Parameter Identifiability of a Multitype Pure-Birth Model of Speciation

Dakota Dragomir,* Elizabeth S. Allman, John A. Rhodes
Department of Mathematics and Statistics, University of Alaska Fairbanks,
PO Box 756660, Fairbanks, AK 99775
*To whom correspondence should be addressed;
E-mail: dbdragomir@alaska.edu.

January 17, 2023

Keywords: Diversification model, Multitype branching process

Abstract: Diversification models describe the random growth of evolutionary trees, modeling the historical relationships of species through speciation and extinction events. One class of such models allows for independently changing traits, or types, of the species within the tree, upon which speciation and extinction rates depend. Although identifiability of parameters is necessary to justify parameter estimation with a model, it has not been formally established for these models, despite their adoption for inference. This work establishes generic identifiability up to label swapping for the parameters of one of the simpler forms of such a model, a multitype pure birth model of speciation, from an asymptotic distribution derived from a single tree observation as its depth goes to infinity. Crucially for

applications to available data, no observation of types is needed at any internal points in the tree, nor even at the leaves.

1 Introduction

Species diversification models are used in Biology to make inferences about historical speciation and extinction rates over the time since a group of species, or *taxa*, evolved from a common ancestor. By providing information on rates of speciation and extinction, inference with these models seeks to give insight into the evolutionary dynamics leading to the present diversity of life. These models have a long history, starting with the constant-rate pure-birth model of Yule (1925), and a fairly large literature has developed.

Diversification models describe a process beginning with a single lineage at some time in the past, which as time progresses may speciate or go extinct. When a speciation occurs the edge bifurcates into two edges, with the number of lineages increasing by 1. When an extinction occurs, the lineage ends, and the number of lineages decreases by 1. After either event, the process continues forward, independently on all lineages, producing a growing tree structure until the present time is reached. This tree, which has both topological and metric structure, constitutes an observation. (In applications, it may be necessary to consider the *reconstructed tree*, which is obtained by removing all tree edges with no descendents at the present (Nee et al., 1994; Harvey et al., 1994).)

Two basic sorts of these models have found common use in empirical studies. In the first, the speciation and extinction rates are functions of time, and apply to all taxon lineages present at any moment. This can be thought of as modeling exogenous factors, such as environmental conditions, that affect all taxa in the tree identically. Since all lineages behave in the same probabilistic way at any moment, it is not hard to show that the exact branching pattern

of the tree-structure is irrelevant, with all the information in a tree observation being captured by the number of lineages as a function of time. Thus the work on *time-dependent birth-death models* by Kendall (1948) is foundational.

In the second sort of diversification model, which we call the *multitype birth-death tree model*, lineages are assigned one of a finite number of *types* at each moment, with the model's speciation and extinction rates dependent only on the type. Over time, however, species may change types at fixed switching rates. This models endogenous factors, such as a particular biological trait a taxon may possess, including, for instance, a morphological feature, behavior, or whether a particular gene is present and active in an organism. A given type might correlate with faster or slower speciation than another, and/or affect the extinction rate. For these models the branching structure of a tree observation does matter, as taxa present at a given time may each have different types, and thus different tendencies to speciate or go extinct.

The Binary State-specific Speciation and Extinction (BiSSE) model of Maddison et al. (2007) formalized the multitype framework for biological applications. Multitype (MuSSE) and quantitative-type (QuaSSE) variants of the model were subsequently proposed by FitzJohn (2012). Although these works assumed the type is observed for the extant taxa at the leaves of a tree, we consider the multitype birth-death tree model with no type information observable for any lineage at any time, as type observations are unnecessary for our results. Indeed, the usefulness of these models to infer correlation between observed types and diversification rates from data with type information for extant taxa has been called into question (Rabosky and Goldberg, 2015).

Many other diversification models have been proposed, combining or extending these basic frameworks, with Stadler (2013) offering one review. New variants continue to be developed, (e.g., Cantalapiedra et al., 2014; Maliet et al.,

2019; Stadler, 2019; Rasmussen and Stadler, 2019; Barido-Sottani et al., 2020).

When these models are used for inference, the data is taken as a single tree assumed to show the true evolutionary relationships of the taxa. (In practice, this tree itself must be inferred, usually from sequence data using phylogenetic and/or phylogenomic methods which we do not discuss here.) Multiple trees which one can reasonably hypothesize were generated with the same parameter values are simply not available. If the tree is sufficiently large, researchers hope it provides enough information to infer the speciation and extinction parameters reasonably well. More precisely, it has been implicitly assumed that the inference is statistically consistent, in the sense that as the number of taxa increases toward infinity (i.e., the tree grows larger), the probability of inferring model parameters arbitrarily close to the generating ones approaches 1. Establishing such a result, however, requires showing identifiability of the model parameters: A distribution derived from an observation of a single tree has a limit, as the number of taxa approaches infinity, that uniquely determines all parameter values.

Of course a full proof of the statistical consistency of a particular estimator requires additional arguments. For instance, the standard results on the consistency of maximum likelihood assume the availability of multiple independent samples, and therefore cannot be applied. Leroux's result (Leroux, 1992) on the consistency of maximum likelihood inference from a single sequence of observation from a Hidden Markov Model is analogous to what is need for applications of these diversification models. Nonetheless, establishing parameter identifiability is the first step toward this goal.

Recent work has shown that the first type of diversification model, with time-dependent rates, does not in fact have identifiable parameters (Louca and Pennell, 2020), calling into question the conclusions of many empirical studies.

This non-identifiability result, which holds even if one allows for identification to be based on arbitrarily many independent tree observations with the same underlying rate parameters, was compellingly illustrated by construction of examples of wildly different rate functions producing identical tree distributions. An instance of this lack of identifiability had in fact appeared earlier, in an argument in which speciation rates were modified and extinction rates set to zero without changing the model distribution (Nee et al., 1994).

Little work, however, has addressed identifiability questions for multitype birth-death tree models. The strongest results on parameter identifiability for a pure birth model focus on a tree's topological features but assume the types of both leaf nodes and their parents are observed (Popovic and Rivas, 2016). In biological applications, however, the type of a leaf of the tree may be observable, but the type of the parent nodes is virtually never known. Thus no identifiability result relevant to typical data analyses has been produced. A recent paper of O'Meara and Beaulieu (2021), which broadly discusses current issues with diversification models in evolutionary biology in light of the Louca and Pennell (2020) result, argues that multitype birth-death tree models are likely to be identifiable — provided their rates are time-independent — but is careful to indicate this has not yet been established. And as the community has seen for time-dependent models, formal mathematical analysis is essential to settle the question.

One might hope that the analysis of multitype birth-death tree models would be simpler than for a time-dependent rate model, as its parameter space is finite dimensional. On the other hand, while trees produced by the time-dependent rate models can be summarized by the counts of lineages through time with no loss, this is not true for the multitype models, where the full tree structure carries additional information. Effectively extracting information from a tree

with both topological and metric structure requires a new approach.

In this paper, we investigate parameter identifiability of the *multiype pure-birth tree (MPBT) model* with any finite number of types. We thus restrict extinction rates for all classes to be zero. This model has also been called the multitype Yule model (Popovic and Rivas, 2016). We assume only that the metric tree is observable, with no information on the types either at points internal to the tree or at the leaves. More formally, we establish *generic identifiability of parameters up to label swapping*. "Generic" means the result holds if we exclude parameters lying in a measure-zero subset of the parameter space. We give an explicit characterization of such a measure-zero exceptional set, as the zero set of a certain polynomial. "Up to label swapping" means that there are certain symmetries of the parameter space, arising from interchanging types so that their corresponding speciation and switching rates are also interchanged, that have no effect on the model's behavior. Generic identifiability up to label swapping is often the strongest form of identifiability that holds in models with hidden variables (Allman et al., 2009), and since we treat the types as unobservable, its appearance here is not surprising.

