Selective inference after variable selection via multiscale bootstrap

Yoshikazu Terada
Graduate School of Engineering Science, Osaka University
1-3 Machikaneyama-cho, Toyonaka, Osaka 560-8531, Japan
e-mail: terada@sigmath.es.osaka-u.ac.jp

and

Hidetoshi Shimodaira
Graduate School of Informatics, Kyoto University
Yoshida Honmachi, Sakyo-ku, Kyoto, 606-8501, Japan
e-mail: shimo@i.kyoto-u.ac.jp

Abstract: A general resampling approach is considered for selective inference problem after variable selection in regression analysis. Even after variable selection, it is important to know whether the selected variables are actually useful by showing p-values and confidence intervals of regression coefficients. In the classical approach, significance levels for the selected variables are usually computed by t-test but they are subject to selection bias. In order to adjust the bias in this post-selection inference, most existing studies of selective inference consider the specific variable selection algorithm such as Lasso for which the selection event can be explicitly represented as a simple region in the space of the response variable. Thus, the existing approach cannot handle more complicated algorithm such as MCP (minimax concave penalty). Moreover, most existing approaches set an event, that a specific model is selected, as the selection event. This selection event is too restrictive and may reduce the statistical power, because the hypothesis selection with a specific variable only depends on whether the variable is selected or not. In this study, we consider more appropriate selection event such that the variable is selected, and propose a new bootstrap method to compute an approximately unbiased selective p-value for the selected variable. Our method is applicable to a wide class of variable selection algorithms. In addition, the computational cost of our method is the same order as the classical bootstrap method. Through the numerical experiments, we show the usefulness of our selective inference approach.

1. Introduction

In the classical statistical inference, it is assumed that specification of a hypothesis is independent of obtained data. In recent years, since big and complicated data have been common in various fields, it is difficult to set hypotheses in advance. Thus, in modern data analysis, we commonly find useful hypotheses from obtained data using exploratory data analysis, and then we perform the classical inference for the selected hypotheses. However, we ignore the effects of

* Jointly affiliated at Mathematical Statistics Team, RIKEN Center for Advanced Intelligence Project (AIP), 1-4-1 Nihonbashi, Chuo-ku, Tokyo 103-0027, Japan.
the hypothesis selection in the classical inference, and thus this naive approach will not provide a valid statistical inference. Therefore, recently, the selective (or post-selection) inference, which deals with the hypothesis selection effect appropriately, has drawn considerable attention.

In this paper, we focus on the selective inference after the variable selection in regression analysis. The most simple and intuitive approach of selective inference is proposed by Cox (1975) and called data splitting. In data splitting, an i.i.d. sample is divided into two subsamples: one is used for the variable selection and the other is used for the inference of the selected variables. However, this approach reduces available data for both variable selection and inference. Fithian, Sun and Taylor (2014) provides the theoretical foundation to consider the optimality of the selective inference in the sense of the statistical power. In Berk et al. (2013), without assuming a specific variable selection method, the valid selective inference after variable selection for the submodel parameters is developed on the regression problem. Importantly, Berk et al. (2013) also introduces both “submodel view” and “full model view” of the targets of selective inference after variable selection. Under the setting of Berk et al. (2013), Lee et al. (2016) characterizes the selection event in which a specific model is selected by the lasso (Tibshirani, 1996). More precisely, this selection event can be represented as a union of polyhedra in the space of the response variable. In addition, based on this fact, Lee et al. (2016) proposes the exact selective inference for the variable selection via the lasso. Tibshirani et al. (2016) develops a general framework to perform selective inference after any selection event that can be represented as a response vector falling into a polyhedral set. Tibshirani et al. (2018) proves that this selective inference is asymptotically valid even for non-normal error distributions. On the other hand, the exact selective inference approaches such as Lee et al. (2016) and Tibshirani et al. (2016) assume that the selection event can be explicitly represented as a simple region in the space of the response variable. Thus, the existing approaches cannot handle more complicated algorithm such as MCP (minimax concave penalty; Zhang, 2010) and SCAD (smoothly clipped absolute deviation; Fan and Li, 2001). Although the selective inference of Berk et al. (2013) can be applied separately from the variable selection method, it is difficult to apply this method for the case in which the number of variables over 20. Moreover, as described in Berk et al. (2013), this selective inference is generally conservative for particular variable selection methods.

Recently, Terada and Shimodaira (2017) extends the general hypothesis testing framework, called problem of regions (Efron and Tibshirani, 1998), to the selective inference, and propose a new selective inference approach via multi-scale bootstrap of Shimodaira (2002, 2004, 2008). Using this approach, we can compute an approximately unbiased selective p-value. Moreover, Terada and Shimodaira (2017) provides the theoretical justification for this approach in two asymptotic theories. In this framework, we consider the general setting in which the hypothesis and the selection event can be represented as general regions in some parameter space. It is not necessary to know the shapes of these regions, and thus this approach can be applied widely. In fact, Shimodaira and Terada
(2019) describes an application of this approach for testing trees and edges in phylogenetics. In the usual multiscale bootstrap method, we change the sample size of bootstrap samples and then compute a bias corrected $p$-value using geometric quantities (curvature and signed distance of the region) estimated from the scaling-law of the bootstrap probability. However, the selective inference algorithm via multiscale bootstrap cannot be directly applied to selective inference after variable selection since the shape of the selective region is unwillingly related to the sample size in the variable selection problem. To overcome this difficulty, we propose the use of the resampling of the residuals with scale change. The advantage of our method is that we can apply our method to almost any variable selection algorithm. In addition, the computational cost of our method is the same order as the classical bootstrap method.

This paper is organized as follows. In Section 2, we give a brief exposition of multiscale bootstrap and the general selective inference via multiscale bootstrap. Section 3 presents some preliminaries. In Section 4, we develop a new selective inference algorithm via multiscale bootstrap in regression analysis. In Section 5, the usefulness of our approach is demonstrated through the numerical experiments.

