Selective inference after feature selection via multiscale bootstrap

Yoshikazu Terada[1] · Hidetoshi Shimodaira[2,3]

Abstract

It is common to show the confidence intervals or p-values of selected features, or predictor variables in regression, but they often involve selection bias. The selective inference approach solves this bias by conditioning on the selection event. Most existing studies of selective inference consider a specific algorithm, such as Lasso, for feature selection, and thus they have difficulties in handling more complicated algorithms. Moreover, existing studies often consider unnecessarily restrictive events, leading to over-conditioning and lower statistical power. Our novel and widely applicable resampling method via multiscale bootstrap addresses these issues to compute an approximately unbiased selective p-value for the selected features. As a simplification of the proposed method, we also develop a simpler method via the classical bootstrap. We prove that the p-value computed by our multiscale bootstrap method is more accurate than the classical bootstrap method. Furthermore, numerical experiments demonstrate that our algorithm works well even for more complicated feature selection methods such as non-convex regularization.

✉ Yoshikazu Terada
terada@sigmath.es.osaka-u.ac.jp

Hidetoshi Shimodaira
shimo@i.kyoto-u.ac.jp

[1] Graduate School of Engineering Science, Osaka University, 1-3 Machikaneyama-cho, Toyonaka, Osaka 560-8531, Japan

[2] Graduate School of Informatics, Kyoto University, Yoshida Honmachi, Sakyo-ku, Kyoto 606-8501, Japan

[3] Jointly affiliated at RIKEN Center for Advanced Intelligence Project (AIP), 1-4-1 Nihonbashi, Chuo-ku, Tokyo 103-0027, Japan

1 Introduction

In the classical statistical inference, the specification of a hypothesis is assumed to be 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 often ignore the effects of the hypothesis selection in the classical inference, and thus this naive approach will not provide a valid statistical inference. Recently, the selective (or post-selection) inference, which deals with the hypothesis selection effect appropriately, has drawn considerable attention not only in the statistical community but also in the machine learning community (e.g., Yang et al. 2016; Suzumura et al. 2017; Slim et al. 2019; Lim et al. 2020).

In this paper, we focus on the selective inference after the feature selection, i.e., predictor variable selection, in regression analysis. The most intuitive and straightforward 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 feature selection, and the other is used for the inference of the selected features. However, this approach reduces available data for both feature selection and inference. Fithian et al. (2014) provides the theoretical foundation to consider the optimality of the selective inference in the sense of statistical power. In Berk et al. (2013), without assuming a specific feature selection method, a valid selective inference after feature 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 feature 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 Lasso (Tibshirani 1996). More precisely, this selection event is 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 feature selection via Lasso. The significance levels conditioned on the selection event are computed by truncated normal distributions, justified by the *polyhedral lemma*. Tibshirani et al. (2016) develops a general framework to perform selective inference after any selection event that is 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 is explicitly represented as a union of polyhedra in the space of the response variable. Although the idea of the polyhedral lemma is, in fact, valid for any selective sets beyond a union of polyhedra (Liu et al. 2018), the existing approaches have computational difficulties in handling more complicated algorithms with non-convex penalties such as MCP (minimax concave penalty; Zhang 2010) and SCAD (smoothly clipped absolute deviation; Fan and Li 2001), where the selective sets

become more complicated than the ordinary Lasso. Although the selective inference of Berk et al. (2013) is not limited to specific feature selection methods, the computation cost may be prohibitive for the number of variables over 20. In addition, it controls type-I errors simultaneously under all submodels, thus leading to very conservative confidence intervals. Moreover, most existing selective inference with the full-model view is unnecessarily *over-conditioning* and lower statistical power because the inference is conditioned on a selected model, whereas it could be minimally conditioned on a selected feature. The selective set of the minimally conditioning event becomes more complicated and computationally difficult, and thus its valid post-selection inference is implemented recently by Liu et al. (2018) first time but only for the ordinary Lasso case.

