Ancestral state reconstruction with large numbers of sequences and edge-length estimation

Lam Si Tung Ho · Edward Susko

Received: 31 March 2021 / Revised: 18 October 2021 / Accepted: 3 January 2022 / Published online: 21 February 2022
© The Author(s), under exclusive licence to Springer-Verlag GmbH Germany, part of Springer Nature 2022

Abstract
Likelihood-based methods are widely considered the best approaches for reconstructing ancestral states. Although much effort has been made to study properties of these methods, previous works often assume that both the tree topology and edge lengths are known. In some scenarios the tree topology might be reasonably well known for the taxa under study. When sequence length is much smaller than the number of species, however, edge lengths are not likely to be accurately estimated. We study the consistency of the maximum likelihood and empirical Bayes estimators of the ancestral state of discrete traits in such settings under a star tree. We prove that the likelihood-based reconstruction is consistent under symmetric models but can be inconsistent under non-symmetric models. We show, however, that a simple consistent estimator for the ancestral states is available under non-symmetric models. The results illustrate that likelihood methods can unexpectedly have undesirable properties as the number of sequences considered gets very large. Broader implications of the results are discussed.

Keywords Ancestral state reconstruction · Maximum likelihood estimator · Empirical Bayes estimator · Consistency · Evolution · Phylogenetics

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. ES was supported by a Discovery Grant awarded by the Natural Sciences and Engineering Research Council of Canada.

✉ Lam Si Tung Ho
Lam.Ho@dal.ca

Edward Susko
Edward.Susko@dal.ca

1 Department of Mathematics and Statistics, Dalhousie University, Halifax, Nova Scotia, Canada
Ancestral state reconstruction is an important problem in evolutionary biology (Madison 1994; Felsenstein 2004; Liberles 2007). Reconstructing the ancestral states helps answer many questions about macroevolution including the evolution of phenotypes (Finarelli and Flynn 2006; Odom et al. 2014) and the origin of epidemics (Lemey et al. 2009; Faria et al. 2014; Gill et al. 2017). Ancestral reconstruction has also been used to determine which types of substitutions frequently occur in pseudogenes (Gojobori et al. 1982) and to study the optimal growth temperature of the Last Universal Common Ancestor (Gaucher et al. 2003). More broadly, the degree to which ancestral sequences can be accurately reconstructed is indirectly related to the efficiency of tree reconstruction methods, molecular dating and inference about adaptive evolution in molecular settings. This is because such methods, in effect, consider weighted averages of probabilities over ancestral sequences.

With a relatively large number of sites and a small to moderate number of sequences, edge lengths and even tree topology can be estimated accurately from sequence data. For this reason and for simplicity, previous studies on the theory of ancestral state reconstruction methods often assume that the tree topology and edge lengths are known (Ané 2008; Royer-Carenzi et al. 2013; Fan and Roch 2018; Ho et al. 2019). However, large numbers of sequences are increasingly available for a wide variety of species, and phylogenies based on hundreds or thousands of taxa are becoming commonplace. Increasing the number of taxa gives more information but to avoid having large amounts of missing data, the number of characters considered often needs to be kept small. Some theoretical results are available in such settings and suggest challenges, particularly for edge-length estimation. For instance, for $n$ species, the required sequence length for accurately reconstructing the tree topology is a power of $\log n$ (Erdős et al. 1999; Erdős et al. 1999) while the required sequence length for reconstructing both the tree topology and edge lengths is a power of $n$ (Dinh et al. 2018). In such settings, treating edge-lengths as known is problematic yet little effort has been made to study the problem of ancestral state reconstruction without edge lengths, especially for likelihood-based methods. In this paper, we will focus on this problem for discrete traits.

The simplest ancestral reconstruction method for discrete traits is majority rule, which estimates the state at the root by the most frequent state appearing at the leaves. Maximum parsimony, on the other hand, utilizes the information from the tree topology. This method estimates the root value by minimizing the number of changes needed to explain the evolution of the character along the tree. Maximum parsimony can have strong biases in the presence of compositional bias, however (Collins et al. 1994; Eyre-Walker 1998). The maximum likelihood estimator (MLE) and Bayesian inference maximize the likelihood function and the posterior distribution respectively for reconstructing the ancestral state. These likelihood-based methods employ the information from both the tree topology and edge lengths. Intuitively, utilizing more information can be expected to result in a more efficient estimation method. Moreover, in standard settings, likelihood methods are known to have the optimality property of being asymptotically minimum variance among approximately unbiased estimators (Bickel and Doksum 2007, §5.4.3). Therefore, it is not surprising that they are often
considered the best approaches for ancestral state reconstruction. The setting considered here however is non-standard at least in that the number of parameters increases as the number of taxa increases.

Consistency is often considered a base criterion for judging whether an estimation method is good or not. An estimator is consistent if it converges to the true value as the number of observations increases to infinity. In the present setting, consistency arises if we can recover the true ancestral state when we have an infinite number of species. When the tree topology and edge lengths are known, Fan and Roch (2018) provides a necessary and sufficient condition, called “big bang”, for the existence of a consistent estimator of the ancestral state on bounded-height trees. It is worth noticing that a direct consequence of Proposition 6 in Steel and Rodrigo (2008) is that the MLE is consistent if there exists a consistent estimator for the ancestral state (assuming that the evolution model is known). This result confirms that the MLE is a reasonable estimator of the ancestral state in this scenario and that the “big bang” condition is a necessary and sufficient condition for the consistency of the MLE.

A natural hypothesis is that likelihood-based methods are also the best ancestral state reconstruction methods when edge lengths are unknown. To investigate this, we consider a simple scenario where discrete traits evolve along a star tree according to a proportional model. Although estimation of ancestral frequencies is usually of greatest interest under non-stationary models (Susko and Roger 2013), we assume a simple stationary setting. We show that in this setting the MLE and Empirical Bayes estimator (sometimes also referred to as the Maximum A Posteriori (MAP) estimator) converge upon the same solution. Consequently, it suffices to consider the MLE. The MLE is shown to be consistent under symmetric models but there exists a zone of inconsistency under non-symmetric models. As a consequence, the “big bang” condition in Fan and Roch (2018) is no longer a sufficient condition for the consistency of the MLE. We also uncover that when the edge lengths are unknown, the MLE is not the best ancestral reconstruction method. Specifically, we present a simple new estimator for the ancestral state that is consistent under some mild conditions. We show that the MLE is not always consistent under the same conditions. Therefore, the proposed estimator is better than the MLE in this scenario.

2 Settings

Throughout this paper, we will focus on star trees whose edge lengths are unknown. A star tree is a tree such that all taxa are direct descendants of the root (Fig. 1). Let $n$ be the number of leaves and $t = (t_k)_{k=1}^n$ be the (unknown) edge lengths. In this setting, we observe a sequence of $N$ sites at each leaf. We assume that characters at these sites evolve independently along the tree according to the proportional model. That is, the evolution of characters follows a finite-state continuous-time Markov process with the following transition probabilities

$$P_{ij}(t) = \pi_j [1 - \exp(-\mu t)] + 1_{i=j} \exp(-\mu t), \quad \forall i, j \in \{1, 2, \ldots, c\}.$$ 

Here, $c$ is the number of possible states and $\pi = (\pi_1, \pi_2, \ldots, \pi_c)$ is the stationary distribution of the process. In practice, to avoid problems of confounding,
$\mu = [\sum_j \pi_j (1 - \pi_j)]^{-1}$ is used so that edge lengths are interpretable as expected numbers of substitutions. We consider the re-parameterization $s = \exp[-\mu t]$. So, the transition probabilities become

$$P_{ij}(s) = \pi_j [1 - s] + 1_{\{i=j\}} s.$$  

We are interested in reconstructing the ancestral states $\rho^*$ at the root of the tree. Without loss of generality, we can assume that $\rho^* = (1, \ldots, 1, \ldots, c, \ldots, c)$. When the edge lengths are unknown, $(\rho^*, \pi)$ are not identifiable under the proportional model (Gascuel and Steel 2020). Therefore, unless mentioned otherwise, we assume that the stationary distribution $\pi > 0$ is known. The setting is a proxy for the situation where frequencies can be estimated from a larger tree and the star portion is a local subtree of focus for ancestral reconstruction. We also make the following assumption.

**Assumption 1** Define

$$\bar{s} = \frac{1}{n} \sum_{k=1}^{n} s_k.$$  

We assume that $\lim \inf \bar{s} > 0$.

The reason for Assumption 1 is to guarantee that the majority of edge lengths are not so large ($s = 0$ corresponds to $t = \infty$) that the tip data is almost independent of the root data. A trivial scenario where Assumption 1 is satisfied is when $t_k = t < \infty$ for all $k = 1, 2, \ldots, n$. We describe a non-trivial example where these assumptions hold in the following Lemma.

**Lemma 1** Let $t_1, t_2, \ldots, t_n$ be independent and identically distributed (iid) random variables on $\mathbb{R}^+$ with finite mean. Then, the star tree with edge lengths $t = (t_k)_{k=1}^{n}$ satisfies Assumption 1 as $n$ increases to infinity.
For two sequences $\rho = (\rho_1, \rho_2, \ldots, \rho_N)$ and $y = (y_1, y_2, \ldots, y_N)$, we define

$$P_{\rho y}(s) = \prod_{l=1}^{N} P_{\rho_l y_l}(s), \quad \text{and} \quad \bar{P}_{\rho y} = \frac{1}{n} \sum_{k=1}^{n} P_{\rho y}(s_k).$$

We assume below that $\lim \bar{s} > 0$ and that $\bar{P}_{\rho^* y}$ converges to a limit, which we denote $p_{\rho^* y}$.

Note that $\bar{s}^j = \sum_{k=1}^{n} s_k^j / n = \int s^j \, dF_n(s)$ where $F_n$ is the distribution function placing weight $1/n$ on $s_1, \ldots, s_n$. Since $\bar{P}_{\rho^* y}$ is a linear transformation of the $s^j$, with fixed coefficients that do not change with $n$, its limit exists as long as the limiting $s^j$ exist. The linear transformation $\bar{P}_{\rho^* y}$ has positive coefficients. Thus this limit, $p_{\rho^* y}$, is also positive.

The assumption that limits exist is made without loss of generality. The Helly subsequence principle implies that there exists a subsequence of $F_n$ that converges, implying that any linear combination of $s^j$ converges. Consequently, there always exists a subsequence for which limits of $\bar{s}$ and $\bar{P}_{\rho^* y}$ exist. Thus if inconsistency is shown assuming limits exist, then it applies to this convergent subsequence. Inconsistency for a subsequence is sufficient to meet the definition of inconsistency.

