Syst. Biol. 70(4):774–785, 2021

DOI:10.1093/sysbio/syaa100
Advance Access publication December 30, 2020

The Asymptotic Behavior of Bootstrap Support Values in Molecular Phylogenetics

Jun Huang[1,2], Yuting Liu[1,*], Tianqi Zhu[3,*] and Ziheng Yang[2,*]

[1]*Department of Mathematics, Beijing Jiaotong University, Beijing, 100044, China;*
[2]*Department of Genetics, Evolution and Environment, University College London, Gower Street, London WC1E 6BT, UK;*
[3]*National Center for Mathematics and Interdisciplinary Sciences, Key Laboratory of Random Complex Structures, Data Science, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100000, China*
[*]*Correspondence to be sent to: Department of Mathematics, Beijing Jiaotong University, Beijing, 100044, China; E-mail: ytliu@bjtu.edu.cn (Yuting Liu); Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, China; E-mail: zhutq@amss.ac.cn (Tianqi Zhu); and Department of Genetics, Evolution and Environment, University College London, Gower Street, London WC1E 6BT, UK; E-mail: z.yang@ucl.ac.uk (Ziheng Yang)*

Received 14 August 2020; reviews returned 9 December 2020; accepted 15 December 2020
Associate Editor: Mark Holder

Abstract.—The phylogenetic bootstrap is the most commonly used method for assessing statistical confidence in estimated phylogenies by non-Bayesian methods such as maximum parsimony and maximum likelihood (ML). It is observed that bootstrap support tends to be high in large genomic data sets whether or not the inferred trees and clades are correct. Here, we study the asymptotic behavior of bootstrap support for the ML tree in large data sets when the competing phylogenetic trees are equally right or equally wrong. We consider phylogenetic reconstruction as a problem of statistical model selection when the compared models are nonnested and misspecified. The bootstrap is found to have qualitatively different dynamics from Bayesian inference and does not exhibit the polarized behavior of posterior model probabilities, consistent with the empirical observation that the bootstrap is more conservative than Bayesian probabilities. Nevertheless, bootstrap support similarly shows fluctuations among large data sets, with no convergence to a point value, when the compared models are equally right or equally wrong. Thus, in large data sets strong support for wrong trees or models is likely to occur. Our analysis provides a partial explanation for the high bootstrap support values for incorrect clades observed in empirical data analysis. [Bootstrap; model selection; star-tree paradox; support value.]

Introduction

Recently, Yang and Zhu (2018) characterized the asymptotic behaviors of Bayesian model selection in large data sets. When two models are both right or are equally wrong and indistinct, the posterior model probability varies among data sets according to a statistical distribution such as $\mathbb{U}(0,1)$, whereas one might expect it to converge to the point value $\frac{1}{2}$. Even more disturbingly, when the two models are equally wrong and distinct, the posterior model probability approaches $\sim$100% in some data sets and 0% in others. This polarized behavior may be a major reason for the observation that in Bayesian analysis of large phylogenetic data sets, posterior probabilities for trees or clades are often close to 100%, whether or not the relationships are correct. Note that in this article, we take the view that phylogeny reconstruction is a problem of statistical model selection, rather than one of parameter estimation under a well-specified model: different tree topologies are different statistical models while branch lengths on a given tree topology are parameters in the model (Yang 1996, 1997, 2000).

For non-Bayesian methods including maximum parsimony (Fitch 19710, neighbor joining (Saitou and Nei 1987), and maximum likelihood (ML, Felsenstein 1981), confidence for inferred trees or clades is most often assessed using Felsenstein's (1985) phylogenetic bootstrap. An interesting question is whether bootstrap exhibits similar behaviors as the posterior model probabilities. In modern phylogenomic studies, both posterior probabilities and bootstrap support values are often very high, whether or not the clades or trees are correct. Such results lead to widespread mistrust for such support values in large data sets. For example, Chan et al. (2020) wrote that "high bootstrap support did not necessarily reflect congruence or support for the correct topology. This study reiterates findings of some previous studies, which demonstrated that traditional bootstrap values can produce positively misleading measures of support in large phylogenomic data sets."

Bootstrap was originally developed by Efron (1979) to calculate the standard error for a parameter, by resampling the original data and studying the variation among the bootstrap resample data sets. It has since been used to conduct all sorts of analyses in Frequentist statistics, such as correction for bias, calculation of standard errors, construction of confidence intervals, and performing significance tests (Efron and Tibshirani 1993; Davison and Hinkley 1997). In phylogenetics, bootstrap was introduced by Felsenstein (1985) to assess the confidence in estimated phylogenetic trees. Although it follows the same operational procedure of resampling data points from the observed data set, bootstrap in phylogenetics differs from its use in bias correction or in confidence interval construction, in that a statistical interpretation has been illusory despite numerous efforts (Zharkikh and Li 1992; Felsenstein and Kishino 1993; Hillis and Bull 1993; Berry and Gascuel 1996; Efron et al. 1996; Holmes 2003; Susko 2009). Modifications to the procedure have also been made, including the complete-and-partial bootstrap (Zharkikh and Li, 1995), correction for first-order biases (Susko, 2010), or adjustment for short branches (Lemoine et al., 2018). These correct for the perceived bias in the procedure or to make it agree better with standard ideas of confidence levels and hypothesis testing. Its interpretation aside, phylogenetic bootstrap is the most widely used procedure for assessing the confidence in estimated phylogenies by non-Bayesian methods. Felsenstein's 1985 paper is a citation classic in all sciences. For Bayesian methods, the posterior probability for the inferred tree provides a natural measure of uncertainty (Rannala and Yang, 1996), and bootstrap is in theory not needed in Bayesian inference. However, the sensitivity of Bayesian model choice to the prior (O'Hagan and Forster, 2004) and the polarized behavior of Bayesian model selection under model misspecification (Yang and Zhu, 2018) have prompted the application of bootstrap in Bayesian model selection as well, leading to methods such as Bayesian bagging (Rubin, 1981; Weng, 1989; Huggins and Miller, 2020). It is important to study the asymptotic behavior of phylogenetic bootstrap. Earlier simulation studies suggest that the phylogenetic bootstrap may be conservative, and that 70% (instead of 95%) means strong support (e.g., Hillis and Bull, 1993). It has been noted that bootstrap support is numerically less extreme than posterior model probabilities (e.g., Huelsenbeck and Rannala, 2004; Yang and Rannala, 2005).

In this article, we explore the asymptotic behavior of phylogenetic bootstrap when the data size increases. We consider phylogenetic reconstruction as a statistical model selection problem, and treat phylogenetic trees as nonnested statistical models (rather than different values of a parameter in a well-specified model). We present an asymptotic theory for bootstrap model probability under different scenarios in the Appendix and in the main paper illustrate the theory using canonical problems that are analytically tractable. We discuss phylogenetic reconstruction problems in the case of three or four taxa to illustrate the general theory.

Summary of Analytical Results

Following Felsenstein and Kishino (1993) and Efron et al. (1996), we consider bootstrap as a general approach to assessing the confidence in the selected model in a model-selection problem.

Bootstrap in Model Selection

The data are an independently and identically distributed (i.i.d.) sample of size n, $x=\{x_1,\ldots,x_n\}$, from the true data-generating model $g(X)$. We compare K models, H_j, $j=1,\ldots,K$. Model H_j specifies the density $f_j(X|\theta_j)$ with parameters θ_j. Let $\hat{\theta}_j$ be the MLE of θ_j under model H_j given data x. When $n\to\infty$, $\hat{\theta}_j\to\theta_{j*}$, where θ_{j*} minimizes the Kullback–Leibler (K–L) divergence from model H_j to the true model,

$$D_j=\int g(X)\log\frac{g(X)}{f_j(X|\theta_{j*})}\mathrm{d}X. \quad (1)$$

If H_j is correct, θ_{j*} will be the *true parameter values*, with $D_j=0$. Otherwise if H_j is wrong, θ_{j*} will be the *best-fitting* or *pseudo-true parameter values*, with $D_j>0$. In this article, we focus on the case where all K models have the same K–L divergence to the true model. Two models f_1 and f_2 are said to be equally right if $D_1=D_2=0$, and equally wrong if $D_1=D_2>0$. If two models are unidentifiable at their pseudo-true parameter values, that is, if

$$f_1(X|\theta_{1*})=f_2(X|\theta_{2*}) \quad \text{for almost every } X, \quad (2)$$

they are said to be indistinct. This can occur when both models are right (with $D_1=D_2=0$) or when both are wrong (with $D_1=D_2>0$). Otherwise if equation (2) does not hold for some X of nonzero measure, the models are said to be distinct. This can occur only if both models are wrong (with $D_1=D_2>0$).

The model selected by ML is the one that achieves the greatest log likelihood, $\ell_j(\hat{\theta}_j)$ = $\log f_j(x|\hat{\theta}_j)$. To assess the confidence on the selected model, we calculate the bootstrap probability. Let $x_b^*=\{x_{b1}^*,\ldots,x_{bn}^*\}$ be a bootstrap sample, formed by resampling with replacement n times from the original data x. Let $\hat{\theta}_b^*$ be the MLE from a bootstrap sample x_b^*. Here, we follow the convention of using the superscript * to indicate a bootstrap sample, and the subscript $_*$ for the true or pseudotrue parameter values. We assume that θ_{j*}, $\hat{\theta}_j$, and $\hat{\theta}_j^*$ are inner points in the parameter space. The proportion of bootstrap replicates in which model j is the optimal model is the bootstrap probability or bootstrap support P_j for model j. For example, in the case of two models, the bootstrap probability for model H_1 is

$$\begin{aligned}P_1(x)&=\mathbb{P}\left\{\log f_1(x^*|\hat{\theta}_1^*)>\log f_2(x^*|\hat{\theta}_2^*)|x\right\}\\&\approx\frac{1}{B}\sum_b \mathbb{I}_{\ell_1(\hat{\theta}_1^*)>\ell_2(\hat{\theta}_2^*)},\end{aligned} \quad (3)$$

where $\ell_j(\hat{\theta}_j^*)=\log f_j(x^*|\hat{\theta}_j^*)$ is the log likelihood value for model j, calculated at the MLE $(\hat{\theta}_j^*)$ and where the indicator function $\mathbb{I}_A$ is 1 if A is true or 0 otherwise. Note that P_1 is a function of x and is a random variable. We are interested in the asymptotic distribution of P_1 when x varies.

