frontiers in Ecology and Evolution

METHODS
published: 24 May 2019
doi: 10.3389/fevo.2019.00174

Selective Inference for Testing Trees and Edges in Phylogenetics

Hidetoshi Shimodaira[1,2]* *and Yoshikazu Terada*[2,3]

[1] Graduate School of Informatics, Kyoto University, Kyoto, Japan, [2] Mathematical Statistics Team, RIKEN Center for Advanced Intelligence Project, Tokyo, Japan, [3] Graduate School of Engineering Science, Osaka University, Osaka, Japan

Selective inference is considered for testing trees and edges in phylogenetic tree selection from molecular sequences. This improves the previously proposed approximately unbiased test by adjusting the selection bias when testing many trees and edges at the same time. The newly proposed selective inference p-value is useful for testing selected edges to claim that they are significantly supported if $p > 1-\alpha$, whereas the non-selective p-value is still useful for testing candidate trees to claim that they are rejected if $p < \alpha$. The selective p-value controls the type-I error conditioned on the selection event, whereas the non-selective p-value controls it unconditionally. The selective and non-selective approximately unbiased p-values are computed from two geometric quantities called signed distance and mean curvature of the region representing tree or edge of interest in the space of probability distributions. These two geometric quantities are estimated by fitting a model of scaling-law to the non-parametric multiscale bootstrap probabilities. Our general method is applicable to a wider class of problems; phylogenetic tree selection is an example of model selection, and it is interpreted as the variable selection of multiple regression, where each edge corresponds to each predictor. Our method is illustrated in a previously controversial phylogenetic analysis of human, rabbit and mouse.

Keywords: statistical hypothesis testing, multiple testing, selection bias, model selection, Akaike information criterion, bootstrap resampling, hierarchical clustering, variable selection

OPEN ACCESS

Edited by:
Rodney L. Honeycutt, Pepperdine University, United States

Reviewed by:
Carlos Garcia-Verdugo, Jardín Botánico Canario "Viera y Clavijo", Spain
Michael S. Brewer, University of California, Berkeley, United States

***Correspondence:**
Hidetoshi Shimodaira
shimo@i.kyoto-u.ac.jp

Specialty section:
This article was submitted to Phylogenetics, Phylogenomics, and Systematics, a section of the journal Frontiers in Ecology and Evolution

Received: *31 January 2019*
Accepted: *30 April 2019*
Published: *24 May 2019*

Citation:
Shimodaira H and Terada Y (2019) Selective Inference for Testing Trees and Edges in Phylogenetics. Front. Ecol. Evol. 7:174. doi: 10.3389/fevo.2019.00174

1. INTRODUCTION

A phylogenetic tree is a diagram showing evolutionary relationships among species, and a tree topology is a graph obtained from the phylogentic tree by ignoring the branch lengths. The primary objective of any phylogenetic analysis is to approximate a topology that reflects the evolution history of the group of organisms under study. Branches of the tree are also referred to as edges in the tree topology. Given a rooted tree topology, or a unrooted tree topology with an outgroup, each edge splits the tree so that it defines the clade consisting of all the descendant species. Therefore, edges in a tree topology represent clades of species. Because the phylogenetic tree is commonly inferred from molecular sequences, it is crucial to assess the statistical confidence of the inference. In phylogenetics, it is a common practice to compute confidence levels for tree topologies and edges. For example, the bootstrap probability (Felsenstein, 1985) is the most commonly used confidence measure, and other methods such as the Shimodaira-Hasegawa test (Shimodaira and Hasegawa, 1999) and the multiscale bootstrap method (Shimodaira, 2002) are also often used. However, these conventional methods are limited in how well they address the issue of multiplicity when there are many alternative topologies and edges. Herein, we discuss a new approach, selective inference (SI), that is designed to address the issue of multiplicity.

For illustrating the idea of selective inference, we first look at a simple example of 1-dimensional normal random variable Z with unknown mean $\theta \in \mathbb{R}$ and variance 1:

$$Z \sim N(\theta, 1). \tag{1}$$

Observing $Z = z$, we would like to test the null hypothesis $H_0 : \theta \leq 0$ against the alternative hypothesis $H_1 : \theta > 0$. We denote the cumulative distribution function of $N(0,1)$ as $\Phi(x)$ and define the upper tail probability as $\bar{\Phi}(x) = 1 - \Phi(x) = \Phi(-x)$. Then, the ordinary (i.e., non-selective) inference leads to the p-value of the one-tailed z-test as

$$p(z) := P(Z > z \mid \theta = 0) = \bar{\Phi}(z). \tag{2}$$

What happens when we test many hypotheses at the same time? Consider random variables $Z_i \sim N(\theta_i, 1)$, $i = 1, \ldots, K_{\text{all}}$, not necessarily independent, with null hypotheses $\theta_i \leq 0$, where K_{true} hypotheses are actually true. To control the number of falsely rejecting the K_{true} hypotheses, there are several multiplicity adjusted approaches such as the family-wise error rate (FWER) and the false discovery rate (FDR). Instead of testing all the K_{all} hypotheses, selective inference (SI) allows for K_{select} hypotheses with $z_i > c_i$ for constants c_i specified in advance. This kind of selection is very common in practice (e.g., publication bias), and it is called as the *file drawer problem* by Rosenthal (1979). Instead of controlling the multiplicity of testing, SI alleviates it by reducing the number of tests. The mathematical formulation of SI is easier than FWER and FDR in the sense that hypotheses can be considered separately instead of simultaneously. Therefore, we simply write $z > c$ by dropping the index i for one of the hypotheses. In selective inference, the selection bias is adjusted by considering the conditional probability given the selection event, which leads to the following p-value (Fithian et al., 2014; Tian and Taylor, 2018)

$$p(z, c) := P(Z > z \mid Z > c, \theta = 0) = \bar{\Phi}(z)/\bar{\Phi}(c), \tag{3}$$

where $p(z)$ of Equation (2) is divided by the selection probability $P(Z > c \mid \theta = 0) = \bar{\Phi}(c)$. In the case of $c = 0$, this corresponds to the two-tailed z-test, because the selection probability is $\bar{\Phi}(0) = 0.5$ and $p(z, c) = 2p(z)$. For significance level α (we use $\alpha = 0.05$ unless otherwise stated), it properly controls the type-I error conditioned on the selection event as $P(p(Z, c) < \alpha \mid Z > c, \theta = 0) = \alpha$, while the non-selective p-value violates the type-I error as $P(p(Z) < \alpha \mid Z > c, \theta = 0) = \alpha/\bar{\Phi}(c) > \alpha$. The selection bias can be very large when $\bar{\Phi}(c) \ll 1$ (i.e., $c \gg 0$), or $K_{\text{select}} \ll K_{\text{all}}$.

Selective inference has been mostly developed for inferences after model selection (Taylor and Tibshirani, 2015; Tibshirani et al., 2016), particularly variable selection in regression settings such as lasso (Tibshirani, 1996). Recently, Terada and Shimodaira (2017) developed a general method for selective inference by adjusting the selection bias in the approximately unbiased (AU) p-value computed by the multiscale bootstrap method (Shimodaira, 2002, 2004, 2008). This new method can be used to compute, for example, confidence intervals of regression coefficients in lasso (**Figure 1**). In this paper, we apply this method to phylogenetic inference for computing proper confidence levels of tree topologies (dendrograms) and edges (clades or clusters) of species. As far as we know, this is the first attempt to consider selective inference in phylogenetics. Our selective inference method is implemented in software *scaleboot* (Shimodaira, 2019) working jointly with *CONSEL* (Shimodaira and Hasegawa, 2001) for phylogenetics, and it is also implemented in a new version of *pvclust* (Suzuki and Shimodaira, 2006) for hierarchical clustering, where only edges appeared in the observed tree are "selected" for computing p-values. Although our argument is based on the rigorous theory of mathematical statistics in Terada and Shimodaira (2017), a self-contained illustration is presented in this paper for the theory as well as the algorithm of selective inference.

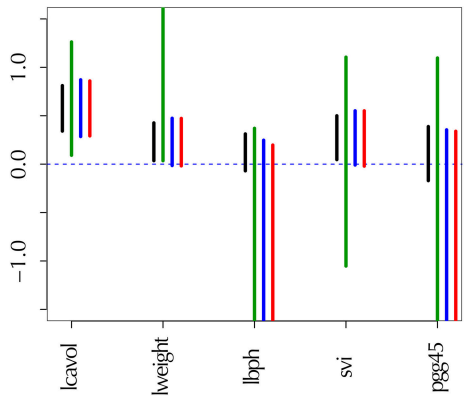

FIGURE 1 | Confidence intervals of regression coefficients for selected variables by lasso; see section 6.8 for details. All intervals are computed for confidence level $1 - \alpha$ at $\alpha = 0.01$. (Black) the ordinary confidence interval $[L_j^{\text{ordinary}}, U_j^{\text{ordinary}}]$. (Green) the selective confidence interval $[L_j^{\text{model}}, U_j^{\text{model}}]$ under the selected model. (Blue) the selective confidence interval $[L_j^{\text{variable}}, U_j^{\text{variable}}]$ under the selection event that variable j is selected. (Red) the multiscale bootstrap version of selective confidence interval $[\hat{L}_j^{\text{variable}}, \hat{U}_j^{\text{variable}}]$ under the selection event that variable j is selected.

Phylogenetic tree selection is an example of model selection. Since each tree can be specified as a combination of edges, tree selection can be interpreted as the variable selection of multiple regression, where edges correspond to the predictors of regression (Shimodaira, 2001; Shimodaira and Hasegawa, 2005). Because all candidate trees have the same number of model parameters, the maximum likelihood (ML) tree is obtained by comparing log-likelihood values of trees (Felsenstein, 1981). In order to adjust the model complexity by the number of parameters in general model selection, we compare Akaike Information Criterion (AIC) values of candidate models (Akaike, 1974). AIC is used in phylogenetics for selecting the substitution model (Posada and Buckley, 2004). There are several modifications of AIC that allow for model selection. These include the precise estimation of the complexity term known as Takeuchi Information Criterion (Burnham and Anderson, 2002; Konishi and Kitagawa, 2008), and adaptations for incomplete data (Shimodaira and Maeda, 2018) and covariate-shift data (Shimodaira, 2000). AIC and all these modifications are derived

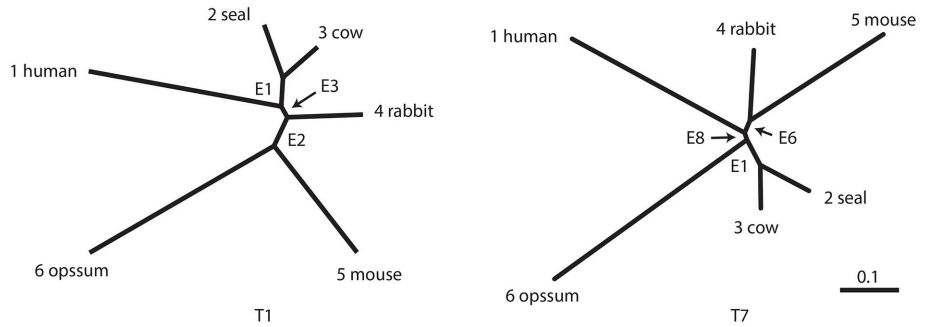

FIGURE 2 | Examples of two unrooted trees T1 and T7. Branch lengths represent ML estimates of parameters (expected number of substitutions per site). T1 includes edges E1, E2, and E3 and T7 includes E1, E6, and E8.

for estimating the expected Kullback-Leibler divergence between the unknown true distribution and the estimated probability distribution on the premise that the model is misspecified. When using regression model for prediction purpose, it may be sufficient to find only the best model which minimizes the AIC value. Considering random variations of dataset, however, it is obvious in phylogenetics that the ML tree does not necessarily represent the true history of evolution. Therefore, Kishino and Hasegawa (1989) proposed a statistical test whether two log-likelihood values differ significantly (also known as *Kishino-Hasegawa* test). The log-likelihood difference is often not significant, because its variance can be very large for non-nested models when the divergence between two probability distributions is large; see Equation (26) in section 6.1. The same idea of model selection test whether two AIC values differ significantly has been proposed independently in statistics (Linhart, 1988) and econometrics (Vuong, 1989). Another method of model selection test (Efron, 1984) allows for the comparison of two regression models with an adjusted bootstrap confidence interval corresponding to the AU p-value. For testing which model is better than the other, the null hypothesis in the model selection test is that the two models are equally good in terms of the expected value of AIC on the premise that both models are misspecified. Note that the null hypothesis is whether the model is correctly specified or not in the traditional hypothesis testing methods including the likelihood ratio test for nested models and the modified likelihood ratio test for non-nested models (Cox, 1962). The model selection test is very different from these traditional settings. For comparing AIC values of more than two models, a multiple comparisons method is introduced to the model selection test (Shimodaira, 1998; Shimodaira and Hasegawa, 1999), which computes the confidence set of models. But the multiple comparisons method is conservative by nature, leading to more false negatives than expected, because it considers the worst scenario, called the least favorable configuration. On the other hand, the model selection test (designed for two models) and bootstrap probability (Felsenstein, 1985) lead to more false positives than expected when comparing more than two models (Shimodaira and Hasegawa, 1999; Shimodaira, 2002). The AU p-value mentioned earlier has been developed for solving this problem, and we are going to upgrade it for selective inference.

2. PHYLOGENETIC INFERENCE

For illustrating phylogenetic inference methods, we analyze a dataset consisting of mitochondrial protein sequences of six mammalian species with $n = 3,414$ amino acids (n is treated as sample size). The taxa are labeled as 1=*Homo sapiens* (human), 2=*Phoca vitulina* (seal), 3=*Bos taurus* (cow), 4=*Oryctolagus cuniculus* (rabbit), 5=*Mus musculus* (mouse), and 6=*Didelphis virginiana* (opossum). The dataset will be denoted as $\mathcal{X}_n = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$. The software package PAML (Yang, 1997) was used to calculate the site-wise log-likelihoods for trees. The mtREV model (Adachi and Hasegawa, 1996) was used for amino acid substitutions, and the site-heterogeneity was modeled by the discrete-gamma distribution (Yang, 1996). The dataset and evolutionary model are similar to previous publications (Shimodaira and Hasegawa, 1999; Shimodaira, 2001, 2002), thus allowing our proposed method to be easily compared with conventional methods.

