Theoretical Population Biology 148 (2022) 22–27

Contents lists available at ScienceDirect

Theoretical Population Biology

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

When can we reconstruct the ancestral state? A unified theory

Lam Si Tung Ho [a,*], Vu Dinh [b]

[a] Department of Mathematics and Statistics, Dalhousie University, Halifax, Nova Scotia, Canada
[b] Department of Mathematical Sciences, University of Delaware, USA

ARTICLE INFO

Article history:
Received 18 November 2021
Available online 24 September 2022

Keywords:
Ancestral state reconstruction
Consistency
Big bang condition
Phylogenetic comparative methods

ABSTRACT

Ancestral state reconstruction is one of the most important tasks in evolutionary biology. Conditions under which we can reliably reconstruct the ancestral state have been studied for both discrete and continuous traits. However, the connection between these results is unclear, and it seems that each model needs different conditions. In this work, we provide a unifying theory on the consistency of ancestral state reconstruction for various types of trait evolution models. Notably, we show that for a sequence of nested trees with bounded heights, the necessary and sufficient conditions for the existence of a consistent ancestral state reconstruction method under discrete models, the Brownian motion model, and the threshold model are equivalent. When tree heights are unbounded, we provide a simple counter-example to show that this equivalence is no longer valid.

1. Introduction

The evolution of biological features, such as genotypes and phenotypes, is often assumed to follow Markov processes along a phylogenetic tree (Edwards, 1970; Felsenstein, 2004). Under these models, each internal node in the tree depicts a speciation event when an ancestral lineage splits into two new ones. The descendant lineages inherit the ancestral state of their most recent common ancestor and then evolve independently from each other (Steel, 2016). One important task in evolutionary biology is reconstructing the ancestral state from observations at the leaves of a given tree. This problem, usually referred to as *ancestral state reconstruction* or *root reconstruction*, helps answer many questions about the underlying evolutionary process and directly affects the efficiency and accuracy of other phylogenetic estimates (Maddison, 1994; Liberles, 2007; Thornton, 2004; Ho and Susko, 2022). One important application is to infer the origin of epidemics (Faria et al., 2014; Gill et al., 2017).

In recent years, evolutionary data for a wide variety of species are increasingly available, and the problem of ancestral reconstruction based on thousands of leaves is becoming commonplace. It is well-known that sampled data at the leaves of the tree cannot be considered independent since closely related species are expected to have similar characteristics (Felsenstein, 1985). Previous works in the field indicate that in this setting, basic statistical properties should not be taken for granted (Ané, 2008; Li et al., 2008; Ho and Ané, 2013, 2014; Ané et al., 2017; Ho and Susko, 2022). For example, one of the most desired properties of good estimation methods is consistency (which dictates that the estimator converges to the true value as the number of leaves increases), but even rigorous methods such as Maximum likelihood estimator (MLE) could be inconsistent in phylogenetic settings. Characterizing conditions under which the ancestral state can be reliably estimated has become an active research direction.

Perhaps, Ané (2008) provides the most notable result for reconstructing the ancestral state of continuous traits. In this paper, the author derives a necessary and sufficient condition for the consistency of the MLE under the Brownian motion (BM) model. The condition involves the covariance matrix $\mathbf{V}_n$ whose components are the times of shared ancestry between leaves, that is, the element in ith-row and jth column, $[\mathbf{V}_n]_{ij}$, is the length shared by the paths from the root to the leaves i and j. Specifically, the MLE is consistent if and only if $\mathbf{1}^\top \mathbf{V}_n^{-1}\mathbf{1} \to \infty$. For discrete models, Fan and Roch (2018) show that under a certain root density assumption, referred to as the big bang condition, it is possible to identify a subset of leaves that are sufficiently independent. This enables the derivations of a necessary and sufficient condition for the existence of a consistent ancestral state reconstruction method under discrete models.

Despite the usefulness of these results, the connection between them is unclear. In principle, they consider different stochastic processes, study distinct aspects of the problem and focus on conditions with seemingly unrelated mathematical formulations. For example, Ané (2008) specifically studies the MLE, while Fan and Roch (2018) consider a more abstract question of the existence of a consistent estimator. It seems that the consistency property for each trait evolution model needs different conditions. In this work, we attempt to bridge this gap by showing that when the sequences of trees are nested and

* Corresponding author.
E-mail addresses: lam.ho@dal.ca (L.S.T. Ho), vucdinh@udel.edu (V. Dinh).

https://doi.org/10.1016/j.tpb.2022.09.001
0040-5809/

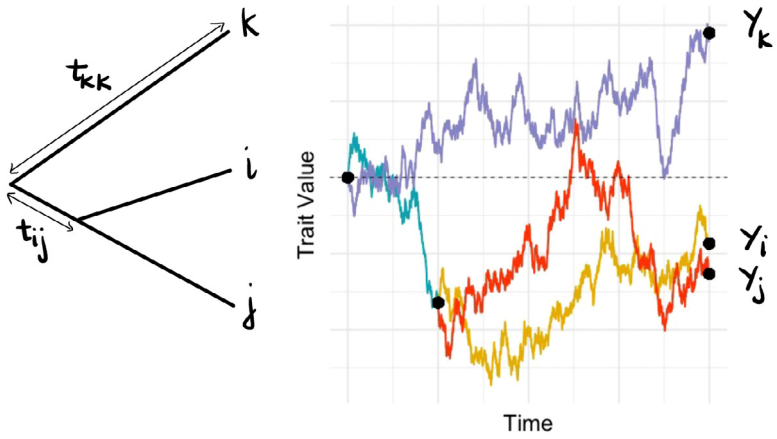

Fig. 1. Visualization of a BM process on a tree (right). The distance from the root to the most recent common ancestor of leaves i and j is t_{ij}, and the distance from the root to leaf k is t_{kk}.