Our explicit generic conditions are stated as four assumptions throughout the paper, as need for each arises for specific arguments. Briefly, they are that speciation rates for all types are positive and distinct (Assumptions 1 and 4), all switching rates between types are positive (Assumption 2), and that a certain matrix with entries in the speciation and switching rates is nonsingular (Assumption 3). The first few of these are intuitive and plausible assumptions. Although the meaning of the last condition is less clear outside the setting of the formal mathematical proof, we illustrate that in a few special cases it also imposes a natural condition.

Our arguments draw on several earlier studies. The first is the work of

Athreya (1968) on Multitype Continuous Time Markov Branching Processes. In fact, these models and the MPBT model have the same underlying structure. But much of the classical branching process literature allows only for observing type counts over time, and not for observing the tree structure indicating the branching of specific lineages. The MPBT model, in contrast, treats the tree structure as observable, with type information hidden. Thus while providing an important tool in this work, the results of Athreya are not immediately applicable to the MPBT model.

The second result crucial to our work is a general theorem on identifiability up to label swapping of parameters of a mixture model of product distributions (Allman et al., 2009). In applying this to the MPBT model, we consider the joint distribution of edge lengths around a node on a uniformly-at-random chosen edge of a random tree, as the random tree grows arbitrarily large. Due to conditional independence of edge lengths, conditioned on the type of the shared node, this joint distribution takes the form of a mixture distribution (over types) of product distributions. Although additional work is necessary to show parameter identifiability, this theorem is a crucial ingredient in our argument.

Although we do not address the multitype birth-death tree model with nonzero extinction rates here, we believe that our approach provides a pathway toward a more general result.

Some applications of multitype birth-death models also attempt to choose an appropriate number of types based on the data, with several Bayesian software packages supporting this, (e.g. Rabosky, 2014; Barido-Sottani et al., 2020). While this is an important element of some data analyses, it is not addressed in this work, where we fix the number of types. Choosing the number of types amounts to choosing among a family of nested models, each with generically identifiable parameters, where one may expect any finite data set to be naively

better fit with each increase in the number of types. While in the theoretical world of exact distributions one could choose the smallest number of types giving an exact fit, the finiteness of data necessitates the use of more sophisticated approaches to model adequacy.

This paper is structured as follows. In Section 2 we provide a more formal definition of the MPBT model, and begin its analysis by deriving formulas related to the generation of a single edge in the tree in Section 3. Section 4 uses the results of Athreya (1968) to obtain asymptotic results on the distribution of types across lineages in the tree at times increasingly distant from the root of the tree. Then, in Section 5, we bring these ingredients together, and apply the theorem of Allman et al. (2009) to obtain our main results. Concluding remarks appear in Section 6.

2 Model definition

In this section we formalize the Multitype Pure-Birth Tree model, in a form useful for our analysis.

Let m be a positive integer denoting the number of types, and denote the set of types by $[m] = \{1, 2, \ldots, m\}$.

The parameter space of the MBDT model with m types is all 3-tuples $(\boldsymbol{\pi}, \boldsymbol{\lambda}, S)$ described as follows:

A *root distribution* $\boldsymbol{\pi} = (\pi_1, \pi_2, \ldots, \pi_m)$, with $\pi_i \geq 0$, $\sum_i \pi_i = 1$ gives probabilities π_i of type i being chosen for the tree root. A vector $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \ldots, \lambda_m)$ with non-negative entries gives *speciation rates* λ_i for type i. An $m \times m$ matrix $S = (s_{ij})$ with non-negative off-diagonal entries and rows summing to 0 gives scalar *type switching rates* s_{ij} from type i to type j, $i \neq j$. Note that S is determined by the $m^2 - m$ independent scalar switching rates.

2.1 The edge process model

We first describe how an edge of a tree is produced under the model. As edges of the tree are produced independently conditioned on their starting types, a description of a single edge is sufficient.

We view an edge as growing with time, randomly changing the type of its leading point as it does so. At any time the edge may speciate, at a rate λ_i determined by its current type i. When speciation occurs, the edge ceases to grow, and in the full model two new edge processes are started for its descendent edges. However, in formalizing the edge process we describe the speciation of an edge as the process entering an absorbing state, for mathematical convenience.

For each type $i \in [m]$, define two states i_-, i_+. At any time, state i_- indicates that the current leading point of the edge has type i and that the edge has not yet speciated. The absorbing state i_+ represents that a speciation has occurred and at the time of speciation the leading point had type i. The parameter s_{ij}, $i \neq j$, is thus a rate of change from state i_- to state j_-, while λ_i is the rate of change from state i_- to i_+. No other instantaneous state changes are allowed.

Definition 1. The *m-type pure-birth edge process* $E_\tau = E_\tau(\widetilde{\boldsymbol{\pi}}, \boldsymbol{\lambda}, S)$ with $\widetilde{\pi}_i \geq 0$, $\sum_i \widetilde{\pi}_i = 1$, is the $2m$-state continuous-time Markov process over $\tau \in [0, \infty)$ with states

$$1_-, 2_-, \ldots, m_-, 1_+, 2_+, \ldots, m_+,$$

initial state distribution $(\widetilde{\boldsymbol{\pi}}, \mathbf{0}) \in \mathbb{R}^{2m}$, and $2m \times 2m$ transition rate matrix

$$Q := \begin{bmatrix} S - \operatorname{diag}(\boldsymbol{\lambda}) & \operatorname{diag}(\boldsymbol{\lambda}) \\ \mathbf{0} & \mathbf{0} \end{bmatrix},$$

where the rows and columns of Q are ordered by states as above. Here $\mathbf{0}$ is a vector or matrix of 0s, and $\operatorname{diag}(\boldsymbol{\lambda})$ is the diagonal matrix formed from vector

$\boldsymbol{\lambda}$.

The transition probability matrix associated to E_τ is

$$P(\tau) = \exp(Q\tau),$$

with $P_{ij}(\tau)$ giving the probability that an edge is in state j at time τ given that it was in state i at time 0.

Definition 2. The *speciation time* $\mathcal{T}$ associated to $E_\tau(\widetilde{\boldsymbol{\pi}}, \boldsymbol{\lambda}, S)$ is the $[0, \infty]$-valued random variable

$$\mathcal{T} = \inf\left(\{\tau \geq 0 \mid E_\tau \in \{1_+, 2_+, \ldots, m_+\}\} \cup \infty\right).$$

A realization of the edge process that reaches a "+" state is viewed as an edge of length $\mathcal{T}$, the time at which a speciation occurs. Each point (time τ) along the edge is "colored" by type i if the process is in state i_- (or state i_+ at its endpoint) at that time. Under mild assumptions, the edge length is finite with probability 1, as is shown below. Although for the MPBT model colors on edges are ultimately hidden, they play an important role in our arguments.

The terminal edges of the tree are produced by terminating edge processes at a specific time, before they may have reached an absorbing state. Formally defining such a *truncated edge process* and the colored edge it produces, is straightforward.