Suppose now that consistency has been shown assuming that limits exist. We argue that inconsistency in such a setting gives a contradiction. If $\hat{\rho}$ is inconsistent, then there must be a subsequence, $\{n_k\}$, such that $\hat{\rho} \neq \rho^*$ with probability that does not converge to 0 as $k \to \infty$. But as discussed above, there will always exist a further subsequence for which the limits of $\bar{s}$ and $\bar{P}_{\rho^* y}$ exist. Since consistency applies to this further subsequence, then for this further subsequence $\hat{\rho} = \rho^*$ with probability converging to 1. But this contradicts that the probability of $\hat{\rho} \neq \rho^*$ does not converge to 0 in the first subsequence.

3 Ancestral state reconstruction

The majority rule and maximum parsimony reconstruction methods have been studied extensively in the scenario where edge lengths are unknown because these methods do not take into account this information (Maddison 1995; Gascuel and Steel 2010; Mossel and Steel 2014; Herbst and Fischer 2018; Herbst et al. 2019). On the other hand, little is known about asymptotic properties of likelihood-based methods in this scenario.

3.1 Maximum likelihood estimator

Given its good performance in a wide variety of settings, it a reasonable hypothesis that the MLE is a good method for reconstructing the ancestral state. Since the edge lengths are unknown, we need to estimate both the ancestral states and edge lengths jointly. Let $y_k$ be the sequence observed at the $k$-th leaf. The MLE is defined by
\[ \hat{\rho} = \arg\max_\rho \left( \max_s \sum_{k=1}^n \log P_{\rho y_k}(s_k) \right). \]

For an \( N \)-dimensional pattern \( y \), let \( n_y \) be the number of times \( y \) is observed at the leaves. That is,

\[ n_y = \sum_{k=1}^n 1_{\{y_k = y\}}. \]

We define

\[ \ell(\rho) = \sum_{k=1}^n \log P_{\rho y_k}(\hat{s}(\rho, y_k)) = \sum_y n_y \log P_{\rho y}(\hat{s}(\rho, y)) \]

where the sum is over all possible patterns, and \( \hat{s}(\rho, y) = \arg\max_s P_{\rho y}(s) \). Then, we have

\[ \hat{\rho} = \arg\max_\rho \frac{1}{n} \ell(\rho). \]

Then

\[ \mathbb{E} \left[ \frac{1}{n} \ell(\rho) \right] = \sum_y \tilde{P}_{\rho^* y} \log P_{\rho y}(\hat{s}(\rho, y)) \to \sum_y p_{\rho^* y} \log P_{\rho y}(\hat{s}(\rho, y)). \tag{1} \]

Here, we consider the sequence \( (s_k)_{k=1}^n \) as deterministic. We define

\[ e(\rho) = \sum_y p_{\rho^* y} \log P_{\rho y}(\hat{s}(\rho, y)). \]

The limit (1) prompts an immediate question: Is there a connection between the MLE \( \hat{\rho} \) and the function \( e(\rho) \)? The answer is yes. When the number of leaves is large, the MLE is a maximizer of \( e(\rho) \) with high probability. Specifically,

**Lemma 2** Let \( \mathcal{H} \) be the set of maximum points of \( e(\rho) \). Then, under Assumption 1,

\[ \lim_{n \to \infty} \Pr(\hat{\rho} \in \mathcal{H}) = 1. \]

The proof of Lemma 2 can be found in the Appendix A. The main idea is showing that \( \text{Var}[\ell(\rho)/n] \to 0 \). Thus,

\[ \frac{1}{n} \ell(\rho) \to_p \lim_n \mathbb{E} \left[ \frac{1}{n} \ell(\rho) \right] = e(\rho). \]
Since \( \mathcal{H} \) is the set of maximum points of \( e(\rho) \), for any \( \rho \notin \mathcal{H} \) and \( \rho_M \in \mathcal{H}, \ell(\rho_M)/n - \ell(\rho)/n > 0 \) with probability converging to 1. Therefore, \( \hat{\rho} \in \mathcal{H} \) with probability converging to 1.

A direct consequence of Lemma 2 is that if \( \rho^* \notin \mathcal{H} \), then the MLE is inconsistent. Another one is that if \( e(\rho) \) has a unique maximizer \( \rho_M \), then the MLE of the ancestral sequence converges to \( \rho_M \).

### 3.2 Empirical Bayes estimator

The empirical Bayes estimator is more widely used in practice than the MLE. For estimation of edge-lengths, it uses the more conventional likelihood that averages over the unobserved sequences at internal nodes:

\[
\hat{s} = \arg\max_s \sum_{\rho} P_\pi(\rho) \prod_{k=1}^n P_{\rho y_k}(s_k)
\]

where \( P_\pi \) is the probability according to the stationary distribution \( \pi \). The estimator of the ancestral sequence is then obtained as the sequence having the largest conditional probability, given the data and calculated using the estimated edge-lengths:

\[
\hat{\rho}_B = \arg\max_{\rho} P(\rho \mid (y_k)_{k=1}^n, \hat{s}) = \arg\max_{\rho} P_\pi(\rho) \prod_{k=1}^n P_{\rho y_k}(\hat{s}_k).
\]

Here, the empirical Bayes estimator estimates edge lengths first, then we estimate ancestral states conditional on these estimated values. Will this make any difference? A short answer is no. Just like the MLE, the empirical Bayes estimator will eventually be in the set \( \mathcal{H} \) of maximizers of \( e(\rho) \).

**Theorem 1** *Under Assumption 1,*

\[
\lim_{n \to \infty} \Pr(\hat{\rho}_B \in \mathcal{H}) = 1.
\]

A detailed proof is in Appendix A. In brief, we show that the ratio of the posterior for \( \rho \) to the posterior for \( \rho_M \) converges to 0 if \( \rho \notin \mathcal{H} \); here \( \rho_M \), which depends on the observed data, gives the largest \( \ell(\rho) \) among elements of \( \mathcal{H} \). This posterior ratio can be expressed as \( T_1/T_2 \) where \( T_1 \) and \( T_2 \) are defined below, with

\[
T_1 = \frac{P_\pi(\rho) \prod_{k=1}^n P_{\rho y_k}(\hat{s}_k)}{P_\pi(\rho_M) \exp[\ell(\rho_M)]} \leq \frac{P_\pi(\rho)}{P_\pi(\rho_M)} \exp\left\{-n \frac{\ell(\rho_M)}{n} - \frac{\ell(\rho)}{n}\right\} \to 0.
\]

The inequality comes from the fact that the edge-lengths, \( \hat{s}_k \), maximizing the conventional likelihood used for empirical Bayes give smaller \( \prod_{k=1}^n P_{\rho y_k}(\hat{s}_k) \) than the edge-lengths that give \( \ell(\rho) \). Using that property and that \( \hat{s}_k \) gives the largest conven-
tional likelihood, \( \sum_\rho P_\pi(\rho) \prod_{k=1}^n P_{\rho y_k}(s_k) \), we show after some manipulation that

\[
T_2 = \frac{P_\pi(\rho_M) \prod_{k=1}^n P_{\rho_M y_k}(\hat{s}_k)}{P_\pi(\rho_M) \exp[\ell(\rho_M)]} \geq 1 - \sum_{\rho \neq \rho_M} \frac{P_\pi(\rho)}{P_\pi(\rho_M)} \exp \left\{ -n \left[ \frac{\ell(\rho_M)}{n} - \frac{\ell(\rho)}{n} \right] \right\} \rightarrow_p 1.
\]

Thus the posterior ratio \( T_1 / T_2 \rightarrow_p 0 \). Since \( \hat{\rho}_B \) maximizes the posterior with probability converging to 1, it must be the case that \( \hat{\rho}_B \in \mathcal{H} \) with probability converging to 1.

A consequence of Theorem 1 is that both the empirical Bayes estimator and the MLE will be consistent if \( \mathcal{H} = \{ \rho^* \} \) and will be inconsistent if \( \rho^* \notin \mathcal{H} \). Since the consistency property of these estimators is similar, we will focus on the MLE from now on.

### 3.3 Consistency of the MLE

For a fixed ancestral sequence \( \rho \) and an observed sequence \( y \), let \( n_{\rho y}(i, j) \) be the number of times the putative ancestral state was \( \rho_l = i \) with corresponding observed value equal to \( y_l = j \). We have

\[
\log P_{\rho y}(s) = \sum_{i=1}^c \sum_{j=1}^c n_{\rho y}(i, j) \log P_{ij}(s) = \sum_{i=1}^c n_{\rho y}(i, i) \log[\pi_i + (1 - \pi_i)s] + \sum_{i=1}^c \sum_{j \neq i} n_{\rho y}(i, j) \log[\pi_j(1 - s)]
\]

\[
= \sum_{i=1}^c n_{\rho y}(i, i) \log[\pi_i + (1 - \pi_i)s] + \left[ N - \sum_{i=1}^c n_{\rho y}(i, i) \right] \log(1 - s) + \sum_{i=1}^c \sum_{j=1}^c n_{\rho y}(i, j) \log(\pi_j) - \sum_{i=1}^c n_{\rho y}(i, i) \log(\pi_i).
\]

The third term depends on \( \rho \) and \( y \) through \( \sum_i n_{\rho y}(i, j) \), which is the number of \( y_l = j \). Since it depends on \( y \) alone, letting \( C(y) = \sum_{i=1}^c \sum_{j=1}^c n_{\rho y}(i, j) \log(\pi_j) \), we have that

\[
\log P_{\rho y}(s) = \sum_{i=1}^c n_{\rho y}(i, i) \log[\pi_i + (1 - \pi_i)s] + \left[ N - \sum_{i=1}^c n_{\rho y}(i, i) \right] \log(1 - s)
\]

\[
- \sum_{i=1}^c n_{\rho y}(i, i) \log(\pi_i) + C(y).
\]

Maximizing \( \log P_{\rho y}(s) \), we obtain:

\[\text{Springer}\]
Lemma 3  

If

\[ \sum_{i=1}^{c} \frac{n_{\rho}(i, i)}{\pi_i} \leq N, \]

then \( \hat{s}(\rho, y) = 0 \).

- If

\[ \sum_{i=1}^{c} n_{\rho}(i, i) = N, \]

then \( \hat{s}(\rho, y) = 1 \).

- Otherwise,

\[ \sum_{i=1}^{c} \frac{n_{\rho}(i, i)}{\pi_i + (1 - \pi_i)\hat{s}(\rho, y)} = N. \]

First, let us consider symmetric models, that is \( \pi_1 = \pi_2 = \ldots = \pi_c = 1/c \). In this scenario, the MLE for the ancestral states is consistent. It is worth noticing that under symmetric models, the MLE and the ancestral state reconstructed by Maximum parsimony are the same when there is only 1 site (\( N = 1 \)) (Tuffley and Steel 1997).

Theorem 2  

Suppose that Assumption 1 holds. For symmetric models, we have

\[ \lim_{n \to \infty} \Pr(\hat{\rho} = \rho^*) = 1. \]

A detailed proof is in Appendix A. In brief, we first argue that for the symmetric model, up to an additive constant that does not depend on \( \rho \), the likelihood contribution for sequence \( k \) depends on \( X(k) \), a vector indicating whether the ancestral character states match (\( X(k)_i = 1 \)) or not (\( X(k)_i = 0 \)). Let \( g(X(k)) \) denote this contribution, then

\[ E[\ell(\rho)/n] = \frac{1}{n} \sum_{k=1}^{n} E[g(X(k))] + C = \frac{1}{n} \sum_{k=1}^{n} h_\rho(s_k) + C \]

where \( C \) is a constant and the second expression for \( E[g(X(k))] \) is used to emphasize dependence on \( \rho \) and the transformed edge-length \( s_k \).