have bounded heights, which corresponds to the natural setting where data from new species are continually being collected, the two geometric conditions of Ané (2008) and Fan and Roch (2018) are equivalent. As a consequence, they are the necessary and sufficient condition for the existence of a consistent ancestral state reconstruction method under the BM model and discrete models. We also show that the results extend to the threshold model (Felsenstein, 2012; Revell, 2014), thus providing a unifying perspective on the consistency of ancestral state reconstruction procedures across a wide range of popular phylogenetic Markov processes. Finally, we give a simple counter-example to show that when tree heights are unbounded, these conditions are not equivalent and neither of them is a sufficient condition for the existence of a consistent ancestral state reconstruction method under discrete models.

2. Settings

We consider a sequence of nested rooted trees $\mathbb{T}_n$, meaning $\mathbb{T}_{n-1}$ is a subtree of $\mathbb{T}_n$ for all n. It is worth noticing that this setting represents the situation when more species are continually sampled and added to the data set. This is a common setup for theoretical studies of trait evolution models (Fan and Roch, 2018; Ho and Ané, 2013). Without loss of generality, we assume that $\mathbb{T}_n$ has n species and the root of all trees is the same species. We denote the observed trait values at the leaves of $\mathbb{T}_n$ by $\mathbf{Y}_n = (Y_k)_{k=1}^n$. Furthermore, we assume that distances from this root to the leaves are uniformly bounded by H. The goal of ancestral state reconstruction is to estimate the trait value of this root from the trait values at the leaves.

In this paper, we study three different types of trait evolution models: BM, discrete, and threshold models. As we already discussed, all three (types of) models follow Markovian dynamics along phylogenetic trees where at each internal node, descendant lineages inherit the value from the parent lineage just prior to the speciation event. Conditional of their starting value, each lineage then evolves independently of the sister lineages. We focus on the consistency property of ancestral state reconstruction methods.

Definition 1. Let μ be the ancestral state. An estimator $\hat{\mu}_n$ constructing from n observations is consistent if and only if for any $\epsilon > 0$, we have

$$\lim_{n\to\infty} \mathbb{P}\left\{|\hat{\mu}_n - \mu| > \epsilon\right\} \to 0.$$

In other words, an estimator $\hat{\mu}_n$ is consistent if it converges (in probability) to the true ancestral state μ as the number of observations increases to infinity.

Brownian motion model. The BM model assumes that a continuous phenotype evolves along a tree according to a Brownian motion. Under the BM model, the observations $\mathbf{Y}_n = (Y_k)_{k=1}^n$ follow a Gaussian distribution $\mathcal{N}(\mu\mathbf{1}, \sigma^2\mathbf{V}_n)$. Here, μ is the ancestral state, σ^2 is the variance of the BM, and $\mathbf{V}_n = (t_{ij})$ depends on the tree where t_{ij} is the distance from the root to the most recent common ancestor of leaves i and j (Ané, 2008). We visualize the evolution of a trait along a tree under the BM model in Fig. 1.

Maximum likelihood estimator (MLE) is the most popular method for reconstructing the ancestral state. Under the BM model, the MLE has an analytic formula

$$\hat{\mu}_n = (\mathbf{1}^\top \mathbf{V}_n^{-1}\mathbf{1})^{-1}(\mathbf{1}^\top \mathbf{V}_n^{-1}\mathbf{Y}_n).$$

Ané (2008) provides the following necessary and sufficient condition for the consistency of the MLE:

Lemma 1 (*Ané (2008)*)**.** *Under the BM model, the MLE of the ancestral state is consistent if and only if* $\mathbf{1}^\top \mathbf{V}_n^{-1}\mathbf{1} \to +\infty$.

Discrete models. These models assume that traits evolve along the tree according to a continuous-time Markov chain with finite state–space. Fan and Roch (2018) focus on models that satisfy the "initial-state identifiability": all rows of the transition probability matrix $\mathbf{P}_t$ of the Markov chain are distinct for all t. Throughout the paper, we also require that for two states i, j (do not necessarily distinct), $P_{ij}(t) > 0$ for some $t > 0$. We refer to models satisfying those two conditions as *regular discrete models*. It is worth noticing that popular evolution models, such as two-state, Jukes–Cantor (Jukes and Cantor, 1969), and GTR (Lanave et al., 1984; Tavaré, 1986) (with positive transition rates) are regular discrete models. Fan and Roch (2018) derive a necessary and sufficient condition for the existence of a consistent estimator for the ancestral state, called the big bang condition. To understand the big bang condition, let us first introduce some notations. For a tree $\mathbb{T}$, a truncated tree at level s of $\mathbb{T}$, denoted by $\mathbb{T}(s)$, is the tree obtained by truncating $\mathbb{T}$ at distance s from the root. We denote the set of leaves of a tree $\mathbb{T}$ by $\partial\mathbb{T}$ and denote the cardinality of a set A by $|A|$.

Definition 2 (*Big Bang Condition*)**.** A sequence of nested trees $(\mathbb{T}_n)_{n=1}^\infty$ satisfies the big bang condition if for all $s > 0$, we have $|\partial\mathbb{T}_n(s)| \to \infty$ as $n \to \infty$.

