On the convergence of the maximum likelihood estimator for the transition rate under a 2-state symmetric model

Lam Si Tung Ho1 · Vu Dinh2 · Frederick A. Matsen IV3 · Marc A. Suchard4

Abstract
Maximum likelihood estimators are used extensively to estimate unknown parameters of stochastic trait evolution models on phylogenetic trees. Although the MLE has been proven to converge to the true value in the independent-sample case, we cannot appeal to this result because trait values of different species are correlated due to shared evolutionary history. In this paper, we consider a 2-state symmetric model for a single binary trait and investigate the theoretical properties of the MLE for the transition rate in the large-tree limit. Here, the large-tree limit is a theoretical scenario where the number of taxa increases to infinity and we can observe the trait values for all species. Specifically, we prove that the MLE converges to the true value under some regularity conditions. These conditions ensure that the tree shape is not too irregular, and holds for many practical scenarios such as trees with bounded edges, trees generated from the Yule (pure birth) process, and trees generated from the coalescent point process. Our result also provides an upper bound for the distance between the MLE and the true value.

Keywords Maximum likelihood estimator · 2-state symmetric model · Trait evolution · Phylogenetics · Yule process · Coalescent point process

Mathematics Subject Classification Primary 62F12; Secondary 92D15 · 92B10

Lam Si Tung Ho and Vu Dinh have equally contributed to this work.

LSTH was supported by startup funds from Dalhousie University, the Canada Research Chairs program, and the Natural Sciences and Engineering Research Council of Canada (NSERC) Discovery Grant RGPIN-2018-05447. FAM was supported by CISE-1564137, and in part by a Faculty Scholar grant from the Howard Hughes Medical Institute and the Simons Foundation. MAS was supported by National Science Foundation Grant DMS1264153 and National Institutes of Health grant R01 AI107034.

Extended author information available on the last page of the article
1 Introduction

The maximum likelihood estimator (MLE) is frequently used to estimate unknown parameters in trait evolution models on phylogenetic trees. To safely use this machinery, it is important to know that the MLE is consistent: that is, the estimate converges to the true value as we observe trait values for more species. However, traditional statistical theory only guarantees the consistency of the MLE when observations are independent and identically distributed. In contrast, trait values of biological species are not independent because species are related to each other according to a phylogenetic tree. Therefore, traditional consistency results are not directly applicable to trait evolution studies. For this paper we consider the scenario where only a single trait is observed and the number of species increases to infinity. This is different from the setting where many traits/characters are observed for the same set of species; this alternate setting typically assumes independence between traits.

Binary traits studied by comparative biologists come in various types, including morphological traits (whether a fruit fly has curly wings), behavioral traits (whether an antelope hides from predators), and geographical traits (whether a species is aquatic). Here we study the consistency property of the MLE for estimating the transition rate of a binary trait evolved along a phylogenetic tree according to a 2-state symmetric Markov process. Under this 2-state symmetric model, there is an unique rate of switching back and forth between the two states of a phenotype, which is the parameter of interest. We show that under two mild conditions, the MLE of this transition rate is consistent, that is, the estimate converges to the correct value as we observe the trait values of more species. These conditions ensure that edge lengths of the tree are not too small and the pairwise distances between leaves are neither too small nor too large. We verify that these conditions holds for many practical cases including trees with bounded edges, trees generated from the Yule process (Yule 1925), and trees generated from the coalescent point process (Lambert and Stadler 2013).

The consistency of the MLE does not always hold under trait evolution models. Indeed, for estimating the ancestral state of the 2-state symmetric model, Li et al. (2008) point out that the MLE can be inconsistent. As a consequence, the estimate of the ancestral state may not be close to the true value no matter how many species have been sampled. Several efforts have also been made to investigate the consistency of the MLE under evolution models of a continuous trait. Ané (2008) points out that unlike traditional linear regression, the MLE of the coefficients of phylogenetic linear regression under the Brownian motion model can be inconsistent. Additionally, Sagitov and Bartoszek (2012) show that the sample mean is an inconsistent estimator for the ancestral state under this model when the tree is generated from the Yule process. For the Ornstein–Uhlenbeck model, Ho and Ané (2013) show that if the height of the phylogenetic tree is bounded as more observations are collected, then the MLE of the selective optimum is not consistent. Moreover, they discovered that in this scenario, no consistent estimator for the selective optimum exists. Recently, Ané et al. (2017) provide a necessary and sufficient condition for consistency of the MLE under the Ornstein–Uhlenbeck model. Although the problem of reconstructing the ancestral state under the 2-state symmetric model has been studied extensively
On the convergence of the maximum likelihood estimator…

Fig. 1 Example of rerooting a 4-taxa tree

(Tuffley and Steel 1997; Mooers and Schluter 1999; Li et al. 2008; Mossel and Steel 2014), it remains unknown whether the transition rate can be estimated consistently.

In this paper, we show that the MLE is a consistent estimator for the transition rate of the 2-state symmetric model under simple conditions. We start with introducing the 2-state symmetric model for binary traits and derive several statistical properties of this model in Sect. 2. In Sect. 3, we state two necessary conditions for the consistency of the MLE of the transition rate and provide a detailed proof for this result. Section 4 verifies these conditions for several practical scenarios and illustrates our result through a simulation.

2 Properties of the 2-state symmetric model

Let $\mu$ be the transition rate of a 2-state symmetric Markov process. Then, the transition probability matrix has an analytical form:

$$
P_{\mu}(t) = \begin{pmatrix}
\frac{1}{2} + \frac{1}{2} e^{-2\mu t} & \frac{1}{2} - \frac{1}{2} e^{-2\mu t} \\
\frac{1}{2} - \frac{1}{2} e^{-2\mu t} & \frac{1}{2} + \frac{1}{2} e^{-2\mu t}
\end{pmatrix}.
$$

(1)

Hence, the probability that the process switches state after $t$ unit of time is $\frac{1}{2} - \frac{1}{2} e^{-2\mu t}$.

In this paper, the term phylogenetic tree (or phylogeny) refers to a bifurcating rooted tree with leaves labeled by a set of taxon (species) names. We can reroot a phylogenetic tree by moving the root to another location along the tree (see Fig. 1). The evolution of a binary trait along a tree is modeled using the 2-state symmetric Markov process as follows. At each node in the phylogeny, the children inherit the trait value of their parent and the trait of each child evolves independently of one another.

