Frequentist and Bayesian measures of confidence via multiscale bootstrap for testing three regions

Hidetoshi Shimodaira

Abstract   A new computation method of frequentist p values and Bayesian posterior probabilities based on the bootstrap probability is discussed for the multivariate normal model with unknown expectation parameter vector. The null hypothesis is represented as an arbitrary-shaped region of the parameter vector. We introduce new functional forms for the scaling-law of bootstrap probability so that the multiscale bootstrap method, which was designed for a one-sided test, can also compute confidence measures of a two-sided test, extending applicability to a wider class of hypotheses. Parameter estimation for the scaling-law is improved by the two-step multiscale bootstrap and also by including higher order terms. Model selection is important not only as a motivating application of our method, but also as an essential ingredient in the method. A compromise between frequentist and Bayesian is attempted by showing that the Bayesian posterior probability with a noninformative prior is interpreted as a frequentist p value of "zero-sided" test.

Keywords   Approximately unbiased tests · Bootstrap probability · Bias correction · Hypothesis testing · Model selection · Probability matching priors · Problem of regions · Scaling-law

1 Introduction

Let $Y = (Y_1, \ldots, Y_{m+1})$ be a random vector of dimension $m + 1$ for some integer $m \geq 1$, and $y = (y_1, \ldots, y_{m+1}) \in \mathbb{R}^{m+1}$ be its observed value. Our argument is based on the multivariate normal model with unknown mean vector $\mu = (\mu_1, \ldots, \mu_{m+1}) \in \mathbb{R}^{m+1}$ and covariance identity I_{m+1},

H. Shimodaira (✉)
Department of Mathematical and Computing Sciences, Tokyo Institute of Technology,
2-12-1 Ookayama, Meguro-ku, Tokyo 152-8552, Japan
e-mail: shimo@is.titech.ac.jp

$$Y \sim N_{m+1}(\mu, I_{m+1}), \tag{1}$$

where the probability with respect to (1) will be denoted as $P(\cdot|\mu)$. Let $H_0 \subset \mathbb{R}^{m+1}$ be an arbitrary-shaped region. The subject of this paper is to compute measures of confidence for testing the null hypothesis $\mu \in H_0$. Observing y, we compute a frequentist p value, denoted $p(H_0|y)$, and also a Bayesian posterior probability $\pi(H_0|y)$ with a noninformative prior density $\pi(\mu)$ of μ.

This is the *problem of regions* discussed in literature (Efron et al. 1996; Efron and Tibshirani 1998; Shimodaira 2002, 2004, 2008). The confidence measures were calculated by the bootstrap methods for complicated application problems such as the variable selection of regression analysis and phylogenetic tree selection of molecular evolution. These model selection problems are motivating applications for the issues discussed in this paper, and the normal model of (1) is a simplification of reality. Let $\mathcal{X} = \{x_1, \ldots, x_n\}$ be a sample of size n in application problems. We assume there exists a transformation, depending on n, from $\mathcal{X}$ to y so that Y is approximately normalized. We assume only the existence of such a transformation, and do not have to consider its details. Since we work only on the transformed variable Y in this paper for developing the theory, readers may refer to the literature above for the examples of applications. Before the problem formulation is given in Sect. 2, our methodology is illustrated in simple examples below in this section.

The simplest example of H_0 would be the half space of $\mathbb{R}^{m+1}$,

$$H_0' : \mu_{m+1} \leq 0, \tag{2}$$

where the notation H_0', instead of H_0, is used to distinguish this case from another example given in (3). Only μ_{m+1} is involved in this H_0', and one-dimensional normal model $Y_{m+1} \sim N(\mu_{m+1}, 1)$ is considered. Taking $\mu_{m+1} > 0$ as an alternative hypothesis and denoting the cumulative distribution function of the standard normal as $\Phi(\cdot)$ with density $\phi(\cdot)$, the unbiased frequentist p value is given as $p(H_0'|y) = \Phi(-y_{m+1})$.

A slightly complex example of H_0 is

$$H_0 : -d \leq \mu_{m+1} \leq 0 \tag{3}$$

for $d > 0$. The rejection regions are $y_{m+1} > c$ and $y_{m+1} < -d - c$ with a critical constant c, which is obtained as a solution of the equation

$$\Phi(-c) + \Phi(-d - c) = \alpha \tag{4}$$

for a specified significance level $0 < \alpha < 1$. The left hand side of (4) is the rejection probability $P(Y_{m+1} > c \vee Y_{m+1} < -d - c|\mu)$ when μ is on the boundary of H_0, i.e., $\mu_{m+1} = 0$ or $\mu_{m+1} = -d$. The frequentist p value is defined as the infimum of α such that H_0 can be rejected. This becomes $p(H_0|y) = \Phi(-y_{m+1}) + \Phi(-d - y_{m+1})$ for $y_{m+1} \geq -d/2$ and $p(H_0|y) = \Phi(y_{m+1}) + \Phi(d + y_{m+1})$ for $y_{m+1} \leq -d/2$. For a numerical example

$$d = 1, \quad y_{m+1} = -0.1, \tag{5}$$

we obtain $p(H_0'|y) = 0.540$ and $p(H_0|y) = 0.724$.

These two simple cases of H_0 and H_0' exhibit what Efron and Tibshirani (1998) called *paradox* of frequentist p values. Our simple examples of (2) and (3) suffice for this purpose, although they had actually used the spherical shell example explained later in Sect. 4. Efron and Tibshirani (1998) indicated that a confidence measure should be monotonically increasing in the order of set inclusion of the hypothesis. Noting $H_0 \subset H_0'$, therefore, it should be $p(H_0|y) \le p(H_0'|y)$, but it is not. This kind of "paradox" cannot occur with Bayesian methods, and $\pi(H_0|y) \le \pi(H_0'|y)$ holds always. Considering the flat prior $\pi(\mu) = \text{const}$, say, the posterior distribution of μ given y becomes

$$\mu|y \sim N_{m+1}(y, I_{m+1}), \tag{6}$$

and the posterior probabilities for the case (5) are $\pi(H_0'|y) = \Phi(-y_{m+1}) = 0.540$ and $\pi(H_0|y) = \Phi(-y_{m+1}) - \Phi(-d - y_{m+1}) = 0.356$. The "paradox" of frequentist p values may be nothing surprise for a frequentist statistician, but a natural consequence of the fact that $p(H_0'|y)$ is for a one-sided test and $p(H_0|y)$ is for a two-sided test. The power of testing is higher, i.e., p values are smaller, for an appropriately formulated one-sided test than a two-sided test. In this paper, we do not intend to argue the philosophical question of whether to be frequentist or to be Bayesian, but discuss only computation of these two confidence measures.

Computation of the confidence measures is made by the bootstrap resampling of Efron (1979). Let $\mathcal{X}^* = \{x_1^*, \ldots, x_{n'}^*\}$ be a bootstrap sample of size n' obtained by resampling with replacement from $\mathcal{X}$. The idea of bootstrap probability, which is introduced first by Felsenstein (1985) to phylogenetic inference, is to generate $\mathcal{X}^*$ many times, say B, and count the frequency C that a hypothesis of interest is supported by the bootstrap samples. The bootstrap probability is computed as C/B. Recalling the transformation to get y from $\mathcal{X}$, we get Y^* by applying the same transformation to $\mathcal{X}^*$. For typical problems, the variance of Y^* is approximately proportional to the factor

$$\sigma^2 = \frac{n}{n'}$$

as mentioned in Shimodaira (2008). Although we generate $\mathcal{X}^*$ in practice, we only work on Y^* in this paper. More specifically, we formally consider the parametric bootstrap

$$Y^*|y \sim N_{m+1}(y, \sigma^2 I_{m+1}), \tag{7}$$

which is analogous to (1) but the scale σ is introduced for multiscale bootstrap. The bootstrap probability is defined as

$$\alpha_{\sigma^2}(H_0|y) = P_{\sigma^2}(Y^* \in H_0|y), \tag{8}$$

where $P_{\sigma^2}(\cdot|y)$ denotes the probability with respect to (7). For computing a crude confidence measure, we set $\sigma = 1$, or $n' = n$ in terms of $\mathcal{X}^*$, so that the distribution (7) for Y^* is equivalent to the posterior (6) for μ. Thus the bootstrap probability is interpreted as $\alpha_1(H_0|y) = \pi(H_0|y)$ for any H_0 under the flat prior. In the multiscale bootstrap of Shimodaira (2002, 2004, 2008), however, we may intentionally alter the scale from $\sigma = 1$, or to change n' from n, for computing $p(H_0|y)$. Let $\sigma_1, \ldots, \sigma_M$ be
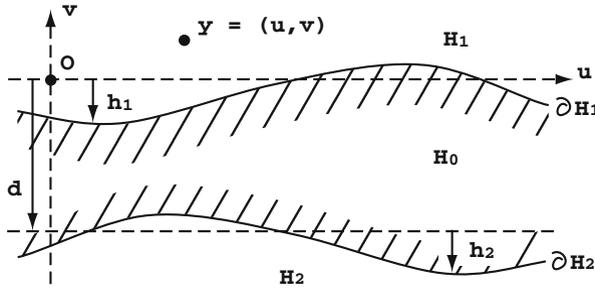
Fig. 1 Region $H_0 \subset \mathbb{R}^{m+1}$ is the *shaded area* between surfaces ∂H_1 and ∂H_2

M different values of scale, which we specify in advance. In our numerical examples, $M = 13$ scales are equally spaced in log-scale between $\sigma_1 = 1/3$ and $\sigma_{13} = 3$. For each $i = 1, \ldots, M$, we generate $\mathcal{X}^*$ with scale σ_i for B_i times, and observe the frequency C_i. The observed bootstrap probability is $\hat{\alpha}_{\sigma_i^2} = C_i/B_i$.