The number of unrooted trees for six taxa is 105. These trees are reordered by their likelihood values and labeled as T1, T2, ..., T105. T1 is the ML tree as shown in **Figure 2** and its tree topology is represented as (((1(23))4)56). There are three internal branches (we call them as edges) in T1, which are labeled as E1, E2, and E3. For example, E1 splits the six taxa as {23|1456} and the partition of six taxa is represented as $-++---$, where $+/-$ indicates taxa $1, \ldots, 6$ from left to right and $++$ indicates the clade {23} (we set $-$ for taxon 6, since it is treated as the outgroup). There are 25 edges in total, and each tree is specified by selecting three edges from them, although not all the combinations of three edges are allowed.

The result of phylogenetic analysis is summarized in **Table 1** for trees and **Table 2** for edges. Three types of p-values are computed for each tree as well as for each edge. BP is the bootstrap probability (Felsenstein, 1985) and AU is the approximately unbiased p-value (Shimodaira, 2002). Bootstrap probabilities are computed by the non-parametric bootstrap

TABLE 1 | Three types of p-values (BP, AU, SI) and geometric quantities (β_0, β_1) for the best 20 trees.

Tree	BP	AU	SI	β_0	β_1	Topology	Edges
T1†	0.559 (0.001)	0.752 (0.001)	0.372 (0.001)	-0.41 (0.00)	0.27 (0.00)	(((1(23))4)56)	E1, E2, E3
T2	0.304 (0.000)	0.467 (0.001)	0.798 (0.001)	0.30 (0.00)	0.22 (0.00)	((1((23)4))56)	E1 ,E2, E4
T3	**0.038** (0.000)	0.126 (0.002)	0.202 (0.003)	1.46 (0.01)	0.32 (0.00)	(((14)(23))56)	E1, E2, E5
T4	**0.014** (0.000)	0.081 (0.002)	0.124 (0.003)	1.79 (0.01)	0.40 (0.01)	((1(23))(45)6)	E1, E3, E6
T5	**0.032** (0.000)	0.127 (0.002)	0.199 (0.003)	1.50 (0.01)	0.36 (0.00)	(1((23)(45))6)	E1, E6, E7
T6	**0.005** (0.000)	**0.032** (0.002)	0.050 (0.002)	2.21 (0.02)	0.35 (0.01)	(1(((23)4)5)6)	E1, E4, E7
T7‡	**0.015** (0.000)	0.100 (0.003)	0.150 (0.003)	1.72 (0.01)	0.44 (0.01)	((1(45))(23)6)	E1, E6, E8
T8	**0.001** (0.000)	**0.011** (0.001)	**0.016** (0.002)	2.74 (0.03)	0.43 (0.02)	((15)((23)4)6)	E1, E4, E9
T9	**0.000** (0.000)	**0.001** (0.000)	**0.001** (0.000)	3.67 (0.09)	0.46 (0.04)	(((1(23))5)46)	E1, E3, E10
T10	**0.002** (0.000)	**0.022** (0.002)	**0.033** (0.002)	2.43 (0.02)	0.42 (0.01)	(((15)4)(23)6)	E1, E8, E9
T11	**0.000** (0.000)	**0.004** (0.001)	**0.006** (0.002)	3.14 (0.07)	0.51 (0.03)	(((14)5)(23)6)	E1, E5, E8
T12	**0.000** (0.000)	**0.000** (0.000)	**0.001** (0.000)	3.78 (0.09)	0.41 (0.04)	(((15)(23))46)	E1, E9, E10
T13	**0.000** (0.000)	**0.000** (0.000)	**0.001** (0.001)	3.96 (0.19)	0.54 (0.09)	(1(((23)5)4)6)	E1, E7, E11
T14	**0.000** (0.000)	**0.000** (0.000)	**0.000** (0.000)	4.66 (0.31)	0.65 (0.12)	((14)((23)5)6)	E1, E5, E11
T15	**0.000** (0.000)	**0.000** (0.000)	**0.000** (0.000)	5.28 (0.34)	0.43 (0.11)	((1((23)5))46)	E1, E10, E11
T16	**0.000** (0.000)	**0.000** (0.000)	**0.001** (0.000)	3.63 (0.04)	0.23 (0.01)	((((13)2)4)56)	E2, E3, E12
T17	**0.000** (0.000)	**0.000** (0.000)	**0.000** (0.000)	3.81 (0.04)	0.22 (0.01)	((((12)3)4)56)	E2, E3, E13
T18	**0.000** (0.000)	**0.000** (0.000)	**0.000** (0.000)	4.33 (0.10)	0.34 (0.03)	(((13)2)(45)6)	E3, E6, E12
T19	**0.000** (0.000)	**0.000** (0.000)	**0.000** (0.000)	4.36 (0.11)	0.32 (0.04)	(((12)3)(45)6)	E3, E6, E13
T20	**0.000** (0.000)	**0.000** (0.000)	**0.000** (0.000)	3.90 (0.12)	0.44 (0.05)	(((1(45))2)36)	E6, E8, E14

Standard errors are shown in parentheses. Boldface indicates significance ($p < 0.05$) for the null hypothesis that the tree is true (outside mode). For the rest of trees (T21, . . . , T105), p-values are very small ($p < 0.001$). † T1 is the ML tree, i.e., the tree selected by the ML method based on the dataset of Shimodaira and Hasegawa (1999). ‡ T7 is presumably the true tree as suggested by later researches; see section 4.3.

resampling (Efron, 1979) described in section 6.1. The theory and the algorithm of BP and AU will be reviewed in section 3. Since we are testing many trees and edges at the same time, there is potentially a danger of selection bias. The issue of selection bias has been discussed in Shimodaira and Hasegawa (1999) for introducing the method of multiple comparisons of log-likelihoods (also known as *Shimodaira-Hasegawa test*) and in Shimodaira (2002) for introducing AU test. However, these conventional methods are only taking care of the multiplicity of comparing many log-likelihood values for computing just one p-value instead of many p-values at the same time. Therefore, we intend to further adjust the AU p-value by introducing the selective inference p-value, denoted as SI. The theory and the algorithm of SI will be explained in section 4 based on the geometric theory given in section 3. After presenting the methods, we will revisit the phyloegnetic inference in section 4.3.

For developing the geometric theory in sections 3 and 4, we formulate tree selection as a mathematical formulation known as *the problem of regions* (Efron et al., 1996; Efron and Tibshirani, 1998). For better understanding the geometric nature of the theory, the problem of regions is explained below for phylogenetic inference, although the algorithm is simple enough to be implemented without understanding the theory. Considering the space of probability distributions (Amari and Nagaoka, 2007), the parametric models for trees are represented as manifolds in the space. The dataset (or the empirical distribution) can also be represented as a "data point" X in the space, and the ML estimates for trees are represented as projections to the manifolds. This is illustrated in the visualization of probability distributions of **Figure 3A** using log-likelihood vectors of models (Shimodaira, 2001), where models are simply indicated as red lines from the origin; see section 6.2 for details. This visualization may be called as *model map*. The point X is actually reconstructed as the minimum full model containing all the trees as submodels, and the Kullback-Leibler divergence between probability distributions is represented as the squared distance between points; see Equation (27). Computation of X is analogous to the Bayesian model averaging, but based on the ML method. For each tree, we can think of a region in the space so that this tree becomes the ML tree when X is included in the region. The regions for T1, T2, and T3 are illustrated in **Figure 3B**, and the region for E2 is the union of these three regions.

In **Figure 3A**, X is very far from any of the tree models, suggesting that all the models are wrong; the likelihood ratio statistic for testing T1 against the full model is 113.4, which is highly significant as χ^2_8 (Shimodaira, 2001, section 5). Instead of testing whether tree models are correct or not, we test whether models are significantly better than the others. As seen in **Figure 3B**, X is in the region for T1, meaning that the model for T1 is better than those for the other trees. For convenience, observing X in the region for T1, we state that T1 is *supported* by the data. Similarly, X is in the region for E2 that consists of the three regions for T1, T2, T3, thus indicating that E2 is *supported* by the data. Although T1 and E2 are supported by the data, there is still uncertainty as to whether the true evolutionary history of lineages is depicted because the location of X fluctuates randomly. Therefore, statistical

TABLE 2 | Three types of p-values (BP, AU, SI) and geometric quantities (β_0, β_1) for all the 25 edges of six taxa.

Edge	BP	AU	SI	β_0	β_1	Clade
E1†‡	**1.000** (0.000)	**1.000** (0.000)	**1.000** (0.000)	-3.87 (0.03)	0.16 (0.01)	−++−−−
E2†	0.930 (0.000)	**0.956** (0.001)	0.903 (0.001)	-1.59 (0.00)	0.12 (0.00)	++++−−
E3†	0.580 (0.001)	0.719 (0.001)	0.338 (0.001)	-0.39 (0.00)	0.19 (0.00)	+++−−−
E4	0.318 (0.000)	0.435 (0.001)	0.775 (0.001)	0.32 (0.00)	0.16 (0.00)	−+++−−
E5	**0.037** (0.000)	0.124 (0.002)	0.198 (0.002)	1.47 (0.01)	0.32 (0.00)	+−−+−−
E6‡	0.060 (0.000)	0.074 (0.001)	0.141 (0.002)	1.50 (0.00)	0.05 (0.00)	−−−++−
E7	**0.038** (0.000)	0.091 (0.002)	0.154 (0.002)	1.56 (0.01)	0.22 (0.00)	−++++−
E8‡	**0.018** (0.000)	0.068 (0.002)	0.110 (0.003)	1.80 (0.01)	0.31 (0.01)	+−−++−
E9	**0.003** (0.000)	**0.014** (0.001)	**0.023** (0.002)	2.48 (0.02)	0.27 (0.02)	+−−−+−
E10	**0.000** (0.000)	**0.000** (0.000)	**0.001** (0.000)	3.72 (0.07)	0.29 (0.03)	+++−+−
E11	**0.000** (0.000)	**0.000** (0.000)	**0.000** (0.000)	4.31 (0.10)	0.35 (0.03)	−++−+−
E12	**0.000** (0.000)	**0.000** (0.000)	**0.000** (0.000)	3.68 (0.05)	0.17 (0.02)	+−+−−−
E13	**0.000** (0.000)	**0.000** (0.000)	**0.000** (0.000)	3.90 (0.04)	0.15 (0.02)	++−−−−
E14	**0.000** (0.000)	**0.000** (0.000)	**0.000** (0.000)	4.03 (0.09)	0.30 (0.04)	++−++−
E15	**0.000** (0.000)	**0.000** (0.000)	**0.000** (0.000)	4.03 (0.13)	0.38 (0.06)	+−+++−
E16	**0.000** (0.000)	**0.000** (0.000)	**0.000** (0.000)	4.44 (0.05)	0.12 (0.01)	−+−+−−
E17	**0.000** (0.000)	**0.000** (0.000)	**0.000** (0.000)	4.70 (0.07)	0.19 (0.02)	++−+−−
E18	**0.000** (0.000)	**0.000** (0.000)	**0.000** (0.000)	3.94 (0.09)	0.26 (0.04)	−+−++−
E19	**0.000** (0.000)	**0.000** (0.000)	**0.000** (0.000)	5.23 (0.43)	0.57 (0.13)	−−+−−−
E20	**0.000** (0.000)	**0.000** (0.000)	**0.000** (0.000)	5.66 (0.29)	0.28 (0.09)	+−+−−−
E21	**0.000** (0.000)	**0.000** (0.000)	**0.000** (0.000)	6.38 (0.33)	0.24 (0.08)	−−+++−
E22	**0.000** (0.000)	**0.000** (0.000)	**0.000** (0.000)	5.62 (0.21)	0.17 (0.07)	−−+−+−
E23	**0.000** (0.000)	**0.000** (0.000)	**0.000** (0.000)	4.86 (0.43)	0.70 (0.13)	−+−−+−
E24	**0.000** (0.000)	**0.000** (0.000)	**0.000** (0.000)	5.61 (0.17)	0.23 (0.04)	+−+−+−
E25	**0.000** (0.000)	**0.000** (0.000)	**0.000** (0.000)	6.32 (0.71)	0.52 (0.20)	++−−+−

Standard errors are shown in parentheses. Boldface without underline indicates significance ($p < 0.05$) for the null hypothesis that the edge is true (outside mode). Boldface with underline indicates significance ($p > 0.95$) for the null hypothesis that the edge is not *true (inside mode).* † *Edges included in T1.* ‡ *Edges included in T7.*

confidence of the outcome needs to be assessed. A mathematical procedure for statistically evaluating the outcome is provided in the following sections.

3. NON-SELECTIVE INFERENCE FOR THE PROBLEM OF REGIONS

3.1. The Problem of Regions

For developing the theory, we consider $(m + 1)$-dimensional multivariate normal random vector $\boldsymbol{Y}$, $m \geq 0$, with unknown mean vector $\boldsymbol{\mu} \in \mathbb{R}^{m+1}$ and the identity variance matrix $\boldsymbol{I}_{m+1}$:

$$\boldsymbol{Y} \sim N_{m+1}(\boldsymbol{\mu}, \boldsymbol{I}_{m+1}). \tag{4}$$

A region of interest such as tree and edge is denoted as $\mathcal{R} \subset \mathbb{R}^{m+1}$, and its complement set is denoted as $\mathcal{R}^C = \mathbb{R}^{m+1} \setminus \mathcal{R}$. There are K_{all} regions $\mathcal{R}_i$, $i = 1, \ldots, K_{\text{all}}$, and we simply write $\mathcal{R}$ for one of them by dropping the index i. Observing $\boldsymbol{Y} = \boldsymbol{y}$, the null hypothesis $H_0 : \boldsymbol{\mu} \in \mathcal{R}$ is tested against the alternative hypothesis $H_1 : \boldsymbol{\mu} \in \mathcal{R}^C$. This setting is called *problem of regions*, and the geometric theory for non-selective inference for slightly generalized settings (e.g., exponential family of distributions) has been discussed in Efron and Tibshirani (1998) and Shimodaira (2004). This theory allows arbitrary shape of $\mathcal{R}$ without assuming a particular shape such as half-space or sphere, and only requires the expression (29) of section 6.3.