We then show that \( g(x) \) satisfies a sort of monotonicity property: \( g(x) - g(x_-) \geq 0 \), where \( x_- = [x_1, \ldots, x_{j-1}, 0, x_{j+1}, \ldots, x_c] \), with strict inequality for most \( x \). This property allows us to use a property of Bernoulli expectations (Lemma 5) to show that if \( \rho \neq \rho^* \), then \( h_{\tilde{\rho}}(s) - h_\rho(s) > 0 \) for any \( s \in (0, 1) \) whenever \( \tilde{\rho} \) changes one of the \( \rho_j \) to \( \rho^*_j \). Consequently

\[ E[\ell(\tilde{\rho})/n] - E[\ell(\rho)/n] = \frac{1}{n} \sum_{k=1}^{n} [h_{\tilde{\rho}}(s_k) - h_\rho(s_k)] > 0 \]
Because \( \lim \tilde{s} = 0 \), a positive proportion of \( s_k \) are expected to be bounded away from 0. This allows us to conclude that the inequality holds even in the limit. Thus \( \hat{\rho} \) can be found with \( e(\hat{\rho}) - e(\rho) > 0 \) whenever \( \rho \neq \rho^* \). Since \( \hat{\rho} \) is a maximizer of \( e(\rho) \) for all large \( n \), it must be the case that \( \hat{\rho} = \rho^* \) with probability converging to 1.

Theorem 2 suggests that we may be able to reconstruct the ancestral states with high accuracy even when the edge lengths are unknown. This result aligns with similar findings in Gascuel and Steel (2010). Since the MLE is consistent under symmetric models, would it also work well under non-symmetric models? Unfortunately, it is not the case. To see this, let us consider a single site scenario (i.e. \( N = 1 \)). Suppose \( \rho^* = \rho \). For any \( \rho = a \), we have

\[
\hat{s}(a, y) = \begin{cases} 
1 & \text{if } y = a \\
0 & \text{if } y \neq a,
\end{cases} \quad \text{and } \quad \log P_{\rho y}(\hat{s}(a, y)) = \begin{cases} 
0 & \text{if } y = a \\
\log(\pi_y) & \text{if } y \neq a,
\end{cases}
\]

By (1), we have

\[
e(a) = \sum_{y \neq a} p_{ry} \log(\pi_y) = \left( \sum_y p_{ry} \log(\pi_y) \right) - p_{ra} \log(\pi_a)
\]

By Lemma 2, the MLE is inconsistent if

\[
e(\rho) < e(a) \iff p_{ra} \log(\pi_a) < p_{rr} \log(\pi_r)
\]

\[
\iff \lim_n [1 - \tilde{s}] \pi_a \log(\pi_a) < \lim_n [\pi_r + (1 - \pi_r)\tilde{s}] \log(\pi_r)
\]

\[
\iff \lim_n \tilde{s} < \frac{\pi_r \log(\pi_r) - \pi_a \log(\pi_a)}{\pi_r \log(\pi_r) - \pi_a \log(\pi_a) - \log(\pi_r)}.
\]

It is worth noticing that (3) cannot hold under symmetric models (\( \pi_r = \pi_a \)), which aligns with Theorem 2. On the other hand, when \( \pi_r \log(\pi_r) > \pi_a \log(\pi_a) \), (3) holds when \( \lim_n \tilde{s} \) is sufficiently small. Since the function \( \pi \log(\pi) \) is minimized at \( \pi = e^{-1} \), there exists a zone of inconsistency of the MLE unless \( \pi_r = e^{-1} \). We generalize this argument to obtain the zone of inconsistency for the case when we have more than 1 site.

**Theorem 3** Assume that \( \rho^* = (r, \ldots, r) \). Then, there exists \( \lim_n \tilde{s} > 0 \) and \( \pi \) such that the MLE is inconsistent. A sufficient condition for inconsistency is that \( v(\pi_a) > v(\pi_r) \), where

\[
v(p) = E \left[ \left\{ \hat{p} \log \left( \frac{\hat{p}}{p} \right) + (1 - \hat{p}) \log \left( \frac{1 - \hat{p}}{1 - p} \right) \right\} I\{\hat{p} > p\} \right],
\]

calculated with \( \hat{p} = X/N \) and \( X \sim \text{binomial}(N, p) \).

A detailed proof is in Appendix A. In brief, by Lemma 2, a sufficient condition for inconsistency is \( \rho^* \notin \mathcal{H} \). So, we seek a condition that there exists \( \rho \) such that
For simplicity, we focus on the scenario where \( \rho^* = (r, \ldots, r) \) and \( \rho = (a, \ldots, a) \). We show that \( [e(\rho) - e(\rho^*)]/N \) simplifies to

\[
\lim_{n \to \infty} \frac{1}{n} \sum_{k=1}^{n} \{w(P_{\rho a}(s_k), \pi_a) - w(P_{\rho r}(s_k), \pi_r)\}
\]

where

\[
w(p, \pi) = E \left\{ \hat{p} \log \left( \frac{\hat{p}}{\pi} \right) + (1 - \hat{p}) \log \left( \frac{1 - \hat{p}}{1 - \pi} \right) \right\} I\{\hat{p} > \pi\}.
\]

When \( s_k = 0 \) for all \( k \) this becomes

\[
\frac{e(\rho) - e(\rho^*)}{N} = v(\pi_a) - v(\pi_r).
\]

Thus if \( v(\pi_a) - v(\pi_r) > 0 \), by continuity there exists \( s_k \) sufficiently small that \( e(\rho) - e(\rho^*) > 0 \).

Figure 2 visualizes the zone of inconsistency described in Theorem 3. Perhaps surprisingly, with a small number of sites, a preferred alternative ancestral state can have a smaller stationary frequency than the true ancestral state. As the sequence length gets larger, however, the region approaches the region for parsimony where the alternative stationary frequency is larger than that of the true ancestral state.

### 3.4 A simple consistent estimator

Although Theorem 3 seems to eliminate our hope of reconstructing the ancestral states accurately under non-symmetric models, we note that this negative result was only shown to apply to the case when the ancestral state is the same for all sites. However, such a case is rare in practice when the number of sites is large. Assuming that there are at least two distinct ancestral states, we propose a simple estimator that can estimate the ancestral states consistently. Let \( y_{kl} \) be the character state at site \( l \) of the \( k \)-th species. For \( j \in \{1, 2, \ldots, c\} \), denote

\[
\hat{\pi}_{jl} = \frac{1}{n} \sum_{k=1}^{n} I\{y_{kl} = j\}, \quad V(j, l) = \hat{\pi}_{jl} - \sum_{v \neq j} \hat{\pi}_{jv} = \frac{\hat{\pi}_{jl}}{N - 1} - \sum_{v \neq j} \hat{\pi}_{jv}.
\]

We estimate the ancestral state at site \( l \) by

\[
\rho_{l}^{(D)} = \arg \max_j V(j, l).
\]

The motivation for this estimate is that for the proportional model,

\[
E[\hat{\pi}_{jl}] = \frac{1}{n} \sum_{k=1}^{n} (\pi_j - \pi_j s_k + I\{\rho_{l}^* = j\} s_k) = \pi_j - \pi_j \bar{s} + I\{\rho_{l}^* = j\} \bar{s}.
\]  (4)
Fig. 2 Areas in black indicate frequencies of pairs of true ancestral character states and alternative character states for which there exists \( \lim \tilde{r} > 0 \) such that the alternative state is preferred to the true ancestral state. It follows that the observed frequency of \( \rho_l^* \) at site \( l \) is expected to be larger than the average observed frequency of \( \rho_l^* \) over other sites, unless they all have \( \rho_l^* \) as their ancestral character state.

**Theorem 4** If Assumption 1 holds, \( N = O(\sqrt{n/\log n}) \), and there are at least two distinct ancestral states (that is, \( \rho_l^* \neq \rho_1^* \) for some site \( l \)). Then,

\[
\Pr(\rho_l^{(D)} = \rho^*) \to 1.
\]

A natural question is whether this Theorem also holds for the MLE. Interestingly, this is not the case. In other words, there exists a zone of inconsistency for the MLE. Hence, the MLE is not the best ancestral reconstruction method when edge lengths are unknown and the number of sites is small relative to the number of sequences.
To see this, let us consider the situation when the true ancestral state $\rho^*$ has the form $(\rho^*_1, \rho^*_2)$ such that $\rho^*_1 \neq \rho^*_2$. Again, we want to investigate the set of maximum points of $e(\rho)$. For this simple scenario, we can derive an analytic formula for $E[\ell(\rho_1, \rho_2)/n]$.

**Lemma 4** Denote

$$f(\pi) = I[\pi < 1/2] \log[4\pi(1 - \pi)].$$

We have

$$E \left[ \frac{\ell(\rho_1, \rho_2)}{n} \right] = C_0 - \tilde{P}_{\rho_1^*\rho_1} f(\pi_{\rho_1}) - \tilde{P}_{\rho_2^*\rho_2} f(\pi_{\rho_2}) + \tilde{P}_{(\rho_1^*, \rho_2^*)}(\rho_{\rho_1}, \rho_{\rho_2}) \{ f(\pi_{\rho_1}) + f(\pi_{\rho_2}) - \log[\pi_{\rho_1}\pi_{\rho_2}] \}$$

where

$$C_0 = \sum_x \{ \tilde{P}_{\rho_1^*x} + \tilde{P}_{\rho_2^*x} \} \log[\pi_x]$$

is independent of $(\rho_1, \rho_2)$.