How can we use the observed $\hat{\alpha}_{\sigma_1^2}, \ldots, \hat{\alpha}_{\sigma_M^2}$ for computing $p(H_0|y)$? Let us assume that H_0 can be expressed as (3) but we are unable to observe the values of y_{m+1} and d. Nevertheless, by fitting the model $\alpha_{\sigma^2}(H_0|y) = \Phi(-y_{m+1}/\sigma) - \Phi(-(d + y_{m+1})/\sigma)$ to the observed $\hat{\alpha}_{\sigma_1^2}, \ldots, \hat{\alpha}_{\sigma_M^2}$, we may compute an estimate $\hat{\varphi}$ of the parameter vector $\varphi = (y_{m+1}, d)$ with constraints $d > 0$ and $y_{m+1} > -d/2$. The confidence measures are then computed as $p(H_0|y) = \Phi(-\hat{y}_{m+1}) + \Phi(-\hat{d} - \hat{y}_{m+1})$ and $\pi(H_0|y) = \Phi(-\hat{y}_{m+1}) - \Phi(-\hat{d} - \hat{y}_{m+1})$. In case we are not sure which of (2) and (3) is the reality, we may also fit $\alpha_{\sigma^2}(H_0'|y) = \Phi(-y_{m+1}/\sigma)$ to the observed $\hat{\alpha}_{\sigma_i^2}$'s and compare the AIC values (Akaike 1974) for model selection. In practice, we prepare collection of such models describing the scaling-law of bootstrap probability, and choose the model which minimizes the AIC value.

2 Formulation of the problem

The examples in Sect. 1 were very simple because the boundary surfaces of the regions are flat. In the following sections, we work on generalizations of (2) and (3) by allowing curved boundary surfaces. For convenience, we denote $y = (u, v)$ with $u = (y_1, \ldots, y_m)$ and $v = y_{m+1}$. Similarly, we denote $\mu = (\theta, \mu_{m+1})$ with $\theta = (\mu_1, \ldots, \mu_m) \in \mathbb{R}^m$. As shown in Fig. 1, we consider the region of the form $H_0 = \{(\theta, \mu_{m+1}) \mid -d - h_2(\theta) \leq \mu_{m+1} \leq -h_1(\theta), \theta \in \mathbb{R}^m\}$, where $h_1(\theta)$ and $h_2(\theta)$ are arbitrary functions of θ. This region will reduce to (3) if $h_1(\theta) = h_2(\theta) = 0$ for all θ. The region may be abbreviated as

$$H_0 : -d - h_2(\theta) \leq \mu_{m+1} \leq -h_1(\theta). \tag{9}$$

Two other regions $H_1 : \mu_{m+1} \geq -h_1(\theta)$ and $H_2 : \mu_{m+1} \leq -d - h_2(\theta)$ as well as two boundary surfaces $\partial H_1 : \mu_{m+1} = -h_1(\theta)$ and $\partial H_2 : \mu_{m+1} = -d - h_2(\theta)$ are also shown in Fig. 1. We define $H_0' = H_0 \cup H_2$, or equivalently as

$$H_0' : \mu_{m+1} \leq -h_1(\theta). \tag{10}$$

The boundary surfaces of the hypotheses are $\partial H_0 = \partial H_1 \cup \partial H_2$ for the region H_0, and $\partial H_0' = \partial H_1$ for the region H_0'.

We do not have to specify the functional forms of h_1 and h_2 for our theory, but assume that the magnitude of h_1 and h_2 is very small. Technically speaking, h_1 and h_2 are *nearly flat* in the sense of Shimodaira (2008). Introducing an artificial parameter λ, a function h is called nearly flat when $\sup_{\theta \in \mathbb{R}^m} |h(\theta)| = O(\lambda)$ and L^1-norms of h and its Fourier transform are bounded. We develop asymptotic theory as $\lambda \to 0$, which is analogous to $n \to \infty$ with the relation $\lambda = 1/\sqrt{n}$.

The whole parameter space is partitioned into two regions as $H_0' \cup H_1 = \mathbb{R}^{m+1}$ or three regions as $H_0 \cup H_1 \cup H_2 = \mathbb{R}^{m+1}$. These partitions are treated as disjoint in this paper by ignoring measure-zero sets such as $H_0' \cap H_1 = \partial H_1$. Bootstrap methods for computing frequentist confidence measures in the case of two regions are well developed in the literature as reviewed in Sect. 3. The main contribution of our paper is then given in Sect. 4 for the case of three regions. In Sect. 5, this new computation method is used also for Bayesian measures of Efron and Tibshirani (1998). Note that the flat prior $\pi(\mu) = $ const in the previous section was in fact carefully chosen so that $\pi(H_0'|y) = p(H_0'|y)$ for (2). This same $\pi(\mu)$ led to $\pi(H_0|y) \neq p(H_0|y)$ for (3). Our definition of H_0 given in (9) is a simplest formulation, yet with a reasonable generality for applications, to observe such an interesting difference between the frequentist and Bayesian confidence measures.

Multiscale bootstrap computation of the confidence measures for the three regions case is described in Sect. 6. Simulation study and some discussions are given in Sects. 7 and 8, respectively. Mathematical proofs are mostly given in Appendix.

3 Frequentist measures of confidence for testing two regions

In this section, we review the multiscale bootstrap of Shimodaira (2008) for computing a frequentist p value of "one-sided" test of H_0'. Let $z = -\Phi^{-1}(\alpha)$ be the inverse function of $\alpha = \Phi(-z)$. The bootstrap z value of H_0', defined as $z_{\sigma^2}(H_0'|y) = -\Phi^{-1}(\alpha_{\sigma^2}(H_0'|y))$, is convenient to work with. By multiplying σ to it, $\sigma z_{\sigma^2}(H_0'|y)$ is called the normalized bootstrap z value. Theorem 1 of Shimodaira (2008), as reproduced below, states that the z value of $p(H_0'|y)$ is obtained by extrapolating the normalized bootstrap z value to $\sigma^2 = -1$, or equivalently $n' = -n$ in terms of $\mathcal{X}^*$.

Theorem 1 *Let H_0' be a region of* (10) *with nearly flat h_1. Given H_0' and y, consider the normalized bootstrap z value as a function of σ^2; We denote it by $\psi(\sigma^2) = -\sigma \Phi^{-1}(\alpha_{\sigma^2}(H_0'|y))$. Let us define a frequentist p value as*

$$p(H_0'|y) = \Phi(-\psi(-1)), \tag{11}$$

and assume that the right hand side exists. Then for $\mu \in \partial H_0'$ and $0 < \alpha < 1$,

$$P(p(H_0'|Y) < \alpha|\mu) = \alpha + O(\lambda^3), \tag{12}$$

meaning that the coverage error, i.e., the difference between the rejection probability and α, vanishes asymptotically as λ → 0, and that the p value, or the associated hypothesis testing, is "similar on the boundary" asymptotically.

Proof Here we show only an outline of the proof by allowing the coverage error of $O(\lambda^2)$, instead of $O(\lambda^3)$, in (12). This is a brief summary of the argument given in Shimodaira (2008). First define the expectation operator $\mathcal{E}_{\sigma^2}$ for a nearly flat function h as

$$(\mathcal{E}_{\sigma^2} h)(u) := E_{\sigma^2}(h(U^*)|u),$$

where $E_{\sigma^2}(\cdot)$ on the right hand side denotes the expectation with respect to (7), that is, for $Y^* = (U^*, V^*)$ with

$$U^*|u \sim N_m(u, \sigma^2 I_m), \quad V^*|v \sim N(v, \sigma^2).$$

Using the expectation operator, we next define two quantities

$$z_1 = -\frac{v + \mathcal{E}_{\sigma^2} h_1(u)}{\sigma}, \quad \epsilon_1 = -\frac{h_1(U^*) - \mathcal{E}_{\sigma^2} h_1(u)}{\sigma},$$

and work on the bootstrap probability as

$$\begin{aligned}
\alpha_{\sigma^2}(H_0'|y) &= P_{\sigma^2}(V^* \le -h_1(U^*)|y) \\
&= E_{\sigma^2}\left(\Phi(z_1 + \epsilon_1)|u\right) \\
&= E_{\sigma^2}\left(\Phi(z_1) + \phi(z_1)\epsilon_1|u\right) + O(\lambda^2) \\
&= \Phi(z_1) + O(\lambda^2).
\end{aligned} \tag{13}$$

