(t)$ is the probability that a lineage alive at time t has no descendants sampled at time 0. The antiderivative of λ_p is denoted

$$\Lambda_p(\tau) = \int_0^\tau \lambda_p(u)\,\mathrm{d}u.$$

The function E satisfies the ordinary differential equation

$$\frac{\mathrm{d}E}{\mathrm{d}\tau} = -\mu + (\lambda + \mu) \cdot E + E^2\lambda, \tag{3}$$

with initial condition $E(0) = 1 - \rho$. The solution to (3) is (Morlon et al., 2011)

$$E(t) = 1 - \frac{e^{\int_0^t \lambda(u)-\mu(u)\,\mathrm{d}u}}{\rho^{-1} + \int_0^t \lambda(u) e^{\int_0^u \lambda(v)-\mu(v)\,\mathrm{d}v}\,\mathrm{d}u}. \tag{4}$$

Note that $E(t)$ is continuous, even if λ and μ are not.

The pulled rate function completely characterizes the likelihood of an extant timetree. Specifically, by equation (34) of LP,

$$L^{(\lambda_p)}(\tau_1, \ldots, \tau_n) \propto e^{-\Lambda_p(\tau_o)} \prod_{j=1}^{n} \lambda_p(\tau_j) e^{-\Lambda_p(\tau_j)}, \quad 0 \leq \tau_n \leq \cdots \leq \tau_1 < \tau_o. \tag{5}$$

By implication, any two BD parameter triples $(\lambda_1, \mu_1, \rho_1)$ and $(\lambda_2, \mu_2, \rho_2)$ which generate the same λ_p via (2) are indistinguishable. LP's contribution is to show that this phenomenon emerges in a surprisingly general class of models. Restated in our notation, their main result is:

Theorem (LP)**.** *Given an extant timetree on n taxa with origin τ_o, let $\mathcal{C}^1_+[0, \tau_o]$ denote the space of all functions which are strictly positive and continuously differentiable on $[0, \tau_o]$, and let*

$$\mathcal{U} = \left\{(\lambda, \mu, \rho) : \lambda, \mu \in \mathcal{C}^1_+, \rho \in (0, 1]\right\}$$

be the set of all BD parameterizations derived from this space. Then the BD model $\mathcal{B}_{\mathcal{U}}$ is unidentifiable.

Importantly, the theorem holds for any sample size n, and also if ρ is fixed. LP's proof is constructive and provides, for any given BD model, a set of infinitely many "congruent" models which all have the same likelihood. As LP argue in their discussion, this result has disturbing implications for the reliability of statistical estimates obtained from BD models, which have been widely reported in phylogenetics, phylodynamics, paleogenetics, and related fields.

3.2 Piecewise constant models are identifiable

In this section, we state and prove our main results.

Definition 2. Let

$$C_+^{\oplus K}[0,\tau_o] = \left\{ \sum_{k=1}^{K} a_k \mathbf{1}_{[t_{k-1},t_k)}(t) : \mathbf{a} \in \mathbb{R}_{>0}^K, 0 = t_0 < t_1 < \cdots < t_K = \tau_o \right\}$$

be the set of all positive piecewise constant functions with K pieces defined on $[0,\tau_o]$.

Note from the definition that $C_+^{\oplus K}[0,\tau_o]$ encompasses all possible piecewise constant functions with K breakpoints; the breakpoints are not considered fixed or known in advance, and can vary from one function to another.

Next, we define the class of BD parameterizations that forms the basis of our identifiability proof. In the definition and in what follows, we assume that the sampling fraction $\rho \in (0,1]$ is a fixed, known parameter. This is necessary because if ρ is allowed to vary, then as noted in the introduction, Stadler (2009) has shown that even the *constant rates* BD model is unidentifiable.

Definition 3. Let

$$\mathcal{I}_{K,\rho} = \left\{ (\lambda,\mu,\rho) : \lambda,\mu \in C_+^{\oplus K}[0,\tau_o] \right\}$$

be the space of all piecewise-constant BD parameterizations with rate functions in $C_+^{\oplus K}[0,\tau_o]$ and fixed sampling fraction $\rho \in (0,1]$.

The following is our main result.

Theorem 4. *If $n > 8K$, then the BD model $\mathcal{B}_{\mathcal{I}_{K,\rho}}$ is identifiable.*

The theorem is proved in Section 3.3. Unpacking the result, it asserts that *both* the positions of the breakpoints (the vector $\mathbf{t}$ in Definition 2) as well as the levels of each piece (the vector $\mathbf{a}$ in the definition) of both $\lambda(t)$ and $\mu(t)$ are estimable given sufficient data, and that these breakpoints are not assumed to be shared between the two rate functions, or indeed between any two functions in the piecewise constant model space considered by the theorem. If, as is common in practice, we do assume that $\lambda(t)$ and $\mu(t)$ are defined on the same set of breakpoints (while still allowing this set to vary between different parameterizations in $\mathcal{I}_{K,\rho}$), then easy modifications to the proof show that $n > 4K$ suffices for identifiability.

Note that the theorem only establishes a *sufficient* condition for identifiability. It does not imply that piecewise models are unidentifiable if n is

below the stated bounds; in other words, we do not know if our bound is tight. In our view, the main message is that piecewise constant models are identifiable if $\mathcal{O}(K)$ tips are sampled. Further implications of the theorem are discussed in Section 4.

Several extensions and conjectures follow naturally from Theorem 4. Since it is possible to uniformly approximate a regular function class over a compact set using step functions, identifiable BD models are in some sense dense in the space of all BD models. A prototypical result is:

Theorem 5. *Let $\rho \in (0, 1]$ be fixed, let*

$$\mathcal{F} = \left\{ f \in \mathcal{C}^1_+[0, \tau_o] : \|f'\|_\infty < B \right\}$$

be the set of positive, continuously differentiable functions with bounded first derivative over $[0, \tau_o]$, and let $\Theta_{\mathcal{F},\rho} = \{(\lambda, \mu, \rho) : \lambda, \mu \in \mathcal{F}\}$ denote the resulting BD parameter space. Then

1. $\mathcal{B}_{\Theta_{\mathcal{F},\rho}}$ is unidentifiable; and
2. There exists a set of functions $\mathcal{G}$ defined over $[0, \tau_o]$ such that for any $\epsilon > 0$,
   (a) $\sup_{f \in \mathcal{F}} \inf_{g \in \mathcal{G}} \|f - g\|_\infty < \epsilon$, and
   (b) $\mathcal{B}_{\Theta_{\mathcal{G},\rho}}$ is identifiable if $n > 8B\tau_o/\epsilon$.