Recently, Terada and Shimodaira (2017) extends the general hypothesis testing framework, called *the problem of regions* (Efron and Tibshirani 1998), to the selective inference, and proposes a new selective inference approach via *multiscale bootstrap* of Shimodaira (2002, 2004, 2008). This approach is not based on the polyhedral lemma, and we can easily compute approximately unbiased selective *p*-values for hypotheses conditioned on complicated selective sets. 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 are represented as regions in some parameter space. This approach can be widely applied because we do not need to know the shapes of these regions, but only need to prepare functions that tell whether these regions include a realization of the parameter estimate. In fact, Shimodaira and Terada (2019) describes an application of this approach for testing trees and edges in phylogenetics. Moreover, based on our idea described in this paper and Terada and Shimodaira (2017), Lim et al. (2020) develops the powerful selective inference after feature selection using the Hilbert–Schmidt Independence Criterion and the Maximum Mean Discrepancy.

In the original form of 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 of the hypothesis and selection event. However, this multiscale bootstrap method cannot be directly applied to selective inference after feature selection since the shape of the selective region is unwillingly related to the sample size in the feature 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 it can be applied to almost any feature selection algorithm. In addition, the computational complexity of our method is the same order as the classical bootstrap method.

This paper is organized as follows. In Section 2, we describe the setting of selective inference after feature selection. In Section 3, we give a brief exposition of multiscale bootstrap and the general selective inference via multiscale bootstrap. 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 numerical experiments.

2 Selective inference after feature selection

We briefly describe the selective inference after feature selection in linear regression; we will give the setting in Section 4 with details. Let $Z = (Z_1, \ldots, Z_n)^T$ be the response variable with mean $\xi \in \mathbb{R}^n$ and variance $\tau^2 I_n$. Let $x_1, \ldots, x_p \in \mathbb{R}^n$ be non-random features (i.e., predictor variables) and $X = (x_1, \ldots, x_p) = (x_{ij})_{n \times p}$. At first, we need to clarify the target of statistical inference. If we assume the following first-order correctness:

$$\exists \beta^* \in \mathbb{R}^p; \ \xi = X\beta^*,$$

the target of the estimators is clearly the "true" coefficients β^*. However, in real data analysis, it is difficult to assume this correctness, as mentioned in Box's famous quote "All models are wrong". Even under the first-order correctness, the selected models may be wrong. Without the first-order correctness, the target of inference is ambiguous. Berk et al. (2013) clarifies the target of statistical inference in linear regression, and there are two possible types of the target:

– Let $M \subseteq \{1, \ldots, p\}$ be a specified set of features. We consider a submodel using the features in M. The target for the submodel view with respect to M is

$$\beta^{(M)} := \arg\min_{b_M \in \mathbb{R}^{\#(M)}} \mathbb{E}\big[\|Z - X_M b_M\|^2\big] = (X_M^T X_M)^{-1} X_M^T \xi,$$

where X_M is the predictor matrix consisting of the features of M. The true value of $\beta_j^{(M)}$ ($j \in M$) in submodel view depends on M.

– The target for the full-model view is $\beta = \beta^{(M)}$ with $M = \{1, \ldots, p\}$. Thus

$$\beta = (\beta_1, \ldots, \beta_p) = \arg\min_{b \in \mathbb{R}^p} \mathbb{E}\big[\|Z - Xb\|^2\big] = (X^T X)^{-1} X^T \xi.$$

The true value of β_j in the full-model view does not depend on M. In this paper, the target for our method is the coefficients in the full-model view, while both views are discussed below.