2. An overview of multiscale bootstrap

First, we describe the basic idea of multiscale bootstrap (Shimodaira, 2002, 2004, 2008), and also briefly introduce the general selective inference framework proposed by Terada and Shimodaira (2017). The general statistical inference problem, in which the hypothesis is represented by a general region in some parameter space, is called the problem of regions (Efron and Tibshirani, 1998). This problem is an abstraction of many applications. For example, this framework is useful for assigning a confidence level for each clade in the obtained phylogenetic tree (Felsenstein, 1985; Efron, Halloran and Holmes, 1996).

Let $X_n = (X_1, \ldots, X_n)$ be a data with sample size $n$. In the problem of regions, it is assumed that there exists a transform $f_n$ of $X_n$ such that the transformed data follows the $(m+1)$-dimensional Gaussian distribution with unknown parameter $\mu$ and covariance identity $I_{m+1}$:

$$Y := f_n(X_n) \sim N_{m+1}(\mu, I_{m+1}).$$

Typically, $f_n$ involves multiplying $\sqrt{n}$ to a form of sample average so that the covariance matrix of $Y$ is properly rescaled. Here, the $(m+1)$-dimensional space of $Y$ will be referred to as the model space in this paper. Figure 1 shows the image of the model space. In addition, let $y$ be an observed value of $Y$, and suppose that a bootstrap sample $X'_n = (X'_1, \ldots, X'_n)$ with sample size $n'$ can be represented as a realization of the following Gaussian distribution in the model space:

$$Y^* = f_n(X'_n) \sim N(y, \sigma^2 I_{m+1}), \quad \sigma^2 = \frac{n}{n'}.$$
We will denote by $P_{\sigma^2}(\cdot|y)$ the probability measure of a bootstrap sample $Y^*$ with scale $\sigma > 0$. This framework is a simplification of reality, and is justified by the central limit theorem in many situations.

Let $H \subset \mathbb{R}^{m+1}$ be a general region and let us consider $H_0 : \mu \in H$ as a hypothesis. It is assumed that the region $H$ can be locally represented as $H = \{(u, v) \mid v \leq -h(u), \; u \in \mathbb{R}^m\}$ using some continuous function $h : \mathbb{R}^m \to \mathbb{R}$. Let $\partial H := \{(u, v) \mid v = -h(u), \; u \in \mathbb{R}^m\}$ be the boundary surface of the region $H$. In this setting, our main propose is to compute an approximately unbiased $p$-value $p(H|y)$ for the hypothesis $H_0 : \mu \in H$ and a given significance level $\alpha > 0$:

$$\forall \mu \in \partial H; \; P(p(H|y) < \alpha \mid \mu) \approx \alpha. \quad (1)$$

In the above equation, the difference between $P(p(H|y) < \alpha \mid \mu)$ and $\alpha$ is called bias (or error). The bootstrap probability $p_{\text{BP}}(H|y) := P(Y^* \in H|y)$ can be considered as the most simple $p$-value satisfying (1) (e.g., see Efron, Halloran and Holmes (1996)). More formally, in the classical large sample theory, if the region $H$ has a smooth boundary surface, the bootstrap probability $p_{\text{BP}}(H|y)$ has the first-order accuracy: $\forall \mu \in \partial H; \; P(p_{\text{BP}}(H|y) < \alpha \mid \mu) = \alpha + O(n^{-1/2})$. However, in many practical situations, the bootstrap probability $p_{\text{BP}}$ often has a serious bias.

### 2.1. Basic idea of multiscale bootstrap

To obtain more accurate $p$-values, the geometric quantities related to the data point $y$ and the region $H$ play a key role. In fact, Efron and Tibshirani (1998) shows that we can compute a more accurate $p$-value using the signed distance $v(y)$ from the data point $y$ to the region $H$. More precisely, the $p$-value $p_{\text{sign}}(H|y) := P_1(v(Y^*) \geq v(y) \mid \hat{\mu}(y))$ is proposed, where $\hat{\mu}(y)$ is the projected point of $y$ onto $\partial H$. This $p$-value $p_{\text{sign}}(H|y)$ has the third-order accuracy (Efron, 1985; Efron and Tibshirani, 1998).

However, in most practical situations, it is difficult to access the model space and to obtain the explicit formula of the hypothesis region in the model space. Thus, we cannot compute the signed distance $v(y)$ in general. To overcome this
difficulty, Shimodaira (2002, 2004, 2008) propose a new bootstrap method, called multiscale bootstrap. In multiscale bootstrap, the geometric quantities such as the signed distance \( v(y) \) and the mean curvature of \( \partial H \) at \( \mu(y) \) are estimated based on the scaling law of the bootstrap probabilities, and an accurate \( p \)-value is computed based on these estimated quantities. Let \( \alpha_{\sigma^2}(H|y) = P_{\sigma^2}(Y^* \in H|y) \) be the bootstrap probability with scale \( \sigma > 0 \). We will denote by \( \Phi(x) \) the cumulative distribution function of the standard normal distribution, and write \( \Phi(x) = 1 - \Phi(x) \). For a general region \( H \), we define the normalized bootstrap \( z \)-value as \( \psi_{\sigma^2}(H|y) := \sigma \Phi^{-1}(\alpha_{\sigma^2}(H|y)) \). Then, we have the following scaling law of the bootstrap probabilities:

\[
\psi_{\sigma^2}(H|y) = v(y) + \gamma(y)\sigma^2 + O_p(n^{-1}),
\]

