論文 / 著書情報
Article / Book Information

題目(和文)	
Title(English)	Speedy double bootstrap method and its application for assessing the statistical reliability of estimated phylogenetic trees
著者(和文)	任愛珍
Author(English)	aizhen ren
出典(和文)	学位:博士(理学), 学位授与機関:東京工業大学, 報告番号:甲第9266号, 授与年月日:2013年9月25日, 学位の種別:課程博士, 審査員:渡辺 治,秋山 泰,間瀬 茂,三好 直人,杉山 将
Citation(English)	Degree:Doctor (Science), Conferring organization: Tokyo Institute of Technology, Report number:甲第9266号, Conferred date:2013/9/25, Degree Type:Course doctor, Examiner:,,,,
学位種別(和文)	博士論文
Type(English)	Doctoral Thesis

Speedy double bootstrap method and its application for assessing the statistical reliability of estimated phylogenetic trees

A dissertation presented

by

Aizhen Ren

to

Department of Mathematical and Computing Sciences

for the degree of

Doctor of Science

Tokyo Institute of Technology

September 2013

Thesis advisor

Osamu Watanabe and Yutaka Akiyama

Author

Aizhen Ren

Abstract

The bootstrap method (BP-method) is a well-known computational approach for assessing phylogenetic trees and, more generally, the reliability of statistical models. However, BP-method is known to be biased under certain circumstances, calling into question the accuracy of the method. Several advanced bootstrap methods have been developed to achieve higher accuracy, one of which is the double bootstrap approach (DBP-method), but the computational burden of this method has precluded its application to practical problems of phylogenetic tree selection. In practical model selection problems where the maximum-likelihood criterion is used, we address this issue by proposing the "speedy double bootstrap method" (sDBP-method), which circumvents the second-tier resampling step of regular DBP-method. We also develop an implementation of the regular DBP-method for comparison with our "speedy" method. Then, we evaluate sDBP-method and DBP-method using biological data, as a consequence, the sDBP-method suffers no significant loss of accuracy compared with regular DBP-method, and is significantly faster. Via a series of simulations, we also compare the statistical properties and computational efficiency of sDBP-method with three other bootstrap methods (BP, DBP, and a multiscale bootstrap method (AU)). We find that the rejection probability of the third-order methods (sDBP, DBP, and AU) is similar, and that sDBP-method is apparently faster than DBP-method. Finally, we develop an R package that allows researchers to easily apply sDBP-method to the problem of phylogenetic tree selection.

Contents

Acknowledgments

Writing this doctoral work has been a wonderful and often overwhelming experience.

I must acknowledge my supervisor, Professor Osamu Watanabe. Throughout these years he has been encouraging me a lot to do this challenging research. I am very appreciative of his consideration during my progress toward this degree. Professor Osamu watanabe also is the leader of JSPS Global COE program "Computationism as a Foundation for the Sciences (CompView in short)", and he provided me an opportunity to work as a research assistant of CompView and I attended a lot of CompView seminars and others CompView events on a rugular basis. While I was a research assistant of CompView, I also learned a lot from Professor Osamu Watanabe, for example, the passion for doing research and being friendly to others. I also would like to express my appreciation to CompView, for financial support through my research works.

I owe my gratitute to my another supervisor Professor Yutaka Akiyama for his enthusiastic guidance and support. Ever since he taught me what an avoided crossing was, I have been stimulated and inspired by his constant flow of good ideas. I will always admire Professor Yutaka Akiyama's ability to cut through reams of statistics with a single visual explanation, and I very much appreciate him accepting me as a student and his generous guidance for my final paper. He has also known when (and how) to give me a little push forward when needed.

I would like to thank my former supervisor Professor Hidetoshi Shimodaira for giving me the opportunity to study in Tokyo Institute of Technology and guiding me to do research.

I am also grateful to Assistant Professor Takashi Ishida for his helpful, valuable advices and feedbacks, checking and correcting my papers.

Moreover, I would like to acknowledge the assistance of Kanako Ozeki and Yukie Omura, who were always cheerful in their handling of my various paperwork.

Throughout my four and a half years, I was supported for many semesters by the Rotary Yoneyama Memorial Foundation.

I would also like to thank Yuri Matsuzaki, Masahito Ohue and all other students and post-docs, both past and present, of Professor Akiyama's laboratory. They comprise a superb research group and offer many fascinating conversations on machine learning, although Akiyama lab is Bioinformatic lab.

In particular, I would like to thank Wuyun Gaowa, a PhD student in the Tokyo University of Foreign Studies. I admire the unfailing determination she displayed toward her research. It would also be remiss not to thank Summer A. Seay for our lunches together and interesting conversations in English.

Finally, I wish to express a collective thanks to all of my family, but especially to my husband and daughter, who would always cheer me up when I was depressed. I dedicate this thesis to them.

Dedicated to my husband Bai yin zhuang,
and my daughter Yilina,

Chapter 1

Introduction

1.1 Background

1.1.1 Model selection

The statistical inference of a data set is intended to determine the distribution from which the data are drawn. For a given data set, we always assume that a stochastic model or statistical model can express the distribution. The statistician R.A. Fisher defined the supposition of a model as a specification. A statistical model is a collection of probability distribution functions or probability density functions (collectively referred to as distributions for brevity). These can be divided into parametric models, non-parametric models, and semi-parametric models. A parametric model is a collection of distributions, each of which is indexed by a unique finite-dimensional parameter: $\mathcal{P} = \{P_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Theta\}$, where $\boldsymbol{\theta}$ is a parameter and $\Theta \subseteq \mathbb{R}^d$ is the feasible parameter space, which is a subset of d-dimensional Euclidean space.

Most statistical tests can be described in the form of a statistical model. For example, the Student's t-test for comparing the mean of two groups can be formulated as the determination of whether an estimated parameter in the model is different from 0. Parametric models sometimes use phylogenetic tree topologies. Using a method proposed in this thesis, we will assess the reliability of such topologies. Therefore, we focus on parametric models.

However, for a single set of observed data, there are several parametric models that can be considered. Thus, we must select a statistical model from a set of candidate models. This is known as the model selection problem, and is discussed in books by Linhart and Zucchinni (1986) [23] and Shimodaira (2004) [40].

Since Akaike (1974) [2] advocated the Akaike information criterion (AIC) for model selection, the minimum AIC estimate (MAICE), which minimizes AIC, has been widely used as a simple and practical estimate of the best model. Additional information criteria have been proposed, for example, Schwartz's (1978) [35] Bayesian information criterion (BIC) and Hannan and Quinn's (1979) [17] HQ criterion.

However, Shimodaira (1998) [41] stated that, by considering the sampling error of AIC, a set of good models can be constructed rather than a single model being chosen. This set is called a confidence set of models, and includes the minimum risk model, with an error rate below the specified significance level. The result for each model is described by its p-value, from which the confidence set is immediately obtained. A variant of Gupta's [15] subset selection procedure was devised, in which a standardized difference in AIC is calculated for every pair of models. Shimodaira (1998) [41] used multiple comparison techniques to calculate the reliability of model i,

and later proposed a similar approach, known as the SH-test [38], which used multiple comparison statistics discussed in Nagata and Yoshida (1997) [27] and an iterative bootstrap technique to calculate the reliability of phylogenetic trees. Let i index the topologies, and L_i be the maximum log-likelihood under the probabilistic model specified by the i-th phylogenetic tree topology. We have M candidate topologies, labeled $1, 2, \cdots, M$. The SH-test procedure can then be explained as follows:

Step 1 Calculate the test statistic $T_i = \max\{L_1 - L_i, \cdots, L_M - L_i\}$ for $i = 1, \cdots, M$.

Step 2 Generate N bootstrap replicates of vector $(L_1, \cdots, L_M)$. The replicates $\tilde{L}_{i\cdot b}, i = 1, \cdots, M, b = 1, \cdots, N$ are stored in an $M \times N$ array. For the resampling of $\tilde{L}_{i\cdot b}$, the RELL method and the normal approximation method of Kishino et al. (1990) [20] are computationally useful.

Step 3 Subtract the average of each row from the array entries. We now have the array $\tilde{R}_{i\cdot b} = \tilde{L}_{i\cdot b} - N^{-1}\sum_{a=1}^{N} \tilde{L}_{i\cdot a}$. This step is called "centering", and $\tilde{R}_{i\cdot b}$ is regarded as a replicate of L_i generated under the least favorable configuration (l.f.c.), i.e., $E(L_1) = \cdots = E(L_M)$.

Step 4 For each column $(b = 1, \cdots, N)$ of the array, calculate $\tilde{S}_{i\cdot b} = \max\{\tilde{R}_{1\cdot b} - \tilde{R}_{i\cdot b}, \cdots, \tilde{R}_{M\cdot b} - \tilde{R}_{i\cdot b}\}$ and replace the entries with these values. $\tilde{S}_{i\cdot b}$ is a replicate of T_i.

Step 5 For each row $(i = 1, \cdots, M)$ of the array, count the number of entries that exceed T_i, then calculate the p-value according to $P_i =$ (number of $\{\tilde{S}_{i\cdot b} > T_i\})/N$.

The above procedure is well-known in the field of phylogenetic tree selection, because it corrects the selection bias that occurs when T_i is assumed to be normally distributed. In fact, the asymptotic distribution of T_i is not the normal distribution (Barlow et al. 1972 [5]), and is known to be very complicated. In Step 4 above, the definition of $\tilde{S}_{i\cdot b}$ takes into account the possibility that any of $j = 1, \cdots, M$ could be the maximum-likelihood topology. However, the SH-test is very conservative, giving a bigger confidence set because Step 3 assumes the least favorable configuration $E(L_1) = \cdots = E(L_M)$.

On the other hand, prior to the development of the SH-test, Felsenstein (1985) [12] used the bootstrap method (BP-method) to propose the regular bootstrap probability (BP-probability), which can be expressed as follows. For model i,

Step 1 Generate N bootstrap replicates of vector $(L_1, \cdots, L_M)$. The replicates $\tilde{L}_{i\cdot b}, i = 1, \cdots, M, b = 1, \cdots, N$ are stored in an $M \times N$ array.

Step 2 Calculate the bootstrap probability BP_i as follows:

$$BP_i = (\text{number } of\{i == \text{argmax}\{\tilde{L}_{1\cdot b}, \cdots, \tilde{L}_{M\cdot b}\}\})/N \tag{1.1}$$

This procedure for calculating BP_i is very simple, even in complicated scenarios. Therefore, as a consequence, methods such as Felsenstein's BP-method have become the principal technique for complicated practical applications. Unfortunately, the BP-method is computationally expensive in most applications. A number of methods, e.g., the RELL approximation (Kishino et al. 1990 [20]), have thus been developed to reduce the computational complexity. Moreover, the regular bootstrap probability

(BP-probability) is known to be biased under certain circumstance. Because of this limitation of BP-probability, this thesis focuses on improving the accuracy of it. Thus in the next subsection, we briefly explain the concept of phylogenetic trees and their inference problems.

Our work is aimed not only at phylogenetic trees, but also practical model selection problems where the maximum-likelihood criterion is used. The problem of statistical model selection has been widely researched in various fields. This has included the assessment of the statistical reliability of phylogenetic trees (Felsenstein 1983 [14] and Felsenstein 1985 [12]), which we will discuss in this thesis, assessing uncertainty in hierarchical cluster analysis (Suzuki and Shimodaira 2006 [43]), where the BP-method was used, determining the number of hidden units for multilayer neural networks, based on a generalized AIC (Murata et al. 1994 [25]), and assessing the statistical reliability of Linear Non-Gaussian Acyclic Models [36]) (Komatsu et al. 2010 [21]), which again used the BP-method.

1.1.2 Phylogenetic tree selection problem

The analytical methods used in the field of molecular phylogenetics are important basic tools for reconstructing the evolutionary history (phylogenetic relationships) of molecules and organisms. Molecular phylogenetic methods are primarily used in the context of biological systematics, but they find applications in a wide variety of fields, such as community ecology (Webb et al. 2002 [45]), biogeography (Wiley 1981 [46]), and proteomics, including the inference of protein–protein interactions (Pazos and Valencia 2001 [28]). Many methods of phylogenetic reconstruction have been

developed and are in regular use (Felsenstein 2004 [13]). However, those based on maximum-likelihood estimation have proved most effective for reconstructing phylogenies using molecular sequence data (e.g., DNA, protein). Early work on this application of maximum-likelihood was conducted by Felsenstein (1981) [11], whose approach involved computing the maximum-likelihood value for many topologies, and selecting that with the highest likelihood (the maximum-likelihood (ML) tree) as the most probable candidate for the true topology.

1.2 Problem statement

As mentioned before, the regular BP-probability is only accurate to the leading order (first order), and it is known that under some circumstances it is a biased estimator of the exact p-value (Hillis and Bull 1993 [19]; Standerson and Wojciechowski 2000 [34]). Thus, some advanced bootstrap techniques, such as the multiscale bootstrap method (AU-method) (Shimodaira 2002 [37]) and the double bootstrap method (DBP-method) (Hall 1992 [16]; Efron and Tibshirani 1998 [9]), have been proposed to achieve third-order accuracy.

However, the prohibitive computational burden imposed by the otherwise promising DBP-method has curtailed its use in most practical contexts. In practical problems such as assessing the reliability of phylogenetic trees, implementing the DBP-method and overcoming its computational load is a problem. And another method is multiscale bootstrap method (it is known for AU). It also has third order accuracy and has less complexity than double bootstrap method. This method has been shown successful in many real world applications. However, the problem of multiscale boot-

strap method is that the complexity is still large. Therefor, we devote some effort to solving the time-consumming of DBP-method. Indeed, we use a simple method that circumvents second-tier resampling in the DBP-method, thus reducing its computational complexity. We call this method is the "speedy double bootstrap method" (sDBP-method). The basic idea of speedy double bootstrap method appears in the technical report Efron and Tibshirani (1997) [8]. Becaue, in phylogenetic case, when use their basic idea to computing the p-value, have to solve the projection and signed distance problem, in their technical report does not metioned about these problems. Thus we proposed speedy double bootstrap method for phylogenetic tree case, but our method not limited to phylogenetic trees case.

To date, there has been no comparison of the statistical properties and computational efficiency of the sDBP-method (acturally in this case sDBP-method is the basic idear in technical report Efron and Tibshirani (1997)[8]) and the DBP-method, AU-method and BP-method. This has motivated us to conduct simulations to evaluate these four bootstrap methods.

Finally, in the phylogenetic tree selection problem, we develop an R package (available via CRAN, the official R archive) to implement the sDBP-method proposed this thesis Chapter 2. This package will enable scientists and engineers, as well as statisticians, to easily apply our sDBP-method.

1.3 Contributions

The contributions of this thesis naturally fall into four parts, which are mainly based on the content of a prior publication Ren et al. (2013) [32] and Ren et al.

(2013)[33]. Each contribution is relatively dependent. The contributions can be summarized as follows.

- we presented the sDBP-method for assessing the confidence levels of phylogenetic trees, and also developed a DBP-method procedure for comparison. We evaluate sDBP and DBP-method using biological data. According to our work, the sDBP-method provides improvements in accuracy over the BP-method, and substantial improvements in speed over the DBP-method. Our calculations also showed that the application of the sDBP-method is not confined to general tree selection problems; rather, it is appropriate for general model selection problems using the maximum-likelihood criterion. This result is described in Chapter 2 and Chapter 3. This is the primary contribution of this thesis. These results is also described in detail in Ren et al. (2013) [32].

- We present a novel implementation of the DBP-method for phylogenetic tree selection problems. As a consequence, this enables the DBP-method to be practically applied in the context of molecular phylogenetics for the first time. This result is described in Chapter 2. This result is also described in detail in Ren et al. (2013) [32].

- To evaluate the statistical properties and computational efficiency of sDBP-method, we perform simulation-based tests using simple examples to illustrate the differences between the p-values generated by sDBP, DBP,AU and BP-method. We also compare the computational speed of these methods. The numerical simulations help us to understand the difference in rejection probabilities given by the four bootstrap methods. This result is described in Chapter

4.

- Finally, in the phylogenetic tree selection problem, we develop an R package, named SDBP to implement the sDBP-method, allowing researchers to easily apply sDBP-method. This result is described in Chapter 5 and Ren et al. (2013) [33].

1.4 Outline of the thesis

This thesis consists of four main chapters that are relatively dependent, as well as our final conclusions.

In Chapter 2, we state our primary contribution, namely speedy double bootstrap method for assessing the reliability of phylogenetic trees. After an introduction in Section 2.1, we explain the DBP-method in Section 2.2. In Section 2.3, we use the PAVA (Pool Adjacent Violators Algorithm)[4] to calculate the log-likelihood vector's projection to the boundary of each hypothesis. And we also proposed signed distance using log-likelihood vector. Through these endeavors, we propose the sDBP-method for assessing the statistical reliability of phylogenetic trees. For comparison, in Section 2.4 we present the DBP-method for assessing the statistical reliability of phylogenetic trees. Finally, in Section 2.5, we discuss the advantages and disadvantages of the proposed methods.