Due to the time-homogeneous Markov formulation of the edge process, we may equivalently produce an edge either from a single process reaching a "+" state, or by starting the process, truncating it before it enters a "+" state, starting a new process in the final state of the truncated one, and then conjoining the edges produced. Likewise, to produce an edge from the truncated process, we may allow the process to continue to a later time, and then truncate the

edge that was produced to an initial segment.

2.2 The multitype pure-birth tree model

We now define the MPBT model, as a generative model producing a tree. Let $T > 0$ be the depth (length of all paths from root to any tip) of the tree to be sampled.

1. The process begins with a root node. With parameters $(\boldsymbol{\pi}, \boldsymbol{\lambda}, S)$, generate from an edge process a colored descendent edge from the root to a node of type i, the only current tip of the tree.

   If the length of this edge is $\geq T$, truncate it to length T, and go to Step 4.

   Otherwise, at this node attach two descendent edges of length 0, with points on them colored by i. The tree now has 2 tips.

2. If the tree currently has k tips, for each tip generate a descendent edge via independent edge processes with parameters $(\mathbf{e}_i, \boldsymbol{\lambda}, S)$, where i is the type of the tip and $\mathbf{e}_i$ the standard basis vector in $\mathbb{R}^m$. Truncate all edge processes at the time τ when the first reaches a "+" state. The colored edges for each tip are conjoined to the edges (possibly of length 0) leading to the tip.

   If the path length from the root to a tip of the tree is $\geq T$, truncate all terminal edges so that all paths from root to leaves have length T, and go to Step 4.

   Otherwise, at the tip that arose from reaching state j_+, we attach two descendent edges of length 0 with points on them colored by j.

3. Go to step 2.

4. Uncolor all edges to obtain a sampled tree.

An example simulation of a colored tree from a binary-type model is shown in Figure 1, with the color hidden in Figure 2.

Remark. Inherent in the model are several notions of time. For an individual edge process, τ is a time variable, with $\tau = 0$ at the parental node in the edge. For the tree generation process overall, we use t as the time variable, with $t = 0$ at the root. If the edge process starting at the root enters a "+" state at time $\tau = \mathcal{T}_0$, then that root edge has length $\ell = \mathcal{T}_0$ and at its child node $t = \mathcal{T}_0$. Then if the edge process for an edge descending from the first speciation produces an edge of length $\mathcal{T}_1$, then at its child node $t = \mathcal{T}_0 + \mathcal{T}_1$. In general, a point on any edge e at time τ has

$$t = \tau + \sum_{\tilde{e} \text{ above } e} \mathcal{T}_{\tilde{e}}.$$

We can thus view a random tree as growing with time t, as its terminal edges lengthen while changing type, and speciate.

Remark. While we have defined the MPBT model as starting with a single edge descending from the root node, it is equally common to define diversification model starting at a bifurcating root. The modifications to the definition that are necessary to do so are straightforward, and working in that context would have no substantive impact on the arguments which follow.

Remark. Even if $T \to \infty$, a single observed tree does not allow for the identification of $\boldsymbol{\pi}$, so we focus on identifying the pair $(\boldsymbol{\lambda}, S)$. This factor of the parameter space can be identified with the non-negative orthant of $\mathbb{R}^{m^2}$.

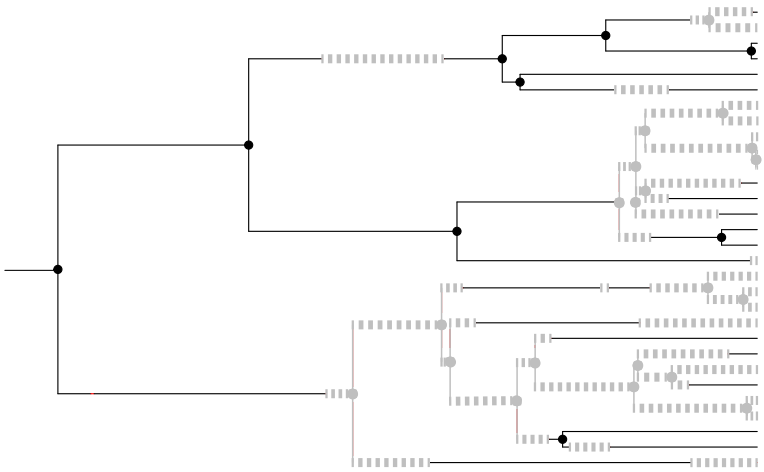

Figure 1: A finite-length colored tree generated by the binary-type pure birth tree model, before colors are hidden. Here black represents type 1 and red type 2, with $\lambda_1 = 0.1$, $\lambda_2 = 0.5$, $s_{12} = 0.1$, $s_{21} = 0.2$. Only the uncolored tree is observed.

3 The edge process

For parameters $(\boldsymbol{\lambda}, S)$, let $D = \text{diag}(\boldsymbol{\lambda})$ and $U = S - D$, so that the edge process E_τ has Markov rate matrix

$$Q = \begin{bmatrix} U & D \\ \mathbf{0} & \mathbf{0} \end{bmatrix}.$$

Lemma 1. *The transition probability matrix for E_τ is*

$$P(\tau) = \begin{bmatrix} \exp(U\tau) & f(U\tau)D\tau \\ \mathbf{0} & I \end{bmatrix},$$

where $f(A) = \sum_{n=0}^{\infty} \frac{1}{(n+1)!} A^n$ satisfies $f(A)A = \exp(A) - I$.

Proof. For $n \geq 1$

$$Q^n = \begin{bmatrix} U^n & U^{n-1}D \\ \mathbf{0} & \mathbf{0} \end{bmatrix},$$

so

$$P(\tau) = I + \sum_{n=1}^{\infty} \frac{1}{n!} \begin{bmatrix} U^n & U^{n-1}D \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \tau^n = \begin{bmatrix} \sum_{n=0}^{\infty} \frac{1}{n!} U^n \tau^n & \sum_{n=1}^{\infty} \frac{1}{n!} U^{n-1} D \tau^n \\ \mathbf{0} & I \end{bmatrix}.$$

□

For technical reasons we impose the following assumption, which is also biologically plausible.

Assumption 1. The speciation rates λ_i are positive for all i.

Lemma 2. *Let $(\boldsymbol{\lambda}, S)$ be parameters for a MPBT edge process satisfying Assumption 1. Then U is non-singular and all eigenvalues of U have negative real part.*

Proof. The assumption implies that U is strictly diagonally dominant, that is, the absolute value of each diagonal entry is strictly greater than the sum of the absolute values of all other entries in its row. Thus U is non-singular (Horn and Johnson, 2012). Since the diagonal entries are also negative, by the Gershgorin Circle Theorem every eigenvalue of U will have negative real part. □

Proposition 3. *Let F_i denote the cdf of the speciation time $\mathcal{T}$ conditioned on $E_0 = i_-$, and $\mathbf{1}$ be the vector of* 1*s. Then F_i is given by the i-th entry of*

$$\mathbf{1} - \exp(U\tau)\mathbf{1}.$$

Moreover, under Assumption 1, $\mathcal{T}$ is finite with probability 1.