The third equation is obtained by the Taylor series around z_1, and the last equation is obtained by $E_{\sigma^2}(\epsilon_1|u) = 0$. Rearranging (13), we then get the scaling-law of the normalized bootstrap z value as

$$\psi(\sigma^2) = v + \mathcal{E}_{\sigma^2} h_1(u) + O(\lambda^2). \tag{14}$$

On the other hand, Eq. (5.10) of Shimodaira (2008) shows, by utilizing Fourier transforms of surfaces, that (12) holds with coverage error $O(\lambda^2)$ for a p value defined as

$$p(H_0'|y) = \Phi(-v - \mathcal{E}_{-1} h_1(u)) + O(\lambda^2). \tag{15}$$

The proof completes by combining (14) and (15).   □

A hypothesis testing is to reject H_0' when observing $p(H_0'|y) < \alpha$ for a specified significance level, say, $\alpha = 0.05$, and otherwise not to reject H_0'. The left hand side of (12) is the rejection probability, which should be $\le \alpha$ for $\mu \in H_0'$ and $\ge \alpha$ for $\mu \notin H_0'$ to claim the unbiasedness of the test. On the other hand, the test is called similar on the

boundary when the rejection probability is equal to α for $\mu \in \partial H_0'$. In this paper, we implicitly assume that $p(H_0'|y)$ is decreasing as y moves away from H_0'. The rejection probability increases continuously as μ moves away from H_0'. This assumption is justified when λ is sufficiently small so that the behavior of $p(H_0'|y)$ is not very different from that for (2). Therefore, (12) implies that the p value is approximately unbiased asymptotically as $\lambda \to 0$.

We can think of a procedure for calculating $p(H_0'|y)$ based on (11). In the procedure, the functional form of $\psi(\sigma^2)$ should be estimated from the observed $\hat{\alpha}_{\sigma_i^2}$'s using parametric models. Then an approximately unbiased p value is computed by extrapolating $\psi(\sigma^2)$ to $\sigma^2 = -1$. Our procedure works fine for the particular H_0' of (2), because $\psi(\sigma^2) = y_{m+1}$ and $p(H_0'|y) = \Phi(-y_{m+1}) = \Phi(-\psi(-1))$. Our procedure works fine also for any H_0' of (10) when the boundary surface $\partial H_0'$ is smooth. The model is given as $\psi(\sigma^2) = \beta_0 + \beta_1 \sigma^2 + \beta_2 \sigma^4 + \beta_3 \sigma^6 + \cdots$ using parameters $\varphi = (\beta_0, \beta_1, \ldots)$, and thus an approximately unbiased p value can be computed by $p(H_0'|y) = \Phi(-\beta_0 + \hat{\beta}_1 - \hat{\beta}_2 + \hat{\beta}_3 - \cdots)$. It may be interesting to know that the parameters are interpreted as geometric quantities; β_0 is the distance from y to the surface $\partial H_0'$, β_1 is the mean curvature of the surface, and β_j, $j \geq 2$, is related to $2j$th derivatives of h_1.

However, the series expansion above does not converge, i.e., $\psi(-1)$ does not exist, when $\partial H_0'$ is nonsmooth. For example, $\psi(\sigma^2) = \beta_0 + \beta_1 \sqrt{\sigma^2}$ serves as a good approximating model for cone-shaped H_0', for which $\psi(-1)$ does not take a value of $\mathbb{R}$. This observation agrees with the fact that an unbiased test does not exist for cone-shaped H_0' as indicated in the argument of Lehmann (1952). Instead of (11), the modified procedure of Shimodaira (2008) calculates a p value defined as

$$p_k(H_0'|y) = \Phi \left\{ -\sum_{j=0}^{k-1} \frac{(-1 - \sigma_0^2)^j}{j!} \frac{\partial^j \psi(\sigma^2)}{\partial (\sigma^2)^j} \Big|_{\sigma_0^2} \right\} \tag{16}$$

for an integer $k > 0$ and a real number $\sigma_0^2 > 0$. This is to extrapolate $\psi(\sigma^2)$ back to $\sigma^2 = -1$ by using the first k terms of the Taylor series around σ_0^2. The coverage error in (12) should reduce as k increases, but then the rejection region violates the desired property called monotonicity in the sense of Lehmann (1952) and Perlman and Wu (1999, 2003). For taking the balance, we chose $k = 3$ and $\sigma_0^2 = 1$ for numerical examples in this paper.

4 Frequentist measures of confidence for testing three regions

The following theorem is our main result for computing a frequentist p value of "two-sided" test of H_0. The proof is given in Appendix A.1.

Theorem 2 *Let H_0 be a region of* (9) *with nearly flat h_1 and h_2. Given H_0 and y, consider the approximately unbiased p value $p(H_i|y)$ by applying Theorem 1 to H_i for $i = 1, 2$. Assuming these two p values exist, let us define a frequentist p value of H_0 as*
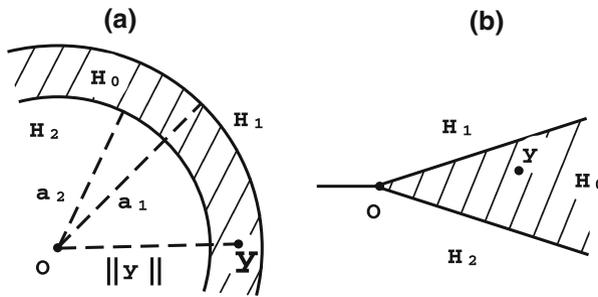
Fig. 2 **a** Spherical shell region. **b** Cone-shaped region (Sect. 7)

$$p(H_0|y) = 1 - |p(H_1|y) - p(H_2|y)|. \tag{17}$$

For example, (17) holds for the exact p value of (3) defined in Sect. 1. Then for $\mu \in \partial H_0 = \partial H_1 \cup \partial H_2$ and $0 < \alpha < 1$,

$$P(p(H_0|Y) < \alpha|\mu) = \alpha + O(\lambda^2), \tag{18}$$

meaning that $p(H_0|y)$ is approximately unbiased asymptotically as $\lambda \to 0$.

For illustrating the methodology, let us work on the spherical shell example of Efron and Tibshirani (1998), for which we can still compute the exact p values to verify our methods. The region of interest is $H_0 : a_2 \leq \|\mu\| \leq a_1$ as shown in Fig. 2a. We consider the case, say,

$$m + 1 = 4, \quad a_1 = 6, \quad a_2 = 5, \quad \|y\| = 5.9,$$

so that this region is analogous to (5) except for the curvature. The exact p value for $H_1 :$ $\|\mu\| \geq a_1$ is easily calculated knowing that $\|Y\|^2$ is distributed as the chi-square distribution with degrees of freedom $m+1$ and noncentrality $\|\mu\|^2$. Writing this random variable as $\chi^2_{m+1}(\|\mu\|^2)$, the exact p value is $p(H_1|y) = P(\chi^2_{m+1}(a_1^2) \leq \|y\|^2) = 0.362$, that is, the probability of observing $\|Y\| \leq \|y\|$ for $\|\mu\| = a_1$. On the other hand, the exact p value for $H_2 : \|\mu\| \leq a_2$ is $p(H_2|y) = P(\chi^2_{m+1}(a_2^2) \geq \|y\|^2) = 0.267$. In a similar way as for (3), the exact p value for H_0 is computed numerically as $p(H_0|y) = 0.907$, although the procedure is a bit complicated as explained below. We first consider the critical constants c_1 and c_2 for the rejection regions $R_1 = \{y \mid \|y\| < a_1 - c_1\}$ and $R_2 = \{y \mid \|y\| > a_2 + c_2\}$. By equating the rejection probability to α for $\mu \in \partial H_0$, that is, $P(\chi^2_{m+1}(a_i^2) < (a_1 - c_1)^2) + P(\chi^2_{m+1}(a_i^2) > (a_2 + c_2)^2) = \alpha$ for $i = 1, 2$, we may get the solution numerically as $c_1 = 1.331$ and $c_2 = 1.903$ for $\alpha = 0.05$, say. The p value is defined as the infimum of α such that H_0 can be rejected.

We first check if Theorem 2 is ever usable by computing (17) with the exact values of $p(H_1|y)$ and $p(H_2|y)$. This gives $p(H_0|y) = 1 - (0.362 - 0.267) = 0.905$, which agrees extremely well to the exact $p(H_0|y) = 0.907$. The spherical shell is approximated by (9) only locally in a neighborhood of y. Nevertheless, Theorem 2 worked fine.