In Chapter 3, we evaluate sDBP, DBP, AU and BP-method using biological data. In Section 3.1, we analyze the mammalian mitochondrial protein sequences of 6 species. In Section 3.2, we analyze the mammalian mitochondrial amino acid se-

quences and 12S and 16S rRNA genes of 20 species. For each section we use our experiment results to compare the sDBP and DBP-values using the paired t-test. We also investigate the time taken to calculate a p-value for a single tree, we compare four methods: DBP, sDBP,AU and BP-method. We conducted two separate sets of analyses in both experiments. In Section 3.3, we compare the experimental results from these species and give a detail conclutions of comparing result. In Section 3.4, we discusse the our proposed signed distance and Euclidian signed distance.

In Chapter 4, our focus switches to simulations and data analysis. We use simulations based on artificial data to investigate the relative statistical performance and computational efficiency of the four different bootstrap methods (sDBP, DBP, AU and BP). In Section 4.1, we design and perform a simulation from the Normal Model, and use our simulation results to compare the rejection probabilities for sDBP, DBP, AU and BP-method. Graphically, the rejection probability of sDBP, DBP and AU-method are similar, whereas that of BP-method is noticeably different. Then, in Section 4.2, we apply statistical tests that is multiple comparison to compare the difference in rejection probabilities between the first-order accurate BP and the third-order accurate methods. Finally, in Section 4.3, we compare the computational speed of the four methods.

In Chapter 5, we develop an easy-to-use R package, named SDBP, for assessing the reliability of estimated phylogenetic trees based on sDBP-method. In Section 5.1, we explain the implementation of our package, then in Section 5.2, we describe its usage when applied to the mammalian mitochondrial acid sequences and 12S rRNA and 16S rRNA genes for 20 species.

Finally, in Chapter 6, we summarize and discuss the results from Chapters 2 to 5 and provide an outlook on possible future directions of research based on the methods and results presented here.

Chapter 2

Speedy double bootstrap method for assessing reliability of phylogenetic trees

In this chapter, we state our primary contribution, namely speedy double bootstrap method for assessing reliability of phylogenetic trees. After introduction in Section 2.1, we brief review about the DBP-method in Efron and Tibshirani (1998) [9]. Then in Section 2.3 we propose the procedure of sDBP-method in assessing the statistical reliability of phylogenetic trees, for comparison we also propose the procedure of DBP-method in Section 2.4. Finally, in Section 2.5, we discuss the advantages and disadvantages of the proposed methods.

2.1 Introduction

As mentioned in Subsection 1.1.2, early work on phylogenetic trees selection of maximum likelihood was conducted by Felsenstein (1981) [11], whose approach involved computing the maximum likelihood value for many topologies, and selecting the topology with the highest likelihood (the maximum likelihood (ML) tree) as the most probable candidate for the true topology.

It must be noted that the maximum likelihood values are dependent on the particular characteristics of a random variable: the molecular sequences that constitute the underlying data for phylogeny reconstruction. Thus, some analysis of the statistical reliability of the estimated ML tree or multiple alternative trees should be undertaken. Statistical hypothesis testing is commonly used for this purpose, and the 'bootstrapping' technique is a well-known computational method for calculating reliability when a simple mathematical formula is difficult to derive. Bootstrapping is a resampling method that approximates a random sample by creating a bootstrap sample, generated by random sampling with replacement from the original single data set. In the context of phylogenetic tree selection, Felsenstein (1985) [12] proposed the use of bootstrapping to place confidence intervals on phylogenies. He defined the p-value of a tree according to a frequency called the bootstrap probability (BP-probability); the proportion of bootstrap pseudoreplicates of the original data set in which the tree is found to be optimal. However, it has been shown to be first order accurate (Efron and Tibshirani 1998 [9]). And the double bootstrap (Hall 1992 [16]; Efron and Tibshirani 1998 [9]) has been shown to be third order accurate and may hold great potential as a measure of phylogenetic tree support. However, the method imposes

huge computation burdens and has yet to be applied in the context of molecular phylogenetics. To overcome this computational difficulty we propose a 'speedy' double bootstrap method to compute the reliability of phylogenetic trees. For comparison, we also developed a procedure to implement the regular double bootstrap (Hall 1992 [16]; Efron and Tibshirani 1998 [9]).

2.2 The double bootstrap method

In this study, homologous sites of aligned molecular sequence data are regarded as the units of sampling, and we use DNA data as the example for the following methodological descriptions. Suppose we have m homologous sequences, each with n nucleotide sites. These data can be represented as a $m \times n$ matrix $\mathbf{X} = (x_{jh}\} = \{\mathbf{x}_1, \cdots, \mathbf{x}_n)$, where $\mathbf{x}_h$ is the value of the h-th site and x_{jh} is one of the four deoxyribonucleotides (T, C, A, or G).

$$\begin{array}{rcccccc} \textit{Species } 1 & : & x_{11} & x_{12} & \cdots & x_{1n} \\ \textit{Species } 2 & : & x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & & & & \vdots & \\ \textit{Species } m & : & x_{m1} & x_{m2} & \cdots & x_{mn} \end{array} \tag{2.1}$$

The log-likelihood can be expressed as

$$l(\boldsymbol{\theta}; \mathbf{X}) = \sum_{h=1}^{n} logf(\mathbf{x}_h; \boldsymbol{\theta}) \tag{2.2}$$

where $f(\mathbf{x}_h; \boldsymbol{\theta}) = f(x_{1h}, x_{2h}, \cdots, x_{mh}; \boldsymbol{\theta})$ is the probability that at a particular homologous site, species 1 has base x_{1h}, species 2 has x_{2h} and species m has x_{mh}. The vector $\boldsymbol{\theta}$ denotes unknown parameters such as the edge lengths (branch lengths) of a tree, and the base substitution rates along these branches. Here we assume that the base substitution rates have already been estimated, so $\boldsymbol{\theta}$ denotes only the unknown edge lengths. For a given tree topology, $\boldsymbol{\theta}$ is estimated by maximizing the log-likelihood, and the maximum log-likelihood of any tree topology i is given by

$$l_i(\hat{\boldsymbol{\theta}}_i(\mathbf{X}); \mathbf{X}) = \sum_{h=1}^{n} logf_i(\mathbf{x}_h; \hat{\boldsymbol{\theta}}_i) \tag{2.3}$$

The topology with the highest value of $l_i(\hat{\boldsymbol{\theta}}_i(\mathbf{X}); \mathbf{X}), i = 1, \cdots, K$ is the maximum likelihood phylogenetic tree (T_{ML}) and for data set $\mathbf{X}$, and is thus the most likely candidate for the best topology described later in this Section. To define null hypotheses for performing model comparisons, we must consider the true distribution (it is unknown) for a random variable $\mathbf{x}$ can be expressed as

$$q(\mathbf{x}) \tag{2.4}$$

And the expectation of $l_i(\hat{\boldsymbol{\theta}}_i(\mathbf{X}); \mathbf{X}), i = 1, \cdots, K$ with respect to

$$(\mathbf{x_1}, \cdots, \mathbf{x_n}) \overset{i.i.d.}{\sim} q(\cdot) \tag{2.5}$$

can be expressed as

$$\mu_i \quad = \quad E_q[l_i(\hat{\boldsymbol{\theta}}_i(\mathbf{X}); \mathbf{X})] \tag{2.6}$$

Through equation (2.6), we can say that the μ_i is unknown quantity, because it include unknown distribution $q(\mathbf{x})$. And the best topology we denote k^* can be defined as

$$k^* := \text{argmax}_{i \in \{1, \cdots, K\}} \text{E}_q(l_i(\hat{\boldsymbol{\theta}}_i(\mathbf{X}); \mathbf{X})) \tag{2.7}$$

Therefor, if we assume that tree T_1 is the best topology, the null and alternative hypotheses will be

$$H_1 : \mu_1 = max_{i=1,\cdots,K} \ \mu_i \quad vs. \quad H_1^A : \ \text{others} \tag{2.8}$$

and we must continue performing these comparisons as many times as is necessary, assuming in turn that tree $T_i, i = 2, \cdots, K$ is the best topology. Note that the null hypothesis H_1 involves multiple comparisons: as can be seen from (2.8), the null contains $k - 1$ hypotheses such that

$$H_{1j} : \mu_1 \geq \mu_j, j = 2, \cdots, K \tag{2.9}$$

The null hypothesis H_1 is a polyhedral convex cone and ∂H_1, which is boundary of H_1 is nonsmooth at the vertex as well as on the faces of dimensions less than $K-1$. Shimodaira and Hasegawa (1999) [38] proposed a multiple comparisons procedure (the SH-test) to test H_1, but this was shown to be overly conservative and a different method was designed (the AU test), which uses a multiscale bootstrap technique to obtain third-order accurate p-values for testing the null hypothesis. Other authors (Hall 1992 [16]; Efron and Tibshirani 1998 [9]) had previously developed a double bootstrap method that was also able to provide third-order accurate p-values, but due to high computational requirements this method has not been adopted for phylogenetic applications.

At this juncture it is necessary to briefly review the double bootstrap method. The third-order accurate p-values was first proposed by Efron (1985) [7] for the multivariate normal model, which can be represented as

$$\mathbf{Y} \overset{i.i.d.}{\sim} N_t(\boldsymbol{\eta}, I_t) \tag{2.10}$$

This normal model is a simplification of reality. Let $\mathcal{H} \subset \mathbb{R}^t$ be an arbitrarily-shaped region with smooth boundaries denoted by $\partial\mathcal{H}$. We want to calculate a p-value $p(\boldsymbol{y})$ for testing the null hypothesis $\boldsymbol{\eta} \in \mathcal{H}$. According to Efron (1985) [7], when the true parameter $\boldsymbol{\eta}$ is on the boundary surface $\partial\mathcal{H}$, the third-order accurate p-value can be expressed as

$$p(\boldsymbol{y}) = 1 - \Phi(d - c) \tag{2.11}$$

where d is the signed distance from $\boldsymbol{y}$ to $\hat{\boldsymbol{\eta}}(\boldsymbol{y})$, with a positive or negative sign when

$\boldsymbol{y}$ is, respectively, outside or inside $\mathcal{H}$. The point $\hat{\boldsymbol{\eta}}(\boldsymbol{y})$ is the closest point to $\boldsymbol{y}$ (in Euclidean distance) on the surface $\partial\mathcal{H}$, and c in formula (2.11) is a quantity related to the curvature of $\partial\mathcal{H}$ at point $\hat{\boldsymbol{\eta}}(y)$. The double bootstrap method of Hall (1992) [16] and Efron and Tibshirani (1998) [9] begins with a first tier of bootstrap resampling from the multivariate normal model with distribution

$$\mathbf{Y}^* \overset{i.i.d.}{\sim} N_t(\hat{\boldsymbol{\eta}}(\boldsymbol{y}), I_t) \tag{2.12}$$

A second tier of resampling is carried out for each of these vectors $\mathbf{Y}^*$, as well as for $\mathbf{Y}$, with the following distributions

$$\begin{aligned} \mathbf{Y}^{**} &\overset{i.i.d.}{\sim} N_t(\mathbf{Y}^*, I_t) \\ \mathbf{Y}^{**} &\overset{i.i.d.}{\sim} N_t(\mathbf{Y}, I_t) \end{aligned} \tag{2.13}$$

The second tier quantities in each case are as follows

$$\tilde{p}^* = P(\boldsymbol{y}^{**} \in \mathcal{H}; \boldsymbol{y}^*); \quad \tilde{p} = P(\boldsymbol{y}^{**} \in \mathcal{H}; \boldsymbol{y}) \tag{2.14}$$

Then, according to Hall (1992) [16] and Efron and Tibshirani (1998) [9], the third-order accurate p-value (2.11) obtained by the double bootstrap method can be expressed as

$$1 - \Phi(d - c) = P(\tilde{p}^* < \tilde{p}; \hat{\boldsymbol{\eta}}(\boldsymbol{y})) + O(n^{-3/2}) \tag{2.15}$$

Although the double bootstrap has third-order accuracy, formula (2.15) suggests that it requires enormous numbers of bootstrap pseudoreplicates (many more than would

be practically feasible in most cases), and in addition, computation of $\hat{\boldsymbol{\eta}}(\boldsymbol{y})$ is known to be difficult. However, we propose a manipulation of the regular double bootstrap that will greatly speed its implementation and thus facilitate its application to real phylogenetic problems. Our method relies on use of formula (2.16) below (technical report Efron and Tibshirani 1997 [8]), and on computation of $\hat{\boldsymbol{\eta}}(\boldsymbol{y})$ using the PAVA (pool adjacent violators algorithm method) (Ayer et al. 1955 [4]; Zhao 2007 [49]), for assessment of $\mathcal{H}$. To avoid of confusion, we call our proposed approach the 'speedy' double bootstrap method. In the context of this section, the technical report Efron and Tibshirani (1997) [8] showed that $1-\Phi(d-c)$ in formula (2.11) can be stated as follows

$$1-\Phi(d-c) = P(d^* > d; \hat{\boldsymbol{\eta}}(\boldsymbol{y})) + O(n^{-3/2}) \tag{2.16}$$

where d^* is the signed distance from $\boldsymbol{y}^* \sim N_t(\hat{\boldsymbol{\eta}}(\boldsymbol{y}), I_t)$ to $\partial\mathcal{H}$. Formula (2.15) and (2.16) lead immediately to next resultant formula, when the true parameter $\boldsymbol{\eta}$ is on the boundary surface $\partial\mathcal{H}$.

$$P(\tilde{p}^* < \tilde{p}; \hat{\boldsymbol{\eta}}(\boldsymbol{y})) = P(d^* > d; \hat{\boldsymbol{\eta}}(\boldsymbol{y})) + O(n^{-3/2}) \tag{2.17}$$

It is shown that double bootstrap probability also equals to third order accurate p-value $P(d^* > d; \hat{\boldsymbol{\eta}}(\boldsymbol{y}))$, the error being $O(n^{-3/2})$. The formula $P(d^* > d; \hat{\boldsymbol{\eta}}(\boldsymbol{y}))$ indicates that if we can calculate d^* and d using $\boldsymbol{y}^*$ and $\boldsymbol{y}$ respectively, then we do not need to resample from $\boldsymbol{y}^*$ and $\boldsymbol{y}$.

Now, we return to the problem of phylogenetic trees, as seen in H_1 and vector $(l_1, l_2, \cdots, l_K)$. Practically, in addition to the difficulty of computing $\hat{\boldsymbol{\eta}}(\boldsymbol{y})$, calculation of d^* and d is also problematic. However, in the case of H_1, d can be analogous to $\max_{j=2,\cdots,K} l_j - l_1$ (Shimodaira and Hasegawa 2005 [39]) and as already mentioned, computation of $\hat{\boldsymbol{\eta}}(\boldsymbol{y})$ can be achieved basically using the PAVA method, its detail computation will be shown in next section. Furthermore, d^* can be analogous, d^*'s analogous will be considered in the following section.