Proof. Since $\mathcal{T}$ is the time E_τ first enters any of the absorbing states j_+, F_i is the sum across the i_- row of the upper right $m \times m$ block of $P(\tau)$. From Lemma

1, using that $D\mathbf{1} = -U\mathbf{1}$, the column vector of the F_is is therefore given by

$$f(U\tau)D\tau\mathbf{1} = -f(U\tau)U\tau\mathbf{1} = \mathbf{1} - \exp(U\tau)\mathbf{1}.$$

Under Assumption 1, by Lemma 2 the eigenvalues of U have negative real parts, so $\lim_{\tau\to\infty}\exp(U\tau) = \mathbf{0}$. Thus $\lim_{\tau\to\infty} F_i(\tau) = 1$ for each i, implying that $\mathcal{T}$ is finite with probability 1. □

Proposition 4. *Let $P_{i_-,j_+} = \lim_{\tau\to\infty} P_{i_-,j_+}(\tau)$ denote the asymptotic probability of transition to j_+ conditioned on $E_0 = i_-$. Then under Assumption 1, P_{i_-,j_+} is the (i,j)-entry of $-U^{-1}D$.*

Proof. The matrix $P_{-,+}(\tau)$ with entries $P_{i_-,j_+}(\tau)$ is the upper right $m \times m$ block of $P(\tau)$, so by Lemma 1,

$$P_{-,+}(\tau) = f(U\tau)D\tau = (\exp(U\tau) - I)U^{-1}D,$$

using that U is non-singular by Lemma 2. But $\lim_{\tau\to\infty}\exp(U\tau) = \mathbf{0}$ because U's eigenvalues have negative real parts. Thus

$$P_{-,+} = \lim_{\tau\to\infty}(\exp(U\tau) - I)U^{-1}D = (\mathbf{0} - I)U^{-1}D = -U^{-1}D.$$

□

4 Type Counting Process

Another ingredient of our approach to establishing the identifiability of MPBT model parameters is an analysis of an associated classical branching process, in which only the type counts are observed. More specifically, it records the number of edges of the tree which have each type as a function of time, but retains no

information on the topology of the tree. We call this the *type counting process*, and in this section use established results to determine the asymptotic behavior of the relative frequencies of each type.

Definition 5. For $i \in [m]$, let N_t^i denote the number of edges in a colored random tree arising from the colored MPBT model that exist at time t and are of type i at that moment. The *type counting process* N_t is the $(\mathbb{Z}^{\geq 0})^m$-valued continuous-time stochastic process over $[0, \infty)$ defined by $N_t := (N_t^1, N_t^2, \ldots, N_t^m)$. The *relative frequency process* is $R_t = N_t / (\sum_{i=1}^m N_t^i)$, provided the denominator is non-zero.

The asymptotics of the relative frequencies follow from results of Athreya (1968) on *multitype continuous-time Markov branching processes*, specifically Theorems 1 and 2 of that work, which are paraphrased below as Theorem 7. Such a model can be described as a process where individuals of type i live an exponentially-distributed length of time (whose rate only depends on type) and on death may be replaced by individuals of any type according to a distribution over $(\mathbb{Z}^{\geq 0})^m$.

To place the type counting process of the MPBT model into this framework, both speciation and change in type are viewed as deaths. Speciation results in replacement by 2 individuals of the same type, and change in type results in replacement by an individual of a different type. Since a speciation "death" of a type i individual occurs with rate λ_i, and a type change "death" of a type i individual followed by replacement with type j occurs with rate s_{ij}, the combined rate of death for type i is $\lambda_i + \sum_{j \neq i} s_{ij}$. When a death occurs, it is a speciation with probability

$$\frac{\lambda_i}{\lambda_i + \sum_{j \neq i} s_{ij}},$$

and a change to type j with probability

$$\frac{s_{ij}}{\lambda_i + \sum_{j \neq i} s_{ij}}$$

Basic properties of the type counting process are summarized in the following.

Lemma 3. *The type counting process N_t of the MPBT model is a strong Markov, continuous-time, m-type branching process, where each type i death has an offspring distribution defined by the multivariable probability generating function*

$$h_i(x_1, x_2, \ldots, x_m) = \frac{\lambda_i}{\lambda_i + \sum_{j \neq i} s_{ij}} x_i^2 + \sum_{j \neq i} \frac{s_{ij}}{\lambda_i + \sum_{j \neq i} s_{ij}} x_j.$$

We introduce yet another matrix defined in terms of the MPBT model parameters, as its leading eigenvalue and corresponding eigenvector plays a large role in the counting process's behavior.

Definition 6. Given parameters $(\boldsymbol{\lambda}, S)$ of the MPBT model, let

$$A = S + D.$$

A *leading eigenvalue of A* is an eigenvalue, ω, with the largest real part, and a *normalized leading left eigenvector* of A, is a left eigenvector for ω with $\sum_i u_i = 1$.

The matrix A is the infinitesimal generator of the conditional expectation of the N_is. More precisely,

$$\exp(At) = M_t = (m_{ij}(t))$$

with

$$m_{ij}(t) = \mathbb{E}\left[N_t^j | N_0 = \mathbf{e}_i\right],$$

where $\mathbf{e}_i$ is the i-th standard basis vector.

We will shortly show ω and $\mathbf{u}$ are uniquely determined, under an additional assumption.

Assumption 2. The off-diagonal entries of S are positive, i.e., $s_{ij} > 0$ for $i \neq j$.

Lemma 4. *For parameters $(\boldsymbol{\lambda}, S)$ of the MPBT model satisfying Assumption 2,*

1. *$M_t = \exp(At)$ has positive entries for $t > 0$.*
2. *A has a unique leading eigenvalue ω, which is both simple and real. Moreover the corresponding normalized left eigenvector $\mathbf{u}$ can be chosen to have all positive components.*

Proof. Fix $t > 0$. Then, using Assumption 2, A has positive off-diagonal entries, so there is a real k such that $B = At + kI$ has positive entries. Since B, kI commute, it follows that $e^{At} = e^{B-kI} = e^{-k}e^B$. Since B has positive entries, e^B does as well. Thus, e^{At} has positive entries.

The Perron-Frobenius Theorem applied to B shows it has a unique dominant (i.e., of maximal absolute value) eigenvalue ω which is also positive and simple, with a unique normalized left eigenvector $\mathbf{u}$ whose components are all positive. Since A has the same eigenvectors, and eigenvalues shifted by $-k$ and scaled by $1/t$, the second claim follows. □

Key properties of the counting process follow from the following more general theorem on classical branching processes.

Theorem 7. *(Athreya, 1968) Let X_t be a strong Markov, continuous-time, m-type branching process over $[0, \infty)$ which takes values in $(\mathbb{Z}^{\geq 0})^k$. Let $M_t =$*

$\exp(At)$ *be the conditional expectation matrix. Let* $h_i(x_1, ..., x_k)$ *be the offspring probability generating function for type* i*.*

If M_{t_0} *has positive entries for some* $t_0 > 0$*, and* $h_i(s)$ *is of degree* > 1 *for all* i*, then as* $t \to \infty$*,*

$$X_t e^{-\omega t} \xrightarrow{a.s.} W\mathbf{u},$$

where W *is a non-negative random variable,* ω *is the leading eigenvalue of* A*, and* $\mathbf{u}$ *is the positive normalized left eigenvector of* A *associated with* ω*.*

Moreover, if $\boldsymbol{\xi}^i = (\xi^i_j)$ *are random variables with generating functions* h_i*, then*

$$\mathbb{E}\left[\xi^i_j \log(\xi^i_j)\right] < \infty \tag{1}$$

for all i, j *if and only if for all* i

$$\mathbb{P}(W = 0 \mid X_0 = \mathbf{e}_i) = \mathbb{P}(X_t = 0 \text{ for some } t \mid X_0 = \mathbf{e}_i).$$

Corollary 8. *Consider the counting process associated to the MPBT model for parameters* $(\boldsymbol{\pi}, \boldsymbol{\lambda}, S)$*. Then under Assumptions 1 and 2,* $\sum N^i_t$ *is non-zero and as* $t \to \infty$*,*

$$R_t \xrightarrow{a.s.} \mathbf{u},$$

where u *is the positive normalized leading left eigenvector of* A*.*

Proof. Using the assumptions and Lemmas 3 and 4, the hypotheses of Theorem 7 are met, including inequality (1). Thus

$$N_t e^{-\omega t} \xrightarrow{a.s.} W\mathbf{u},$$

where ω is the leading eigenvalue of A, $\mathbf{u}$ is its positive normalized left eigenvector, and W is a non-negative random variable.