We next think of the situation that bootstrap probabilities of H_1 and H_2 are available but not their exact p values. We apply the procedure of Sect. 3 separately to the two regions for calculating the approximately unbiased p values. A simple model $\psi(\sigma^2) = \beta_0 + \beta_1\sigma^2$ with parameters $\varphi = (\beta_0, \beta_1)$ is considered for

$$\alpha_{\sigma^2}(H_0'|y) = \Phi(-\psi(\sigma^2)/\sigma). \tag{19}$$

Let $\psi_i(\sigma^2)$ be the normalized bootstrap z value of H_i for $i = 1, 2$. By assuming the simple model for $\psi_i(\sigma^2)$, we fit $\alpha_{\sigma^2}(H_i|y) = \Phi(-\psi_i(\sigma^2)/\sigma)$ to the observed multiscale bootstrap probabilities of H_i for estimating the parameters. The actual estimation was done using the method described in Sect. 6.3, but we would like to forget the details for the moment. We get $\hat{\beta}_0 = 0.101$, $\hat{\beta}_1 = -0.258$ for H_1, and similarly $\hat{\beta}_0 = 0.889$, $\hat{\beta}_1 = 0.286$ for H_2. β_0's are interpreted as the distances from y to the boundary surfaces, and the estimates agree well to the exact values $\beta_0 = 0.1$ for H_1 and $\beta_0 = 0.9$ for H_2. Using these estimated values, the approximately unbiased p values are computed by (11) as $p(H_1|y) = \Phi(-0.101 - 0.258) = 0.360$ and $p(H_2|y) = \Phi(-0.889 + 0.286) = 0.273$, and thus (17) gives $p(H_0|y) = 1 - (0.360 - 0.273) = 0.913$, which again agrees well to the exact $p(H_0|y) = 0.907$.

We finally think of a more practical situation, where the bootstrap probabilities are not available for H_1 and H_2, but only for H_0. This situation is plausible in applications where many regions are involved and we are not sure which of them can be treated as H_1 or H_2 in a neighborhood of y (see Efron et al. 1996) for an illustration. We consider a simple model $\psi_1(\sigma^2) = \beta_0 + \beta_1\sigma^2$, $\psi_2(\sigma^2) = d - \beta_0 - \beta_1\sigma^2$ with parameters $\varphi = (\beta_0, \beta_1, d)$ for

$$\alpha_{\sigma^2}(H_0|y) = 1 - (\Phi(-\psi_1(\sigma^2)/\sigma) + \Phi(-\psi_2(\sigma^2)/\sigma)) \tag{20}$$

by assuming that the two surfaces are curved in the same direction with the same magnitude of curvature $|\beta_1|$. For estimating φ, (20) is fitted to the observed multiscale bootstrap probabilities of H_0 with constraints $\beta_0 < d/2$ and $d > 0$, and $\hat{\varphi}$ is obtained as $\hat{\beta}_0 = 0.089$, $\hat{\beta}_1 = -0.199$, $\hat{d} = 0.995$. Using these estimated values, the approximately unbiased p values are computed by (11) as $p(H_1|y) = \Phi(-0.089 - 0.199) = 0.387$ and $p(H_2|y) = \Phi(-0.995 + 0.089 + 0.199) = 0.240$ and thus (17) gives $p(H_0|y) = 1 - (0.387 - 0.240) = 0.853$. This is not very close to the exact $p(H_0|y) = 0.907$, partly because the model is too simple. However, it is a great improvement over $\alpha_1(H_0|y) = P(a_1^2 \leq \chi_{m+1}^2(\|y\|^2) \leq a_2^2) = 0.320$.

5 Bayesian measures of confidence

Choosing a good prior density is essential for Bayesian inference. We consider a version of noninformative prior with which the posterior probability behaves similarly as a frequentist p value.

First note that the sum of bootstrap probabilities of disjoint partitions of the whole parameter space is always 1. For the two regions case, $\alpha_{\sigma^2}(H_0'|y) + \alpha_{\sigma^2}(H_1|y) = 1$,

and thus $\sigma z_{\sigma^2}(H_0'|y) + \sigma z_{\sigma^2}(H_1|y) = 0$. Therefore $p(H_0'|y) + p(H_1|y) = 1$ for the approximately unbiased p values computed by (11), suggesting that we may think of a prior so that $p(H_0'|y) = \pi(H_0'|y)$. This was the idea of Efron and Tibshirani (1998) to define a Bayesian measure of confidence of H_0. Since each of H_1 and H_2 can be treated as H_0' by changing the coordinates, we may assume a prior satisfying

$$\pi(H_i|y) = p(H_i|y), \quad i = 1, 2. \tag{21}$$

It follows from $\sum_{i=0}^2 \pi(H_i|y) = 1$ that

$$\pi(H_0|y) = 1 - (p(H_1|y) + p(H_2|y)). \tag{22}$$

Priors satisfying (21) are called probability matching priors. The theory has been developed in literature (Peers 1965; Tibshirani 1989; Datta and Mukerjee 2004) for posterior quantiles of a single parameter of interest. The examples are the flat prior $\pi(\mu) = $ const for the flat boundary case in Sect. 1, and $\pi(\mu) \propto \|\mu\|^{-m}$ for the spherical shell case in Sect. 4.

Our multiscale bootstrap method provides a new computation to $\pi(H_0|y)$. We may simply compute (22) with the $p(H_1|y)$ and $p(H_2|y)$ used for computing $p(H_0|y)$ of (17). Although we implicitly assumed the matching prior, we do not have to know the functional form of $\pi(\mu)$. For the spherical shell example, we may use the exact $p(H_1|y)$ and $p(H_2|y)$ to get $p(H_0|y) = 1 - (0.362 + 0.267) = 0.371$, or more practically, use only bootstrap probabilities of H_0 to get $p(H_0|y) = 1 - (0.387 + 0.240) = 0.373$.

6 Estimating parametric models for the scaling-law of bootstrap probabilities

6.1 One-step multiscale bootstrap

We first recall the estimation procedure of Shimodaira (2002, 2008) before describing our new proposals for improving the estimation accuracy in the following sections.

Let $f(\sigma^2|\varphi)$ be a parametric model of bootstrap probability such as (19) for H_0' or (20) for H_0. As already mentioned in Sect. 1, the model is fitted to the observed C_i/B_i, $i = 1, \ldots, M$. Since C_i is distributed as binomial with probability $f(\sigma_i^2|\varphi)$ and B_i trials, the log-likelihood function is $\ell(\varphi) = \sum_{i=1}^M \{C_i \log f(\sigma_i^2|\varphi) + (B_i - C_i) \log(1 - f(\sigma_i^2|\varphi))\}$. The maximum likelihood estimate $\hat{\varphi}$ is computed numerically for each model. Let $\dim \varphi$ denote the number of parameters. Then $AIC = -2\ell(\hat{\varphi}) + 2\dim \varphi$ may be compared for selecting a best model among several candidate models.

6.2 Two-step multiscale bootstrap

Shimodaira (2004) has devised a generalization of the multiscale bootstrap. This is called multistep-multiscale bootstrap, and the usual multiscale bootstrap is called one-step multiscale bootstrap. Our new proposal here is to utilize the two-step multiscale bootstrap for improving the estimation accuracy of φ, although the two-step method

was originally used for replacing the normal model of (1) with the exponential family of distributions.

Recalling that $\mathcal{X}^*$ is obtained by resampling from $\mathcal{X}$, we may resample again from $\mathcal{X}^*$, instead of $\mathcal{X}$, to get a bootstrap sample of size n'', and denote it as $\mathcal{X}^{**} = \{x_1^{**}, \ldots, x_{n''}^{**}\}$. We formally consider the parametric bootstrap

$$Y^{**}|y^* \sim N_{m+1}(y^*, (\tau^2 - \sigma^2)I_{m+1}),$$

where τ is a new scale defined by $\tau^2 - \sigma^2 = n/n''$. We denote the probability and the expectation by $P_{\sigma^2,\tau^2}(\cdot|y)$ and $E_{\sigma^2,\tau^2}(\cdot|y)$, respectively. In Shimodaira (2004), only the marginal distribution $Y^{**}|y \sim N_{m+1}(y, \tau^2 I_{m+1})$ is considered to detect the nonnormality, utilizing the fact that $P_{\sigma^2,\tau^2}(Y^{**} \in H_0|y) = \alpha_{\tau^2}(H_0|y)$ should have the same functional form as $P_{\sigma^2,\tau^2}(Y^* \in H_0|y) = \alpha_{\sigma^2}(H_0|y)$ for the normal model. Here we also consider the joint distribution of (Y^*, Y^{**}) given y. It is $2m + 2$-dimensional multivariate normal with $Cov(Y^*, Y^{**}|y) = \sigma^2 I_{m+1}$. Then, the joint bootstrap probability is defined as

$$\alpha_{\sigma^2,\tau^2}(H_0|y) = P_{\sigma^2,\tau^2}(Y^* \in H_0 \wedge Y^{**} \in H_0|y).$$

Note that only one Y^{**} is generated from each Y^* here, whereas thousands of Y^{**}'s may be generated from each Y^* in the double bootstrap method.