Proof. The first claim follows from LP, because their congruence classes include e.g. smooth perturbations of constant-rate BD models. For the second, if $f \in \mathcal{F}$ then

$$\begin{aligned}
&\left| \sum_{k=0}^{K-1} f(\tau_o k/K) \mathbf{1}_{[k/K,(k+1)/K)}(x/\tau_o) - f(x) \right| \\
&\quad \leq \max_k \sup_{x/\tau_o \in [k/K,(k+1)/K)} |f(x) - f(\tau_o k/K)| \\
&\quad \leq \max_k \sup_{x/\tau_o \in [k/K,(k+1)/K)} B|x - \tau_o k/K| \\
&\quad = B\tau_o/K.
\end{aligned}$$

Letting $K = B\tau_o/\epsilon$ yields the claim. □

An obvious caveat to this theorem is that the sample size needed to have (provable) identifiability grows rapidly as $\epsilon \to 0$.

Another possible extension relates to estimating birth-death models using polynomials. Since constant functions are polynomials of degree zero, it is natural to conjecture that identifiability holds for higher degrees as well.

Conjecture 6. *Let $P_{d,+}^{(\oplus K)}[0, \tau_o]$ be the set of non-negative, piecewise polynomials (splines) of order d with $K-1$ internal knots defined over $[0, \tau_o]$, and let $\Theta_{P_{d,+}^{(\oplus K)},\rho}[0, \tau_o]$ be the corresponding BD parameter space, again for fixed ρ. Then the BD model $\mathcal{B}_{\Theta_{P_{d,+}^{(\oplus K)},\rho}[0,\tau_o]}$ is identifiable if $n > 8K(1+d)$.*

The conjecture would seem to imply that n grows with d, but this would be offset by having to use fewer pieces to obtain a good approximation. We are unable to prove this conjecture because substantial difficulties arise when trying to extend our proof technique to non-constant functions. Specifically, we do not know how to bound the sign change complexity of spline-based BD models (see Lemma 13) except when $d = 0$.

3.3 Proof of Theorem 4

Our proof derives from a general technique developed by Bhaskar and Song (2014) for establishing identifiability of rate functions in coalescent-type models. We follow their method closely, reproducing their results where necessary for completeness of exposition.

To build the necessary connections between the BD and coalescent models, we first notice that $L^{(\lambda_p)}$ in (5) can be re-written as

$$L(\tau_1, \ldots, \tau_n) \propto \prod_{j=1}^{n} j\lambda_p(\tau_j) e^{-j[\Lambda_p(\tau_j) - \Lambda_p(\tau_{j+1})]} \tag{6}$$

where we defined $\tau_{n+1} \equiv 0$. This is a coalescent-type likelihood where the rate of merging (backwards in time) when there are j remaining lineages in the tree is $j-1$ instead of the usual $\binom{j}{2}$. Hence, retrospectively, lineages are each killed at rate $\lambda_p(\tau)$, except for one which must survive past time τ_o.

Our strategy for establishing identifiability is to construct a vector of invariants which, for a sufficiently large sample size, uniquely identifies the pulled rate function λ_p. To that end, given any pulled rate function λ_p and

sample size n, we form an associated moment vector $\mathbf{c}^{(\lambda_p)} \in \mathbb{R}^n$, with entries

$$c_j^{(\lambda_p)} = \int_0^{\tau_o} e^{-j\Lambda_p(\tau)}\,\mathrm{d}\tau, \quad 1 \le j \le n. \tag{7}$$

Proposition 7. *Suppose that $L^{(\lambda_p^{(1)})}$ and $L^{(\lambda_p^{(2)})}$ are equal almost everywhere. Then $\mathbf{c}^{(\lambda_p^{(1)})} = \mathbf{c}^{(\lambda_p^{(2)})}$.*

Proof. In this proof we will refer to the likelihood function for multiple sample sizes, so we let $L_n^{(\lambda_p)}(\tau_{n1}, \ldots, \tau_{nn})$ be the likelihood of an extant timetree with $n+1$ sampled tips, where the merger times are $\tau_o \ge \tau_{n1} \ge \tau_{n2} \ge \cdots \ge \tau_{nn} \ge 0$. Expectation of a functional $f : \mathbb{R}^n \to \mathbb{R}$ with respect to $L_n^{(\lambda_p)}$ is denoted by $\mathbb{E}_{\lambda_p} f$; by definition, if $L_n^{(\lambda_p^{(1)})} = L_n^{(\lambda_p^{(2)})}$ a.e., then $\mathbb{E}_{\lambda_p^{(1)}} f = \mathbb{E}_{\lambda_p^{(2)}} f$ for all measurable f.

We use some results from Kamm et al. (2017) on moments of the truncated coalescent process, replacing each occurrence of the coalescent rate $\binom{j}{2}/N_e(\tau)$ with its corresponding rate in the BD model, $(j-1)\lambda_p(\tau)$. The expected value $\mathbb{E}_{\lambda_p}(\tau_o - \tau_{n1}) = \tau_o - \mathbb{E}_{\lambda_p}\tau_{n1}$ is written in their notation as $f_{n+1}^{\tau_o}(n+1 \mid \mathcal{A}_{\tau_o}^{(\lambda_p)} = 1)$, where $\mathcal{A}_\tau^{(\lambda_p)}$ is the birth-death process analogue of the coalescent ancestral process, i.e. a pure death process on $\{n+1, \ldots, 1\}$, which begins at state $n+1$ and transitions from state $j+1$ to state j at rate $j\lambda_p(\tau)$. By formulas (3) and (5) of Kamm et al. (2017), we have

$$\begin{aligned}
\tau_o - \mathbb{E}_{\lambda_p}\tau_{n1} &= f_{n+1}^{\tau_o}(n+1 \mid \mathcal{A}_{\tau_o}^{(\lambda_p)} = 1) \\
&= \frac{f_{n+1}^{\tau_o}(n+1)}{\mathbb{P}_{n+1}(\mathcal{A}_{\tau_o}^{(\lambda_p)} = 1)} \\
&= \frac{\tau_o - \sum_{k=1}^{n} \frac{k}{n+1} f_{n+1}^{\tau_o}(k)}{\mathbb{P}_{n+1}(\mathcal{A}_{\tau_o}^{(\lambda_p)} = 1)},
\end{aligned}$$

where $f_{n+1}^{\tau_o}(k)$ is defined below. The quantity $\mathbb{P}_{n+1}(\mathcal{A}_{\tau_o}^{(\lambda_p)} = 1)$ is the probability that the *unconditioned* birth-death process reaches a common ancestor before time τ_o, meaning it is exactly the normalizing constant in equation (6). Rearranging the preceding display and defining $d_n^{(\lambda_p)} = \mathbb{P}_{n+1}(\mathcal{A}_{\tau_o}^{(\lambda_p)} = 1)$, we obtain

$$\sum_{k=1}^{n} \frac{k}{n+1} f_{n+1}^{\tau_o}(k) = \tau_o[1 - d_n^{(\lambda_p)}] + d_n^{(\lambda_p)}\mathbb{E}_{\lambda_p}\tau_{n1}. \tag{8}$$