2.3 The speedy double bootstrap procedure for assessing reliability of phylogenetic trees

We propose a simple double bootstrap method for assessing reliability of phylogenetic trees, that significantly mitigates the challenges of the regular double bootstrap. First we find a vector corresponding to $\hat{\boldsymbol{\eta}}(\boldsymbol{y})$ in formula (2.12). According to Kishino et al. (1990) [20], the vector

$$\boldsymbol{l} = (l_1(\hat{\theta}_1), \cdots, l_K(\hat{\theta}_K)) = (l_1, \cdots, l_K) \tag{2.18}$$

asymptotically follows a multivariate normal distribution, the mean vector of which is

$$\boldsymbol{\mu} = (\mu_1, \cdots, \mu_K) \tag{2.19}$$

Note that, $\boldsymbol{l}$ of formula (2.18) is an unrestricted maximum likelihood estimate for $\boldsymbol{\mu}$. Assuming $\mu_1 = max_{i=1,\cdots,K}\ \mu_i$ is the same as in H_1, under this restriction, the

restricted estimator for $\boldsymbol{\mu}$ can be estimated using the PAVA method and expressed as

$$\hat{\boldsymbol{\mu}} = (\hat{\mu}_1, \cdots, \hat{\mu}_K) \tag{2.20}$$

Then we excise a subset W of the numerical set $\{1, \cdots, K\}$, including element 1, so that

$$\begin{aligned} \hat{\mu}_1 &= \frac{\sum_{j \in W} l_j}{\#W} \\ \hat{\mu}_j &= min(\hat{\mu}_1, l_j), \quad j \in \{2, \cdots, K\} \end{aligned} \tag{2.21}$$

The symbol $\#W$ denotes the number of set W, and vector $\hat{\boldsymbol{\mu}} = (\hat{\mu}_1, \cdots, \hat{\mu}_K)$ corresponds to $\hat{\boldsymbol{\eta}}(\boldsymbol{y})$ in formula (2.12). About the detail of computing $\hat{\boldsymbol{\mu}}$ is in the Tim Robertson's book [18]. For computing $\hat{\boldsymbol{\mu}}$, we also can use the minimum lower sets algorithm in book [18], althorgh this algolithm is for order $l_1 = \min_{j=1,\cdots,15} l_j$. However this algorithm is time-consumming. Therefor we do not use it to compute the projection $\hat{\boldsymbol{\mu}}$. On the other hand, the covariance matrix of vector $(l_1, l_2, \cdots, l_K)$ can be estimated by $\Sigma = (\sigma_{ij})$, with σ_{ij} given by

$$\frac{n}{n-1} \sum_{h=1}^{n} \left[log f_i(\mathbf{x}_h; \hat{\boldsymbol{\theta}}_i) - \frac{1}{n} \sum_{h=1}^{n} log f_i(\mathbf{x}_h; \hat{\boldsymbol{\theta}}_i) \right] \times \left[log f_j(\mathbf{x}_h; \hat{\boldsymbol{\theta}}_j) - \frac{1}{n} \sum_{h=1}^{n} log f_j(\mathbf{x}_h; \hat{\boldsymbol{\theta}}_j) \right] \tag{2.22}$$

Then we need to calculate another quantity, corresponding to d^* in formula (2.16). For this, we must generate $B1$ bootstrap pseudoreplicates of vector $(\hat{\mu}_1, \hat{\mu}_2, \cdots, \hat{\mu}_K)$ in formula (2.20). The pseudoreplicates $(\hat{\mu}_1^{*(b1)}, \cdots, \hat{\mu}_K^{*(b1)}), b1 = 1, \cdots, B1$ are sampled

from

$$(\hat{\mu}_1^{*(b1)}, \cdots, \hat{\mu}_K^{*(b1)})^T \overset{i.i.d.}{\sim} N_K((\hat{\mu}_1, \hat{\mu}_2, \cdots, \hat{\mu}_K)^T, \Sigma) \tag{2.23}$$

where T represents transpose, and Σ is used as above. Vectors $(\hat{\mu}_1^{*(b1)}, \cdots, \hat{\mu}_K^{*(b1)})$ constitute the first-order (first-tier) bootstrap pseudoreplicates. Now, d^* in formula (2.16) can be presented as

$$\max_{j=2,\cdots,K} \hat{\mu}_j^{*(b1)} - \hat{\mu}_1^{*(b1)} \tag{2.24}$$

The following summary of the discussion above serves as a convenient step-by-step outline of our proposed procedure to test the null hypothesis H_1, dubbed the sDBP-test (see Figure 1).

sDBP-test

Step 1 Generate $B1$ bootstrap pseudoreplicates of the vector $(\hat{\mu}_1, \hat{\mu}_2, \cdots, \hat{\mu}_K)$ in (2.20). These pseudoreplicates

$$(\hat{\mu}_1^{*(b1)}, \cdots, \hat{\mu}_K^{*(b1)}), b1 = 1, \cdots, B1 \tag{2.25}$$

are sampled from

$$(\hat{\mu}_1^{*(b1)}, \cdots, \hat{\mu}_K^{*(b1)})^T \overset{i.i.d.}{\sim} N_K((\hat{\mu}_1, \hat{\mu}_2, \cdots, \hat{\mu}_K)^T, \Sigma) \tag{2.26}$$

Step 2 For each vector $(\hat{\mu}_1^{*(b1)}, \cdots, \hat{\mu}_K^{*(b1)})$ of Step 1, calculate

$$\max_{j=2,\cdots,K} \hat{\mu}_j^{*(b1)} - \hat{\mu}_1^{*(b1)}, \quad b1 = 1, \cdots, B1 \tag{2.27}$$

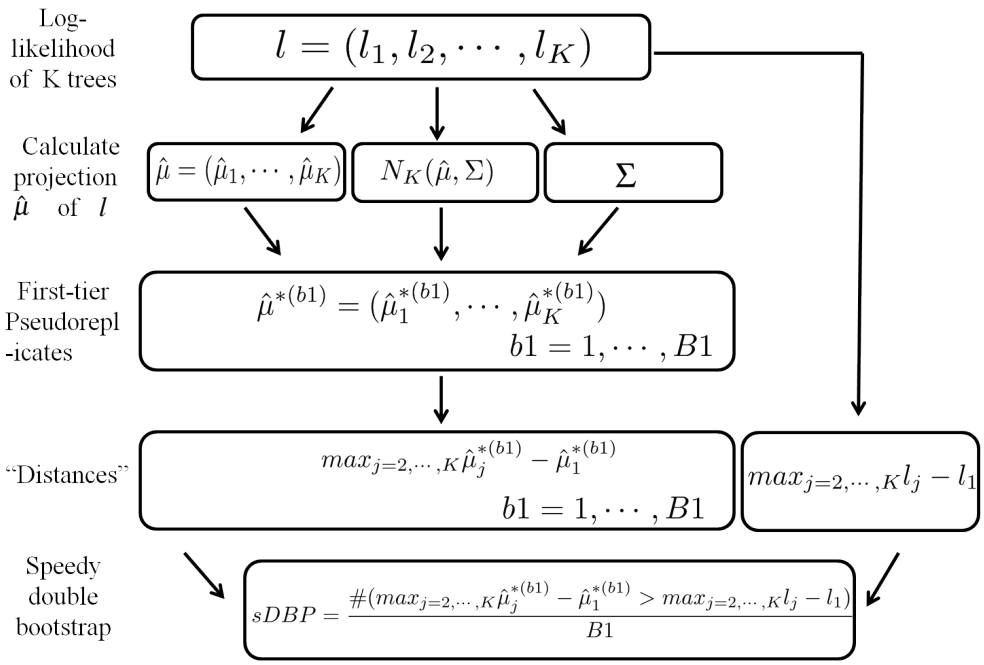

Figure 1: Steps of the sDBP-test for Tree-1 Flow diagram to illustrate the steps in our speedy double bootstrap method (the sDBP-test of H_1). The "Distances" are analogy distances d^* and d in formula (2.16), Σ is calculate from Equation (2.22).

Step 3 For the vector $(l_1, l_2, \cdots, l_K)$ calculate

$$\max_{j=2,\cdots,K} l_j - l_1 \tag{2.28}$$

Step 4 Calculate the p-value for H_1, defined below and denoted $sDBP$

$$sDBP = \frac{\#(\max_{j=2,\cdots,K} \hat{\mu}_j^{*(b1)} - \hat{\mu}_1^{*(b1)} > \max_{j=2,\cdots,K} l_j - l_1)}{B1} \tag{2.29}$$

In exactly the same way as shown for H_1, we can apply the sDBP-test to all other hypotheses $H_k, k = 2, \cdots, K$.

2.4 The double bootstrap procedure for assessing reliability of phylogenetic trees

To properly assess the utility of our sDBP-test, it is necessary to compare our results with those generated using the standard double bootstrap procedure. To this end, we propose the following protocol for application of the regular double bootstrap to test the null hypothesis H_1, dubbed the DBP-test (see Figure 2).

DBP-test

Step 1 Generate $B1$ bootstrap pseudoreplicates of the vector $(\hat{\mu}_1, \hat{\mu}_2, \cdots, \hat{\mu}_K)$. These pseudoreplicates $(\hat{\mu}_1^{*(b1)}, \cdots, \hat{\mu}_K^{*(b1)}), b1 = 1, \cdots, B1$ are sampled from

$$(\hat{\mu}_1^{*(b1)}, \cdots, \hat{\mu}_K^{*(b1)})^T \overset{i.i.d.}{\sim} N_K((\hat{\mu}_1, \hat{\mu}_2, \cdots, \hat{\mu}_K)^T, \Sigma) \tag{2.30}$$

This step is identical to Step 1 of the sDBP-test.

Step 2 Generate $B2$ bootstrap pseudoreplicates of each vector

$$(\hat{\mu}_1^{*(b1)}, \hat{\mu}_2^{*(b1)}, \cdots, \hat{\mu}_K^{*(b1)}) \tag{2.31}$$

from Step 1. These pseudoreplicates $(\hat{\mu}_1^{**(b2)}, \cdots, \hat{\mu}_K^{**(b2)}), b2 = 1, \cdots, B2$ are sampled from

$$(\hat{\mu}_1^{**(b2)}, \cdots, \hat{\mu}_K^{**(b2)})^T \overset{i.i.d.}{\sim} N_K((\hat{\mu}_1^{*(b1)}, \cdots, \hat{\mu}_K^{*(b1)})^T, \Sigma), \quad b1 = 1, \cdots, B1 \tag{2.32}$$

Vectors $(\hat{\mu}_1^{**^{(b2)}}, \cdots, \hat{\mu}_K^{**^{(b2)}})$ constitute the second-order (second-tier) bootstrap pseudoreplicates. Calculate the bootstrap probability $BP^{*(b1)}$ for each $(\hat{\mu}_1^{*^{(b1)}}, \hat{\mu}_2^{*^{(b1)}}, \cdots, \hat{\mu}_K^{*^{(b1)}})$, as follows

$$BP^{*^{(b1)}} = \frac{\#(\text{argmax}(\hat{\mu}_1^{**^{(b2)}}, \cdots, \hat{\mu}_K^{**^{(b2)}}) == 1)}{B2}, \quad b1 = 1, \cdots, B1 \tag{2.33}$$

Step 3 Generate $B2$ bootstrap pseudoreplicates of the vector $(l_1, l_2, \cdots, l_K)$. These pseudoreplicates $(l_1^{**^{(b2)}}, \cdots, l_K^{**^{(b2)}}), b2 = 1, \cdots, B2$ are sampled from

$$(l_1^{**^{(b2)}}, \cdots, l_K^{**^{(b2)}})^T \overset{i.i.d.}{\sim} N_K((l_1, l_2, \cdots, l_K)^T, \Sigma) \tag{2.34}$$

Now calculate the bootstrap probability BP, as follows

$$BP = \frac{\#(\text{argmax}(l_1^{**^{(b2)}}, \cdots, l_K^{**^{(b2)}}) == 1)}{B2} \tag{2.35}$$

Step 3 is the step for calculating traditional BP.

Step 4 Calculate the p-value for H_1, defined below and denoted DBP

$$DBP = \frac{\#(BP^{*^{(b1)}} < BP)}{B1} \tag{2.36}$$

Similarly, we can apply the DBP-test to all other hypotheses $H_k, k = 2, \cdots, K$.

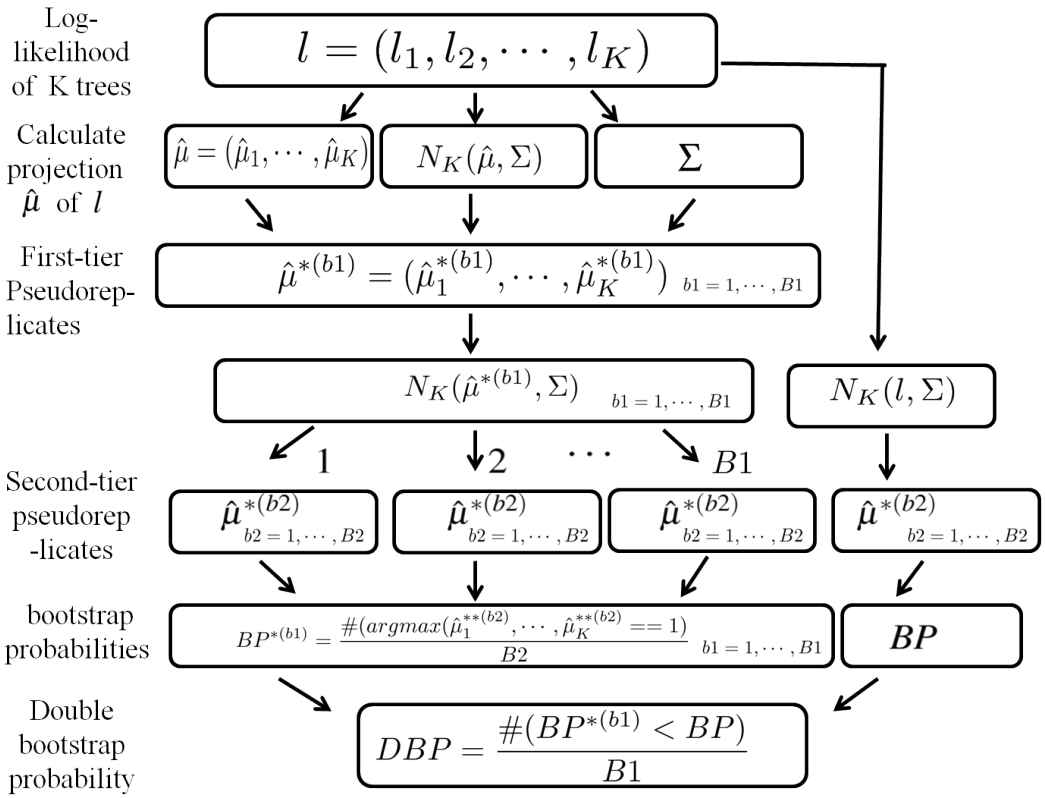

Figure 2: The steps of the DBP-test for Tree-1 Flow diagram to illustrate the steps in our implementation of the regular double bootstrap method (the DBP-test of H_1). Σ is calculate from Equation (2.22), BP is traditional bootstrap proportion.

2.5 Discussions

The maximum likelihood inference $(l_1, l_2, \cdots, l_K)$ for phylogenetic trees is also expressed approximately as (2.10), but the covariance matrix is not identity. This reduces again to the identity matrix case by applying a linear transformation to $(l_1, l_2, \cdots, l_K)$ (Shimodaira and Hasegawa 2005 [39]). Thus, we can use the sDBP-test and the DBP-test for the general phylogenetic tree selection problem.

Let us define G as the region occupied by H_1 in formula (2.8), and ∂H_1 is then

the boundary of the region G. From $\triangle l_i = \max_{j \neq i} - l_i, i = 1, \cdots, K$, for example, the following can be determined: If $\triangle l_1 = (\max_{j=2,\cdots,K} l_j - l_1) > 0$, then we can establish that $\boldsymbol{l} \notin G$. If $\triangle l_1 = (\max_{j=2,\cdots,K} l_j - l_1) < 0$, then we can establish that $\boldsymbol{l} \in G$. If $\triangle l_1 = (\max_{j=2,\cdots,K} l_j - l_1) = 0$, then we can establish that $\boldsymbol{l} \in \partial G$. Therefore, there are analogs in nature for d^*, d and formula (2.27) and (2.28). However, in case of hypothesis H_1, the shape of the null is a polyhedral convex cone and ∂H_1 is nonsmooth at the vertex as well as on the faces of dimensions less than $K - 1$. As already mentioned, the double bootstrap method (Hall 1992 [16]; Efron and Tibshirani 1998 [9]) assumes that the boundary of the region is a smooth surface. Regions with nonsmooth boundaries, in particular, may lead to serious difficulties as discussed by Perlman and Wu (1999, 2003) [29, 30], and further study is needed in this regard.

Our sDBP-test has several advantages over competing methods (DBP-test, AU-test and BP-test). The estimator $\hat{\boldsymbol{\mu}}$ estimates the particular parameter configuration of a given molecular dataset when we assume that the true parameter lies on the boundary ∂H_1. This allows calculation of higher-order accurate p-values by the sDBP-test, an advantageous feature shared with the DBP-test but lacking in the AU-test and the BP-test. Furthermore, because the sDBP-values is not dependent on the BP values, it is not susceptible to the potential biases associated with the BP values. This is a unique advantage of the sDBP-test, and an additional argument for its superiority. However, the sDBP-test is impractical when $\hat{\boldsymbol{\mu}}$ or the signed distance are difficult to estimate. In these cases, other methods of tree selection should be used instead.

Our calculations show that the sDBP-test is not confined to general tree selection

problems, as the algorithm and theory can be used across various fields. Therefore, our work opens the door to many practical model selection problems that use the maximum-likelihood criterion.

Chapter 3

Evaluation of speedy double bootstrap method using biological data

In this chapter, we evaluate our proposed method sDBP-test and DBP-test using real biological data. In Section 3.1, we analyse the mammlian mitochondrial protein sequences for 6 species. In Section 3.2, we analyse mammlian mitochondrial amino acid sequences and 12S and 16S rRNA genes for 20 species. In Section 3.3, we compare the experimental results from these species. In Section 3.4, we discusse the our proposed signed distance and Euclidian signed distance.