In phylogenetics, the models under comparison are the tree topologies for the given set of taxa, while each data point corresponds to one site or one column in the alignment. While the bootstrap is applicable as long as the inference method is statistically consistent (Felsenstein, 1985), we focus on ML in this article. In phylogenetics, bootstrap is commonly used to attach support values for clades or splits on the phylogeny, calculated as the proportion of bootstrap trees that contain the splits. Here, we focus on the bootstrap probability for the whole model. In the case of simple trees with three or four species with only one internal branch, the two measures are equivalent. We assume that the number of bootstrap replicates B is large so that the sampling errors due to limited number of bootstrap replicates is negligible.

The Asymptotic Behavior of Bootstrap Model Selection Under Different Scenarios

We develop an asymptotic theory of bootstrap model selection in the Appendix. In general, when equally right or equally wrong models are compared, bootstrap model probabilities have a nondegenerate distribution. In the case of two equally wrong and distinct models, the bootstrap model probability P_1 has the distribution $\mathbb{U}(0,1)$.

The case of two equally wrong and distinct models with no parameters provides valuable insights into the differences between bootstrap and Bayesian methods. The log-likelihood ratio between the two models is

$$\Delta \equiv \log\frac{f_1(x)}{f_2(x)}, \quad \Delta^* \equiv \log\frac{f_1(x^*)}{f_2(x^*)}, \tag{4}$$

for the original data x and the bootstrap resample data x^*, respectively. Each of these is a sum of n i.i.d. terms. Thus $\mathbb{E}(\Delta) = n\mathbb{E}\log f_1(X) - n\mathbb{E}\log f_2(X) = n(D_2 - D_1) = 0$ (Equation 1). Let

$$\sigma^2 = \mathbb{V}\left\{\log\frac{f_1(X)}{f_2(X)}\right\} = \int g(X)\left[\log\frac{f_1(X)}{f_2(X)}\right]^2 \mathrm{d}X. \tag{5}$$

When $n \to \infty$, $\Delta \sim \mathbb{N}(0, n\sigma^2)$ and $\Delta^*|x \sim \mathbb{N}(\Delta, n\sigma^2)$, according to the central limit theorem. Thus

$$P_1 = \mathbb{P}\{\Delta^* > 0|x\} \approx \Phi\left(\tfrac{\Delta}{\sqrt{n}\sigma}\right) \to \mathbb{U}(0,1), \tag{6}$$

where Φ is the cumulative distribution function (CDF) for $\mathbb{N}(0,1)$.

In Bayesian comparison of two equally wrong models with no parameters, Δ is the log Bayes factor. With equal prior probabilities ($\frac{1}{2}$ for each model), this is related to the posterior model probability through $\Delta = \log\frac{P_1}{1-P_1}$ or $P_1 = \frac{\mathrm{e}^{\Delta}}{\mathrm{e}^{\Delta}+1}$. As Δ behaves like a random walk when n increases, it is nearly impossible for Δ to be in a small interval around 0, say, $-5 < \Delta < 5$ which corresponds to $0.007 < P_1 < 0.993$. In other words, for large n, the posterior probability will be 0 in half of the data sets and 1 in the other half. This polarized behavior also occurs when the compared models, equally wrong and distinct, have parameters as the Bayes factor is dominated by the random-walk term (Yang and Zhu, 2018). The analysis here suggests that bootstrap probability has a qualitatively different behavior, as it contrasts Δ^* for the bootstrap sample with Δ for the original data.

Illustrative Examples

We present several simple examples to illustrate the asymptotic behavior of bootstrap model probability under different scenarios when the data size $n \to \infty$. In the first two examples, two models are equally wrong and distinct, and the bootstrap probability P_1 varies among data sets like a random number, $P_1 \sim \mathbb{U}(0,1)$ (Equation 6).

Problem 1 Fair-coin paradox, with equally wrong models and no parameters. Suppose a coin is fair with the true probability of heads to be $p = 0.5$, and we flip the coin n times to compare two models $H_1 : p = 0.4$ and $H_2 : p = 0.6$. The data set is $x = \{x_1, \ldots, x_n\}$, where x_i takes the value 1 for heads and 0 for tails, and has the Bernoulli distribution. The data can be summarized as the proportion of heads in n tosses, $\bar{x}$, which is approximately normal $\mathbb{N}(\frac{1}{2}, \frac{1}{4n})$. H_1 is favored if $\bar{x} < \frac{1}{2}$, and this happens in half of the data sets.

Given x, the bootstrap sample $x_b^* = \{x_{b1}^*, \cdots, x_{bn}^*\}$, where x_{bi}^* is a Bernoulli variable with probability $\bar{x}$, can be summarized as the bootstrap sample mean $\bar{x}^*$, which is approximately normal, with $\bar{x}^*|x \sim \mathbb{N}(\bar{x}, \frac{\bar{x}(1-\bar{x})}{n}) \approx \mathbb{N}(\bar{x}, \frac{1}{4n})$. The bootstrap sample x_b^* favors model H_1 if and only if $\bar{x}^* < 1/2$. Thus,

$$P_1 = \mathbb{P}\{\bar{x}^* < \tfrac{1}{2}|x\} \approx \Phi\left(\frac{1/2 - \bar{x}}{\sqrt{1/(4n)}}\right) \to \mathbb{U}(0,1), \quad \text{as } n \to \infty. \tag{7}$$

Thus P_1 varies like a random number among data sets (Fig. 1a). Alternatively, we have $\Delta = \ell_1 - \ell_2 = 2n(\bar{x} - \frac{1}{2})\log\frac{0.4}{0.6} \sim \mathbb{N}(0, n\sigma^2)$ and $\Delta^*|x \sim \mathbb{N}(\Delta, n\sigma^2)$, with $\sigma = \log\frac{0.4}{0.6}$, so that Equation 6 gives $P_1 \sim \mathbb{U}(0,1)$.

Problem 2 Normal distribution, equally wrong and distinct models with free parameters. Suppose the true model is $\mathbb{N}(0,1)$ and we consider $H_1 : \mathbb{N}(\mu, 1/\tau_1)$ and $H_2 : \mathbb{N}(\mu, 1/\tau_2)$, where μ is a free parameter while the precisions τ_1 and τ_2 are given with $\log(\tau_2/\tau_1) = \tau_2 - \tau_1$ so that the two models are equally wrong ($D_1 = D_2 > 0$) (Yang and Zhu, 2018). We use $\tau_1 = 0.25$ and $\tau_2 = 2.58666$. Under each model, the pseudo-true parameter value is $\mu_* = 0$ and H_1 and H_2 are two equally wrong and distinct models. Note that H_1 is overdispersed and H_2 is underdispersed. Under the model $\mathbb{N}(\mu, 1/\tau)$ with known τ, the log likelihood is

$$\ell = -\frac{n}{2}\log(2\pi) + \frac{n}{2}\log\tau - \frac{\tau}{2}\sum_{i=1}^{n}(x_i - \mu)^2, \tag{8}$$

with $\hat{\mu} = \bar{x}$. Thus, $\ell_1 > \ell_2$ if and only if $(\tau_2 - \tau_1)\sum_{i=1}^{n}(x_i - \bar{x})^2 > n\log(\tau_2/\tau_1)$ or if and only if $s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 > 1$. We have $ns^2 \sim \chi^2_{n-1} \approx \mathbb{N}(n-1, 2(n-1))$.

Given x, the bootstrap sample x^* favors H_1 if the sample variance $s^{2*} > 1$. We have $ns^{2*}/s^2|x \sim \chi^2_{n-1} \approx \mathbb{N}(n-1, 2(n-1))$. For large n, re-sampling from the empirical distribution represented by the observed data x is approximately equivalent to sampling from the