The problem of regions is well described by geometric quantities (**Figure 4**). Let $\hat{\boldsymbol{\mu}}$ be the projection of $\boldsymbol{y}$ to the boundary surface $\partial\mathcal{R}$ defined as

$$\hat{\boldsymbol{\mu}} = \operatorname*{arg\,min}_{\boldsymbol{\mu} \in \partial\mathcal{R}} \|\boldsymbol{y} - \boldsymbol{\mu}\|,$$

and β_0 be the *signed distance* defined as $\beta_0 = \|\boldsymbol{y} - \hat{\boldsymbol{\mu}}\| > 0$ for $\boldsymbol{y} \in \mathcal{R}^C$ and $\beta_0 = -\|\boldsymbol{y} - \hat{\boldsymbol{\mu}}\| \leq 0$ for $\boldsymbol{y} \in \mathcal{R}$; see **Figures 4A,B**, respectively. A large β_0 indicates the evidence for rejecting $H_0 : \boldsymbol{\mu} \in \mathcal{R}$, but computation of p-value will also depend on the shape of $\mathcal{R}$. There should be many parameters for defining the shape, but we only need the *mean curvature* of $\partial\mathcal{R}$ at $\hat{\boldsymbol{\mu}}$, which represents the amount of surface bending. It is denoted as $\beta_1 \in \mathbb{R}$, and defined in (30).

Geometric quantities β_0 and β_1 of regions for trees (T1, . . . , T105) and edges (E1, . . . , E25) are plotted in **Figure 5**, and these values are also found in **Tables 1**, **2**. Although the phylogenetic model of evolution for the molecular dataset $\mathcal{X}_n = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$ is different from the multivariate normal model (4) for $\boldsymbol{y}$, the multiscale bootstrap method of section 3.4 estimates β_0 and β_1 using the non-parametric bootstrap probabilities (section 6.1) with bootstrap replicates $\mathcal{X}^*_{n'}$ for several values of sample size n'.

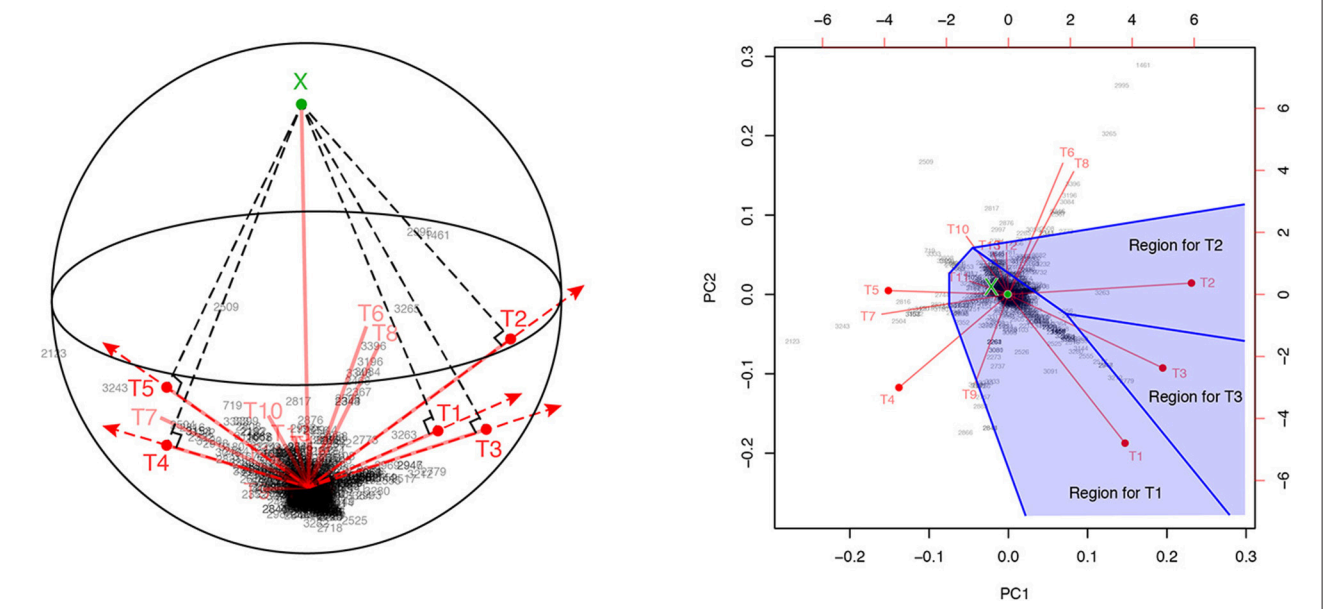

FIGURE 3 | Model map: Visualization of ML estimates of probability distributions for the best 15 trees. The origin represents the star-shaped tree topology (obtained by reducing the internal branches to zero length). Sites of amino acid sequences $t = 1, \ldots, n$ (black numbers) and probability distributions for trees T1, . . . ,T15 (red segments) are drawn by biplot of PCA. Auxiliary lines are drawn by hand. **(A)** 3-dimensional visualization using PC1, PC2, and PC3. The reconstructed data point X is also shown (green point). The ML estimates are represented as the end points of the red segments (shown by red points only for the best five trees), and they are placed on the sphere with the origin and X being placed at the poles. **(B)** The top-view of model map. Regions for the best three trees Ti, $i = 1, 2, 3$ (blue shaded regions) are illustrated; Ti will be the ML tree if X is included in the region for Ti.

3.2. Bootstrap Probability

For simulating (4) from $\boldsymbol{y}$, we may generate replicates $\boldsymbol{Y}^*$ from the bootstrap distribution (**Figure 4C**)

$$\boldsymbol{Y}^* \sim N_{m+1}(\boldsymbol{y}, \boldsymbol{I}_{m+1}), \tag{5}$$

and define bootstrap probability (BP) of $\mathcal{R}$ as the probability of $\boldsymbol{Y}^*$ being included in the region $\mathcal{R}$:

$$\mathrm{BP}(\mathcal{R}|\boldsymbol{y}) := P(\boldsymbol{Y}^* \in \mathcal{R}|\boldsymbol{y}). \tag{6}$$

$\mathrm{BP}(\mathcal{R}|\boldsymbol{y})$ can be interpreted as the Bayesian posterior probability $P(\boldsymbol{\mu} \in \mathcal{R}|\boldsymbol{y})$, because, by assuming the flat prior distribution $\pi(\boldsymbol{\mu}) = \text{constant}$, the posterior distribution $\boldsymbol{\mu}|\boldsymbol{y} \sim N_{m+1}(\boldsymbol{y}, \boldsymbol{I}_{m+1})$ is identical to the distribution of $\boldsymbol{Y}^*$ in (5). An interesting consequence of the geometric theory of Efron and Tibshirani (1998) is that BP can be expressed as

$$\mathrm{BP}(\mathcal{R}|\boldsymbol{y}) \simeq \bar{\Phi}(\beta_0 + \beta_1), \tag{7}$$

where $\simeq$ indicates the *second order asymptotic accuracy*, meaning that the equality is correct up to $O_p(n^{-1/2})$ with error of order $O_p(n^{-1})$; see section 6.3.

For understanding the formula (7), assume that $\mathcal{R}$ is a half space so that $\partial\mathcal{R}$ is flat and $\beta_1 = 0$. Since we only have to look at the axis orthogonal to $\partial\mathcal{R}$, the distribution of signed distance is identified as (1) with $\beta_0 = z$. The bootstrap distribution for (1) is $Z^* \sim N(z, 1)$, and bootstrap probability is expressed as $P(Z^* \le 0|z) = \bar{\Phi}(z)$. Therefore, we have $\mathrm{BP}(\mathcal{R}|\boldsymbol{y}) = \bar{\Phi}(\beta_0)$. For general $\mathcal{R}$ with curved $\partial\mathcal{R}$, the formula (7) adjusts the bias caused by β_1. As seen in **Figure 4C**, $\mathcal{R}$ becomes smaller for $\beta_1 > 0$ than $\beta_1 = 0$, and BP becomes smaller.

BP of $\mathcal{R}^C$ is closely related to BP of $\mathcal{R}$. From the definition,

$$\mathrm{BP}(\mathcal{R}^C|\boldsymbol{y}) = 1 - \mathrm{BP}(\mathcal{R}|\boldsymbol{y}) \simeq 1 - \bar{\Phi}(\beta_0 + \beta_1) = \bar{\Phi}(-\beta_0 - \beta_1). \tag{8}$$

The last expression also implies that the signed distance and the mean curvature of $\mathcal{R}^C$ is $-\beta_0$ and $-\beta_1$, respectively; this relation is also obtained by reversing the sign of v in (29).

3.3. Approximately Unbiased Test

Although $\mathrm{BP}(\mathcal{R}|\boldsymbol{y})$ may work as a Bayesian confidence measure, we would like to have a frequentist confidence measure for testing $H_0 : \boldsymbol{\mu} \in \mathcal{R}$ against $H_1 : \boldsymbol{\mu} \in \mathcal{R}^C$. The signed distance of $\boldsymbol{Y}$ is denoted as $\beta_0(\boldsymbol{Y})$, and consider the region $\{\boldsymbol{Y} \mid \beta_0(\boldsymbol{Y}) > \beta_0\}$ in which the signed distance is larger than the observed value $\beta_0 = \beta_0(\boldsymbol{y})$. Similar to (2), we then define an approximately unbiased (AU) p-value as

$$\mathrm{AU}(\mathcal{R}|\boldsymbol{y}) := P(\beta_0(\boldsymbol{Y}) > \beta_0 \mid \boldsymbol{\mu} = \hat{\boldsymbol{\mu}}) = \mathrm{BP}(\{\boldsymbol{Y} \mid \beta_0(\boldsymbol{Y}) > \beta_0\}|\hat{\boldsymbol{\mu}}), \tag{9}$$

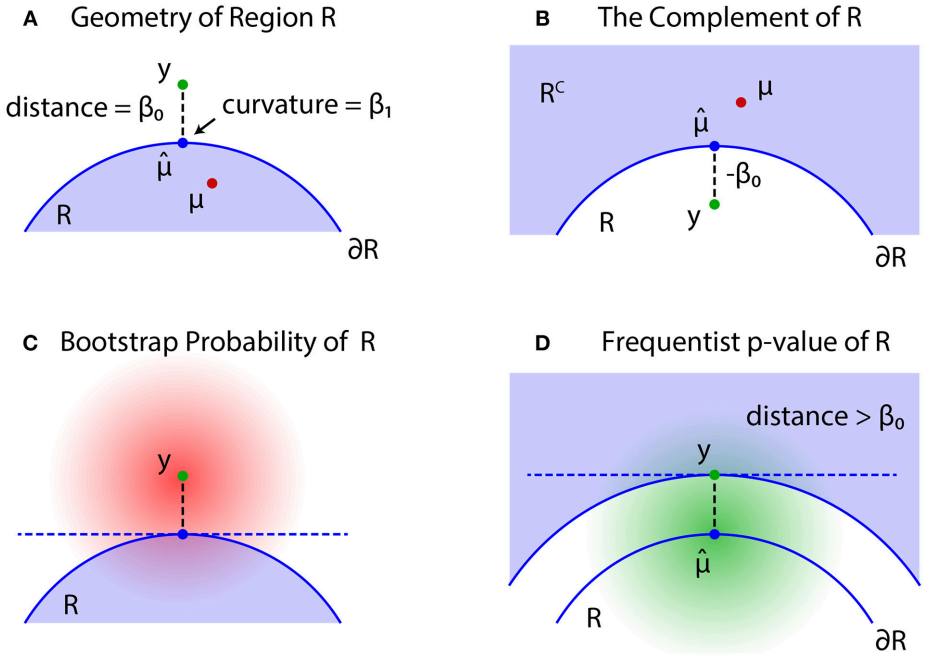

FIGURE 4 | Problem of regions. **(A)** $\beta_0 > 0$ when $\boldsymbol{y} \in \mathcal{R}^C$, then select the null hypothesis $\boldsymbol{\mu} \in \mathcal{R}$. **(B)** $\beta_0 \leq 0$ when $\boldsymbol{y} \in \mathcal{R}$, then select the null hypothesis $\boldsymbol{\mu} \in \mathcal{R}^C$. **(C)** The bootstrap distribution of $\boldsymbol{Y}^* \sim N_{m+1}(\boldsymbol{y}, \boldsymbol{I}_{m+1})$ (red shaded distribution). **(D)** The null distribution of $\boldsymbol{Y} \sim N_{m+1}(\hat{\boldsymbol{\mu}}, \boldsymbol{I}_{m+1})$ (green shaded distribution).

where the probability is calculated for $\boldsymbol{Y} \sim N_{m+1}(\hat{\boldsymbol{\mu}}, \boldsymbol{I}_{m+1})$ as illustrated in **Figure 4D**. The shape of the region $\{\boldsymbol{Y} \mid \beta_0(\boldsymbol{Y}) > \beta_0\}$ is very similar to the shape of $\mathcal{R}^C$; the difference is in fact only $O_p(n^{-1})$. Let us think of a point $\boldsymbol{y}'$ with signed distance $-\beta_0$ (shown as $\boldsymbol{y}$ in **Figure 4B**). Then we have

$$\mathrm{AU}(\mathcal{R}|\boldsymbol{y}) \simeq \mathrm{BP}(\mathcal{R}^C|\boldsymbol{y}') \simeq \bar{\Phi}(\beta_0 - \beta_1), \tag{10}$$

where the last expression is obtained by substituting $(-\beta_0, \beta_1)$ for (β_0, β_1) in (8). This formula computes AU from (β_0, β_1). An intuitive interpretation of (10) is explained in section 6.4.