3.1 Analysis of mammalian mitochondrial protein sequences for 6 species

3.1.1 Explaining the data

To apply our methods to a biological data set, we reanalyzed the mammalian mt protein sequences from Shimodaira and Hasegawa (1999) [38] using the sDBP-test and the DBP-test. The mammalian protein data set includes sequences of $n = 3414$ amino acids from 6 mammalian species (human, seal, cow, rabbit, mouse and opossum). The clade $\{seal, cow\}$ was significantly supported in preliminary analyses, so only the 15 bifurcating trees that included this clade were considered in our comparisons (Shimodaira and Hasegawa 1999 [38]). The site-wise log-likelihoods obtained by Shimodaira and Hasegawa (1999) [38] for these trees were used to calculate their respective sDBP and DBP values. According to Shimodaira (2002) [37], site-wise log-likelihoods were originally calculated using the software package PAML (Yang 1997 [48]), applying the mtREV model (Adachi and Hasegawa 1996 [1]) of amino acid substitution, and modeling rate heterogeneity among sites with the discrete-gamma distribution (Yang 1996 [47]). We compared our sDBP-test and DBP-test results with each other, and with those reported by Shimodaira (2002) [37] for BP and the AU-test (see Table 1).

3.1.2 Result of the experiment

Performance measures and computer specifications

First, mean square errors were used to assess whether sDBP-test is a reasonable approximation of DBP-test, and to compare its accuracy with the other two tests (BP and AU). Second, to determine whether there was any significant difference between the sDBP and DBP results, we used the availability of paired data for each phylogenetic tree, and applied the paired t-test (which is actually a t-test on a sample of differences). These analyses were performed using the software R 2.9.0 (R Core Team 2012 [31]) on a personal computer with the following specifications: 2.50 GHz CPU (Core (TM) i5-2520M CPU) and 8.00 GB RAM.

Table 1 presents the results of our sDBP value calculations for the 15 phylogenetic trees analyzed in this study, along with values reported by Shimodaira (2002) [37] for traditional BP analyses and the AU-test. The confidence sets of trees obtained by the sDBP-test and the DBP-test at $\alpha = 0.05$ were $\{1, 2, 3, 4, 5, 6, 7\}$ and $\{1, 2, 3, 5, 7\}$, respectively (Table 1). The sDBP tree set was thus slightly larger than the set selected by DBP-test. Tree 7 is the most strongly supported as T_{ML} by recent analyses incorporating additional sequence data (e.g., Cao et al. 2000 [6]; Madsen et al. 2001 [24]; Murphy et al. 2001 [26]), and our results for this tree indicate that sDBP=0.084>0.05 and DBP=0.056 $>$ 0.05. Our conclusions are thus not in contradiction with the latest data. Based on the values in Table 1, mean square errors between DBP-test and the other three test (sDBP, BP and AU) were, respectively, 0.001, 0.003 and 0.003. Thus, the sDBP-test apparently provides a good approximation of the regular DBP-test, with the sDBP-DBP comparison having the lowest error (0.001). In addition, comparison of sDBP and DBP-test results using the paired t-test returned a p-value of 0.101, providing no evidence of a significant difference between the methods.

Table 1: Comparison of four different p-values from analyses of fifteen mammalian trees, based on protein sequence data from Shimodara and Hasegawa (1999) [38]. The p-values that are NOT significant at $\alpha = 0.05$ are emphasized in bold type.

Tree[a]	$\triangle l_i$	BP_i^b	DBP_i^c	$sDBP_i^d$	AU_i^e	Tree form[f]
1	-2.7	**0.579**	**0.607**	**0.576**	**0.789**	(((1(23))4)56)
2	2.7	**0.312**	**0.458**	**0.401**	**0.516**	((1((23)4))56)
3	7.4	0.036	**0.167**	**0.235**	**0.114**	(((14)(23))56)
4	17.6	0.013	0.041	**0.116**	**0.075**	((1(23))(45)6)
5	18.9	0.035	**0.082**	**0.110**	**0.128**	(1((23)(45))6)
6	20.1	0.005	0.031	**0.069**	0.029	(1(((23)4)5)6)
7	20.6	0.017	**0.056**	**0.084**	**0.101**	((1(45))(23)6)
8	22.2	0.001	0.007	0.042	0.009	((15)((23)4)6)
9	25.4	0.000	0.002	0.022	0.000	(((1(23))5)46)
10	26.3	0.003	0.011	0.023	0.028	(((15)4)(23)6)
11	28.9	0.000	0.003	0.013	0.003	(((14)5)(23)6)
12	31.6	0.000	0.001	0.004	0.001	(((15)(23))46)
13	31.7	0.000	0.002	0.005	0.001	(1(((23)5)4)6)
14	34.7	0.000	0.003	0.001	0.005	((14)((23)5)6)
15	36.2	0.000	0.001	0.000	0.002	((1((23)5))46)

[a]Trees are numbered by increasing order of $\triangle l_i = max_{j \neq i} l_j - l_i$, the difference between the log-likelihood value for a given tree and the largest value among all other trees.
[b]Bootstrap probability, calculated from 10000 pseudoreplicates (from Shimodaira (2002)).
[c]Double bootstrap probability, calculated from 25 million pseudoreplicates (B1 = 5 $\times$1000, $B2 = 5 \times 1000$).
[d]Speedy double bootstrap probability, calculated from 10000 pseudoreplicates (B1 = 10000).
[e]Multiscale bootstrap probability, calculated from 100000 pseudoreplicates (AU-test; from Shimodaira (2002)).
[f]Taxon labels: 1 = human, 2 = seal, 3 = cow, 4 = rabbit, 5 = mouse, 6 = opossum.

3.1.3 Comparison of computational speed

Computational measures

For the sDBP, DBP, AU and BP-test, we measured the time taken to calculate a p-value for Tree 1, based on the site-wise log-likelihood data. We used the RELL approximation method with the BP-test (Kishino et al., 1990), and conducted two separate sets of analyses. In the first set, we applied the sDBP-test with $B1 = 10^3$ pseudoreplicates, the DBP-test with $B1 = 10^3$ and $B2 = 10^3$ pseudoreplicates, and

the BP-test with 10^3 pseudoreplicates. In the second set, we applied the sDBP-test with $B1 = 5\times10^3$ pseudoreplicates, the DBP-test with $B1 = 5\times10^3$ and $B2 = 5\times10^3$ pseudoreplicates, and the BP-test with 5×10^3 pseudoreplicates. For both sets, the BP-test was the fastest, followed by the sDBP-test, the AU-test then the DBP-test. For the first set of calculations (lower numbers of pseudoreplicates) the sDBP-test was 1021-fold faster than the DBP-test, and this advantage improved substantially for the second set (higher pseudoreplication), with the sDBP-test being 5076-fold faster than the DBP-test.

Table 2: Comparison of the BP, DBP, sDBP and AU methods, regarding their speed for computing a p-value for tree-1.

	BP	DBP	sDBP	AU	Speed increase (sDBP/DBP)
Time (secs)[a]	0.69	715	0.73	3.72	1021-fold
Time (secs)[b]	3.52	17921	3.53	14.39	5076-fold

[a] Case of $B1 = 10^3$, $B2 = 10^3$ pseudoreplicates
[b] Case of $B1 = 5 \times 10^3$, $B2 = 5 \times 10^3$ pseudoreplicates

3.2 Analysis of mammalian mitochondrial amino acid sequences and 12S and 16S rRNA genes for 20 species

3.2.1 Explaining the data

In this case, we analyzed the amino acid sequences of the mammalian mitochondrial protein-coding genes and the DNA sequences of the 12S and 16S rRNA genes using the sDBP-test and the DBP-test. We included 20 mammalian species in these

analyses, belonging to eight major clades: Primates, Lagomorpha, Rodentia, Fereungulata, Chiroptera, Soricomorpha, Marsupialia and Monotremata (see Table 3). We used the representatives from Marsupialia and Monotremata as outgroups. In cases where a major clade was represented by more than two species, the following relationships were applied, based on the results of Cao et al. (2000) [6] : in the Fereuungulata, ((domestic cat, harbor seal), (horse, (Indian rhinoceros, white rhinoceros)), (cow, blue whale)) ; in the Primates, (((human, chimpanzee), western gorilla), Sumatran orangutan); and in the outgroup, ((American opossum, wallaroo), platypus). We also assumed the relationship (Fereuungulata, (Chiroptera, Soricomorpha)), based once again on strong support for these groupings in Cao et al. (2000) [6]. We can thus divide these taxa into five monophyletic groups: Primates (group I), Lagomorpha (group II), Rodentia (group III), (Fereuungulata, (Chiroptera, Soricomorpha)) (group IV), (Marsupialia, Monotremata) (group V). Finally all of the 15 unrooted trees (see Table 3 footnote *c*) compatible with these five groups were considered in our comparisons. These 15 unrooted trees of 20 species are shown in Table 3, respectively.

The 12 proteins coded for in the mammalian mitochondrial genome are ND1, ND2, COX1, COX2, ATP8, ATP6, COX3, ND3, ND4L, ND4, ND5, and CYTB. Amino acid alignments for these proteins were constructed using ClustalW version 1.83 (Thompson et al. 1994 [44]), and all positions with gaps were excluded from analyses. This resulted in a final total of 3593 amino acids in the alignment. DNA sequences for the small (12S) and large (16S) mitochondrial rRNA genes were also aligned using ClustalW. As with the amino acid sequences, alignment positions with

gaps were excluded from the analyses leaving a total of 870 and 1416 sites for the 12S and 16S genes, respectively. Analysis of the amino acid alignment was conducted using the CodeML program of the PAML package (Yang 1997 [48]) , applying the Empirical+F model and the mtREV24.dat rate matrix (Adachi and Hasegawa 1996 [1]), and modeling rate heterogeneity among sites with the discrete gamma distribution (Yang 1996 [47]). The CodeML program provides site-wise log-likelihoods for each of the 3593 amino acids in the alignment. The 12S and 16S rRNA gene sequences were analyzed using the BaseML program of the PAML package (Yang 1997 [48]), applying the REV model and modeling rate heterogeneity among sites with the discrete gamma distribution. The BaseML program provides site-wise log-likelihoods for each of the bases in the 12S and 16S alignments. These site-wise log-likelihood scores for the amino acids and for the two rRNA gene sequences were summed to determine the best ML tree and to calculate the sDBP, DBP, AU and BP for the 15 candidate trees.

3.2.2 Result of experiment

Performance measures and computer specifications

About performance measures and computer specifications are same as Subsection 3.1.2. Table 4 presents results of our sDBP-test and DBP-test calculations for the 15 phylogenetic trees analyzed in this study, alongside values for BP- and AU- values. The confidence sets of trees obtained, respectively, by the sDBP-test and the DBP-test at $\alpha = 0.05$ were $\{1, 2, 3, 4, 5, 6, 7, 8, 9\}$ and $\{1, 2, 3, 4, 5, 6, 8, 9\}$ (Table 4). The

Table 3: The 20 mammalian species used in this study, including GenBank accession numbers for sequence data, and the major clade membership and group membership (Groups I - V) of each species.

GenBank accession number[a]	Binomial scientific name	Common name	Major clade[b]	Group[c]
NC_001807	*Homo sapiens*	human	Primates	I
NC_001643	*Pan troglodytes*	chimpanzee	Primates	I
NC_001645	*Gorilla gorilla*	western gorilla	Primates	I
NC_002083	*Pongo abelii*	Sumatran orangutan	Primates	I
NC_001913	*Oryctolagus cuniculus*	rabbit	Lagomorpha	II
NC_002658	*Thryonomys swinderianus*	greater cane rat	Rodentia	III
NC_005089	*Mus musculus*	house mouse	Rodentia	III
NC_001700	*Felis catus*	domestic cat	Fereuungulata	IV
NC_001325	*Phoca vitulina*	harbor seal	Fereuungulata	IV
NC_001640	*Equus caballus*	horse	Fereuungulata	IV
X97336	*Rhinoceros unicornis*	Indian rhinoceros	Fereuungulata	IV
NC_001808	*Ceratotherium simum*	white rhinoceros	Fereuungulata	IV
NC_001601	*Balaenoptera musculus*	blue whale	Fereuungulata	IV
NC_006853	*Bos taurus*	cow	Fereuungulata	IV
AF061340	*Artibeus jamaicensis*	Jamaican fruit-eating bat	Chiroptera	IV
AB042770	*Pteropus dasymallus*	Ryukyu flying fox	Chiroptera	IV
Y19192	*Talpa europaea*	European mole	Soricomorpha	IV
NC_001610[d]	*Didelphis virginiana*	North American opossum	Marsupialia	V
Y10524[d]	*Macropus robustus*	wallaroo	Marsupialia	V
NC_000891[d]	*Ornithorhynchus anatinus*	platypus	Monotremata	V

[a]NCBI (GenBank) accession number.
[b]The major clade that the species belongs to.
[c]The group (I - V) that the species belongs to. The 15 unrooted trees compatible with these five groups are t1: (((I,IV),II),III,V), t2: ((I,IV),(II,III),V), t3: (((I,II),IV),III,V), t4: ((I,(II,III)),IV,V), t5: ((I,(IV,II)),III,V), t6: (I,(IV,(II,III)),V), t7: (((I,IV),III),II,V), t8: (((I,II),III),IV,V), t9: (((I,III),II),IV,V), t10: (I,((IV,II),III),V), t11: ((I,III),(IV,II),V), t12: ((I,II),(IV,III),V), t13: (I,((IV,III),II),V), t14: (((I,III),IV),II,V), t15: ((I,(IV,III)),II,V).
[d]Outgroup species.

sDBP tree set was thus slightly larger than the set selected by DBP-test. Tree-4 is most strongly supported as the T_{ML} by previous studies that have included a more comprehensive set of clades and species than we have used here (e.g., Cao et al. 2000 [6]; Madsen et al. 2001 [24]; Murphy et al. 2001 [26]). Our results for this tree indicate that sDBP=0.327> 0.05 and DBP=0.319 > 0.05, and our conclusions are thus not in contradiction with the latest data. Based on the values in Table 4, mean square errors between DBP-test and the other three tests (sDBP, BP and AU) were, respectively, 0.003, 0.022 and 0.006. Thus, the sDBP-test apparently provides a good approximation of the regular DBP, with the sDBP-DBP comparison having

Table 4: sDBP, DBP, AU and BP results based on final amino acid and DNA sequence alignments, for each of the 15 candidate trees of 20 mammalian species. The p-values that are NOT significant at $\alpha = 0.05$ are emphasized in bold type.

Tree form[a]	$\triangle l_i$	BP_i^b	DBP_i^c	$sDBP_i^d$	AU_i^e
1:((((((1,2),3),4),(((8,9),(10,(11,12)),(13,14)),((15,16),17))),5),(6,7),((18,19),20))	-5.5	**0.474**	**0.688**	**0.748**	**0.881**
2:(((((1,2),3),4),(((8,9),(10,(11,12)),(13,14)),((15,16),17))),(5,(6,7)),((18,19),20))	5.5	**0.189**	**0.458**	**0.436**	**0.524**
3:((((((1,2),3),4),5),(((8,9),(10,(11,12)),(13,14)),((15,16),17))),(6,7),((18,19),20))	8.0	**0.115**	**0.332**	**0.384**	**0.375**
4:(((((1,2),3),4),(5,(6,7))),(((8,9),(10,(11,12)),(13,14)),((15,16),17)),((18,19),20))	11.6	**0.094**	**0.319**	**0.327**	**0.429**
5:(((((1,2),3),4),((((8,9),(10,(11,12)),(13,14)),((15,16),17)),5)),(6,7),((18,19),20))	11.9	0.035	**0.197**	**0.307**	**0.193**
6:((((1,2),3),4),((((8,9),(10,(11,12)),(13,14)),((15,16),17)),(5,(6,7))),((18,19),20))	14.5	0.042	**0.224**	**0.263**	**0.271**
7:((((((1,2),3),4),(((8,9),(10,(11,12)),(13,14)),((15,16),17))),(6,7)),5,((18,19),20))	17.9	0.001	0.022	**0.164**	0.014
8:((((((1,2),3),4),5),(6,7)),(((8,9),(10,(11,12)),(13,14)),((15,16),17)),((18,19),20))	18.0	0.028	**0.201**	**0.223**	**0.332**
9:((((((1,2),3),4),(6,7)),5),(((8,9),(10,(11,12)),(13,14)),((15,16),17)),((18,19),20))	20.5	0.021	**0.142**	**0.181**	**0.205**
10:((((1,2),3),4),(((((8,9),(10,(11,12)),(13,14)),((15,16),17)),5),(6,7)),((18,19),20))	30.0	0.001	0.019	0.047	0.022
11:(((((1,2),3),4),(6,7)),((((8,9),(10,(11,12)),(13,14)),((15,16),17)),5),((18,19),20))	32.1	0.000	0.022	0.033	0.014
12:(((((1,2),3),4),5),((((8,9),(10,(11,12)),(13,14)),((15,16),17)),(6,7)),((18,19),20))	36.1	0.000	0.032	0.011	0.000
13:((((1,2),3),4),(((((8,9),(10,(11,12)),(13,14)),((15,16),17)),(6,7)),5),((18,19),20))	36.7	0.000	0.032	0.011	0.000
14:((((((1,2),3),4),(6,7)),(((8,9),(10,(11,12)),(13,14)),((15,16),17))),5,((18,19),20))	37.0	0.000	0.035	0.012	0.000
15:(((((1,2),3),4),((((8,9),(10,(11,12)),(13,14)),((15,16),17)),(6,7))),5,((18,19),20))	44.0	0.000	0.045	0.000	0.000

[a]Trees are numbered by increasing order of $\triangle l_i = \max_{j \neq i} l_j - l_i$, the difference between the log-likelihood value of a given tree and the largest value among all other trees. Species labels: 1 = human, 2 = chimpanzee, 3= western gorilla, 4 = Sumatran orangutan, 5 = rabbit , 6 = greater cane rat, 7 = house mouse, 8 = domestic cat, 9 = harbor seal, 10 = horse, 11 = Indian rhinoceros, 12 = white rhinoceros,13 = blue whale, 14 = cow, 15 = Jamaican fruit-eating bat, 16 = Ryukyu flying fox, 17 = European mole, 18 = North American opossum, 19 = wallaroo, 20 = platypus.
[b]Bootstrap probability, calculated from B1=10000 pseudoreplicates.
[c]Double bootstrap probability, calculated from 1 million pseudoreplicates (B1 = 1000, B2 = 1000).
[d]Speedy double bootstrap probability, calculated from B1=10000 pseudoreplicates.
[e]Multiscale bootstrap probability, calculated from B1=10000 pseudoreplicates.

the lowest error (0.003). In addition, comparison of sDBP and DBP-test results using the paired t-test returned a p-value of 0.079, providing no evidence of a significant difference between these methods.