Let $T$ be a phylogenetic tree with $n$ leaves, and $Y$ be the trait values at the leaves of $T$. We assume that the ancestral state at the root $\rho$ of $T$ follows a stationary distribution, which is a Bernoulli distribution with success probability $1/2$. Let $E$ be the set of all edges of $T$. The joint probability distribution of $Y$ is

$$
\mathbb{P}_{T,\mu}(Y) := \mathbb{P}(Y \mid T, \mu) = \frac{1}{2} \sum_y \left( \prod_{(u,v) \in E} [P_{\mu}(d_{uv})]_{y_u y_v} \right)
$$

where $y$ ranges over all extensions of $Y$ to the internal nodes of the tree, $y_u$ denotes the assigned state of node $u$ by $y$, $d_{uv}$ is the edge length of $(u, v)$, and $[P_{\mu}(t)]_{kl}$ is the element at $k$-th row and $l$-th column of matrix $P_{\mu}(t)$. We define the log-likelihood...
function as $\ell_{T,\mu}(Y) = \log P_{T,\mu}(Y)$. Throughout this paper, we denote the true transition rate with $\mu^*$ and assume that $\mu^* \in [\underline{\mu}, \overline{\mu}]$ where $\underline{\mu}, \overline{\mu}$ are two known positive numbers. Define

$$R_{T,\mu}(Y) = \ell_{T,\mu^*}(Y) - \ell_{T,\mu}(Y).$$

Then the Kullback–Leibler divergence from $P_{T,\mu}$ to $P_{T,\mu^*}$ is

$$\text{KL}(P_{T,\mu^*} \parallel P_{T,\mu}) = \mathbb{E}[\ell_{T,\mu^*}(Y) - \ell_{T,\mu}(Y)] = \mathbb{E}[R_{T,\mu}(Y)]$$

where $\mathbb{E}$ is the expectation with respect to $P_{T,\mu^*}$.

**Lemma 1** Let $T'$ be the phylogenetic tree obtained by rerooting a tree $T$, then

$$\ell_{T',\mu}(Y) = \ell_{T,\mu}(Y)$$

for any trait values $Y$ at the leaves.

This lemma, sometimes called the “pulley principle”, is a direct consequence of the fact that the ancestral state follows a stationary distribution (Felsenstein 1981). We also have the following lemmas:

**Lemma 2** Let $T$ be a rooted tree with root $\rho$, and $Y$ be the trait values at the leaves of $T$ generated under the 2-state symmetric model. Let $h$ be a function such that $h(Y) = h(1 - Y)$, where $1$ denotes the all-ones vector. Then, $h(Y)$ and the trait value at $\rho$ are independent.

The following two lemmas concern the regularity of the log-likelihood function. In particular, Lemma 3 shows that the log-likelihood function of a binary tree can be bounded by the sum of the log-likelihood functions of its two subtrees.

**Lemma 3** Let $\rho_1$ and $\rho_2$ be two direct descendants of $\rho$, and $T_1$ and $T_2$ the two subtrees descending from them. Let $Y_1, Y_2$ be the observations at the leaves of $T_1$ and $T_2$. We have

$$|\ell_{T,\mu}(Y) - \ell_{T_1,\mu}(Y_1) - \ell_{T_2,\mu}(Y_2)| \leq \max \left\{ \log 2, \log \frac{1}{1 - e^{-2\mu d}} \right\},$$

where $d$ is the tree distance between $\rho_1$ and $\rho_2$.

**Lemma 4** (Uniform Lipschitz) There exists $C > 0$ such that

$$\left| \frac{1}{n} \ell_{T,\mu_1}(Y) - \frac{1}{n} \ell_{T,\mu_2}(Y) \right| \leq C|\mu_1 - \mu_2| \quad \forall \mu_1, \mu_2 \in [\underline{\mu}, \overline{\mu}], \forall T$$

where $n$ denotes the number of leaves of $T$.

The Proofs of Lemmas 2, 3, and 4 are provided in the “Appendix A”. Henceforth we will also use $\rho$ to denote the trait value at the root by an abuse of notation.
3 Convergence of the MLE of the transition rate

The MLE of $\mu$ is defined as follows:

$$\hat{\mu} = \arg\max_{\mu \in [\mu, \mu]} \ell_{T, \mu}(Y).$$

In this section, we will state our main result regarding the consistency of $\hat{\mu}$. We will need to make two assumptions to ensure that the shape of $T$ is not too irregular. Let us define

$$f(x) = \max \left\{ \log 2, \log \frac{1}{1 - e^{-2x}} \right\} \quad \text{and} \quad S_{T, \mu} = \sum_{e \in E(T)} f(\mu t_e)^2.$$

where $t_e$ denotes the edge length of edge $e$.

**Assumption 1** $S_{T, \mu} = O(\gamma n^\gamma)$ for some $1 \leq \gamma < 2$. That is, there exists a universal constant $c < \infty$ such that $S_{T, \mu} < cn^\gamma$.

**Assumption 2** There exist $\Omega(n)$ pairs of leaves such that the paths connecting each pair are pairwise disconnected and their lengths are bounded in some fixed range $[d, \bar{d}]$. Here, $\Omega(n)$ denotes a quantity that is greater than $cn$ for a positive constant $c$ and all $n$.

Assumption 1 makes sure that the edge lengths of $T$ are not too small. It is worth noticing that $1/(1 - e^{-2x}) \leq 1/x$ when $x$ is small enough. Therefore, this assumption holds when the smallest edge length is $\Omega(e^{-n(\gamma-1)/2})$. On the other hand, Assumption 2 guarantees that the pairwise distances between leaves of $T$ do not vary too extremely. In Sect. 4, we will verify these assumptions for several common tree models. Although we employ rerooting in proofs below, Assumption 1 is for the original root of the tree.

**Theorem 1** Under Assumptions 1 and 2, for any $\delta > 0$, $\forall \alpha \in \left( \max \left\{ \frac{\log(2)}{\log(3/2)}, \gamma \right\}, 2 \right)$, there exists a constant $C_{\delta, \alpha, d, \bar{d}, \mu, \mu} > 0$ such that

$$|\hat{\mu} - \mu^*| \leq C_{\delta, \alpha, d, \bar{d}, \mu, \mu} \left( \frac{\sqrt{\log n}}{n^{(2-\alpha)/6}} \right)$$

with probability $1 - \delta$.