Let $g(\sigma^2, \tau^2|\varphi)$ be a parametric model of $\alpha_{\sigma^2,\tau^2}(H_0'|y)$ or $\alpha_{\sigma^2,\tau^2}(H_0|y)$. To work on specific forms of $g(\sigma^2, \tau^2|\varphi)$, we need some notations. Let (X', X'') be distributed as bivariate normal with mean $(0, 0)$, variance $V(X') = V(X'') = 1$, and covariance $Cov(X', X'') = \rho$. The distribution function is denoted as $\Phi_\rho(a_1, b_1) = P(X' \leq a_1 \wedge X'' \leq b_1)$, where the joint density is explicitly given as $\phi_\rho(a_1, b_1) = (1 - \rho^2)^{-1/2}\phi((1-\rho^2)^{-1/2}(b_1 - \rho a_1))\phi(a_1)$. We also define $\Phi_\rho(a_1, b_1; a_2, b_2) = P(a_2 \leq X' \leq a_1 \wedge b_2 \leq X'' \leq b_1) = \Phi_\rho(a_1, b_1) - \Phi_\rho(a_2, b_1) - \Phi_\rho(a_1, b_2) + \Phi_\rho(a_2, b_2)$. Then a generalization of (14) is given as follows. The proof is in Appendix A.2.

Lemma 1 *For sufficiently small λ, the joint bootstrap probabilities for H_0' and H_0 are expressed asymptotically as*

$$\alpha_{\sigma^2,\tau^2}(H_0'|y) = \Phi_\rho(z_1, w_1) + O(\lambda^2), \tag{23}$$

$$\alpha_{\sigma^2,\tau^2}(H_0|y) = \Phi_\rho(z_1, w_1; z_2, w_2) + O(\lambda^2), \tag{24}$$

where $z_1 = -(v + \mathcal{E}_{\sigma^2}h_1(u))/\sigma$, $w_1 = -(v + \mathcal{E}_{\tau^2}h_1(u))/\tau$, $z_2 = -(v + d + \mathcal{E}_{\sigma^2}h_2(u))/\sigma$, $w_2 = -(v + d + \mathcal{E}_{\tau^2}h_2(u))/\tau$, and $\rho = \sigma/\tau$.

Thus $g(\sigma^2, \tau^2|\varphi)$ is specified for H_0' as (23) with $z_1 = -\psi(\sigma^2)/\sigma$, $w_1 = -\psi(\tau^2)/\tau$ using the ψ function of (19). Similarly, $g(\sigma^2, \tau^2|\varphi)$ is specified for H_0 as (24) with $z_1 = \psi_1(\sigma^2)/\sigma$, $w_1 = \psi_1(\tau^2)/\tau$, $z_2 = -\psi_2(\sigma^2)/\sigma$, $w_2 = -\psi_2(\tau^2)/\tau$ using ψ_1 and ψ_2 functions of (20).

We may specify M sets of (σ, τ), denoted as $(\sigma_1, \tau_1), \ldots, (\sigma_M, \tau_M)$. In our numerical examples, $\sigma_1, \ldots, \sigma_{13}$ are specified as mentioned in Sect. 1 and τ_i's are specified

so that $\tau_i^2 - \sigma_i^2 = 1$ holds always, meaning $n'' = n$. For each $i = 1, \ldots, M$, we generate (Y^*, Y^{**}) with (σ_i, τ_i) many times, say $B_i = 10,000$, and observe the frequencies $C_i = \#(Y^* \in H_0)$, $D_i = \#(Y^{**} \in H_0)$, and $E_i = \#(Y^* \in H_0 \wedge Y^{**} \in H_0)$. The log-likelihood function becomes $\ell(\varphi) = \sum_{i=1}^{M} \{E_i \log g(\sigma_i^2, \tau_i^2 | \varphi) + (C_i - E_i) \log (f(\sigma_i^2 | \varphi) - g(\sigma_i^2, \tau_i^2 | \varphi)) + (D_i - E_i) \log(f(\tau_i^2 | \varphi) - g(\sigma_i^2, \tau_i^2 | \varphi)) + (B_i - C_i - D_i + E_i) \log(1 - f(\sigma_i^2 | \varphi) - f(\tau_i^2 | \varphi) + g(\sigma_i^2, \tau_i^2 | \varphi))\}$. We have used this two-step multiscale bootstrap, instead of the one-step method, in all the numerical examples. The one-step method had difficulty in distinguishing H_0 with very small d from H_0 with moderate d but heavily curved ∂H_1. The two-step method avoids this identifiability issue because a small value of E_i indicates that d is small.

6.3 Higher-order terms of bootstrap probabilities for testing two regions

The asymptotic errors of the scaling law of the bootstrap probabilities in (13) and (23) are of order $O(\lambda^2)$. As shown in the following lemma, the errors can be reduced to $O(\lambda^3)$ by introducing correction terms of order $O(\lambda^2)$ for improving the parametric model $g(\sigma^2, \tau^2 | \varphi)$ of H_0'. The proof is given in Appendix A.3.

Lemma 2 *For sufficiently small λ, the bootstrap probabilities for H_0' are expressed asymptotically as*

$$\alpha_{\sigma^2}(H_0' | y) = \Phi(z_1 + \Delta z_1) + O(\lambda^3) \tag{25}$$

$$\alpha_{\tau^2}(H_0' | y) = \Phi(w_1 + \Delta w_1) + O(\lambda^3) \tag{26}$$

$$\alpha_{\sigma^2, \tau^2}(H_0' | y) = \Phi_{\rho + \Delta \rho}(z_1 + \Delta z_1, w_1 + \Delta w_1) + O(\lambda^3), \tag{27}$$

where z_1, w_1, and ρ are those defined in Lemma 1, and the higher order correction terms are defined as $\Delta z_1 = -\frac{1}{2} z_1 E_{\sigma^2, \tau^2}(\epsilon_1^2 | u)$, $\Delta w_1 = -\frac{1}{2} w_1 E_{\sigma^2, \tau^2}(\delta_1^2 | u)$, and $\Delta \rho = -\frac{1}{2} \left(\rho E_{\sigma^2, \tau^2}(\epsilon_1^2 | u) + \rho E_{\sigma^2, \tau^2}(\delta_1^2 | u) - 2 E_{\sigma^2, \tau^2}(\epsilon_1 \delta_1 | u) \right)$ using

$$\epsilon_1 = -\frac{h_1(U^*) - \mathcal{E}_{\sigma^2} h_1(u)}{\sigma}, \quad \delta_1 = -\frac{h_1(U^{**}) - \mathcal{E}_{\tau^2} h_1(u)}{\tau}. \tag{28}$$

For deriving a very simple model for $\Delta \rho$, we think of a situation $h(u) = (A/\sqrt{m}) \|u\| + (B/m) \|u\|^2$ and $\theta = 0$, and consider asymptotics as $m \to \infty$. This formulation is only for convenience of derivation. The two values A and B will be specified later by looking at the functional form of $f(\sigma^2 | \varphi)$. A straightforward, yet tedious, calculation (the details are not shown) gives $\psi(\sigma^2) = \text{const} + A\sigma + B\sigma^2 + O(m^{-1})$ and

$$\Delta \rho = -\frac{1}{2m} \left(A^2 \rho(1 - \rho) + 2B^2 \rho(\tau^2 - \sigma^2) + 2AB\sigma(1 - \rho^2) \right) + O(m^{-3/2}).$$

This correction term was in fact already used for the simple model $\psi(\sigma^2) = \beta_0 + \beta_1 \sigma^2$ of the spherical shell example in Sect. 4, where the parameter was actually $\varphi = (\beta_0, \beta_1, m)$ instead of $\varphi = (\beta_0, \beta_1)$. We did not change the $\psi(\sigma^2)$ for adjusting

Δz_1 and Δw_1, meaning that $z_1 + \Delta z_1$, instead of z_1, was modeled as $-\psi(\sigma^2)/\sigma$. Comparing the coefficients of $\psi(\sigma^2)$, we get $A = 0$ and $B = \beta_1$, and thus $\Delta \rho = -(\beta_1)^2(\sigma/\tau)(\tau^2 - \sigma^2)/m$. When (19) was fitted to H_1, the estimated parameter $\hat{m} = 2.83$ was close to the true value $m = 3$.

For the numerical example mentioned above, we have also fitted the same model but $\Delta \rho = 0$ being fixed. The estimated parameters are $\hat{\beta}_0 = 0.101$, $\hat{\beta}_1 = -0.256$, and the p value is $p(H_1|y) = \Phi(-0.101 - 0.256) = 0.361$. These values are not much different from those shown in Sect. 4. However, the AIC value improved greatly by the introduction of $\Delta \rho$, and the AIC difference was 96.67, mostly because improved fitting for the joint bootstrap probability of (27). My experience suggests that consideration of the $\Delta \rho$ term is useful for choosing a reasonable model of $\psi(\sigma^2)$.

7 Simulation study

7.1 Overview and results

Let us consider a cone-shaped region H_0 in $\mathbb{R}^2$ with the angle at the vertex being $2\pi/10$ as shown in Fig. 2b. This cone can be interpreted, locally in a neighborhood of y with appropriate coordinates, as H_0 of (9) when y is close to one of the edges but far from the vertex. The same cone can also be interpreted as H_0' of (10) when y is close to the vertex. In this section, the cone is labeled either by H_0 or H_0' depending on which view we are taking.