By Lemma 3.3 of Kamm et al., the summands in (8) are given by

$$f_{n+1}^{\tau_o}(k) = \sum_{m=2}^{n+1} W_{km}^{(n+1)} c_{m-1}^{(\lambda_p)}, \tag{9}$$

where the vector $\mathbf{c}^{(\lambda_p)}$ was defined in (7), and the matrix $\mathbf{W}^{(n)} \in \mathbb{Q}^{(n-1)\times(n-1)}$ is defined by Polanski and Kimmel (2003, formula 10) to have entries

$$W_{bj}^{(n)} = \sum_{k=2}^{j} j(j-1)\binom{n-k}{b-1}\frac{(n-b-1)!(b-1)!}{(n-1)!}A_{kj}^{(n)}. \tag{10}$$

(Here, we follow the convention in the literature of indexing the second axis of $\mathbf{W}^{(n)}$ by $2,\ldots,n$ instead of the usual $1,\ldots,n-1$.) In the above display, the matrix $\mathbf{A}^{(n)}$ holds combinatorial coefficients which again have to be modified from their original definition (formula 6 of Polanski and Kimmel, 2003) to reflect the coalescence rate of the BD process:

$$A_{kj}^{(n)} = \prod_{\ell=k,\ell\neq j}^{n} \frac{\ell-1}{\ell-j}, \quad 2 \le k \le j \le n.$$

Now from (5), we have

$$L_m^{(\lambda_p)}(\tau_1,\ldots,\tau_m) \propto L_n^{(\lambda_p)}(\tau_o,\ldots,\tau_o,\tau_1,\ldots,\tau_m)$$

for any $1 \le m \le n$. Thus, given any $L_n^{(\lambda_p)}$, we may use the above procedure to calculate the moment vector

$$\mathbf{e}^{(\lambda_p)} = (\mathbb{E}_{\lambda_p}\tau_{11}, \mathbb{E}_{\lambda_p}\tau_{21}, \ldots, \mathbb{E}_{\lambda_p}\tau_{n1})^{\intercal}.$$

Define the lower-triangular matrix $\mathbf{B} = (b_{ij}) \in \mathbb{R}^{n\times n}$ to have entries

$$b_{ij} = \sum_{b=1}^{i} \frac{b}{i+1} W_{bj}^{(i+1)}, \quad 1 \le i \le n, \quad 2 \le j \le i+1,$$

where the second axis of $\mathbf{B}$ is indexed in the same manner as $\mathbf{W}^{(n)}$. In Appendix A, we derive a closed-form expression for the entries of $\mathbf{B}$, which shows in particular that the diagonal entries $b_{i,i+1} = i(-1)^{i+1}$. Therefore, $\mathbf{B}$ is invertible, so that by (8),

$$\mathbf{c}^{(\lambda_p)} = \mathbf{B}^{-1}\left(\tau_o[\mathbf{I} - \mathrm{diag}(\mathbf{d}^{(\lambda_p)})] + \mathrm{diag}(\mathbf{d}^{(\lambda_p)})\mathbf{e}^{(\lambda_p)}\right). \tag{11}$$

Finally, suppose that $L_n^{(\lambda_p^{(1)})}$ and $L_n^{(\lambda_p^{(2)})}$ are two BD model likelihoods which are equal almost everywhere. Then there exists

$$0 = t_0 < t_1 < \cdots < t_K = \tau_o$$

such that $L_n^{(\lambda_p^{(1)})} - L_n^{(\lambda_p^{(2)})}$ is continuous on open rectangles of the form

$$\mathcal{R} = (t_{i_1}, t_{i_1+1}) \times \cdots \times (t_{i_n}, t_{i_n+1}) \subset \mathbb{R}^n,$$

and equals zero almost everywhere on each such $\mathcal{R}$. Therefore, the pre-image

$$(L_n^{(\lambda_p^{(1)})} - L_n^{(\lambda_p^{(2)})})^{-1}(\mathbb{R}^n \backslash \{\mathbf{0}\}) \cap \mathcal{R}$$

is an open set of zero measure; the only such set is $\emptyset$. Hence, $L_n^{(\lambda_p^{(1)})} = L_n^{(\lambda_p^{(2)})}$ everywhere on $\mathcal{R}$. In particular, this implies that for all $1 \leq m \leq n$, the BD likelihoods $L_m^{(\lambda_p^{(i)})}$ are equal almost everywhere on $\mathbb{R}^m$. Therefore, the vectors $\mathbf{d}^{(\lambda_p^{(i)})}$ and $\mathbf{e}^{(\lambda_p^{(i)})}$, which are defined entirely in terms integrals of $L_m^{(\lambda_p^{(i)})}$, $1 \leq m \leq n$, are equal for $i = 1, 2$. Equation (11) then implies that $\mathbf{c}^{(\lambda_p^{(1)})} = \mathbf{c}^{(\lambda_p^{(2)})}$. □

Contrapositively, if $\mathbf{c}^{(\lambda_p^{(1)})} \neq \mathbf{c}^{(\lambda_p^{(2)})}$, then $L^{(\lambda_p^{(1)})}$ and $L^{(\lambda_p^{(2)})}$ differ on a set of positive measure. The rest of the proof amounts to showing that if $\lambda_p^{(1)}$ and $\lambda_p^{(2)}$ are generated by piecewise constant BD models, and n is sufficiently large, then they have different moment vectors.

The next theorem is restated for completeness.

Theorem (Generalized rule of signs, Bhaskar and Song, 2014; Jameson, 2006)**.** *Let $f : \mathcal{D} \to \mathbb{R}$ be a piecewise-continuous function defined on some domain $\mathcal{D} \subset \mathbb{R}$, which is not identically zero and has a finite number $\sigma(f)$ of sign changes. Then the function*

$$G(x) = \int_{\mathcal{D}} f(t) e^{-tx} \, \mathrm{d}t$$

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

Informally, f is said to have a sign change any time it crosses zero, including by jump discontinuities. For a precise statement, refer to Definition 3 of Bhaskar and Song.