From now on, we will use $\ell_{T, \mu}$ and $R_{T, \mu}$ as short for $\ell_{T, \mu}(Y)$ and $R_{T, \mu}(Y)$. Recall that $\ell_{T, \mu}$ is the log-likelihood function and $R_{T, \mu} = \ell_{T, \mu^*} - \ell_{T, \mu}$ where $\mu^*$ is the true value of the transition rate. The main ideas of the Proof of Theorem 1 can be outlined as follows. For a fixed value of $\mu$, we can view the function $R_{T, \mu}$ as the evidence to distinguish between $\mu$ and $\mu^*$. We will prove that as the number of leaves approaches infinity, the information to distinguish $\mu$ and $\mu^*$, characterized by the KL divergence between $P_{T, \mu}$ to $P_{T, \mu^*}$ (which is equal to $\mathbb{E}(R_{T, \mu})$) increases linearly Lemma 6, while
the associated uncertainty, characterized by the variance of $R_{T, \mu}$, only increases sub-quadratically Lemma 9. It is worth noting that while the results of Lemmas 6 and 9 are straightforward for independent and identically distributed data, the analyses for phylogenetic traits are more complicated due to the correlations between trait values at the leaves. To overcome this issue, we use the independent phylogenetic contrasts, introduced in the next section, to obtain a lower bound on the information. On the other hand, Lemma 3 shows that the log-likelihood function of a binary tree can be bounded by the sum of the log-likelihood functions of its two subtrees, which allows us to exploit the sparse structure of the tree to derive an upper bound on uncertainty through an induction argument. Finally, we obtain a uniform bound on the difference between the log-likelihood functions and its expected values Lemma 10, which enables us to derive an analysis of convergence of the MLE.

3.1 Lower bound on information

Phylogenetic contrasts Letting $i$ and $j$ be two different species, we define $C_{ij} = Y_i - Y_j$ to be a contrast between the two species. This is a popular notion introduced by Felsenstein (1985) for computing the likelihood function under the Brownian motion model. Ané et al. (2017) use independent contrasts to construct consistent estimators for the covariance parameters of the Ornstein–Uhlenbeck model. Here, we will introduce a notion of a squared-contrast $C_{ij}^2$, which is simply the square of a contrast $C_{ij}$, and show that squared-contrasts have the same independence properties as independent contrasts. Let $A_{ij}$ be the state of the most recent common ancestor of $i$ and $j$, $d_{ij}$ the tree distance from $i$ to $j$, and $p_{ij}$ the path on the tree connecting $i$ and $j$. By symmetry of the model, we have

$$P(C_{ij}^2 = y \mid A_{ij} = 0) = P(C_{ij}^2 = y \mid A_{ij} = 1) = \begin{cases} \frac{1}{2}(1 + e^{-2\mu d_{ij}}), & \text{if } y = 0 \\ \frac{1}{2}(1 - e^{-2\mu d_{ij}}), & \text{if } y = 1 \end{cases} \quad (2)$$

Therefore, $C_{ij}^2$ and $A_{ij}$ are independent.

Lemma 5 (Sequence of contrasts) Let $\{C_{ik,jk}\}_{k=1}^m$ be a sequence of contrasts such that any two paths in $\{p_{ik,jk}\}_{k=1}^m$ have no common node. Then $\{C_{ik,jk}^2\}_{k=1}^m$ are independent.

Proof We have, using $E(\cdot)$ here to denote expectation over ancestral states,

$$P(C_{i1,j1}^2 = y_1, C_{i2,j2}^2 = y_2, \ldots, C_{im,jm}^2 = y_m)$$

$$= E\left(P(C_{i1,j1}^2 = y_1, C_{i2,j2}^2 = y_2, \ldots, C_{im,jm}^2 = y_m \mid \{A_{ik,jk}\}_{k=1}^m)\right)$$

$$= E\left(\prod_{k=1}^m P(C_{ik,jk}^2 = y_k \mid \{A_{ik,jk}\}_{k=1}^m)\right) = E\left(\prod_{k=1}^m P(C_{ik,jk}^2 = y_k \mid A_{ik,jk})\right)$$

$$= E\left(\prod_{k=1}^m P(C_{ik,jk}^2 = y_k)\right) = \prod_{k=1}^m E(P(C_{ik,jk}^2 = y_k), \hat{\text{Springer}}
where the third equality comes from the Markov property. This completes the proof.

The independent squared-contrasts allow us to derive the following lower bound on the information.

**Lemma 6** Under Assumption 2, there exists \( C_{d,d,\mu} > 0 \) such that
\[
\mathbb{E}[R_{T,\mu}] \geq C_{d,d,\mu} n |\mu - \mu^*|^2
\]
for all \( \mu \in [\underline{\mu}, \overline{\mu}] \).

**Proof** Under Assumption 2, there exists a set of \( m = \Omega(n) \) pairwise disjoint pairs (that is, the paths connecting each pair are pairwise disconnected) \((Y_{k,1}, Y_{k,2})_{k=1}^m\). Consider the corresponding set of squared-contrasts \( C_{ik,jk}^2 = (Y_{k,1} - Y_{k,2})^2 \). By Lemma 5, \((C_{ik,jk}^2)_{k=1}^m\) are independent. Let \( Q_{k,\mu} \) be the distribution of \( C_{ik,jk}^2 \) corresponding to parameter \( \mu \). The total variation from \( Q_{k,\mu^*} \) to \( Q_{k,\mu} \) is
\[
\text{TV}(Q_{k,\mu^*} \| Q_{k,\mu}) = \frac{1}{2} \sum_{x \in \{0,1\}} |Q_{k,\mu^*}(x) - Q_{k,\mu}(x)|
\]
\[
= |e^{-2\mu^* d_{ik,jk}} - e^{-2\mu d_{ik,jk}}| \geq C_{d,d,\mu} |\mu^* - \mu|.
\]
By the data processing inequality (Theorem 9 in Van Erven and Harremoës 2014), the fact that \((C_{ik,jk}^2)_{k=1}^m\) are independent, and Pinsker’s inequalities, we have
\[
\mathbb{E}[R_{T,\mu}] = \text{KL}(P_{T,\mu^*} \| P_{T,\mu}) \geq \sum_{k=1}^m \text{KL}(Q_{k,\mu^*} \| Q_{k,\mu})
\]
\[
\geq 2 \sum_{k=1}^m [\text{TV}(Q_{k,\mu^*} \| Q_{k,\mu})]^2 \geq m C_{d,d,\mu} |\mu^* - \mu|^2.
\]
Note that \( m = \Omega(n) \) which completes the proof.

### 3.2 Upper bound on the uncertainty

The result of Sect. 3.1 indicates that as the number of leaves increases, the information to distinguish \( \mu \) and \( \mu^* \) also increases linearly. In order to prove that the MLE can successfully reconstruct the true parameter \( \mu \), we need to ensure that such information is not confounded by the uncertainty associated with the log-likelihood function due to randomness in the data. In other words, we want to guarantee that as \( n \to \infty \), the variance of \( R_{T,\mu} \) only grows sub-quadratically, that is, there exist \( \alpha < 2 \) and \( C > 0 \) such that \( \text{Var} R_{T,\mu} \leq C n^\alpha \). To obtain the bound, we need the following Lemma, the proof of which appears in “Appendix A”.

