The Asymptotic Behavior of Bootstrap Support Values in Molecular Phylogenetics

JUN HUANG1,2, YUTING LIU1,*, TIANQI ZHU3,* AND ZIHENG YANG2,*

1Department of Mathematics, Beijing Jiaotong University, Beijing, 100044, China;
2Department of Genetics, Evolution and Environment, University College London, Gower Street, London WC1E 6BT, UK;
3National 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: ytlu@bjtu.edu.cn (Yuting Liu);
Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, China; E-mail: zhu@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 $\mathcal{U}(0, 1)$, 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 phylogenetic 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 1970), 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 model probabilities (e.g., Huelsenbeck and Hillis, 1993) 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 = {x1, ..., xn}, from the true data-generating model g(X). We compare K models, Hj, j = 1, ..., K. Model Hj specifies the density fj(X|θj) with parameters θj. Let ˆθj be the MLE of θj under model Hj given data x. When n → ∞, ˆθj → θj, where θj minimizes the Kullback–Leibler (K–L) divergence from model Hj to the true model,

\[ D_j = \int g(X) \log \frac{g(X)}{f_j(X|\hat{\theta}_j)} dX. \]  

If Hj is correct, ˆθj will be the true parameter values, with Dj = 0. Otherwise if Hj is wrong, ˆθj will be the best-fitting or pseudo-true parameter values, with Dj > 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 fj and f2 are said to be equally right if D1 = D2 = 0, and equally wrong if D1 = D2 > 0. If two models are unidentifiable at their pseudo-true parameter values, that is, if

\[ f_1(X|\theta_{1*}) = f_2(X|\theta_{2*}) \]  

for almost every X, (2) they are said to be indistinct. This can occur when both models are right (with D1 = D2 = 0) or when both are wrong (with D1 = D2 > 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 D1 = D2 > 0).

The model selected by ML, is the one that achieves the greatest log likelihood, ij(ˆθj) = logfj(x|ˆθj). To assess the confidence on the selected model, we calculate the bootstrap probability. Let x∗ = {x∗,1, ..., x∗,n} be a bootstrap sample, formed by resampling with replacement n times from the original data x. Let ˆθ∗ be the MLE from a bootstrap sample x∗. 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 ˆθ1, ˆθ2, and ˆθ∗ 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 Pj for model j. For example, in the case of two models, the bootstrap probability for model H1 is

\[ P_1(x) = \mathbb{P} \left\{ \log f_1(x|\hat{\theta}_1^*) > \log f_2(x|\hat{\theta}_2^*) | x \right\} \approx \frac{1}{B} \sum_b I_{\ell_1(\hat{\theta}_1^*) > \ell_2(\hat{\theta}_2^*)}, \]  

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 \( I_A \) 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 \( \hat{P}_1 \) 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.

\textbf{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 \( U(0,1) \).

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

\[ \Delta \equiv \log \frac{f_1(x)}{f_2(x)}, \quad \Delta^* \equiv \log \frac{f_1(x^*)}{f_2(x^*)}, \]  

(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 \( E(\Delta) = n E \log f_1(X) - n E \log f_2(X) = n(D_2 - D_1) = 0 \) (Equation 1). Let

\[ \sigma^2 = \text{Var} \left( \log \frac{f_1(X)}{f_2(X)} \right) = \int g(x) \left( \log \frac{f_1(X)}{f_2(X)} \right)^2 \, dx. \]  

(5)

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

\[ P_1 = P(\Delta > 0|x) \approx \Phi \left( \frac{\Delta}{\sqrt{n \sigma^2}} \right) \to U(0,1). \]  

(6)

where \( \Phi \) is the cumulative distribution function (CDF) for \( N(0,1) \).

In Bayesian comparison of two equally wrong models with no parameters, \( \Delta \) is the log Bayes factor. With equal prior probabilities (1/2 for each model), this is related to the posterior model probability through

\[ \Delta = \log \frac{P_1}{1-P_1} \text{ or } P_1 = \frac{e^\Delta}{e^{\Delta} + 1}. \]  

As \( \Delta \) behaves like a random walk when \( n \) increases, it is nearly impossible for \( \Delta \) 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 \( \Delta^* \) for the bootstrap sample with \( \Delta \) for the original data.