A sufficient condition for inconsistency of the MLE is that there exists $(\rho_1, \rho_2)$ such that $e(\rho_1, \rho_2) > e(\rho_1^*, \rho_2^*)$. First, we will apply Lemma 4 to derive the formula for

$$Q_{\rho_1, \rho_2}(s) := E \left[ \frac{1}{n} \ell(\rho_1, \rho_2) \right] - E \left[ \frac{1}{n} \ell(\rho_1^*, \rho_2^*) \right].$$

We define $\delta_{xy} = 1_{\{x = y\}}$. For the proportional model, $P_{xy}(s) = (\delta_{xy} - \pi_y)s + \pi_y$, so

$$\tilde{P}_{(\rho_1^*, \rho_2^*)}(\rho_{\rho_1}, \rho_{\rho_2}) = (\delta_{\rho_1^*\rho_1} - \pi_{\rho_1})(\delta_{\rho_2^*\rho_2} - \pi_{\rho_2})\bar{s}^2 + \{ (\delta_{\rho_1^*\rho_1} - \pi_{\rho_1})\pi_{\rho_2} + (\delta_{\rho_2^*\rho_2} - \pi_{\rho_2})\pi_{\rho_1} \} s + \pi_{\rho_1}\pi_{\rho_2}$$

where

$$\bar{s}^2 = \frac{1}{n} \sum_{k=1}^{n} s_k^2.$$


\[ B(\rho_1, \rho_2) = - (\delta_{\rho_1} \rho_1 - \pi_{\rho_1}) f (\pi_{\rho_1}) - (\delta_{\rho_2} \rho_2 - \pi_{\rho_2}) f (\pi_{\rho_2}) + (\delta_{\rho_1} \rho_1 - \pi_{\rho_1}) \pi_{\rho_2} [f (\pi_{\rho_1}) + f (\pi_{\rho_2}) - \log (\pi_{\rho_1} \pi_{\rho_2})] + (\delta_{\rho_2} \rho_2 - \pi_{\rho_2}) \pi_{\rho_1} [f (\pi_{\rho_1}) + f (\pi_{\rho_2}) - \log (\pi_{\rho_1} \pi_{\rho_2})] \]

(7)

and

\[ C(\rho_1, \rho_2) = - \pi_{\rho_1} f (\pi_{\rho_1}) - \pi_{\rho_2} f (\pi_{\rho_2}) + \pi_{\rho_1} \pi_{\rho_2} [f (\pi_{\rho_1}) + f (\pi_{\rho_2}) - \log (\pi_{\rho_1} \pi_{\rho_2})] \]

(8)

Thus the difference in expected log likelihoods \( Q_{\rho_1\rho_2}(s) \) satisfies that

\[ Q_{\rho_1\rho_2}(s) = A_D(\rho_1, \rho_2) \bar{s}^2 + B_D(\rho_1, \rho_2) \bar{s} + C_D(\rho_1, \rho_2), \]

where

\[ A_D(\rho_1, \rho_2) = A(\rho_1, \rho_2) - A(\rho_1^*, \rho_2^*) \]

\[ B_D(\rho_1, \rho_2) = B(\rho_1, \rho_2) - B(\rho_1^*, \rho_2^*) \]

\[ C_D(\rho_1, \rho_2) = C(\rho_1, \rho_2) - C(\rho_1^*, \rho_2^*). \]

If \( Q_{\rho_1\rho_2}(0) = C_D(\rho_1, \rho_2) > 0 \), then for \( \bar{s}^2 \) sufficiently small, the expected log likelihood for \( (\rho_1, \rho_2) \) is larger than for \( (\rho_1^*, \rho_2^*) \) and estimation is inconsistent. Note that \( (\bar{s})^2 \leq \bar{s}^2 \).

For simplicity, let us focus on the case that all edge-lengths are equal. That is, \( s_k = s \) for \( k = 1, 2, \ldots, n \). In this situation,

\[ e(\rho_1, \rho_2) - e(\rho_1^*, \rho_2^*) = Q_{\rho_1\rho_2}(s) = A_D(\rho_1, \rho_2) s^2 + B_D(\rho_1, \rho_2) s + C_D(\rho_1, \rho_2). \]

Therefore, the MLE is inconsistent if \( Q_{\rho_1\rho_2}(s) > 0 \). Next, we will establish the condition for the inconsistency of the MLE in this scenario.

Note that \( P_{\rho_1 \rho_2}^*(1) = \delta_{\rho_1 \rho_2} \). So, Lemma 4 gives that \( Q_{\rho_1\rho_2}(1) = \log [\pi_{\rho_1} \pi_{\rho_2}] < 0 \) with \( \rho_1^* \neq \rho_1 \) and \( \rho_2^* \neq \rho_2 \). If \( \rho_1^* \neq \rho_1 \) but \( \rho_2^* = \rho_2 \) then \( Q_{\rho_1\rho_2}(1) = - f(\pi_{\rho_2}) + \log [\pi_{\rho_1} \pi_{\rho_2}^*] \). If \( \pi_{\rho_2} \geq 1/2, f(\pi_{\rho_2}) = 0 \) and \( Q_{\rho_1\rho_2}(1) = \log [\pi_{\rho_1} \pi_{\rho_2}^*] < 0 \). If \( \pi_{\rho_2} < 1/2, \) then

\[ Q_{\rho_1\rho_2}(1) = - \log (4\pi_{\rho_2} (1 - \pi_{\rho_2}^*]) + \log [\pi_{\rho_1} \pi_{\rho_2^*}] < - \log (4(1 - \pi_{\rho_2}^*)) < 0. \]

Thus \( Q_{\rho_1\rho_2}(1) < 0 \) for \( \rho_1^* \neq \rho_1 \) but \( \rho_2^* = \rho_2 \). By symmetry \( Q_{\rho_1\rho_2}(1) < 0 \) for \( \rho_1^* = \rho_1 \) but \( \rho_2^* \neq \rho_2 \). In summary, \( Q_{\rho_1\rho_2}(1) < 0 \) for \( (\rho_1, \rho_2) \neq (\rho_1^*, \rho_2^*). \)

If \( Q_{\rho_1\rho_2}(0) > 0 \), then since \( Q_{\rho_1\rho_2}(1) < 0 \) and \( Q_{\rho_1\rho_2}(s) \) is quadratic, \( Q_{\rho_1\rho_2}(s) \) will have exactly one root in \( s^* \in [0, 1] \). Moreover, for \( s < s^* \), \( Q_{\rho_1\rho_2}(s) > 0 \) and \( Q_{\rho_1\rho_2}(s) < 0 \) for \( s > s^* \). Thus, the MLE is inconsistent for \( s < s^* \) whenever there exists \( (\rho_1, \rho_2) \) with \( Q_{\rho_1\rho_2}(0) > 0 \). In Table 1 we tabulate the values of \( \pi \) in the intersection of the unit simplex and \( \{0.1, \ldots, 0.9\}^4 \) for which the MLE can be inconsistent. The \( (\rho_1, \rho_2) \) indicated are the ones giving the largest \( Q_{\rho_1\rho_2}(0) \). In
Table 1 The values of \( \pi \) in the intersection of the unit simplex and \([0.1, \ldots, 0.9]^4\) for which the MLE can be inconsistent. Here \((\rho_1^*, \rho_2^*) = (A, C)\) and \((\rho_1, \rho_2)\) give ancestral character states having a larger likelihood than \((\rho_1^*, \rho_2^*)\) when all edge lengths are equal and the common edge length is larger than \(t^*\).

| \(\pi_A\) | \(\pi_C\) | \(\pi_G\) | \(\pi_T\) | \(\rho_1\) | \(\rho_2\) | \(t^*\) |
|-----|-----|-----|-----|-----|-----|-----|
| 0.1  | 0.1  | 0.2  | 0.6  | G   | T   | 2.2  |
| 0.1  | 0.1  | 0.3  | 0.5  | G   | T   | 2.3  |
| 0.1  | 0.1  | 0.4  | 0.4  | G   | T   | 2.4  |
| 0.1  | 0.2  | 0.1  | 0.6  | C   | T   | 2.3  |
| 0.1  | 0.2  | 0.2  | 0.5  | C   | T   | 2.8  |
| 0.1  | 0.2  | 0.3  | 0.4  | G   | T   | 2.6  |
| 0.1  | 0.3  | 0.1  | 0.5  | C   | T   | 2.1  |
| 0.1  | 0.3  | 0.2  | 0.4  | C   | T   | 2.4  |
| 0.1  | 0.3  | 0.3  | 0.3  | C   | G   | 2.6  |
| 0.2  | 0.2  | 0.1  | 0.5  | A   | T   | 3.4  |
| 0.2  | 0.2  | 0.3  | 0.3  | G   | T   | 3.5  |
| 0.1  | 0.4  | 0.1  | 0.4  | C   | T   | 2.1  |
| 0.1  | 0.4  | 0.2  | 0.3  | C   | T   | 2.3  |
| 0.2  | 0.3  | 0.1  | 0.4  | C   | T   | 3.0  |
| 0.2  | 0.3  | 0.2  | 0.3  | C   | T   | 3.5  |
| 0.1  | 0.5  | 0.1  | 0.3  | C   | T   | 2.0  |
| 0.1  | 0.5  | 0.2  | 0.2  | C   | G   | 2.3  |
| 0.2  | 0.4  | 0.1  | 0.3  | C   | T   | 2.9  |
| 0.2  | 0.4  | 0.2  | 0.2  | A   | G   | 4.0  |
| 0.1  | 0.6  | 0.1  | 0.2  | C   | T   | 2.0  |
| 0.1  | 0.7  | 0.1  | 0.1  | A   | G   | 2.1  |

the case that all edge-lengths are equal (that is, \(s_k = \exp[-\mu t], \forall k\)), the MLE is inconsistent if \(t > t^* := -\log(s^*)/\mu\).

The above discussion assumed constant \(s_k = s\). In the cases that the \(s_k\) vary,

\[
Q_{\rho_1, \rho_2}(s) = AD(\rho_1, \rho_2)[\bar{s}^2 - (\bar{s})^2] + AD(\rho_1, \rho_2)(\bar{s})^2 + BD(\rho_1, \rho_2)\bar{s} + CD(\rho_1, \rho_2),
\]

If \(\lim \bar{s}\) corresponds to a value of \(s\) that gives rise to inconsistency in the single \(s_k\) case, then for sufficiently small variation of the \(s_k\) (small \(\bar{s}^2 - \bar{s}^2\)) inconsistency will arise. However, in the calculations leading to Table 1 we always found that \(AD(\rho_1, \rho_2) < 0\). Since inconsistency only arises when \(Q_{\rho_1, \rho_2}(s) > 0\), the result suggests that trees with more variable edge-lengths are less likely to lead to inconsistency, all other things being equal.
4 Discussion and conclusion

Likelihood-based methods, in particular the MLE, are often considered the best methods for ancestral state reconstruction. In this paper, we studied the consistency of the MLE for the problem of reconstructing the ancestral state of discrete traits on star trees whose edge lengths are unknown. We proved that the MLE is consistent under symmetric models but can be inconsistent under non-symmetric models. It is worth noticing that Theorem 1 implies that the empirical Bayes estimator will have the same difficulties as the MLE. This is a little surprising given the findings of Shaw et al. (2019) who showed that when the number of sites is large relative to the number of taxa, maximum likelihood estimation treating ancestral states as parameters can lead to inconsistent topological and edge-length estimation. By contrast the approach that averages over ancestral states, which is more analogous to empirical Bayes, does not suffer from such difficulties. We see here that with large numbers of sequences and small numbers of sites, the maximum likelihood and empirical Bayes approaches are more comparable.

The results were for the setting of ancestral reconstruction for the root of a star tree. This is for a simpler framework but results likely apply to multifurcations or “big bang” settings. Indeed, although not shown here, the results for the simple difference of frequency estimator can be extended directly to multifurcations. Outside of these settings, ancestral reconstructions will be more variable and not converge. Nevertheless, the results suggest that, more broadly, biases in ancestral reconstructions can become a significant difficulty with large numbers of sequences.