3.2.3 Comparisons of computational speed

Computational measures

About this experiment's computational measures are same as Subsection 3.1.3. Results of the two sets of analyses conducted to compare computational speed between the sDBP, DBP, AU and BP tests are shown in Table 5. In both sets the BP-test was the fastest, followed by the sDBP-test, then the DBP-test. In the first set of

calculations (lower numbers of pseudoreplicates) the sDBP-test was 371-fold faster than DBP-test, and this advantage improved substantially in the second set (higher pseudoreplication), in which the sDBP-test was 3893-fold faster than DBP-test.

Table 5: Comparison of the BP, DBP, sDBP and AU tests, regarding their speed to compute a p-value for tree-1.

	BP	DBP	sDBP	AU	Speed increase (sDBP/DBP)
Time (secs)[a]	0.448	900.02	2.42	5.91	371-fold
Time (secs)[b]	2.20	23164	5.95	24.44	3893-fold

[a] ($B1 = 10^3, B2 = 10^3$ pseudoreplicates)
[b] ($B1 = 5 \times 10^3, B2 = 5 \times 10^3$ pseudoreplicates)

3.3 Discussion of the two experimental results

Based on our comparison of the speedy double bootstrap method with other approaches for estimating the reliability of phylogenetic trees (regular double bootstrap, multiscale bootstrap (AU-test) and traditional bootstrap probability) we recommend the sDBP-test for general tree selection problems. This method is computationally less burdensome than the AU-test or the DBP-test and is the most close to the BP-test. According to the results of the experiments, initially, the confidence sets of the trees obtained by the sDBP-test and the DBP-test at $\alpha = 0.05$ both included the tree that was most strongly supported as T_{ML} by recent analyses incorporating additional sequence data (e.g., Cao et al. 2000 [6]; Madsen et al. 2001 [24]; Murphy et al. 2001 [26]). For the 6 species, Tree 7 in Table 1 was most strongly supported as T_{ML} by recent analyses incorporating additional sequence data, whereas for the 20 species, Tree 4 in Table 4 was the most strongly supported as T_{ML} by recent analyses incorpo-

rating additional sequence data. Tree 7 in Table 1 and Tree 4 in Table 4 are different at species level, but not at Group level (in Table 3). At Group level, both are the same as T_{ML} by recent analyses incorporating additional sequence data (e.g., Cao et al. 2000 [6]; Madsen et al. 2001 [24]; Murphy et al. 2001 [26]). As mentioned in Subsection 1.1.1, for a confidence set of models, our sDBP-test gives a confidence set of candidate trees, and includes the best topology $\max_{i=1,\cdots,15} \mu_i$ where μ_i is defined by equation 2.6, with an error rate below the 0.05 level. This is one of the statistical decision problems mentioned briefly by Lehmann (1959) [22]. Thus, our sDBP tree set does not immediately give the work for straightly gives the best topology.

In both cases, the sDBP tree set was slightly larger than that selected by the DBP-test. However, the sDBP tree set for the case of 20 species was larger than the sDBP tree set for 6 species. This result might indicate a trend when the number of species becomes large and the number of candidate trees is small, but further experiments would be needed to prove this. Finally, when the computational burden is large, the sDBP-test was much faster than the DBP-test and much closer to the BP-test.

3.4 About the signed distance

In our sDBP-test for the both experiment, the signed distance has been defined as $d_1 = \max_{j=2,\cdots,15} l_j - l_1$ for phylogenetic Tree 1, similarly for phylogenetic Tree $i, i = 2, \cdots, 15$ the signed distance has been defined as $d_i = \max_{j=1,2,\cdots,i-1,i+1,\cdots,15} l_j - l_i$. However, in Efron and Tibushirani's technical report [8], they defined the signed distance as Euclidian signed distance. In this way, we define the Euclidian signed

distance for phylogenetic Tree 1 as follow.

$$d_{1(Eu)} = \begin{cases} -\sqrt{(l_1 - \hat{\mu}_1)^2 + \cdots + (l_{15} - \hat{\mu}_{15})^2}, & l_1 > l_j, j = 2, \cdots, 15 \\ \\ \sqrt{(l_1 - \hat{\mu}_1)^2 + \cdots + (l_{15} - \hat{\mu}_{15})^2}, & \text{otherwise} \end{cases} \tag{3.1}$$

where $\hat{\boldsymbol{\mu}} = (\hat{\mu}_1, \cdots, \hat{\mu}_{15})$ is the projection of log-likelihood vector $\boldsymbol{l}$ to the boundary ∂H_1. For other phylogenetic tree $j, j = 2, \cdots, 15$, we can similarly define its Euclidian signed distance $d_{j(Eu)}$.

Here our purpose is to compare $\text{sDBP}^{\text{j}}_{\text{Pro}}$ and $\text{sDBP}^{\text{j}}_{\text{Eu}}, j = 1 \cdots, 15$ for the 20 species, the symbol $\text{sDBP}^{\text{j}}_{\text{Pro}}$ is the p-value using our proposed signed distance d_j and its replications $d_j^{*(b1)}, b1 = 1, \cdots, 10000$ to calculate itself for 15 trees and symbol $\text{sDBP}^{\text{j}}_{\text{Eu}}$ are the p-value using Euclidian signed distance $d_{j(Eu)}$ and its replications $d_{j(Eu)}^{*(b1)}, b1 = 1, \cdots, 10000$ to calculate itself for 15 trees. For this purpose, we calculate $\text{sDBP}^{\text{j}}_{\text{Eu}}$ and plot the data $(\text{sDBP}^{\text{j}}_{\text{Pro}}, \text{sDBP}^{\text{j}}_{\text{Eu}})$. The results of $\text{sDBP}^{\text{j}}_{\text{Eu}}$ values for 15 phylogenetic trees are $0.520, 0.440, 0.393, 0.345, 0.323, 0.273, 0.184, 0.232, 0.195, 0.051,$ $0.038, 0.010, 0.010, 0.014, 0.000$. And the plot is shown in Figure 3, through Figure 3, we can see that except in upper right point that is pair $(\text{sDBP}^{1}_{\text{Pro}}, \text{sDBP}^{1}_{\text{Eu}})$ at Tree 1, the value sDBP_{Eu} is almost equal sDBP_{Pro}. In addition, comparison of $\text{sDBP}^{\text{j}}_{\text{Eu}}$ and $\text{sDBP}^{\text{j}}_{\text{Pro}}, j = 1, \cdots, 15$ values we also using the paired t-test returned a p-value 0.629. Therefor we can see that, there are no evidence of a significant difference between the values. Thus according our this experimental result, the velues $\text{sDBP}^{\text{j}}_{\text{Eu}}, j = 1, \cdots, 15$ appear that they are good apporximations of $\text{sDBP}^{\text{j}}_{\text{Pro}}, j = 1, \cdots, 15$. However, for computing $\text{sDBP}^{\text{j}}_{\text{Eu}}, j = 1, c \cdots, 15$, we need to calculate the projection of $\boldsymbol{l}$ as well

as projections of bootstrap replications $\hat{\boldsymbol{\mu}}^{*(b1)}, b1 = 10000$. These lead its complexity larger than sDBP_{Pro}, so we used sDBP_{Pro} as the p-vealue in our speedy double bootstrap method.

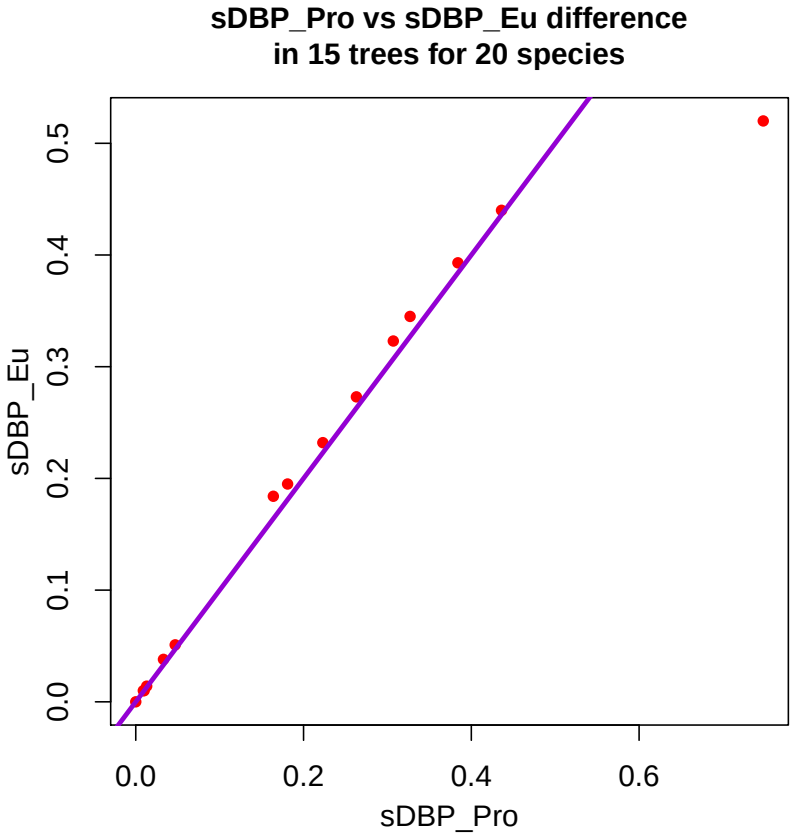

Figure 3: Plots of sDBP_{Eu} and sDBP_{Pro} for 15 trees in 20 species case

Chapter 4

Evaluation of the rejection probabilities for four bootstrap methods

As mentioned Chapter 1, there are no one give a comparison that compare the statistical properties and computational efficiency of the sDBP-method with three other bootstrap methods (DBP, AU and BP-method) in artificial simulation. This motivated us to give a simulation to evaluate this four bootstrap methods. In this chapter, we use simulations based on artificial data to investigate the relative statistical performance and computational efficiency of the four different bootstrap methods (sDBP, DBP, AU and BP). Here we emphasize that in this Chapter the sDBP-method is same as the basic idea of technical report Efron and Tibshirani (1997) [8]. In Section 4.1, we design and perform a simulation from the Normal Model, and use our simulation results to compare the rejection probabilities for sDBP, DBP, AU and

BP-method. Graphically, the rejection probability of sDBP, DBP and AU-method are similar, whereas that of BP-method is noticeably different. Then, in Section 4.2, we apply statistical tests that is multiple comparison to compare the difference in rejection probabilities between the first-order accurate BP and the third-order accurate methods. Finally, in Section 4.3, we compare the computational speed of the four methods. Note that in this chapter the signed distance is the signed Euclidean distance used by the "region problem" in Efron (1985) [7]. Unless otherwise specified, the original analyses described below were performed using the software R 2.9.0 (R Core Team 2012 [31]) on the TSUBAME (a large-scale supercomputer at Tokyo Institute of Technology) with the following specifications: 2.0 GHz CPU (Intel Xeon 7550, 64 core) and 128.00 GB RAM.

4.1 Simulation from the normal model data

4.1.1 Setting up of the Simulation

As a first step, for the normal model $\boldsymbol{y} \sim N_2(\boldsymbol{\eta}, I_2)$ we formulated the following null hypothesis regarding parameter $\boldsymbol{\eta}$

$$H_0 : \boldsymbol{\eta} \in R(h), R(h) = \{\boldsymbol{\eta} = (\eta_1, \eta_2) : \eta_2 \geq \theta/\eta_1, \eta_1 \geq 0\} \tag{4.1}$$

We set the value of θ to be 0.1 and we investigated 7 different curvature values representing increasing order, denote $Ci, i = 1, \cdots, 7$.

C1	C2	C3	C4	C5	C6	C7
0.1970	0.4525	0.6711	1.0289	1.2807	1.9056	2.2361

these curvature values are at 7 different parameters $(\eta_1, 0.1/\eta_1)$

$$\eta_1 = 1,\ 0.75,\ 0.65,\ 0.55,\ 0.5,\ 0.4,\ \sqrt{0.1}$$

For each curvature value, we generate 10^5 numbers data $\boldsymbol{y}$ from the normal distribution $N_2(\boldsymbol{\eta}, I_2)$. We refer to this procedure as " Experiment 1 ", and the resultant 7 sets of simulated data allow us to test the null hypothesis 7 times at a range of different curvatures. We repeated this procedure at two alternative values of θ $(\theta = 1, 10)$, adjusting the null hypothesis accordingly. In each case, 10^5 simulations on data $y_i, i = 1, \cdots, 10^5$ were performed for each of 7 different curvature values. For $\theta = 1$ $(\eta_2 = 1/\eta_1)$, referred to as " Experiment 2 ", the set of curvature values is represented as

C 1	C 2	C 3	C 4	C 5	C 6	C 7
0.0160	0.0462	0.0727	0.1232	0.2283	0.4522	0.7071

these curvature values are at 7 different parameters $(\eta_1, 1/\eta_1)$

$$\eta_1 = 5,\ 3.5,\ 3,\ 2.5,\ 2,\ 1.5,\ 1$$

For $\theta = 10$ $(\eta_2 = 10/\eta_1)$, referred to as " Experiment 3 ", the set of curvature values is represented as

C1	C2	C3	C4	C5	C6	C7
0.0548	0.0671	0.0828	0.1281	0.1582	0.1906	0.2224

these curvature values are at 7 different parameters $(\eta_1, 10/\eta_1)$

$$\eta_1 = 7,\ 6.5,\ 6,\ 5,\ 4.5,\ 4,\ \sqrt{10}$$

4.1.2 Step of the Simulation

We will summary our simulation procedure using $\theta = 0.1$ and $\eta_1 = 1$ in "Experiment 1". First we will compute the DBP values and then compute the sDBP values.