Given any pulled rate function λ_p, we define its time-rescaled rate function

$$\tilde{\lambda}_p(x) = \lambda_p(\Lambda_p^{-1}(x)), \quad 0 \le x < \Lambda_p(\tau_o).$$

This transformation is invertible, since if

$$S_{\tilde{\lambda}_p}(t) = \int_0^t \left[\tilde{\lambda}_p(u)\right]^{-1} \mathrm{d}u = \Lambda_p^{-1}(t),$$

then $\tilde{\lambda}_p(S_{\tilde{\lambda}_p}^{-1}(t)) = \lambda_p(t)$. Hence,

$$\tilde{\lambda}_p^{(1)} = \tilde{\lambda}_p^{(2)} \iff \lambda_p^{(1)} = \lambda_p^{(2)}.$$

Then the entries of $\mathbf{c}^{(\lambda_p)}$ can be written as

$$c_j^{(\lambda_p)} = \int_0^{\Lambda_p(\tau_o)} \left[\tilde{\lambda}_p(x)\right]^{-1} e^{-jx}\, \mathrm{d}x.$$

Proposition 8. *Suppose that $\lambda_p^{(1)}$ and $\lambda_p^{(2)}$ are two pulled rate functions for which $\mathbf{c}^{(\lambda_p^{(1)})} = \mathbf{c}^{(\lambda_p^{(2)})} \in \mathbb{R}^n$. Then either $\lambda_p^{(1)} = \lambda_p^{(2)}$, or $[\tilde{\lambda}_p^{(1)}]^{-1} - [\tilde{\lambda}_p^{(2)}]^{-1}$ has at least $n-1$ sign changes over the shared domain of $[\tilde{\lambda}_p^{(1)}]^{-1}$ and $[\tilde{\lambda}_p^{(2)}]^{-1}$.*

Proof. Suppose that $\lambda_p^{(1)} \neq \lambda_p^{(2)}$. Then $[\tilde{\lambda}_p^{(1)}]^{-1} - [\tilde{\lambda}_p^{(2)}]^{-1}$ is not identically zero. Assume without loss of generality that $\Lambda_p^{(1)}(\tau_o) \le \Lambda_p^{(2)}(\tau_o)$ so the shared domain of $[\tilde{\lambda}_p^{(1)}]^{-1}$ and $[\tilde{\lambda}_p^{(2)}]^{-1}$ is $[0, \Lambda_p^{(1)}(\tau_o))$. Consider the integral transform given by

$$G(z) = \int_{x=0}^{\Lambda_p^{(2)}(\tau_o)} \left[u_1(x) - \frac{1}{\tilde{\lambda}_p^{(2)}(x)}\right] e^{-zx}\, dx,$$

where

$$u_1(x) = \begin{cases} [\tilde{\lambda}_p^{(1)}(x)]^{-1} & x \in [0, \Lambda_p^{(1)}(\tau_o)) \\ 0 & \text{otherwise.} \end{cases}$$

The supposition implies that $G(z)$ has zeros at $z = 1, \ldots, n$. By the generalized rule of signs, this implies that $u_1 - [\tilde{\lambda}_p^{(2)}]^{-1}$ has at least n sign changes on $[0, \Lambda_p^{(2)}(\tau_o))$. There is at most one sign change on the interval $[\Lambda_p^{(1)}(\tau_o), \Lambda_p^{(2)}(\tau_o))$, caused by a possible jump at $\Lambda_p^{(1)}(\tau_o)$. This implies that $u_1 - [\tilde{\lambda}_p^{(2)}]^{-1}$ has at least $n-1$ sign changes on $[0, \Lambda_p^{(1)}(\tau_o))$. $\square$

Based on the preceding result, we define the following complexity measure on BD model spaces. This is an adaptation of the Definition 4 in Bhaskar and Song to our setting. In the definition, the notation $\lambda_p^{(\theta)}$ is used to denote the pulled rate function corresponding to a particular BD parameterization $\theta = (\lambda, \mu, \rho)$.

Definition 9 (Pulled sign change complexity)**.** Let Θ be a set of BD models. The *pulled sign change complexity* $\mathscr{S}_p(\Theta)$ is defined as

$$\mathscr{S}_p(\Theta) = \sup_{\substack{\theta_1,\theta_2\in\Theta \\ a\geq 0}} \left\{ \sigma(g) : \begin{array}{c} g(x) = \left[\tilde{\lambda}_p^{(\theta_1)}(x)\right]^{-1} - \left[\tilde{\lambda}_p^{(\theta_2)}(x-a)\right]^{-1} \\ \mathrm{dom}(g) = \left[\max\{0,a\}, \min\{\Lambda_p^{(1)}(\tau_o), a + \Lambda_p^{(2)}\}\right) \end{array} \right\},$$

where $\sigma(g)$ denotes the number of sign changes of g.

Using Proposition 8 and the preceding definition, we immediately have the following sample size criterion for the identifiability of BD models.

Theorem 10. *Suppose that $\mathscr{S}_p(\Theta) \leq S$, and that the mapping $\theta \mapsto \lambda_p^{(\theta)}$ is injective over Θ. Then $\mathcal{B}_\Theta$ is identifiable if $n > S + 1$.*

Theorem 10 is a general result that holds for any class of BD models. However, we can only apply it to spaces where the injectivity condition holds and where we know how to bound the pulled sign change complexity. Next, we establish these properties when $\Theta = \mathcal{I}_{K,\rho}$.

Recall that $\tilde{\lambda}_p(x) = \lambda_p(\Lambda_p^{-1}(x))$ for $x \in [0, \Lambda_p(\tau_o))$. By supplemental equation (9) of LP,

$$\frac{\mathrm{d}\lambda_p}{\mathrm{d}t} = \lambda_p \left(\frac{1}{\lambda}\frac{\mathrm{d}\lambda}{\mathrm{d}t} - \mu + \lambda E \right).$$

Hence,

$$\frac{\mathrm{d}\tilde{\lambda}_p}{\mathrm{d}x} = \frac{\mathrm{d}\lambda_p}{\mathrm{d}t}(\Lambda_p^{-1}(x)) \times \frac{\mathrm{d}\Lambda_p^{-1}}{\mathrm{d}x}(x) = \frac{1}{\lambda}\frac{\mathrm{d}\lambda}{\mathrm{d}t} - \mu + \lambda E \Big|_{t=\Lambda_p^{-1}(x)}, \tag{12}$$

where in the second equality we used

$$\frac{\mathrm{d}\Lambda_p^{-1}}{\mathrm{d}x} = \frac{1}{\lambda_p(\Lambda_p^{-1}(x))} = [\tilde{\lambda}_p(x)]^{-1}.$$