**Lemma 7** If two functions \( u \) and \( v \) satisfy \( \|u - v\|_\infty \leq c \), then for all \( \omega > 1 \) and random variables \( X \),
\[ \text{Var}[u(X)] \leq \omega \text{Var}[v(X)] + \frac{2\omega}{\omega - 1} c^2. \]

Recalling that,
\[ f(x) = \max \left\{ \log 2, \log \frac{1}{1 - e^{-2x}} \right\}, \quad S_{T,\mu} = \sum_{e \in E(T)} f(\mu e)^2, \]
we have the following estimate that does not depend on Assumption 1.

**Lemma 8** For all \( \alpha \in \left( \frac{\log 2}{\log(3/2)}, 2 \right) \) and \( \beta \in (0, 1) \), there exist \( C_{\alpha,\mu,\overline{\mu}}, \omega_{\alpha,\beta} > 1 \) such that
\[ \text{Var}R_{T,\mu} \leq C_{\alpha,\mu,\overline{\mu}} n^\alpha + \frac{8\omega_{\alpha,\beta}}{\omega_{\alpha,\beta} - 1} n^\beta S_{T,\mu} \]
for all trees \( T \) with \( n \geq 2 \) taxa and \( \mu \in [\mu, \overline{\mu}] \).

**Proof** We will prove the result by induction in the number of taxa \( n \). For \( n = 2 \), we apply Lemma 4 to obtain
\[ R_{T,\mu} \leq 2C|\mu^* - \mu| \leq 2C|\overline{\mu} - \mu| \]
where \( C \) is the constant in Lemma 4. Thus \( \text{Var}R_{T,\mu} \leq 4C^2|\overline{\mu} - \mu|^2 \). Assume that the statement is valid for all \( k < n \). We will prove that it is also valid for \( k = n \). Now, let \( T \) be a bifurcating tree with \( n \) taxa. Lipton and Tarjan (1979) show that there exists an edge \( e = (I_1, I_2) \) of \( T \) such that if we reroot \( T \) at the middle of this edge, the two subtrees \( T_1 \) and \( T_2 \) stemming from \( I_1 \) and \( I_2 \) have no more than \( 2n/3 \) leaves. Let \( I \) be the middle point of edge \( e \) and \( T' \) be the tree obtained by rerooting \( T \) to \( I \). Denote by \( n_1 \) and \( n_2 \) the number of leaves of \( T_1 \) and \( T_2 \). According to Lemmas 1 and 2,
\[ \text{Var}R_{T,\mu} = \text{Var}R_{T',\mu} = \text{Var}(R_{T',\mu} | I). \]

By Lemma 3, we have \( |R_{T',\mu} - R_{T_1,\mu} - R_{T_2,\mu}| \leq 2f(\mu e) \). We apply Lemma 7 to obtain
\[ \text{Var}R_{T,\mu} = \text{Var}(R_{T',\mu} | I) \leq \omega \text{Var}(R_{T_1,\mu} + R_{T_2,\mu} | I) + \frac{8\omega}{\omega - 1} f(\mu e)^2. \]

Let \( Y_1 \) and \( Y_2 \) be the observations at the leaves of \( T_1 \) and \( T_2 \) respectively. Since \( Y_1 \) and \( Y_2 \) are independent conditional on \( I \), we have \( \text{Var}(R_{T_1,\mu} + R_{T_2,\mu} | I) = \text{Var}(R_{T_1,\mu} | I) + \text{Var}(R_{T_2,\mu} | I) \). By Lemma 2, we deduce that
\[ \text{Var}R_{T,\mu} \leq \omega[\text{Var}R_{T_1,\mu} + \text{Var}R_{T_2,\mu}] + \frac{8\omega}{\omega - 1} f(\mu e)^2. \]
Recalling that the trees $T_1$ and $T_2$ have no more than $2n/3$ leaves, using the induction hypothesis for $T_1$ and $T_2$, we have

$$
\text{Var} R_{T, \mu} \leq \omega C_\alpha \left[ n_1^\alpha + n_2^\alpha \right] + \frac{8\omega^2}{\omega - 1} \left[ n_1^\beta S_{T_1} + n_2^\beta S_{T_2} \right] + \frac{8\omega}{\omega - 1} f(\mu t_e)^2
$$

$$
\leq \omega C_\alpha \frac{2^{\alpha+1}}{3^\alpha} n^\alpha + \frac{8\omega}{\omega - 1} \left[ \omega \frac{2^\beta}{3^\beta} (S_{T_1} + S_{T_2}) + f(\mu t_e)^2 \right] n^\beta.
$$

Let $e_1$ and $e_2$ be the edges stemming from the root of $T$. By definition, we have

$$
S_{T} - [S_{T_1} + S_{T_2} + f(\mu t_e)^2] = f(\mu t_{e_1})^2 + f(\mu t_{e_2})^2 - f(\mu (t_{e_1} + t_{e_2}))^2 \geq 0
$$

where the last inequality comes from the fact that $f$ is a non-increasing function. Thus, if we choose $\omega$ such that

$$
1 < \omega \leq \min \left\{ \left( \frac{3}{2} \right)^\beta, \frac{1}{2} \left( \frac{3}{2} \right)^\alpha \right\},
$$

then

$$
\text{Var} R_{T, \mu} \leq C_\alpha n^\alpha + \frac{8\omega}{\omega - 1} n^\beta S_{T},
$$

which completes the proof.

A combination of Lemma 8 and Assumption 1 gives rise to the desired bound:

**Lemma 9** (Sub-quadratic upper bound on variance) Under Assumption 1, $\forall \alpha \in \left( \max \left\{ \frac{\log(2)}{\log(3/2)}, \gamma \right\}, 2 \right)$, there exists $C_{\alpha, \gamma} > 0$ such that

$$
\text{Var} R_{T, \mu} \leq C_{\alpha, \gamma} n^\alpha
$$

for all $n \geq 2$ and $\mu \in [\mu, \overline{\mu}]$.

**Proof** By Lemma 8 and Assumption 1, we have

$$
\text{Var} R_{T, \mu} \leq C_\alpha n^\alpha + C_{\alpha, \beta, \gamma} n^{\beta + \gamma}, \quad \forall \alpha \in \left( \frac{\log(2)}{\log(3/2)}, 2 \right), \beta \in (0, 1).
$$

Note that $1 \leq \gamma < 2$. So, $\forall \alpha \in \left( \max \left\{ \frac{\log(2)}{\log(3/2)}, \gamma \right\}, 2 \right)$, we can choose $\beta = \alpha - \gamma$.