\textbf{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, \( P_1 \sim U(0,1) \) (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, \( \hat{x} \), which is approximately normal \( N(\frac{1}{2}, \frac{1}{n}) \). \( H_1 \) is favored if \( \hat{x} < \frac{1}{2} \), and this happens in half of the data sets.

Given \( x \), the bootstrap sample \( x^*_i = \{x^*_1, \ldots, x^*_n\} \), where \( x^*_i \) is a Bernoulli variable with probability \( \hat{x} \), can be summarized as the bootstrap sample mean \( \bar{x} \), which is approximately normal, with \( \bar{x}^* \sim \text{normal} \left( \bar{x}, \frac{1}{n} \right) \). Under each model, the pseudo-true parameter value \( \mu^* = x^* \) favors model \( H_1 \) if and only if \( x^* < 1/2 \). Thus,

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

(7)

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

**Problem 2 Normal distribution, equally wrong and distinct models with free parameters.** Suppose the true model is \( N(0,1) \) and we consider \( H_1 : N(\mu_1, 1/\tau_1) \) and \( H_2 : N(\mu_2, 1/\tau_2) \), where \( \mu \) is a free parameter while the precisions are fixed \( \tau_1 \) and \( \tau_2 \) 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 \( \mu^* = \hat{x} \) 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 \( N(\mu, 1/\tau) \) with known \( \tau \), the log likelihood is

\[ \ell = -\frac{n}{2} \log(2\pi) + \frac{n}{2} \log \tau - \frac{\tau}{2} \sum_{i=1}^n (x_i - \mu)^2. \]  

(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 \( s^2 = \frac{1}{\hat{\tau}} \sum_{i=1}^n (x_i - \bar{x})^2 > 1 \). We have \( ns^2 \sim \chi^2_{n-1} \approx N(n-1,2(n-1)) \).

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

This is confirmed in Figure 1b.

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 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). 

b) Problem 2 in which the true model is \( N(0,1) \) while the two compared models, \( N(\mu,1/\tau_1) \) and \( N(\mu,1/\tau_2) \) with \( \tau_1 < 1 < \tau_2 \), are equally right and distinct, and each has one free parameter. 

c) Problem 3 (the fair-balance paradox) where the true model is \( N(0,1) \) and the two compared models, \( N(\mu,1/\tau), \mu < 0 \) versus \( N(\mu,1/\tau), \mu > 0 \), 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 \( N(\mu,1) \) while the two compared models, \( N(\mu) \) versus \( N(0,1/\tau) \), are both right. Black dashed line is for the expensive simulation generating \( x \) and \( x^* \), the red dashed line is for simulation generating \( x \) and \( x^* \), 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 \( n = 10^4, B = 3 \times 10^4, \) and \( R = 10^4 \) for problem 1, \( n = 10^3, B = 3 \times 10^3, \) and \( R = 10^3 \) for problem 2, \( n = 10^3, B = 3 \times 10^3, \) and \( R = 10^3 \) for problem 3, and \( n = 10^3, B = 10^3, \) and \( R = 10^3 \) for problem 4.

The continuous distribution \( N(x,s^2) \). Thus, 

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

This is confirmed in Figure 1b.

Alternatively, we have \( \Delta = \ell_1 - \ell_2 = \frac{n}{2}(\tau_2 - \tau_1) s^2 - 1 \sim N(0,ns^2) \) and \( \Delta^* | \Delta \sim N(\Delta, ns^2) \), with \( s = \sqrt{2} (\tau_2 - \tau_1) \), 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 \( N(0,1) \) and the two compared models are \( N(\mu,1/\tau), \mu < 0 \) and \( N(\mu,1/\tau), \mu > 0 \), with \( \tau \) given. If \( \tau = 1 \), the two models are equally right. If \( \tau \neq 1 \), 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 \( s < 0 \). As
The simulation is done in two ways. In the first, data are generated as
\[ n \sim \mathcal{N}(0, 1/(nt)) \] and \( x^* \mid x \sim \mathcal{N}(\bar{x}, 1/(nt)) \), we have
\[
P_1 = \mathbb{P}(x^* < 0 \mid x) = \Phi(-\sqrt{\bar{x}^2/n}) \to U(0, 1), \quad \text{as } n \to \infty.
\] (10)

This is confirmed in Figure 1c.