Cones in $\mathbb{R}^2$ appear in the problem of multiple comparisons of three elements X_0, X_1, X_2, say, and H_i corresponds to the hypothesis that the mean of X_i is the largest among the three (DuPreez et al. 1985; Perlman and Wu 2003; Shimodaira 2008). The angle at the vertex is related to the covariance structure of the elements. Although an unbiased test does not exist for this region, we would like to see how our methods work for reducing the coverage error.

Contour lines at the levels 0.05 and 0.95 are drawn in Fig. 3 for several types of confidence measures. We use symbol $p(y)$ in general for the value of a confidence measure at y. The rejection region of the cone and that of the complement of the cone are $R = \{y|p(y) < 0.05\}$ and $R' = \{y|p(y) > 0.95\}$, respectively, at $\alpha = 0.05$. We observe that $p(y)$ decreases as y moves away from the cone in Fig. 3a–c. On the other hand, Figs. 4 and 5 show the rejection probability. For an unbiased test, it should be 5% for all $\mu \in \partial H_0$ so that the coverage error is zero.

In Fig. 3a, $p(y) = \alpha_1(H_0|y)$ is computed by the bootstrap samples of $\sigma^2 = 1$. This bootstrap probability, labeled as BP in Fig. 4, is heavily biased near the vertex, and this tendency is enhanced when the angle becomes $2\pi/20$ in Fig. 5.

In Fig. 3b, $p(y) = p(H_0'|y)$ is computed by regarding the cone as H_0' of (10). The dent of R and the bump of R' become larger than those of Fig. 3a near the vertex, confirming what we observed in Fig. 4 of Shimodaira (2008). As seen in Figs. 4 and 5, the coverage error of $p(H_0'|y)$, labeled as "one sided" there, is smaller than that of BP.

In Fig. 3c, $p(H_0|y)$ is also computed by regarding the cone as H_0 of (9), and then one of $p(H_0'|y)$ and $p(H_0|y)$ is selected as $p(y)$ by comparing the AIC values at each y. This $p(y)$, labeled as "two sided Freq" in Figs. 4 and 5, improves greatly on the

(a)

(b)

(c)
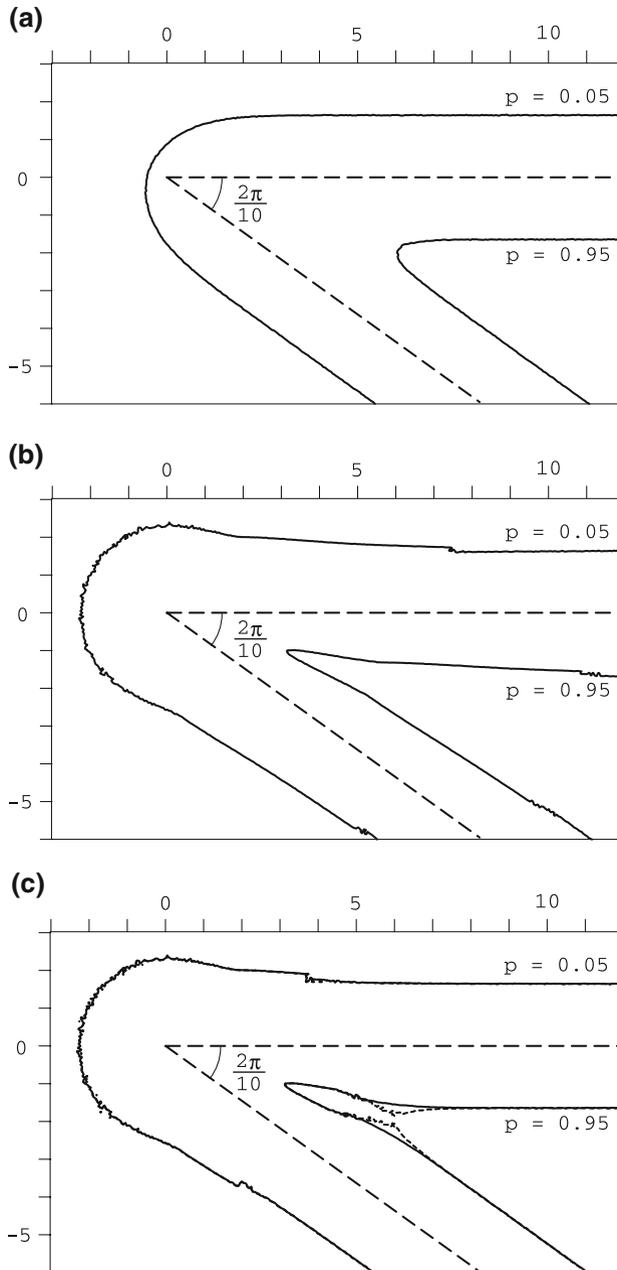
Fig. 3 *Contour lines* $p(y) = 0.05$ and $p(y) = 0.95$. The cone-shaped region H_0 is rotated so that one of the edges is placed along the x-axis. *Solid curves* are drawn for **a** the bootstrap probability with $\sigma^2 = 1$, and for **b** the frequentist p value for "one-sided" test. In **c**, $p(y)$ is switched to the frequentist p value for "two-sided" test when appropriate. The *dotted curve* in **c** is for the Bayesian posterior probability
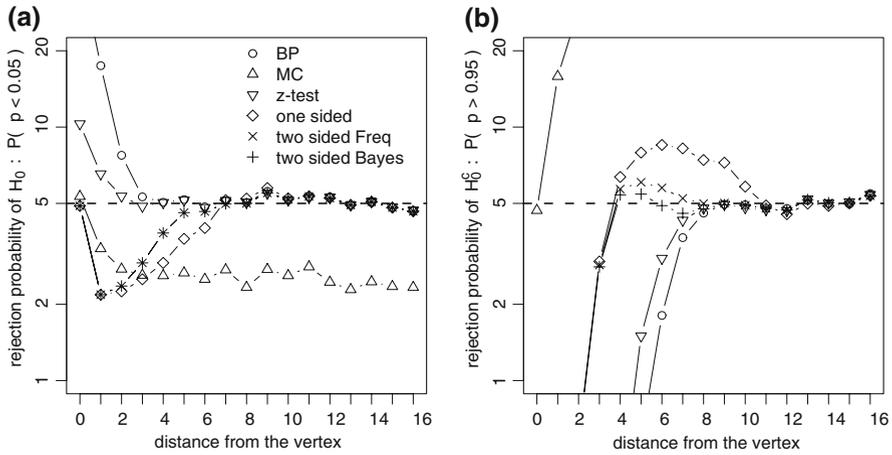
Fig. 4 **a** Rejection probability of the cone and **b** that of the complement of the cone. The angle at the vertex is $2\pi/10$
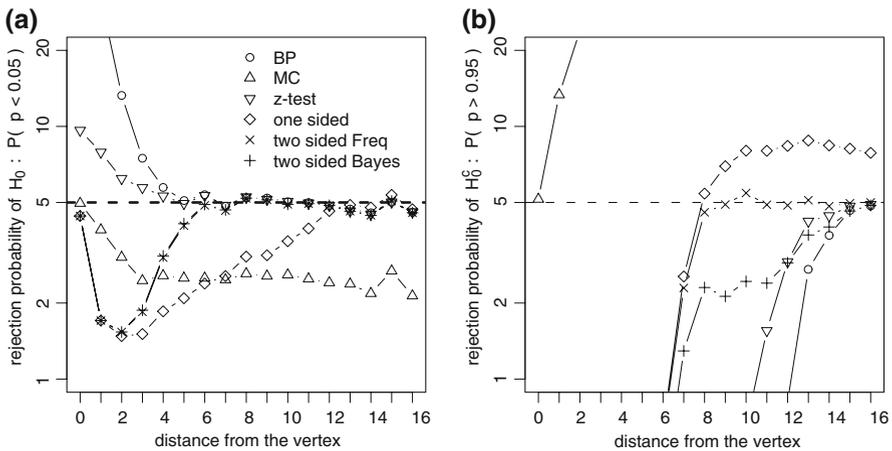
Fig. 5 **a** Rejection probability of the cone and **b** that of the complement of the cone. The angle at the vertex is $2\pi/20$

one-sided p value. The coverage error is almost zero except for small $\|\mu\|$'s, verifying what we attempted in this paper. The corresponding Bayesian posterior probability, labeled as "two sided Bayes," performs similarly.

The coverage error was further reduced near the vertex by setting simply $p(y) = p(H_0|y)$ without the model selection (the result is not shown here). However, this $p(y)$ increases as y moves away from the cone in some area, and then the shapes of R and R' are rather unreasonable in the sense of monotonicity mentioned at the last paragraph of Sect. 3.