Now by (2),

$$\lambda E|_{t=\Lambda_p^{-1}(x)} = -\tilde{\lambda}_p(x) + \lambda(\Lambda_p^{-1}(x)), \tag{13}$$

If λ and μ are constant, then $\mathrm{d}\lambda/\mathrm{d}t = 0$, and we obtain from (12) and (13) the first-order ordinary differential equation

$$\begin{aligned} \frac{\mathrm{d}\tilde{\lambda}_p}{\mathrm{d}x} &= \lambda - \mu - \tilde{\lambda}_p \\ \tilde{\lambda}_p(0) &= \rho\lambda. \end{aligned} \tag{14}$$

The solution to this differential equation is

$$\tilde{\lambda}_p(x) = (\lambda - \mu)(1 - e^{-x}) + \rho\lambda e^{-x}, \quad x \in [0, \Lambda_p(\tau_o)). \tag{15}$$

More generally, if λ and μ are constant over some interval $[t, t')$, then

$$\tilde{\lambda}_p(x) = (\lambda - \mu)(1 - e^{-(x-\Lambda_p(t))}) + \lambda_p(t)e^{-(x-\Lambda_p(t))}, \quad x \in [\Lambda_p(t), \Lambda_p(t')). \tag{16}$$

Proposition 11. *Let θ_1, θ_2 be two different models in $\mathcal{I}_{K,\rho}$ with pulled rate functions $\lambda_p^{(1)}$ and $\lambda_p^{(2)}$. Then $\lambda_p^{(1)} \neq \lambda_p^{(2)}$.*

Proof. Let $(\lambda_1, \mu_1) \neq (\lambda_2, \mu_2)$ be two different models in $\mathcal{I}_{K,\rho}$. Then there is a non-empty interval $[t, t') \subset [0, \tau_o]$ such that

1. $(\lambda_1(s), \mu_1(s)) = (\lambda_2(s), \mu_2(s))$ for all $0 < s < t$; and
2. $\lambda_1, \mu_1, \lambda_2, \mu_2$ are all constant over $[t, t')$ and $(\lambda_1(s), \mu_1(s)) \neq (\lambda_2(s), \mu_2(s))$ for all $t \leq s < t'$.

(Note that we could have $t = 0$, in which case condition 1 becomes vacuous.) To show $\lambda_p^{(1)} \neq \lambda_p^{(2)}$, it is sufficient to show that $\tilde{\lambda}_p^{(1)} \neq \tilde{\lambda}_p^{(2)}$. We assume that $\Lambda_p^{(1)}(t') = \Lambda_p^{(2)}(t')$, since if not the conclusion is immediate. By (16), for all $x \in [\Lambda_p^{(1)}(t), \Lambda_p^{(1)}(t'))$, we have

$$\tilde{\lambda}_p^{(2)}(x) - \tilde{\lambda}_p^{(1)}(x) = c_1 e^{-[x-\Lambda_p^{(1)}(t)]} + c_2,$$

where

$$\begin{aligned} c_1 &= \lambda_p^{(1)}(t) - \lambda_p^{(2)}(t) - \lambda_2 + \mu_2 + \lambda_1 - \mu_1 \\ c_2 &= \lambda_2 - \mu_2 - \lambda_1 + \mu_1. \end{aligned}$$

Suppose $c_2 = 0$. Let $\varepsilon = E^{(1)}(t) = E^{(2)}(t) \in (0, 1)$, where the equality follows from condition 1 and the facts that a) $E(0) = \rho$ across all models, and b) $E(t)$ is continuous (cf. equation 4). Then $c_1 = \lambda_p^{(1)}(t) - \lambda_p^{(2)}(t) = [\lambda_1(t) - \lambda_2(t)](1 - \varepsilon)$. If $c_1 = 0$ then this would contradict condition 2. □

Remark. The preceding result makes crucial use of the fact that all models in $\mathcal{I}_{K,\rho}$ are constrained to have the same sampling fraction ρ. Without this assumption, the proposition would not even hold for $K = 1$ (Stadler, 2009).

Next, we bound $\mathscr{S}_p(\mathcal{I}_{K,\rho})$. First, let

$$\mathcal{S}_K = \bigcup_{\rho \in (0,1]} \mathcal{I}_{K,\rho}$$

be the space of all K-piecewise constant BD models with unconstrained sampling proportions. As remarked above, this space is not identifiable, since in particular Proposition 11 does not hold for it. Nevertheless, it follows directly from Definition 9 that $\mathscr{S}_p(\mathcal{I}_{K,\rho}) \leq \mathscr{S}_p(\mathcal{S}_K)$, so bounding the pulled sign change complexity of $\mathcal{S}_K$ is all that is required for our purposes.

We first show that $\mathscr{S}_p(\mathcal{S}_K)$ can be bounded in terms of the simpler quantity $\mathscr{S}_p(\mathcal{S}_1)$.

Lemma 12. *The pulled sign change complexity of $\mathcal{S}_K$ is bounded by the pulled sign change complexity of $\mathcal{S}_1$ as*

$$\mathscr{S}_p(\mathcal{S}_K) \leq (4K - 1) + 4K\mathscr{S}_p(\mathcal{S}_1).$$

Proof. Let $\lambda_p^{(i)}$ be the pulled rate function corresponding to $(\lambda_i, \mu_i, \rho_i)$ for $i = 1, 2$. According to Definition 9, we need to bound all sign changes of

$$\left[\tilde{\lambda}_p^{(1)}(x)\right]^{-1} - \left[\tilde{\lambda}_p^{(2)}(x - a)\right]^{-1} \tag{17}$$

over the domain $x \in [m, M)$, where

$$\begin{aligned} m &= \max\{0, a\} \\ M &= \min\{\Lambda_p^{(1)}(\tau_o), a + \Lambda_p^{(2)}(\tau_o)\}. \end{aligned}$$

Enlarging the domain of (17) can only increase the number of sign changes, and the largest possible domain occurs when $a = 0$ and $\Lambda_p^{(1)}(\tau_o) = \Lambda_p^{(2)}(\tau_o)$, so we assume these conditions hold for the rest of the proof.

If $\lambda_i, \mu_i \in C_+^{\oplus K}$, then we can place them onto a common set of $2K$ breakpoints

$$0 = t_0^{(i)} < t_1^{(i)} < \cdots < t_{2K}^{(i)} = \tau_o.$$

Let

$$\mathcal{X} = \left\{\Lambda_p^{(i)}(t_k^{(i)}) : 1 \leq i \leq 2, 0 \leq k \leq 2K\right\},$$

and sort the points in $\mathcal{X}$ to form a partition

$$0 = x_0 < \cdots < x_{4K} = \Lambda_p^{(1)}(\tau_o) = \Lambda_p^{(2)}(\tau_o).$$

Allowing for possible jump discontinuities at $x_1, x_2, \ldots, x_{4K-1}$, the number of sign changes of (17) is at most $4K-1$ plus the number of sign changes on each interval (x_j, x_{j+1}).