**Problem 4 Normal-distribution example with an infinite spike at \( 1/2 \) in the \( P_1 \) distribution.** The true model is \( \mathcal{N}(0, 1) \) and the two compared models are \( \mathcal{N}(u, 1) \) and \( \mathcal{N}(0, 1/\tau) \). In \( H_1 \), \( \mu_\star = 0 \) while in \( H_2 \), \( \tau_\star = 1 \), 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
\[
\ell_1(\hat{\mu}) = -\frac{1}{2} \sum (x_i - \bar{x})^2 = -\frac{1}{2} ns^2,
\]
\[
\ell_2(\hat{\tau}) = -\frac{n}{2} \log \left( \frac{1}{n} \sum x_i^2 \right) - \frac{n}{2} \log(s^2 + \bar{x}^2) - \frac{n}{2}.
\]
Thus, \( \ell_1 > \ell_2 \) if and only if
\[
x^2 > e^{s^2/2} - s^2 \approx 1 + (s^2 - 1) + \frac{1}{2} (s^2 - 1)^2 - s^2 = \frac{1}{2} (s^2 - 1)^2,
\] (11)
or if and only if
\[
|x| > \sqrt{s^2 - 1}.
\] (13)

A large deviation of \( \bar{x} \) from \( x \) supports \( H_1 \), whereas a large deviation of \( s^2 \) from \( 1 \) favors \( H_2 \). Also \( x \sim \mathcal{N}(0, \frac{1}{2}) \) and \( s^2 \sim \frac{1}{2} \mathcal{X}^2_{n-1} \approx \mathcal{N}(\frac{n-1}{2}, \frac{2(n-1)}{n}) \) or \( \frac{1}{\sqrt{2}} (s^2 - 1) \approx \mathcal{N}(0, \frac{1}{n}) \), 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 \( x^* \mid x \sim \mathcal{N}(\bar{x}, s^2/n) \approx \mathcal{N}(\bar{x}, \frac{1}{n}) \) and \( \frac{1}{\sqrt{2}} (s^2 - 1) \) \( x \sim \mathcal{N}(\frac{1}{\sqrt{2}} (s^2 - 1), \frac{1}{2}) \) and \( x^* \) and \( s^2 \) are conditionally independent. Let \( z_1 = \sqrt{n} \bar{x} \) and \( z_2 = \sqrt{n} s^2 \), with \( z_1 \) and \( z_2 \) from \( \mathcal{N}(0, 1) \). Let \( z_1^* = \sqrt{n} x^* \) and \( z_2^* = \sqrt{n} s^2 \), with \( z_1^* \sim \mathcal{N}(z_1, 1) \) and \( z_2^* \sim \mathcal{N}(z_2, 1) \) to be conditionally i.i.d. Then,
\[
P_1 = \mathbb{P}(|z_1^*| > |z_2^*| \mid x).
\] (14)

This problem is analyzed in the Supplementary information text available on Dryad at https://doi.org/10.5061/dryad.7mtcfxpwr. The limiting distribution of \( P_1 \) when \( n \to \infty \) is
\[
f(P_1) = -\log(2|P_1 - 1|).
\] (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 \( \frac{1}{3}\pi \). This is confirmed by simulation in Figure 1d. The simulation is done in two ways. In the first, data \( x \) is sampled from \( \mathcal{N}(0, 1) \), and given \( x \) bootstrap samples \( x^*_b \) are generated, with \( x^* \) and \( s^2 \) calculated to apply Equation 14. In the second approach, \( x \sim \mathcal{N}(0, 1/n) \) and \( ns^2 \sim \chi^2_{n-1} \) 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 \( \mathcal{N}(\mu, \Sigma) \), with mean vector \( \mu = (\mu_1, \ldots, \mu_{K-1}) \) where \( \mu_1 = \cdots = \mu_{K-1} = 0 \) and variance matrix \( \Sigma \) which has 1 on the diagonal and \(-1/(K - 1)\) on the off-diagonal. The data are an i.i.d. sample of size \( n, x = (x_i), i = 1, \ldots, n; j = 1, \ldots, K - 1 \). Also let \( x_jK = -(x_1 + \cdots + x_{K-1}) \) 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, \( \Sigma \). 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 \( x^* = (x^*_j) \), with
\[
x^*_j = \frac{1}{n} \sum x_{ij}, \quad x^*_j = \frac{1}{n} \sum x^*_{ij}, \quad j = 1, \ldots, K,
\] (16)
be the sample means from data set \( x \) and from bootstrap sample \( x^* \), respectively. Then \( \bar{x} \sim \mathcal{N}(\mu, \Sigma) \) and approximately \( x^* \sim \mathcal{N}(\bar{x}, \Sigma/n) \). Without the constraint under each model \( H_j; \mu_j > \mu_k \), the MLEs of \( \mu \) 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
\[
P_1 = \mathbb{P}(x^*_1 > x^*_2, \ldots, x^*_K > x^*_K \mid x).
\] (17)