In non-selective inference, *p*-values are computed using formula (10). If $\mathrm{AU}(\mathcal{R}|\boldsymbol{y}) < \alpha$, the null hypothesis $H_0 : \boldsymbol{\mu} \in \mathcal{R}$ is rejected and the alternative hypothesis $H_1 : \boldsymbol{\mu} \in \mathcal{R}^C$ is accepted. This test procedure is approximately unbiased, because it controls the non-selective type-I error as

$$P\big(\mathrm{AU}(\mathcal{R}|\boldsymbol{Y}) < \alpha \mid \boldsymbol{\mu} \in \partial\mathcal{R}\big) \simeq \alpha, \tag{11}$$

and the rejection probability increases as $\boldsymbol{\mu}$ moves away from $\mathcal{R}$, while it decreases as $\boldsymbol{\mu}$ moves into $\mathcal{R}$.

Exchanging the roles of $\mathcal{R}$ and $\mathcal{R}^C$ also allows for another hypothesis testing. AU of $\mathcal{R}^C$ is obtained from (9) by reversing the inequality as $\mathrm{AU}(\mathcal{R}^C|\boldsymbol{y}) = \mathrm{BP}(\{\boldsymbol{Y} \mid \beta_0(\boldsymbol{Y}) < \beta_0\}|\hat{\boldsymbol{\mu}}) = 1 - \mathrm{AU}(\mathcal{R}|\boldsymbol{y})$. This is also confirmed by substituting $(-\beta_0, -\beta_1)$, i.e., the geometric quantities of $\mathcal{R}^C$, for (β_0, β_1) in (10) as

$$\mathrm{AU}(\mathcal{R}^C|\boldsymbol{y}) \simeq \bar{\Phi}(-\beta_0 + \beta_1) \simeq 1 - \mathrm{AU}(\mathcal{R}|\boldsymbol{y}). \tag{12}$$

If $\mathrm{AU}(\mathcal{R}^C|\boldsymbol{y}) < \alpha$ or equivalently $\mathrm{AU}(\mathcal{R}|\boldsymbol{y}) > 1 - \alpha$, then we reject $H_0 : \boldsymbol{\mu} \in \mathcal{R}^C$ and accept $H_1 : \boldsymbol{\mu} \in \mathcal{R}$.

3.4. Multiscale Bootstrap

In order to estimate β_0 and β_1 from bootstrap probabilities, we consider a generalization of (5) as

$$\boldsymbol{Y}^* \sim N_{m+1}(\boldsymbol{y}, \sigma^2 \boldsymbol{I}_{m+1}), \tag{13}$$

for a variance $\sigma^2 > 0$, and define multiscale bootstrap probability of $\mathcal{R}$ as

$$\mathrm{BP}_{\sigma^2}(\mathcal{R}|\boldsymbol{y}) := P_{\sigma^2}(\boldsymbol{Y}^* \in \mathcal{R}|\boldsymbol{y}), \tag{14}$$

where P_{σ^2} indicates the probability with respect to (13).

Although our theory is based on the multivariate normal model, the actual implementation of the algorithm uses the non-parametric bootstrap probabilities in section 6.1. To fill the gap between the two models, we consider a non-linear transformation $\boldsymbol{f}_n$ so that the multivariate normal model holds at least approximately for $\boldsymbol{y} = \boldsymbol{f}_n(\mathcal{X}_n)$ and $\boldsymbol{Y}^* = \boldsymbol{f}_n(\mathcal{X}^*_{n'})$. An example of $\boldsymbol{f}_n$ is given in (25) for phylogenetic inference. Surprisingly, a specification of $\boldsymbol{f}_n$ is *not required* for computing *p*-values, but we simply assume the existence of such a transformation; this property may be called as "bootstrap trick." For phylogenetic inference, we compute the non-parametric bootstrap probabilities by (24) and substitute these values for (14) with $\sigma^2 = n/n'$.

For estimating β_0 and β_1, we need to have a scaling law which explains how BP_{σ^2} depends on the scale σ. We rescale (13) by multiplying σ^{-1} so that $\sigma^{-1}\boldsymbol{Y}^* \sim N_{m+1}(\sigma^{-1}\boldsymbol{y}, \boldsymbol{I}_{m+1})$ has the variance $\sigma^2 = 1$. $\boldsymbol{y}$ and $\mathcal{R}$ are now resaled by the factor σ^{-1}, which amounts to signed distance $\beta_0\sigma^{-1}$ and mean curvature

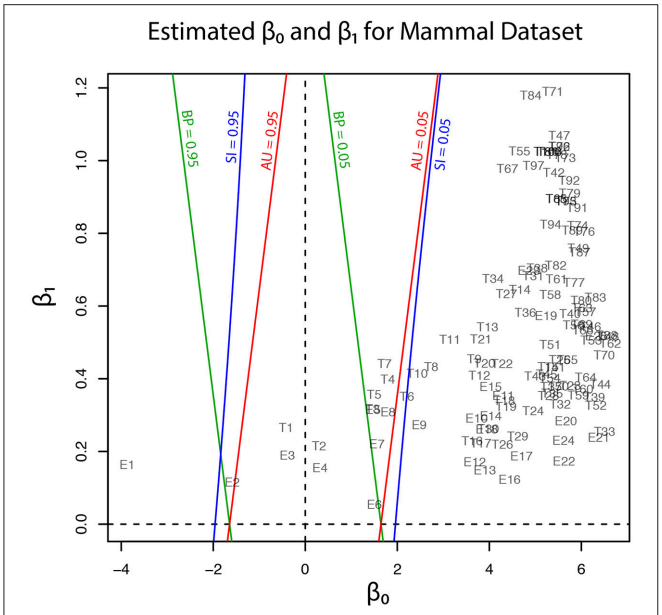

FIGURE 5 | Geometric quantities of regions (β_0 and β_1) for trees and edges are estimated by the multiscale bootstrap method (section 3.4). The three types of p-value (BP, AU, SI) are computed from β_0 and β_1, and their contour lines are drawn at $p = 0.05$ and 0.95.

$\beta_1\sigma$ (Shimodaira, 2004). Therefore, by substituting $(\beta_0\sigma^{-1}, \beta_1\sigma)$ for (β_0, β_1) in (7), we obtain

$$\mathrm{BP}_{\sigma^2}(\mathcal{R}|\boldsymbol{y}) \simeq \bar{\Phi}(\beta_0\sigma^{-1} + \beta_1\sigma). \tag{15}$$

For better illustrating how BP_{σ^2} depends on σ^2, we define

$$\psi_{\sigma^2}(\mathcal{R}|\boldsymbol{y}) := \sigma\bar{\Phi}^{-1}(\mathrm{BP}_{\sigma^2}(\mathcal{R}|\boldsymbol{y})) \simeq \beta_0 + \beta_1\sigma^2. \tag{16}$$

We can estimate β_0 and β_1 as regression coefficients by fitting the linear model (16) in terms of σ^2 to the observed values of non-parametric bootstrap probabilities (**Figure 6**). Interestingly, (10) is rewritten as $\mathrm{AU}(\mathcal{R}|\boldsymbol{y}) \simeq \bar{\Phi}(\psi_{-1}(\mathcal{R}|\boldsymbol{y}))$ by formally letting $\sigma^2 = -1$ in the last expression of (16), meaning that AU corresponds to $n' = -n$. Although σ^2 should be positive in (15), we can think of negative σ^2 in $\beta_0 + \beta_1\sigma^2$. See section 6.5 for details of model fitting and extrapolation to negative σ^2.

4. SELECTIVE INFERENCE FOR THE PROBLEM OF REGIONS

4.1. Approximately Unbiased Test for Selective Inference

In order to argue selective inference for the problem of regions, we have to specify the selection event. Let us consider a selective region $\mathcal{S} \subset \mathcal{R}^{m+1}$ so that we perform the hypothesis testing only when $\boldsymbol{y} \in \mathcal{S}$. Terada and Shimodaira (2017) considered a general shape of $\mathcal{S}$, but here we treat only two special cases of $\mathcal{S} = \mathcal{R}^C$ and $\mathcal{S} = \mathcal{R}$; see section 6.6. Our problem is formulated as follows. Observing $\boldsymbol{Y} = \boldsymbol{y}$ from the multivariate normal model (4), we first check whether $\boldsymbol{y} \in \mathcal{R}^C$ or $\boldsymbol{y} \in \mathcal{R}$. If $\boldsymbol{y} \in \mathcal{R}^C$ and we are interested in the null hypothesis $H_0 : \boldsymbol{\mu} \in \mathcal{R}$, then we may test it against the alternative hypothesis $H_1 : \boldsymbol{\mu} \in \mathcal{R}^C$. If $\boldsymbol{y} \in \mathcal{R}$ and we are interested in the null hypothesis $H_0 : \boldsymbol{\mu} \in \mathcal{R}^C$, then we may test it against the alternative hypothesis $H_1 : \boldsymbol{\mu} \in \mathcal{R}$. In this paper, the former case ($\boldsymbol{y} \in \mathcal{R}^C$, and so $\beta_0 > 0$) is called as *outside mode*, and the latter case ($\boldsymbol{y} \in \mathcal{R}$, and so $\beta_0 \le 0$) is called as *inside mode*. We do not know which of the two modes of testing is performed until we observe $\boldsymbol{y}$.

Let us consider the outside mode by assuming that $\boldsymbol{y} \in \mathcal{R}^C$, where $\beta_0 > 0$. Recalling that $p(z, c) = p(z)/\bar{\Phi}(c)$ in section 1, we divide $\mathrm{AU}(\mathcal{R}|\boldsymbol{y})$ by the selection probability to define a selective inference p-value as

$$\mathrm{SI}(\mathcal{R}|\boldsymbol{y}) := \frac{P(\beta_0(\boldsymbol{Y}) > \beta_0 \mid \boldsymbol{\mu} = \hat{\boldsymbol{\mu}})}{P(\boldsymbol{Y} \in \mathcal{R}^C \mid \boldsymbol{\mu} = \hat{\boldsymbol{\mu}})} = \frac{\mathrm{AU}(\mathcal{R}|\boldsymbol{y})}{\mathrm{BP}(\mathcal{R}^C|\hat{\boldsymbol{\mu}})}. \tag{17}$$

From the definition, $\mathrm{SI}(\mathcal{R}|\boldsymbol{y}) \in (0, 1)$, because $\{\boldsymbol{Y} \mid \beta_0(\boldsymbol{Y}) > \beta_0\} \subset \mathcal{R}^C$ for $\beta_0 > 0$. This p-value is computed from (β_0, β_1) by

$$\mathrm{SI}(\mathcal{R}|\boldsymbol{y}) \simeq \frac{\bar{\Phi}(\beta_0 - \beta_1)}{\bar{\Phi}(-\beta_1)}, \tag{18}$$

where $\mathrm{BP}(\mathcal{R}^C|\hat{\boldsymbol{\mu}}) = \bar{\Phi}(-\beta_1)$ is obtained by substituting $(0, \beta_1)$ for (β_0, β_1) in (8). An intuitive justification of (18) is explained in section 6.4.

For the outside mode of selective inference, p-values are computed using formula (18). If $\mathrm{SI}(\mathcal{R}|\boldsymbol{y}) < \alpha$, then reject $H_0 : \boldsymbol{\mu} \in \mathcal{R}$ and accept $H_1 : \boldsymbol{\mu} \in \mathcal{R}^C$. This test procedure is approximately unbiased, because it controls the selective type-I error as

$$P\big(\mathrm{SI}(\mathcal{R}|\boldsymbol{Y}) < \alpha \mid \boldsymbol{Y} \in \mathcal{R}^C, \boldsymbol{\mu} \in \partial\mathcal{R}\big) \simeq \alpha, \tag{19}$$

and the rejection probability increases as $\boldsymbol{\mu}$ moves away from $\mathcal{R}$, while it decreases as $\boldsymbol{\mu}$ moves into $\mathcal{R}$.

Now we consider the inside mode by assuming that $\boldsymbol{y} \in \mathcal{R}$, where $\beta_0 \le 0$. SI of $\mathcal{R}^C$ is obtained from (17) by exchanging the roles of $\mathcal{R}$ and $\mathcal{R}^C$.

$$\mathrm{SI}(\mathcal{R}^C|\boldsymbol{y}) = \frac{\mathrm{AU}(\mathcal{R}^C|\boldsymbol{y})}{\mathrm{BP}(\mathcal{R}|\hat{\boldsymbol{\mu}})} \simeq \frac{\bar{\Phi}(-\beta_0 + \beta_1)}{\bar{\Phi}(\beta_1)}. \tag{20}$$

For the inside mode of selective inference, p-values are computed using formula (20). If $\mathrm{SI}(\mathcal{R}^C|\boldsymbol{y}) < \alpha$, then reject $H_0 : \boldsymbol{\mu} \in \mathcal{R}^C$ and accept $H_1 : \boldsymbol{\mu} \in \mathcal{R}$. Unlike the non-selective p-value $\mathrm{AU}(\mathcal{R}^C|\boldsymbol{y})$, $\mathrm{SI}(\mathcal{R}^C|\boldsymbol{y}) < \alpha$ is *not* equivalent to $\mathrm{SI}(\mathcal{R}|\boldsymbol{y}) > 1 - \alpha$, because $\mathrm{SI}(\mathcal{R}|\boldsymbol{y}) + \mathrm{SI}(\mathcal{R}^C|\boldsymbol{y}) \neq 1$. For convenience, we define

$$\mathrm{SI}'(\mathcal{R}|\boldsymbol{y}) := \begin{cases} \mathrm{SI}(\mathcal{R}|\boldsymbol{y}) & \boldsymbol{y} \in \mathcal{R}^C \\ 1 - \mathrm{SI}(\mathcal{R}^C|\boldsymbol{y}) & \boldsymbol{y} \in \mathcal{R} \end{cases} \tag{21}$$

so that $\mathrm{SI}' > 1 - \alpha$ implies $\mathrm{SI}(\mathcal{R}^C|\boldsymbol{y}) < \alpha$. In our numerical examples of **Figure 5**, **Tables 1**, **2**, SI' is simply denoted as SI. We do not need to consider (21) for BP and AU, because $\mathrm{BP}'(\mathcal{R}|\boldsymbol{y}) = \mathrm{BP}(\mathcal{R}|\boldsymbol{y})$ and $\mathrm{AU}'(\mathcal{R}|\boldsymbol{y}) = \mathrm{AU}(\mathcal{R}|\boldsymbol{y})$ from (8) and (12).