Step 1: Generating the data

15ptGenerate 10^5 numbers data $\boldsymbol{y}$ from following distribution

$$N_2(\boldsymbol{\eta}, I_2) \tag{4.2}$$

Step 2: Calculation of the four competing probabilities

For each of the 10^5 numbers data $\boldsymbol{y}$, we conducted first-tier bootstrap sampling from the normal distribution

$$\mathbf{Y}^*|\hat{\boldsymbol{\eta}}(\boldsymbol{y}) \sim N_2(\hat{\boldsymbol{\eta}}(\boldsymbol{y}), I_2) \tag{4.3}$$

where $\hat{\boldsymbol{\eta}}(\boldsymbol{y})$ is the projection of data $\boldsymbol{y}$ on the boundary curve $\eta_1\eta_2 = 0.1$. Through this resampling, we can get first-tier bootstrap pseudoreplicates $\boldsymbol{y}^*_{b1}, b1 = 1, \cdots, 10^3$ and second-tier bootstrap sampling from the normal distribution

$$\begin{aligned} \mathbf{Y}^{**}|\boldsymbol{y}^*_{b1} &\sim N_2(\boldsymbol{y}^*_{b1}, I_2), b1 = 1, \cdots, 10^3 \\ \mathbf{Y}^{**}|\boldsymbol{y} &\sim N_2(\boldsymbol{y}, I_2) \end{aligned} \tag{4.4}$$

We generated $B1 = 10^3$ first-tier bootstrap pseudoreplicates $\boldsymbol{y}^*_{b1}, b1 = 1, \cdots, B1$. For each of these $\boldsymbol{y}^*_{b1}$ and the single $\boldsymbol{y}$, we then used formula (4.4) to generate $B2 = 10^3$ second-tier bootstrap pseudoreplicates $\boldsymbol{y}^{**}_{b2}, b2 = 1, \cdots, B2$, and calculated the frequency of the event $\boldsymbol{y}^{**} \in R(h)$. We denote these frequencies as $BP(\boldsymbol{y}^*_{b1}), BP(\boldsymbol{y})$ and can express as

$$BP(\boldsymbol{y}^*_{b1}) = \frac{\#(\boldsymbol{y}^{**}_{b2} \in R(h))}{B2}, \quad b1 = 1, \cdots, 1000 \tag{4.5}$$

$$BP(\boldsymbol{y}) = \frac{\#(\boldsymbol{y}^{**}_{b2} \in R(h))}{B2} \tag{4.6}$$

The double bootstrap probability is then calculated as follows

$$DBP = \frac{\#(BP(y^*_{b1}) \leq BP(y))}{B1} \tag{4.7}$$

Then we compte the sDBP. First determine the signed Euclidean distance d^*_{b1} between each $\boldsymbol{y}^*_{b1}, b1 = 1, \cdots, B1$ and its respective projection $\hat{\boldsymbol{\eta}}(\boldsymbol{y}^*_{b1}), b1 = 1, \cdots, B1$. Also calculate the signed Euclidean distance d between the single $\boldsymbol{y}$ and its projection $\hat{\boldsymbol{\eta}}(\boldsymbol{y})$. For example, the signed distance between $\boldsymbol{y} = (y_1, y_2)$ and its projection $\hat{\boldsymbol{\eta}}(\boldsymbol{y}) = (\hat{\eta}(y)_1, \hat{\eta}(y)_2)$ computing with analytical method (not the PAVA method) is

given by

$$d = \begin{cases} -\sqrt{(y_1 - \hat{\eta}(y)_1)^2 + (y_2 - \hat{\eta}(y)_2)^2}, & y_2 > 0.1/y_1,\ y_1 > 0 \\ \\ \sqrt{(y_1 - \hat{\eta}(y)_1)^2 + (y_2 - \hat{\eta}(y)_2)^2}, & \text{otherwise} \end{cases} \tag{4.8}$$

Finally, the sDBP can be calculated as follows

$$sDBP = \frac{\#(d^*_{b1} \geq d)}{B1} \tag{4.9}$$

The traditional bootstrap probability is calculated using formula (4.6). We have omitted a description of how AU values were calculated since this is adequately discussed elsewhere, and interested readers are referred to Shimodaira (2002) [37] for a full explanation of this method.

Note that calculations have only been shown for one of the 10^5 numbers data y based on $\theta = 0.1$ and $\eta_1 = 1$ in "Experiment 1". However, calculations of the four competing probabilities from simulated data based on other curvature values and other values of θ were also carried out as described above.

Step 3: Proportion of each type of probability that is ≤ 0.05

In the three experiments, 10^5 separate calculations of sDBP-probability (sDBP), DBP-probability (DBP), AU-probability (AU) and BP-probability (BP) were carried out for each of the 7 curvature value at 7 parameter values. In each case we can thus

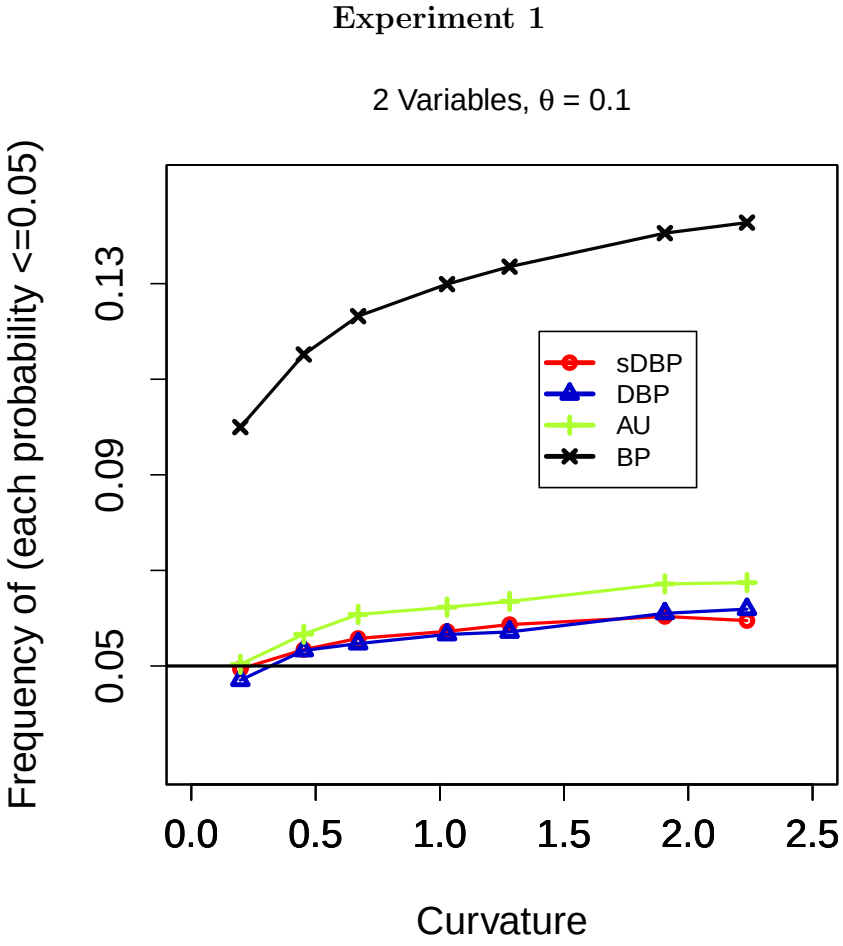

Figure 4: Plots of rejection probability (proportion of probability ≤ 0.05) vs curvature for the four different types of probability (sDBP, DBP, AU and BP), in Experiment 1 (see text for details).

obtain the proportion of each probability that is less than 0.05, as follows

$$\#(sDBP_i \leq 0.05)/10^5 \quad , \quad \#(DBP_i \leq 0.05)/10^5 \tag{4.10}$$

$$\#(AU_i \leq 0.05)/10^5 \quad , \quad \#(BP_i \leq 0.05)/10^5$$

The proportion of each probability $<= 0.05$ (i.e., $\#(\text{probability} <= 0.05)/10^5$) is

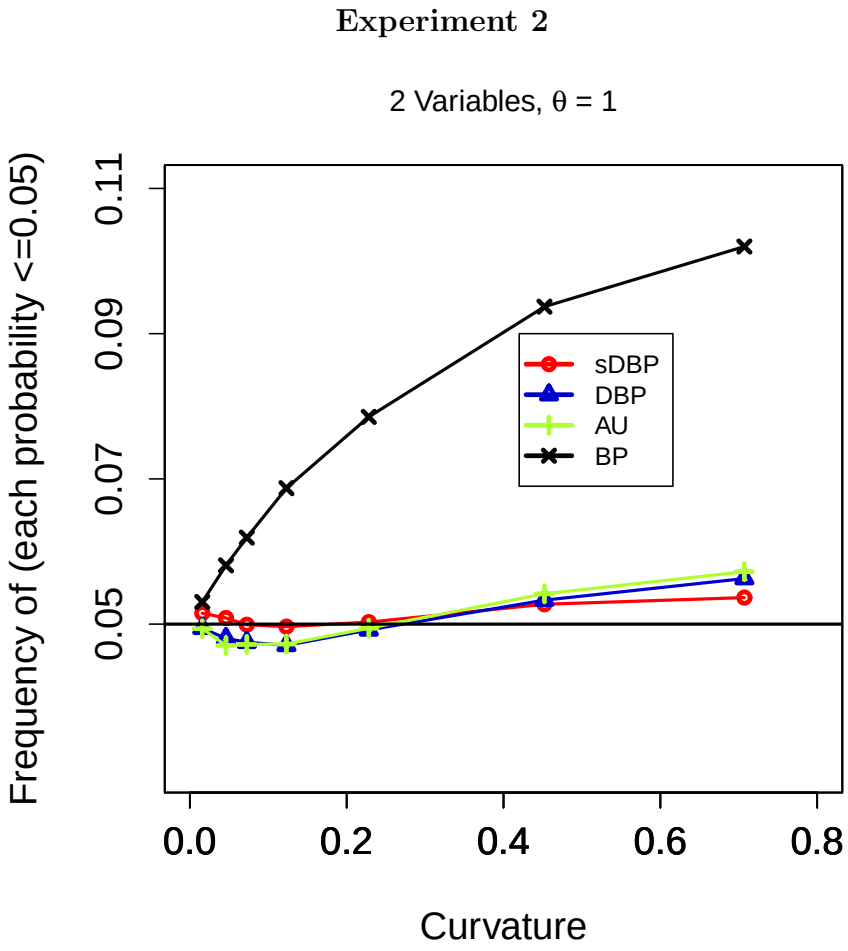

Figure 5: Plots of rejection probability (proportion of probability ≤ 0.05) vs curvature for the four different types of probability (sDBP, DBP, AU and BP), in Experiment 2 (see text for details).

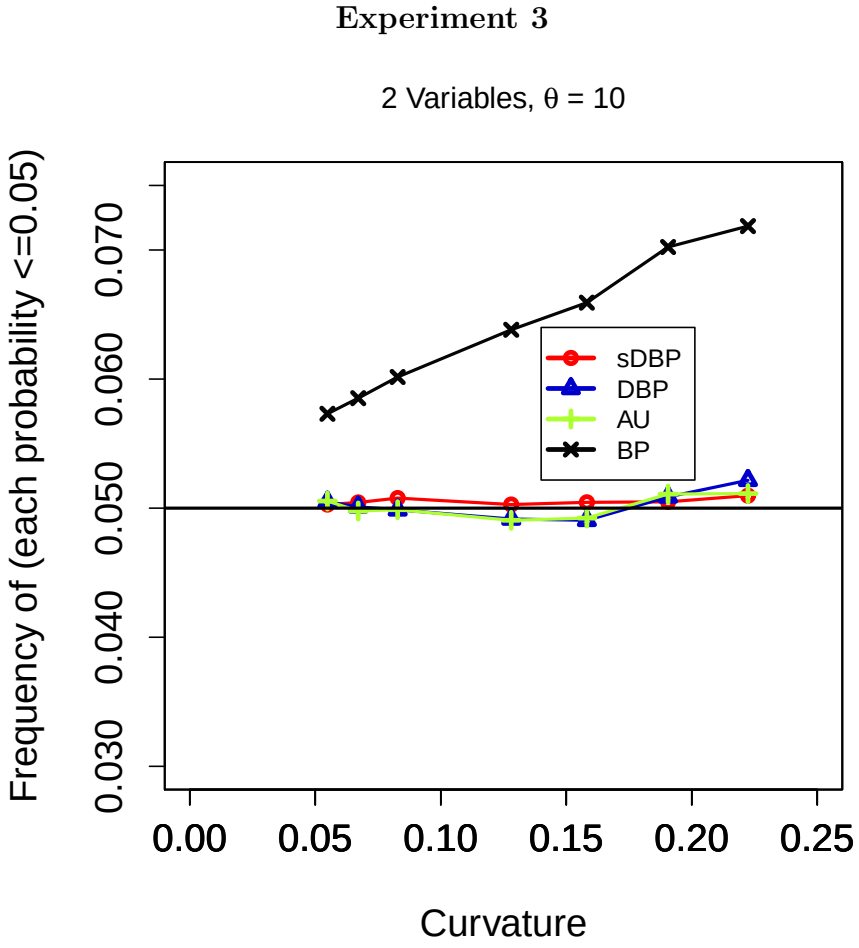

Figure 6: Plots of rejection probability (proportion of probability ≤ 0.05) vs curvature for the four different types of probability (sDBP, DBP, AU and BP), in Experiment 3 (see text for details).

called the rejection probability, and the difference between the rejection probability and 0.05 is known as the bias of the rejection probability.

In each of the three experiments, each probability of (4.10) was paired with its corresponding curvature value and plotted as shown in Figure 4, 5 and 6. The Figure 4, 5 and 6 indicates that almost all cases the third order accurate probabilities (sDBP, DBP and AU) perform substantially better than the traditional bootstrap probability. According to Figure 4, 5 and 6 at bigger curvature values, for example bigger than 0.1, each experiment, sDBP, DBP and AU are similar while BP is different.

4.2 Statistical tests for rejection probability

The performance of the four methods is compared graphically using Figure 4, 5 and 6. We now apply statistical tests to compare the difference in rejection probabilities between the first order accurate BP and the third order accurate methods and between the third order accurate methods. We will explain the statistical tests for Experiment 1 at $\eta_1 = 1$. We denote the rejection probabilities among sDBP, DBP, AU and BP as $\bar{x}_1, \bar{x}_2, \bar{x}_3, \bar{x}_4$ respectively. Because the rejection probability for sDBP was calculated from 10^5 numbers, the 10^5 times rejection probability $\bar{x}_1$ of sDBP follows the binomial distribution with size 10^5 and the probability of success on each trial is μ_{sDBP}, which is true rejection probability of sDBP. The size 10^5 is very large so the rejection probability $\bar{x}_1$ of sDBP approximately follows the normal distribution with mean μ_{sDBP} and deviation $\sqrt{\mu_{sDBP}(1-\mu_{sDBP})/10^5}$. We denote $\mu_{sDBP} = \mu_1, \mu_{DBP} =$

$\mu_2, \mu_{AU} = \mu_3, \mu_{BP} = \mu_4$.

$$\bar{x}_1 \sim N(\mu_1, \mu_1(1-\mu_1)/10^5) \tag{4.11}$$

As with the rejection probability for sDBP, the rejection probabilities among the DBP, AU and BP are approximately normal.

$$\bar{x}_2 \sim N(\mu_2, \mu_2(1-\mu_2)/10^5) \tag{4.12}$$

$$\bar{x}_3 \sim N(\mu_3, \mu_3(1-\mu_3)/10^5)$$

$$\bar{x}_4 \sim N(\mu_4, \mu_4(1-\mu_4)/10^5)$$

The null hypotheses are

$$H_{ij}: \ \mu_i = \mu_j, \text{for } (4,j), j = 1,2,3 \text{ and all combinations } (i,j), i,j = 1,2,3. \tag{4.13}$$

Therefor, the family of subset null hypotheses is

$$\mathcal{F} = \{H_{41}, H_{42}, H_{43}, H_{12}, H_{13}, H_{23}\} \tag{4.14}$$

For null hypotheses H_{ij}, the alternative hypotheses are denoted by

$$H_{ij}^A: \ \mu_i \neq \mu_j, \text{for } (4,j), j = 1,2,3 \text{ and all combinations } (i,j), i,j = 1,2,3. \tag{4.15}$$

Under the null hypothesis H_{ij}, the statistic denoted by t_{ij} is approximately normal.

$$t_{ij} = \frac{\bar{x}_i - \bar{x}_j}{\sqrt{(\bar{x}_i(1-\bar{x}_i) + \bar{x}_j(1-\bar{x}_j))/10^5}} \sim N(0,1), \tag{4.16}$$

where $\sqrt{(\bar{x}_i(1-\bar{x}_i) + \bar{x}_j(1-\bar{x}_j))/10^5}$ is the deviation of sample rather than the population. Thus the half of p-value of the hypothesis H_{ij} can be calculated by $1-\Phi(|t_{ij}|)$, which is denoted by p_{ij}. We take the significance level as 0.05, because we simultaneously do six tests so we adjusted the significance levels by the Holm method. The statistic values are $t_{41} = 43.2422, t_{42} = 45.6055, t_{43} = 42.2985, t_{12} = 2.4532, t_{13} = 0.9760, t_{23} = 3.4290$. The half of p-values in ascending order are $p_{42} = 0.0000, p_{41} = 0.0000, p_{43} = 0.0000, p_{23} = 0.0003, p_{12} = 0.0071, p_{13} = 0.1645$. These values are compared with the half of significance level of $0.05/12, 0.05/10, 0.05/8, 0.05/6, 0.05/4, 0.05/2$ $(0.0042, 0.0050, 0.0063, 0.0083, 0.0125, 0.0250)$, the results are TRUE, TRUE, TRUE, TRUE, TRUE, FALSE where TRUE denote rejected the hypothesis. According to this results, we can say that the difference in rejection probabilities between the first order accurate BP and the third order accurate methods sDBP, DBP and AU is statistically significant at curvature 1. However the difference in rejection probabilities among the DBP and AU, sDBP and DBP are statistically significant. The sDBP and AU are not statistically significant. Based on Figure 4 of experiment 1 at curvature 1, sDBP, DBP, AU are similar while BP is different. On the other hand, at the curvature 1, the deviations of statistics $t_{41}, t_{42}, t_{43}, t_{12}, t_{13}, t_{23}$ are $\sqrt{(\bar{x}_i(1-\bar{x}_i) + \bar{x}_j(1-\bar{x}_j))/10^5}$, with values $0.0012, 0.0012, 0.0012, 0.0010, 0.0010, 0.0010$, respectively. Because these deviations are quite small, they do not sufficiently reflect the result of Figure 4 mentioned above. However, if we assume the deviations of

t_{ij} are larger than the maximum deviation value 0.0012, which is 0.003 for all the deviations at each curvature, using 0.003 we obtain that the difference in rejection probabilities between the BP and the third order accurate methods is statistically significant, and we cannot say that the difference in rejection probabilities among the third order accurate methods are statistically significant for any curvature value (exclude curvature 7, see Table 6). The constant deviation 0.003 is obtained through trial and error.