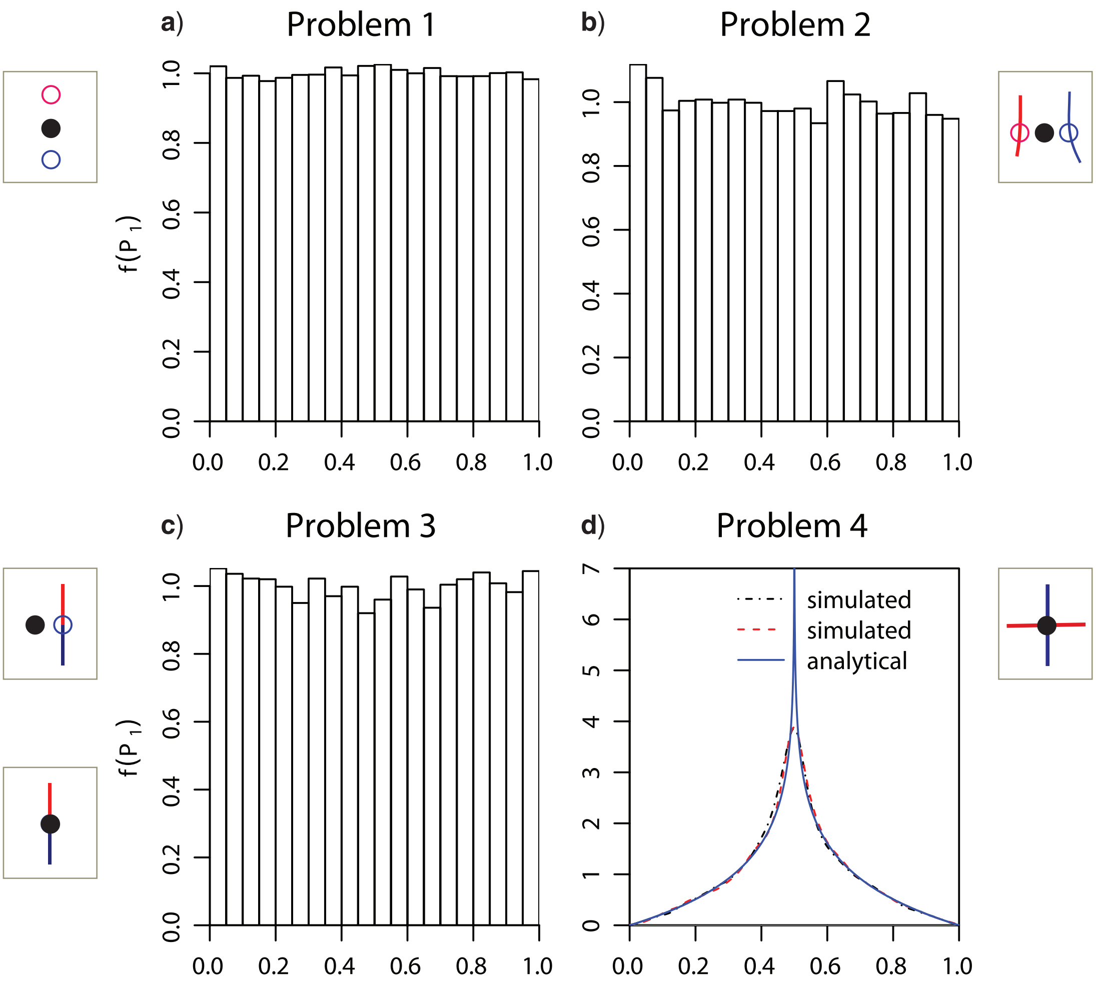

FIGURE 1. Histogram/density of bootstrap model probability P_1 in problems that involve comparisons of two models. The insets next to the density plots illustrate the type of the inference problem represented by the example, where the black circle represents the true model, the two empty circles represent the best-fitting parameter values under the two models, and the two lines represent the parameter space for the two models (see Yang and Zhu, 2018, Fig. 2). a) Problem 1 (the fair-coin paradox) in which a fair coin (with $p=0.5$) is tossed n times to compare two models: $p=0.4$ and $p=0.6$. Here the two models are equally wrong and distinct and have no free parameters. b) Problem 2 in which the true model is $\mathbb{N}(0,1)$ while the two compared models, $\mathbb{N}(\mu,1/\tau_1)$ and $\mathbb{N}(\mu,1/\tau_2)$ with $\tau_1<1<\tau_2$, are equally wrong and distinct, and each has one free parameter. c) Problem 3 (the fair-balance paradox) where the true model is $\mathbb{N}(0,1)$ and the two compared models, $\mathbb{N}(\mu,1/\tau)$, $\mu<0$ versus $\mathbb{N}(\mu,1/\tau)$, $\mu>0$, are equally right (if $\tau=1$) or equally wrong and indistinct (if $\tau\neq 1$). d) Problem 4 (equally right models). The true model is $\mathbb{N}(0,1)$ while the two compared models, $\mathbb{N}(\mu,1)$ versus $\mathbb{N}(0,1/\tau)$, are both right. Black dashed line is for the expensive simulation generating x and x^*, the red dashed line is for simulation generating $\bar{x}$ and s^2, while blue solid line is for the analytical approximation by Equation 14. The insets characterize the problems, with the true models represented as filled circles and the pseudo-true parameter values as empty circles, while the lines represent the parameter space for each model. The settings are $n=10^5, B=3\times10^4$, and $R=10^5$ for problem 1, $n=10^4, B=3\times10^4$, and $R=10^4$ for problem 2, $n=10^4, B=3\times10^4$, and $R=10^4$ for problem 3, and $n=10^4, B=10^3$, and $R=10^4$ for problem 4.

continuous distribution $\mathbb{N}(\bar{x},s^2)$. Thus,

$$P_1=\mathbb{P}\{s^{2*}>1|x\}\approx\Phi\left(\frac{((n-1)/n)s^2-1}{\sqrt{2(n-1)/n^2}}\right)\to\mathbb{U}(0,1),\quad \text{as } n\to\infty. \tag{9}$$

This is confirmed in Figure 1b.

Alternatively, we have $\Delta=\ell_1-\ell_2=\frac{n}{2}(\tau_2-\tau_1)(s^2-1)\sim\mathbb{N}(0,n\sigma^2)$ and $\Delta^*|\Delta\sim\mathbb{N}(\Delta,n\sigma^2)$, with $\sigma=\frac{1}{\sqrt{2}}(\tau_2-\tau_1)$, so that $P_1=\mathbb{P}\{\Delta^*>0|x\}\sim\mathbb{U}(0,1)$.

If the two compared models are both right (with $D_1=D_2=0$) or are equally wrong and indistinct (with $D_1=D_2>0$), then P_1 varies among data sets according to a nondegenerate distribution, which may and may not be $\mathbb{U}(0,1)$), as illustrated in the next two examples.

Problem 3 Fair-balance paradox with two equally right or equally wrong and indistinct models. The true model is $\mathbb{N}(0,1)$ and the two compared models are $\mathbb{N}(\mu,1/\tau)$, $\mu<0$ and $\mathbb{N}(\mu,1/\tau)$, $\mu>0$, with τ given. If $\tau=1$, the two models are equally right. If $\tau\neq1$, the two models are equally wrong (because of the assumed incorrect variance) and indistinct (because the pseudo-true parameter value $\mu_*=0$ under each model). Model 1 is favored if and only if the sample mean $\bar{x}<0$. As

$\bar{x} \sim \mathbb{N}(0, 1/(n\tau))$ and $\bar{x}^*|x \sim \mathbb{N}(\bar{x}, 1/(n\tau))$, we have

$$P_1 = \mathbb{P}\{\bar{x}^* < 0|x\} = \Phi(-\sqrt{n\tau}\bar{x}) \to \mathbb{U}(0,1), \quad \text{as } n \to \infty. \tag{10}$$

This is confirmed in Figure 1c.

Problem 4 Normal-distribution example with an infinite spike at $\frac{1}{2}$ in the P_1 distribution. The true model is $\mathbb{N}(0,1)$ and the two compared models are $\mathbb{N}(\mu, 1)$ and $\mathbb{N}(0, 1/\tau)$. In H_1, $\mu_* = 0$ while in H_2, $\tau_* = 1$, so the two models are equally right. The data x may be summarized as the sample mean $\bar{x}$ and sample variance $s^2 = \frac{1}{n}\sum_i (x_i - \bar{x})^2$. The MLE of the parameter is $\hat{\mu} = \bar{x}$ under H_1 and $\hat{\tau} = n/\sum x_i^2 = 1/(s^2 + \bar{x}^2)$ under H_2. The log-likelihood values are

$$\begin{aligned} \ell_1(\hat{\mu}) &= -\tfrac{1}{2}\sum (x_i - \bar{x})^2 = -\tfrac{1}{2} n s^2, \\ \ell_2(\hat{\tau}) &= -\tfrac{n}{2}\log\left(\tfrac{1}{n}\sum x_i^2\right) - \tfrac{n}{2} = -\tfrac{n}{2}\log(s^2 + \bar{x}^2) - \tfrac{n}{2}. \end{aligned} \tag{11}$$

Thus, $\ell_1 > \ell_2$ if and only if

$$\bar{x}^2 > e^{s^2 - 1} - s^2 \approx 1 + (s^2 - 1) + \tfrac{1}{2}(s^2 - 1)^2 - s^2 = \tfrac{1}{2}(s^2 - 1)^2, \tag{12}$$

or if and only if

$$|\bar{x}| > \frac{1}{\sqrt{2}}|s^2 - 1|. \tag{13}$$

A large deviation of $\bar{x}$ from 0 supports H_1, whereas a large deviation of s^2 from 1 favors H_2. Also $\bar{x} \sim \mathbb{N}(0, \frac{1}{n})$ and $s^2 \sim \frac{1}{n}\chi^2_{n-1} \approx \mathbb{N}(\frac{n-1}{n}, \frac{2(n-1)}{n^2})$ or $\frac{1}{\sqrt{2}}(s^2 - 1) \sim \mathbb{N}(0, \frac{1}{n})$, and $\bar{x}$ and s^2 are independent. Thus, Equation 13 holds and H_1 is the selected model in half of the data sets.

Given x, we have $\bar{x}^*|x \sim \mathbb{N}(\bar{x}, s^2/n) \approx \mathbb{N}(\bar{x}, \frac{1}{n})$ and $\frac{1}{\sqrt{2}}(s^{2*} - 1)\Big|x \sim \mathbb{N}(\frac{1}{\sqrt{2}}(s^2 - 1), \frac{1}{n})$ and $\bar{x}^*$ and s^{2*} are conditionally independent. Let $z_1 = \sqrt{n}\bar{x}$ and $z_2 = \sqrt{\frac{n}{2}}(s^2 - 1)$, with z_1 and z_2 from $\mathbb{N}(0,1)$. Let $z_1^* = \sqrt{n}\bar{x}^*$ and $z_2^* = \sqrt{\frac{n}{2}}(s^{2*} - 1)$, with $z_1^*|x \sim \mathbb{N}(z_1, 1)$ and $z_2^*|x \sim \mathbb{N}(z_2, 1)$ to be conditionally i.i.d. Then,

$$P_1 = \mathbb{P}\{|z_1^*| > |z_2^*| \big| x\}. \tag{14}$$

This problem is analyzed in the Supplementary information text available on Dryad at https://doi.org/10.5061/dryad.7m0cfxprw. The limiting distribution of P_1 when $n \to \infty$ is

$$f(P_1) = -\log|2P_1 - 1|. \tag{15}$$

The density is symmetrical around $\frac{1}{2}$, is 0 at 0 and 1, and has an infinite spike at $\frac{1}{2}$, with the mean $\frac{1}{2}$ and variance $\frac{1}{36}$. This is confirmed by simulation in Figure 1d. The simulation is done in two ways. In the first, data x is sampled from $\mathbb{N}(0,1)$, and given x bootstrap samples x_b^* are generated, with $\bar{x}^*$ and s^{2*} calculated to apply Equation 14. In the second approach, $\bar{x} \sim \mathbb{N}(0, 1/n)$ and $ns^2 \sim \chi^2_{n-1}$ are sampled, and then $\bar{x}^* \sim \mathbb{N}(\bar{x}, s^2/n)$ and $ns^{2*}/s^2 \sim \chi^2_{n-1}$ are generated to select the model for the bootstrap sample using Equation 14. Both approaches produce the same results as Equation 15.