Since the random variable $\sum N_t^i$ is non-decreasing, the probability of extinction is zero:

$$\mathbb{P}(N_t = 0 \text{ for some } t \mid N_0 = \mathbf{e}_i) = 0.$$

Thus we find $\mathbb{P}(W = 0 \mid X_0 = \mathbf{e}_i) = 0$, implying $\mathbb{P}(W = 0) = 0$ regardless of $\boldsymbol{\pi}$. Then by the continuous mapping theorem,

$$R_t^i = \frac{N_t^i}{\sum_i N_t^i} = \frac{N_t^i e^{-\omega t}}{\sum_i N_t^i e^{-\omega t}} \xrightarrow{a.s.} \frac{W u_i}{W} = u_i$$

for each i. □

Remark. In studying diversification models with a single type but time-dependent rates of speciation and extinction, it is common to consider the random function giving the the number of lineages through time in a tree. This loses no information on parameters from the full tree, as each change in its value (speciation or extinction) is equally likely to have occurred on any lineage, and the growth of this function is thus highly informative on parameter values. For the multitype pure-birth model, however, the function $\sum_i N_t^i$ should not capture all information in the tree, as speciation may not be equally likely on all lineages. Corollary 8 indicates its growth is determined only by ω, the largest eigenvalue of A.

5 Identifiability of the MPBT model

Using the distributions of edge lengths and relative frequencies of each type of edge in a tree at a given time found in Sections 3 and 4, we are ready to establish identifiability of the MPBT parameters. To do so, we consider an asymptotic joint distribution of the lengths of 3 edges around a common node in the tree (see Figure 2). We seek to show that from this distribution the model

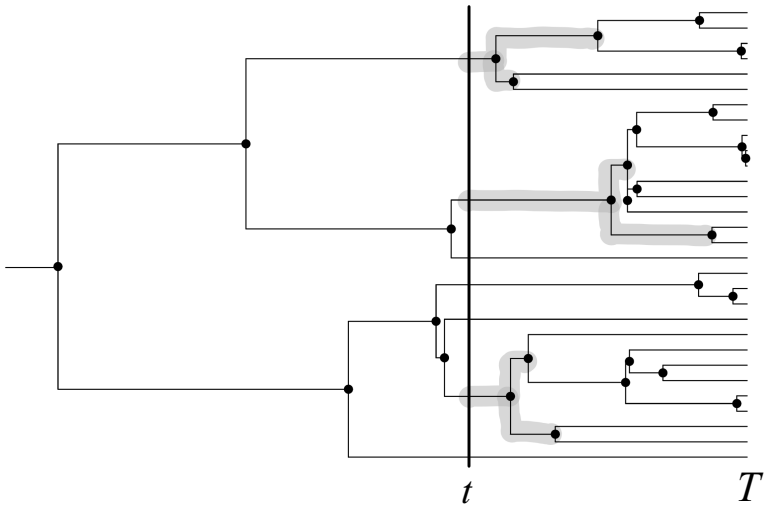

Figure 2: The uncolored tree of Figure 1, of depth T, generated by the binary-type pure birth model. The blue line at t determines several highlighted triples of edges whose lengths are possible draws from the probability distribution G_t of Definition 10 of Section 5.

parameters $(\boldsymbol{\lambda}, S)$ can be determined, up to label swapping.

Due to the conditional independence of the lengths of three edges sharing a common node, given that node's type, this distribution is a mixture of product distribution, with the mixing distribution and the components of the products closely related to distributions previously computed. This structure allows for the application of the following theorem, to obtain unmixed distributions of edge lengths conditioned on the type of the parental node. Thus even though we have no observation of type at any point in the tree, we can extract a distribution that is conditioned on type.

The following is a variant of Theorem 8 of Allman et al. (2009), with the hypotheses modified as discussed on p. 3116 of that paper.

Theorem 9. *(Allman et al., 2009) For* $1 \leq i \leq m$*, let*

$$\mu_i = \prod_{k=1}^{3} \mu_i^k$$

be a product of 3 *independent, absolutely continuous distributions* μ_i^k *on* $\mathbb{R}$*. With* $\pi_i > 0$*, let* $(\pi_1, \pi_2, \ldots, \pi_m)$ *be a distribution on* $[m]$*. For each* k*, suppose the set of distributions* $\{\mu_i^k\}_{i=1}^m$ *has the property that every subset of* r_k *elements is linearly independent, and that*

$$r_1 + r_2 + r_3 \geq 2m + 2.$$

Then, up to label swapping in i*, the* μ_i^k *and* π_i *are determined by the mixture distribution*

$$P = \sum_{i=1}^{m} \pi_i \mu_i = \sum_{i=1}^{m} \pi_i \prod_{k=1}^{3} \mu_i^k.$$

More precisely, P *determines distributions* ν_i^k *and* $(p_1, p_2, \ldots, p_m)$ *such that for some permutation* σ *of the set* $[m]$*,*

$$\mu_i^k = \nu_{\sigma(i)}^k \text{ and } \pi_i = p_{\sigma(i)}.$$

To apply this theorem, we make a further technical assumption, denoting the vector of 1s by $\mathbf{1}$.

Assumption 3. Parameters $(\boldsymbol{\lambda}, S)$ are such that the $m \times m$ matrix

$$M = M(\boldsymbol{\lambda}, S) = \begin{pmatrix} \mathbf{1} & U\mathbf{1} & U^2\mathbf{1} & \ldots & U^{m-1}\mathbf{1} \end{pmatrix}$$

is non-singular.

While the role of this assumption in our arguments will be clear in our proofs of Lemma 6 and Theorem 12 below, to understand its implications concretely,

consider first the case $m = 2$. Then

$$U = \begin{pmatrix} -s_{12} - \lambda_1 & s_{12} \\ s_{21} & -s_{21} - \lambda_2 \end{pmatrix},$$

so

$$M = \begin{pmatrix} 1 & -\lambda_1 \\ 1 & -\lambda_2 \end{pmatrix}.$$

The non-singularity of M thus is equivalent to $\lambda_1 \neq \lambda_2$. That these speciation rates would need to be different for parameters to be identifiable is intuitively clear, since otherwise type changes governed by S would have no impact on the structure of the uncolored tree.

For general m, Assumption 3 is equivalent to the non-vanishing of $\det M$, a degree $\sum_{i=1}^{m-1} i = \binom{m}{2}$ polynomial in the m^2 independent entries of $\boldsymbol{\lambda}, S$. Its non-vanishing thus excludes an algebraic variety of codimension 1, a set of Lebesque measure 0 in the unrestricted parameter space. An explicit calculation in the $m = 3$ case shows the polynomial to be an irreducible polynomial in the λ_i and s_{ij}, $i \neq j$.