### 3.3 Concentration bound

We are now ready to prove the concentration inequality for the 2-state symmetric model.
Lemma 10 (Concentration bound) Under Assumption 1, for any $\delta > 0$ and $\forall \alpha \in \left(\max \left\{ \frac{\log(2)}{\log(3/2)}, \gamma \right\}, 2 \right)$, there exists $C_{\delta, \alpha, \mu, \pi}$ such that

$$\left| \frac{1}{n} R_{T, \mu} - \mathbb{E} \left[ \frac{1}{n} R_{T, \mu} \right] \right| \leq \frac{C_{\delta, \alpha, \mu, \pi, \gamma}}{n^{(2-\alpha)/3}} \quad \forall \mu \in [\underline{\mu}, \overline{\mu}]$$

with probability at least $1 - \delta$.

Proof Applying Chebyshev’s inequality, we obtain

$$\mathbb{P} \left[ \left| \frac{1}{n} R_{T, \mu} - \mathbb{E} \left[ \frac{1}{n} R_{T, \mu} \right] \right| \geq \eta \right] \leq \frac{\text{Var}(R_{T, \mu})}{n^2 \eta^2}$$

for any $\mu \in [\underline{\mu}, \overline{\mu}]$. On the other hand, by Lemma 4, we have

$$\left| \frac{1}{n} R_{T, \mu_1} - \frac{1}{n} R_{T, \mu_2} \right| = \left| \frac{1}{n} \ell_{T, \mu_1} - \frac{1}{n} \ell_{T, \mu_2} \right| \leq C_{\mu, \pi} |\mu_1 - \mu_2|, \quad \forall \mu_1, \mu_2 \in [\underline{\mu}, \overline{\mu}].$$

Therefore, if

$$\left| \frac{1}{n} R_{T, \mu_0} - \mathbb{E} \left[ \frac{1}{n} R_{T, \mu_0} \right] \right| \geq \eta,$$

then

$$\left| \frac{1}{n} R_{T, \mu} - \mathbb{E} \left[ \frac{1}{n} R_{T, \mu} \right] \right| \geq \frac{\eta}{2},$$

for all

$$\mu \in \left[ \mu_0 - \frac{\eta}{4C_{\mu, \pi}}, \mu_0 + \frac{\eta}{4C_{\mu, \pi}} \right] \cap [\underline{\mu}, \overline{\mu}].$$

Define

$$\mu_k = \mu + k \frac{\eta}{4C_{\mu, \pi}}, \quad k = 1, 2, \ldots, \left\lfloor \frac{4C_{\mu, \pi} (\overline{\mu} - \mu)}{\eta} \right\rfloor.$$

We have

$$\mathbb{P} \left[ \exists \mu \in [\underline{\mu}, \overline{\mu}] : \left| \frac{1}{n} R_{T, \mu} - \mathbb{E} \left[ \frac{1}{n} R_{T, \mu} \right] \right| \geq \eta \right] \leq \mathbb{P} \left[ \bigcup_k \left\{ \left| \frac{1}{n} R_{T, \mu_k} - \mathbb{E} \left[ \frac{1}{n} R_{T, \mu_k} \right] \right| \geq \frac{\eta}{2} \right\} \right]$$
\[
\leq \sum_k \mathbb{P} \left[ \left| \frac{1}{n} R_{T,\mu_k} - \mathbb{E} \left[ \frac{1}{n} R_{T,\mu_k} \right] \right| \geq \frac{\eta}{2} \right]
\]
\[
\leq \sum_k \frac{\text{Var}(R_{T,\mu_k})}{n^2 \eta^2}
\]
\[
\leq 8C_{\mu,\bar{\mu}}(\bar{\mu} - \mu)C_{\alpha,\gamma} n^{\alpha-2} \eta^{-3}.
\]

The last inequality comes from Lemma 9. We complete the proof by picking
\[
\eta = \left( \frac{8C_{\mu,\bar{\mu}}(\bar{\mu} - \mu)C_{\alpha,\gamma}}{\delta n^{2-\alpha}} \right)^{1/3}.
\]

3.4 Proof of Theorem 1

From Lemma 6, we have
\[
C_{d, \vec{d}, \pi} n |\hat{\mu} - \mu^*|^2 \leq \mathbb{E}_{\mu^*}[R_{T,\hat{\mu}}] = \mathbb{E}_{\mu^*} \left[ \ell_{T,\mu^*} \right] - \mathbb{E}_{\mu^*} \left[ \ell_{T,\hat{\mu}} \right]
\]
where \( E_{\mu^*} \) is the expectation with respect to \( P_{T,\mu^*} \).

Note that
\[
\frac{1}{n} \ell_{T,\hat{\mu}} - \frac{1}{n} \ell_{T,\mu^*} \geq 0.
\]

By Lemma 10, with probability \( 1 - \delta \), we obtain
\[
C_{d, \vec{d}, \pi} |\hat{\mu} - \mu^*|^2 \leq \left( \frac{1}{n} \ell_{T,\hat{\mu}} - \mathbb{E}_{\mu^*} \left[ \frac{1}{n} \ell_{T,\hat{\mu}} \right] \right) - \left( \frac{1}{n} \ell_{T,\mu^*} - \mathbb{E}_{\mu^*} \left[ \frac{1}{n} \ell_{T,\mu^*} \right] \right)
\]
\[
\leq \frac{C_{\delta,\alpha,\mu,\bar{\mu},\gamma}}{n^{(2-\alpha)/3}},
\]
which completes the proof.

4 Applications

In this section, we discuss applications of Theorem 1 for several practical scenarios. We first consider trees with edges of bounded length.

Theorem 2 (Trees with bounded edges) If edge lengths of \( \mathbb{T} \) are bounded from below and above, then the MLE is consistent.

Proof Since all edge lengths are bounded from below, Assumption 1 holds with \( \gamma = 1 \). Next, we will show that Assumption 2 is satisfied. We select \( \lfloor n/2 \rfloor \) pairs of leaves using the following procedure:
1. Pick a cherry as a pair.
2. Remove the cherry from the tree as well as the edge immediately above it.
3. Repeat step 1 and 2 until the tree has 0 or 1 leaves.

Denote $C = (i_k, j_k)_{k=1}^m$ be the set of pairs returned by the procedure where $m = \lfloor n/2 \rfloor$. It is obvious that the paths connecting each pair are pairwise disjoint. Let $e$ and $\bar{e}$ be the lower and upper bound of edge lengths, so we have a lower bound for distances between each pair $d_{i_k j_k} \geq 2e$. Moreover,

$$\sum_{k=1}^m d_{i_k j_k} \leq (2n - 2)\bar{e}. \tag{3}$$

We will prove by contradiction that the set $B = \{k : d_{i_k j_k} \leq 8\bar{e}\}$ has at least $n/8$ elements. Assume that $|B| < n/8$. Note that

$$|C| = \lfloor n/2 \rfloor \geq \frac{n-1}{2}.$$

We deduce that

$$|C \setminus B| > \frac{n-1}{2} - \frac{n}{8} = \frac{3n-4}{8}.$$

Therefore

$$\sum_{k=1}^m d_{i_k j_k} \geq \sum_{k \in C \setminus B} d_{i_k j_k} > 8\bar{e} \frac{3n-4}{8} = (3n-4)\bar{e} \geq (2n-2)\bar{e}, \quad \forall n \geq 2$$

which contradicts (3). Therefore, $|B| \geq n/8$. \hfill \Box

Next, we apply our result to trees generated from a pure-birth process (Yule 1925). This is a classical tree-generating process which assumes that lineages give birth independently to one another with the same birth rate.