The effects of adding taxa or sites to an existing alignment has long been of interest (Graybeal 1998; Yang 1998; Pollock et al. 2002; Zwickl and Hillis 2002). Studies tend to conclude that phylogenetic accuracy is improved by additional taxon sampling. Most such research considers relatively small numbers of taxa, however. The results here do not deal directly with phylogenetic estimation which deserves additional study. Consider comparison of two conflicting topologies differing in $S_1S_2|S_3S_4$ versus $S_1S_3|S_2S_4$ where $S_1, \ldots, S_4$ are subtrees each with large numbers of taxa that give rise to separate clades. Label the corresponding ancestral nodes 1, \ldots, 4. The likelihoods for $S_1S_2|S_3S_4$ and $S_1S_3|S_2S_4$ can be expressed as

$$L(S_1S_2|S_3S_4) = \sum_{\rho} P[\rho_1, \ldots, \rho_4; 12|34] \prod_j P[S_j|\rho_j]$$

$$L(S_1S_3|S_2S_4) = \sum_{\rho} P[\rho_1, \ldots, \rho_4; 13|24] \prod_j P[S_j|\rho_j]$$

where $P[\rho_1, \ldots, \rho_4; 12|34]$ is the likelihood of the data $\rho_1, \ldots, \rho_4$ at ancestral nodes 1, \ldots, 4, calculated for the topology 12|34 for those nodes. Similarly, $P[\rho_1, \ldots, \rho_4; 13|24]$ gives the likelihood for the topology 13|24.

The arguments here suggest that $P[S_j|\rho_j]$ will be relatively large for only one $\rho_j$, say $\rho_j'$. Dividing gives

$$\frac{L(S_1S_2|S_3S_4)}{\prod_j P[S_j|\rho_j']} = \sum_{\rho} P[\rho_1, \ldots, \rho_4; 12|34] \prod_j P[S_j|\rho_j] \approx P[\rho_1', \ldots, \rho_4'; 12|34]$$
\[
\frac{L(S_1S_3|S_2S_4)}{\prod_j P[S_j|\rho'_j]} = \sum_{\rho} P[\rho_1, \ldots, \rho_4; 13|24] \prod_j P[S_j|\rho_j] \approx P[\rho'_1, \ldots, \rho'_4; 13|24]
\]

Thus, when estimation is consistent, comparison of the likelihoods for \(S_1S_2|S_3S_4\) and \(S_1S_3|S_2S_4\) is approximately the same as comparing the likelihoods for \(12|34\) and \(13|24\) having known what the ancestral data is. However, the inconsistency results here suggest that \(\rho'_j \neq \rho^*\) is a real possibility. In such a case, comparison of \(S_1S_2|S_3S_4\) and \(S_1S_3|S_2S_4\) is effectively the same as comparison of the two four-taxon topologies, \(12|34\) and \(13|24\), but using the wrong data at the tips.

Our results open an interesting direction for future research. An immediate avenue is extending these results beyond star trees. Since likelihood-based methods are no longer the undisputed best options when edge lengths are unknown, one long-term goal is to develop a better alternative approach. On the other hand, the inconsistency of likelihood-based methods under non-symmetric models may come from the fact that the number of unknown parameters grows with the number of species. A similar phenomenon has been shown in other statistical settings and is sometimes referred to as the Neyman-Scott problem after Neyman and Scott (1948) who showed that estimation of a measurement error variance \(\sigma^2\) can be inconsistent when a small number \(m\) of repeated measures, \(X_{i1}, \ldots, X_{im} \sim N(\mu_i, \sigma^2)\) are obtained for a large number \(i = 1, \ldots, n\) of objects. We hypothesize that if edge lengths are generated from a common distribution, then likelihood-based methods will be consistent. Motivation for this approach comes in part from the fact that treating the \(\mu_i\) as random in the Neyman-Scott problem can lead to consistent estimation of \(\sigma^2\) (Kiefer and Wolfowitz 1956). The setting for ancestral reconstruction is complicated by a wide range of dependence structures however and so this question remains open.

Acknowledgements We would like to thank Dr. Vu Dinh for the fruitful discussions about the topic.

A Technical details

In this section, we provide detailed proofs of our results. First, we prove the following useful property of Bernoulli expectations.

**Lemma 5 (A property of Bernoulli expectations)** Suppose that \(X_1, \ldots, X_M\) are independent and \(X_i \sim Bernoulli(p_i)\). For any \(j \in \{1, 2, \ldots, M\}\), we define

\[
x_{-} = [x_1, \ldots, x_{j-1}, 0, x_{j+1}, \ldots, x_M]
\]

where \(x \in \{0, 1\}^M\). Let \(g(x)\) be any function such that \(g(x) - g(x_{-}) \geq 0\) with strict inequality for at least one \(x\) that arises with positive probability. Then

\[
E_p[g(X)] - E_{p'}[g(X)] > 0, \quad \forall p > p'
\]

where \(E_p[g(X)]\) is the expected value when \(X = (X_1, X_2, \ldots, X_M)\) and \(p_j = p\).
Proof

\[
E[g(X)] = \sum_{x|x_j=1} g(x) P(X_j = 1) \prod_{i|i \neq j} P(X_i = x_i) \\
+ \sum_{x|x_j=0} g(x) P(X_j = 0) \prod_{i|i \neq j} P(X_i = x_i)
\]

\[
= \sum_{x|x_j=1} [g(x) - g(x_-)] P(X_j = 1) \prod_{i|i \neq j} P(X_i = x_i) \\
+ \sum_{x|x_j=0} g(x) \prod_{i|i \neq j} P(X_i = x_i)
\]

Therefore,

\[
E_p[g(x)] - E_{p'}[g(x)] = \sum_{x|x_j \geq 1} [g(x) - g(x_-)] [P_p(X_j = 1) \\
- P_{p'}(X_j = 1)] \prod_{i|i \neq j} P(X_i = x_i) \geq 0
\]

with strict inequality if \( g(x) - g(x_-) > 0 \) and \( \prod_{i|i \neq j} P(X_i = x_i) > 0 \) for at least one \( x \).

Now, we are ready to prove our Lemmas and Theorems in the paper.

Proof of Lemma 1 Since \((s_k)_{k=1}^\eta \in (0, 1)\), we have \( 0 < E(s_1) < +\infty \). By the Strong Law of Large Numbers, \( \tilde{s} \) converges to \( E(s_1) \) almost surely. Thus, Assumption 1 holds.

Proof of Lemma 2 Let \( g(y) = \log P_{\rho y}(\hat{s}(\rho, y)) \). Then,

\[
\text{Var}[\ell(\rho)/n] = n^{-2} \sum_y \text{Var}(n_y) g(y)^2 + n^{-2} \sum_{y'y'} \text{Cov}(n_y, n_{y'}) g(y) g(y').
\]

Multinomial calculations give that \( \text{Var}(n_y) = \sum_k P_{\rho^*y}(s_k)[1 - P_{\rho^*y}(s_k)] \leq \sum_k P_{\rho^*y}(s_k) \text{ and } \text{Cov}(n_y, n_{y'}) = -\sum_k P_{\rho^*y}(s_k) P_{\rho^*y'}(s_k) < 0 \). Thus

\[
\text{Var}[\ell(\rho)/n] \leq n^{-2} \sum_y \sum_k P_{\rho^*y}(s_k) g(y)^2 = n^{-1} \sum_y \bar{P}_{\rho^*y}[\log P_{\rho y}(\hat{s}(\rho, y))]^2 \to 0.
\]

Note that \( P_{\rho y}(\hat{s}(\rho, y)) > 0 \) because it is the maximized likelihood. Since \( \text{Var}[\ell(\rho)/n] \to 0 \), then for each \( \rho, \ell(\rho)/n \to_p \lim_n E[\ell(\rho)/n] = e(\rho) \). Since there are a finite collection of possible \( \rho \), then for any \( \epsilon > 0 \), with probability converging to 1, \(|\ell(\rho)/n - e(\rho)| < \epsilon \) for all \( \rho \). Let \( \rho_M \) be a maximizer of \( e(\rho) \) and let \( 3\epsilon = \min_{\rho \neq \rho_M} \{e(\rho_M) - e(\rho)\} > 0 \). Then \(|\ell(\rho)/n - e(\rho)| < \epsilon \) for all \( \rho \) implies
\[
\ell(\rho_M)/n - \ell(\rho)/n > e(\rho_M) - \epsilon - (e(\rho) + \epsilon)
\]
\[
e(\rho_M) - e(\rho) - 2\epsilon \geq 3\epsilon - 2\epsilon > 0
\]
for all \(\rho \notin \mathcal{H}\). In conclusion, \(\hat{\rho} \in \mathcal{H}\), with probability converging to 1. \(\square\)

**Proof of Theorem 1** Let \(\rho_M = \rho_M(y)\) be the element of \(\mathcal{H}\) giving the largest \(\ell(\rho)\) for a given \(y\). The empirical Bayes estimator \(\hat{\rho}_B\) is also the maximizer of

\[
\frac{P_\pi(\rho) \prod_{k=1}^n P_{\rho y_k}(\hat{s}_k)}{P_\pi(\rho_M) \prod_{k=1}^n P_{\rho_M y_k}(\hat{s}_k)} = \frac{P_\pi(\rho) \prod_{k=1}^n P_{\rho y_k}(\hat{s}_k)}{P_\pi(\rho_M) \exp[\ell(\rho_M)]} \times \frac{P_\pi(\rho_M) \prod_{k=1}^n P_{\rho_M y_k}(\hat{s}_k)}{P_\pi(\rho_M) \exp[\ell(\rho_M)]} = : T_1/T_2.
\]

Since this function equals 1 when \(\rho = \rho_M\), it suffices to show that it converges to 0 for \(\rho \notin \mathcal{H}\). Since \(\hat{s}(\rho, y_k)\) is the maximizer of \(P_{\rho y_k}(s)\) and since \(\lim_n \{\ell(\rho_M)/n - \ell(\rho)/n\} > 0\),

\[
T_1 \leq \frac{P_\pi(\rho) \prod_{k=1}^n P_{\rho y_k}(\hat{s}(\rho, y_k))}{P_\pi(\rho_M) \exp[\ell(\rho_M)]} = \frac{P_\pi(\rho) \exp[\ell(\rho)]}{P_\pi(\rho_M) \exp[\ell(\rho_M)]} = \frac{P_\pi(\rho)}{P_\pi(\rho_M)} \exp \left\{ -n \left[ \frac{\ell(\rho_M)}{n} - \frac{\ell(\rho)}{n} \right] \right\} \to_p 0.
\]