The non-vanishing of $\det M$ always requires that the vector $\boldsymbol{\lambda} = -U\mathbf{1}$ not be a multiple of $\mathbf{1}$ (so that the first two columns of M are linearly independent), and hence that not all λ_i are the same. However, the additional restrictions it imposes on the parameters are more opaque to intuition without considering special cases.

For instance, when $m = 3$, if all the s_{ij} are equal, so the type switching behavior is identical for all types, the polynomial simplifies considerably, and factors as

$$(\lambda_1 - \lambda_2)(\lambda_2 - \lambda_3)(\lambda_3 - \lambda_1).$$

Non vanishing of the polynomial, then requires that the three λ_i be distinct, as

one would expect is needed for identifiability, for otherwise several types would behave identically. However, for other choices of the s_{ij}, two of the λ_i can be equal without the polynomial vanishing.

Next, we define the joint edge length distribution for several edges of a tree.

Definition 10. For some $t < T$, consider the following three random variables: Sample an (uncolored) tree of depth T under the MPBT model. From among the edges of the tree existing at time t choose one uniformly at random. Then with $t_b \in (t, T)$, the time at which that edge speciates, let $\ell_t^0 = t_b - t$ denote the time interval until it speciates, and let ℓ_t^1 and ℓ_t^2, respectively denote the lengths of the immediate descendent edges (where the edges are designated 1,2 uniformly at random). Then the joint distribution of these three variables ℓ_t^0, ℓ_t^1, and ℓ_t^2 is

$$G_{T,t}(\tau_0, \tau_1, \tau_2) := \mathbb{P}\left(\ell_t^0 \leq \tau_0, \ell_t^1 \leq \tau_1, \ell_t^2 \leq \tau_2 \mid \ell_t^1, \ell_t^2 < T - t - \ell_t^0\right).$$

We call $G_{T,t}$ the *joint distribution of edge lengths around a node.*

The three edge lengths used in the definition of $G_{T,t}$ are depicted in Figure 2, for $t = T/2$. The conditioning in the definition of $G_{T,t}$ ensures it only considers edges in which the edge process has led to speciation, that is, the edge processes for the parental and child edges are not truncated.

Lemma 5. *Under Assumptions 1 and 2, as $T \to \infty$, the joint distribution $G_{T,T/2}$ at time $T/2$ of edge lengths around a node on a tree of depth T converges to*

$$G_\infty = \sum_i \sum_j u_i P_{i_-,j_+}(\tau_0) F_j(\tau_1) F_j(\tau_2), \tag{2}$$

where F_j, P_{i_-,j_+}, and u_i are defined in Propositions 3, 4, and Lemma 4, respectively.

Proof. Note that the event E which is conditioned upon in the definition of $G_{T,T/2}$ excludes edge lengths resulting from truncated edge processes, so that all edge lengths under consideration are in fact speciation times $\mathcal{T}$. Thus

$$\begin{aligned}
&\lim_{T\to\infty} G_{T,T/2}(\tau_0,\tau_1,\tau_2) \\
&\qquad = \lim_{T\to\infty} \mathbb{P}(\mathcal{T}^0_{T/2}\le\tau_0, \mathcal{T}^1_{T/2}\le\tau_1, \mathcal{T}^2_{T/2}\le\tau_2 \mid \mathcal{T}^1_{T/2}, \mathcal{T}^2_{T/2} < T/2 - \mathcal{T}^0_{T/2}) \\
&\qquad = \lim_{T\to\infty} \Big(\mathbb{P}(\mathcal{T}^0_{T/2}\le\tau_0, \mathcal{T}^1_{T/2}\le\tau_1, \mathcal{T}^2_{T/2}\le\tau_2) + \epsilon_T(\tau_0,\tau_1,\tau_2)\Big),
\end{aligned}$$

where the function ϵ_T is the difference of the conditional and non-conditional probabilities above. But since the probability of $E \to 1$ as $T \to \infty$, it follows that $\epsilon_T \to 0$. We henceforth focus on $\mathbb{P}(\mathcal{T}^0_{T/2}\le\tau_0, \mathcal{T}^1_{T/2}\le\tau_1, \mathcal{T}^2_{T/2}\le\tau_2)$ rather than $G_{T,T/2}$.

Letting A_i denote the event that the uniformly-at-random chosen edge is of type i at time $\frac{T}{2}$ and B_j denote the event that that edge speciates in color j, and recalling that edge processes around a node are independent when conditioned on the type of that node, we have

$$\begin{aligned}
&\mathbb{P}(\mathcal{T}^0_{T/2}\le\tau_0, \mathcal{T}^1_{T/2}\le\tau_1, \mathcal{T}^2_{T/2}\le\tau_2) \\
&\qquad = \sum_i \sum_j \mathbb{P}(\mathcal{T}^0_{T/2}\le\tau_0, \mathcal{T}^1_{T/2}\le\tau_1, \mathcal{T}^2_{T/2}\le\tau_2 \mid A_i, B_j)\mathbb{P}(A_i, B_j) \\
&\qquad = \sum_i \sum_j \mathbb{P}(\mathcal{T}^0_{T/2}\le\tau_0, B_j \mid A_i)\mathbb{P}(\mathcal{T}^1_{T/2}\le\tau_1 \mid B_j)\mathbb{P}(\mathcal{T}^2_{T/2}\le\tau_2 \mid B_j)\mathbb{P}(A_i) \\
&\qquad = \sum_i \sum_j P_{i_-,j_+}(\tau_0)F_j(\tau_1)F_j(\tau_2)\mathbb{P}(A_i).
\end{aligned}$$

In this last expression, the only dependence on T is in $\mathbb{P}(A_i)$. But by Corollary 8, $\mathbb{P}(A_i) = \mathbb{E}[R^i_{T/2}] \to u_i$ as $T \to \infty$, yielding equation (2). □

Remark. While the specific time $T/2$ is used in this Lemma, our arguments

would be essentially unchanged if this were replaced by any function $f(T)$ with $f(T)$ and $T - f(T) \to \infty$ as $T \to \infty$.

This immediately gives that G_∞ is a finite mixture of product distributions.

Corollary 11. *The asymptotic joint distribution of edge lengths around a node, G_∞ can be expressed as a m-component mixture of products of 3 univariate distributions:*

$$G_\infty = \sum_{j=1}^{m} \pi_j \prod_{k=1}^{3} \mu_j^k,$$

where $\pi_j = \sum_i P_{i_-,j_+} u_i$, $\mu_j^1 = \frac{\sum_i P_{i_-,j_+}(\tau) u_i}{\sum_i P_{i_-,j_+} u_i}$, $\mu_j^2 = \mu_j^3 = F_j(\tau)$, and P_{i_-,j_+} is as defined in Proposition 4.

In order to apply Theorem 9 to G_∞, we need to verify that some of the univariate distributions in its decomposition above are linearly independent. To do so, the following lemma is needed.

We now introduce an additional assumption, which holds for generic parameters.

Assumption 4. The speciation parameters satisfy $\lambda_i \neq \lambda_j$ for all $i \neq j$.

Lemma 6. *Suppose Assumption 1,2, 3, and 4 hold, and consider the sets of univariate distributions $\{\mu_j^k\}_{j=1}^m$ defined in Corollary 11. For $k = 1$, every pair of functions in this set is linearly independent, while for $k = 2, 3$ the full set is linearly independent.*

Proof. Since $\{\mu_j^2\}_j = \{\mu_j^3\}_j$, we need only consider the cases $k = 1, 2$.