Problem 5 Multivariate normal-distribution example. The true model is the $(K-1)$-variate normal distribution $\mathbb{N}(\mu, \Sigma)$, with mean vector $\mu = (\mu_1, \ldots, \mu_{K-1})$ where $\mu_1 = \cdots = \mu_{K-1} = 0$ and variance matrix Σ which has 1 on the diagonal and $-1/(K-1)$ on the off-diagonal. The data are an i.i.d. sample of size n, $x = \{x_{ij}\}$, $i = 1, \ldots, n; j = 1, \ldots, K-1$. Also let $x_{iK} = -(x_{i1} + \cdots + x_{i,K-1})$ and $\mu_K = -(\mu_1 + \cdots + \mu_{K-1})$. We use the data to compare K models. Model H_j, $j = 1, \ldots, K$, assumes $\mu_j > \mu_k$ for any $k \neq j$. The model has $K-1$ free parameters: $\mu_1, \ldots, \mu_K$ with the constraint $\mu_1 + \cdots + \mu_K = 0$. The variance is assumed to be known, $c\Sigma$. The models are equally right if $c = 1$ and equally wrong if $c \neq 1$. An alternative formulation of the problem is to have only one parameter in model H_j: $\mu_j > \mu_k$ with $\mu_k = -\mu_j/(K-1)$ for all $k \neq j$.

Let $\bar{x} = \{\bar{x}_j\}$ and $\bar{x}^* = \{\bar{x}_j^*\}$, with

$$\bar{x}_j = \frac{1}{n}\sum_i x_{ij}, \quad \bar{x}_j^* = \frac{1}{n}\sum_i x_{ij}^*, \quad j = 1, \cdots, K, \tag{16}$$

be the sample means from data set x and from bootstrap sample x^*, respectively. Then $\bar{x} \sim \mathbb{N}(\mu, \frac{1}{n}\Sigma)$ and approximately $\bar{x}^*|x \sim \mathbb{N}(\bar{x}, \frac{1}{n}\Sigma)$. Without the constraint under each model H_j: $\mu_j > \mu_k$, the MLEs of μ are the sample means. With the constraint, H_j is the selected model if $\bar{x}_j$ is the greatest among $\bar{x}_1, \ldots, \bar{x}_K$. The bootstrap probability for model H_1 given data x is

$$P_1 = \mathbb{P}(\bar{x}_1^* > \bar{x}_2^*, \ldots, \bar{x}_1^* > \bar{x}_K^* | x). \tag{17}$$

Now for any $j \neq k$,

$$\sigma_{jk}^2 = \mathbb{V}(\bar{x}_j - \bar{x}_k) = \frac{2}{n} - 2 \cdot \frac{1}{n} \cdot \left(-\frac{1}{K-1}\right) = \frac{2}{n} \cdot \frac{K}{K-1}. \tag{18}$$

Let $z = (z_2, \ldots, z_K)^T$ and $z^* = (z_2^*, \ldots, z_K^*)^T$, with $z_j = \frac{\bar{x}_1 - \bar{x}_j}{\sigma_{1j}}$ and $z_j^* = \frac{\bar{x}_1^* - \bar{x}_j^*}{\sigma_{1j}}$, $j = 2, \ldots K$. We have

$$\begin{aligned} &\mathbb{V}(z_j) = 1, \\ &\mathrm{Cor}(z_j, z_k) = \mathrm{Cov}(\bar{x}_1 - \bar{x}_j, \bar{x}_1 - \bar{x}_k)/(\sigma_{1j}\sigma_{1k}) \\ &\qquad = [\mathbb{V}(\bar{x}_1) - 2\mathrm{Cov}(\bar{x}_1, \bar{x}_j) + \mathrm{Cov}(\bar{x}_i, \bar{x}_j)]/(\sigma_{1j}\sigma_{1k}) \\ &\qquad = \tfrac{1}{n}\left(1 + \tfrac{1}{K-1}\right) \Big/ \left(\tfrac{2}{n}\tfrac{K}{K-1}\right) = \tfrac{1}{2}. \end{aligned} \tag{19}$$

Thus, $z \sim \mathbb{N}(0, \Sigma_0)$ and $z^*|x \sim \mathbb{N}(z, \Sigma_0)$, where Σ_0 is a $(K-1) \times (K-1)$ variance matrix with 1 on the diagonal and $\frac{1}{2}$ on the off-diagonal. Thus,

$$P_1 = \mathbb{P}(z_2^* > 0, \ldots, z_K^* > 0|x) = \Phi(z_2, \ldots, z_K). \tag{20}$$

TABLE 1. Proportions of data replicates with very high bootstrap probability (P_1) in the multivariate normal example (Problem 5)

K	2	3	4	5
$\mathbb{P}\{P_1>0.90\}$	0.100	0.023	0.008	0.004
$\mathbb{P}\{P_1>0.95\}$	0.050	0.008	0.003	0.001
$\mathbb{P}\{P_1>0.99\}$	0.010	0.001	0.000	0.000

As $\bar{x}_j^*-\bar{x}_k^*=(\bar{x}_1^*-\bar{x}_k^*)-(\bar{x}_1^*-\bar{x}_j^*)$, the bootstrap probabilities for all K models given data x are

$$\begin{bmatrix} P_1 \\ P_2 \\ \vdots \\ P_K \end{bmatrix} = \begin{bmatrix} \Phi(z_2,z_3,\ldots,z_K) \\ \Phi(-z_2,z_3-z_2,\ldots,z_K-z_2) \\ \vdots \\ \Phi(-z_K,z_2-z_K,\ldots,z_{K-1}-z_K) \end{bmatrix} \qquad (21)$$

For example, in the case of $K=3$, a fast way of simulating the limiting distribution of (P_1,P_2,P_3) is thus to generate $(z_2,z_3)\sim\mathbb{N}(\binom{0}{0},\left(\begin{smallmatrix}1 & 0.5\\ 0.5 & 1\end{smallmatrix}\right))$ and then calculate (P_1,P_2,P_3) by Equation 21. This is confirmed by the slow simulation of generating x and then x^* in Figure 2. The joint distribution of (P_1,P_2,P_3) has peaks at the three corners, and is nearly flat around the center. By symmetry P_1 has mean $\frac{1}{3}$, and by numerical integration using Equation 20, P_1 has SD = 0.25904. The probability that one of the models is strongly supported is close to 0 (Table 1). Figure 3a,b shows the marginal distribution of P_1 when $K=3$ and 6.

Even though (P_1,P_2,P_3) do not converge to the point value $(\frac{1}{3},\frac{1}{3},\frac{1}{3})$, extreme bootstrap support values are not highly frequent. Bootstrap probabilities are thus qualitatively different from Bayesian model probabilities.

BOOTSTRAP IN PHYLOGENETICS

We consider ML reconstruction of phylogenies of three or four species (Fig. 4), under the JC model (Jukes and Cantor, 1969). We simulate data to verify the asymptotic theory and compare with Bayesian results from Yang and Zhu (2018).

Case A (Fig. 5A and A′) involves equally right models. This is the star-tree paradox analyzed previously (Lewis et al., 2005; Yang and Rannala, 2005; Yang, 2007a; Susko, 2008). We use the rooted star tree T_0 for three species with $t=0.2$ (Fig. 4a) to generate data sets to compare the three binary trees. The JC model (Jukes and Cantor, 1969) is used both to generate and to analyze the data. The molecular clock (rate constancy over time) is assumed as well, so that the parameters in each binary tree are the two node ages (t_0 and t_1), measured by the expected number of nucleotide changes per site. The best-fitting parameter values are $t_{0*}=0$ and $t_{1*}=0.2$ for each of the three binary trees, so that the three binary trees are equally right models.

Case B (Fig. 5B and B′) involves equally wrong models that are indistinct. This is similar to case A except that the JC+Γ model (Jukes and Cantor, 1969; Yang, 1993) is used to generate data, with different sites in the sequence evolving at variable rates according to the gamma distribution with shape parameter $\alpha=1$. The data are then analyzed using JC (equivalent to JC+Γ with $\alpha=\infty$), giving $t_{0*}=0$ and $t_{1*}=0.16441$ as the pseudo-true parameter values for each binary tree. The binary trees are equally wrong and indistinct models ($D_1=D_2=D_3>0$).

Case C (Fig. 5C&C′) involves equally wrong and distinct models. Like case B, the simulation model is JC+Γ with $\alpha=1$ and the analysis model is JC. However, the molecular clock is not assumed and unrooted trees are used. The true tree is the unrooted star tree T_0 of Fig. 4B, with $t_1=t_2=t_3=t_4=0.2$, with $t_{0*}=0.01037$ and $t_{i*}=0.16409$, $i=1,\cdots,4$ for the binary trees (Fig. 4B). As $t_{0*}>0$, the three binary trees are equally wrong and distinct models ($D_1=D_2=D_3>0$).

In cases A and B, the data for the three species have a multinomial distribution with five categories corresponding to the five site patterns xxx, xxy, xyx, yxx, and xyz, where x,y,z are any distinct nucleotides. Let the frequencies of the informative site patterns xxy, xyx, yxx be $\bar{x}_1$, $\bar{x}_2$, and $\bar{x}_3$, while that for the two uninformative patterns xxx and xyz be $\bar{x}_0$. With the star tree being the true tree, the probabilities for the three informative site patterns are identical, with $p_1=p_2=p_3$. Tree 1 specifies $p_1>p_2=p_3$. Given data x, tree j is the ML tree if $\bar{x}_j$ is the greatest among $\bar{x}_1,\bar{x}_2$, and $\bar{x}_3$ (Yang, 2000). Then $\bar{x}=(\bar{x}_1,\bar{x}_2,\bar{x}_3)$ is approximately normal, with mean (p,p,p), and variance $p(1-p)/n$ and covariance $-p^2/n$. Applying a multivariate normal approximation to the multinomial distribution, we see that the problem has the same mathematical structure as problem 5. Thus the bootstrap distribution for cases A and B should be identical to that in problem 5. We wrote a C program to simulate and analyze data for cases A and B. Given branch lengths t_0 and t_1, the probabilities for the five site patterns are calculated according to the JC model (Yang, 1994), and the data x are then generated by sampling from the multinomial distribution. Given data x, bootstrap dataset x^* is sampled using the observed site-pattern frequencies in x. Then tree j is the ML tree for data x^* if $\bar{x}_j^*$ is the largest among $(\bar{x}_1^*,\bar{x}_2^*,\bar{x}_3^*)$.