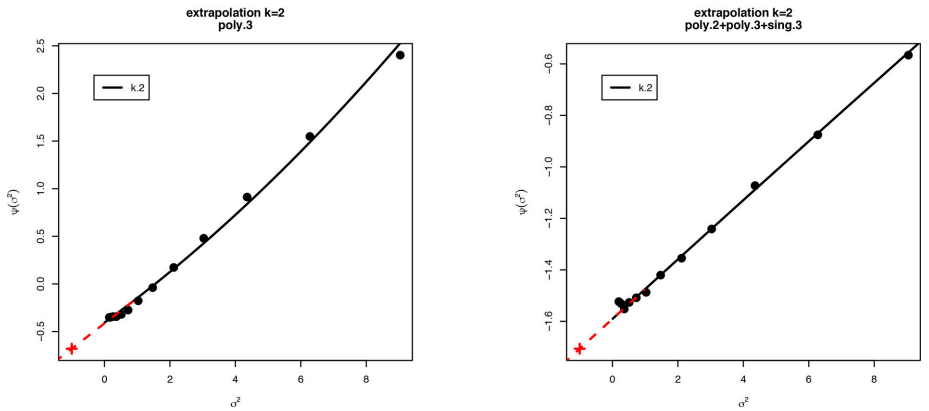

FIGURE 6 | Multiscale bootstrap for **(A)** tree T1 and **(B)** edge E2. $\psi_{\sigma^2}(\mathcal{R}|\boldsymbol{y})$ is computed by the non-parametric bootstrap probabilities for several $\sigma^2 = n/n'$ values, then β_0 and β_1 are estimated as the intercept and the slope, respectively. See section 6.5 for details.

4.2. Shortcut Computation of SI

We can compute SI from BP and AU. This will be useful for reanalyzing the results of previously published researches. Let us write $\mathrm{BP} = \mathrm{BP}(\mathcal{R}|\boldsymbol{y})$ and $\mathrm{AU} = \mathrm{AU}(\mathcal{R}|\boldsymbol{y})$. From (7) and (10), we have

$$\beta_0 = \tfrac{1}{2}\Big(\bar{\Phi}^{-1}(\mathrm{BP}) + \bar{\Phi}^{-1}(\mathrm{AU})\Big)$$
$$\beta_1 = \tfrac{1}{2}\Big(\bar{\Phi}^{-1}(\mathrm{BP}) - \bar{\Phi}^{-1}(\mathrm{AU})\Big).$$

We can compute SI from β_0 and β_1 by (18) or (20). More directly, we may compute

$$\mathrm{SI}(\mathcal{R}|\boldsymbol{y}) = \frac{\mathrm{AU}}{\bar{\Phi}\Big\{\tfrac{1}{2}\Big(\bar{\Phi}^{-1}(\mathrm{AU}) - \bar{\Phi}^{-1}(\mathrm{BP})\Big)\Big\}}$$
$$\mathrm{SI}(\mathcal{R}^C|\boldsymbol{y}) = \frac{1-\mathrm{AU}}{\bar{\Phi}\Big\{\tfrac{1}{2}\Big(\bar{\Phi}^{-1}(\mathrm{BP}) - \bar{\Phi}^{-1}(\mathrm{AU})\Big)\Big\}}.$$

4.3. Revisiting the Phylogenetic Inference

In this section, the analytical procedure outlined in section 2 is used to determine relationships among human, mouse, and rabbit. The question is: Which of mouse or human is closer to rabbit? The traditional view (Novacek, 1992) is actually supporting E6, the clade of rabbit and mouse, which is consistent with T4, T5, and T7. Based on molecular analysis, Graur et al. (1996) strongly suggested that rabbit is closer to human than mouse, thus supporting E2, which is consistent with T1, T2, and T3. However, Halanych (1998) criticized it by pointing out that E2 is an artifact caused by the *long branch attraction* (LBA) between mouse and opossum. In addition, Shimodaira and Hasegawa (1999) and Shimodaira (2002) suggested that T7 is not rejected by multiplicity adjusted tests. Shimodaira and Hasegawa (2005) showed that T7 becomes the ML tree by resolving the LBA using a larger dataset with more taxa. Although T1 is the ML tree based on the dataset with fewer taxa, T7 is presumably the true tree as indicated by later researches. With these observations in mind, we retrospectively interpret *p*-values in **Tables 1**, **2**.

The results are shown below for the two test modes (inside and outside) as defined in section 4.1. The extent of multiplicity and selection bias depends on the number of regions under consideration, thus these numbers are considered for interpreting the results. The numbers of regions related to trees and edges are summarized in **Table 3**; see section 6.7 for details.

In inside mode, the null hypothesis $H_0 : \boldsymbol{\mu} \in \mathcal{R}_i^C$ is tested against the alternative hypothesis $H_1 : \boldsymbol{\mu} \in \mathcal{R}_i$ for $\boldsymbol{y} \in \mathcal{R}_i$ (i.e., $\beta_0 \le 0$). This applies to the regions for T1, E1, E2, and E3, and they are *supported* by the data in the sense mentioned in the last paragraph of section 2. When H_0 is rejected by a test procedure, it is claimed that $\mathcal{R}_i$ is *significantly supported* by the data, indicating H_1 holds true. For convenience, the null hypothesis H_0 is said like E1 is not true, and the alternative hypothesis H_1 is said like E1 is true; then rejection of H_0 implies that E1 is true. This procedure looks unusual, but makes sense when both $\mathcal{R}_i$ and $\mathcal{R}_i^C$ are regions with nonzero volume. Note that selection bias can be very large in the sense that $K_{\mathrm{select}}/K_{\mathrm{all}} \approx 0$ for many taxa, and non-selective tests may lead to many false positives because $K_{\mathrm{true}}/K_{\mathrm{all}} \approx 1$. Therefore selective inference should be used in inside mode.

In outside mode, the null hypothesis $H_0 : \boldsymbol{\mu} \in \mathcal{R}_i$ is tested against the alternative hypothesis $H_1 : \boldsymbol{\mu} \in \mathcal{R}_i^C$ for $\boldsymbol{y} \in \mathcal{R}_i^C$ (i.e., $\beta_0 > 0$). This applies to the regions for T2, ..., T105, and E4, ..., E25, and they are *not supported* by the data. When H_0 is rejected by a test procedure, it is claimed that $\mathcal{R}_i$ is rejected.

TABLE 3 | The number of regions for trees and edges. The number of taxa is $N = 6$.

	Inside mode		Outside mode	
	Tree	Edge	Tree	Edge
K_{select}	1	3	104	22
K_{true}	104	22	1	3
K_{all}	105	25	105	25

For convenience, the null hypothesis is said like T9 is true, and the alternative hypothesis is said like T9 is not true; rejection of H_0 implies that T9 is not true. This is more or less a typical test procedure. Note that selection bias is minor in the sense that $K_{\text{select}}/K_{\text{all}} \approx 1$ for many taxa, and non-selective tests may result in few false positives because $K_{\text{true}}/K_{\text{all}} \approx 0$. Therefore selective inference is not much beneficial in outside mode.

In addition to p-values for some trees and edges, estimated geometric quantities are also shown in the tables. We confirm that the sign of β_0 is estimated correctly for all the trees and edges. The estimated β_1 values are all positive, indicating the regions are convex. This is not surprising, because the regions are expressed as intersections of half spaces at least locally (**Figure 3B**).

Now p-values are examined in inside mode. (T1, E3) BP, AU, SI are all $p \leq 0.95$. This indicates that T1 and E3 are *not* significantly supported. There are nothing claimed to be definite. (E1) BP, AU, SI are all $p > 0.95$, indicating E1 is significantly supported. Since E1 is associated with the best 15 trees T1, ..., T15, some of them are significantly better than the rest of trees T16, ..., T105. Significance for edges is common in phylogenetics as well as in hierarchical clustering (Suzuki and Shimodaira, 2006). (E2) The results split for this presumably wrong edge. AU > 0.95 suggests E2 is significantly supported, whereas BP, SI ≤ 0.95 are not significant. AU tends to violate the selective type-I error, leading to false positives or overconfidence in wrong trees/edges, whereas SI is approximately unbiased for the selected hypothesis. This overconfidence is explained by the inequality AU $>$ SI (meant SI$'$ here) for $\boldsymbol{y} \in \mathcal{R}$, which is obtained by comparing (12) and (20). Therefore SI is preferable to AU in inside mode. BP is safer than AU in the sense that BP $<$ AU for $\beta_1 > 0$, but BP is not guaranteed for controlling type-I error in a frequentist sense. The two inequalities (SI, BP $<$ AU) are verified as relative positions of the contour lines at $p = 0.95$ in **Figure 5**. The three p-values can be very different from each other for large β_1.

Next p-values are examined in outside mode. (T2, E4, E6) BP, AU, SI are all $p \geq 0.05$. They are *not* rejected, and there are nothing claimed to be definite. (T8, T9, ..., T105, E9,..., E25) BP, AU, SI are all $p < 0.05$. These trees and edges are rejected. (T7, E8) The results split for these presumably true tree and edge. BP < 0.05 suggests T7 and E8 are rejected, whereas AU, SI ≥ 0.05 are not significant. AU is approximately unbiased for controlling the type-I error when H_0 is specified in advance (Shimodaira, 2002). Since BP $<$ AU for $\beta_1 > 0$, BP violates the type-I error, which results in overconfidence in non-rejected wrong trees. Therefore BP should be avoided in outside mode. Inequality AU $<$ SI can be shown for $\boldsymbol{y} \in \mathcal{R}^C$ by comparing (10) and (18). Since the null hypothesis $H_0 : \boldsymbol{\mu} \in \mathcal{R}$ is chosen after looking at $\boldsymbol{y} \in \mathcal{R}^C$, AU is not approximately unbiased for controlling the selective type-I error, whereas SI adjusts this selection bias. The two inequalities (BP $<$ AU $<$ SI) are verified as relative positions of the contour lines at $p = 0.05$ in **Figure 5**. AU and SI behave similarly (Note: $K_{\text{select}}/K_{\text{all}} \approx 1$), while BP is very different from AU and SI for large β_1. It is arguable which of AU and SI is appropriate: AU is preferable to SI in tree selection ($K_{\text{true}} = 1$), because the multiplicity of testing is controlled as FWER $= P(\text{reject any true null}) = P(\text{AU}(\mathcal{R}_{\text{true tree}}|\boldsymbol{Y}) < \alpha \mid \boldsymbol{\mu} \in \mathcal{R}_{\text{true tree}}) \leq \alpha$. The FWER is multiplied by $K_{\text{true}} \geq 1$ for edge selection, and SI does not fix it either. For testing edges in outside mode, AU may be used for screening purpose with a small α value such as α/K_{true}.

5. CONCLUSION

We have developed a new method for computing selective inference p-values from multiscale bootstrap probabilities, and applied this new method to phylogenetics. It is demonstrated through theory and a real-data analysis that selective inference p-values are in particular useful for testing selected edges (i.e., clades or clusters of species) to claim that they are supported significantly if $p > 1 - \alpha$. On the other hand, the previously proposed non-selective version of approximately unbiased p-values are still useful for testing candidate trees to claim that they are rejected if $p < \alpha$. Although we focused on phylogenetics, our general theory of selective inference may be applied to other model selection problems, or more general selection problems.

6. REMARKS

6.1. Bootstrap Resampling of Log-Likelihoods

Non-parametric bootstrap is often time consuming for recomputing the maximum likelihood (ML) estimates for bootstrap replicates. Kishino et al. (1990) considered the resampling of estimated log-likelihoods (RELL) method for reducing the computation. Let $\mathcal{X}_n = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$ be the dataset of sample size n, where $\boldsymbol{x}_t$ is the site-pattern of amino acids at site t for $t = 1, \ldots, n$. By resampling $\boldsymbol{x}_t$ from $\mathcal{X}_n$ with replacement, we obtain a bootstrap replicate $\mathcal{X}^*_{n'} = (\boldsymbol{x}^*_1, \ldots, \boldsymbol{x}^*_{n'})$ of sample size n'. Although $n' = n$ for the ordinary bootstrap, we will use several $n' > 0$ values for the multiscale bootstrap. The parametric model of probability distribution for tree Ti is $p_i(\boldsymbol{x}; \boldsymbol{\theta}_i)$ for $i = 1, \ldots, 105$, and the log-likelihood function is $\ell_i(\boldsymbol{\theta}_i; \mathcal{X}_n) = \sum_{t=1}^n \log p_i(\boldsymbol{x}_t; \boldsymbol{\theta}_i)$. Computation of the ML estimate $\hat{\boldsymbol{\theta}}_i = \arg\max_{\boldsymbol{\theta}_i} \ell_i(\boldsymbol{\theta}_i; \mathcal{X}_n)$ is time consuming, so we do not recalculate $\hat{\boldsymbol{\theta}}^*_i = \arg\max_{\boldsymbol{\theta}_i} \ell_i(\boldsymbol{\theta}_i; \mathcal{X}^*_{n'})$ for bootstrap replicates. Define the site-wise log-likelihood at site t for tree Ti as

$$\xi_{ti} = \log p_i(\boldsymbol{x}_t; \hat{\boldsymbol{\theta}}_i), \quad t = 1, \ldots, n, \; i = 1, \ldots, 105, \tag{22}$$

so that the log-likelihood value for tree Ti is written as $\ell_i(\hat{\boldsymbol{\theta}}_i; \mathcal{X}_n) = \sum_{t=1}^n \xi_{ti}$. The bootstrap replicate of the log-likelihood value is approximated as

$$\ell_i(\hat{\boldsymbol{\theta}}^*_i; \mathcal{X}^*_{n'}) \approx \ell_i(\hat{\boldsymbol{\theta}}_i; \mathcal{X}^*_{n'}) = \sum_{t=1}^n w^*_t \xi_{ti}, \tag{23}$$

where w^*_t is the number of times $\boldsymbol{x}_t$ appears in $\mathcal{X}^*_{n'}$. The accuracy of this approximation as well as the higher-order term is given in Equations (4) and (5) of Shimodaira (2001). Once $\ell_i(\hat{\boldsymbol{\theta}}^*_i; \mathcal{X}^*_{n'})$, $i = 1, \ldots, 105$, are computed by (23), its ML tree is T$\hat{i}^*$ with $\hat{i}^* = \arg\max_{i=1,\ldots,105} \ell_i(\hat{\boldsymbol{\theta}}^*_i; \mathcal{X}^*_{n'})$.