Consider first the case $k = 2$. Consider the vector F of functions $\mu_j^2 = F_j$. Then by Proposition 3,

$$F = \mathbf{1} - \exp(U\tau)\mathbf{1}.$$

Suppose $\mathbf{c}^T F = 0$ for some vector $\mathbf{c}$. Since $\frac{d^n}{d\tau^n} F(0) = -U^n \mathbf{1}$, it follows that $\mathbf{c}^T M = \mathbf{0}$ where M is defined in Assumption 3. Since M is non-singular, $\mathbf{c} = \mathbf{0}$,

so the entries of F are independent.

For $k = 1$, it is enough to show the independence of each pair of functions

$$\nu_j(\tau) = (\sum_i P_{i_-,j_+} u_i)\mu_j^1 = \sum_i P_{i_-,j_+}(\tau) u_i.$$

From Lemma 1 the vector G of all ν_j is given by

$$G(\tau)^T = \mathbf{u}^T \sum_{n=1}^{\infty} \frac{1}{n!} U^{n-1} D\tau^n.$$

Suppose $G(\tau)^T \mathbf{c} = 0$ for some vector $\mathbf{c}$. Since $\frac{d^n}{d\tau^n} G(0)^T = \mathbf{u}^T U^{n-1} D$, it follows that

$$\mathbf{u}^T U^{n-1} D\mathbf{c} = 0 \text{ for } n \geq 1.$$

In particular, for $n = 1$ we find $\mathbf{u}^T D\mathbf{c} = 0$. For $n = 2$, since $U = A - 2D$ and $\mathbf{u}^T A = \omega \mathbf{u}^T$, we have

$$\mathbf{u}^T U D\mathbf{c} = \mathbf{u}^T(\omega I - 2D)D\mathbf{c} = 0.$$

To show every pair of the ν_js is independent, consider $\mathbf{c}$ all of whose entries except possibly two are zero. Without loss of generality suppose the exceptions are c_1, c_2. Then the $n = 1, 2$ equations become

$$\begin{pmatrix} u_1\lambda_1 & u_2\lambda_2 \\ u_1(\omega - 2\lambda_1)\lambda_1 & u_2(\omega - 2\lambda_2)\lambda_2 \end{pmatrix} \begin{pmatrix} c_1 \\ c_2 \end{pmatrix} = \mathbf{0}$$

Using $u_1, u_2, \lambda_1, \lambda_2 > 0, \lambda_1 \neq \lambda_2$, computing the determinant of this matrix shows it is non-singular, and hence $c_1 = c_2 = 0$.

□

We now arrive at our main result.

Theorem 12. *Under the explicit generic Assumptions 1, 2 3, and 4, the parameters* $(\boldsymbol{\lambda}, S)$ *of the uncolored Multitype Pure-birth Tree model are identifiable up to label swapping from the asymptotic distribution* G_∞ *of edge lengths around a node.*

Proof. Suppose two parameter choices, $(\boldsymbol{\pi}, \boldsymbol{\lambda}, S)$ and $(\boldsymbol{\pi}^*, \boldsymbol{\lambda}^*, S^*)$, induce the same asymptotic distribution G_∞. Denoting the various distributions of conditional branching times, asymptotic transition probabilities, eigenvectors of matrices, etc. associated to parameters $(\boldsymbol{\pi}, \boldsymbol{\lambda}, S)$ as earlier in this work, we use the same notation with a "$*$" appended to denote the corresponding entities associated to parameters $(\boldsymbol{\pi}^*, \boldsymbol{\lambda}^*, S^*)$.

By Theorem 9, Corollary 11, and Lemma 6 the distributions π_i, μ_i^k, for $1 \leq i \leq m$, $1 \leq k \leq 3$ are determined from $G_\infty = G_\infty^*$, up to label swapping in i. Thus $F_i^*(\tau) = F_{\sigma(i)}(\tau)$ for some permutation σ.

Using Proposition 3 the equations $F_i^*(\tau) = F_{\sigma(i)}(\tau)$ for all j can be represented in matrix form as

$$\mathbf{1} - e^{U^*\tau}\mathbf{1} = \Sigma(\mathbf{1} - e^{U\tau}\mathbf{1}), \tag{3}$$

where Σ is the permutation matrix representing σ. Equating coefficients of the MacLauren series yields for $n = 1, 2, 3, \ldots$ that

$$(U^*)^n\mathbf{1} = \Sigma U^n \mathbf{1}. \tag{4}$$

Using equation (4) and the definition of M, M^* in Assumption 3 shows

$$M^* = \Sigma M. \tag{5}$$

Equation (4) further implies

$$\begin{aligned} U^*M^* &= \begin{pmatrix} U^*\mathbf{1} & (U^*)^2\mathbf{1} & (U^*)^3\mathbf{1} & \dots & (U^*)^m\mathbf{1} \end{pmatrix} \\ &= \begin{pmatrix} \Sigma U\mathbf{1} & \Sigma U^2\mathbf{1} & \Sigma U^3\mathbf{1} & \dots & \Sigma U^m\mathbf{1} \end{pmatrix} \\ &= \Sigma UM. \end{aligned}$$

Using equation (5) then yields

$$U^*\Sigma M = \Sigma UM,$$

and since M is non-singular,

$$U^*\Sigma = \Sigma U.$$

Since $U = S - D$ and each row of S adds to 0, multiplying the last equation by $\mathbf{1}$ on the right gives $\boldsymbol{\lambda}^* = \Sigma\boldsymbol{\lambda}$. Since this implies $D^*\Sigma = \Sigma D$, it follows that $S^* = \Sigma S\Sigma^T$ as well. Thus the parameters differ only up to label swapping. □

Remark. Theorem 12 establishes that an asymptotic distribution, as tree depth $\to \infty$ associated to the MPBT model yields parameter identifiability. This suggests that with a sample of many trees of arbitrarily large size, there is potential for statistically consistent inference, where "consistency" would mean as both the number of trees and the tree depth go to infinity. However, this is not the framework in which data analysis with this model is performed, since while a tree may be large, only one tree observation is available (Maddison et al., 2007).

Fortunately, a minor modification to the proofs above again yields identifiability of parameters from an asymptotic distribution derived from a single observation, as the depth of the tree goes to infinity. Indeed, modify Definition 10 so that G_t is the distribution of edge lengths around a node from single

growing tree. The proof of Lemma 5, then, is modified only in its last line, as $\mathbb{P}(A_i) = R^i_{T/2}$, a random variable rather than its expected value. Nonetheless, by Corollary 8, we again find $\mathbb{P}(A_i) \to u_i$, so the conclusion is unchanged.

6 Discussion

Theorem 12, and Remark 5, show that parameters $(\boldsymbol{\lambda}, S)$ of the MPBT model can be identified from an asymptotic distribution as the tree depth grows, whether or not the number of sampled trees grows. Although this is not sufficient to conclude that estimation of parameters by maximum likelihood (ML) from a single tree, as suggested by Maddison et al. (2007), is statistically consistent, it does at least indicate that is a possibility. A similar question on ML inference of parameters for a hidden Markov model from a single sequence of observations was addressed by Leroux (1992), with the consistency of ML estimation established as the sequence length goes to infinity.

For applications, it would be highly desirable to extend our identifiability result to a model incorporating constant extinction rates for each type. In most biological settings, the obtainable "data," however is not the tree with edges stopping at extinction events, but rather the pruned tree in which all edges with no extant descendants are removed.