where \( \gamma(y) \) is the mean curvature of the boundary surface at \( \mu \). This scaling law can be considered as the simple linear regression model \( \beta_{H,0} + \beta_{H,1}\sigma^2 \) with \( \sigma^2 \) as the predictor. We will denote by \( \varphi_H(\sigma^2|\beta_H) \) the model for the normalized bootstrap \( z \)-value, such as \( \beta_{H,0} + \beta_{H,1}\sigma^2 \) with parameter \( \beta_H = (\beta_{H,0}, \beta_{H,1}) \). We note that the bootstrap probabilities with several scales can be computed by using the bootstrap samples with different sample sizes, say \( n' = [0.5n_1, \cdots, 1.0n_1, \cdots, 1.5n_1] \). Let \( B \) be the number of bootstrap replicates, and \( C_H = \#\{Y^* \in H\} \) be the frequency to be \( Y^* \in H \). Let \( \hat{\psi}_{\sigma^2}(H|y) \) be the estimated normalized bootstrap \( z \)-value by using the estimated bootstrap probability \( \hat{\alpha}_{\sigma^2}(H|y) = C_H/B \). We can estimate the values of \( v(y) \) and \( \gamma(y) \) by the simple regression for \( \{(\sigma_j^2, \hat{\psi}_{\sigma_j^2}(H|y))\} \). Shimodaira (2002) proposes the following \( p \)-value:

\[
p_{AU}(H|y) := \Phi(\varphi_H(-1|\beta_H)) = \Phi(v(y) - \gamma(y)) + O_p(n^{-1}).
\]

This \( p \)-value \( p_{AU}(H|y) \) has the second-order accuracy (Shimodaira, 2004; Efron and Tibshirani, 1998):

\[
\forall \mu \in \partial H; \quad P(p_{AU}(H|Y) < \alpha | \mu) = \alpha + O(n^{-1}).
\]

It becomes third-order accurate erring only \( O(n^{-3/2}) \) when \( \varphi_H(\sigma^2|\beta_H) \) is properly estimated from observed values of \( \psi_{\sigma^2}(H|y) \) including terms of order \( O_p(n^{-1}) \).

In the classical large sample theory, the shape of \( H \) in the model space is magnified by \( \sqrt{n} \), and thus the key point is that the smooth boundary surface \( \partial H \) approaches a flat surface in a neighborhood of any point on \( \partial H \). In contrast, for non-smooth surfaces, this key property is not satisfied. For example, if the region \( H \) is cone-shaped, the shape of \( H \) is scale invariant in a neighborhood of the vertex of \( H \). To deal with general regions with non-smooth boundary surfaces, Shimodaira (2008) develops a new theoretical frame work, called the asymptotic theory of nearly flat surfaces. The brief introduction of this theory is provided in Appendix A.

### 2.2. General selective inference via multiscale bootstrap

Here, we describe an extended framework of the problem of regions for the selective inference. In the model space, two regions \( H = \{(u,v) \mid v \leq -h(u), \ u \in \mathbb{R} \} \)
\( \mathbb{R}^m \) and \( S = \{(u, v) \mid v > -s(u), \ u \in \mathbb{R}^m \} \) are considered. Suppose that the selection event can be represented as \( y \in S \), and we consider the selective inference in which the hypothesis \( H_0 : \mu \in H \) is selected whenever \( y \in S \). For a technical reason, \( \partial H \) and \( \partial S \) are assumed to be nearly parallel in the sense that the first derivatives of \( h \) and \( s \) differ only \( O(n^{-1}) \). In this setting, for a given significant level \( \alpha \), we want to compute selective \( p \)-values \( p(H|S, y) \) satisfying
\[
\forall \mu \in \partial H; \quad \frac{P(p(H|S, Y) < \alpha \mid \mu)}{P(Y \in S \mid \mu)} \approx \alpha.
\]

Terada and Shimodaira (2017) proposes the following approximately unbiased selective \( p \)-value \( p_{SI}(H|S, y) \) for regions \( H \) and \( S \) with smooth boundary surfaces:
\[
p_{SI}(H|S, y) := \frac{\Phi(\varphi_H(-1|\beta_H))}{\Phi(\varphi_H(-1|\beta_H) + \varphi_S(0|\beta_S))}.
\]

In the classical large sample theory, this selective \( p \)-value \( p_{SI}(H|S, y) \) has the second-order accuracy:
\[
\forall \mu \in \partial H; \quad \frac{P(p_{SI}(H|S, Y) < \alpha \mid \mu)}{P(Y \in S \mid \mu)} = \alpha + O(n^{-1}).
\]

In Terada and Shimodaira (2017), the selective \( p \)-value for the regions with non-smooth boundary surfaces is also proposed, and the theoretical justification of this \( p \)-value is provided using the asymptotic theory of nearly flat surfaces. For more details about the case in which the regions \( H \) and \( S \) have possibly non-smooth boundary surfaces, see Appendix A. The detailed calculation is provided as Algorithm 1. In the step 4 of this algorithm, (B) is used when the hypothesis region has the non-smooth boundaries.

**Algorithm 1** Computing approximately unbiased \( p \)-values for general regions \( H \) and \( S \)

1. Specify several \( n' \in \mathbb{N} \) values, and set \( \sigma^2 = n/n' \) for each \( n' \). Set the number of bootstrap replicates \( B \), say, 10,000.
2. For each \( n' \), perform bootstrap resampling to generate \( Y^* \) for \( B \) times and compute \( \alpha_{a2}(H|y) = C_H/B \) and \( \alpha_{a2}(S|y) = C_S/B \) by counting the frequencies \( C_H = \#\{Y^* \in H \} \) and \( C_S = \#\{Y^* \in S \} \). We actually work on \( X^*_n \) instead of \( Y^* \). Compute \( \psi_{a2}(H|y) = \sigma^-1(\alpha_{a2}(H|y)) \) and \( \psi_{a2}(S|y) = \sigma^-1(\alpha_{a2}(S|y)) \).
3. Estimate parameters \( \beta_H(y) \) and \( \beta_S(y) \) by fitting models \( \psi_{a2}(H|y) = \varphi_H(\sigma^2|\beta_H) \) and \( \psi_{a2}(S|y) = \varphi_S(\sigma^2|\beta_S) \), respectively.
4. Approximately unbiased \( p \)-values of non-selective inference \( p_{AU} \) and of selective inference \( p_{SI} \) are computed by one of (A) and (B) below.