Table 6: The difference in rejection probabilities between the first order accurate BP and the third order accurate methods (sDBP, DBP, AU) and the difference in rejection probabilities between the third order accurate methods (sDBP, DBP, AU) for experiment 1.

	Curvature 1[a] 0.1970	Curvature 2 0.4525	Curvature 3 0.6711	Curvature 4 1.0289	Curvature 5 1.2807	Curvature 6 1.9056	Curvature 7 2.2361
Statistic: t_{ij} with deviation 0.003[b]	t_{41}=16.8667	20.5912	22.4800	24.2200	24.9633	26.74	27.76
	t_{42}=17.6500	20.6473	22.8400	24.4367	25.4767	26.52	26.96
	t_{43}=16.5500	19.5106	20.8067	22.5367	23.3500	24.47	25.11
	t_{12}=0.7833	0.0561	0.3600	0.2167	0.5133	0.22	0.80
	t_{13}=0.3167	1.0806	1.6733	1.6833	1.6133	2.27	2.65
	t_{23}=1.1000	1.1367	2.0333	1.9000	2.1267	2.05	1.85
Ordered p_{ij}^{c}	p_{42}= 0.0000<0.0042 : TRUE	p_{42}= 0.0000<0.0042 : TRUE	p_{42}= 0.0000<0.0042 : TRUE	p_{42}= 0.0000<0.0042 : TRUE	p_{42}= 0.0000<0.0042 : TRUE	p_{41}= 0.0000<0.0042 : TRUE	p_{41}= 0.0000<0.0042 : TRUE
	p_{41}= 0.0000<0.0050 : TRUE	p_{41}= 0.0000<0.0050 : TRUE	p_{41}= 0.0000<0.0050 : TRUE	p_{41}= 0.0000<0.0050 : TRUE	p_{41}= 0.0000<0.0050 : TRUE	p_{42}= 0.0000<0.0050 : TRUE	p_{42} = 0.0000<0.0050 : TRUE
	p_{43}= 0.0000<0.0063 : TRUE	p_{43}= 0.0000<0.0063 : TRUE	p_{43} = 0.0000<0.0063 : TRUE	p_{43} = 0.0000<0.0063 : TRUE	p_{43} = 0.0000 <0.0063 : TRUE	p_{43} = 0.0000<0.0063 : TRUE	p_{43} = 0.0000<0.0063 : TRUE
	p_{23}= 0.1357<0.0083 : FALSE	p_{23}= 0.1278<0.0083 : FALSE	p_{23} = 0.0210<0.0083 : FALSE	p_{23} = 0.0287<0.0083 : FALSE	p_{23} = 0.0167 <0.0083 : FALSE	p_{13} = 0.0116<0.0083 : FALSE	p_{13} = 0.0040<0.0083 : TRUE
	p_{12}= 0.2167<0.0125 : FALSE	p_{13}= 0.1399<0.0125 : FALSE	p_{13} = 0.0471<0.0125 : FALSE	p_{13} = 0.0462<0.0125 : FALSE	p_{13} = 0.0533 <0.0125 : FALSE	p_{23} = 0.0202<0.0125 : FALSE	p_{23} = 0.0322<0.0125 : FALSE
	p_{13}= 0.3757<0.0250 : FALSE	p_{12}= 0.4776<0.0250 : FALSE	p_{12}= 0.3594<0.0250 : FALSE	p_{12}= 0.4142<0.0250 : FALSE	p_{12}= 0.3039 <0.0250 : FALSE	p_{12}= 0.4129<0.0250 : FALSE	p_{12} = 0.2119<0.0250 : FALSE

[a] Curvatures are numbered in ascending order.
[b] The constant deviation 0.003 is obtained through trial and error.
[c] The half of p-values are in ascending order.

For experiment 2 and experiment 3, we also apply the statistical tests to compare the difference in rejection probabilities between the first order accurate BP and the third order accurate methods and between the third order accurate methods. We also assume all of the deviation of statistic t_{ij} are 0.003. We summarized the tests results at

the Table 7 and Table 8. For experiment 2, at curvatures 3,4,5,6 and 7, the difference

Table 7: The difference in rejection probabilities between the first order accurate BP and the third order accurate methods (sDBP, DBP, AU) and the difference in rejection probabilities between the third order accurate methods (sDBP, DBP, AU) for experiment 2.

	Curvature 1[a] 0.0160	Curvature 2 0.0462	Curvature 3 0.0727	Curvature 4 0.1232	Curvature 5 0.2283	Curvature 6 0.4522	Curvature 7 0.7071
Statistic: t_{ij} with deviation 0.003[b]	t_{41}=0.5167 t_{42}=1.2067 t_{43}=1.2267 t_{12}=0.6900 t_{13}=0.7100 t_{23}=0.0200	2.4167 3.3533 3.6767 0.9367 1.2600 0.3233	4.0000 4.8033 4.9033 0.8033 0.9033 0.1000	6.3567 7.2167 7.1567 0.8600 0.8000 0.0600	9.4233 9.7800 9.7033 0.3567 0.2800 0.0767	13.6633 13.4733 13.1833 0.1900 0.4800 0.2900	16.1167 15.2433 14.9267 0.8733 1.1900 0.3167
Ordered p_{ij}^{c}	p_{43}= 0.1100<0.0042 : FALSE	p_{43}= 0.0001<0.0042 : TRUE	p_{43}= 0.0000<0.0042 : TRUE	p_{42}= 0.0000<0.0042 : TRUE	p_{42}= 0.0000 <0.0042 : TRUE	p_{41}= 0.0000<0.0042 : TRUE	p_{41}= 0.0000 <0.0042 : TRUE
	p_{42}= 0.1138<0.0050 : FALSE	p_{42}= 0.0004<0.0050 : TRUE	p_{42}= 0.0000<0.0050 : TRUE	p_{43}= 0.0000<0.0050 : TRUE	p_{43}= 0.0000 <0.0050 : TRUE	p_{42}= 0.0000<0.0050 : TRUE	p_{42} = <0.0050 : TRUE
	p_{13}= 0.2389<0.0063 : FALSE	p_{41}= 0.0078<0.0063 : FALSE	p_{41} = 0.0000<0.0063 : TRUE	p_{41} = 0.0000<0.0063 : TRUE	p_{41} = 0.0000 <0.0063 : TRUE	p_{43} = 0.0000<0.0063 : TRUE	p_{43} = 0.0000 <0.0063 : TRUE
	p_{12}= 0.2451<0.0083 : FALSE	p_{13}= 0.1038<0.0083 : FALSE	p_{13} = 0.1832<0.0083 : FALSE	p_{12} = 0.1949<0.0083 : FALSE	p_{12} = 0.3607 <0.0083 : FALSE	p_{13} = 0.3156<0.0083 : FALSE	p_{13} = 0.1170<0.0083 : FALSE
	p_{41}= 0.3027<0.0125 : FALSE	p_{12}= 0.1745<0.0125 : FALSE	p_{12} = 0.2109<0.0125 : FALSE	p_{13} = 0.2119<0.0125 : FALSE	p_{13} = 0.3897 <0.0125 : FALSE	p_{23} = 0.3859<0.0125 : FALSE	p_{12} = 0.1912<0.0125 : FALSE
	p_{23}= 0.4920<0.0250 : FALSE	p_{23}= 0.3732<0.0250 : FALSE	p_{23}= 0.4602<0.0250 : FALSE	p_{23}= 0.4761<0.0250 : FALSE	p_{23}= 0.4694 <0.0250 : FALSE	p_{12}= 0.4247<0.0250 : FALSE	p_{23} = 0.3757<0.0250 : FALSE

[a] Curvatures are numbered in ascending order.
[b] The constant deviation 0.003 is obtained through trial and error.
[c] The half of p-values are in ascending order.

in rejection probabilities between the first order accurate BP and the third order accurate methods is statistically significant, and we cannot say that the difference in rejection probabilities among all of the third order accurate methods are statistically significant (see Table 7). At curvature 1, the four methods no statistically significant difference, and at curvature 2, the pair BP and AU and the pair BP and DBP have statistically significant difference, and the pair BP and sDBP and third order accurate methods have no statistical significance. For experiment 3, at curvatures 2,3,4,5,6 and 7, we can say that the difference in rejection probabilities between the first order accurate BP and the third order accurate methods is statistically significant, and we cannot say that the difference in rejection probabilities among the third order

accurate methods are statistically significant. At curvature 1, the four methods have no statistically significant difference.

Table 8: The difference in rejection probabilities between the first order accurate BP and the third order accurate methods (sDBP, DBP, AU) and the difference in rejection probabilities between the third order accurate methods (sDBP, DBP, AU) for experiment 3.

	Curvature 1[a] 0.0548	Curvature 2 0.0671	Curvature 3 0.0828	Curvature 4 0.1281	Curvature 5 0.1582	Curvature 6 0.1906	Curvature 7 0.2224
Statistic: t_{ij} with deviation 0.003[b]	t_{41}=2.3433	2.6933	3.1267	4.5133	5.1567	6.5800	6.9600
	t_{42}=2.2733	2.8067	3.4300	4.8867	5.6200	6.4533	6.5633
	t_{43}=2.2533	2.9133	3.4267	4.9200	5.5633	6.3800	6.9067
	t_{12}=0.0700	0.1133	0.3033	0.3733	0.4633	0.1267	0.3967
	t_{13}=0.0900	0.2200	0.3000	0.4067	0.4067	0.2000	0.0533
	t_{23}=0.0200	0.1067	0.0033	0.0333	0.0567	0.0733	0.3433
Ordered p_{ij}^{c}	p_{41}= 0.0096<0.0042 : FALSE	p_{43}= 0.0018<0.0042 : TRUE	p_{42}= 0.0003<0.0042 : TRUE	p_{43}= 0.0000<0.0042 : TRUE	p_{42}= 0.0000 <0.0042 : TRUE	p_{41}= 0.0000<0.0042 : TRUE	p_{41}= 0.0000<0.0042 : TRUE
	p_{42}= 0.0115<0.0050 : FALSE	p_{42}= 0.0025<0.0050 : TRUE	p_{43}= 0.0003<0.0050 : TRUE	p_{42}= 0.0000<0.0050 : TRUE	p_{43}= 0.0000<0.0050 : TRUE	p_{42}= 0.0000<0.0050 : TRUE	p_{43} = 0.0000<0.0050 : TRUE
	p_{43}= 0.0121<0.0063 : FALSE	p_{41}= 0.0035<0.0063 : TRUE	p_{41} = 0.0009<0.0063 : TRUE	p_{41} = 0.0000<0.0063 : TRUE	p_{41} = 0.0000 <0.0063 : TRUE	p_{43} = 0.0000<0.0063 : TRUE	p_{42} = 0.0000<0.0063 : TRUE
	p_{13}= 0.4641<0.0083 : FALSE	p_{13}= 0.4129<0.0083 : FALSE	p_{12} = 0.3808<0.0083 : FALSE	p_{13} = 0.3421<0.0083 : FALSE	p_{13} = 0.3216 <0.0083 : FALSE	p_{13} = 0.4207<0.0083 : FALSE	p_{12} = 0.3458<0.0083 : FALSE
	p_{12}= 0.4721<0.0125 : FALSE	p_{12}= 0.4549<0.0125 : FALSE	p_{13} = 0.3821<0.0125 : FALSE	p_{12} = 0.3545<0.0125 : FALSE	p_{12} = 0.3421 <0.0125 : FALSE	p_{12} = 0.4496<0.0125 : FALSE	p_{23} = 0.3657<0.0125 : FALSE
	p_{23}= 0.4920<0.0250 : FALSE	p_{23}= 0.4575<0.0250 : FALSE	p_{23}= 0.4987<0.0250 : FALSE	p_{23}= 0.4867<0.0250 : FALSE	p_{23}= 0.4774 <0.0250 : FALSE	p_{23}= 0.4708<0.0250 : FALSE	p_{13} = 0.4787<0.0250 : FALSE

[a] Curvatures are numbered in ascending order.
[b] The constant deviation 0.003 is obtained through trial and error.
[c] The half of p-values are in ascending order.

4.3 Comparisons of computational speed

Original analyses for this part of the study were performed using the software R 2.9.0 on a personal computer with the following specifications: 2.50 GHz CPU (Core (TM) i5-2520M CPU) and 8.00 GB RAM. For the sDBP, DBP, AU and the BP-method, we measured the time taken to calculate a p-value in Experiment 1 (based on curvature value $C7 = 2.2361$). We conducted two separate sets of analyses. In the first set, we applied the sDBP-method with $B1 = 10^4$ pseudoreplicates, the DBP with $B1 = 10^4$ and $B2 = 10^4$ pseudoreplicates, and the BP with 10^4 pseudoreplicates, AU

with 10^4 pseudoreplicates and the scale values are $0.7, 0.8, 1, 1.2, 1.4, 1.6, 1.8, 2$. In the second set, we applied the sDBP-method with $B1 = 10^5$ pseudoreplicates, the DBP with $B1 = 10^5$ and $B2 = 10^5$ pseudoreplicates, and the BP with 10^5 pseudoreplicates, AU with 10^5 pseudoreplicates and the scale values are $0.7, 0.8, 1, 1.2, 1.4, 1.6, 1.8, 2$. Results of the two sets of analyses are shown in Table 9. In both sets the BP-method was the fastest, followed by the AU, sDBP, then the DBP-method. In the first set of calculations (lower numbers of pseudoreplicates) the sDBP-method was 9.2291-fold faster than DBP-method, and this advantage improved substantially in the second set (higher pseudoreplicates), in which the sDBP-method was 63.997-fold faster than DBP-method.

Table 9: Comparison of the BP, DBP, sDBP and AU methods, regarding their speed to compute a p-value for Experiment 1 (based on curvature value $C7 = 2.2361$).

	BP	DBP	sDBP	AU	Speed increase (DBP/sDBP)
Time (secs)[a]	0.30	29.81	3.23	1.31	9.2291-fold
Time (secs)[b]	0.32	3131.37	48.93	1.67	63.997-fold

[a] ($B1 = B2 = 10^4$ pseudoreplicates)
[b] ($B1 = B2 = 10^5$ pseudoreplicates)

Chapter 5

Implementation of speedy double bootstrap method for phylogenetic trees

In this chapter, we develop an easy-to-use R program package for assessing the reliability of estimated phylogenetic trees based on our sDBP-test. Because it is well known that a good statistical method is not in itself sufficient, also need to develop an easy-to-use computer tool. We thus develop the R package named SDBP (available via our website and CRAN, the official R package archive) so that researchers can easily apply the sDBP-test. In Section 5.1, after a brief introduction, we explain the implementation of our package. In Section 5.2, we then describe its usage with the biological data from Section 3.2.

5.1 Introduction

In the phylogenetic tree selection problem, our work has been shown that the sDBP-test has comparable accuracy to the DBP-test and is much more computationally efficient. We thus develop the R package named SDBP, which is an implementation of our sDBP-test on a statistical software R to assess the reliability of phylogenetic trees. We are confident that biologists, who may not have advanced computer skills, will benefit from our sDBP-test and SDBP package.

R is a language and environment for statistical computing and graphics. It is an open-source GNU project based on the S language and environment developed at Bell Laboratories (formerly AT&T, now Lucent Technologies) by John Chambers and colleagues. R can be considered as a different implementation of S. There are some important differences, but much of the code written for S runs under R without alteration. We can summarize why we implemented our method in R as follows. At first, R provides a wide variety of statistical (linear and non-linear modeling, classical statistical tests, time-series analysis, classification, clustering, $\cdots$) and graphical techniques, and is highly extensible. In addition, it is important that R is not only applicable to statistical fields of research, but also to the biological field. Genome analysis, including GneABEL (Aulchenko 2007 [3]), and areas related to biotechnology also have a great many applicable R packages. Finally, R is available under the terms of the Free Software Foundation's GNU General Public License in source code form. It can be compiled and run on a wide variety of UNIX platforms and similar systems (including FreeBSD and Linux), Windows, and MacOS.