In case C for four species, the informative site patterns are $xxyy$, $xyxy$ and $xyyx$ while there are 11 uninformative patterns. The binary tree has only five parameters, such that the model achieves a better fit to the observed data by having a positive internal branch length. As a result, the three binary trees are distinct models (with $t_{0*}>0$). Case C thus differs from problem 5, but has a similar symmetry in that the K-L distance between any pair of models is the same. From the general theory, the distribution of bootstrap probabilities (P_1,P_2,P_3) is the same as that in problem 5. We simulated data using EVOLVER, and generated bootstrap resample data using SEQBOOT. The data are then analyzed using BASEML in PAML (Yang, 2007b).

Our theory predicts that the limiting distribution is the same in all three cases, with the mean 1/3 and SD 0.25904. This is confirmed by the simulation (Fig. 5), which gave

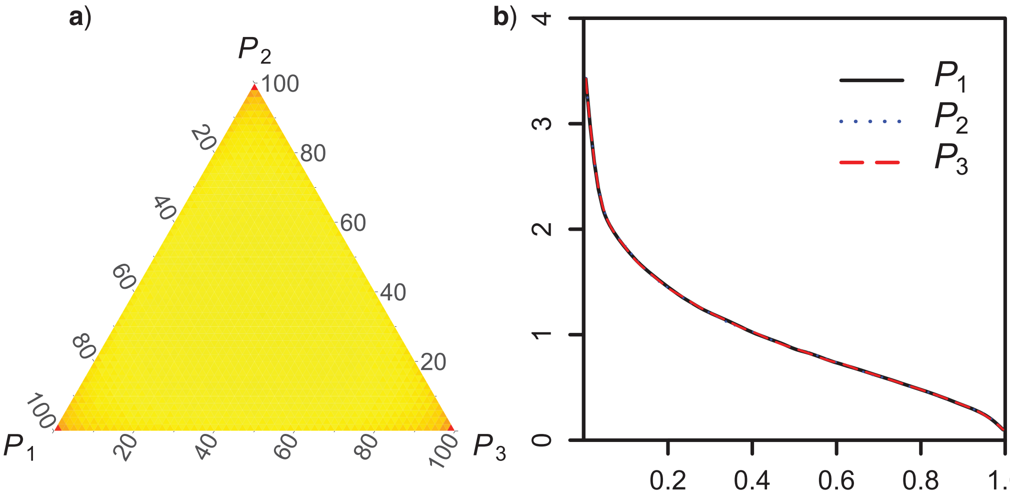

FIGURE 2. Marginal and joint distributions of P_1, P_2, P_3 for Problem 5 (the multivariate normal example with $K=3$). The three corners in the plots correspond to points (1, 0, 0), (0, 1, 0), and (0, 0, 1), while the center is $(\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$. The number of replicates is $R=10^6$, with $n=10^6$ and $B=10^3$.

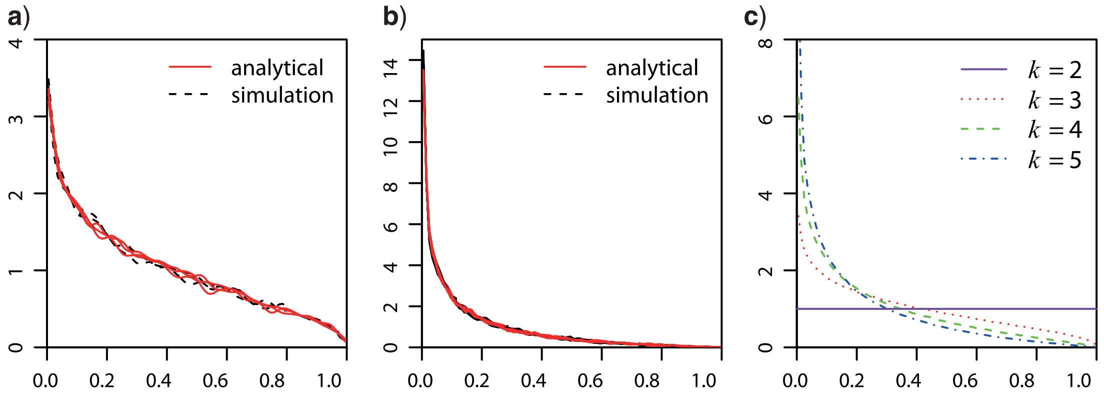

FIGURE 3. Marginal distribution of P_1 in comparisons of K equally right or equally wrong and indistinct models based on the normal distribution of Problem 5 ($K=3$ in a and 6 in B). The sample size is $n=10^4$. The number of simulated replicates is $R=10^4$, with $B=10^3$, but the "theoretical" distribution is based on simulating 10^6 replicates and using Equation 21.

the mean of P_1 as 1/3 and the SD as 0.259. The bootstrap probabilities have modes at the corners, and roughly uniformly distributed around the center. While in case C, the Bayesian posterior probabilities show extreme polarized behavior, concentrated on three points: (1, 0, 0), (0, 1, 0), and (0, 0, 1) (Yang and Zhu, 2018, Fig. 4C&4C′), bootstrap probabilities are much more moderate and have a nondegenerate distribution.

We calculated the proportions of data sets in which the bootstrap and posterior probabilities for the three binary trees are extremely low or extremely high (Table 2). As the three trees are equally right or equally wrong, both extremely low and extremely high support values are undesirable. Overall posterior probabilities are much more extreme, with extremely low and extremely high values to be commonly observed. In contrast, bootstrap probabilities are much more moderate, with extreme values to be rare. For example, at the sequence length $n=10^5$, $\mathbb{E}(P_{\max}) = 0.647$ using bootstrap method and 0.964 for the Bayesian method. If $P_{\max} > 0.95$, one of the models is strongly favored, and this occurs in 2.9% of data sets for the bootstrap and 85.3% for the Bayesian. It is much less likely to see high bootstrap support for equally wrong models than high posterior probabilities for them.

We conducted another simulation generating sequence data on a four-species tree under JC+Γ with α = 1 and analyzing them under JC, as in case C above (Fig. 5C&C′), but instead of the star tree, we used the binary unrooted tree T_1 with a very short internal branch: ((*a*: 0.2, *b*: 0.2) : 0.002, *c*: 0.2, *d*: 0.2) (see Fig. 4B). In this case one of the binary trees is correct while the other two are wrong. The results are summarized in Table 3. In large but finite data sets, the Bayesian method

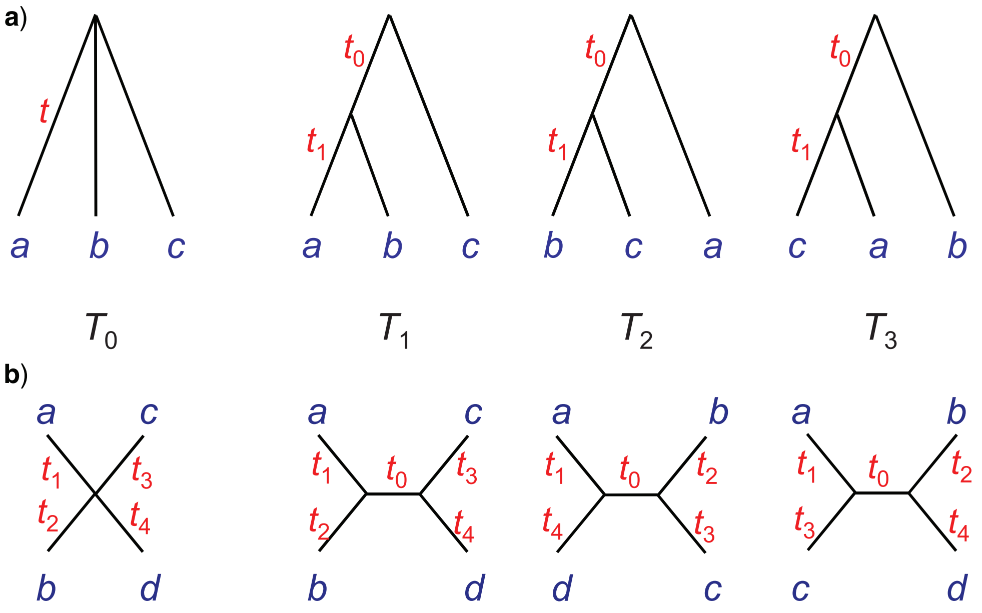

FIGURE 4. The star tree T_0 and three binary rooted trees T_1, T_2, and T_3 for a) three or b) four species. Branch length parameters, shown next to the branches, are measured by the expected number of changes per site. The star tree is used to generate data, which are analyzed by ML to compare the three binary trees.