Lemma 2 (*Fan and Roch (2018)*)**.** *Under regular discrete models, there exists a consistent estimator for the ancestral state if and only if the big bang condition holds.*

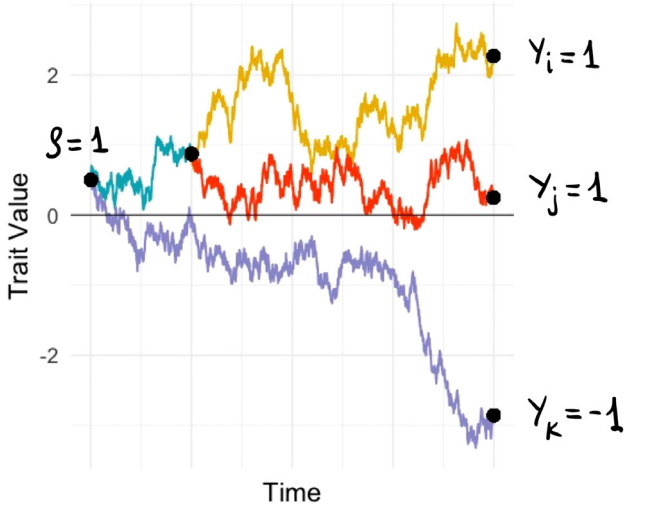

Fig. 2. Visualization of an underlying BM process and corresponding observations under the threshold model.

We note that the "downstream disjointness" condition in Fan and Roch (2018) is not satisfied for regular discrete models and can be removed.

Threshold model. Threshold model (Wright, 1934a,b; Felsenstein, 2005, 2012; Revell, 2014) assumes that a binary phenotype (± 1) is driven by an underlying process that evolves along a tree according to a BM. Let $\mathbf{Z}_n$ be the underlying process and $\mathbf{Y}_n$ be the observations at the leaves of the tree. Under threshold model, $\mathbf{Z}_n \sim \mathcal{N}(\mu, \sigma^2\mathbf{V}_n)$ and

$$Y_i = \begin{cases} 1 & Z_i \geq 0 \\ -1 & Z_i < 0. \end{cases}$$

Fig. 2 visualizes the evolution of the underlying BM process and the corresponding observations.

We want to estimate the ancestral state at the root $\rho = \text{sign}(\mu)$. To the best of our knowledge, there are no theoretical results for the problem of reconstructing the ancestral state under this threshold model.

3. Necessary and sufficient condition for consistency of ancestral state reconstruction

While the results of Lemmas 1 and 2 are very useful, the connection between them is unclear. The derived conditions of the two models focus on seemingly unrelated mathematical formulations, and it seems that the consistency property for each model needs to be studied separately. In this work, we aim to bridge this gap by showing that in our setting, the two geometric conditions for discrete and continuous models are equivalent. We then extend the results to threshold models to showcase the generalizability of the result across a wide range of popular phylogenetic Markov processes.

Theorem 1. *Under our settings (a sequence of nested trees with bounded heights),*

- *The big bang condition is equivalent with the condition* $\mathbf{1}^\top\mathbf{V}_n^{-1}\mathbf{1} \to +\infty$.
- *These conditions are the necessary and sufficient condition for the existence of a consistent estimator for the ancestral state under the BM, regular discrete, and threshold models.*

The flow of our proofs is as follows. First, we prove that the condition $\mathbf{1}^\top\mathbf{V}_n^{-1}\mathbf{1} \to +\infty$ is a necessary condition for the existence of a consistent estimator for the ancestral state under the BM model (Theorem 2). Together with Lemma 1, we conclude that $\mathbf{1}^\top\mathbf{V}_n^{-1}\mathbf{1} \to +\infty$ is also the necessary and sufficient condition. Next, we show that the big bang condition is also the necessary and sufficient condition for the existence of a consistent estimator for the ancestral state under the BM model (Theorem 3). Therefore, big bang condition and the condition $\mathbf{1}^\top\mathbf{V}_n^{-1}\mathbf{1} \to +\infty$ are equivalent. Finally, we prove that under the threshold model, the condition $\mathbf{1}^\top\mathbf{V}_n^{-1}\mathbf{1} \to +\infty$ is a necessary condition (Theorem 4) and the big bang condition is the sufficient condition (Theorem 5) for the existence of a consistent estimator for the ancestral state .

3.1. Equivalence of consistency condition for discrete and continuous models

Theorem 2. *Under the BM model, a necessary condition for the existence of a consistent estimator for the ancestral state is* $\mathbf{1}^\top\mathbf{V}_n^{-1}\mathbf{1} \to +\infty$.

Proof. We only need to prove that if there exists a constant $C > 0$ such that $\mathbf{1}^\top\mathbf{V}_n^{-1}\mathbf{1} \leq C$ for all n, then there is no consistent estimator for the ancestral state. Let P_{μ,σ^2} be the joint distribution of the observations $\mathbf{Y}_n$ under the BM model with mean μ and variance σ^2. We have

$$\text{KL}(P_{\mu_1,\sigma^2}, P_{\mu_2,\sigma^2}) = \frac{1}{2\sigma^2}\mathbf{1}^\top\mathbf{V}_n^{-1}\mathbf{1}(\mu_1-\mu_2)^2 \leq \frac{C}{2\sigma^2}(\mu_1-\mu_2)^2.$$