Now, we describe the basic concepts of selective inference in regression. Let α be a significant level in selective inference. When the null hypothesis H_0 is selected based on data, we should control the selective type I error rate:

$$P_{H_0}(H_0 \text{ is rejected} \mid H_0 \text{ is selected}) \le \alpha, \tag{1}$$

where P_{H_0} is a probability distribution under H_0. The event $\{H_0 \text{ is selected}\}$ is called the selection event. After feature selection based on data, we obtain the selected model $\hat{M} \subseteq \{1, \ldots, p\}$. Then, depending on whether the target is in the submodel view or the full-model view, the null hypotheses for $j \in \hat{M}$ is $H_0 : \beta_j^{(\hat{M})} = 0$ or $H_0 : \beta_j = 0$, respectively. Here $\beta_j^{(\hat{M})} = 0$ is a simplified notation of $\beta_j^{(M)} = 0$ with $\hat{M} = M$. Moreover, we may consider two different types of events $\{j \in \hat{M}\}$ or $\{\hat{M} = M\}$ as the selection event.

However, the event $\{j \in \hat{M}\}$ is not an appropriate selection event for the hypothesis of submodel view $H_0 : \beta_j^{(\hat{M})} = 0$, because this hypothesis depends on the other selected features $\hat{M} \setminus \{j\}$ and thus the probability (1) does not make sense. Therefore, for the hypothesis of submodel view, the event $\{\hat{M} = M\}$ or more restrictive events are appropriate as a selection event; the event $\{\hat{M} = M, \text{sign}(\hat{\beta}^{(M)}) = s\}$ is sometimes considered as the selection event (e.g., Lee et al. 2016) for computational reason.

On the other hand, we can consider the two different types of conditioning $\{j \in \hat{M}\}$ and $\{\hat{M} = M\}$ $(j \in M)$ as a selection event for the hypothesis of full-model view $H_0 : \beta_j = 0$. Since both of these two events are appropriate, we may wonder which of them is more desirable. This is answered by the argument of the monotonicity of selective error in Proposition 3 of Fithian et al. (2014). Here we see its adaptation to our setting. Since $\{j \in \hat{M}\} = \bigcup_{M : j \in M} \{\hat{M} = M\}$, we have

$$P_{H_0}(H_0 : \beta_j = 0 \text{ is rejected} \mid j \in \hat{M})$$

$$= \sum_{M : j \in M} P_{H_0}(H_0 : \beta_j = 0 \text{ is rejected} \mid \hat{M} = M) \frac{P_{H_0}(\hat{M} = M)}{P_{H_0}(j \in \hat{M})}$$

$$\leq \max_{M : j \in M} P_{H_0}(H_0 : \beta_j = 0 \text{ is rejected} \mid \hat{M} = M).$$

If we control the selective type-I error $P_{H_0}(H_0 : \beta_j = 0 \text{ is rejected} \mid \hat{M} = M)$ at level α for all models M $(j \in M)$, the selective type-I error $P_{H_0}(H_0 : \beta_j = 0 \text{ is rejected} \mid j \in \hat{M})$ is also automatically controlled at level α. Thus, this monotonicity tells us that the over-conditioning leads to a loss of information and that, for the hypothesis $H_0 : \beta_j = 0$, the minimal selection event $\{j \in \hat{M}\}$ (i.e., the minimally conditioning and thus the maximal event set) is the most desirable in the sense of statistical power.

3 An overview of multiscale bootstrap

First, we describe the framework of the problem of regions in Section 3.1, followed by the basic idea of multiscale bootstrap for non-selective inference (Shimodaira 2002, 2004, 2008) in Section 3.2. Then we briefly introduce the general selective inference framework proposed by Terada and Shimodaira (2017) in Section 3.3.

3.1 The problem of regions

The general statistical inference framework, 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 framework is an abstraction of many applications, e.g., phylogenetic inference, in which a confidence level is assigned for each clade of the estimated phylogenetic tree (Felsenstein 1985; Efron et al. 1996).
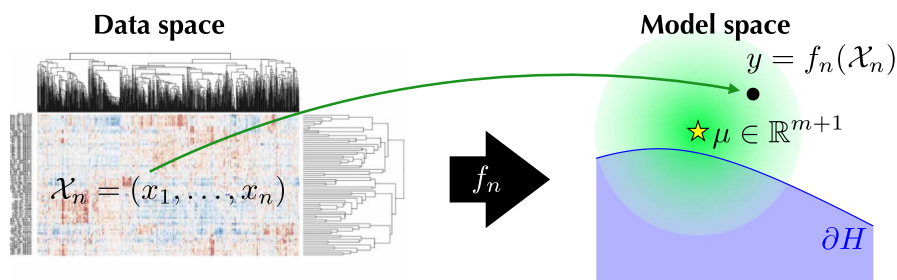
Fig. 1 Relationship between the data space and the model space