For a single type, parameter identifiability of a model with pruning was essentially considered by Nee et al. (1994), where it was shown that the lineages-through-time function's rate of change allowed the speciation and extinction rates to be determined, by separately considering the time regimes much earlier than the tree tips, and near the tree tips. An analysis combining the insight from Nee et al. (1994) with the mixture distribution framework used in this work might be successful in showing parameters can be recovered from a single large tree observation for the multitype birth-death model.

We emphasize that our work here in no way suggests that a multitype model incorporating arbitrary time-dependence in its rates will have identifiable parameters. Indeed, the issues that Louca and Pennell (2020) raised are likely to only be compounded in such a setting, unless the time-dependence is restricted to some specific form. Results such as those of Legried and Terhorst (2022a,b) in the single-type case, which show identifiabiilty for piecewise constant and polynomial time-dependent rates, can be expected to generalize to more types.

Another interesting identifiability question for multitype tree models concerns what information on parameters is contained in the tree topology alone, or from weaker metric information than precise branch lengths. While our analysis depends heavily on metric features of the tree, that of Popovic and Rivas (2016) required no metric information. However, it did use type observations at the tips of the tree, and at their parental nodes. While types at tree tips may be observed in some biological studies, types of the parental nodes are generally not observable, as data is generally collected only from the taxa extant at the present. Even if ancient DNA or other trait data from earlier times is available, it is unlikely to be from the time of the last speciation.

Authors' Contributions

All authors contributed equally to this work,

Author Disclosure Statement

The authors declare they have no competing financial interests.

Funding Information

ESA and JAR were supported in part by NSF Grant DMS-2051760.

References

Allman, E. S., Matias, C., and Rhodes, J. A. Identifiability of parameters in latent structure models with many observed variables. *Ann. Statist.*, 37(6A): 3099–3132, 12 2009. doi: 10.1214/09-AOS689. URL https://doi.org/10.1214/09-AOS689.

Athreya, K. B. Some results on multitype continuous time markov branching processes. *The Annals of Mathematical Statistics*, 39(2):347 – 357, 1968. doi: 10.1214/aoms/1177698395. URL https://doi.org/10.1214/aoms/1177698395.

Barido-Sottani, J., Vaughan, T. G., and Stadler, T. A multi-type birth-death model for Bayesian inference of lineage-specific birth and death rates. *Syst. Biol.*, 2020. https://doi.org/10.1093/sysbio/syaa016.

Cantalapiedra, J., FitzJohn, R., Kuhn, T., Hernández Fernández, M., DeMiguel, D., Azanza, B., Morales, J., and A.Ø. Mooers. Dietary innovations spurred the diversification of ruminants during the Caenozoic. *Proc. Royal Society B: Biol. Sci.*, 281(1776):20132746, 2014.

FitzJohn, R. G. Diversitree: comparative phylogenetic analyses of diversification in R. *Methods in Ecology and Evolution*, 3:1084–1092, 2012.

Harvey, P. H., May, R. M., and Nee, S. Phylogenies without fossils. *Evolution*, 48(3):523–529, 1994. ISSN 00143820, 15585646. URL http://www.jstor.org/stable/2410466.

Horn, R. and Johnson, C. *Matrix Analysis*. Cambridge University Press, 2012. ISBN 9781139788885. URL https://books.google.com/books?id=O7sgAwAAQBAJ.

Kendall, D. G. 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.

Legried, B. and Terhorst, J. A class of identifiable phylogenetic birth & death models. *Proceedings of the National Academy of Sciences*, 119(35): e2119513119, 2022a. doi: 10.1073/pnas.2119513119.

Legried, B. and Terhorst, J. Identifiability and inference of phylogenetic birth-death models. https://www.biorxiv.org/content/10.1101/2022.08.26.505438v1, 2022b.

Leroux, B. G. Maximum-likelihood estimation for hidden Markov models. *Stochastic Processes and their Applications*, 40(1):127–143, 1992. ISSN 0304-4149. doi: https://doi.org/10.1016/0304-4149(92)90141-C. URL https://www.sciencedirect.com/science/article/pii/030441499290141C.

Louca, S. and Pennell, M. W. Extant timetrees are consistent with a myriad of diversification histories. *Nature*, 580(7804):502–505, Apr 2020. ISSN 1476-4687. doi: 10.1038/s41586-020-2176-1. URL https://doi.org/10.1038/s41586-020-2176-1.

Maddison, W. P., Midford, P. E., and Otto, S. P. Estimating a binary character's effect on speciation and extinction. *Systematic Biology*, 56(5):701–710, 10 2007. ISSN 1063-5157. doi: 10.1080/10635150701607033. URL https://doi.org/10.1080/10635150701607033.

Maliet, O., Hartig, F., and Morlon, H. A model with many small shifts for estimating species-specific diversification rates. *Nature Ecology & Evolution*, 3(7):1086–1092, 2019. doi: 10.1038/s41559-019-0908-0. URL https://doi.org/10.1038/s41559-019-0908-0.

Nee, S., May, R. M., and Harvey, P. H. The reconstructed evolutionary process. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 344(1309):305–311, 1994. doi: 10.1098/rstb.1994.0068. URL https://royalsocietypublishing.org/doi/abs/10.1098/rstb.1994.0068.

O'Meara, B. and Beaulieu, J. Potential survival of some, but not all, diversification methods. https://ecoevorxiv.org/w5nvd, 2021.

Popovic, L. and Rivas, M. Topology and inference for Yule trees with multiple states. *J. Math. Biol.*, 73(5):1251–1291, 2016. doi: 10.1007/s00285-016-0992-6.

Rabosky, D. L. Automatic detection of key innovations, rate shifts, and diversity-dependence on phylogenetic trees. *PLOS ONE*, 9(2):1–15, 02 2014. doi: 10.1371/journal.pone.0089543. URL https://doi.org/10.1371/journal.pone.0089543.

Rabosky, D. L. and Goldberg, E. E. Model inadequacy and mistaken inferences of trait-dependent speciation. *Systematic Biology*, 64(2):340–355, 01 2015. ISSN 1063-5157. doi: 10.1093/sysbio/syu131. URL https://doi.org/10.1093/sysbio/syu131.

Rasmussen, D. A. and Stadler, T. Coupling adaptive molecular evolution to phylodynamics using fitness-dependent birth-death models. *eLife*, 8:e45562, aug 2019. ISSN 2050-084X. doi: 10.7554/eLife.45562. URL https://doi.org/10.7554/eLife.45562.

Stadler, T. Species-specific diversification. *Nature Ecology & Evolution*, 3(7):1003–1004, 2019. doi: 10.1038/s41559-019-0923-1. URL https://doi.org/10.1038/s41559-019-0923-1.

Stadler, T. Recovering speciation and extinction dynamics based on phylogenies. *Journal of Evolutionary Biology*, 26(6):1203–1219, 2013. doi: https://doi.org/10.1111/jeb.12139. URL `https://onlinelibrary.wiley.com/doi/abs/10.1111/jeb.12139`.

Yule, G. U. A mathematical theory of evolution, based on the conclusions of Dr. J. C. Willis, F.R.S. *Philosophical Transactions of the Royal Society of London. Series B, Containing Papers of a Biological Character*, 213(402-410):21–87, 1925. doi: 10.1098/rstb.1925.0002. URL `https://royalsocietypublishing.org/doi/abs/10.1098/rstb.1925.0002`.