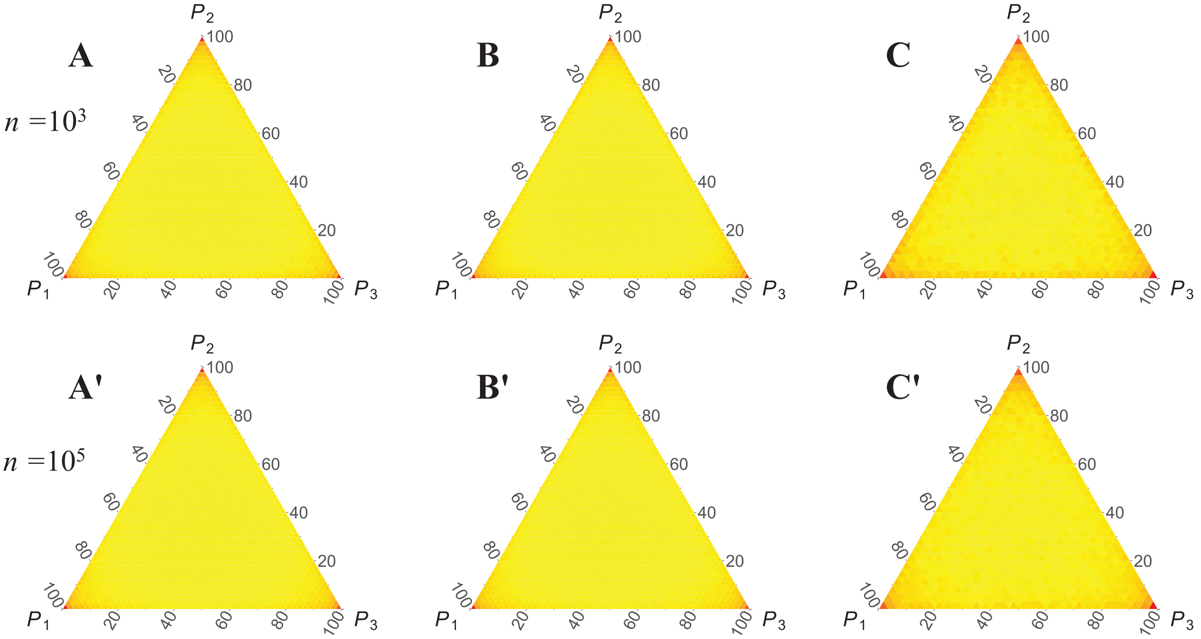

FIGURE 5. The joint distribution of the bootstrap model probabilities in the star-tree problem. The star tree T_0 of Fig. 4 is used to simulate data (sequence alignments of $n = 10^3$ or 10^5 sites), and ML is used to compare the three binary trees T_1, T_2, and T_3 to calculate their bootstrap probabilities (P_1, P_2, P_3). In (A) and (A′), the true tree is the star tree T_0 for the three species of Fig. 4A, with $t = 0.2$. Both the simulation and analysis models are JC, and the three binary trees are equally right models. In (B) and (B′), the true tree is the star tree T_0 for three species of Fig. 4A, with $t = 0.2$. The simulation model is JC+Γ (with $\alpha = 1$), and the analysis model is JC. The three binary trees represent equally wrong and indistinct models. In (C) and (C′), the true tree is the star tree T_0 for four species of Fig. 4B, with $t_1 = t_2 = t_3 = t_4 = 0.2$. The simulation model is JC+Γ ($\alpha = 1$) and the analysis model is JC. The three binary trees represent equally wrong and distinct models. The number of bootstrap samples $B = 1000$ and the number of replicates is $R = 10^6$ for three-species trees and 10^5 for four-species trees. Our theoretical analysis predicts the same limiting distribution (when $n \to \infty$) for the three cases, which is also the same distribution as in problem 5 (Fig. 2A).

TABLE 2. Proportions of data sets with extreme bootstrap or posterior (in parentheses) probabilities for the three binary trees in the star-tree simulation

n	$\mathbb{P}\{P_{\min} < 1\%\}$	$\mathbb{P}\{P_{\min} < 5\%\}$	$\mathbb{P}\{P_{\max} > 95\%\}$	$\mathbb{P}\{P_{\max} > 99\%\}$	$\mathbb{E}(P_{\min})$	$\mathbb{E}(P_{\max})$
10^3	0.119 (0.234)	0.391 (0.550)	0.028 (0.205)	0.001 (0.079)	0.094 (0.067)	0.644 (0.754)
10^4	0.123 (0.812)	0.386 (0.931)	0.030 (0.606)	0.002 (0.450)	0.093 (0.011)	0.653 (0.897)
10^5	0.113 (0.979)	0.383 (0.992)	0.029 (0.853)	0.004 (0.773)	0.093 (0.001)	0.647 (0.964)

Notes: $P_{\min} = \min(P_1, P_2, P_3)$ and $P_{\max} = \max(P_1, P_2, P_3)$. Data are generated under JC+Γ with $\alpha = 1$, using the star tree for four species ((*a*: 0.2, *b*: 0.2) : 0.002, *c*: 0.2, *d*: 0.2), and analyzed under JC. The number of replicates is $R = 10^3$ and the number of bootstrap samples is $B = 10^3$. The probability density of (P_1, P_2, P_3) is shown in Fig. 5C&C′ for $n = 10^3$ and 10^5, respectively. Posterior tree probabilities from the Bayesian analysis are shown in parentheses, from Yang and Zhu (2018, Supplementary Table S1 available on Dryad).

was noted to produce strong support for the wrong trees frequently (Yang and Zhu, 2018, table S2), but this is found to be rare for the bootstrap method (table 3).

TABLE 3. Proportions of data sets with strong bootstrap or posterior (in parentheses) support for wrong trees in simulated data sets for four species

n	$\mathbb{P}\{P_1 < 1\%\}$	$\mathbb{P}\{P_1 < 5\%\}$	$\mathbb{P}\{P_{23} > 95\%\}$	$\mathbb{P}\{P_{23} > 99\%\}$
10^3	0.031 (0.083)	0.109 (0.225)	0.019 (0.113)	0.002 (0.038)
10^4	0.009 (0.250)	0.044 (0.337)	0.005 (0.266)	0.000 (0.166)
10^5	0.000 (0.102)	0.001 (0.120)	0.000 (0.115)	0.000 (0.097)

Note: P_1 is the probability for the true tree, while P_2 and P_3 are for the two wrong trees, with $P_{23} = \max\{P_2, P_3\}$. Data were generated under JC+Γ with $\alpha = 1$ on the unrooted tree T_1 for four species: $((a:0.2, b:0.2):0.002, c:0.2, d:0.2)$, and analyzed under JC. The number of simulated replicates is $R = 10^3$, with $B = 10^3$. Posterior tree probabilities from the Bayesian analysis are shown in parentheses, from Yang and Zhu (2018, Supplementary Table S2 available on Dryad).

DISCUSSION

As mentioned in the Introduction, the interpretation of bootstrap in model selection in general and in phylogenetics in particular is controversial. A number of studies have attempted to give bootstrap a Bayesian interpretation, that is, the bootstrap probability for a tree is the probability that the tree is correct. For example, Hastie et al. (2009, p.272) wrote that "[i]n this sense, the bootstrap distribution represents an (approximate) nonparametric, noninformative posterior distribution for our parameter." The plug-in principle for bootstrap appears to support this interpretation: bootstrap probability $\mathbb{P}\{\Delta^* > 0|x\}$ is an estimate of $\mathbb{P}\{\Delta > 0\}$, which is the probability that the ML tree is correct. In phylogenetics, such an interpretation was suggested by Efron et al. (1996), although the prior for the corresponding Bayesian analysis assumes infinite branch lengths and appears to be implausible biologically (Yang, 2014, p.176).

Our analysis suggests qualitatively different asymptotic behaviors between bootstrap and posterior probabilities for models or trees. The greatest difference occurs in the case of comparing equally wrong and distinct models. In that case, the posterior model probabilities show extreme polarized behavior, with ~100% for one model and 0 for others. This behavior occurs because the log marginal likelihood ratio for two models (or the log Bayes factor) (Δ) is dominated by a random-walk term that deviates from 0 at the rate of $\sqrt{n}$ when n increases (Yang and Zhu, 2018), so that for large n there is a vanishingly small probability for the log marginal likelihood ratio to stay in the neighborhood of 0 (or for the posterior model probability to stay in the neighborhood of $\frac{1}{2}$ away from both 0 and 1). Bootstrap probabilities show a different behavior. While the log likelihood ratio for the bootstrap data set (Δ^*) also increases like a random walk when n increases, this is compared with the log likelihood ratio for the original data set (Δ) when the bootstrap model probability is calculated. As a result, whether the models are distinct or indistinct does not matter anymore.

This property of the bootstrap probability, that it does not exhibit the polarized behavior in comparisons of equally wrong and distinct models (or trees) and that it does not often produce strong support for wrong models in large but finite data sets, should be considered an advantage over Bayesian posterior probability. However, neither bootstrap nor posterior probabilities for models converge to a point value when equally wrong or equally right models are compared. Given that our models of sequence evolution must always be wrong and simplistic, those results suggest that caution is needed to interpret strong support values (in particular, high posterior probabilities) for trees or clades in analyses of phylogenomic data sets. For the present, it is not so clear how the problems discussed in (Yang and Zhu, 2018) and in this paper can be mitigated. We note that robust model selection, in particular Bayesian selection of misspecified models or Bayesian nonparametrics, is a very active area of research in statistics, and several ideas have been suggested to make the inference less sensitive to misspecification of the likelihood model, including bootstrap resampling combined with Bayesian model selection (Huggins and Miller 2020) and flattening of the likelihood function (Watson and Holmes 2016). Further research is needed to assess the utility of those ideas to our problem of phylogeny reconstruction.

SUPPLEMENTARY DATA

Data available from the Dryad Digital Repository: https://doi.org/10.5061/dryad.7m0cfxprw.

Acknowledgments

We thank Zhi-ming Ma and Bin Wang for discussions. We are grateful to two anonymous reviewers, the Associate Editor (Mark Holder), and the Editor (Bryan Carstens) for comments.

Funding

This work was supported by Biotechnology and Biological Sciences Research Council grant (BB/P006493/1) and a BBSRC equipment grant (BB/R01356X/1) to Z.Y. J.H.'s visit to UCL is supported by China Scholarship Council (CSC). Natural Science Foundation grants 31671370 and 32070685 and a grant from the Youth Innovation Promotion Association of Chinese Academy of Sciences (201901) to T.Z.

Appendix. Asymptotic Theory for Bootstrap Probability in Model Selection

We use ML to compare K models, $H_j: X \sim f_j(X|\theta_j), j = 1,\ldots,K$. The data set, $x = \{x_1,\ldots,x_n\}$, is an i.i.d. sample of from the true model $g(X)$. Given x, we generate a bootstrap sample x^* and analyze it using ML. The bootstrap probability for model H_1 is the probability that model H_1 has higher log likelihood than other models in the bootstrap sample.

The case of two equally wrong and distinct models. We decompose the log-likelihood ratio between models H_1 and H_2 for the bootstrap data set x^* into several components, and study their dynamics when $n \to \infty$.

$$\Delta^* \equiv \log\frac{f_1(x^*|\hat{\theta}_1^*)}{f_2(x^*|\hat{\theta}_2^*)} = \log\frac{f_1(x^*|\hat{\theta}_1^*)}{f_1(x^*|\hat{\theta}_1)} - \log\frac{f_2(x^*|\hat{\theta}_2^*)}{f_2(x^*|\hat{\theta}_2)} + \log\frac{f_1(x^*|\hat{\theta}_1)}{f_2(x^*|\hat{\theta}_2)} \equiv \Delta A_1 - \Delta A_2 + \Delta_*^*. \tag{A1}$$

Model H_1 is the selected model in the bootstrap sample if and only if $\Delta^* > 0$, so that the bootstrap probability for H_1 given data x is $P_1 \equiv \mathbb{P}\{\Delta^* > 0|x\}$. We are interested in the distribution of P_1 when x varies. First we consider the case where H_1 and H_2 are equally wrong and distinct. We show that ΔA_1 and ΔA_2 are $O_p(1)$ while Δ_*^* is $O_p(n^{1/2})$, so that Δ^* is dominated by Δ_*^*.