(A) Compute \( p \)-values by \( p_{AU}(H|y) = \Phi(z_H) \) and \( p_{SI}(H|S, y) = \Phi(z_H)/\Phi(z_H + z_S) \), where \( z_H = \varphi_H(-1|\beta_H(y)) \) and \( z_S = \varphi_S(0|\beta_S(y)) \).

(B) Specify \( k \in \mathbb{N}, \sigma^2_0, \sigma^2_1 > 0 \) (e.g., \( k = 3 \) and \( \sigma^2_1 = \sigma^2_0 = 1 \)). Compute \( p \)-values by \( p_{AU,k}(H|y) = \Phi(z_{H,k}) \) and \( p_{SI,k}(H|S, y) = \Phi(z_{H,k})/\Phi(z_{H,k} + z_{S,k}) \), where \( z_{H,k} = \varphi_{H,k}(-1|\beta_H(y), \sigma^2_k) \) and \( z_{S,k} = \varphi_{S,k}(0|\beta_S(y), \sigma^2_k) \) computed by formula (A.1) in Appendix A.
3. Preliminaries

Here, we describe the setting of the selective inference after variable selection in regression analysis. We employ the general assumption used in Berk et al. (2013), Lee et al. (2016), and Tibshirani et al. (2016). Consider the response variable \( Y = (Y_1, \ldots, Y_n) \) drawn from the multivariate Gaussian distribution:

\[
Y \sim N(\mu, \sigma^2 I_n),
\]

where \( \mu \in \mathbb{R}^n \) is an unknown parameter, \( I_n \) is the \( n \)-dimensional identity matrix, and \( \sigma^2 \) is assumed to be known. We will denote by \( y \in \mathbb{R}^n \) the observed value of \( Y \). Let \( X = (x_1, \ldots, x_p) = (x_{ij})_{n \times p} \) be a non-random full column matrix whose columns represent the predictors. Note that the error variance \( \sigma^2 \) can be estimated if \( \mu \) is modeled as a function of predictors \( x_1, \ldots, x_p \in \mathbb{R}^n \). Let \( \hat{M} \subseteq \{1, \ldots, p\} \) be the set of selected variables by applying a specific variable selection method to \( (X, y) \in \mathbb{R}^n \times \mathbb{R}^n \). Let \( \hat{s}_j \in \{\pm\} \) be the sign of the estimated coefficient \( \hat{\beta}_j \) of variable \( j \) by a specific variable selection method such as the lasso and MCP. Let \( \hat{s}_M \) denote the vector of the estimated signs of the coefficients of the selected variables.

In most existing approaches such as Berk et al. (2013) and Lee et al. (2016), the selection event is set as that a specific model \( M \subseteq \{1, \ldots, p\} \) is selected, that is \( \{\hat{M} = M\} \). Due to computational issue, Lee et al. (2016) and Tibshirani et al. (2016) mainly consider the more specific selection event that a specific model \( M \subseteq \{1, \ldots, p\} \) with specific signs \( s_M \in \{\pm\}^{\#(M)} \) is selected, that is, \( \{\hat{M} = M, \hat{s}_M = s_M\} \). Moreover, in Berk et al. (2013), the target of the selective inference is the submodel parameters:

\[
\beta^{(M)} := \arg \min_{b_M \in \mathbb{R}^{\#(M)}} \mathbb{E} \left[ \|Y - X_M b_M\|^2 \right] = (X_M^T X_M)^{-1} X_M^T \mu,
\]

where \( X_M \) is the predictor matrix consisting of the selected variables, and \( M \subseteq \{1, \ldots, p\} \) is the set of the selected variables. In Lee et al. (2016) and Tibshirani et al. (2016), the following full-model parameter can be treated as the target of the selective inference:

\[
\beta = (\beta_1, \ldots, \beta_p) = \arg \min_{b \in \mathbb{R}^p} \mathbb{E} \left[ \|Y - X b\|^2 \right] = (X^T X)^{-1} X^T \mu.
\]

In most practical situations, we are interested in whether the variable \( j \) is useful or not whenever the variable \( j \) is selected. More precisely, whenever the variable \( j \) is selected and \( \hat{\beta}_j \gtrless 0 \), the hypothesis \( H_0 : \beta_j \lessgtr 0 \) is tested. Thus, the selection event \( \{\hat{M} = M, \hat{s}_M = s_M\} \) is too restrictive, and this over-conditioning could reduce the statistical power. Recently, Liu, Markovic and Tibshirani (2018) suggests the use of the selection event \( \{ j \in \hat{M} \} \) in the sense of the statistical power. This idea is related to the monotonicity of the selective error provided in Fithian, Sun and Taylor (2014). Here, we also consider the minimal selection event \( \{ j \in \hat{M}, \hat{s}_j = s_j \} \) where \( \hat{s}_j \in \{\pm\} \). In this setting, the target of the selective inference is the full-model parameter \( \beta \). Hence, the main purpose of the
selective inference is to compute the selective p-value \( p_j(y) \), which satisfies the following inequality under the null hypothesis \( H_0: \beta_j \preceq 0 \):

\[
P(p_j(Y) < \alpha \mid j \in \hat{M} \text{ and } \hat{\beta}_j \preceq 0) \leq \alpha.
\]

Ideally, the selective p-value should satisfy the unbiasedness as follows: under \( H_0: \beta_j = 0 \)

\[
P(p_j(Y) < \alpha \mid j \in \hat{M} \text{ and } \hat{s}_j = s_j) = \alpha,
\]

where \( s_j \in \{\pm\} \).