On the other hand,

\[
T_2 = \frac{P_\pi(\rho_M) \prod_{k=1}^n P_{\rho_M y_k}(\hat{s}_k)}{P_\pi(\rho_M) \exp[\ell(\rho_M)]} = \sum_\rho \frac{P_\pi(\rho) \prod_{k=1}^n P_{\rho y_k}(\hat{s}_k)}{P_\pi(\rho_M) \exp[\ell(\rho_M)]} - \sum_{\rho \neq \rho_M} \frac{P_\pi(\rho) \prod_{k=1}^n P_{\rho y_k}(\hat{s}_k)}{P_\pi(\rho_M) \exp[\ell(\rho_M)]}.
\]

Since \(\hat{s}\) maximizes the first sum and since \(\hat{s}(\rho, y_k)\) is the maximizer of \(P_{\rho y_k}(s)\),

\[
T_2 \geq \sum_\rho \frac{P_\pi(\rho) \prod_{k=1}^n P_{\rho y_k}(\hat{s}(\rho, y_k))}{P_\pi(\rho_M) \exp[\ell(\rho_M)]} - \sum_{\rho \neq \rho_M} \frac{P_\pi(\rho) \prod_{k=1}^n P_{\rho y_k}(\hat{s}(\rho, y_k))}{P_\pi(\rho_M) \exp[\ell(\rho_M)]}
\]
\[
\geq \frac{P_\pi(\rho_M) \prod_{k=1}^n P_{\rho_M y_k}(\hat{s}(\rho_M, y_k))}{P_\pi(\rho_M) \exp[\ell(\rho_M)]} - \sum_{\rho \neq \rho_M} \frac{P_\pi(\rho) \prod_{k=1}^n P_{\rho y_k}(\hat{s}(\rho, y_k))}{P_\pi(\rho_M) \exp[\ell(\rho_M)]}
\]
\[
= 1 - \sum_{\rho \neq \rho_M} \frac{P_\pi(\rho) \prod_{k=1}^n P_{\rho y_k}(\hat{s}(\rho, y_k))}{P_\pi(\rho_M) \exp[\ell(\rho_M)]} \to_p 1.
\]

Hence, \(T_1/T_2 \to_p 0\). \(\square\)
Proof of Lemma 3  Taking derivatives of both sides of (2), we obtain

\[
\frac{\partial}{\partial s} \log P_{\rho y}(s) = \sum_i n_{\rho_y}(i, i) \frac{1 - \pi_i}{\pi_i + (1 - \pi_i)s} - \frac{N - \sum_i n_{\rho_y}(i, i)}{1 - s}
\]

\[
= \sum_i n_{\rho_y}(i, i) \frac{(1 - \pi_i)(1 - s) + \pi_i + (1 - \pi_i)s}{[\pi_i + (1 - \pi_i)s](1 - s)} - \frac{N}{1 - s}
\]

(9)

Taking derivatives of (9),

\[
\frac{\partial^2}{\partial s^2} \log P_{\rho y}(s) = -\sum_i n_{\rho_y}(i, i) \frac{(1 - \pi_i)^2}{(\pi_i + (1 - \pi_i)s)^2} - \frac{N - \sum_i n_{\rho_y}(i, i)}{(1 - s)^2} < 0.
\]

Thus, \( \log P_{\rho y}(s) \) is concave as a function of \( s \).

When \( \sum_i n_{\rho_y}(i, i) = N \), up to an additive constant,

\[
\log P_{\rho y}(s) = \sum_i n_{\rho_y}(i, i) \log[\pi_i + (1 - \pi_i)s],
\]

an increasing function of \( s \). So \( \log P_{\rho y}(s) \) achieves its maximum at \( s = 1 \). Hence \( \hat{s}(\rho, y) = 1 \).

If \( \sum_i n_{\rho_y}(i, i) < N \), then \( N - \sum_i n_{\rho_y}(i, i) > 0 \) in the second term of (2), so \( s = 1 \) gives \( \log P_{\rho y}(s) = -\infty \). Thus \( \hat{s}(\rho, y) < 1 \) in this case. Since \( \log P_{\rho y}(s) \) is concave, \( \hat{s}(\rho, y) = 0 \) if and only if

\[
\frac{\partial}{\partial s} \log P_{\rho y}(0) \leq 0.
\]

Substituting \( s = 0 \) in (10) gives \( \hat{s}(\rho, y) = 0 \) if and only if

\[
\sum_i \frac{n_{\rho_y}(i, i)}{\pi_i} \leq N.
\]

If \( \sum_i n_{\rho_y}(i, i)/\pi_i > N \) and \( \sum_i n_{\rho_y}(i, i) < N \), then \( \hat{s}(\rho, y) \) is obtained by setting (10) to 0 which gives

\[
\sum_i \frac{n_{\rho_y}(i, i)}{\pi_i + (1 - \pi_i)\hat{s}(\rho, y)} = N.
\]
Proof of Theorem 2} Note that $\sum_i n_{\rho y}(i, i) = \sum_{l=1}^N x_l$, here $x_l = 1$ if $y_l = \rho_l$ and 0 otherwise. For the symmetric model, (2) gives that

$$
\log P_{\rho y}(\hat{s}) = \sum_l x_l \cdot \log[1/c + (1 - 1/c)\hat{s}] + (N - \sum_l x_l) \log(1 - \hat{s}) - \sum_l x_l \log(1/c) + C(y)
$$

It follows that $\hat{s} = \hat{s}(x)$ is a function of $(\rho, y)$ through $x$ alone. As a function of $x$, $\log P_{\rho y}(\hat{s}) - C(y)$ is

$$
g(x) = F(\hat{s}(x), x) = \sum_l x_l \log[1/c + (1 - 1/c)\hat{s}(x)] + (N - \sum_l x_l) \log[1 - \hat{s}(x)] - \sum_l x_l \log(1/c).
$$

Thus

$$
\frac{\partial}{\partial x_j} g(x) = \frac{\partial}{\partial s} F(\hat{s}(x), x) \cdot \frac{\partial \hat{s}}{\partial x_j} + \frac{\partial}{\partial x_j} F(\hat{s}(x), x) = \frac{\partial \hat{s}}{\partial x_j} \cdot \frac{\partial}{\partial s} \log P_{\rho y}(\hat{s}) + \log[1/c + (1 - 1/c)\hat{s}] - \log(1 - \hat{s}) - \log[1/c].
$$

If $0 < \hat{s} < 1$, then $\frac{\partial}{\partial s} \log P_{\rho y}(\hat{s}) = 0$. so

$$
\frac{\partial}{\partial x_j} g(x) = \log[1 + (c - 1)\hat{s}] - \log(1 - \hat{s}) > 0.
$$

If $\hat{s} = 0$,

$$
\frac{\partial}{\partial x_j} g(x) = \frac{\partial \hat{s}}{\partial x_j} \cdot \frac{\partial}{\partial s} \log P_{\rho y}(0).
$$

From Lemma 3, $\hat{s} = 0$ whenever $\sum_i n_{\rho y}(i, i)/(1/c) = c \sum_i x_l \leq N$. Thus $\frac{\partial \hat{s}}{\partial x_j} = 0$ for $\hat{s} = 0$ and $c \sum_i x_l < N$. If $\hat{s} = 0$ and $c \sum_i x_l = \sum_i n_{\rho y}(i, i)/(1/c) = N$, then, by Equation (10), we deduce that $\frac{\partial}{\partial s} \log P_{\rho y}(0) = 0$. Therefore, $\frac{\partial}{\partial x_j} g(x) = 0$ when $\hat{s} = 0$. Thus, for $0 \leq \hat{s} < 1$, $\frac{\partial}{\partial x_j} g(x) \geq 0$ with strict inequality for any $x$ such that $0 < \hat{s} < 1$. If $\hat{s} = 1$ then $\sum_i n_{\rho y}(i, i) = N$ or each of the $y_l = \rho_l$. In this case $x_j = 1$ and cannot be increased. In conclusion, $g(x) - g(x_-) \geq 0$, where $x_- = [x_1, \ldots, x_{j-1}, 0, x_{j+1}, \ldots, x_c]$, with strict inequality for all $x$ with $\hat{s} > 0$.  

 Springer
By Equation (2), an alternative expression for $E[\ell(\rho)/n]$ is given by

$$E[\ell(\rho)/n] = \frac{1}{n} \sum_{k=1}^{n} E[\log P_{\rho} Y_k (\hat{s}(\rho, Y_k))] = \frac{1}{n} \sum_{k=1}^{n} E[g(X^{(k)})] + \frac{1}{n} \sum_{k=1}^{n} E[C(Y_k)].$$

where $E[g(X^{(k)})]$ is the expected value with respect to the sequence observed at the $k$th leaf. Here, $X^{(k)} = (X_1, X_2, \ldots, X_N)$ where, independently, $X_i \sim$ Bernoulli($P_{\rho_i} \rho_i(s_k)$).

Consider a $\rho$ with $\rho_j \neq \rho_j^*$. We denote $h_\rho (s_k) = E[g(X^{(k)})]$ and define $\tilde{\rho}$ as $\tilde{\rho}_i = \rho_i$ for $i \neq j$ and $\tilde{\rho}_j = \rho_j^*$. Since, for the symmetric model, $P_{\rho_i}^{\rho_i^*}(s) > P_{\rho_j}^{\rho_j^*}(s), \forall s \in (0,1)$ and $g(x) - g(x-) \geq 0$ with strict inequality for all $x$ with $\hat{s} > 0$, Lemma 5 (the property of Bernoulli expectations) gives that $h_\tilde{\rho}(s) > h_\rho(s)$ for all $s \in (0,1]$.

We have that $e(\rho) = e(\rho)$ is the limiting value of $\sum_{k=1}^{n}[h_\tilde{\rho}(s_k) - h_\rho(s_k)]/n$ and $h_\tilde{\rho}(s) - h_\rho(s) > 0$ for $s \in (0,1]$. Thus if sufficiently many of the $s_k \geq U$ for some $U > 0$ but possibly small, then $e(\rho) > 0$. Formally, for any $U \in (0,1]$, $s_U = \text{argmin}_{s \in [U,1]} h_\rho(s) - h_\rho(s) > 0$, by continuity with respect to $s$. Denoting the proportion of $s_k \geq U$ as $p_U$,

$$\bar{s} \leq \sum_{k|s_k \geq U} \frac{1}{n} + \frac{1}{n} \sum_{k|s_k < U} U \leq p_U + U.$$

Since $0 < \lim \bar{s}$, choosing $U = \lim \bar{s}/2$, gives that $\lim \inf p_U \geq \lim \bar{s}/2$. Thus, for sufficiently large $n$,

$$e(\rho) = \lim \frac{1}{n} \sum_{k} h_\tilde{\rho}(s_k) - h_\rho(s_k) \geq \lim \frac{1}{n} \sum_{k|s_k \geq U} h_\tilde{\rho}(s_k) - h_\rho(s_k) \geq [h_\tilde{\rho}(s_U) - h_\rho(s_U)] \lim \inf p_U \geq \frac{[h_\tilde{\rho}(s_U) - h_\rho(s_U)] \lim \bar{s}}{2} > 0.$$

In other words, $e(\rho)$ can always be increased by changing a $\rho_j$ that is not equal to the true ancestral state, if such $\rho_j$ exists. Thus $\rho_M = \rho^*$.