Here, $\text{KL}(P_{\mu_1,\sigma^2}, P_{\mu_2,\sigma^2})$ denotes the Kullback–Leibler divergence from P_{μ_2,σ^2} to P_{μ_1,σ^2}. Let $d_{\text{TV}}(P_{\mu_1,\sigma^2}, P_{\mu_2,\sigma^2})$ be the total variation distance between P_{μ_1,σ^2} and P_{μ_2,σ^2}. That is, $d_{\text{TV}}(P_{\mu_1,\sigma^2}, P_{\mu_2,\sigma^2}) = \sup_{\mathcal{A}} |P_{\mu_1,\sigma^2}(\mathcal{A}) - P_{\mu_2,\sigma^2}(\mathcal{A})|$. Applying Vajda's inequality (Vajda, 1970), we have

$$\text{KL}(P_{\mu_1,\sigma^2}, P_{\mu_2,\sigma^2}) \geq \log\left(\frac{1+d_{\text{TV}}(P_{\mu_1,\sigma^2}, P_{\mu_2,\sigma^2})}{1-d_{\text{TV}}(P_{\mu_1,\sigma^2}, P_{\mu_2,\sigma^2})}\right) - \frac{2d_{\text{TV}}(P_{\mu_1,\sigma^2}, P_{\mu_2,\sigma^2})}{d_{\text{TV}}(P_{\mu_1,\sigma^2}, P_{\mu_2,\sigma^2})+1}.$$

Since $\text{KL}(P_{\mu_1,\sigma^2}, P_{\mu_2,\sigma^2})$ is bounded from above, we deduce that $d_{\text{TV}}(P_{\mu_1,\sigma^2}, P_{\mu_2,\sigma^2}) \leq d_0 < 1$. Assume that there exists a consistent estimator $\hat{\mu}_n$ for the ancestral state. For a sufficiently small ϵ, we define $\mathcal{A}_\epsilon = \{|\hat{\mu}_n - \mu_1| \leq \epsilon\}$. By the definition of consistency, we have

$$P_{\mu_1,\sigma^2}(\mathcal{A}_\epsilon) \to 1 \quad \text{and} \quad P_{\mu_2,\sigma^2}(\mathcal{A}_\epsilon) \to 0,$$

where $\mu_1 \neq \mu_2$. This is a contradiction with the fact that $d_{\text{TV}}(P_{\mu_1,\sigma^2}, P_{\mu_2,\sigma^2}) \leq d_0 < 1$. Therefore, there is no consistent estimator for the ancestral state. □

Theorem 3. *Under the BM model, there exists a consistent estimator for the ancestral state if and only if the big bang condition holds.*

Proof. First, we will prove that if the big bang condition does not hold, then there exists a constant $C > 0$ such that $\mathbf{1}^\top\mathbf{V}_n^{-1}\mathbf{1} \leq C$. When the big bang condition does not hold, there exist $s > 0$, $K > 0$ and $N > 0$ such that $|\partial\mathbb{T}_n(s)| = K$ for all $n \geq N$. Let ϵ be the smallest distance from the root to the internal nodes and leaves of $\mathbb{T}_N(s)$. We note that, by this construction, $\mathbb{T}_n(\epsilon)$ is a fixed ultrametric star tree with height equal to ϵ for all $n \geq N$. Let $I_1, I_2, \ldots, I_\ell$ be the leaves of $\mathbb{T}_n(\epsilon)$ and $\mathbb{S}_1, \mathbb{S}_2, \ldots, \mathbb{S}_\ell$ be the subtree of $\mathbb{T}_n$ stemming from $I_1, I_2, \ldots, I_\ell$. Then,

$$\mathbf{V}_n = \begin{pmatrix} \mathbf{V}_{\mathbb{S}_1} + \epsilon\mathbf{1}\mathbf{1}^\top & 0 & \cdots & 0 \\ 0 & \mathbf{V}_{\mathbb{S}_2} + \epsilon\mathbf{1}\mathbf{1}^\top & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \mathbf{V}_{\mathbb{S}_\ell} + \epsilon\mathbf{1}\mathbf{1}^\top \end{pmatrix}$$

where $\mathbf{V}_{\mathbb{S}_1}, \mathbf{V}_{\mathbb{S}_2}, \ldots, \mathbf{V}_{\mathbb{S}_\ell}$ are the covariance matrices of the leaves of $\mathbb{S}_1, \mathbb{S}_2, \ldots, \mathbb{S}_\ell$ respectively. By the Woodbury matrix identity, we have

$$\begin{aligned}\mathbf{1}^\top \mathbf{V}_n^{-1}\mathbf{1} &= \sum_{i=1}^{\ell} \mathbf{1}^\top (\mathbf{V}_{\mathbb{S}_i} + \epsilon \mathbf{1}\mathbf{1}^\top)^{-1}\mathbf{1} \\ &= \sum_{i=1}^{\ell} \frac{1}{(\mathbf{1}^\top \mathbf{V}_{\mathbb{S}_i}^{-1}\mathbf{1})^{-1} + \epsilon} \le \frac{\ell}{\epsilon}.\end{aligned}$$

Next, we will prove that if the big bang condition holds, then there exists a consistent estimator for the ancestral state. By the big bang condition, for any positive integer m, there exists $k_m > k_{m-1}$ such that $|\partial \mathbb{T}_{k_m}(1/m)| \ge m$ with a convention that $k_0 = 0$. Thus, there exists a subtree of m leaves of $\mathbb{T}_{k_m}$ such that distances from the root to all internal nodes are less than $1/m$. Let $Y_{1,m}, \ldots, Y_{m,m}$ be the leaves of this subtree. We define our estimator as follows:

$$\hat{\mu}_n = \frac{Y_{1,m} + Y_{2,m} + \cdots + Y_{m,m}}{m}, \quad k_m \le n < k_{m+1}.$$