4. Selective inference after model selection via multiscale bootstrap

In this section, we develop a new algorithm to compute the approximately unbiased selective p-value which satisfies approximately the equation (2). For the variable selection via the lasso, the selection event \( \{j \in \hat{M}, \hat{s}_j = s_j\} (s_j \in \{\pm\}) \) can be represented as a union of polyhedra in the \( n \)-dimensional space of the response variable (Lee et al., 2016). The left panel of Figure 2 shows the relationship between the selected model by the lasso and the corresponding region in the space of the response vector when \( n = 2 \). In contrast, for more complicated variable selection methods such as MCP and SCAD, the region \( S \) of the selective event \( \{j \in \hat{M}, \hat{s}_j = s_j\} \) will be complicated, and the explicit shape of the selective region \( S \) may not be obtained. Thus, it is difficult to consider the exact selective inference such as Lee et al. (2016) and Tibshirani et al. (2016). The right panel of Figure 2 shows the relationship between the selected model by MCP and the corresponding region in the space of the response vector when \( n = 2 \). Note that since no explicit representation of the selection event, we numerically evaluate which variables are selected for each point in the right panel of Figure 2.

The advantage of the selective inference described in Section 2 is that it is not necessary to know the shapes of the hypothesis and selective regions. However, it is assumed that the hypothesis and selective regions can be represented as specific regions, which are independent of the sample size \( n \), in the model space. For the selective inference in regression analysis, however, the shape of selective region inevitably depends on \( n \) because it is the dimension of the model space. Hence, the general selective inference approach described as Algorithm 1 cannot be directly applied.

At first, we remark that it is assumed that \( Y \sim N_n(\mu, \sigma^2 I_n) \) and that the selection event can be represented as the region of the space of \( Y \). Then, it is realized that the normalized space of \( Y/\sigma \) can be considered as the model space described in Section 2. As shown in Figure 2, the selective region \( S \) which represents the selection event \( \{j \in \hat{M}, \hat{s}_j = s_j\} \) has generally a non-smooth boundary surface. In contrast, for \( \eta \in \mathbb{R} \), the hypothesis \( H_0: \beta_j \preceq \eta \) can be
represented as the following region in the space of $Y$:

$$H = \{ y \in \mathbb{R}^n \mid a_j^T y \leq \eta \}, \quad A = \begin{bmatrix} a_1^T \\ \vdots \\ a_p^T \end{bmatrix} = (X^T X)^{-1} X^T.$$

Thus, the hypothesis region has a flat boundary surface. Hence, we may use Step 3 - 4 of Algorithm 1 to compute the selective $p$-value if the bootstrap probabilities related to the selective region can be computed at several scales. With the normality of response $Y$, the parametric bootstrap method by sampling from $N(y, \gamma^2 I_n)$ can be applied to the computation of bootstrap probabilities $\alpha_{\gamma^2}(S \mid y)$ at several scales $\gamma > 0$. To relax the Gaussian assumption, we may consider the resampling of residuals with scale change. More formally, we resample the scaled residuals to compute $\alpha_{\gamma^2}(S \mid y)$ at several $\gamma > 0$ as follows. Let $\tilde{\beta}^{(LS)} = (X^T X)^{-1} X^T y$ be the least-squares estimator based on the full model. Write $\hat{\epsilon} := y - X \hat{\beta}^{(LS)}$ and $(h_1, \ldots, h_n) := \text{diag}(X (X^T X)^{-1} X^T)$. Then, the adjusted residuals $\hat{\epsilon} = (\hat{\epsilon}_1, \ldots, \hat{\epsilon}_n)^T$ are defined as $\hat{\epsilon}_i = \epsilon_i / \sqrt{1 - h_{ii}}$. To compute the bootstrap probability $\alpha_{\gamma^2}(S \mid y)$ at $\gamma > 0$, we use the following bootstrap sample:

$$y^*_\gamma = X \tilde{\beta}^{(LS)} + \gamma \hat{\epsilon}^*,$$

where $\hat{\epsilon}^* = (\hat{\epsilon}_1^*, \ldots, \hat{\epsilon}_n^*)^T$ is a bootstrap sample with size $n$ from $(\hat{\epsilon}_1, \ldots, \hat{\epsilon}_n)$. Here, since the hypothesis region of $H_0 : \beta_j \leq 0$ has the flat boundary surface, we have

$$\psi_{\gamma^2}(H \mid y) = v(y \mid H) = a_j^T y / \|a_j\|_2,$$

where $v(y \mid H)$ is the signed distance from $y$ to the hypothesis region $H$. Thus, it is not necessary to compute bootstrap probabilities related to the hypothesis.
region. Combining these facts, we propose Algorithm 2 to compute an approximately unbiased selective p-value for the selected variable \( j \in \hat{M} \). It is worth noting that Algorithm 2 can be applied to almost any variable selection method. Moreover, in the asymptotic theory of nearly flat surfaces, the proposed p-value \( p_{SI}(H|S,y) \) has the second-order accuracy (Terada and Shimodaira, 2017). The computational cost is the same order as the classical bootstrap method. Thus, this algorithm works even for large \( p \) such as \( p > 20 \).