\begin{proof}[Proof of Theorem 3] It is sufficient for us to focus on $\rho = (a, \ldots, a)$. Let $\hat{\rho}_a = n_{\rho y}(a, a)/N$ and $\hat{s} = \hat{s}(\rho, y)$. When $1 > \hat{\rho}_a > \pi_a$, by Lemma 3, we have

$$\frac{n_{\rho y}(a, a)}{\pi_a + (1 - \pi_a)\hat{s}} = N \iff \hat{\rho}_a = \pi_a + (1 - \pi_a)\hat{s} \quad (11)$$

which has solution $\hat{s} = (\hat{\rho}_a - \pi_a)/(1 - \pi_a)$. Substituting in Equation (2) gives that when $1 > \hat{\rho}_a > \pi_a$,

$$\log \frac{P_{\rho y}(\hat{s})}{N} = \hat{\rho}_a \log \left\{ \pi_a + (1 - \pi_a) \cdot \frac{\hat{\rho}_a - \pi_a}{1 - \pi_a} \right\} + (1 - \hat{\rho}_a) \log \left\{ 1 - \frac{\hat{\rho}_a - \pi_a}{1 - \pi_a} \right\}$$
\[-\hat{p}_a \log(\pi_a) + C(y)/N\]
\[= \hat{p}_a \log \left\{ \frac{\hat{p}_a}{\pi_a} \right\} + (1 - \hat{p}_a) \log \left\{ \frac{1 - \hat{p}_a}{1 - \pi_a} \right\} + C(y)/N \tag{12} \]

In the case that $\hat{p}_a = 1, \hat{s} = 1$, which when substituted in Equation (2) gives Equation (12) where we adopt the convention that $0 \cdot \log(0) = 0$. In the case that $\hat{p}_a \leq \pi_a$, \[\log P_{\rho y}(\hat{s}) = C(y)/N\]. Thus, for the $k$th sequence,
\[E \left[ \log P_{\rho y_k}(\hat{s}_k) \right] = \hat{p}_a \log \left\{ \frac{\hat{p}_a}{\pi_a} \right\} + (1 - \hat{p}_a) \log \left\{ \frac{1 - \hat{p}_a}{1 - \pi_a} \right\} \cdot I \{\hat{p}_a > \pi_a\} \]
\[+ E[C(y)/N], \]
where $I \{\cdot\}$ is the indicator function. For the $k$th sequence, $N \hat{p}_a \sim \text{binomial}(N, P_{ra}(s_k))$. Let
\[w(\rho, \pi) = E \left[ \hat{p} \log \left\{ \frac{\hat{p}}{\pi} \right\} + (1 - \hat{p}) \log \left\{ \frac{1 - \hat{p}}{1 - \pi} \right\} \right] \cdot I \{\hat{p} > \pi\}, \]
where $N \hat{p} \sim \text{binomial}(N, p)$. Since $e(\rho)$ is the limiting average value of $E[\log P_{\rho y_k}(\hat{s}_k)]$,
\[e(\rho) - e(\rho^*) = \lim \frac{1}{n} \sum_{k=1}^{n} \left\{ w(P_{ra}(s_k), \pi_a) - w(P_{rr}(s_k), \pi_r) \right\}. \]
If $s_k = 0$ for all $k$, then $w(P_{ra}(s_k), \pi_a) = v(\pi_a)$ and $w(P_{rr}(s_k), \pi_r) = v(\pi_r)$, so
\[e(\rho) - e(\rho^*) = v(\pi_a) - v(\pi_r) \]
If this value is positive, then by continuity, $e(\rho) - e(\rho^*)$ will be positive for choices of $s$ for which $\lim \hat{s}$ is sufficiently small. \qed

**Proof of Theorem 4** By Equation (4)
\[E[V(j, l)] = \hat{s}[I\{\rho_l^* = j\} - \sum_{v \neq l} I\{\rho_v^* = j\}]/(k - 1). \]
Since there are at least two distinct ancestral character states,
\[E[V(\rho_l^*, l)] \geq \hat{s} \left[ 1 - \frac{N - 2}{N - 1} \right] = \frac{\hat{s}}{N - 1} \to \lim \frac{\hat{s}}{N - 1} > 0. \]
On the other hand, for $j \neq \rho_l^*$, \(E[V(j, l)] \leq 0\). We have
\[P[\rho^{(D)} \neq \rho^*; s] = P \left[ \bigcup_{l=1}^{N} (\hat{\rho}_l^{(D)} \neq \rho_l^*); s \right] \leq \sum_{l=1}^{N} P[\hat{\rho}_l^{(D)} \neq \rho_l^*; s]. \tag{13} \]
For any $\epsilon_n(s)$,

$$P[\rho_l^{(D)} = \rho_l^*; s] = P \left[ \bigcap_{j \neq \rho_l^*} \{ V(\rho_l^*, l) > V(j, l); s \} \right]$$

$$\geq P \left[ \bigcap_{j \neq \rho_l^*} \{ V(j, l) \leq \epsilon_n(s) \} \bigcap \{ V(\rho_l^*, l) > \epsilon_n(s); s \} \right]$$

$$= 1 - P \left[ \bigcup_{j \neq \rho_l^*} \{ V(j, l) > \epsilon_n(s) \} \bigcup \{ V(\rho_l^*, l) \leq \epsilon_n(s); s \} \right]$$

$$\geq 1 - P[V(\rho_l^*, l) \leq \epsilon_n(s); s] - \sum_{j \neq \rho_l^*} P[V(j, l) > \epsilon_n(s); s].$$

So

$$P[\rho_l^{(D)} \neq \rho_l^*; s] = 1 - P[\rho_l^{(D)} = \rho_l^*; s] \leq P[V(\rho_l^*, l) \leq \epsilon_n(s); s]$$

$$+ \sum_{j \neq \rho_l^*} P[V(j, l) > \epsilon_n(s); s]. \quad (14)$$

Now

$$V(j, l) = \frac{1}{n} \sum_{k=1}^{n} \left[ I\{y_{kl} = j\} - \frac{1}{N-1} \sum_{v \neq l} I\{y_{kv} = j\} \right] = \frac{1}{n} \sum_{k=1}^{n} V_k$$

where the $V_k \in [-1, 1]$ are independent. Thus, using Hoeffding’s Inequality, for $E[V(\rho_l^*, l)] - \epsilon_n > 0$, we have

$$P[V(\rho_l^*, l) \leq \epsilon_n(s); s] = P\{ - V(\rho_l^*, l) - E[-V(\rho_l^*, l)] \geq E[V(\rho_l^*, l)] - \epsilon_n(s); s \}$$

$$\leq \exp[-n(E[V(\rho_l^*, l)] - \epsilon_n(s))^2]. \quad (15)$$

Consider $\epsilon_n(s) = \bar{s}/[2(N - 1)] > 0$. Since $E[V(\rho_l^*, l)] \geq \bar{s}/(N - 1)$, then

$$E[V(\rho_l^*, l)] - \epsilon_n(t) \geq \frac{\bar{s}_n(t)}{2(N - 1)} > 0.$$

Substituting in (15)

$$P[V(\rho_l^*, l) \leq \epsilon_n(s); s] \leq \exp[-n\bar{s}_n^2/[4(N - 1)^2]] \quad (16)$$

For $j \neq \rho_l^*$, $E[V(j, l)] \leq 0$, so

\[ Springer\]
\[ P[V(j, l) > \epsilon_n(s); s] \leq P[V(j, l) - E[V(j, l)] > \epsilon_n(s); s] \leq \exp[-n\epsilon_n(s)^2] = \exp[-n\tilde{s}^2/[4(N - 1)^2]] \quad (17) \]

Combining (13), (14), (16) and (17) gives
\[ P[\rho^{(D)} \neq \rho^*; s] \leq cN \exp[-n\tilde{s}^2/[4(N - 1)^2]] \leq cN \exp[-n\tilde{s}^2/[4N^2]]. \quad (18) \]

Substituting \( N = \sqrt{sn/(2\log n)} \) in (18) gives
\[ P[\rho^{(D)} \neq \rho^*; s] \leq c\sqrt{sn/(2\log n)} \cdot \exp[-(\log n)/2] = c\sqrt{\tilde{s}/2\log n}. \]

So consistency holds with \( N = O(\sqrt{n}/\log n). \quad \square \)

**Proof of Lemma 4** Consider the MLE of edge-lengths when likelihoods are calculated with fixed ancestral states, \((\rho_1, \rho_2)\), at the two sites. For a sequence with character states \((x, y)\) at the two sites, when \((x, y) = (\rho_1, \rho_2)\), the MLE is \(\hat{\delta} = 1\), giving log likelihood contribution \(\log(1) = 0\). When \(x \neq \rho_1\) and \(y \neq \rho_2\), \(\hat{\delta} = 0\) and the log likelihood contribution is \(\log(\pi_x \pi_y)\). When \(x = \rho_1\) but \(y \neq \rho_2\), the log likelihood contribution for the sequence is
\[ \log[\pi_{\rho_1} + (1 - \pi_{\rho_1})\hat{\delta}][1 - \hat{\delta}]\pi_y] = \log(\pi_{\rho_1} + (1 - \pi_{\rho_1})\hat{\delta}) + \log(1 - \hat{\delta}) + \log(\pi_y) \quad (19) \]
where \(\hat{\delta}\) is the maximizer of \(F(s) = \log[(\pi_{\rho_1} + (1 - \pi_{\rho_1})s) + \log(1 - s)\]. Here
\[ F'(s) = \frac{1 - \pi_{\rho_1}}{\pi_{\rho_1} + (1 - \pi_{\rho_1})s} - \frac{1}{1 - s}. \]

Note that \(F'(s) = 0\) has a unique solution in \((-\infty, 1]\):
\[ s_0 = \frac{1 - 2\pi_{\rho_1}}{2(1 - \pi_{\rho_1})} \]
and \(F(s) > 0 \Leftrightarrow s < s_0\). If \(\pi_{\rho_1} \geq 1/2\), then \(F'(s) < 0\) for all \(s \in [0, 1]\). So \(\hat{\delta} = 0\) and the log likelihood contribution (19) is \(\log(\pi_{\rho_1}) + \log(\pi_y)\). If \(\pi_{\rho_1} < 1/2\), \(F(s)\) achieves maximum at \(s_0\). Thus, \(\hat{\delta} = s_0\). Substituting in (19) gives log likelihood contribution
\[ \log[1/2] - \log[2(1 - \pi_{\rho_1})] + \log[\pi_y] = -\log[4(1 - \pi_{\rho_1})] + \log[\pi_y] \quad (20) \]
By symmetry, if \(x \neq \rho_1\) but \(y = \rho_2\), the log likelihood contributions are \(\log(\pi_{\rho_2}) + \log[\pi_x]\) when \(\pi_{\rho_2} \geq 1/2\) and \(-\log[4(1 - \pi_{\rho_2})] + \log[\pi_x]\) when \(\pi_{\rho_2} < 1/2\).