For each i and j, there exists an integer $0 \le k(i,j) < 2K$ such that

$$(x_j, x_{j+1}) \subset \left(\Lambda_p^{(i)}(t_{k(i,j)}^{(i)}), \Lambda_p^{(i)}(t_{k(i,j)+1}^{(i)})\right), \quad i = 1, 2.$$

Therefore, there exists a BD parameterization $\theta_{ij} = (\lambda_{ij}, \mu_{ij}, \rho_{ij}) \in \mathcal{S}_1$ such that

$$\lambda_p^{(\theta_{ij})}(s - t_{k(i,j)}^{(i)}) = \lambda_p^{(i)}(s), \quad s \in (t_{k(i,j)}^{(i)}, t_{k(i,j)+1}^{(i)});$$

concretely, the initial sampling fraction is

$$\rho_{ij} = 1 - E^{(i)}(t_{k(i,j)}^{(i)}).$$

Then

$$\left[\tilde{\lambda}_p^{(i)}(x)\right]^{-1} = \left[\lambda_p(I^{(i)}(x))\right]^{-1} = \left[\tilde{\lambda}_p^{(\theta_{ij})}\left(x - \Lambda_p^{(i)}(t_{k(i,j)}^{(i)})\right)\right]^{-1}, \quad i = 1, 2.$$

So within (x_j, x_{j+1}), the number of sign changes of $[\tilde{\lambda}_p^{(1)}(x)]^{-1} - [\tilde{\lambda}_p^{(2)}(x)]^{-1}$ is at most the number of sign changes of

$$\left[\tilde{\lambda}_p^{(\theta_{1j})}\left(x - \Lambda_p^{(1)}(t_{k(1,j)}^{(1)})\right)\right]^{-1} - \left[\tilde{\lambda}_p^{(\theta_{2j})}\left(x - \Lambda_p^{(2)}(t_{k(i,j)}^{(2)})\right)\right]^{-1},$$

which is bounded above by $\mathscr{S}_p(\mathcal{S}_1)$. Hence, the number of sign changes is at most $(4K-1) + 4K\mathscr{S}_p(\mathcal{S}_1)$. □

We conclude the proof by bounding $\mathscr{S}(\mathcal{S}_1)$.

Lemma 13. *Let $(\lambda_1, \mu_1, \rho_1), (\lambda_2, \mu_2, \rho_2) \in \mathcal{S}_1$, with corresponding pulled rate functions $\lambda_p^{(1)}$ and $\lambda_p^{(2)}$, and let*

$$g(x) = \left[\tilde{\lambda}_p^{(1)}(x)\right]^{-1} - \left[\tilde{\lambda}_p^{(2)}(x-a)\right]^{-1},$$

where $a \ge 0$ is arbitrary and the domain of g is as indicated in Definition 9. Then $\sigma(g) \le 1$.

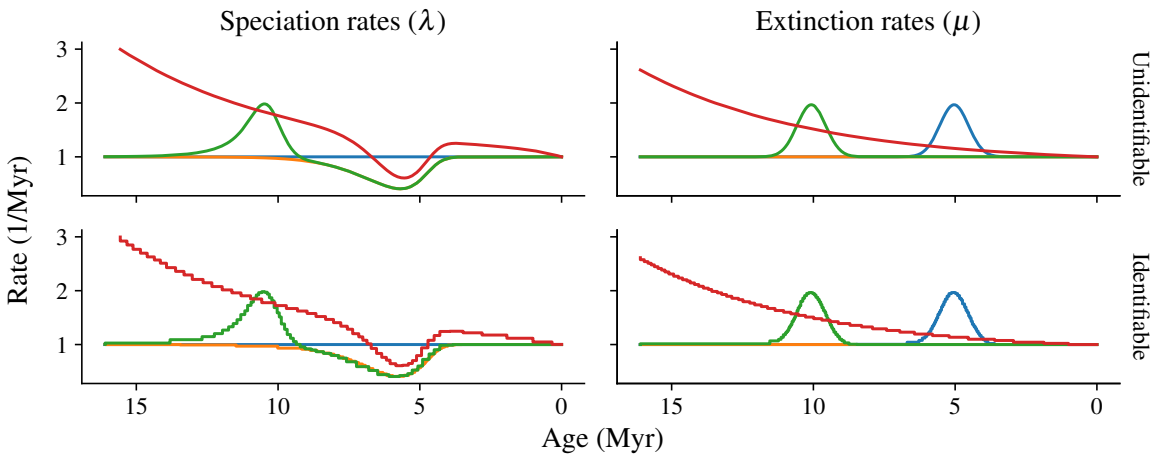

Figure 2: Unidentifiable vs. identifiable BD models. The top row contains four indistinguishable models exhibited in LP Figure 1. In the bottom row, we approximated these models using piecewise constant functions using $K = 50$ pieces. The models in the bottom row are identifiable given sufficiently many samples. All models are assumed to have the same ρ.

Proof. We have

$$g(x) = \frac{\tilde{\lambda}_p^{(2)}(x-a) - \tilde{\lambda}_p^{(1)}(x)}{\tilde{\lambda}_p^{(1)}(x)\tilde{\lambda}_p^{(2)}(x-a)},$$

so the number of sign changes of g is at most the number of zeros of $\tilde{\lambda}_p^{(2)}(x-a) - \tilde{\lambda}_p^{(1)}(x)$. By (15), the function $\tilde{\lambda}_p^{(2)}(x-a) - \tilde{\lambda}_p^{(1)}(x)$ has the form $c_1 e^{-x} + c_2$ for some c_1, c_2 that depend on λ_i, μ_i, ρ_i, and a. Since this function is always monotone, $\tilde{\lambda}_p^{(2)}(x-a) - \tilde{\lambda}_p^{(1)}(x)$ and hence g has at most one zero. □

By the preceding lemma, $\mathscr{S}_p(\mathcal{S}_1) \leq 1$, so that by Lemma 12,

$$\mathscr{S}_p(\mathcal{S}_K) \leq 8K - 1.$$

Finally, the claim follows from Theorem 10.

4 Discussion

In this paper, we proved that piecewise-constant BD models are identifiable from extant timetrees with a sufficient number of tips. We also showed that,

under mild assumptions, unidentifiable BD models of the type considered by LP can be approximated to within arbitrary accuracy by identifiable BD models. Based on these results, we conjecture, but are unable to prove, that spline-based BD models are similarly identifiable.