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

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

Typically, f_n involves multiplying the factor $\sqrt{n}$ 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 $y \in \mathbb{R}^{m+1}$ be an observed value of Y, and suppose that a bootstrap sample $\mathcal{X}_{n'}^* = (X_1^*, \ldots, X_{n'}^*)$ with sample size n' is represented as a realization of the following Gaussian distribution in the model space:

$$Y^* = f_n(\mathcal{X}_{n'}^*) \sim N_{m+1}(y, \sigma^2 I_{m+1}), \quad \sigma^2 = \frac{n}{n'}.$$

We will denote by $P_{\sigma^2}(\cdot|y)$ the probability measure of the 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 null hypothesis. It is assumed that the region H, in a neighborhood of the model space, is locally represented as $H = \{(u, v) \mid v \leq -h(u), \ u \in \mathbb{R}^m, v \in \mathbb{R}\}$ 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 goal is to compute an approximately unbiased p-value $p(H|y)$ for the null hypothesis $H_0 : \mu \in H$ against the alternative hypothesis $H_1 : \mu \notin H$. The approximately unbiased p-value should satisfy

$$\forall \mu \in \partial H; P(p(H|Y) < \alpha \mid \mu) \approx \alpha \tag{2}$$

for a given significance level $\alpha > 0$. In other words, the p-value is approximately distributed as uniform over $(0,1)$, i.e., $p(H|Y) \sim U(0,1)$ when $\mu \in \partial H$. The difference between $P(p(H|Y) < \alpha \mid \mu)$ and α in (2) is called *bias* (or error). The bootstrap probability

$$p_{\mathrm{BP}}(H|y) := P_1(Y^* \in H|y)$$

is considered as the most simple p-value satisfying (2); see Efron et al. (1996); Efron and Tibshirani (1998). More formally, in the classical large sample theory, if the region H has a smooth boundary surface, the bootstrap probability $p_{\mathrm{BP}}(H|y)$ has the first-order accuracy: $\forall \mu \in \partial H; P(p_{\mathrm{BP}}(H|Y) < \alpha \mid \mu) = \alpha + O(n^{-1/2})$. However, in many practical situations, the bootstrap probability p_{BP} often has a non-negligible bias.

3.2 Basic idea of multiscale bootstrap

To obtain more accurate p-values, geometric quantities in the model space, such as distance and curvature, play a key role. In fact, Efron and Tibshirani (1998) shows that we can compute a more accurate p-value using the signed distance $v(H|y)$ from the data point y to the region H. More precisely, the p-value $p_{\mathrm{sign}}(H|y) := P_1(v(H|Y^*) \geq v(H|y) \mid \hat{\mu}(y))$ is proposed, where $\hat{\mu}(y) \in \partial H$ is the projection point of y onto ∂H. This p-value $p_{\mathrm{sign}}(H|y)$ 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 $v(H|y)$ 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 $v(H|y)$ and the mean curvature of ∂H are estimated based on the scaling law of the bootstrap probabilities, and an accurate p-value is computed based on these estimated quantities.