The non-parametric bootstrap probability of tree Ti is obtained as follows. We generate B bootstrap replicates $\boldsymbol{X}^{*b}_{n'}$, $b = 1,\ldots,B$. In this paper, we used $B = 10^5$. For each $\boldsymbol{X}^{*b}_{n'}$, the ML tree $\hat{\mathrm{T}i}^{*b}$ is computed by the method described above. Then we count the frequency that Ti becomes the ML tree in the B replicates. The non-parametric bootstrap probability of tree Ti is computed by

$$\mathrm{BP}(\mathrm{T}i, n') = \#\{\hat{i}^{*b} = i,\ b = 1,\ldots,B\}/B. \tag{24}$$

The non-parametric bootstrap probability of a edge is computed by summing $\mathrm{BP}(\mathrm{T}i, n')$ over the associated trees.

An example of the transformation $\boldsymbol{Y}^* = \boldsymbol{f}_n(\mathcal{X}^*_{n'})$ mentioned in section 3.4 is

$$\boldsymbol{Y}^* = \boldsymbol{V}_n^{-1/2}\boldsymbol{L}^*_{n'}, \tag{25}$$

where $\boldsymbol{L}^*_{n'} = (1/n')(\ell^*_1,\ldots,\ell^*_{105})^T$ with $\ell^*_i = \ell_i(\hat{\boldsymbol{\theta}}^*_i; \mathcal{X}^*_{n'})$ and $\boldsymbol{V}_n$ is the variance matrix of $\boldsymbol{L}^*_n$. According to the approximation (23) and the central limit theorem, (13) holds well for sufficiently large n and n' with $m = 104$ and $\sigma^2 = n/n'$. It also follows from the above argument that $\mathrm{var}(\ell^*_i - \ell^*_j) \approx (n'/n)\|\boldsymbol{\xi}_i - \boldsymbol{\xi}_j\|^2$, and thus the variance of log-likelihood difference is

$$\mathrm{var}\Big(\ell_i(\hat{\boldsymbol{\theta}}_i; \mathcal{X}_n) - \ell_j(\hat{\boldsymbol{\theta}}_j; \mathcal{X}_n)\Big) \approx \|\boldsymbol{\xi}_i - \boldsymbol{\xi}_j\|^2, \tag{26}$$

which gives another insight into the visualization of section 6.2, where the variance can be interpreted as the divergence between the two models; see Equation (27). This approximation holds well when the two predictive distributions $p_i(\boldsymbol{x}; \hat{\boldsymbol{\theta}}_i)$, $p_j(\boldsymbol{x}; \hat{\boldsymbol{\theta}}_j)$ are not very close to each other. When they are close to each other, however, the higher-order term ignored in (26) becomes dominant, and there is a difficulty for deriving the limiting distribution of the log-likelihood difference in the model selection test (Shimodaira, 1997; Schennach and Wilhelm, 2017).

6.2. Visualization of Probability Models

For representing the probability distribution of tree Ti, we define $\boldsymbol{\xi}_i := (\xi_{1i},\ldots,\xi_{ni})^T \in \mathbb{R}^n$ from (22) for $i = 1,\ldots,15$. The idea behind the visualization of **Figure 3** is that locations of $\boldsymbol{\xi}_i$ in $\mathbb{R}^n$ will represent locations of $p_i(\boldsymbol{x}; \hat{\boldsymbol{\theta}}_i)$ in the space of probability distributions. Let $D_{\mathrm{KL}}(p_i\|p_j)$ be the Kullback-Leibler divergence between the two distributions. For sufficiently small $(1/n)\|\boldsymbol{\xi}_i - \boldsymbol{\xi}_j\|^2$, the squared distance in $\mathbb{R}^n$ approximates n times Jeffreys divergence

$$\|\boldsymbol{\xi}_i - \boldsymbol{\xi}_j\|^2 \approx n \times \Big(D_{\mathrm{KL}}(p_i(\boldsymbol{x}; \hat{\boldsymbol{\theta}}_i)\|p_j(\boldsymbol{x}; \hat{\boldsymbol{\theta}}_j)) + D_{\mathrm{KL}}(p_j(\boldsymbol{x}; \hat{\boldsymbol{\theta}}_j)\|p_i(\boldsymbol{x}; \hat{\boldsymbol{\theta}}_i))\Big) \tag{27}$$

for non-nested models (Shimodaira, 2001, section 6). When a model p_0 is nested in p_i, it becomes $\|\boldsymbol{\xi}_i - \boldsymbol{\xi}_0\|^2 \approx 2n \times D_{\mathrm{KL}}(p_i(\boldsymbol{x}; \hat{\boldsymbol{\theta}}_i)\|p_0(\boldsymbol{x}; \hat{\boldsymbol{\theta}}_0)) \approx 2 \times (\ell_i(\hat{\boldsymbol{\theta}}_i; \mathcal{X}_n) - \ell_0(\hat{\boldsymbol{\theta}}_0; \mathcal{X}_n))$. We explain three different visualizations of **Figure 7**. There are only minor differences between the plots, and the visualization is not sensitive to the details.

For dimensionality reduction, we have to specify the origin $\boldsymbol{c} \in \mathbb{R}^n$ and consider vectors $\boldsymbol{a}_i := \boldsymbol{\xi}_i - \boldsymbol{c}$. A naive choice would be the average $\boldsymbol{c} = \sum_{i=1}^{15} \boldsymbol{\xi}_i/15$. By applying PCA without centering and scaling (e.g., `prcomp` with option `center=FALSE, scale=FALSE` in R) to the matrix $(\boldsymbol{a}_1,\ldots,\boldsymbol{a}_{15})$, we obtain the visualization of $\boldsymbol{\xi}_i$ as the axes (red arrows) of biplot in **Figure 7A**.

For computing the "data point" X in **Figure 3**, we need more models. Let tree T106 be the star topology with no internal branch (completely unresolved tree), and T107,..., T131 be partially resolved tree topologies with only one internal branch corresponding to E1,..., E25, whereas T1,..., T105 are fully resolved trees (bifurcating trees). Then define $\boldsymbol{\eta}_i := \boldsymbol{\xi}_{106+i}$, $i = 0,\ldots,25$. Now we take $\boldsymbol{c} = \boldsymbol{\eta}_0$ for computing $\boldsymbol{a}_i = \boldsymbol{\xi}_i - \boldsymbol{\eta}_0$ and $\boldsymbol{b}_i = \boldsymbol{\eta}_i - \boldsymbol{\eta}_0$. There is hierarchy of models: $\boldsymbol{\eta}_0$ is the submodel nested in all the other models, and $\boldsymbol{\eta}_1, \boldsymbol{\eta}_2, \boldsymbol{\eta}_3$, for example, are submodels of $\boldsymbol{\xi}_1$ (T1 includes E1, E2, E3). By combining these non-nested models, we can reconstruct a comprehensive model in which all the other models are nested as submodels (Shimodaira, 2001, Equation 10 in section 5). The idea is analogous to reconstructing the full model $y = \beta_1 x_1 + \cdots + \beta_{25} x_{25} + \epsilon$ of multiple regression from submodels $y = \beta_1 x_1 + \epsilon, \ldots, y = \beta_{25} x_{25} + \epsilon$. Thus we call it as "full model" in this paper, and the ML estimate of the full model is indicated as the data point X; it is also said "super model" in Shimodaira and Hasegawa (2005). Let $\boldsymbol{B} = (\boldsymbol{b}_1,\ldots,\boldsymbol{b}_{25}) \in \mathbb{R}^{n\times 25}$ and $\boldsymbol{d} = (\|\boldsymbol{b}_1\|^2,\ldots,\|\boldsymbol{b}_{25}\|^2)^T \in \mathbb{R}^{25}$, then the vector for the full model is computed approximately by

$$\boldsymbol{a}_X = \boldsymbol{B}(\boldsymbol{B}^T\boldsymbol{B})^{-1}\boldsymbol{d}. \tag{28}$$

For the visualization of the best 15 trees, we may use only $\boldsymbol{b}_1,\ldots,\boldsymbol{b}_{11}$, because they include E1 and two more edges from E2,...,E11. In **Figures 3**, **7B**, we actually modified the above computation slightly so that the star topology T106 is replaced by T107, the partially resolved tree corresponding to E1 (T107 is also said star topology by treating clade (23) as a leaf of the tree), and the 10 partially resolved trees for E2,..., E11 are replaced by those for (E1,E2),..., (E1,E11), respectively; the origin becomes the maximal model nested in all the 15 trees, and X becomes the minimal full model containing all the 15 trees. Just before applying PCA in **Figure 7B**, $\boldsymbol{a}_1,\ldots,\boldsymbol{a}_{15}$ are projected to the space orthogonal to $\boldsymbol{a}_X$, so that the plot becomes the "top-view" of **Figure 3A** with $\boldsymbol{a}_X$ being at the origin.

In **Figure 7C**, we attempted a even simpler computation without using ML estimates for partially resolved trees. We used $\boldsymbol{B} = (\boldsymbol{a}_1,\ldots,\boldsymbol{a}_{15})$ and $\boldsymbol{d} = (\|\boldsymbol{a}_1\|^2,\ldots,\|\boldsymbol{a}_{15}\|^2)^T$, and taking the largest 10 singular values for computing the inverse in (28). The orthogonal projection to $\boldsymbol{a}_X$ is applied before PCA.

6.3. Asymptotic Theory of Smooth Surfaces

For expressing the shape of the region $\mathcal{R} \subset \mathbb{R}^{m+1}$, we use a local coordinate system $(\boldsymbol{u}, v) \in \mathbb{R}^{m+1}$ with $\boldsymbol{u} \in \mathbb{R}^m, v \in \mathbb{R}$. In a neighborhood of $\boldsymbol{y}$, the region is expressed as

$$\mathcal{R} = \{(\boldsymbol{u}, v) \mid v \le -h(\boldsymbol{u}),\ \boldsymbol{u} \in \mathbb{R}^m\}, \tag{29}$$

where h is a smooth function; see Shimodaira (2008) for the theory of non-smooth surfaces. The boundary surface $\partial\mathcal{R}$ is expressed as $v = -h(\boldsymbol{u})$, $\boldsymbol{u} \in \mathbb{R}^m$. We can choose the coordinates

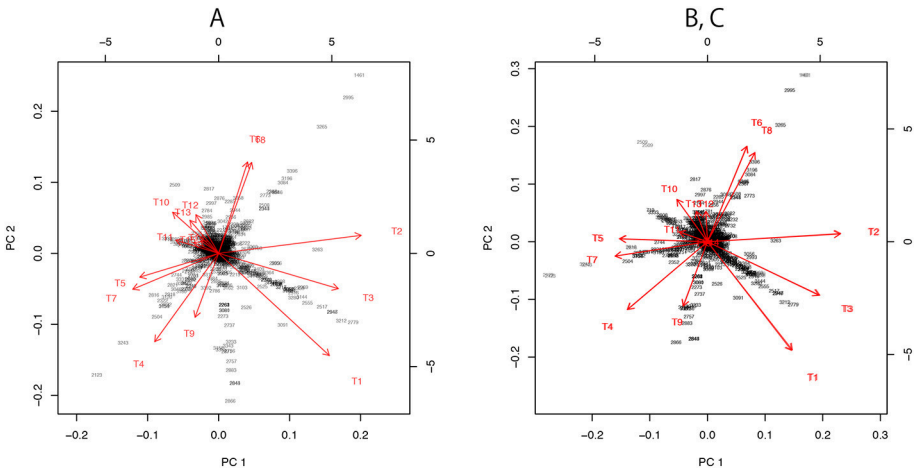

FIGURE 7 | Three versions the visualization of probability distributions for the best 15 trees drawn using different sets of models. **(A)** Only the 15 bifurcating trees. **(B)** 15 bifurcating trees + 10 partially resolved trees + 1 star topology. This is the same plot as **Figure 3B**. **(C)** 15 bifurcating trees + 1 star topology. Note that **(B,C)** are superimposed, since their plots are almost indistinguishable.

so that $\boldsymbol{y} = (\boldsymbol{0}, \beta_0)$ (i.e., $\boldsymbol{u} = (0, \ldots, 0)$ and $v = \beta_0$), and $h(\boldsymbol{0}) = 0$, $\partial h/\partial u_i|_{\boldsymbol{0}} = 0$, $i = 1, \ldots, m$. The projection now becomes the origin $\hat{\boldsymbol{\mu}} = (\boldsymbol{0}, 0)$, and the signed distance is β_0. The mean curvature of surface $\partial\mathcal{R}$ at $\hat{\boldsymbol{\mu}}$ is now defined as

$$\beta_1 = \frac{1}{2}\sum_{i=1}^{m} \frac{\partial^2 h(\boldsymbol{u})}{\partial u_i \partial u_i}\bigg|_{\boldsymbol{0}}, \tag{30}$$

which is interpreted as the trace of the hessian matrix of h. When $\mathcal{R}$ is convex at least locally in the neighborhood, all the eigenvalues of the hessian are non-negative, leading to $\beta_1 \geq 0$, whereas concave $\mathcal{R}$ leads to $\beta_1 \leq 0$. In particular, $\beta_1 = 0$ when $\partial\mathcal{R}$ is flat (i.e., $h(\boldsymbol{u}) \equiv 0$).