In the short time since their publication, LP's findings have generated considerable discussion (e.g., Morlon et al., 2020; Helmstetter et al., 2021; Parry, 2021), with some authors concluding that they "will be dispiriting to evolutionary scientists" seeking to understand the factors affecting speciation and extinction (Pagel, 2020). Our results may serve to lift those spirits, while also illustrating potential subtleties that can arise when reasoning about a limiting concept like identifiability. For example, consider the BD models shown in Figure 2. The top row is reproduced from Figure 1 of LP and shows four color-coded BD models which all have the same pulled birth rate, and hence the same likelihood function. On the bottom, we approximated these functions over the domain $[0, 16]$ using piecewise constant functions. By Theorem 4, these models can, in principle, be distinguished given a sufficiently large timetree. Is the underlying natural process that is modeled in Figure 2 inferable from data? The answer seemingly depends on whether the researcher believes that the piecewise functions shown on the bottom panel can faithfully represent this process. Empirically, we note that it would be nearly impossible to differentiate (using, say, a simple hypothesis test) between one of the $\mathcal{C}_+^1$ models in the figure and its corresponding $C_+^{\oplus 50}$ approximation on the basis of a realistically achievable amount of data.

Still, practitioners should be careful not to overinterpret affirmative identifiability results as conclusive evidence that high quality estimates can be obtained on real problems. As we noted above, identifiability is a minimal regularity criterion that offers no guarantees about the ability to accurately infer model parameters using a finite amount of data. Even in identifiable models, it is often the case that potentially significant amounts of regularization and/or prior information have to be incorporated to obtain sensible results (Stadler et al., 2013; Kim et al., 2015; Bhaskar et al., 2015; Morlon et al., 2020; DeWitt et al., 2021). Having established identifiability, the next step would be to understand the finite-sample accuracy and rate of convergence of piecewise constant estimators in BD models. This is a challenging theoretical problem that will require new ideas and techniques. Fortunately, since several popular software packages (e.g., Bouckaert et al., 2019) already implement the piecewise constant BD model, there are already many simulation studies in the literature to help guide the way.

The reader may wonder whether our result is somehow a byproduct of the fact that we consider piecewise constant—hence discontinuous—rate functions, whereas in LP they are assumed to be continuously differentiable. In our opinion, this is not the main driver. Indeed, we believe that (cf. Conjecture 6) identifiable parameter spaces consisting of smooth functions also exist. Were it true, this conjecture would not contradict LP's result, because the construction they use to generate their congruence classes (specifically, the operator $S[S_o, f]$ defined by supplementary equation 75) is not closed over simple function spaces like fixed-degree polynomials. In other words, even f is e.g. a spline, $S[S_o, f]$ is not. Thus, while there are infinitely large congruence classes of alternative BD parameterizations that are indistinguishable, the conjecture asserts that the intersection between these classes and a sufficiently simple function space consists of at most a single element. LP provide an heuristic argument supporting this conjecture in Section S.3 of their supplement.

Another criticism related to our use of piecewise constant models is that they require a potentially large number of samples to become (provably) identifiable. In some settings (e.g., phylodynamic analysis of a global pandemic) these bounds would be achievable, while in others they would plainly exceed the amount of available data. Using a more efficient representation like splines would lead to BD models with lower sampling requirements, but which currently lack identifiability guarantees.

In follow-up work, Louca et al. (2021) study a more general model where sampling is allowed to occur over time, and show that similar unidentifiability results hold in that setting as well. The coalescent-based methods we used in this paper, which condition on a number of lineages sampled at the present, do not readily extend to this setting, so our results leave open the question of whether piecewise-constant identifiability holds in random sampling models as well. In Section S.2.2 of their supplement, Louca et al. (2021) assert that restricting to piecewise constant model spaces cannot possibly resolve identifiability issues, however their argument is nonrigorous and based only on simulation evidence. Rather than posing a contradiction, we think their findings underscore the conclusions reached above: identifiability is fundamentally a mathematical property that cannot be checked on the basis of simulations alone, and it may have little bearing on one's (in)ability to successfully carry out inference in real-world problems.

Acknowledgments We thank Andy Magee, Sebastian Prillo, and Ed Ionides for helpful comments on a draft of this manuscript. All errors are our own. This research was supported by NSF grants DMS-2052653 and DMS-1646108.

References

A Bhaskar and Yun S Song. Descartes' rule of signs and the identifiability of population demographic models from genomic variation data. *Annals of Statistics*, 42(6):2469–2493, 2014.

A. Bhaskar, Y. X. Rachel Wang, and Y. S. Song. Efficient inference of population size histories and locus-specific mutation rates from large-sample genomic variation data. *Genome Research*, 25(2):268–279, 2015.

Remco Bouckaert, Timothy G Vaughan, Joëlle Barido-Sottani, Sebastián Duchêne, Mathieu Fourment, Alexandra Gavryushkina, Joseph Heled, Graham Jones, Denise Kühnert, Nicola De Maio, et al. BEAST 2.5: An advanced software platform for Bayesian evolutionary analysis. *PLoS computational biology*, 15(4):e1006650, 2019.

Fabien L. Condamine, Jonathan Rolland, and Hélène Morlon. Assessing the causes of diversification slowdowns: temperature-dependent and diversity-dependent models receive equivalent support. *Ecology Letters*, 22(11):1900–1912, 2019. doi: https://doi.org/10.1111/ele.13382. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/ele.13382.

Natalie Cusimano, Tanja Stadler, and Susanne S. Renner. A New Method for Handling Missing Species in Diversification Analysis Applicable to Randomly or Nonrandomly Sampled Phylogenies. *Systematic Biology*, 61(5): 785–792, 02 2012. ISSN 1063-5157. doi: 10.1093/sysbio/sys031. URL https://doi.org/10.1093/sysbio/sys031.

William S DeWitt, Kameron Decker Harris, Aaron P Ragsdale, and Kelley Harris. Nonparametric coalescent inference of mutation spectrum history and demography. *Proceedings of the National Academy of Sciences*, 118 (21), 2021.

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

Alexandra Gavryushkina, David Welch, Tanja Stadler, and Alexei J. Drummond. Bayesian inference of sampled ancestor trees for epidemiology and fossil calibration. *PLoS Computational Biology*, 10(12), 2014.

Andrew J Helmstetter, Sylvain Glemin, Jos Kafer, Rosana Zenil-Ferguson, Herve Sauquet, Hugo de Boer, Leo-Paul MJ Dagallier, Nathan Mazet, Eliette L Reboud, Thomas LP Couvreur, et al. Pulled diversification rates, lineage-through-time plots and modern macroevolutionary modelling. *bioRxiv*, 2021.