7.2 Details of data generation

The contour lines in Fig. 3 are drawn by computing p values at all grid points (300×180) of step size 0.05 in the rectangle area. This huge computation was made possible by parallel processing using up to 700 cpus. The computation takes a few minutes per each grid point per cpu. Our algorithm is implemented in an experimental version of the scaleboot package of Shimodaira (2006) available from CRAN.

The rejection probabilities in Figs. 4 and 5 are computed by generating y according to (1) for 10,000 times, and then counting how many times $p(y) < 0.05$ or $p(y) > 0.95$ is observed. This computation is done for each $\mu \in \partial H_0$ with the distance from the vertex $\|\mu\| = 0, 1, \dots, 16$, i.e., $\mu = (0, 0), (1, 0), \dots, (16, 0)$ in the coordinates of Fig. 3.

For computing $p(H_0'|y)$ and $p(H_0|y)$, the two-step multiscale bootstrap described in Sect. 6.2 was performed with the $M = 13$ sets of scales (σ_i, τ_i), $i = 1, \dots, 13$, specified there. The parametric bootstrap, instead of the resampling, was used for the simulation. For drawing smooth contour lines, the number of bootstrap samples has increased to $B_i = 10^5$, whereas it was $B_i = 10^4$ in all the other results.

For $p(H_0'|y)$, we have considered the singular model of Shimodaira (2008) defined as $\psi(\sigma^2) = \beta_0 + \beta_1/(1 + \beta_2(\sigma - 1))$ for cones, and performed the model fitting method described in Sect. 6.3. From the Taylor series of this $\psi(\sigma^2)$ around $\sigma = 1$, we get $A = \beta_1\beta_2(3 - 2\beta_2)$, $B = \beta_1(\beta_2 - 1)^2$ for computing the higher order correction term $\Delta\rho$. We have also considered submodels by restricting some of $\varphi = (\beta_0, \beta_1, \beta_2, m)$ to specified values, and the minimum AIC model is chosen at each y. The frequentist p value is computed by (16) with $k = 3$ and $\sigma_0^2 = 1$.

For $p(H_0|y)$, we have considered the same singular model for the two surfaces by assuming they are curved in the opposite directions with the same magnitude of curvature. More specifically, the two ψ functions in (20) are defined as $\psi_1(\sigma^2) = \beta_0 + \beta_1/(1 + \beta_2(\sigma - 1))$ and $\psi_2(\sigma^2) = d - \beta_0 + \beta_1/(1 + \beta_2(\sigma - 1))$. The parameters $\varphi = (\beta_0, \beta_1, \beta_2, d)$ are estimated by the model fitting method described in Sect. 6.2. Submodels are also considered and model selection is performed using AIC. The frequentist p value is computed by (17), and the Bayesian posterior probability is computed by (22).

The rejection probabilities of other two commonly used measures (MC and z-test) are shown only for reference purposes. See Shimodaira (2008) for the details of these two measures. The rejection probability of the multiple comparisons, denoted MC here, is always below 5% in Figs. 4a and 5a, and the coverage error becomes zero at the vertex. On the other hand, the rejection probability of the z test is always below 5% in Figs. 4b and 5b, and the coverage error reduces to zero as $\|\mu\| \to \infty$.

8 Concluding remarks

In this paper, we have discussed frequentist and Bayesian measures of confidence for the three regions case, and have proposed a new computation method using the multiscale bootstrap technique. In this method, AIC played an important role for choosing appropriate parametric models of the scaling-law of bootstrap probability. Simulation

study showed that the proposed frequentist measure performs better for controlling the coverage error than the previously proposed multiscale bootstrap designed only for the two regions case.

A generalization of the confidence measures gives a frequentist interpretation of the Bayesian posterior probability as follows. Let us consider the situation of Theorem 2. If we strongly believe that $\mu \notin H_2$, we could use the one-sided p value $p(H_0'|y) = 1 - p(H_1|y)$, instead of the two sided $p(H_0|y)$. Similarly, we might use $1 - p(H_2|y)$ if we believe that $\mu \notin H_1$. By making the choice "adaptively," someone may want to use $p^{(1)}(H_0|y) = 1 - \max(p(H_1|y), p(H_2|y))$, although it is not justified in terms of coverage error. By connecting $p^{(1)}(H_0|y)$ and $p(H_0|y)$ linearly using an index s for the number of "sides," we get

$$p^{(s)}(H_0|y) = \pi(H_0|y) + s \min(p(H_1|y), p(H_2|y)).$$

It is easily verified that $p(H_0|y) = p^{(2)}(H_0|y)$ and $\pi(H_0|y) = p^{(0)}(H_0|y)$, indicating that the Bayesian posterior probability defined in Sect. 5 can be interpreted, interestingly, as a frequentist p value of "zero-sided" test of H_0. This kind of argument might lead to yet another compromise between frequentist and Bayesian.

Our formulation is rather restrictive. We have considered only the three regions case by introducing the surface h_2 in addition to the surface h_1 of the two regions case. Also these two surfaces are assumed to be nearly parallel to each other. It is worth to elaborate on generalizations of this formulation in future work, but too much of complication may result in unstable computation for estimating the scaling-law of bootstrap probability. AIC will be useful again in such a situation.

Appendix A

A.1 Proof of Theorem 2

First we consider rejection regions of testing H_0 for a specified α by modifying the two rejection regions of (3). Since h_1 and h_2 are nearly flat, the modified regions should be expressed as $R_1 = \{(u, v) \mid v > c - r_1(u), u \in \mathbb{R}^m\}$ and $R_2 = \{(u, v) \mid v < -d - c - r_2(u), u \in \mathbb{R}^m\}$ using nearly flat functions r_1 and r_2. The constant c is the same one as defined in (4). Write $a = \phi(c)$, $b = \phi(c + d)$ for brevity sake. We evaluate the rejection probability for $\mu \in \partial H_1 \cup \partial H_2$. Let $\mu \in \partial H_1$ for a moment, and put $\mu = (\theta, -h_1(\theta))$. By applying the argument of (13) to R_1 but (7) is replaced by (1), we get $P(Y \in R_1|\mu) = 1 - \Phi(c - \mathcal{E}_1 r_1(\theta) + h_1(\theta)) + O(\lambda^2) = \Phi(-c) + a(\mathcal{E}_1 r_1(\theta) - h_1(\theta)) + O(\lambda^2)$. The same argument applied to R_2 gives $P(Y \in R_2|\mu) = \Phi(-d - c - \mathcal{E}_1 r_2(\theta) + h_1(\theta)) + O(\lambda^2) = \Phi(-d - c) + b(-\mathcal{E}_1 r_2(\theta) + h_1(\theta)) + O(\lambda^2)$. Rearranging these two formula with the identity

$$P(Y \in R_1|\mu) + P(Y \in R_2|\mu) = \alpha \tag{29}$$

for an unbiased test, we get an equation $a(\mathcal{E}_1 r_1(\theta) - h_1(\theta)) + b(-\mathcal{E}_1 r_2(\theta) + h_1(\theta)) = O(\lambda^2)$. By exchanging the roles of r_1 and r_2, the equation becomes $b(\mathcal{E}_1 r_1(\theta) - h_2(\theta)) + a(-\mathcal{E}_1 r_2(\theta) + h_2(\theta)) = O(\lambda^2)$ for $\mu \in \partial H_2$ with $\mu = (\theta, -d - h_2(\theta))$. These two equations are expressed as

$$\begin{pmatrix} a & -b \\ -b & a \end{pmatrix} \begin{pmatrix} \mathcal{E}_1 r_1(\theta) \\ \mathcal{E}_1 r_2(\theta) \end{pmatrix} = (a - b) \begin{pmatrix} h_1(\theta) \\ h_2(\theta) \end{pmatrix} + O(\lambda^2). \tag{30}$$

For solving this equation with respect to r_1 and r_2, first apply the inverse matrix of the 2×2 matrix from the left in (30), and then apply the inverse operator of $\mathcal{E}_1$ so that

$$\begin{pmatrix} r_1(u) \\ r_2(u) \end{pmatrix} = \frac{1}{a+b} \begin{pmatrix} a & b \\ b & a \end{pmatrix} \begin{pmatrix} \mathcal{E}_{-1} h_1(u) \\ \mathcal{E}_{-1} h_2(u) \end{pmatrix} + O(\lambda^2). \tag{31}$$

Next we obtain an expression of p value corresponding to the rejection regions. $p(H_0|y)$ is defined as the value of α for which either of $y \in \partial R_1$ and $y \in \partial R_2$ holds. Note that r_1, r_2, and c depend on α. Let us assume $y \in \partial R_1$ and thus $c = v + r_1(u)$ for a moment. Write $a' = \phi(v) = a + O(\lambda)$, $b' = \phi(v + d) = b + O(\lambda)$ for brevity sake. Recalling (4), $p(H_0|y) = \Phi(-c) + \Phi(-d - c) = \Phi(-v - r_1(u)) + \Phi(-d - v - r_1(u)) = \Phi(-v) + \Phi(-d - v) - (a' + b')r_1(u) + O(\lambda^2)$, where $r_1(u)$ in (31) can be expressed as

$$r_1(u) = \frac{a'}{a' + b'} \mathcal{E}_{-1} h_1(u) + \frac{b'}{a' + b'} \mathcal{E}_{-1} h_2(u) + O(\lambda^2).$$

Therefore, $p(H_0|y) = \Phi(-v) + \Phi(-d - v) - a' \mathcal{E}_{-1} h_1(u) - b' \mathcal{E}_{-1} h_2(u) + O(\lambda^2) = \Phi(-v - \mathcal{E}_{-1} h_1(u)) + \Phi(-d - v - \mathcal{E}_{-1} h_2(u)) + O(\lambda^2)$. By applying (15) to H_1 and H_2, respectively, we get $p(H_1|y) = \Phi(v + \mathcal{E}_{-1} h_1(u)) + O(\lambda^2)$ and $p(H_2|y) = \Phi(-v - d - \mathcal{E}_{-1} h_2(u)) + O(\lambda^2)$, and thus $p(H_0|y) = 1 - p(H_1|y) + p(H_2|y) + O(\lambda^2)$. By exchanging the roles of H_1 and H_2, we have $p(H_0|y) = 1 - p(H_2|y) + p(H_1|y) + O(\lambda^2)$ for $y \in \partial R_2$. By taking the minimum of these two expressions of $p(H_0|y)$, we finally obtain (17). This p value satisfies (29) with error $O(\lambda^2)$, and thus (18) holds.