Note that $E(\hat{\mu}_n) = \mu$ and $\mathrm{Cov}(Y_{i,m}, Y_{j,m}) = \sigma^2 t_{ij,m} \le \sigma^2/m$ where $t_{ij,m}$ is the distance from the root to the most recent common ancestor of the leaves $Y_{i,m}$ and $Y_{j,m}$. Therefore,

$$\begin{aligned}\mathrm{Var}(\hat{\mu}_n) &= \frac{1}{m^2}\left(\sum_{i=1}^{m} \mathrm{Var}(Y_{i,m}) + 2\sum_{1\le i<j\le m} \mathrm{Cov}(Y_{i,m}, Y_{j,m})\right) \\ &\le \frac{1}{m^2}\left(mH\sigma^2 + m(m-1)\frac{\sigma^2}{m}\right) \le \frac{H+1}{m}\sigma^2 \to 0.\end{aligned}$$

By Chebyshev's inequality, for all $\epsilon > 0$, we have

$$P(|\hat{\mu}_n - \mu| \ge \epsilon) \le \frac{\mathrm{Var}(\hat{\mu}_n)}{\epsilon^2} \to 0.$$

Hence, $\hat{\mu}_n$ is a consistent estimator. □

Remark 1. We note that the first part of the proof of *Theorem 3* also shows that the condition $\mathbf{1}^\top \mathbf{V}_n^{-1}\mathbf{1} \to +\infty$ implies the big bang condition even without the assumption of bounded tree heights.

3.2. Necessary and sufficient condition for consistency of ancestral state reconstruction for threshold models

Theorem 4. *Assume that $\mathbf{1}^\top \mathbf{V}_n^{-1}\mathbf{1}$ are bounded. Then, there is no consistent estimator for the ancestral state under the threshold model.*

Proof. Let P_{μ,σ^2} and Q_{μ,σ^2} be the joint distribution of $\mathbf{Z}$ and $\mathbf{Y}$ respectively. We have

$$\mathrm{KL}(Q_{\mu_1,\sigma^2}, Q_{\mu_2,\sigma^2}) \le \mathrm{KL}(P_{\mu_1,\sigma^2}, P_{\mu_2,\sigma^2}) = \frac{1}{2\sigma^2}\mathbf{1}^\top \mathbf{V}_n^{-1}\mathbf{1}(\mu_1 - \mu_2)^2.$$

Applying Vajda's inequality (Vajda, 1970), we have

$$\begin{aligned}\mathrm{KL}(Q_{\mu_1,\sigma^2}, Q_{\mu_2,\sigma^2}) \ge \log\left(\frac{1 + d_{\mathrm{TV}}(Q_{\mu_1,\sigma^2}, Q_{\mu_2,\sigma^2})}{1 - d_{\mathrm{TV}}(Q_{\mu_1,\sigma^2}, Q_{\mu_2,\sigma^2})}\right) \\ - \frac{2d_{\mathrm{TV}}(Q_{\mu_1,\sigma^2}, Q_{\mu_2,\sigma^2})}{d_{\mathrm{TV}}(Q_{\mu_1,\sigma^2}, Q_{\mu_2,\sigma^2}) + 1}.\end{aligned}$$

Hence,

$$d_{\mathrm{TV}}(Q_{\mu_1,\sigma^2}, Q_{\mu_2,\sigma^2}) \le c < 1.$$

Assume that there exists a consistent estimator $\hat{\rho}_n$ for the ancestral state $\rho = \mathrm{sign}(\mu)$. Define $\mathcal{A} = \{\hat{\rho}_n = 1\}$, we have

$$Q_{1,1}(\mathcal{A}) \to 1 \quad \text{and} \quad Q_{-1,1}(\mathcal{A}) \to 0,$$

which implies

$$\begin{aligned}d_{\mathrm{TV}}(Q_{1,1}, Q_{-1,1}) &= \sup_A(|Q_{1,1}(A) - Q_{-1,1}(A)|) \\ &\ge |Q_{1,1}(\mathcal{A}) - Q_{-1,1}(\mathcal{A})| \to 1.\end{aligned}$$

This is a contradiction. Therefore, there is no consistent estimator for $\rho = \mathrm{sign}(\mu)$. □

Lemma 3 (*Lancaster (1957)*)**.** *Let (X, Y) be a bivariate normal distribution and two functions f, g such that $E(f(X)^2) < +\infty$ and $E(g(Y)^2) < +\infty$. Then*

$$\frac{|\mathrm{Cov}(f(X), g(Y))|}{\sqrt{\mathrm{Var}(f(X))\mathrm{Var}(g(Y))}} \le \frac{|\mathrm{Cov}(X, Y)|}{\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}}.$$

Theorem 5. *Assume that big bang condition is satisfied. Then, there is a consistent ancestral state reconstruction method under the threshold model.*

Proof. Let k_m be the increasing sequence constructed in the proof of Theorem 3. There exists a subtree of m leaves of $\mathbb{T}_{k_m}$ such that distances from the root to all internal nodes are less than $1/m$. Let $Y_{1,m}, \ldots, Y_{m,m}$ be the leaves of this subtree. Let $t_{i,m}$ be the distance from the root to the leaf $Y_{i,m}$, and $t_{ij,m}$ be the distance from the root to the most recent common ancestor of the leaves $Y_{i,m}$ and $Y_{j,m}$.