Now for any \( j \neq k \),
\[
\sigma^2_{jk} = \sqrt{(\bar{x}_j - \bar{x}_k)^2} = \frac{1}{2} \cdot \frac{1}{n} - \frac{n}{K - 1} = \frac{1}{2} \cdot \frac{n}{K - 1}.
\] (18)

Let \( z = (z_2, \ldots, z_K)^T \) and \( z^* = (z^*_2, \ldots, z^*_K)^T \), with \( z_j = \frac{\bar{x}_j - \bar{x}_1}{\sigma_{1j}} \) and \( z^*_j = \frac{\bar{x}_j^* - \bar{x}_1^*}{\sigma_{1j}}, j = 2, \ldots, K \). We have
\[
\forall(j, z_j), \quad \text{Cor}(z, z_j) = \text{Cov}(\bar{x}_j, z_j)/(\sigma_{1j}\sigma_{1k}) = \frac{[\bar{x}_j - 2\text{Cov}(\bar{x}_1, z_j) + \text{Cov}(\bar{x}_1, z_j)]/(\sigma_{1j}\sigma_{1k})}{\frac{1}{n} + \frac{1}{K - 1}} = \frac{2\bar{x}_j}{n + \frac{1}{K - 1}}.
\] (19)

Thus, \( z \sim \mathcal{N}(0, \Sigma_0) \) and \( z^* \sim \mathcal{N}(z, \Sigma_0) \), where \( \Sigma_0 \) is a \((K - 1) \times (K - 1) \) variance matrix with 1 on the diagonal and \( \frac{1}{2} \) on the off-diagonal. Thus,
\[
P_1 = \mathbb{P}(z^*_1 > 0, \ldots, z^*_K > 0 \mid x) = \Phi(z_2, \ldots, z_K).
\] (20)
As \( x^*)_1 - \bar{x}^*)_1 = (x^*)_1 - \bar{x}^*)_1 - (x^*)_1 - \bar{x}^*)_1 \), the bootstrap probabilities for all \( K \) models given data \( x \) are

\[
\begin{bmatrix}
P_1 \\
P_2 \\
\vdots \\
P_K \\
\end{bmatrix} = \begin{bmatrix}
\Phi(z_2, z_2, \ldots, z_K) \\
\Phi(-z_2, z_2, -z_2, \ldots, z_K, -z_2) \\
\vdots \\
\Phi(-z_K, -z_2, z_2, \ldots, z_K, -z_K). \\
\end{bmatrix}
\]

(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 \((z_2, z_3) \sim N((0,0.5), (1,1))\) 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 \((P_1, P_2, P_3)\) 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.

### Table 1. Proportions of data replicates with very high bootstrap probability \((P_i)\) in the multivariate normal example (Problem 5)

<table>
<thead>
<tr>
<th>( K )</th>
<th>( P[P_i &gt; 0.90] )</th>
<th>( P[P_i &gt; 0.95] )</th>
<th>( P[P_i &gt; 0.99] )</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>0.100</td>
<td>0.008</td>
<td>0.000</td>
</tr>
<tr>
<td>3</td>
<td>0.050</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>4</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>5</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
</tbody>
</table>

### 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 \((t_0, t_1)\) 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+\( \Gamma \) 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+\( \Gamma \)) 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 \((D_1 = D_2 = D_3 = 0.0)\).