Let \(n_{y_1y_2}\) be the number times we observe the sequence \((y_1, y_2)\), we define
\[ n_{y_1}^{(1)} = \sum_{y_2} n_{y_1y_2}, \quad \text{and} \quad n_{y_2}^{(2)} = \sum_{y_1} n_{y_1y_2}, \]

\(\square\) Springer
Substituting the log likelihood contributions gives

\[
\ell(\rho_1, \rho_2) = \sum_{x \neq \rho_1} \sum_{y \neq \rho_2} n_{xy} \log[\pi_x \pi_y] + \sum_{y \neq \rho_2} n_{\rho_1 y} \log[\pi_y] + \sum_{x \neq \rho_1} n_{x \rho_2} \log[\pi_x]
\]

\[
+ \sum_{y \neq \rho_2} n_{\rho_1 y} \{ I[\pi_{\rho_1} \geq 1/2] \log[\pi_{\rho_1}] - I[\pi_{\rho_1} < 1/2] \log[4(1 - \pi_{\rho_1})] \}
\]

\[
+ \sum_{x \neq \rho_1} n_{\rho_1 x} \{ I[\pi_{\rho_2} \geq 1/2] \log[\pi_{\rho_2}] - I[\pi_{\rho_2} < 1/2] \log[4(1 - \pi_{\rho_2})] \}.
\]

(21)

Combining terms involving \(\pi_x\) gives the sum

\[
\sum_{x \neq \rho_1} \sum_{y \neq \rho_2} n_{xy} \log[\pi_x] + \sum_{y \neq \rho_2} n_{x \rho_2} \log[\pi_x]
\]

\[
= \sum_{x \neq \rho_1} \sum_{y} n_{xy} \log[\pi_x]
\]

\[
= \sum_{x \neq \rho_1} n^{(1)}_x \log[\pi_x]
\]

\[
= \sum_{x} n^{(1)}_x \log[\pi_x] - n^{(1)}_{\rho_1} \log[\pi_{\rho_1}].
\]

(22)

Similarly,

\[
\sum_{x \neq \rho_1} \sum_{y \neq \rho_2} n_{xy} \log[\pi_y] + \sum_{y \neq \rho_2} n_{\rho_1 y} \log[\pi_y] = \sum_{y} n^{(2)}_y \log[\pi_y] - n^{(2)}_{\rho_2} \log[\pi_{\rho_2}].
\]

(23)

Substituting (22) and (23) in (21) gives

\[
\ell(\rho_1, \rho_2) = \sum_{x} n^{(1)}_x \log[\pi_x] + \sum_{y} n^{(2)}_y \log[\pi_y] - n^{(1)}_{\rho_1} \log[\pi_{\rho_1}] - n^{(2)}_{\rho_2} \log[\pi_{\rho_2}]
\]

\[
+ [n^{(1)}_{\rho_1} - n_{\rho_1 \rho_2}][I[\pi_{\rho_1} \geq 1/2] \log[\pi_{\rho_1}] - I[\pi_{\rho_1} < 1/2] \log[4(1 - \pi_{\rho_1})]]
\]

\[
+ [n^{(2)}_{\rho_2} - n_{\rho_1 \rho_2}][I[\pi_{\rho_2} \geq 1/2] \log[\pi_{\rho_2}] - I[\pi_{\rho_2} < 1/2] \log[4(1 - \pi_{\rho_2})]]
\]

(24)

Note that

\[
I[\pi \geq 1/2] \log[\pi] - I[\pi < 1/2] \log[4(1 - \pi)]
\]

\[
= \log[\pi] - I[\pi < 1/2] \log[4\pi(1 - \pi)].
\]
Substituting in (24) with \( f(\pi) = I[\pi < 1/2] \log[4\pi(1 - \pi)] \),

\[
\ell(\rho_1, \rho_2) = \sum_x n_x^{(1)} \log[\pi_x] + \sum_y n_y^{(2)} \log[\pi_y] - n_{\rho_1}^{(1)} \log[\pi_{\rho_1}] - n_{\rho_2}^{(2)} \log[\pi_{\rho_2}]
\]

\[
+ [n_{\rho_1}^{(1)} - n_{\rho_1\rho_2}][\log[\pi_{\rho_1}] - f(\pi_{\rho_1})] + [n_{\rho_2}^{(2)} - n_{\rho_1\rho_2}][\log[\pi_{\rho_2}] - f(\pi_{\rho_2})]
\]

\[
= \sum_x n_x^{(1)} \log[\pi_x] + \sum_y n_y^{(2)} \log[\pi_y] - n_{\rho_1}^{(1)} f(\pi_{\rho_1}) - n_{\rho_2}^{(2)} f(\pi_{\rho_2})
\]

\[
+ n_{\rho_1\rho_2} \{ f(\pi_{\rho_1}) + f(\pi_{\rho_2}) - \log[\pi_{\rho_1\rho_2}] \}.
\]

The expectations of \( n_{\rho_1}^{(1)}/n, n_{\rho_2}^{(2)}/n \) and \( n_{\rho_1\rho_2}/n \) are \( \tilde{P}_{\rho_1}^{\pi_1}, \tilde{P}_{\rho_2}^{\pi_2} \) and \( \tilde{P}_{(\rho_1, \rho_2)} \) respectively. So the expected maximized log likelihood is

\[
E[l(\rho_1, \rho_2)/n] = C_0 - \tilde{P}_{\rho_1}^{\pi_1} f(\pi_{\rho_1}) - \tilde{P}_{\rho_2}^{\pi_2} f(\pi_{\rho_2}) + \tilde{P}_{(\rho_1, \rho_2)} \{ f(\pi_{\rho_1}) + f(\pi_{\rho_2}) - \log[\pi_{\rho_1\rho_2}] \}
\]

where

\[
C_0 = \sum_x \tilde{P}_{\rho_1}^{\pi_1} \log[\pi_x] + \sum_y \tilde{P}_{\rho_2}^{\pi_2} \log[\pi_y]
\]

is independent of \((\rho_1, \rho_2)\).

\[\square\]

References

Ané C (2008) Analysis of comparative data with hierarchical autocorrelation. Ann Appl Stat 2(3):1078–1102

Bickel PJ, Doksum K (2007) Mathematical Statistics: Basic Ideas and Selected Topics, vol I. Pearson, New Jersey

Collins TM, Wimberger PH, Naylor GJP (1994) Compositional bias, character state bias, and character-state reconstruction using parsimony. Syst Biol 43(4):482–496

Dinh V, Ho LST, Suchard MA, Matsen FA IV (2018) Consistency and convergence rate of phylogenetic inference via regularization. Ann Stat 46(4):1481

Erdős PL, Steel MA, Székely L, Warnow TJ (1999) A few logs suffice to build (almost) all trees: Part ii. Theoret Comput Sci 221(1–2):77–118

Erdős PL, Steel MA, Székely LA, Warnow TJ (1999) A few logs suffice to build (almost) all trees (i). Random Struct Algorithms 14(2):153–184

Eyre-Walker A (1998) Problems with parsimony in sequences of biased base composition. J Mol Evol 47:686–690

Fan W-TL, Roch S (2018) Necessary and sufficient conditions for consistent root reconstruction in markov models on trees. Electron J Probab 23:1–24

Faria NR, Rambaut A, Suchard MA, Baele G, Bedford T, Ward MJ, Tatem AJ, Sousa JD, 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 (2004) Inferring Phylogenies. Sinauer, Massachusetts

Finarelli JA, Flynn JJ (2006) Ancestral state reconstruction of body size in the Caniformia (Carnivora, Mammalia): the effects of incorporating data from the fossil record. Syst Biol 55(2):301–313

Gascuel O, Steel M (2010) Inferring ancestral sequences in taxon-rich phylogenies. Math Biosci 227(2):125–135

Gascuel O, Steel M (2020) A Darwinian uncertainty principle. Syst Biol 69(3):521–529
Gaucher EA, Thomson JM, Burgan MF, Benner SA (2003) Inferring the palaeoenvironment of ancient bacteria on the basis of resurrected proteins. Nature 425:285–288
Gill MS, Ho LST, Baele G, Lemey P, Suchard MA (2017) A relaxed directional random walk model for phylogenetic trait evolution. Syst Biol 66(3):299–319
Gojobori T, Li W-H, Graur D (1982) Patterns of nucleotide substitution in pseudogenes and functional genes. J Mol Evol 18:360–369
Graybeal A (1998) Is it better to add taxa or characters to a difficulty phylogenetic problem? Syst Biol 47(1):9–17
Herbst L, Fischer M (2018) On the accuracy of ancestral sequence reconstruction for ultrametric trees with parsimony. Bull Math Biol 80(4):864–879
Herbst L, Li H, Steel M (2019) Quantifying the accuracy of ancestral state prediction in a phylogenetic tree under maximum parsimony. J Math Biol 78(6):1953–1979
Ho LST, Dinh V, Nguyen CV (2019) Multi-task learning improves ancestral state reconstruction. Theor Popul Biol 126:33–39
Kiefer J, Wolfowitz J (1956) Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. Ann Math Stat 27(4):887–906
Lemey P, Rambaut A, Drummond AJ, Suchard MA (2009) Bayesian phylogeography finds its roots. PLoS Comput Biol 5(9):e1000520
Liberles D (2007) Ancestral sequence reconstruction. Oxford University Press, Oxford
Maddison DR (1994) Phylogenetic methods for inferring the evolutionary history and processes of change in discretely valued characters. Annu Rev Entomol 39:267–292
Maddison WP (1995) Calculating the probability distributions of ancestral states reconstructed by parsimony on phylogenetic trees. Syst Biol 44(4):474–481
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
Neyman J, Scott EL (1948) Consistent estimates based on partially consistent estimation. Econometrica 16(1):1–32
Odom KJ, Hall ML, Riebel K, Omland KE, Langmore NE (2014) Female song is widespread and ancestral in songbirds. Nat Commun 5(1):1–6
Pollock DD, Zwickl DJ, McGuire JA, Hillis DM (2002) Increased taxon sampling is advantageous for phylogenetic inference. Syst Biol 51(4):664–671
Royer-Carenzi M, Pontarotti P, Didier G (2013) Choosing the best ancestral character state reconstruction method. Math Biosci 242(1):95–109
Shaw DA, Dinh VC, Matsen FA (2019) Joint maximum likelihood of phylogeny and ancestral states is not consistent. Mol Biol Evol 36(10):2352–2357
Steel M, Rodrigo A (2008) Maximum likelihood supertrees. Syst Biol 57(2):243–250
Susko E, Roger AJ (2013) Problems with estimation of ancestral frequencies under stationary models. Syst Biol 62:330–338
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
Yang Z (1998) On the best evolutionary rate for phylogenetic analysis. Syst Biol 47:125–133
Zwickl Derrick J, Hillis DM (2002) Increased taxon sampling greatly reduces phylogenetic error. Syst Biol 51(4):588–598

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.