If there is a sequence of leaves whose distance to the root converges to 0, then their trait values form a trivial consistent estimator of the ancestral state. Formally, denote $\tau_m = \min_i t_{i,m}$ and $s_m = \arg\min_i t_{i,m}$. If there exists a subsequence $\tau_{m_u} \to 0$, then $Y_{s_{m_u},m_u}$ is a trivial consistent estimator for the ancestral state $\rho = \mathrm{sign}(\mu)$.

On the other hand, if there exists $\alpha > 0$ such that $\tau_m \ge \alpha$ for all m, we will prove that

$$\begin{aligned}\hat{\rho}_n &= \mathrm{sign}(\overline{Y}_m) \\ &= \mathrm{sign}\left(\frac{Y_{1,m} + Y_{2,m} + \cdots + Y_{m,m}}{m}\right), \quad k_m \le n < k_{m+1}\end{aligned}$$

is a consistent estimator for the ancestral state. Without the loss of generality, we assume that $\rho = \mathrm{sign}(\mu) = 1$. We have

$$\begin{aligned}E(\overline{Y}_m) &= \frac{1}{m}\sum_{i=1}^{m}\left[P(Z_{i,m} > 0) - P(Z_{i,m} < 0)\right] \\ &= \left(\frac{2}{m}\sum_{i=1}^{m} P(Z_{i,m} > 0)\right) - 1 \\ &= \left(\frac{2}{m}\sum_{i=1}^{m} \Phi\left(\frac{\mu}{\sigma\sqrt{t_{i,m}}}\right)\right) - 1 \ge 2\Phi\left(\frac{\mu}{\sigma\sqrt{H}}\right) - 1 > 0\end{aligned}$$

where Φ is the cumulative distribution function of the standard Normal distribution. Hence,

$$\begin{aligned}P(\hat{\rho}_n = -1) = P(\overline{Y}_m < 0) &= P[\overline{Y}_m - E(\overline{Y}_m) < -E(\overline{Y}_m)] \\ &\le P[|\overline{Y}_m - E(\overline{Y}_m)| \ge E(\overline{Y}_m)] \le \frac{\mathrm{Var}(\overline{Y}_m)}{E(\overline{Y}_m)^2} \\ &\le \frac{\mathrm{Var}(\overline{Y}_m)}{\left[2\Phi\left(\frac{\mu}{\sigma\sqrt{H}}\right) - 1\right]^2}.\end{aligned}$$

Note that $\mathrm{Var}(Y_{i,m}) \le 1$ since $|Y_{i,m}| = 1$. By Lemma 3, we have

$$\begin{aligned}|\mathrm{Cov}(Y_{i,m}, Y_{j,m})| &\le \frac{t_{ij,m}}{\sqrt{t_{i,m}t_{j,m}}}\sqrt{\mathrm{Var}(Y_{i,m})\mathrm{Var}(Y_{j,m})} \\ &\le \frac{t_{ij,m}}{\sqrt{t_{i,m}t_{j,m}}} \le \frac{1}{m\alpha}.\end{aligned}$$

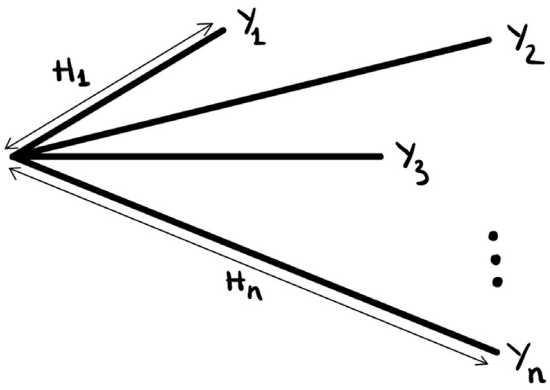

Fig. 3. An n-species star tree.

Therefore,

$$\begin{aligned}\mathrm{Var}(\overline{Y}_m) &= \frac{1}{m^2}\left(\sum_{i=1}^{m}\mathrm{Var}(Y_{i,m}) + 2\sum_{1\le i<j\le m}\mathrm{Cov}(Y_{i,m}, Y_{j,m})\right)\\ &\le \frac{1}{m^2}\left(m + m(m-1)\frac{1}{m\alpha}\right) \le \frac{1+\alpha^{-1}}{m} \to 0.\end{aligned}$$

We conclude that

$$P(\hat{\rho}_n = -1) \le \frac{\mathrm{Var}(\overline{Y}_m)}{\left[2\Phi\left(\frac{\mu}{\sigma\sqrt{H}}\right) - 1\right]^2} \to 0.$$

Thus, $\hat{\rho}_n$ is a consistent estimator. □

Remark 2. We complete the proof of *Theorem* 1 by combining *Lemmas* 1 and 2 with *Theorems* 2, 3, 4, and 5.