Since the transformation $\boldsymbol{y} = \boldsymbol{f}_n(\mathcal{X}_n)$ depends on n, the shape of the region $\mathcal{R}$ actually depends on n, although the dependency is implicit in the notation. As n goes larger, the standard deviation of estimates, in general, reduces at the rate $n^{-1/2}$. For keeping the variance constant in (4), we actually magnifying the space by the factor $n^{1/2}$, meaning that the boundary surface $\partial\mathcal{R}$ approaches flat as $n \to \infty$. More specifically, the magnitude of mean curvature is of order $\beta_1 = O_p(n^{-1/2})$. The magnitude of $\partial^3 h/\partial u_i \partial u_j \partial u_k$ and higher order derivatives is $O_p(n^{-1})$, and we ignore these terms in our asymptotic theory. For keeping $\boldsymbol{\mu} = O(1)$ in (4), we also consider the setting of "local alternatives," meaning that the parameter values approach a origin on the boundary at the rate $n^{-1/2}$.

6.4. Bridging the Problem of Regions to the *Z*-Test

Here we explain the problem of regions in terms of the z-test by bridging the multivariate problem of section 3 to the 1-dimensional case of section 1.

Ideal p-values are uniformly distributed over $p \in (0, 1)$ when the null hypothesis holds. In fact, $\mathrm{AU}(\mathcal{R}|\boldsymbol{Y}) \sim U(0, 1)$ for $\boldsymbol{\mu} \in \partial\mathcal{R}$ as indicated in (11). The statistic $\mathrm{AU}(\mathcal{R}|\boldsymbol{Y})$ may be called *pivotal* in the sense that the distribution does not change when $\boldsymbol{\mu} \in \partial\mathcal{R}$ moves on the surface. Here we ignore the error of $O_p(n^{-1})$, and consider only the second order asymptotic accuracy. From (10), we can write $\mathrm{AU}(\mathcal{R}|\boldsymbol{Y}) \simeq \bar{\Phi}(\beta_0(\boldsymbol{Y}) - \beta_1(\boldsymbol{Y}))$, where the notation such as $\beta_0(\boldsymbol{Y})$ and $\beta_1(\boldsymbol{Y})$ indicates the dependency on $\boldsymbol{Y}$. Since $\beta_1(\boldsymbol{Y}) \simeq \beta_1(\boldsymbol{y}) = \beta_1$, we treat $\beta_1(\boldsymbol{Y})$ as a constant. Now we get the normal pivotal quantity (Efron, 1985) as $\bar{\Phi}^{-1}(\mathrm{AU}(\mathcal{R}|\boldsymbol{Y})) = \beta_0(\boldsymbol{Y}) - \beta_1 \sim N(0, 1)$ for $\boldsymbol{\mu} \in \partial\mathcal{R}$. More generally, it becomes

$$\beta_0(\boldsymbol{Y}) - \beta_1 \sim N(\beta_0(\boldsymbol{\mu}), 1), \quad \boldsymbol{\mu} \in \mathbb{R}^{m+1}. \tag{31}$$

Let us look at the z-test in section 1, and consider substitutions:

$$Z = \beta_0(\boldsymbol{Y}) - \beta_1, \quad \theta = \beta_0(\boldsymbol{\mu}), \quad c = -\beta_1. \tag{32}$$

The 1-dimensional model (1) is now equivalent to (31). The null hypothesis is also equivalent: $\theta \leq 0 \Leftrightarrow \beta_0(\boldsymbol{\mu}) \leq 0 \Leftrightarrow \boldsymbol{\mu} \in \mathcal{R}$. We can easily verify that AU corresponds to $p(z)$, because $p(z) = \bar{\Phi}(z) = \bar{\Phi}(\beta_0(\boldsymbol{y}) - \beta_1) \simeq \mathrm{AU}(\mathcal{R}|\boldsymbol{y})$, which is expected from the way we obtained (31) above. Furthermore, we can derive SI from $p(z, c)$. First verify that the selection event is equivalent: $Z > c \Leftrightarrow \beta_0(\boldsymbol{Y}) - \beta_1 > -\beta_1 \Leftrightarrow \beta_0(\boldsymbol{Y}) > 0 \Leftrightarrow \boldsymbol{Y} \in \mathcal{R}^C$. Finally, we obtain SI as $p(z, c) = p(z)/\bar{\Phi}(c) \simeq \bar{\Phi}(\beta_0(\boldsymbol{y}) - \beta_1)/\bar{\Phi}(-\beta_1) \simeq \mathrm{SI}(\mathcal{R}|\boldsymbol{y})$.

6.5. Model Fitting in Multiscale Bootstrap

We have used thirteen σ^2 values from 1/9 to 9 (equally spaced in log-scale). This range is relatively large, and we observe a slight deviation from the linear model $\beta_0 + \beta_1\sigma^2$ in **Figure 6**. Therefore we fit other models to the observed values of ψ_{σ^2} as implemented in *scaleboot* package (Shimodaira, 2008). For example, poly.k model is $\sum_{i=0}^{k-1} \beta_i \sigma^{2i}$, and sing.3 model is $\beta_0 + \beta_1\sigma^2(1 + \beta_2(\sigma - 1))^{-1}$. In **Figure 6A**, poly.3 is the best model according to AIC

(Akaike, 1974). In **Figure 6B**, poly.2, poly.3, and sing.3 are combined by model averaging with Akaike weights. Then β_0 and β_1 are estimated from the tangent line to the fitted curve of ψ_{σ^2} at $\sigma^2 = 1$. In **Figure 6**, the tangent line is drawn as red line for extrapolating ψ_{σ^2} to $\sigma^2 = -1$. Shimodaira (2008) and Terada and Shimodaira (2017) considered the Taylor expansion of ψ_{σ^2} at $\sigma^2 = 1$ as a generalization of the tangent line for improving the accuracy of AU and SI.

In the implementation of *CONSEL* (Shimodaira and Hasegawa, 2001) and *pvclust* (Suzuki and Shimodaira, 2006), we use a narrower range of σ^2 values (ten σ^{-2} values: 0.5, 0.6, ..., 1.4). Only the linear model $\beta_0 + \beta_1\sigma^2$ is fitted there. The estimated β_0 and β_1 should be very close to those estimated from the tangent line described above. An advantage of using wider range of σ^2 in *scaleboot* is that the standard error of β_0 and β_1 will become smaller.

6.6. General Formula of Selective Inference

Let $\mathcal{H}, \mathcal{S} \subset \mathbb{R}^{m+1}$ be regions for the null hypothesis and the selection event, respectively. We would like to test the null hypothesis $H_0 : \boldsymbol{\mu} \in \mathcal{H}$ against the alternative $H_1 : \boldsymbol{\mu} \in \mathcal{H}^C$ conditioned on the selection event $\boldsymbol{y} \in \mathcal{S}$. We have considered the outside mode $\mathcal{H} = \mathcal{R}, \mathcal{S} = \mathcal{R}^C$ in (18) and the inside mode $\mathcal{H} = \mathcal{R}^C, \mathcal{S} = \mathcal{R}$ in (20). For a general case of $\mathcal{H}, \mathcal{S}$, Terada and Shimodaira (2017) gave a formula of approximately unbiased p-value of selective inference as

$$\mathrm{SI}(\mathcal{H}|\mathcal{S}, \boldsymbol{y}) = \frac{\bar{\Phi}(\beta_0^{\mathcal{H}} - \beta_1^{\mathcal{H}})}{\bar{\Phi}(\beta_0^{\mathcal{S}} + \beta_0^{\mathcal{H}} - \beta_1^{\mathcal{H}})}, \tag{33}$$

where geometric quantities β_0, β_1 are defined for the regions $\mathcal{H}, \mathcal{S}$. We assumed that $\mathcal{H}$ and $\mathcal{S}^C$ are expressed as (29), and two surfaces $\partial\mathcal{H}, \partial\mathcal{S}$ are nearly parallel to each other with tangent planes differing only $O_p(n^{-1/2})$. The last assumption always holds for (18), because $\partial\mathcal{H} = \partial\mathcal{R}$ and $\partial\mathcal{S} = \partial\mathcal{R}^C$ are identical and of course parallel to each other.

Here we explain why we have considered the special case of $\mathcal{S} = \mathcal{H}^C$ for phylogenetic inference. First, we suppose that the selection event satisfies $\mathcal{S} \subset \mathcal{H}^C$, because a reasonable test would not reject H_0 unless $\boldsymbol{y} \in \mathcal{H}^C$. Note that $\boldsymbol{y} \in \mathcal{S} \subset \mathcal{H}^C$ implies $0 \le -\beta_0^{\mathcal{S}} \le \beta_0^{\mathcal{H}}$. Therefore, $\beta_0^{\mathcal{H}} + \beta_0^{\mathcal{S}} \ge 0$ leads to

$$\mathrm{SI}(\mathcal{H}|\mathcal{S}, \boldsymbol{y}) \ge \mathrm{SI}(\mathcal{H}|\boldsymbol{y}), \tag{34}$$

where $\mathrm{SI}(\mathcal{H}|\boldsymbol{y}) := \mathrm{SI}(\mathcal{H}|\mathcal{H}^C, \boldsymbol{y})$ is obtained from (33) by letting $\beta_0^{\mathcal{H}} + \beta_0^{\mathcal{S}} = 0$ for $\mathcal{S} = \mathcal{H}^C$. The p-value $\mathrm{SI}(\mathcal{H}|\mathcal{S}, \boldsymbol{y})$ becomes smaller as $\mathcal{S}$ grows, and $\mathcal{S} = \mathcal{H}^C$ gives the smallest p-value, leading to the most powerful selective test. Therefore the choice $\mathcal{S} = \mathcal{H}^C$ is preferable to any other choice of selection event satisfying $\mathcal{S} \subset \mathcal{H}^C$. This kind of property is mentioned in Fithian et al. (2014) as the monotonicity of selective error in the context of "data curving."

Let us see how these two p-values differ for the case of E2 by specifying $\mathcal{H} = \mathcal{R}^C_{\mathrm{E2}}$ and $\mathcal{S} = \mathcal{R}_{\mathrm{T1}}$. In this case, the two surfaces $\partial\mathcal{H}, \partial\mathcal{S}$ may not be very parallel to each other, thus violating the assumption of $\mathrm{SI}(\mathcal{H}|\mathcal{S}, \boldsymbol{y})$, so we only intend to show the potential difference between the two p-values. The geometric quantities are $\beta_0^{\mathcal{H}} = -\beta_0^{\mathrm{E2}} = 1.59$, $\beta_1^{\mathcal{H}} = -\beta_1^{\mathrm{E2}} = -0.12$, $\beta_0^{\mathcal{S}} = \beta_0^{\mathrm{T1}} = -0.41$; the p-values are calculated using more decimal places than shown. SI of E2 conditioned on selecting T1 is

$$\mathrm{SI}(\mathcal{H}|\mathcal{S}, \boldsymbol{y}) = \frac{\bar{\Phi}(1.59 + 0.12)}{\bar{\Phi}(-0.41 + 1.59 + 0.21)} = 0.448,$$

and it is very different from SI of E2 conditioned on selecting E2

$$\mathrm{SI}(\mathcal{H}|\boldsymbol{y}) = \frac{\bar{\Phi}(1.59 + 0.12)}{\bar{\Phi}(0.12)} = 0.097,$$

where $\mathrm{SI}'(\mathcal{R}^C_{\mathrm{E2}}|\boldsymbol{y}) = 1 - \mathrm{SI}(\mathcal{R}^C_{\mathrm{E2}}|\boldsymbol{y}) = 0.903$ is shown in **Table 2**. As you see, $\mathrm{SI}(\mathcal{H}|\boldsymbol{y})$ is easier to reject H_0 than $\mathrm{SI}(\mathcal{H}|\mathcal{S}, \boldsymbol{y})$.

6.7. Number of Regions for Phylogenetic Inference

The regions $\mathcal{R}_i$, $i = 1, \ldots, K_{\mathrm{all}}$ correspond to trees or edges. In inside and outside modes, the number of total regions is $K_{\mathrm{all}} = 105$ for trees and $K_{\mathrm{all}} = 25$ for edges when the number of taxa is $N = 6$. For general $N \ge 3$, they grow rapidly as $K_{\mathrm{all}} = (2N-5)!/(2^{N-3}(N-3)!)$ for trees and $K_{\mathrm{all}} = 2^{N-1} - (N+1)$ for edges. Next consider the number of selected regions K_{select}. In inside mode, regions with $\boldsymbol{y} \in \mathcal{R}_i$ are selected, and the number is counted as $K_{\mathrm{select}} = 1$ for trees and $K_{\mathrm{select}} = N - 3 = 3$ for edges. In outside mode, regions with $\boldsymbol{y} \notin \mathcal{R}_i$ are selected, and thus the number is K_{all} minus that for inside mode; $K_{\mathrm{select}} = K_{\mathrm{all}} - 1 = 104$ for trees and $K_{\mathrm{select}} = K_{\mathrm{all}} - (N-3) = 22$ for edges. Finally, consider the number of true null hypotheses, denoted as K_{true}. The null hypothesis holds true when $\boldsymbol{\mu} \notin \mathcal{R}_i$ in inside mode and $\boldsymbol{\mu} \in \mathcal{R}_i$ in outside mode, and thus K_{true} is the same as the number of regions with $\boldsymbol{y} \notin \mathcal{R}_i$ in inside mode and $\boldsymbol{y} \in \mathcal{R}_i$ in outside mode (These numbers do not depend on the value of $\boldsymbol{y}$ by ignoring the case of $\boldsymbol{y} \in \partial\mathcal{R}_i$). Therefore, $K_{\mathrm{true}} = K_{\mathrm{all}} - K_{\mathrm{select}}$ for both cases.

6.8. Selective Inference of Lasso Regression

Selective inference is considered for the variable selection of regression analysis. Here, we deal with prostate cancer data (Stamey et al., 1989) in which we predict the level of prostate-specific antigen (PSA) from clinical measures. The dataset is available in the R package *ElemStatLearn* (Halvorsen, 2015). We consider a linear model to the log of PSA (`lpsa`), with 8 predictors such as the log prostate weight (`lweight`), `age`, and so on. All the variables are standardized to have zero mean and unit variance.