**Algorithm 2** Approximately unbiased selective p-value for the selected variable \( j \in \hat{M} 

1. Specify several \( \gamma > 0 \) values. Set the number of bootstrap replicates \( B \), say, 10000.
2. Compute the adjusted residuals \( \hat{e} \) based on the full model.
3. For each \( \gamma \), perform bootstrap resampling to generate \( y_{\gamma}^* \) given by (3) for \( B \) times and compute \( \alpha_{\gamma,z}(S|y) = C_S/B \) by counting the frequency \( C_S = \#\{j \in \hat{M}_s^\gamma\} \) where \( \hat{M}_s^\gamma \) is the set of selected variables by applying the specific algorithm to \((X, y_{\gamma})\). Compute \( \psi_{\gamma,z}(S|y) = \gamma \Phi^{-1}(\alpha_{\gamma,z}(S|y)) \).
4. Estimate parameters \( \theta_S(y) \) by fitting model \( \psi_{\gamma,z}(S|y) = \varphi_S(\gamma^2|\theta_S) \).
5. Compute the selective p-value by \( p_{SI}(H|S,y) = \Phi(z_H)/\Phi(z_H + z_S) \), where \( z_H = a_j^T y/\|a_j\|_2 \) is selected. For each test, we also count how many times, say \( N \), the variable \( j \) is rejected, and the selective rejection probability is estimated by \( R_j/N_j \). The left panel of Figure 3 shows the selective rejection probabilities of these test at each variable for the lasso.

5. Numerical experiments

Here, we show some numerical experiments to demonstrate the usefulness of our method. For the lasso, the exact unbiased selective test conditioned on \( j \in \hat{M} \) and \( \hat{s}_j = s_j \) \( (s_j \in \{\pm\}) \) can be constructed (Lee et al., 2016; Liu, Markovic and Tibshirani, 2018), but it did not work for large \( p \). Here, the lasso is defined as \( \min_{\beta \in \mathbb{R}^p} \|y - X\beta\|^2/2 + \lambda \sum_{j=1}^p |\beta_j| \), where \( \|y\|^2 = \sum_{i=1}^n y_i^2 \). Set \((n, p) = (50, 25) \) and \( \beta = (2, 2, 2, 2, 0, \ldots, 0)^T \in \mathbb{R}^p \). The elements \( x_{ij} \) of the input matrix \( X \) were independently generated from the standard normal distribution \( N(0, 1) \). Then, the response \( y \in \mathbb{R}^n \) was generated as \( y = X\beta + \epsilon \), where \( \epsilon \) was generated from the \( n \)-dimensional standard normal distribution \( \mathcal{N}(0, I_n) \). We simulated 2000 independent datasets. We set the significant level \( \alpha = 5\% \). Here, we used the lasso with the penalty parameter \( \lambda = 10 \) as the variable selection method. In each dataset, we performed the classical \( t \)-test and our approximately unbiased test for each selected variable. We count how many times, say \( N_j \), the variable \( j \) is selected. For each test, we also count how many times, say \( R_j \), the null hypothesis \( H_0 : \beta_j = 0 \) is rejected, and the selective rejection probability is estimated by \( R_j/N_j \). The left panel of Figure 3 shows the selective rejection probabilities of these test at each variable for the lasso.
the variable selection method, and the right panel of Figure 3 shows the selective rejection probabilities in the case of MCP. Note that each variable with the zero coefficient were selected approximately 250 times in this experiment. In this setting, whereas no exact unbiased selective inference is proposed, the selective rejection probabilities of our test for variable 6 to 25 are around 5%. Thus, our method works well not only for the lasso but also for more complicated variable selection methods such as the MCP.

Next, we deal with the prostate cancer data (Stamey et al., 1989), which is available in the R packageElemStatLearn (Halvorsen, 2015). Stamey et al. (1989) studied the relation between the level of prostate-specific antigen (PSA) and 8 clinical measures: the log cancer volume (lcavol), the log prostate weight (lweight), and so on. Here, we consider a linear regression model to the log of PSA (lpsa) with 8 clinical measures. In this application, we prepossessed the data so that each variable has mean zero and unit variance. The main purpose is to provide the selective confidence intervals (CIs) for the coefficients of the 6 selected variables by the lasso with the penalty $\lambda = 5$. Here, we also set $\alpha = 5\%$. We computed four types of confidence intervals with confidence level $1 - \alpha$: the non-selective CI $[L_j^{(t)}, U_j^{(t)}]$ using $t$-distribution, the selective CI $[L_j^{(M,s)}, U_j^{(M,s)}]$ conditioned on $\hat{M} = M, \hat{s}_M = s_M$ (Lee et al., 2016), the selective CI $[L_j^{(vs)}, U_j^{(vs)}]$ conditioned on $j \in \hat{M}, \hat{s}_j = s_j$, and the approximate selective CI $[L_j^{(SI)}, U_j^{(SI)}]$ based on our approximately unbiased $p$-values. We note that first three CIs satisfy the following equations, respectively, for $M \subseteq \{1, \ldots, p\}$, $s = (s_1, \ldots, s_p)^T \in \{\pm\}^p$:

$$
P(\beta_j \in [L_j^{(t)}, U_j^{(t)}]) = 1 - \alpha, \quad P(\beta_j \in [L_j^{(M,s)}, U_j^{(M,s)}] \mid \hat{M} = M, \hat{s} = s) = 1 - \alpha, \quad P(\beta_j \in [L_j^{(vs)}, U_j^{(vs)}] \mid j \in \hat{M}, \hat{s}_j = s_j) = 1 - \alpha.$$

Fig 3. Selective rejection probabilities (in percent) with two standard deviations for each variable (Black: the classical $t$-test, Red: our approximately unbiased test conditioned on $j \in \hat{M}$ and $\hat{s}_j = s_j$).
Figure 4 shows these confidence intervals. From this figure, we can see that our selective CIs \([L_j^{(SI)}, U_j^{(SI)}]\) can approximate the exact selective CIs \([L_j^{(vs)}, U_j^{(vs)}]\) very well. Moreover, the over-conditioning made CIs \([L_j^{(M,s)}, U_j^{(M,s)}]\) wider than \([L_j^{(vs)}, U_j^{(vs)}]\), and this indicates that the selective event \(\{j \in \hat{M}, \hat{s}_j = s_j\}\) is preferable. This is because the selection event \(\{\hat{M} = M, \hat{s} = s\}\) is too restrictive. For more details about an appropriate conditioning, we refer the reader to Liu, Markovic and Tibshirani (2018).