4. Unbounded heights

When the tree heights are unbounded, the equivalence between the condition $\mathbf{1}^\top\mathbf{V}_n^{-1}\mathbf{1} \to +\infty$ and the big bang condition is no longer valid. To see this, let us consider a simple scenario where $(\mathbb{T}_n)_{n=1}^{\infty}$ is a sequence of nested star tree (see Fig. 3). Let H_n be the distance from the root to the nth species. It is trivial that the sequence of trees $(\mathbb{T}_n)_{n=1}^{\infty}$ satisfies the big bang condition. On the other hand,

$$\mathbf{1}^\top\mathbf{V}_n^{-1}\mathbf{1} = \sigma^2\sum_{n=1}^{\infty}\frac{1}{H_n}.$$

Hence, $\mathbf{1}^\top\mathbf{V}_n^{-1}\mathbf{1} \to +\infty$ if and only if

$$\sum_{n=1}^{\infty}\frac{1}{H_n} \to +\infty.$$

Therefore, the condition $\mathbf{1}^\top\mathbf{V}_n^{-1}\mathbf{1} \to +\infty$ and the big bang condition are not equivalent when tree heights are unbounded.

We note that Theorems 2 and 4 do not require the heights of trees are bounded. Therefore, even without the bounded heights condition, $\mathbf{1}^\top\mathbf{V}_n^{-1}\mathbf{1} \to +\infty$ is the necessary condition for the existence of a consistent ancestral state reconstruction method under the BM model and threshold model. On the other hand, the big bang condition is the necessary condition for the existence of a consistent estimator under regular discrete models when tree heights are not bounded (see the proof of Proposition 3.1 in Fan and Roch, 2018). A natural question is when tree heights are unbounded, whether either the big bang condition or the condition $\mathbf{1}^\top\mathbf{V}_n^{-1}\mathbf{1} \to +\infty$ is a sufficient condition for these models. Ané (2008) gives a positive answer for the BM model by showing that the MLE is consistent if $\mathbf{1}^\top\mathbf{V}_n^{-1}\mathbf{1} \to +\infty$. Unfortunately, without additional conditions, neither condition is enough to guarantee that there is a consistent method for reconstructing the ancestral state under regular discrete models. Specifically, we provide a simple counter-example using a sequence of nested star trees and the two-state symmetric model.

Theorem 6. *Consider a sequence of nested star trees $(\mathbb{T}_n)_{n=1}^{\infty}$. Let H_n be the height of $\mathbb{T}_n$ such that $H_n/n \to 0$ and $H_n/\log n \to \infty$. Then,*

$$\sum_{n=1}^{\infty}\frac{1}{H_n} \to +\infty,$$

but there is no consistent ancestral state reconstruction method under the two-state symmetric model.

Proof. Let P_ρ be the joint distribution of the observations $\mathbf{Y}_n = (Y_1, Y_2, \ldots, Y_n)$ under the two-state symmetric model with ancestral state ρ. Denote $p_k = [1 + \exp(-\eta H_k)]/2$ where η is the mutation rate of the binary trait. We have

$$\begin{aligned}\mathrm{KL}(P_1, P_0) &= \sum_{k=1}^{n} p_k\log\left(\frac{p_k}{1-p_k}\right) + (1-p_k)\log\left(\frac{1-p_k}{p_k}\right)\\ &= \sum_{k=1}^{n}(2p_k-1)\log\left(1+\frac{2p_k-1}{1-p_k}\right)\\ &\le \sum_{k=1}^{n}\frac{(2p_k-1)^2}{1-p_k} \le C\sum_{k=1}^{n}(2p_k-1)^2\\ &= C\sum_{k=1}^{n}\exp(-2\eta H_k) = C\sum_{k=1}^{n}\left(\frac{1}{k}\right)^{2\eta H_k/\log k} < +\infty.\end{aligned}$$

By the same arguments of Theorems 2 and 4, we deduce that there is no consistent estimator for the ancestral state. □

5. Conclusion and discussion

In this work, we provide a unified theory for ancestral state reconstruction across different models for a sequence of nested trees with bounded heights. We show that the condition $\mathbf{1}^\top\mathbf{V}_n^{-1}\mathbf{1} \to +\infty$ arose from the study of the BM model is equivalent to the big bang condition for discrete models. Furthermore, these conditions are the necessary and sufficient condition for the existence of a consistent estimator for the ancestral state under the BM, regular discrete, and threshold models.

We note that the "big bang" condition does not hold under the coalescent process because earlier lineages are less likely to be divided into new species. However, it holds under the coalescent point process (Lambert and Stadler, 2013), for which the ages of internal nodes are independent and identically distributed according to a probability distribution in $[0, H]$. Even when the "big bang" condition does not hold, it still provides an insight into how to collect data to maximize the information about the ancestral state. In particular, species stemming close to the root provide more information than species stemming near the present time.

We provide a simple counter-example to show that when tree heights are unbounded, the condition $\mathbf{1}^\top\mathbf{V}_n^{-1}\mathbf{1} \to +\infty$ and the big bang condition are no longer equivalent. Moreover, neither condition is a sufficient condition for the existence of a consistent ancestral state reconstruction method under regular discrete models. It is worth noticing that the condition $\mathbf{1}^\top\mathbf{V}_n^{-1}\mathbf{1} \to +\infty$ is the necessary and sufficient condition for the existence of a consistent estimator for the ancestral state under the BM model without the requirement of bounded tree heights. Establishing a

necessary and sufficient condition for regular discrete models and the threshold model when tree heights are unbounded remains open.

It is worth noticing that the MLE for the ancestral state is consistent under the BM model when the condition $\mathbf{1}^\top \mathbf{V}_n^{-1}\mathbf{1} \to +\infty$ holds (Ané, 2008). Furthermore, by Proposition 6 in Steel and Rodrigo (2008), when evolution dynamics of a regular finite-state discrete model is known, the MLE for the ancestral state is consistent if there exists a consistent ancestral state reconstruction method. However, the consistency of the MLE under the threshold model remains unknown. In some scenarios, there exists a better ancestral state reconstruction method than the MLE (Ho et al., 2019; Ho and Susko, 2022). Therefore, the condition $\mathbf{1}^\top \mathbf{V}_n^{-1}\mathbf{1} \to +\infty$, which also implies the big bang condition, may not be a sufficient condition for the consistency of the MLE under the threshold model.