A.2 Proof of Lemma 1

The argument is very similar to the proof of Theorem 1. Given v, u^*, u^{**}, the joint distribution of $X' = (V^* - v)/\sigma$ and $X'' = (V^{**} - v)/\tau$ is Φ_ρ. Therefore, $P_{\sigma^2,\tau^2}(V^* \le -h_1(u^*) \wedge V^{**} \le -h_1(u^{**})|v, u^*, u^{**}) = P_{\sigma^2,\tau^2}(X' \le z_1 + \epsilon_1 \wedge X'' \le w_1 + \delta_1|v, u^*, u^{**}) = \Phi_\rho(z_1 + \epsilon_1, w_1 + \delta_1)$, where ϵ_1 and δ_1 are defined in (28). Taking the expectation with respect to (U^*, U^{**}), we have $\alpha_{\sigma^2,\tau^2}(H_0'|y) = P_{\sigma^2,\tau^2}(V^* \le -h_1(U^*) \wedge V^{**} \le -h_1(U^{**})|y) = E_{\sigma^2,\tau^2}(\Phi_\rho(z_1 + \epsilon_1, w_1 + \delta_1)|u)$. For proving (23), considering the Taylor series around (z_1, w_1), we obtain

$$E_{\sigma^2,\tau^2}\left(\Phi_\rho(z_1, w_1) + \frac{\partial \Phi_\rho}{\partial z_1} \epsilon_1 + \frac{\partial \Phi_\rho}{\partial w_1} \delta_1 \Big| u\right) + O(\lambda^2) \tag{32}$$

with $E_{\sigma^2,\tau^2}(\epsilon_1|u) = E_{\sigma^2,\tau^2}(\delta_1|u) = 0$ for completing the proof.

Next we show (24). The conditional probability given v, u^*, u^{**} is $P_{\sigma^2, \tau^2}(-d - h_2(u^*) \leq V^* \leq -h_1(u^*) \wedge -d - h_2(u^{**}) \leq V^{**} \leq -h_1(u^{**})|v, u^*, u^{**}) = P_{\sigma^2, \tau^2}(z_2 + \epsilon_2 \leq X' \leq z_1 + \epsilon_1 \wedge w_2 + \delta_2 \leq X'' \leq w_1 + \delta_1|v, u^*, u^{**}) = \Phi_\rho(z_1 + \epsilon_1, w_1 + \delta_1; z_2 + \epsilon_2, w_2 + \delta_2)$, where

$$\epsilon_2 = -\frac{h_2(U^*) - \mathcal{E}_{\sigma^2}h_2(u)}{\sigma}, \quad \delta_2 = -\frac{h_2(U^{**}) - \mathcal{E}_{\tau^2}h_2(u)}{\tau}.$$

Taking the expectation with respect to (U^*, U^{**}), we have $\alpha_{\sigma^2, \tau^2}(H_0|y) = P_{\sigma^2, \tau^2}(-d - h_2(U^*) \leq V^* \leq -h_1(U^*) \wedge -d - h_2(U^{**}) \leq V^{**} \leq -h_1(U^{**})|y) = E_{\sigma^2, \tau^2}(\Phi_\rho(z_1 + \epsilon_1, w_1 + \delta_1; z_2 + \epsilon_2, w_2 + \delta_2)|u)$. We only have to consider the Taylor series

$$E_{\sigma^2, \tau^2}\left(\Phi_\rho(z_1, w_1; z_2, w_2) + \frac{\partial \Phi_\rho}{\partial z_1}\epsilon_1 + \frac{\partial \Phi_\rho}{\partial w_1}\delta_1 + \frac{\partial \Phi_\rho}{\partial z_2}\epsilon_2 + \frac{\partial \Phi_\rho}{\partial w_2}\delta_2|u\right) + O(\lambda^2)$$

with $E_{\sigma^2, \tau^2}(\epsilon_i|u) = E_{\sigma^2, \tau^2}(\delta_i|u) = 0, i = 1, 2$ for completing the proof.

A.3 Proof of Lemma 2

By considering a higher-order term of the Taylor series in (13), we obtain $\alpha_{\sigma^2}(H_0'|y) = E_{\sigma^2}(\Phi(z_1) + \phi(z_1)\epsilon_1 - \phi(z_1)z_1\epsilon_1^2/2|u) + O(\lambda^3) = \Phi(z_1) + \phi(z_1)\Delta z_1 + O(\lambda^3) = \Phi(z_1 + \Delta z_1) + O(\lambda^3)$, proving (25) as well as (26). On the other hand, (27) is shown by considering higher-order terms of the Taylor series in (32) as

$$E_{\sigma^2, \tau^2}\left(\Phi_\rho(z_1, w_1) + \frac{\partial \Phi_\rho}{\partial z_1}\epsilon_1 + \frac{\partial \Phi_\rho}{\partial w_1}\delta_1\right.$$
$$\left. + \frac{1}{2}\left(\frac{\partial^2 \Phi_\rho}{\partial z_1^2}\epsilon_1^2 + 2\frac{\partial^2 \Phi_\rho}{\partial z_1 \partial w_1}\epsilon_1\delta_1 + \frac{\partial^2 \Phi_\rho}{\partial w_1^2}\delta_1^2\right)|u\right) + O(\lambda^3).$$

The proof completes by rearranging the above formula with

$$\frac{\partial^2 \Phi_\rho}{\partial z_1^2} = -z_1 \frac{\partial \Phi_\rho}{\partial z_1} - \rho\phi_\rho(z_1, w_1), \quad \frac{\partial^2 \Phi_\rho}{\partial z_1 \partial w_1} = \phi_\rho(z_1, w_1), \quad \frac{\partial \Phi_\rho}{\partial \rho} = \phi_\rho(z_1, w_1).$$

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control, 19*, 716–723.

Datta, G. S., Mukerjee, R. (2004). *Probability matching priors: higher order asymptotics*. New York: Springer.

DuPreez, J. P., Swanepoel, J. W. H., Venter, J. H., Somerville, P. N. (1985). Some properties of Somerville's multiple range subset selection procedure for three populations. *South African Statistical Journal, 19*, 45–72.

Efron, B. (1979). Bootstrap methods: another look at the jackknife. *Annals of Statistics, 7*, 1–26.

Efron, B., Tibshirani, R. (1998). The problem of regions. *Annals of Statistics, 26*, 1687–1718.

Efron, B., Halloran, E., Holmes, S. (1996). Bootstrap confidence levels for phylogenetic trees. *Proceedings of the National Academy of Sciences of the United States of America, 93*, 13429–13434.

Felsenstein, J. (1985). Confidence limits on phylogenies: An approach using the bootstrap. *Evolution, 39*, 783–791.

Lehmann, E. L. (1952). Testing multiparameter hypotheses. *Annals of Mathematical Statistics, 23*, 541–552.

Peers, H. W. (1965). On confidence points and bayesian probability points in the case of several parameters. *Journal of the Royal Statistical Society: Series B, 27*, 9–16.

Perlman, M. D., Wu, L. (1999). The emperor's new tests. *Statistical Science, 14*, 355–381.

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

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

Shimodaira, H. (2004). Approximately unbiased tests of regions using multistep-multiscale bootstrap resampling. *Annals of Statistics, 32*, 2616–2641.

Shimodaira, H. (2006). Scaleboot: approximately unbiased *p*-values via multiscale bootstrap. R add-on package available from CRAN. http://www.R-project.org.

Shimodaira, H. (2008). Testing regions with nonsmooth boundaries via multiscale bootstrap. *Journal of Statistical Planning and Inference, 138*, 1227–1241.

Tibshirani, R. (1989). Noninformative priors for one parameter of many. *Biometrika, 76*, 604–608.