**Theorem 3** (Yule process) If $\mathbb{T}$ is generated from a Yule process, then the MLE is consistent.

**Proof** Again, we only need to check Assumptions 1 and 2. Let $(t_k)_{k=1}^n$ be the amount of time during which $\mathbb{T}$ has $k$ lineages. Then, $(t_k)_{k=1}^n$ are independent exponential random variables with rate $k\lambda$ where $\lambda$ is the birth rate of the Yule process. Let $e_{\min}$ be the minimum edge length of $\mathbb{T}$. It is sufficient for us to show that $e_{\min} \geq e^{-n(\gamma - 1)/2}$. Indeed, we have

$$\Pr(e_{\min} \geq e^{-n(\gamma-1)/2}) = \Pr(\min_k t_k \geq e^{-n(\gamma-1)/2})$$

$$= \exp\left(-\frac{n(n+1)}{2}\lambda e^{-n(\gamma-1)/2}\right)$$

\text{Springer}
which converges to 1 as \( n \to \infty \). Hence, Assumption 1 is satisfied asymptotically with \( \gamma > 1 \).

Define the age of an internal node to be its distance to the leaves. By Corollary 4 of Ané et al. (2017), there exist \( 0 < c_1 < c_2 < +\infty \) such that the probability of there being \( \Omega(n) \) internal nodes with age between \( c_1 \) and \( c_2 \) is \( 1 - \mathcal{O}(n^{-1}) \). Let \( I \) be the set of internal nodes which have age between \( c_1 \) and \( c_2 \), we select \( \Omega(n) \) pairs of leaves as follows:

1. Start with an internal node \( i \in I \) that has the smallest age. Pick two of its descendants such that \( i \) is their most recent common ancestor to form a pair.
2. Remove \( i \) and its parent node from \( I \).
3. Repeat step 1 and 2 until \( I \) is empty.

This procedure is similar to the procedure for selecting contrasts in Ané et al. (2017, Lemma 3). Hence, it selects at least \( |I|/2 \) pairs, which is of order \( \Omega(n) \). Since these internal nodes have age between \( c_1 \) and \( c_2 \), Assumption 2 is satisfied.

Finally, we consider the coalescent point process (CPP), which generates a random ultrametric tree with a given height \( T \) (see Lambert and Stadler 2013, for more details). Conditioning on the number of species \( n \), the internal node ages \( (t_k)_{k=1}^{n-1} \) of the tree generated from the CPP are independent and identically distributed according to a probability distribution in \([0, T]\). The probability density function \( \phi \) of this common distribution is called the coalescent density. We say that a CPP is regular if its common distribution is not a point mass at 0 and the coalescent density is bounded from above.

**Theorem 4** (Coalescent point process) *If \( T \) is generated from a regular coalescent point process, then the MLE is consistent.*

**Proof** Denote the minimum edge length of \( T \) by \( e_{\text{min}} \). We have

\[
e_{\text{min}} \geq \min \left\{ \frac{\min |t_i - t_j|}{n^3}, \frac{\min t_i}{n}, \frac{\min (T - t_i)}{n} \right\}.
\]

Therefore,

\[
P \left( e_{\text{min}} \leq \frac{1}{n^3} \right) \leq P \left( \min_{1 \leq i < j \leq n-1} |t_i - t_j| \leq \frac{1}{n^3} \right) + P \left( \min_{1 \leq i \leq n-1} t_i \leq \frac{1}{n^3} \right) + P \left( \min_{1 \leq i \leq n-1} (T - t_i) \leq \frac{1}{n^3} \right).
\]

By the results of Section 4 in Jammalamadaka and Janson (1986) (for which the details will be provided in the “Appendix”), we have

\[
P \left( n^2 \min_{1 \leq i < j \leq n-1} |t_i - t_j| \leq \epsilon \right) \to 1 - \exp \left( -c\epsilon \int_0^T \phi^2 \right)
\] (4)

for any \( \epsilon > 0 \). We deduce that
\[
\lim_{n \to \infty} P\left( \min_{1 \leq i < j \leq n-1} |t_i - t_j| \leq \frac{1}{n^3} \right) \leq \lim_{n \to \infty} P\left( \min_{1 \leq i < j \leq n-1} |t_i - t_j| \leq \epsilon \right) = 1 - \exp\left(-c\epsilon \int_0^T \phi^2 \right).
\]

Since \( \epsilon \) can be arbitrary small, we conclude

\[
\lim_{n \to \infty} P\left( \min_{1 \leq i < j \leq n-1} |t_i - t_j| \leq \frac{1}{n^3} \right) = 0.
\]

On the other hand

\[
P\left( \min_{1 \leq i \leq n-1} t_i \leq \frac{1}{n^3} \right) = 1 - P\left( \min_{1 \leq i \leq n-1} t_i \geq \frac{1}{n^3} \right) = 1 - \left( \int_{1/n^3}^T \phi \right)^n.
\]

Note that \( \phi \) is bounded from above by \( M \) in \([0, T]\), thus

\[
\left( \int_{1/n^3}^T \phi \right)^n = \left( 1 - \int_0^{1/n^3} \phi \right)^n \geq \left( 1 - \frac{M}{n^3} \right)^n \to 1.
\]

where \( M \) is the upper bound of \( \phi \). Hence,

\[
\lim_{n \to \infty} P\left( \min_{1 \leq i \leq n-1} t_i \leq \frac{1}{n^3} \right) = 0.
\]

Similarly,

\[
\lim_{n \to \infty} P\left( \min_{1 \leq i \leq n-1} (T - t_i) \leq t \right) = 0.
\]

Therefore,

\[
P\left( e_{\text{min}} \geq \frac{1}{n^3} \right) \to 1.
\]

Hence, Assumption 1 is satisfied.