Data availability

No data was used for the research described in the article.

Acknowledgments

LSTH was supported by startup funds from Dalhousie University, the Canada Research Chairs program, the NSERC Discovery Grant RGPIN-2018-05447, and the NSERC Discovery Launch Supplement DGECR-2018-00181. VD was supported by a startup fund from the University of Delaware and National Science Foundation grant DMS-1951474.

References

Ané, C., 2008. Analysis of comparative data with hierarchical autocorrelation. Ann. Appl. Stat. 2 (3), 1078–1102.
Ané, C., Ho, L.S.T., Roch, S., 2017. Phase transition on the convergence rate of parameter estimation under an Ornstein–Uhlenbeck diffusion on a tree. J. Math. Biol. 74 (1), 355–385.
Edwards, A.W., 1970. Estimation of the branch points of a branching diffusion process. J. R. Stat. Soc. Ser. B Stat. Methodol. 32 (2), 155–164.
Fan, W.-T.L., Roch, S., 2018. Necessary and sufficient conditions for consistent root reconstruction in Markov models on trees. Electron. J. Probab. 23.
Faria, N.R., Rambaut, A., Suchard, M.A., Baele, G., Bedford, T., Ward, M.J., Tatem, A.J., Sousa, J.D., Arinaminpathy, N., Pépin, J., et al., 2014. The early spread and epidemic ignition of HIV-1 in human populations. Science 346 (6205), 56–61.
Felsenstein, J., 1985. Phylogenies and the comparative method. Amer. Nat. 125 (1), 1–15.
Felsenstein, J., 2004. Inferring Phylogenies, Vol. 2. Sinauer associates Sunderland, MA.
Felsenstein, J., 2005. Using the quantitative genetic threshold model for inferences between and within species. Philos. Trans. R. Soc. B 360 (1459), 1427–1434.
Felsenstein, J., 2012. A comparative method for both discrete and continuous characters using the threshold model. Amer. Nat. 179 (2), 145–156.
Gill, M.S., Ho, L.S.T., Baele, G., Lemey, P., Suchard, M.A., 2017. A relaxed directional random walk model for phylogenetic trait evolution. Syst. Biol. 66 (3), 299–319.
Ho, L.S.T., Ané, C., 2013. Asymptotic theory with hierarchical autocorrelation: Ornstein–Uhlenbeck tree models. Ann. Statist. 41 (2), 957–981.
Ho, L.S.T., Ané, C., 2014. Intrinsic inference difficulties for trait evolution with Ornstein–Uhlenbeck models. Methods Ecol. Evol. 5 (11), 1133–1146.
Ho, L.S.T., Dinh, V., Nguyen, C.V., 2019. Multi-task learning improves ancestral state reconstruction. Theor. Popul. Biol. 126, 33–39.
Ho, L.S.T., Susko, E., 2022. Ancestral state reconstruction with large numbers of sequences and edge-length estimation. J. Math. Biol. 84 (4), 1–28.
Jukes, T.H., Cantor, C.R., 1969. Evolution of protein molecules. Mammalian Protein Metab. 3, 21–132.
Lambert, A., Stadler, T., 2013. Birth–death models and coalescent point processes: The shape and probability of reconstructed phylogenies. Theor. Popul. Biol. 90, 113–128.
Lanave, C., Preparata, G., Sacone, C., Serio, G., 1984. A new method for calculating evolutionary substitution rates. J. Mol. Evol. 20 (1), 86–93.
Lancaster, H.O., 1957. Some properties of the bivariate normal distribution considered in the form of a contingency table. Biometrika 44 (1/2), 289–292.
Li, G., Steel, M., Zhang, L., 2008. More taxa are not necessarily better for the reconstruction of ancestral character states. Syst. Biol. 57 (4), 647–653.
Liberles, D.A., 2007. Ancestral Sequence Reconstruction. Oxford University Press on Demand.
Maddison, D.R., 1994. Phylogenetic methods for inferring the evolutionary history and processes of change in discretely valued characters. Annu. Rev. Entomol. 39 (1), 267–292.
Revell, L.J., 2014. Ancestral character estimation under the threshold model from quantitative genetics. Evolution 68 (3), 743–759.
Steel, M., 2016. Phylogeny: Discrete and Random Processes in Evolution. SIAM.
Steel, M., Rodrigo, A., 2008. Maximum likelihood supertrees. Syst. Biol. 57 (2), 243–250.
Tavaré, S., 1986. Some probabilistic and statistical problems in the analysis of DNA sequences. Lect. Math. Life Sci. 17 (2), 57–86.
Thornton, J.W., 2004. Resurrecting ancient genes: Experimental analysis of extinct molecules. Nature Rev. Genet. 5 (5), 366–375.
Vajda, I., 1970. Note on discrimination information and variation (corresp.). IEEE Trans. Inform. Theory 16 (6), 771–773.
Wright, S., 1934a. An analysis of variability in number of digits in an inbred strain of guinea pigs. Genetics 19 (6), 506.
Wright, S., 1934b. The results of crosses between inbred strains of guinea pigs, differing in number of digits. Genetics 19 (6), 537.