The goal is to provide the valid selective inference for the partial regression coefficients of the selected variables by lasso (Tibshirani, 1996). Let n and p be the number of observations and the number of predictors. $\hat{\boldsymbol{M}}$ is the set of selected variables, and $\hat{\boldsymbol{s}}$ represents the signs of the selected regression coefficients. We suppose that regression responses are distributed as $\boldsymbol{Y} \sim N(\boldsymbol{\mu}, \tau^2\boldsymbol{I}_n)$ where $\boldsymbol{\mu} \in \mathbb{R}^n$ and $\tau > 0$. Let e_i be the ith residual. Resampling the scaled residuals σe_i $(i = 1, \ldots, n)$ with several values of scale σ^2, we can apply the multiscale bootstrap method described in section 4

for the selective inference in the regression problem. Here, we note that the target of the inference is the true partial regression coefficients:

$$\boldsymbol{\beta} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{\mu},$$

where $\boldsymbol{X} \in \mathbb{R}^{n\times p}$ is the design matrix. We compute four types of intervals with confidence level $1-\alpha = 0.95$ for selected variable j. $[L_j^{\text{ordinary}}, U_j^{\text{ordinary}}]$ is the non-selective confidence interval obtained via t-distribution. $[L_j^{\text{model}}, U_j^{\text{model}}]$ is the selective confidence interval under the selected model proposed by Lee et al. (2016) and Tibshirani et al. (2016), which is computed by `fixedLassoInf` with `type="full"` in R package *selectiveInference* (Tibshirani et al., 2017). By extending the method of $[L_j^{\text{model}}, U_j^{\text{model}}]$, we also computed $[L_j^{\text{variable}}, U_j^{\text{variable}}]$, which is the selective confidence interval under the selection event that variable j is selected. These three confidence intervals are exact, in the sense that

$$P\left(\beta_j \in [L_j^{\text{ordinary}}, U_j^{\text{ordinary}}]\right) = 1-\alpha,$$
$$P\left(\beta_j \in [L_j^{\text{model}}, U_j^{\text{model}}] \mid \hat{\boldsymbol{M}}, \hat{\boldsymbol{s}}\right) = 1-\alpha,$$
$$P\left(\beta_j \in [L_j^{\text{variable}}, U_j^{\text{variable}}] \mid j \in \hat{\boldsymbol{M}}, \hat{s}_j\right) = 1-\alpha.$$

Note that the selection event of variable j, i.e., $\{j \in \hat{\boldsymbol{M}}, \hat{s}_j\}$ can be represented as a union of polyhedra on $\mathbb{R}^n$, and thus, according to the polyhedral lemma (Lee et al., 2016; Tibshirani et al., 2016), we can compute a valid confidence interval $[L_j^{\text{variable}}, U_j^{\text{variable}}]$. However, this computation is prohibitive for $p > 10$, because all the possible combinations of models with variable j are considered. Therefore, we compute its approximation $[\hat{L}_j^{\text{variable}}, \hat{U}_j^{\text{variable}}]$ by the multiscale bootstrap method of section 4 with much faster computation even for larger p.

We set $\lambda = 10$ as the penalty parameter of lasso, and the following model and signs were selected:

$$\hat{\boldsymbol{M}} = \{\texttt{lcavol}, \texttt{lweight}, \texttt{lbph}, \texttt{svi}, \texttt{pgg45}\},$$
$$\hat{\boldsymbol{s}} = (+,+,+,+,+).$$

The confidence intervals are shown in **Figure 1**. For adjusting the selection bias, the three confidence intervals of selective inference are longer than the ordinary confidence interval. Comparing $[L_j^{\text{model}}, U_j^{\text{model}}]$ and $[L_j^{\text{variable}}, U_j^{\text{variable}}]$, the latter is shorter, and would be preferable. This is because the selection event of the latter is less restrictive as $\{\hat{\boldsymbol{M}}, \hat{\boldsymbol{s}}\} \subseteq \{j \in \hat{\boldsymbol{M}}, \hat{s}_j\}$; see section 6.6 for the reason why larger selection event is better. Finally, we verify that $[\hat{L}_j^{\text{variable}}, \hat{U}_j^{\text{variable}}]$ approximates $[L_j^{\text{variable}}, U_j^{\text{variable}}]$ very well.

AUTHOR CONTRIBUTIONS

HS and YT developed the theory of selective inference. HS programmed the multiscale bootstrap software and conducted the phylogenetic analysis. YT conducted the lasso analysis. HS wrote the manuscript. All authors have approved the final version of the manuscript.

FUNDING

This research was supported in part by JSPS KAKENHI Grant (16H02789 to HS, 16K16024 to YT).

ACKNOWLEDGMENTS

The authors appreciate the feedback from the audience of seminar talk of HS at Department of Statistics, Stanford University. The authors are grateful to Masami Hasegawa for his insightful comments on phylogenetic analysis of mammal species.

REFERENCES

Adachi, J., and Hasegawa, M. (1996). Model of amino acid substitution in proteins encoded by mitochondrial DNA. *J. Mol. Evol.* 42, 459–468. doi: 10.1007/BF02498640

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Trans. Autom. Cont.* 19, 716–723. doi: 10.1109/TAC.1974.1100705

Amari, S.-I., and Nagaoka, H. (2007). *Methods of Information Geometry, Translations of Mathematical Monographs*, Vol. 191. Providence, RI: American Mathematical Society.

Burnham, K. P., and Anderson, D. R. (2002). *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach, 2nd Edn*. New York, NY: Springer-Verlag.

Cox, D. R. (1962). Further results on tests of separate families of hypotheses. *J. R. Stat. Soc. Ser. B (Methodol.).* 24, 406–424. doi: 10.1111/j.2517-6161.1962.tb00468.x

Efron, B. (1979). Bootstrap methods: another look at the jackknife. *Ann. Stat.* 7, 1–26. doi: 10.1214/aos/1176344552

Efron, B. (1984). Comparing non-nested linear models. *J. Am. Stat. Assoc.* 79, 791–803. doi: 10.1080/01621459.1984.10477096

Efron, B. (1985). Bootstrap confidence intervals for a class of parametric problems. *Biometrika* 72, 45–58. doi: 10.1093/biomet/72.1.45

Efron, B., Halloran, E., and Holmes, S. (1996). Bootstrap confidence levels for phylogenetic trees. *Proc. Natl. Acad. Sci. U.S.A.* 93, 13429–13434. doi: 10.1073/pnas.93.23.13429

Efron, B., and Tibshirani, R. (1998). The problem of regions. *Ann. Sta.* 26, 1687–1718. doi: 10.1214/aos/1024691353

Felsenstein, J. (1981). Evolutionary trees from DNA sequences: a maximum likelihood approach. *J. Mol. Evol.* 17, 368–376. doi: 10.1007/BF01734359

Felsenstein, J. (1985). Confidence limits on phylogenies: an approach using the bootstrap. *Evolution* 39, 783–791. doi: 10.1111/j.1558-5646.1985.tb00420.x

Fithian, W., Sun, D., and Taylor, J. (2014). Optimal inference after model selection. *arXiv:1410.2597*.

Graur, D., Duret, L., and Gouy, M. (1996). Phylogenetic position of the order lagomorpha (rabbits, hares and allies). *Nature* 379:333. doi: 10.1038/379333a0

Halanych, K. M. (1998). Lagomorphs misplaced by more characters and fewer taxa. *Syst. Biol.* 47, 138–146. doi: 10.1080/106351598261085

Halvorsen, K. (2015). *ElemStatLearn: data sets, functions and examples from the book: "the elements of statistical learning, data mining, inference, and prediction" by trevor hastie, robert tibshirani and jerome friedman*. R package. Available online at: https://CRAN.R-project.org/package=ElemStatLearn

Kishino, H., and Hasegawa, M. (1989). Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order in Hominoidea. *J. Mol. Evol.* 29, 170–179. doi: 10.1007/BF02100115

Kishino, H., Miyata, T., and Hasegawa, M. (1990). Maximum likelihood inference of protein phylogeny and the origin of chloroplasts. *J. Mol. Evol.* 30, 151–160. doi: 10.1007/BF02109483

Konishi, S., and Kitagawa, G. (2008). *Information Criteria and Statistical Modeling*. New York, NY: Springer Science & Business Media. doi: 10.1007/978-0-387-71887-3

Lee, J. D., Sun, D. L., Sun, Y., and Taylor, J. E. (2016). Exact post-selection inference, with application to the lasso. *Ann. Stat.* 44, 907–927. doi: 10.1214/15-AOS1371

Linhart, H. (1988). A test whether two AIC's differ significantly. *South Afr. Stat. J.* 22, 153–161.

Novacek, M. J. (1992). Mammalian phytogeny: shaking the tree. *Nature* 356, 121–125. doi: 10.1038/356121a0

Posada, D., and Buckley, T. R. (2004). Model selection and model averaging in phylogenetics: advantages of Akaike information criterion and bayesian approaches over likelihood ratio tests. *Syst. Biol.* 53, 793–808. doi: 10.1080/10635150490522304

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychol. Bull.* 86, 638–641. doi: 10.1037/0033-2909.86.3.638

Schennach, S. M., and Wilhelm, D. (2017). A simple parametric model selection test. *J. Am. Stat. Assoc.* 112, 1663–1674. doi: 10.1080/01621459.2016.1224716

Shimodaira, H. (1997). Assessing the error probability of the model selection test. *Ann. Inst. Stat. Math.* 49, 395–410. doi: 10.1023/A:1003140609666

Shimodaira, H. (1998). An application of multiple comparison techniques to model selection. *Ann. Inst. Stat. Math.* 50, 1–13. doi: 10.1023/A:1003483128844

Shimodaira, H. (2000). Improving predictive inference under covariate shift by weighting the log-likelihood function. *J. Stat. Plan. Inference* 90, 227–244. doi: 10.1016/S0378-3758(00)00115-4

Shimodaira, H. (2001). Multiple comparisons of log-likelihoods and combining nonnested models with applications to phylogenetic tree selection. *Commun. Stat. Theory Methods* 30, 1751–1772. doi: 10.1081/STA-100105696

Shimodaira, H. (2002). An approximately unbiased test of phylogenetic tree selection. *Syst. Biol.* 51, 492–508. doi: 10.1080/10635150290069913

Shimodaira, H. (2004). Approximately unbiased tests of regions using multistep-multiscale bootstrap resampling. *Ann. Stat.* 32, 2616–2641. doi: 10.1214/009053604000000823

Shimodaira, H. (2008). Testing regions with nonsmooth boundaries via multiscale bootstrap. *J. Stat. Plan. Inference* 138, 1227–1241. doi: 10.1016/j.jspi.2007.04.001

Shimodaira, H. (2019). *Scaleboot: Approximately Unbiased p-Values via Multiscale Bootstrap*. R package version 1.0-0. Available online at: https://CRAN.R-project.org/package=scaleboot

Shimodaira, H., and Hasegawa, M. (1999). Multiple comparisons of log-likelihoods with applications to phylogenetic inference. *Mol. Biol. Evol.* 16, 1114–1116. doi: 10.1093/oxfordjournals.molbev.a026201

Shimodaira, H., and Hasegawa, M. (2001). CONSEL: for assessing the confidence of phylogenetic tree selection. *Bioinformatics* 17, 1246–1247. doi: 10.1093/bioinformatics/17.12.1246

Shimodaira, H., and Hasegawa, M. (2005). "Assessing the uncertainty in phylogenetic inference," in *Statistical Methods in Molecular Evolution*, Statistics for Biology and Health, ed R. Nielsen (New York, NY: Springer), 463–493. doi: 10.1007/0-387-27733-1_17

Shimodaira, H., and Maeda, H. (2018). An information criterion for model selection with missing data via complete-data divergence. *Ann. Inst. Stat. Math.* 70, 421–438. doi: 10.1007/s10463-016-0592-7

Stamey, T., Kabalin, J., Johnstone, I., Freiha, F., Redwine, E., and Yang, N. (1989). Prostate specific antigen in the diagnosis and treatment of adenocarcinoma of the prostate II. Radical prostatectomy treted patients. *J. Urol.* 16, 1076–1083. doi: 10.1016/S0022-5347(17)41175-X

Suzuki, R., and Shimodaira, H. (2006). Pvclust: an R package for assessing the uncertainty in hierarchical clustering. *Bioinformatics* 22, 1540–1542. doi: 10.1093/bioinformatics/btl117

Taylor, J., and Tibshirani, R. (2015). Statistical learning and selective inference. *Proc. Natl. Acad. Sci. U.S.A.* 112, 7629–7634. doi: 10.1073/pnas.1507583112

Terada, Y., and Shimodaira, H. (2017). Selective inference for the problem of regions via multiscale bootstrap. *arXiv:1711.00949*.

Tian, X., and Taylor, J. (2018). Selective inference with a randomized response. *Ann. Stat.* 46, 679–710. doi: 10.1214/17-AOS1564

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B (Methodol.).* 58, 267–288. doi: 10.1111/j.2517-6161.1996.tb02080.x

Tibshirani, R., Taylor, J., Lockhart, R., and Tibshirani, R. (2016). Exact post-selection inference for sequential regression procedures. *J. Amer. Stat. Assoc.* 111, 600–620. doi: 10.1080/01621459.2015.1108848

Tibshirani, R., Tibshirani, R., Taylor, J., Loftus, J., and Reid, S. (2017). *SelectiveInference: Tools for Post-Selection Inference*. R package version 1.2.4. Available online at: https://CRAN.R-project.org/package=selectiveInference

Vuong, Q. H. (1989). Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica* 57, 307–333. doi: 10.2307/1912557

Yang, Z. (1996). Among-site rate variation and its impact on phylogenetic analyses. *Trends Ecol. Evol.* 11, 367–372. doi: 10.1016/0169-5347(96)10041-0

Yang, Z. (1997). PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput. Appl. Biosci.* 13, 555–556. doi: 10.1093/bioinformatics/13.5.555

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.