Taking the same approach as in Dawid (2011) and Yang and Zhu (2018), we apply Taylor expansion to the log likelihood, $\log f_1(x^*|\theta_1)$, for the bootstrap data set x^* around the MLE $\hat{\theta}_1^*$ and then let $\theta_1 = \hat{\theta}_1$. We have

$$\begin{aligned}\Delta A_1 &= \log f_1(x^*|\hat{\theta}_1^*) - \log f_1(x^*|\hat{\theta}_1)\\ &\approx \tfrac{1}{2}\{(\hat{\theta}_1^* - \hat{\theta}_1)\}^T (nJ_1(\hat{\theta}_1^*))\{(\hat{\theta}_1^* - \hat{\theta}_1)\}\\ &\approx \tfrac{1}{2}\{\sqrt{n}(\hat{\theta}_1 - \theta_{1*})\}^T J_1(\theta_{1*})\{\sqrt{n}(\hat{\theta}_1 - \theta_{1*})\},\end{aligned} \tag{A2}$$

where $J_1(\theta_1) = \mathbb{E}\{-\nabla^2 \log f_1(X|\theta_1)\}$ and ∇^2 is the second derivatives with respect to θ_1. From the plug-in principle, x^* varies given $\hat{\theta}$ as does x given θ_* (Efron and Tibshirani, 1993). We have $\sqrt{n}(\hat{\theta}_1^* - \hat{\theta}_1) \xrightarrow{d} \sqrt{n}(\hat{\theta}_1 - \theta_{1*})$ (Bickel and Freedman, 1981; Cheng and Huang, 2010, Theorem 2), and

$$\sqrt{n}(\hat{\theta}_1 - \theta_{1*}) \sim \mathbb{N}\left(0, [J_1(\theta_{1*})^{-1}]^T I_1(\theta_{1*}) J_1(\theta_{1*})^{-1}\right), \tag{A3}$$

where $I_1(\theta_1) = \mathbb{E}\{\nabla \log f_1(X|\theta_1) \cdot \nabla \log f_1(X|\theta_1)^T\}$ (White, 1982, Theorem 3.2). Thus ΔA_1 is a quadratic form of normal variates and is $O_p(1)$. If H_1 is the true model, $\Delta A_1 \sim \frac{1}{2}\chi_d^2$ where d is the number of parameters in H_1. Similarly $\Delta A_2 = O_p(1)$.

We write the third term in Equation A1 as

$$\Delta_*^* \equiv \log\frac{f_1(x^*|\hat{\theta}_1)}{f_2(x^*|\hat{\theta}_2)} = \sum_{i=1}^n \log\frac{f_1(x_i^*|\hat{\theta}_1)}{f_2(x_i^*|\hat{\theta}_2)} \equiv \sum_{i=1}^n r_i^*(x). \tag{A4}$$

Define two log-likelihood ratios based on the original data x,

$$\begin{aligned}\Delta_* &\equiv \log\frac{f_1(x|\theta_{1*})}{f_2(x|\theta_{2*})},\\ \Delta &\equiv \log\frac{f_1(x|\hat{\theta}_1)}{f_2(x|\hat{\theta}_2)} = \sum_{i=1}^n \log\frac{f_1(x_i|\hat{\theta}_1)}{f_2(x_i|\hat{\theta}_2)} \equiv \sum_{i=1}^n r_i.\end{aligned} \tag{A5}$$

Note that Δ_* is a sum of n i.i.d. terms, so that when $n \to \infty$, $\Delta_* \sim \mathbb{N}(0, n\sigma^2)$, with $\mathbb{E}(\Delta_*) = n(D_1 - D_2) = 0$ (eq. 1) and $\mathbb{V}(\Delta_*) = n\sigma^2$, where

$$\sigma^2 \equiv \mathbb{V}_g\left\{\log\frac{f_1(X|\theta_{1*})}{f_2(X|\theta_{2*})}\right\} = \int g(X)\left[\log\frac{f_1(X|\theta_{1*})}{f_2(X|\theta_{2*})}\right]^2 \mathrm{d}X. \tag{A6}$$

When $n \to \infty$, $\bar{r} = \frac{1}{n}\sum_{i=1}^n r_i \to D_2 - D_1 = 0$ and $s^2 = \frac{1}{n}\sum_{i=1}^n (r_i - \bar{r})^2 \to \sigma^2$, so that $\Delta \sim \mathbb{N}(0, n\sigma^2)$.

Given data x, $\{r_i^*\}$ are conditionally independent, with expectation and variance

$$\begin{aligned}\mathbb{E}(\Delta_*^*|x) &= n\mathbb{E}\left\{\log\frac{f_1(x_1^*|\hat{\theta}_1)}{f_2(x_1^*|\hat{\theta}_2)}\middle| x\right\} \approx n \cdot \frac{1}{n}\sum_{i=1}^n \log\frac{f_1(x_i|\hat{\theta}_1)}{f_2(x_i|\hat{\theta}_2)}\\ &= \sum_{i=1}^n r_i = \Delta,\\ \mathbb{V}(\Delta_*^*|x) &= n\mathbb{V}\left\{\log\frac{f_1(x_1^*|\hat{\theta}_1)}{f_2(x_1^*|\hat{\theta}_2)}\middle| x\right\}\\ &= n\mathbb{E}\{(r_i^* - \mathbb{E}(r_i^*))^2 | x\} \approx n\sigma^2.\end{aligned} \tag{A7}$$

Thus $\Delta_*^*|x \sim \mathbb{N}(\Delta, n\sigma^2)$. The bootstrap probability for H_1 is

$$\begin{aligned}P_1 &= \mathbb{P}\{\Delta^* > 0|x\} = \mathbb{P}\{\Delta A_1 - \Delta A_2 + \Delta_*^* > 0|x\}\\ &\approx \mathbb{P}\{\Delta_*^* > 0|x\} \approx \Phi(\tfrac{\Delta}{\sqrt{n}\sigma}) \sim \mathbb{U}(0,1).\end{aligned} \tag{A8}$$

P_1 varies among data sets like a random number.

The case where there are no free parameters in the compared models has been discussed in the main paper.

We have

$$\Delta=\Delta_*=\log\frac{f_1(x)}{f_2(x)},\quad \Delta^*=\Delta^*_*=\log\frac{f_1(x^*)}{f_2(x^*)}, \tag{A9}$$

with $\Delta \sim \mathbb{N}(0, n\sigma^2)$ and $\Delta^*|\Delta \sim \mathbb{N}(\Delta, n\sigma^2)$, as $n \to \infty$. Thus

$$P_1=\mathbb{P}\{\Delta^*>0|x\}=\Phi\left(\tfrac{\Delta}{\sqrt{n}\sigma}\right)\to \mathbb{U}(0,1). \tag{A10}$$

The case where the two models are equally right or are equally wrong and indistinct, that is, with $f_1(X|\theta_1^*)=f_2(X|\theta_2^*)$ for almost every X. We have $\Delta_*=0$ in Equation A5, and $\Delta=O_p(1)$. As a result, $\Delta^*_*=O_p(1)$, as well as $\Delta A_1=O_p(1)$ and $\Delta A_2=O_p(1)$. From Equation A2, ΔA_1 and ΔA_2 have the same distribution, with $\mathbb{E}(\Delta A_1-\Delta A_2|x)=0$. Thus $\mathbb{E}(\Delta^*|x)=\mathbb{E}(\Delta^*_*|x)=\Delta$. Let F be the CDF of Δ, which has mean 0. Then

$$P_1=\mathbb{P}\{\Delta^*>0|x\}=1-F(-\Delta). \tag{A11}$$

Thus with $n\to\infty$, P_1 converges to a non-degenerate distribution, which is $\mathbb{U}(0,1)$ if and only if $\Delta^*-\Delta$ has the same distribution as $-\Delta$.

DasGupta (2008, Chapter 29) discusses regularity conditions under which T^*-T and $T-\mathbb{E}(T)$ have the same distribution, so that the bootstrap plugin principle can be applied, where T is a statistic or function of data x. If those conditions are not satisfied, the standard bootstrap will fail as T^*-T will not approximate $T-\mathbb{E}(T)$. Problem 4 is one such case, and $\Delta^*-\Delta$ and Δ have different distributions, and the limiting distribution of P_1 is not uniform. As indistinct models are more similar to each other than distinct models and as $P_1 \sim \mathbb{U}(0,1)$ when the two models are distinct (and equally wrong), we conjecture that $\mathbb{V}(P_1)\le\frac{1}{12}$, the variance of $\mathbb{U}(0,1)$.

Problems 3 and 4 are examples of equally right or equally wrong but indistinct models. Problem 3 shows the $\mathbb{U}(0,1)$ distribution, while problem 4 shows a non-uniform distribution.

The case of *K* models. Let the K models be $H_1,\cdots,H_K$, all of which have the same K-L distance to the true model. Define

$$\Delta_{*jk}\equiv\log\frac{f_j(x|\theta_{j*})}{f_k(x|\theta_{k*})},\quad \Delta_{jk}\equiv\sum_{i=1}^{n}\log\frac{f_j(x_i|\hat\theta_j)}{f_k(x_i|\hat\theta_k)} \tag{A12}$$

for data set x and

$$\Delta^*_{*jk}\equiv\sum_{i=1}^{n}\log\frac{f_j(x_i^*|\hat\theta_j)}{f_k(x_i^*|\hat\theta_k)},\quad \Delta^*_{jk}\equiv\sum_{i=1}^{n}\log\frac{f_j(x_i^*|\hat\theta_j^*)}{f_k(x_i^*|\hat\theta_k^*)} \tag{A13}$$

for bootstrap data set x^*.