Since the common distribution is not a point mass at 0, there exist a constant \( c > 0 \) such that \( \int_c^T \phi > 0 \). Let \( I \) be the set of internal nodes of \( \mathbb{T} \) which have age between \( c \) and \( T \). Then, Assumption 2 holds if we can prove that \( |I| \) is of order \( \Omega(n) \) because we can use the same argument as in the Proof of Theorem 3. By strong law of large numbers, we have

\[
\frac{|I|}{n-1} \to \int_c^T \phi > 0,
\]

which completes the proof. \( \square \)
5 Practical implication

The consistency property proved in Theorem 1 suggests that adding more taxa helps to significantly improve the accuracy of the MLE of the transition rate. It is worth noticing that this property does not hold in many scenarios (see Li et al. 2008; Ané 2008; Ho and Ané 2013, 2014). To illustrate our theoretical results, we perform the following simulation using the R package geiger (Harmon et al. 2007; Pennell et al. 2014). We simulate 100 trees (the number of taxa varies from 50 to 2000) according to the Yule process with birth rate $\lambda = 1$. For each tree, we simulate 100 traits under the 2-state symmetric model with the transition rate $\mu = 0.5$. For each of these 100 traits, the MLE for the transition rate is computed separately. The result is summarized in Fig. 2. We can see that the MLEs concentrate more and more around the true transition rate as the number of species increases.

6 Discussion

In this paper, we investigate the convergence of the MLE for the transition rate of a 2-state symmetric model under regularity conditions on tree shape. These conditions ensure that edge lengths of the tree are not too small and the pairwise distances between leaves are not too extreme. For example, these conditions are satisfied when the edge lengths of the tree are bounded from above and below. We have also verified that
trees generated from pure-birth process and coalescent point process satisfy these conditions. On the other hand, whether these sufficient conditions are also necessary conditions remains open.

Our results suggest that adding data in terms of the number of tips tends to improve the accuracy of the MLE of the transition rate in general. To demonstrate this result, we use simulations to confirm that the MLE of the transition rate is consistent when the tree is generated from the Yule process. It is worth noticing that for evolutionary data, the MLE for other quantities of interest in evolutionary studies can be inconsistent (Li et al. 2008; Ané 2008; Ho and Ané 2013, 2014). In these situations, adding more data is a waste of resources because it does not significantly improve the precision of the MLE. Therefore, it is important to study the consistency of the MLE in the context of trait evolution models.

Beside its practical biological application, the paper also investigates many theoretical properties of tree-generating processes. Specifically, in Theorems 3 and 4, we consider the problem of bounding the minimum edge length $e_{\min}$ of a tree generated by Yule and coalescent point processes. The theorems show that generically, $e_{\min}$ is bounded from below by a function of order $n^{-2}$ for the Yule process and $n^{-3}$ for the coalescent point process. Since the minimum edge length plays an important role in many phylogenetic problems, these results may be of independent interest.

Finally, we remark that if we have multiple independent traits which evolve according to the same 2-state Markov process, the distance from the MLE to the true value of the transition rate will decrease linearly with respect to the number of traits. This result comes from the fact that traditional properties of MLE can be applied because these traits are independent. Therefore, incorporating additional traits to the analysis is another way to improve the precision of the MLE.

A Proofs

A.1 Proof of Lemma 2

Note that by symmetry, we have $P(Y = y \mid \rho = 0) = P(1 - Y = y \mid \rho = 1)$. We deduce that

$$P(h(Y) = x \mid \rho = 0) = P(Y \in h^{-1}(x) \mid \rho = 0) = P(1 - Y \in h^{-1}(x) \mid \rho = 1) = P(h(1 - Y) = x \mid \rho = 1) = P(h(Y) = x \mid \rho = 1)$$

which completes the proof.

A.2 Proof of Lemma 3

Denote $P^{(u)}_v = P(Y_u \mid T_u, \mu, \rho_u = v)$ for $u \in \{0, 1\}$, $v \in \{0, 1\}$. We have
\[ P_{\pi_1, \mu}(Y_1) P_{\pi_2, \mu}(Y_2) = \frac{1}{4} \sum_{u, v \in \{0, 1\}} P_u^{(1)} P_v^{(2)}. \]

Moreover
\[ P_{\pi, \mu}(Y) = \frac{1 + e^{-2\mu d}}{4} \sum_{u \in \{0, 1\}} P_u^{(1)} P_u^{(2)} + \frac{1 - e^{-2\mu d}}{4} \sum_{u \in \{0, 1\}} P_u^{(1)} P_{1-u}^{(2)}. \]

Therefore
\[ \frac{1}{1 - e^{-2\mu d}} P_{\pi_1, \mu}(Y_1) P_{\pi_2, \mu}(Y_2) \leq P_{\pi, \mu}(Y) \leq 2 P_{\pi_1, \mu}(Y_1) P_{\pi_2, \mu}(Y_2). \]

### A.3 Proof of Lemma 4

Without loss of generality, we assume that \( \mu_1 < \mu_2 \). By the mean value theorem, there exists \( \tilde{\mu}_{uv} \in (\mu_1, \mu_2) \) for any \( u, v \in \{0, 1\} \) such that
\[ |\log([P_{\mu_1}(t)]_{uv} - \log([P_{\mu_2}(t)]_{uv}| = \frac{t e^{-2\tilde{\mu}_{uv} t}}{[P_{\tilde{\mu}_{uv}(t)]_{uv}} |\mu_1 - \mu_2| \leq \frac{te^{-2\tilde{\mu}_{uv} t}}{1 - e^{-2\tilde{\mu}_{uv} t}} |\mu_1 - \mu_2|. \]

We observe that there exists a \( C_{\mu, \tilde{\mu}} > 0 \) such that
\[ \sup_{t \geq 0; \tilde{\mu}_{uv} \in (\mu, \tilde{\mu})} \frac{te^{-2\tilde{\mu}_{uv} t}}{1 - e^{-2\tilde{\mu}_{uv} t}} \leq C_{\mu, \tilde{\mu}}. \]

Therefore,
\[ |\log([P_{\mu_1}(t)]_{uv} - \log([P_{\mu_2}(t)]_{uv}| \leq C_{\mu, \tilde{\mu}} |\mu_1 - \mu_2|. \]

This implies that
\[ [P_{\mu_1}(t)]_{uv} \leq e^{C_{\mu, \tilde{\mu}} |\mu_1 - \mu_2|} [P_{\mu_2}(t)]_{uv}. \] (5)

Note that
\[ P_{\pi, \mu}(Y) = \frac{1}{2} \sum_{y} \left( \prod_{(u, v) \in E} [P_{\mu}(d_{uv})]_{y_u y_v} \right). \]

By applying (5) for all \( 2n - 3 \) edges on the tree, we deduce that
\[ P_{\pi_1, \mu}(Y) \leq e^{(2n-3)C_{\mu, \tilde{\mu}} |\mu_1 - \mu_2|} P_{\pi_2, \mu}(Y). \]
Hence,
\[ |\ell_{T,\mu_1}(Y) - \ell_{T,\mu_2}(Y)| \leq (2n - 3)C_{\mu_1}\mu_1|\mu_1 - \mu_2|, \]
which validates the lemma.