Case C (Fig. 5C&C') involves equally wrong and distinct models. Like case B, the simulation model is JC+\( \Gamma \) 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 \( t_1 = 0.16409 \), \( i = 1, \ldots, 4 \) 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 \( xx \), \( xy \), \( yx \), \( yy \), and \( yz \), where \( x, y, z \) are any distinct nucleotides. Let the frequencies of the informative site patterns \( xx \), \( xy \), \( yx \), \( yy \), and \( yz \) be \( x_1, x_2, x_3, y_1, x_3, x_3 \), and \( x_3, y_2, x_2, x_3, x_3 \) be \( T_0 = (x_1, x_2, x_3) \) 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 \( x^*_j \) is the largest among \( (x^*_{11}, x^*_{12}, x^*_{13}) \).

In case C for four species, the informative site patterns are \( xxxy, xyxy \) and \( yyyx \) while there are 11 uninformative patterns. The binary tree has only 5 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 PAAML (Yang, 2007b).

Our theory predicts that the limiting distribution is the same in all three cases, with the mean \( \frac{1}{3} \) 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 $n = 10^5$, $E(P_{\max}) = 0.647$ 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+$\Gamma$ with $\alpha = 1$ and analyzing them under JC, as in case C above (Fig. 5C&5C’), 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+$\Gamma$ (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+$\Gamma$ (with $\alpha = 1$) 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).
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).

### 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 \( P[\Delta > 0|x] \) is an estimate of \( P[\Delta > 0] \), 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 \( \approx 100\% \) for one model and 0 for others. This behavior occurs because the log marginal likelihood ratio for two models (or the log Bayes factor) \( \Delta \) 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 \( \Delta^* \) also increases like a random walk when \( n \) increases, this is compared with the log likelihood ratio for the original data set \( \Delta \) 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 the true model $g(X)$. Given $x$, we generate a bootstrap sample $x^*$ and analyze it using ML. The bootstrap probability for model $H_j$ is the probability that model $H_j$ 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 $n \to \infty$.

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

Model $H_1$ is the selected model in the bootstrap sample if and only if $\Delta^* > 0$, so the bootstrap probability for $H_1$ given data $x$ is $P_1 = \mathbb{P}[\Delta^* > 0|x]$. We are interested in the distribution of $P_1$ when $n$ varies. First we consider the case where $H_1$ and $H_2$ are equally wrong and distinct. We show that $\Delta A_1$ and $\Delta A_2$ are $O_p(1)$ while $\Delta^*_s$ is $O_p(n^{1/2})$, so that $\Delta^*$ is dominated by $\Delta^*_s$.

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

$$
\Delta A_1 = \log f_1(x^*|\hat{\theta}_1) - \log f_1(x|\hat{\theta}_1) \\
\approx \frac{1}{2} (\hat{\theta}_1 - \hat{\theta}_1)^T \{n_1 f_1(x|\hat{\theta}_1)\} (\hat{\theta}_1 - \hat{\theta}_1) \\
\approx \frac{1}{2} (\sqrt{n} (\hat{\theta}_1 - \theta_1a)) \sqrt{n} (\hat{\theta}_1 - \theta_1a) + \text{second derivatives with respect to } \hat{\theta}_1.
$$

$x^*$ varies given $\hat{\theta}_1$ as does $x$ given $\theta_1$ (Efron and Tibshirani, 1993). We have $\sqrt{n}(\hat{\theta}_1 - \theta_1a) \xrightarrow{d} \mathcal{N}(0, f_1(x|\theta_1a)^{-1} f_1(x|\theta_1a) f_1(x|\theta_1a)^{-1})$ (Bickel and Freedman, 1981; Cheng and Huang, 2010, Theorem 2), and

$$
\sqrt{n}(\hat{\theta}_1 - \theta_1a) \sim \mathcal{N}(0, f_1(x|\theta_1a)^{-1} f_1(x|\theta_1a) f_1(x|\theta_1a)^{-1})