**Acknowledgments**

This research was supported in part by JSPS KAKENHI Grant (16K16024 to YT, 16H02789 to HS), RIKEN Engineering Network (to YT), and a collaboration grant (to YT) from Organ Technologies Inc.

**References**

Berk, R., Brown, L., Buja, A., Zhang, K. and Zhao, L. (2013). Valid post-selection inference. *Annals of Statistics* **41** 802–837.

Cox, D. R. (1975). A note on data-splitting for the evaluation of significance levels. *Biometrika* **62** 441–444.

Efron, B. (1985). Bootstrap Confidence Intervals for a Class of Parametric Problems. *Biometrika* **72** 45–58.
Efron, B., Halloran, E. and Holmes, S. (1996). Bootstrap confidence levels for phylogenetic trees. *Proc. Natl. Acad. Sci. USA* **93** 13429-13434.

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

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association* **96** 1348–1360.

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

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

Halvorsen, K. (2015). ElemStatLearn: Data Sets, Functions and Examples from the Book: “The Elements of Statistical Learning, Data Mining, Inference, and Prediction” by Trevor Hastie, Robert Tibshirani and Jerome Friedman. R package. [https://CRAN.R-project.org/package=ElemStatLearn](https://CRAN.R-project.org/package=ElemStatLearn).

Lee, J. D., Sun, D. L., Sun, Y. and Taylor, J. E. (2016). Exact post-selection inference, with application to the lasso. *Annals of Statistics* **44** 907–927.

Lehmann, E. L. (1952). Testing multiparameter hypotheses. *Ann. Math. Statistics* **23** 541–552.

Liu, K., Markovic, J. and Tibshirani, R. (2018). More powerful post-selection inference, with application to the Lasso. arXiv.

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. (2008). Testing Regions with Nonsmooth Boundaries via Multiscale Bootstrap. *Journal of Statistical Planning and Inference* **138** 1227–1241.

Shimodaira, H. and Terada, Y. (2019). Selective inference for testing trees and edges in phylogenetics. *Frontiers in Ecology and Evolution* **7**:174 1–15.

Stamey, T., Kabalin, J., Johnstone, I., Freiha, F., Redwine, E. and Yang, N. (1989). Prostate specific antigen in the diagnosis and treatment of adenocarcinoma of the prostate II. Radical prostatectomy treated patients. *Journal of Urology* **141** 1076–1083.

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

Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society: Series B* **58** 267–288.

Tibshirani, R., Taylor, J., Lockhart, R. and Tibshirani, R. (2016). Exact Post-Selection Inference for Sequential Regression Procedures. *Journal of the American Statistical Association* **111** 600–620.

Tibshirani, R., Rinaldo, A., Tibshirani, R. and Wasserman, L. (2018). Uniform Asymptotic Inference and the Bootstrap After Model Selection. *Annals of Statistics* **46** 1255–1287.

Zhang, C. H. (2010). Nearly unbiased variable selection under minimax con-
Appendix A: Asymptotic theory for non-smooth boundary surfaces

Here, we follow the notation of Section 2 in the main paper. In the main paper, we only describe the multiscale bootstrap method for the hypothesis and selective regions with smooth boundaries. In the classical large sample theory, an important point is that the smooth boundary surface of the hypothesis region approaches a flat surface in a neighborhood of any point on its boundary surface. However, this claim cannot be true for regions with non-smooth boundaries since cone-shaped regions are scale invariant in the neighborhood of the vertex. In many practical situations, the hypothesis and selective regions could have non-smooth surfaces. Thus, Shimodaira (2008) develops a new theoretical framework, called the asymptotic theory of nearly flat surfaces. In this theory, we consider the situation that the magnitude of boundary surfaces, say $\lambda$, becomes small, that is, any boundary surfaces approach flat surfaces at least locally in a neighborhood. The artificial parameter $\lambda$ is introduced, and consider the situation of $\lambda \to 0$ instead of $n \to \infty$. More precisely, suppose that the $L^1$-norms $\|h\|_1$ and $\|\tilde{h}\|_1$ of function $h$ and its Fourier transform $\tilde{h}$ are bounded and that the $L^\infty$-norm $\|h\|_\infty$ of $h$ has the same order as $\lambda$. Then, we consider the asymptotic theory as $\lambda \to 0$. Note that $\lambda$ in this theory is corresponding to $1/\sqrt{n}$ in the classical large sample theory.

Even in this theory, the bootstrap probability also has the first-order accuracy:

$$\forall \mu \in \partial H; \ P(p_{\text{BP}}(H|Y) < \alpha | \mu) = \alpha + O(\lambda^{-1}).$$

Write $y = (u,v) \in \mathbb{R}^m \times \mathbb{R}$. Then, the distribution of the bootstrap sample $Y^* = (U^*,V^*)$ with the scale $\sigma^2$ is given as

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

Let $E_{\sigma^2}$ denote the expectation operator related to $U^*$, that is,

$$E_{\sigma^2} h(u) = E_{\sigma^2} [h(U^*)|u] = F^{-1} [e^{-\sigma^2 \|\omega\|^2/2} \tilde{h}](\omega)(u),$$

where $E_{\sigma^2} [\cdot|u]$ is the expectation related to $U^*$ and $F^{-1}$ is the inverse Fourier transform operator. For the normalized bootstrap $z$-value, we have the following scaling-law which is parallel to one of the large sample theory:

$$\psi_{\sigma^2}(H|y) = v + E_{\sigma^2} h(u) + O(\lambda^2).$$

We note that, for $\sigma_1^2, \sigma_2^2 > 0$, it follows that $E_{\sigma_1^2} E_{\sigma_2^2} h(u) = E_{\sigma_1^2 + \sigma_2^2} h(u)$. Hence, at least formally, the expected value with a negative variance is defined as

$$E_{-\sigma_2^2} h(u) = E_{-\sigma_1^2} h(u) = F^{-1} [e^{\sigma^2 \|\omega\|^2/2} \tilde{h}](\omega)(u).$$