L. Francisco Henao Diaz, Luke J. Harmon, Mauro T. C. Sugawara, Eliot T. Miller, and Matthew W. Pennell. Macroevolutionary diversification rates show time dependency. *Proceedings of the National Academy of Sciences*, 116(15):7403–7408, 2019. ISSN 0027-8424. doi: 10.1073/pnas.1818058116. URL https://www.pnas.org/content/116/15/7403.

Sebastian Höhna, Tanja Stadler, Fredrik Ronquist, and Tom Britton. Inferring Speciation and Extinction Rates under Different Sampling Schemes. *Molecular Biology and Evolution*, 28(9):2577–2589, 04 2011. ISSN 0737-4038. doi: 10.1093/molbev/msr095. URL https://doi.org/10.1093/molbev/msr095.

GJO Jameson. Counting zeros of generalised polynomials: Descartes' rule of signs and Laguerre's extensions. *The Mathematical Gazette*, 90(518): 223–234, 2006.

John A. Kamm, Jonathan Terhorst, and Yun S. Song. Efficient computation of the joint sample frequency spectra for multiple populations. *Journal of Computational and Graphical Statistics*, 26(1):182–194, 2017. doi: 10.1080/10618600.2016.1159212.

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.

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.

Stilianos Louca, Angela McLaughlin, Ailene MacPherson, Jeffrey B Joy, and Matthew W Pennell. Fundamental Identifiability Limits in Molecular Epidemiology. *Molecular Biology and Evolution*, 38(9):4010–4024, 05 2021. ISSN 0737-4038. doi: 10.1093/molbev/msab149. URL https://doi.org/10.1093/molbev/msab149.

Mark A McPeek and Jonathan M Brown. Clade age and not diversification rate explains species richness among animal taxa. *The American Naturalist*, 169(4):E97–E106, 2007.

Hélène Morlon. Phylogenetic approaches for studying diversification. *Ecology Letters*, 17(4):508–525, 2014. doi: https://doi.org/10.1111/ele.12251. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/ele.12251.

Hélène Morlon, Matthew D Potts, and Joshua B Plotkin. Inferring the dynamics of diversification: a coalescent approach. *PLoS biology*, 8(9): e1000493, 2010.

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.

Hélène Morlon, Florian Hartig, and Stephane Robin. Prior hypotheses or regularization allow inference of diversification histories from extant timetrees. *bioRxiv*, 2020.

Simon Myers, Charles Fefferman, and Nick Patterson. Can one learn history from the allelic spectrum? *Theor. Popul. Biol.*, 73(3):342–348, 2008.

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

Mark Pagel. Evolutionary trees can't reveal speciation and extinction rates. *Nature*, 580(7804):461–462, April 2020. doi: 10.1038/d41586-020-01021-4. URL https://doi.org/10.1038/d41586-020-01021-4.

Luke A Parry. Evolution: No extinction? no way! *Current Biology*, 31(14): R907–R909, 2021.

Andrzej Polanski and Marek Kimmel. New explicit expressions for relative frequencies of single-nucleotide polymorphisms with application to statistical inference on population growth. *Genetics*, 165(1):427–436, Sep 2003.

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

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, 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.

A Computation of the matrix B

From the binomial identity

$$\sum_{k=0}^{m} \binom{m}{k} \Big/ \binom{n}{k} = \frac{n+1}{n+1-m},$$

we obtain for $k \geq 2$,

$$\begin{aligned}
\sum_{b=1}^{n-1} \frac{\binom{n-k}{b-1}}{\binom{n-1}{b}} &= \sum_{b=1}^{n-1} \frac{\binom{n-k+1}{b} - \binom{n-k}{b}}{\binom{n-1}{b}} \\
&= \sum_{b=0}^{n-1} \frac{\binom{n-k+1}{b}}{\binom{n-1}{b}} - \sum_{b=0}^{n-1} \frac{\binom{n-k}{b}}{\binom{n-1}{b}} \\
&= \sum_{b=0}^{n-k+1} \frac{\binom{n-k+1}{b}}{\binom{n-1}{b}} - \sum_{b=0}^{n-k} \frac{\binom{n-k}{b}}{\binom{n-1}{b}} \\
&= \frac{n}{n-(n-k+1)} - \frac{n}{n-(n-k)} \\
&= \frac{n}{k(k-1)}.
\end{aligned}$$

Hence,

$$\sum_{b=1}^{n-1} \frac{b}{n} W_{bj}^{(n)} = j(j-1) \sum_{k=2}^{j} \frac{A_{kj}^{(n)}}{k(k-1)}.$$

Furthermore,

$$
\begin{aligned}
\sum_{k=2}^{j} \frac{A_{kj}^{(n)}}{k(k-1)} &= \sum_{k=2}^{j} \frac{\prod_{\ell=k,\ell\neq j}^{n} \frac{\ell-1}{\ell-j}}{k(k-1)} \\
&= \prod_{\ell=j+1}^{n} \frac{\ell-1}{\ell-j} \sum_{k=2}^{j} \frac{\prod_{\ell=k}^{j-1} \frac{\ell-1}{\ell-j}}{k(k-1)} \\
&= \prod_{\ell=j+1}^{n} \frac{\ell-1}{\ell-j} \left(\frac{1}{j(j-1)} + \sum_{k=2}^{j-1} \frac{\frac{(j-2)!}{(k-2)!} \times (-1)^{j+k}}{k(k-1)(j-k)!} \right) \\
&= \prod_{\ell=j+1}^{n} \frac{\ell-1}{\ell-j} \left(\frac{1}{j(j-1)} + \frac{1}{j(j-1)} \sum_{k=2}^{j-1} (-1)^{j+k} \binom{j}{k} \right) \\
&= \prod_{\ell=j+1}^{n} \frac{\ell-1}{\ell-j} \left(\frac{1}{j(j-1)} + \frac{(-1)^j}{j(j-1)} \left[-1 + (-1)^{1+j} + j \right] \right) \\
&= \prod_{\ell=j+1}^{n} \frac{\ell-1}{\ell-j} \times \frac{(-1)^j}{j} \\
&= \frac{(n-1)!}{(j-1)!(n-j)!} \times \frac{(-1)^j}{j} \\
&= (-1)^j \frac{(n-1)!}{j!(n-j)!}.
\end{aligned}
$$

We conclude that

$$
\sum_{b=1}^{n-1} \frac{b}{n} W_{bj}^{(n)} = (-1)^j \frac{(n-1)!}{(j-2)!(n-j)!}.
$$