\text{A3}
$$

where $f_1(x|\theta_1) = \mathbb{E} \{\nabla \log f_1(X|\theta_1) \cdot \nabla \log f_1(X|\theta_1)^T\}$ (White, 1982, Theorem 3.2). Thus $\Delta A_1$ is a quadratic form of normal variates and is $O_p(1)$. If $H_1$ is the true model, $\Delta A_1 \sim \frac{n}{2} \chi^2$ 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

$$
\Delta^*_s = \log f_1(x|\theta_1) - \log f_2(x|\theta_1).
$$

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

$$
\Delta_1 = \log f_1(x|\theta_1) - \log f_2(x|\theta_1),
$$

$$
\Delta_2 = \log f_2(x|\theta_1) - \log f_1(x|\theta_1).
$$

Note that $\Delta_1$ is a sum of $n$ i.i.d. terms, so that when $n \to \infty, \Delta_1 \sim \mathcal{N}(0, n \sigma^2)$, with $\mathbb{E}(\Delta_1) = n(D_1 - D_2) = 0$ (eq. 1) and $\mathbb{V}(\Delta_1) = n\sigma^2$, where

$$
\sigma^2 = \mathbb{V}\{\log f_1(X|\theta_1) - \log f_2(X|\theta_1)\} = \int g(X) \left(\log f_1(X|\theta_1) - \log f_2(X|\theta_2)\right)^2 dX.
$$

When $n \to \infty$, $r = \frac{1}{n} \sum_{i=1}^{n} r_i \to D_2 - D_1 = 0$ and $s^2 = \frac{1}{n} \sum_{i=1}^{n} (r_i - r)^2 \to \sigma^2$, so that $\Delta \sim \mathcal{N}(0, n\sigma^2)$. Given data $x$, $r_i^2$ are conditionally independent, with expectation and variance

$$
\mathbb{E}(\Delta_1|x) = n \mathbb{E}\left\{\log f_1(x^*|\hat{\theta}_1) - \log f_2(x^*|\hat{\theta}_1)\right\} \approx n \cdot \frac{1}{n} \sum_{i=1}^{n} \log f_1(x^*|\hat{\theta}_1) - \log f_2(x^*|\hat{\theta}_1)
$$

$$
= \sum_{i=1}^{n} r_i = \Delta.
$$

$$
\mathbb{V}(\Delta_1|x) = n \mathbb{V}\{\log f_1(x^*|\hat{\theta}_1) - \log f_2(x^*|\hat{\theta}_1)\} \approx n \mathbb{E}\{(r_i^2 - \mathbb{E}(r_i^2))|x\} \approx n\sigma^2.
$$

Thus $\Delta_1|x \sim \mathcal{N}(\Delta, n\sigma^2)$. The bootstrap probability for $H_1$ is

$$
P_1 = P[\Delta^*_s > 0|x] = P[\Delta A_1 - \Delta A_2 + \Delta^*_s > 0|x] \approx P[\Delta^*_s > 0|x] \approx \Phi\left(\frac{\Delta}{\sqrt{n}\sigma}\right) \approx \mathbb{U}(0, 1).
$$

$P_1$ 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
\[ \Delta = \Delta_s = \log \frac{f_2(x)}{f_2(x)} \quad \Delta^* = \Delta^*_s = \log \frac{f_1(x^*)}{f_2(x^*)} \]  
(A9)
with \( n \sim N(0, n \sigma^2) \) and \( \Delta^* \mid \Delta \sim N(\Delta, n \sigma^2) \), as \( n \to \infty \). Thus
\[ P_1 = P[\Delta^* > 0 | x] = \Phi \left( \frac{\Delta}{\sqrt{n} \sigma} \right) \to U(0,1). \]  
(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 \( \Delta_s = 0 \) 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, \( \Delta A_1 \) and \( \Delta A_2 \) have the same distribution, with \( \mathbb{E}(\Delta A_1 - \Delta A_2 | x) = 0 \). Thus \( \mathbb{E}(\Delta^* | x) = \mathbb{E}(\Delta^*_s | x) = \Delta \). Let \( F \) be the CDF of \( \Delta \), which has mean 0. Then
\[ P_1 = P[\Delta^* > 0 | x] = 1 - F(-\Delta). \]  
(A11)
Thus with \( n \to \infty \), \( P_1 \) converges to a non-degenerate distribution, which is \( U(0,1) \) if and only if \( \Delta^* - \Delta \) has the same distribution as \( -\Delta \).