Note that $E_{-\sigma_2^2} h(u)$ may not be well-defined in general. For a detailed discussion about $E_{-\sigma_2} h(u)$, we refer the reader to Shimodaira (2008). If $E_{-1} h(u)$ can be
defined, the $p$-value $p_{\text{AU}}(H|y) = \Phi(v + \mathcal{E}_{-1} h(u))$ has the second-order accuracy (Shimodaira, 2008):

$$\forall \mu \in \partial H; \ P(p_{\text{AU}}(H|Y) < \alpha \ | \mu) = \alpha + O(\lambda^2).$$

As with the classical large sample theory, if $\mathcal{E}_{-1} h^2(u)$ also exits, it can be shown that $p_{\text{AU}}(H|y)$ has the third-order accuracy with bias only $O(\lambda^{-3})$.

For the smooth $h$, it follows that $\mathcal{E}_{\sigma^2} h(u) = \sum_{j=0}^{\infty} \sigma^{2j} \beta_j(u)$. That is, letting $\beta_{H,0} = v + \beta_0(u)$ and $\beta_{H,j} = \beta_j(u)$ ($j \geq 1$), $\varphi_H(\sigma^2 | \beta_H)$ can be modeled as $\beta_{H,0} + \beta_{H,1} \sigma^2 + \beta_{H,2} (\sigma^2)^2 + \cdots$. Thus, using a polynomial regression with $\sigma^2$, we can compute the $p$-value $p_{\text{AU}}(H|y)$. In contrast, for a cone-shaped region $H$, it is shown that $\mathcal{E}_{\sigma^2} h(u) = \sum_{j=0}^{\infty} \sigma^{1-j} \beta_j(u)$. Since we have $\beta_j(u) = O(||u||^3)$ as $||u|| \rightarrow 0$, focusing on first two terms, we obtain

$$\mathcal{E}_{\sigma^2} h(u) \approx \beta_0(u) \sigma + \beta_1(u).$$

In this model, we cannot take $\sigma^2 = -1$, and $\mathcal{E}_{-1} h(u)$ does not exist for a cone-shaped region $H$. This observation is related to the important fact proved by Lehmann (1952) that an unbiased test cannot exist for a cone-shaped hypothesis region. Set $\beta_{H,0} = v + \beta_1(u)$ and $\beta_{H,1} = \beta_0(u)$, and then the normalized bootstrap $z$-value $\psi_{\sigma^2}(H|y)$ can be approximated by the model $\beta_{H,0} + \beta_{H,1} \sigma$. Here, we also denote by $\varphi_H(\sigma^2 | \beta_H)$ the model which approximates the normalized bootstrap $z$-value $\psi_{\sigma^2}(H|y)$. For fixed $\sigma_0^2 > 0$, let $\varphi_{H,k}(\sigma^2 | \beta_H, \sigma_0^2)$ be the truncated Taylor expansion of $\varphi_H(\sigma^2 | \beta_H)$ at $\sigma_0^2$:

$$\varphi_{H,k}(\sigma^2 | \beta_H, \sigma_0^2) = \frac{k-1}{j!} \frac{\partial^j \varphi_H(\sigma^2 | \beta_H)}{\partial (\sigma^2)^j} \bigg|_{\sigma^2 = \sigma_0^2}. \quad (A.1)$$

For fixed $k \in \mathbb{N}$, we consider the following $p$-value:

$$p_{\text{AU},k}(H|y) = \Phi(\varphi_{H,k}(-1|\beta_H, \sigma_0^2)).$$

Under some regularity conditions, Shimodaira (2008) proves that

$$\forall \mu \in \partial H; \ P(p_{\text{AU},k}(H|Y) < \alpha \ | \mu) \rightarrow \alpha + O(\lambda^2) \quad \text{as} \quad k \rightarrow \infty.$$

Moreover, for general selective inference with possibly non-smooth boundary surfaces, Terada and Shimodaira (2017) proposes the following selective $p$-value:

$$p_{\text{SI},k}(H|S,y) = \frac{\Phi(\varphi_{H,k}(-1|\beta_H, \sigma_{-1}^2))}{\Phi(\varphi_{H,k}(-1|\beta_H, \sigma_{-1}^2) + \varphi_{S,k}(0|\beta_S, \sigma_0^2))},$$

where $\sigma_{-1}^2, \sigma_0^2 > 0$. In addition, it is shown that the selective $p$-value has the second-order accuracy:

$$\forall \mu \in \partial H; \ \frac{P(p_{\text{SI},k}(H|S,Y) < \alpha \ | \mu)}{P(Y \in S \ | \mu)} \rightarrow \alpha + O(\lambda^2) \quad \text{as} \quad k \rightarrow \infty.$$
In regression analysis, since the hypothesis region has a flat surface, we can simply use the following $p$-value:

$$p_{SI}(H|S, y) = \frac{\Phi(\varphi_H(-1|\beta_H))}{\Phi(\varphi_H(-1|\beta_H) + \varphi_S(0|\beta_S))}.$$

In this setting, Terada and Shimodaira (2017) also proves that the $p$-value $p_{SI}(H|S, y)$ has the second-order accuracy:

$$\forall \mu \in \partial H; \quad \frac{P(p_{SI}(H|S, Y) < \alpha \mid \mu)}{P(Y \in S \mid \mu)} = \alpha + O(\lambda^2).$$