**A.4 Proof of Lemma 7**

For all \(x, y\), we have \(|u(x) - u(y)| \leq |v(x) - v(y)| + 2c\). Let \(Y\) be an independent and identically distributed copy of \(X\), we have
\[
2\text{Var}[u(X)] = \text{E}[u(X)^2] + \text{E}[u(Y)^2] - 2\text{E}[u(X)]\text{E}[u(Y)]
\]
\[
= \text{E}_{X,Y} \left( u(X)^2 + u(Y)^2 - 2u(X)u(Y) \right)
\]
\[
= \text{E}_{X,Y} \left( [u(X) - u(Y)]^2 \right)
\]
\[
\leq \text{E}_{X,Y} \left( [|v(X) - v(Y)| + 2c]^2 \right).
\]

Note that for all \(z, c \in \mathbb{R}\) and \(\omega > 1\),
\[
(z + 2c)^2 \leq \omega z^2 + \frac{4\omega}{\omega - 1}c^2.
\]

Therefore,
\[
2\text{Var}[u(X)] \leq \omega \text{E}_{X,Y} \left( [|v(X) - v(Y)|]^2 \right) + \frac{4\omega}{\omega - 1}c^2
\]
\[
= 2\omega \text{Var}[v(X)] + \frac{4\omega}{\omega - 1}c^2.
\]

**A.5 Proof of Eq. (4)**

In order to establish Eq. (4), we use the following Lemma.

**Lemma 11** (Remark 3.4 in Jammalamadaka and Janson (1986)) Let \(X_1, X_2, \ldots, X_n\) be an i.i.d. sequence of random variables and \(f_n(x, y)\) be an indicator function on \(\mathbb{R}^2\) such that
\[
n^3\text{E}[f_n(X_1, X_2)f_n(X_1, X_3)] \to 0 \quad \text{and} \quad \frac{1}{2}n^2\text{E}[f_n(X_1, X_2)] \to \lambda
\]
for some constant \(\lambda > 0\). Define \(U_n = \sum_{1 \leq i < j \leq n} f_n(X_i, X_j)\).

Then \(U_n \to_d \text{Poisson}(\lambda)\).

We apply this Lemma with \(f_n(x, y) = I[|x - y| < r_n]\) where \(r_n = \epsilon/n^2\) for the sequence \(t_1, t_2, \ldots, t_n\) of the coalescent point process. Note that by Equation (4.3) in Jammalamadaka and Janson (1986),
\[
\frac{1}{2} n^2 E[f_n(t_1, t_2)] \rightarrow c \epsilon \int_0^T \phi(x)^2 \, dx
\]

for some constant \( c > 0 \). On the other hand, we have

\[
E[f_n(t_1, t_2) f_n(t_1, t_3)] = E[(E[f_n(t_1, t_2) | t_1])^2]
\]

\[
= \int_0^T \left( \int_{t-r_n}^{t+r_n} \phi(\tau) \, d\tau \right)^2 \phi(t) \, dt \leq \frac{4 \| \phi \|_\infty^2 \epsilon^2}{n^4}.
\]

Therefore, \( n^3 E[f_n(t_1, t_2) f_n(t_1, t_3)] \rightarrow 0 \). Hence,

\[
P\left( n^2 \min_{1 \leq i < j \leq n-1} |t_i - t_j| \leq \epsilon \right) = P(U_n = 0) \rightarrow 1 - \exp \left( -c \epsilon \int_0^T \phi^2 \right).
\]

References

Ané C (2008) Analysis of comparative data with hierarchical autocorrelation. Ann Appl Stat 2(3):1078–1102
Ané C, Ho LST, Roch S (2017) Phase transition on the convergence rate of parameter estimation under an Ornstein–Uhlenbeck diffusion on a tree. J Math Biol 74(1–2):355–385
Felsenstein J (1981) Evolutionary trees from gene frequencies and quantitative characters: finding maximum likelihood estimates. Evolution 35(6):1229–1242
Felsenstein J (1985) Phylogenies and the comparative method. Am Nat 125(1):1–15
Harmon LJ, Weir JT, Brock CD, Glor RE, Challenger W (2007) GEIGER: investigating evolutionary radiations. Bioinformatics 24(1):129–131
Ho LST, Ané C (2013) Asymptotic theory with hierarchical autocorrelation: Ornstein–Uhlenbeck tree models. Ann Stat 41(2):957–981
Ho LST, Ané C (2014) Intrinsic inference difficulties for trait evolution with Ornstein–Uhlenbeck models. Methods Ecol Evol 5(11):1133–1146
Jammalamadaka SR, Janson S (1986) Limit theorems for a triangular scheme of U-statistics with applications to inter-point distances. Ann Probab 14(4):1347–1358
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
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
Lipton RJ, Tarjan RE (1979) A separator theorem for planar graphs. SIAM J Appl Math 36(2):177–189
Mooers A, Schluter D (1999) Reconstructing ancestor states with maximum likelihood: support for one- and two-rate models. Syst Biol 48(3):623–633
Mossel E, Steel M (2014) Majority rule has transition ratio 4 on Yule trees under a 2-state symmetric model. J Theor Biol 360:315–318
Pennell MW, Eastman JM, Slater GJ, Brown JW, Uyeda JC, FitzJohn RG, Alfaro ME, Harmon LJ (2014) geiger v2.0: an expanded suite of methods for fitting macroevolutionary models to phylogenetic trees. Bioinformatics 30(15):2216–2218
Sagitov S, Bartoszek K (2012) Interspecies correlation for neutrally evolving traits. J Theor Biol 309:11–19
tuffley C, Steel M (1997) Links between maximum likelihood and maximum parsimony under a simple model of site substitution. Bull Math Biol 59(3):581–607
Van Erven T, Harremoës P (2014) Rényi divergence and Kullback–Leibler divergence. IEEE Trans Inf Theory 60(7):3797–3820
Yule GU (1925) A mathematical theory of evolution, based on the conclusions of Dr. JC Willis, FRS. Philos Trans R Soc Lond Ser B 213:21–87
Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Affiliations

Lam Si Tung Ho¹ · Vu Dinh² · Frederick A. Matsen IV³ · Marc A. Suchard⁴

Lam Si Tung Ho
Lam.Ho@dal.ca

Vu Dinh
vucdinh@udel.edu

Frederick A. Matsen IV
ematsen@gmail.com

Marc A. Suchard
msuchard@ucla.edu

1 Department of Mathematics and Statistics, Dalhousie University, Halifax, NS, Canada
2 Department of Mathematical Sciences, University of Delaware, Newark, USA
3 Program in Computational Biology, Fred Hutchinson Cancer Research Center, Seattle, USA
4 Departments of Biomathematics, Biostatistics and Human Genetics, University of California, Los Angeles, USA