DasGupta (2008, Chapter 29) discusses regularity conditions under which \( T^* - T \) and \( T - T_0 \) have the same distribution, so that the bootstrap plug-in 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 \( T - T_0 \). Problem 4 is one such case, and \( \Delta^* - \Delta \) and \( \Delta \) 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 \( P_1 \sim U(0,1) \) when the two models are distinct (and equally wrong), we conjecture that \( \mathbb{V}(P_1) \leq \frac{1}{2^n} \), the variance of \( U(0,1) \).

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

**The case of K models.** Let the K models be \( H_1, \ldots, H_K \), all of which have the same K-L distance to the true model. Define
\[ \Delta_{ijk} = \log \frac{f_j(x | \theta_{jk})}{f_k(x | \theta_{jk})}, \quad \Delta_{jk} = \sum_{i=1}^n \log \frac{f_j(x_i | \theta_{jk})}{f_k(x_i | \theta_{jk})} \]  
(A12)
for data set \( x \) and
\[ \Delta_{ijk}^* = \sum_{i=1}^n \log \frac{f_j(x_i^* | \theta_{jk})}{f_k(x_i^* | \theta_{jk})}, \quad \Delta_{jk}^* = \sum_{i=1}^n \log \frac{f_j(x_i^* | \theta_{jk})}{f_k(x_i^* | \theta_{jk})} \]  
(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, \( \Delta_{jk}^* \) is dominated by \( \Delta_{jk} \) so that \( \Delta_{jk}^* \approx \Delta_{jk} \) while \( \Delta_{jk} \sim N(0, n \sigma_{ijk}^2) \) and \( \Delta_{jk}^* \sim N(\Delta_{jk}, n \sigma_{ijk}^2) \), with \( \sigma_{ijk}^2 = \mathbb{V} \{ \log \frac{f_j(x_i | \theta_{jk})}{f_k(x_i | \theta_{jk})} \} \) (see Equation A6). Given \( x \), there will be a set of bootstrap probabilities \( P_1, \ldots, P_K \). For example
\[ P_1 = P[\Delta_{12} > 0, \ldots, \Delta_{1K} > 0 | x] = P[\Delta_{21} > 0, \ldots, \Delta_{1K} > 0 | x]. \]  
(A14)
Let \( z = (z_2, \ldots, z_K) \) and \( z^* = (z_2^*, \ldots, z_K^*) \), where \( \Delta_j = \frac{\Delta_j}{\sqrt{n \sigma_{ijk}}} \) and \( \Delta_j^* = \frac{\Delta_j^*}{\sqrt{n \sigma_{ijk}}} \). Let
\[ \rho_{jk} = \text{Cov}(z_j, z_k) = \text{Cov}(\Delta_{jk}, \Delta_{jk'}) \]  
(A15)
Thus \( z \sim N(0, \Sigma_0) \) and \( z^* | x \sim N(z, \Sigma_0) \), where \( \Sigma_0 \) is a \((K-1) \times (K-1) \) variance matrix with 1 on the diagonal and \( \rho_{jk} \) on the off-diagonal. We have
\[ P_1 = P(z_j^* > 0, \ldots, z_K^* > 0 | x) = \Phi(-z_2^*, \ldots, -z_K^*). \]  
(A16)
where \( \Phi \) is the \((K-1)\)-variate CDF of \( N(0, \Sigma_0) \). Bootstrap probabilities for the other models, \( P_2, \ldots, P_K \), 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 \( \Sigma_0 \) 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 \( P_1, \ldots, P_K \) can be simulated as follows (see Problem 5). Sample \( z = (z_2, \ldots, z_K) \sim N(0, \Sigma_0) \) where \( \Sigma_0 \) 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
\[ \begin{bmatrix} P_1 \\ P_2 \\ \vdots \\ P_K \end{bmatrix} = \Phi(-z_2^*, -z_2^* + z_2^*, \ldots, -z_K^* + z_K^*, -z_2^*). \]  
(A17)
If the \( K \) models under comparison are equally right or equally wrong and indistinct, \( \Delta_{jk} = O_p(1) \). Then the bootstrap probabilities \( P_1, \ldots, P_K \) 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**