First consider the case where the K models are equally wrong and distinct. As in the case of two models, Δ^*_{jk} is dominated by Δ^*_{*jk} so that $\Delta^*_{jk}\approx\Delta^*_{*jk}$ while $\Delta_{jk}\sim\mathbb{N}(0,n\sigma^2_{jk})$ and $\Delta^*_{*jk}\sim\mathbb{N}(\Delta_{jk},n\sigma^2_{jk})$, with $\sigma^2_{jk}\equiv\mathbb{V}\left\{\log\frac{f_j(X|\theta_{j*})}{f_k(X|\theta_{k*})}\right\}$ (see Equation A6). Given x, there will be a set of bootstrap probabilities $(P_1,\cdots,P_K)$. For example

$$P_1=\mathbb{P}\{\Delta^*_{12}>0,\ldots,\Delta^*_{1K}>0\big|x\}\approx\mathbb{P}\{\Delta^*_{*12}>0,\ldots,\Delta^*_{*1K}>0\big|x\}. \tag{A14}$$

Let $z=\{z_2,\cdots,z_{K-1}\}$ and $z^*=\{z_2^*,\cdots,z_{K-1}^*\}$, where $z_j=\frac{\Delta_{1j}}{\sqrt{n}\sigma_{1j}}$ and $z_j^*=\frac{\Delta^*_{1j}}{\sqrt{n}\sigma_{1j}}$. Let

$$\begin{aligned}\rho_{jk}&=\mathrm{Cor}(z_j,z_k)=\mathrm{Cor}(\Delta_{1j},\Delta_{1k})\\&=\frac{1}{\sigma_{1j}\sigma_{1k}}\mathrm{Cov}\left(\log\frac{f_1(X|\theta_{1*})}{f_j(X|\theta_{j*})},\log\frac{f_1(X|\theta_{1*})}{f_k(X|\theta_{k*})}\right).\end{aligned} \tag{A15}$$

Thus $z\sim\mathbb{N}(0,\Sigma_0)$ and $z^*|x\sim\mathbb{N}(z,\Sigma_0)$, where Σ_0 is a $(K-1)\times(K-1)$ variance matrix with 1 on the diagonal and ρ_{jk} on the off-diagonal. We have

$$P_1=\mathbb{P}(z_2^*>0,\cdots,z_K^*>0\big|x)=\Phi(-z_2,\cdots,-z_K), \tag{A16}$$

where Φ is the $(K-1)$-variate CDF of $\mathbb{N}(0,\Sigma_0)$. Bootstrap probabilities for the other models, $P_2,\cdots,P_K$, are given similarly.

When there is strong symmetry in the problem so that the K-L distance between any two models is the same, the variance matrix Σ_0 will have 1 on the diagonal and $\rho_{jk}=\frac{1}{2}$ on the off-diagonal, and further simplifications are possible. The joint distribution of bootstrap model probabilities $(P_1,\cdots,P_K)$ can be simulated as follows (see Problem 5). Sample $z=\{z_2,\cdots,z_K\}\sim\mathbb{N}(0,\Sigma_0)$ where Σ_0 is $(K-1)\times(K-1)$, with 1 on the diagonal and $\frac{1}{2}$ on the off-diagonal. Let $z_1=-(z_2+\cdots+z_K)$. Then calculate

$$\begin{bmatrix}P_1\\P_2\\\vdots\\P_K\end{bmatrix}=\begin{bmatrix}\Phi(z_2,z_3,\cdots,z_K)\\\Phi(-z_2,z_3-z_2,\cdots,z_K-z_2)\\\vdots\\\Phi(-z_K,z_2-z_K,\cdots,z_{K-1}-z_K)\end{bmatrix}. \tag{A17}$$

If the K models under comparison are equally right or equally wrong and indistinct, $\Delta^*_{jk}=O_p(1)$. Then the bootstrap probabilities $(P_1,\cdots,P_K)$ have a nondegenerate distribution.

In the case where some of the K models are equally wrong and distinct while others are indistinct, the dynamics of bootstrap model probabilities may be complex. A few representative cases which involve comparison of three models are analyzed in Supplementary Table S1 available on Dryad, with the distributions of bootstrap probabilities in Supplementary Figures S1–S3 available on Dryad.

References

Berry V., Gascuel O. 1996. On the interpretation of bootstrap trees: appropriate threshold of clade selection and induced gain. Mol. Biol. Evol. 13:999–1011.

Bickel P.J., Freedman D.A. 1981. Some asymptotic theory for the bootstrap. Ann. Statist. 9:1196–1217.

Chan K.O., Hutter C.R., Wood P.L., J., Grismer L.L., Brown R.M. 2020. Larger, unfiltered datasets are more effective at resolving phylogenetic conflict: introns, exons, and uces resolve ambiguities in golden-backed frogs (anura: Ranidae; genus hylarana). Mol. Phylogenet. Evol. 151:106899.

Cheng, G. and Huang, J. Z. 2010. Bootstrap consistency for general semiparametric M-estimation. Ann. Statist. 38:2884–2915.

DasGupta A. 2008. The bootstrap. In: Asymptotic theory of statistics and probability. New York: Springer. p. 461–497.

Davison A., Hinkley D. 1997. Bootstrap methods and their application. Cambridge, UK: Cambridge University Press.

Dawid A. 2011. Posterior model probabilities. In: Bandyopadhyay P.S., Forster M., editors. Philosophy of statistics. New York: Elsevier. p. 607–630.

Efron B. 1979. Bootstrap methods: another look at the jackknife. Ann. Stat. 7:1–26.

Efron B., Tibshirani R. 1993. An introduction to the bootstrap. London: Chapman and Hall.

Efron B., Halloran E., Holmes S. 1996. Bootstrap confidence levels for phylogenetic trees. Proc. Natl. Acad. Sci. U.S.A. 93:13429–13434.

Felsenstein J. 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. J. Mol. Evol. 17:368–376.

Felsenstein J. 1985. Confidence limits on phylogenies: an approach using the bootstrap. Evolution 39:783–791.

Felsenstein J., Kishino H. 1993. Is there something wrong with the bootstrap on phylogenies? A reply to Hillis and Bull. Syst. Biol. 42:193–200.

Fitch W.M. 1971. Toward defining the course of evolution: minimum change for a specific tree topology. Syst. Zool. 20:406–416.

Hastie T., Tibshirani R., Friedman, J. 2009. The elements of statistical learning: data mining, inference, and prediction. 2nd ed. New York: Springer.

Hillis D.M., Bull J.J. 1993. An empirical test of bootstrapping as a method for assessing confidence in phylogenetic analysis. Syst. Biol. 42:182–192.

Holmes S. 2003. Bootstrapping phylogenetic trees: theory and methods. Stat. Sci. 18:241–255.

Huelsenbeck J., Rannala B. 2004. Frequentist properties of Bayesian posterior probabilities of phylogenetic trees under simple and complex substitution models. Syst. Biol. 53:904–913.

Huggins J.H., Miller J. W. 2020. Robust and reproducible model selection using bagged posteriors. p. arXiv:2007.14845.

Jukes T., Cantor C. 1969. Evolution of protein molecules. In: Munro H., editor, Mammalian protein metabolism. New York: Academic Press. p. 21–123.

Lemoine F., Domelevo Entfellner J.-B., Wilkinson E., Correia D., Davila Felipe M., De Oliveira T., Gascuel O. 2018. Renewing Felsenstein's phylogenetic bootstrap in the era of big data. Nature 556:452–456.

Lewis P., Holder M., Holsinger K. 2005. Polytomies and Bayesian phylogenetic inference. Syst. Biol. 54:241–253.

O'Hagan, A. and Forster, J. 2004. Kendall's Advanced Theory of Statistics: Bayesian Inference. Arnold, London.

Rannala B., Yang Z. 1996. Probability distribution of molecular evolutionary trees: a new method of phylogenetic inference. J. Mol. Evol. 43:304–311.

Rubin D.B. 1981. The Bayesian bootstrap. Ann. Statist. 9:130–134.

Saitou N., Nei M. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol. Biol. Evol. 4:406–425.

Susko E. 2008. On the distributions of bootstrap support and posterior distributions for a star tree. Syst. Biol. 57:602–612.

Susko E. 2009. Bootstrap support is not first-order correct. Syst. Biol. 58:211–223.

Susko E. 2010. First-order correct bootstrap support adjustments for splits that allow hypothesis testing when using maximum likelihood estimation. Mol. Biol. Evol. 27:1621–1629.

Watson J., Holmes C.C. 2016. Approximate models and robust decisions. Stat. Sci. 31:465–489.

Weng, C.-S. 1989. On a second-order asymptotic property of the Bayesian bootstrap mean. Ann. Statist. 17:705–710.

White H. 1982. Maximum likelihood estimation of misspecified models. Econometrica 50:1–25.

Yang Z. 1993. Maximum-likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. Mol. Biol. Evol. 10:1396–1401.

Yang Z. 1994. Statistical properties of the maximum likelihood method of phylogenetic estimation and comparison with distance matrix methods. Syst. Biol. 43:329–342.

Yang Z. 1996. Phylogenetic analysis using parsimony and likelihood methods. J. Mol. Evol. 42:294–307.

Yang Z. 1997. How often do wrong models produce better phylogenies? Mol. Biol. Evol. 14:105–108.

Yang Z. 2000. Complexity of the simplest phylogenetic estimation problem. Proc. R. Soc. B: Biol. Sci. 267:109–116.

Yang Z. 2007a. Fair-balance paradox, star-tree paradox and Bayesian phylogenetics. Mol. Biol. Evol. 24:1639–1655.

Yang Z. 2007b. PAML 4: phylogenetic analysis by maximum likelihood. Mol. Biol. Evol. 24:1586–1591.

Yang Z. 2014. Molecular evolution: a statistical approach. Oxford, England: Oxford University Press.

Yang Z., Rannala B. 2005. Branch-length prior influences Bayesian posterior probability of phylogeny. Syst. Biol. 54:455–470.

Yang Z., Zhu T. 2018. Bayesian selection of misspecified models is overconfident and may cause spurious posterior probabilities for phylogenetic trees. Proc. Natl. Acad. Sci. USA 115:1854–1859.

Zharkikh A., Li W.-H. 1992. Statistical properties of bootstrap estimation of phylogenetic variability from nucleotide sequences. I. four taxa with a molecular clock. Mol. Biol. Evol. 9:119–1147.

Zharkikh A., Li W.-H. 1995. Estimation of confidence in phylogeny: the complete-and-partial bootstrap technique. Mol. Phylogenet. Evol. 4:44–63.