5.2 Implementation

5.2.1 Implementation in R

We have implemented the sDBP-test shown in Section 2.3 for phylogenetic trees as a package in the statistical software R. Our package is named `SDBP`, and calculates p-values for phylogenetic trees. It can be used in combination with several other functions or packages in R. Although our sDBP-test was explained in Section 2.3, we should also briefly explain our sDBP-test for phylogenetic trees here. The number of candidate trees is K, and the vector of maximum log-likelihood for each tree is $(l_1, \cdots, l_K)$. The projection of $(l_1, \cdots, l_K)$ for hypothesis $H_1 : \mu_1 = \max_{i=1,\cdots,K} \ \mu_i$ is $(\hat{\mu}_1, \hat{\mu}_2, \cdots, \hat{\mu}_K)$, as obtained by The PAVA method, and the bootstrap replicates of $(\hat{\mu}_1, \hat{\mu}_2, \cdots, \hat{\mu}_K)$ are obtained by sampling from the normal distribution

$$N_K((\hat{\mu}_1, \hat{\mu}_2, \cdots, \hat{\mu}_K)^T, \Sigma) \tag{5.1}$$

where the covariance matrix Σ can be obtained by equation (2.22). The signed distances are defined by equations (2.27) and (2.28). The source code is written in the R language using the S3 object system, and consists of a number of user-level objects: `sdbp`, `sdbpk`, `bpk`, `bp`, `dbpk` and `mam20`. The SDBP provides three types of p-value: the sDBP (speedy double bootstrap probability), the DBP (double bootstrap probability) and the BP (bootstrap probability). The following subsections describe how to use these user-level objects.

5.2.2 Usage – Using the mammalian mitochondrial acid sequences and 12S and 16S rRNA genes for 20 species

In this subsection, we explain how to use our package with the mammalian mitochondrial acid sequences and 12S and 16S rRNA genes data from Section 3.2. This data set is available as supplementary material, for example mam20files folder, from our website

http://www.bi.cs.titech.ac.jp/sdbp/

We used the software package PAML (Yang 1997 [48]), to calculate the site-wise log-likelihood for each tree. The output files are mam20-conc.lnf for the acid sequences, mam20-12SrRNA.lnf for the 12S rRNA genes and mam20-16SrRNA.lnf for the 16S rRNA. The formats of ".lnf" file is not supported by our package, so it was changed using CONSEL by executing the command "seqmt −−paml mam20-conc.lnf". Thus, we obtained the site-wise log-likelihood matrix in a files mam20-conc.mt for each tree. Similarly we also can obtain file mam20-12SrRNA.mt and file mam20-16SrRNA.mt for each tree. For example, the file mam20-conc.mt saved a matrix whose rows are site-wise log-likelihood scores, with each column representing a candidate tree. Thus, we combined the Three matrices in the file ".mt" files in the row direction. We then stored the combined matrix in the mam20.rda file, which is the data file in SDBP package. The mam20files folder also contains the 15 tree topologies. A more detailed explanation of converting ".lnf" to ".mt" can be found in the original CONSEL paper (Shimodaira and Hasegawa 2001 [42]).

Our SDBP package is built under R version 3.0.0. Therefore, this R version (or later) is needed to install our package. For Windows OS, after booting R, choose the tab **Packages** in the upper toolbar and select the tab **Install Package(s) from zip files** option, then choose the **SDBP_1.0.zip** file downloaded from CRAN.

For UNIX OS, installing the source version package **SDBP_1.0.tar.gz** file downloaded from CRAN, just write the following command on the command line

```
R CMD INSTALL SDBP_1.0.tar.gz
```

and boot R via the command line using the command.

```
R
```

Then, the following on the **R console** command line to load our package (the following command can be typed on both Unix and regular Windows machines):

```
library("SDBP") # load our package
```

And then, read the data named mam15.mt then.

```
> data(mam20) # data named mam20 was loaded
> dim(mam20) # matrix demation
[1] 5879   15
```

Calculating the sDBP-value for each tree requires only the following line.

```
> result <- sdbp.default(mam20)
> result
```

We performed this on a personal computer with the following specifications: 2.50 GHz CPU (Core (TM) i5-2520M CPU) and 8.00 GB RAM. The results are output in decreasing order of maximum log-likelihood.

```
Call:
sdbp.default(dat = mam20)
```

```
Speedy double bootstrap probabilities:
t1     t4     t3     t7     t2     t5
0.7503 0.4281 0.3794 0.3338 0.3054 ...
```

```
> summary(result)
```

The output is

```
$Call:
speedy.default(dat = mam20)

$coefficients

   stdErr p.value
t1 0.0043  0.7503
t4 0.0049  0.4281
t3 0.0048  0.3794
t7 0.0047  0.3338
...
attr(,"class")
[1] "summary.sdbp"
```

When we want to calculate the reliability for one tree, for example tree 2, we can use the command `sdbpk`, with the output shown below.

```
> result1 <- sdbpk(mam20,2)
> result1

Call:
sdbpk(dat = mam20, k = 2)

t2
0.3018
```

Then, calculating the bootstrap probability can use the command `bp`, again shown with the output.

```
> result2 <- bp(mam20)

Call:
bp(dat = mam20)
```

```
Bootstrap probabilities:
t1     t4     t3     t7     t2
0.4887 0.1978 0.1128 0.0882 0.0270 ...
```

Then, calculating the bootstrap probability for a single tree can use the command `bpk(mam20)`, and calculating the double bootstrap probability for a single tree can use command `dbpk(mam20)`.

Availability

The program is freely distributed under GNU General Public License (GPL) and can directly installed from CRAN (http://cran.r.-project.org/), the official R package archive. The instruction and program source code are avaliable at http://www.bi.cs.titech.ac.jp/sdbp/.

Chapter 6

Conclusion

In this chapter, we summarize the thesis with some concluding remarks and provide an outlook on possible future directions of research.

6.1 Concluding remarks

Efron and Tibshirani (1994) [10] stated that 'The BP-method is being developed to take advantage of electronic computation in the practical business of statistical inference.' In fact, a great deal of mathematical theory is needed in the development of bootstrap methods so that they are fully compatible with traditional theories of statistical inference. In phylogenetic tree selection problems, the SH-test and AU-test are two such methodologies. Moreover, in Chapter 2, we presented the sDBP-test for assessing the confidence levels of phylogenetic trees, and also developed a DBP-test procedure for comparison. Our sDBP-test provides improvements in accuracy same the DBP-test, and substantial improvements in speed over the DBP-test. For

the first time, this enables the double bootstrap technique to be practically applied in the context of molecular phylogenetics. Our calculations also showed that the application of the sDBP-test is not confined to general tree selection problems; rather, it is appropriate for general model selection problems using the maximum-likelihood criterion.

In Chapter 3, we applied our methods to assess the reliability of phylogenetic trees. The experimental results showed that our sDBP-test has the greatest utility. As a consequence, the sDBP-method suffers no significant loss of accuracy compared with regular DBP-method, and is significantly faster.

In Chapter 4, we presented a novel comparative investigation of the sDBP-method using simulated data. We performed numerical simulations based on the multivariate normal model for four different bootstrap methods (sDBP, DBP, AU and BP). This study is the first to systematically compare these competing bootstrap probability methods via simulations. We found that the rejection probabilities among third-order methods were similar, and that the sDBP-method is apparently faster than the DBP-method.

In Chapter 5, to allow researchers to apply the sDBP-test easily, we developed an easy-to-use R package. We believe this implementation of the sDBP-method will be of further utility in assessing the reliability of phylogenetic trees. In addition, our implementation of sDBP-test does not involve difficult calculations, such as the optimization of non-linear functions necessary for the AU-test.

6.2 Future research

Our calculations show that the sDBP-test is not confined to general tree selection problems, as the algorithm and theory can be used across various fields. Therefore, our work opens the door to many practical model selection problems that use the maximum-likelihood criterion. The procedure can be recognized as similar to sDBP-test in tree selection problems. However, our sDBP-test cannot be adapted to assess the reliability of individual nodes in a phylogenetic tree. For such individual nodes, the problem is that the PAVA method cannot be applied and the signed distance is not well defined. The same problem occurred in the assessment of uncertainties in hierarchical cluster analysis (Suzuki and Shimodaira 2006 [43]).

We want to devolope sDBP-test for exponential family, so that our sDBP-test can be used for more statistical area. And we want to use it for real word problems, for example inference the protein-protein interactions.

Bibliography

[1] J. Adachi and M. Hasegawa. Model of amino acid substitution in proteins encoded by mitochondrial dna. *Journal of Molecular Evolution*, 42(4):459–468, 1996.

[2] Hirotugu Akaike. A new look at the statistical model identification. *Automatic Control, IEEE Transactions on*, 19(6):716–723, 1974.

[3] Yurii S Aulchenko, Stephan Ripke, Aaron Isaacs, and Cornelia M van Duijn. Genabel: an r library for genome-wide association analysis. *Bioinformatics*, 23(10):1294–1296, 2007.

[4] M. Ayer, H.D. Brunk, G.M. Ewing, W.T. Reid, and E. Silverman. An empirical distribution function for sampling with incomplete information. *The Annals of Mathematical Statistics*, pages 641–647, 1955.

[5] Richard E Barlow, David J Bartholomew, JM Bremner, and HD Brunk. *Statistical inference under order restrictions: The theory and application of isotonic regression*. Wiley New York, 1972.

[6] Y. Cao, M. Fujiwara, M. Nikaido, N. Okada, and M. Hasegawa. Interordinal relationships and timescale of eutherian evolution as inferred from mitochondrial genome data. *Gene*, 259(1):149–158, 2000.

[7] B. Efron. Bootstrap confidence intervals for a class of parametric problems. *Biometrika*, 72(1):45–58, 1985.

[8] B. Efron and R. Tibshirani. The problem of regions. *Stanford Technical Report*, 192, 1997.

[9] B. Efron and R. Tibshirani. The problem of regions. *The Annals of Statistics*, 26(5):1687–1718, 1998.

[10] Bradley Efron and Robert Tibshirani. *An introduction to the bootstrap*, volume 57. Chapman & Hall/CRC, 1994.

[11] J. Felsenstein. Evolutionary trees from dna sequences: A maximum likelihood approach. *Journal of Molecular Evolution*, 17(6):368–376, 1981.

[12] J. Felsenstein. Confidence limits on phylogenies: An approach using the bootstrap. *Evolution*, pages 783–791, 1985.

[13] J. Felsenstein. *Inferring Phylogenies*. Sinauer Associates, Sunderland, Massachusetts, 2004.

[14] Joseph Felsenstein. Statistical inference of phylogenies. *Journal of the Royal Statistical Society. Series A (General)*, pages 246–272, 1983.

[15] Shanti S Gupta and Deng-Yuan Huang. Subset selection procedures for the means and variances of normal populations: Unequal sample sizes case. *Sankhyā: The Indian Journal of Statistics, Series B*, pages 112–128, 1976.

[16] P. Hall. *The bootstrap and Edgeworth expansion*. Springer Verlag, New York, 1992.

[17] Edward J Hannan and Barry G Quinn. The determination of the order of an autoregression. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 190–195, 1979.

[18] W Härdle. Robertson, t., wright, ft and rl dykstra: Order restricted statistical inference. *Statistical Papers*, 30(1):316–316, 1989.

[19] D.M. Hillis and J.J. Bull. An empirical test of bootstrapping as a method for assessing confidence in phylogenetic analysis. *Systematic Biology*, 42(2):182–192, 1993.

[20] H. Kishino, T. Miyata, and M. Hasegawa. Maximum likelihood inference of protein phylogeny and the origin of chloroplasts. *Journal of Molecular Evolution*, 31(2):151–160, 1990.

[21] Yusuke Komatsu, Shohei Shimizu, and Hidetoshi Shimodaira. Assessing statistical reliability of lingam via multiscale bootstrap. In *Artificial Neural Networks–ICANN 2010*, pages 309–314. Springer, 2010.

[22] E. L. Lehmann. *Testing statistical hypotheses*. Wiley, 1959.

[23] H. Linhart and W. Zucchini. *Model selection*. Wiley series in probability and mathematical statistics: Applied probability and statistics. Wiley, 1986.

[24] O. Madsen, M. Scally, C.J. Douady, D.J. Kao, R.W. DeBry, R. Adkins, H.M. Amrine, M.J. Stanhope, W.W. de Jong, and M.S. Springer. Parallel adaptive radiations in two major clades of placental mammals. *Nature*, 409(6820):610–614, 2001.

[25] Noboru Murata, Shuji Yoshizawa, and Shun-ichi Amari. Network information criterion-determining the number of hidden units for an artificial neural network model. *Neural Networks, IEEE Transactions on*, 5(6):865–872, 1994.

[26] W.J. Murphy, E. Eizirik, W.E. Johnson, Y.P. Zhang, O.A. Ryder, and S.J. O'Brien. Molecular phylogenetics and the origins of placental mammals. *Nature*, 409(6820):614–618, 2001.

[27] Y. Nagata and M. Yoshida. *Basic of statistical multiple comparison*. scientist-press, 1997 (In Japanese).

[28] F. Pazos and A. Valencia. Similarity of phylogenetic trees as indicator of protein–protein interaction. *Protein Engineering*, 14(9):609–614, 2001.

[29] M.D. Perlman and L. Wu. The emperor ' s new tests. *Statistical Science*, 14(4):355–369, 1999.

[30] M.D. Perlman and L. Wu. On the validity of the likelihood ratio and maximum likelihood methods. *Journal of Statistical Planning and Inference*, 117(1):59–81, 2003.

[31] R Core Team. R: A language and environment for statistical computing. *R Foundation for Statistical Computing, Vienna, Austria*, 2012.

[32] A.Z. Ren, T.S. Ishida, and Y. Akiyama. Assessing statistical reliability of phylogenetic trees via a speedy double bootstrap method. *Molecular Phylogenetics and Evolution*, 67:429–435, 2013.

[33] A.Z. Ren, T.S. Ishida, and Y. Akiyama. Sdbp: An easy-to-use r package for assessing reliability of estimated phylogenetic trees based on the speedy double bootstrap method. *Mathematical Modeling and Problem Solving*, 2013, accepted.

[34] M.J. Sanderson and M.F. Wojciechowski. Improved bootstrap confidence limits in large-scale phylogenies, with an example from neo-astragalus (leguminosae). *Systematic Biology*, 49(4):671–685, 2000.

[35] Gideon Schwarz. Estimating the dimension of a model. *The annals of statistics*, 6(2):461–464, 1978.

[36] Shohei Shimizu, Patrik O Hoyer, Aapo Hyvärinen, and Antti Kerminen. A linear non-gaussian acyclic model for causal discovery. *The Journal of Machine Learning Research*, 7:2003–2030, 2006.

[37] H. Shimodaira. An approximately unbiased test of phylogenetic tree selection. *Systematic Biology*, 51(3):492–508, 2002.

[38] H. Shimodaira and M. Hasegawa. Multiple comparisons of log-likelihoods with applications to phylogenetic inference. *Molecular Biology and Evolution*, 16:1114–1116, 1999.

[39] H. Shimodaira and M. Hasegawa. Assessing the uncertainty in phylogenetic inference. In R. Nielsen, editor, *Statistical Methods In Molecular Evolution: Statistics for Biology and Health*, chapter 17, pages 463–493. Springer, 2005.

[40] H. Shimodaira, H. Ito, and K. Takeuti. *Model selection: Intersection of prediction testing and estimating*. Frontier of Statistical science. Wiley, 2004 (In Japanese).

[41] Hidetoshi Shimodaira. An application of multiple comparison techniques to model selection. *Annals of the Institute of Statistical Mathematics*, 50(1):1–13, 1998.

[42] Hidetoshi Shimodaira and Masami Hasegawa. Consel: for assessing the confidence of phylogenetic tree selection. *Bioinformatics*, 17(12):1246–1247, 2001.

[43] Ryota Suzuki and Hidetoshi Shimodaira. Pvclust: an r package for assessing the uncertainty in hierarchical clustering. *Bioinformatics*, 22(12):1540–1542, 2006.

[44] J.D. Thompson, D.G. Higgins, and T.J. Gibson. Clustal w: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic acids research*, 22(22):4673–4680, 1994.

[45] C.O. Webb, D.D. Ackerly, M.A. McPeek, and M.J. Donoghue. Phylogenies and community ecology. *Annual Review of Ecology and Systematics*, pages 475–505, 2002.

[46] E.O. Wiley. *Phylogenetics: The Theory and Practice of Phylogenetic Systematics*. Wiley-Interscience, New York, 1981.

[47] Z. Yang. Among-site rate variation and its impact on phylogenetic analyses. *Trends in Ecology and Evolution*, 11(9):367–372, 1996.

[48] Z. Yang. Paml: a program package for phylogenetic analysis by maximum likelihood. *Computer Applications in the Biosciences*, 13(5):555–556, 1997.

[49] H.B. Zhao. Comparing several treatments with a control. *Journal of Statistical Planning and Inference*, 137(9):2996–3006, 2007.