We consider the bootstrap probability with scale $\sigma > 0$

$$\alpha_{\sigma^2}(H|y) := P_{\sigma^2}(Y^* \in H|y),$$

which reduces to $p_{\mathrm{BP}}(H|y)$ when $\sigma = 1$. When we change n' in the data space, and in effect, change σ in the model space, the bootstrap probability changes. This change is simply expressed in terms of the normalized bootstrap z-value defined as

$$\psi_{\sigma^2}(H|y) := \sigma \bar{\Phi}^{-1}(\alpha_{\sigma^2}(H|y)),$$

where $\Phi(x)$ is the cumulative distribution function of the standard normal distribution and $\bar{\Phi}(x) = 1 - \Phi(x)$, i.e., $\bar{\Phi}^{-1}(\alpha)$ is the upper α-value. Shimodaira (2002), Shimodaira (2004) show the following scaling law of the bootstrap probabilities:

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

where $\gamma(H|y)$ is the mean curvature of the boundary surface ∂H at $\hat{\mu}$. This scaling law can be modeled as the simple linear regression $\theta_{H,0} + \theta_{H,1}\sigma^2$ with σ^2 as the predictor. We will denote by $\varphi_H(\sigma^2|\theta_H)$ the model for the normalized bootstrap z-value, such as $\theta_{H,0} + \theta_{H,1}\sigma^2$ with parameter $\theta_H = (\theta_{H,0}, \theta_{H,1})$. The bootstrap probabilities

with several values $\{\sigma_j\}$ of scale can be computed by using the bootstrap samples with different sample sizes, say $n'_j \in \{\lceil 0.5n \rceil, \cdots, \lceil 1.0n \rceil, \cdots, \lceil 1.5n \rceil\}$. Let B be the number of bootstrap replicates, and $C_H = \#\{Y^* \in H\}$ be the frequency to be $Y^* \in H$. Let $\hat{\psi}_{\sigma_j^2}(H|y)$ 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 $v(H|y) = \theta_{H,0}$ and $\gamma(H|y) = \theta_{H,1}$ by the simple regression for the observed $\{(\sigma_j^2, \hat{\psi}_{\sigma_j^2}(H|y))\}$.

Shimodaira (2002) proposes the following p-value:

$$p_{\text{AU}}(H|y) := \bar{\Phi}(\varphi_H(-1|\theta_H)) = \bar{\Phi}(v(H|y) - \gamma(H|y)) + O_p(n^{-1}).$$

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

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

It becomes third-order accurate erring only $O(n^{-3/2})$ when $\varphi_H(\sigma^2|\theta_H)$ 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 property is that the smooth boundary surface ∂H approaches a flat surface in a neighborhood of any point on ∂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. It is well known that there is no unbiased test for a hypothesis region with a non-smooth boundary (Lehmann 1952). To deal with general regions with non-smooth boundary surfaces, Shimodaira (2008) develops a new theoretical framework, called the asymptotic theory of *nearly flat surfaces*. In this framework, we require that the magnitude of boundary surfaces is small. Thus, this framework works well even for non-smooth boundary surfaces when the magnitude of boundary surfaces is not so large, at least locally. We can interpret that the given surface is on the way to approaching the flat surface in this framework. This idea is similar to one behind the local alternative framework (Lehmann 1999), but the rescaling is applied only to the direction normal to the boundary surface while the scale is fixed for the other directions. A brief introduction of this theory is provided in Appendix A.

3.3 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 $H = \{(u, v) \mid v \le -h(u), u \in \mathbb{R}^m, v \in \mathbb{R}\}$ and $S = \{(u, v) \mid v > -s(u), u \in \mathbb{R}^m, v \in \mathbb{R}\}$ are considered. Suppose that the selection event is represented as $\{y \in S\}$, and we consider the selective inference in which the null hypothesis $H_0 : \mu \in H$ is selected if and only if $y \in S$. In this setting, for a given significance level α, we want to compute selective p-values $p(H|S, y)$ satisfying

$$\forall \mu \in \partial H; \frac{P(p(H|S, Y) < \alpha \mid \mu)}{P(Y \in S \mid \mu)} \approx \alpha.$$

In other words, $p(H|S, Y) \sim U(0, 1)$ conditioned on $Y \in S$ when $\mu \in \partial H$. Terada and Shimodaira (2017) proposes the following approximately unbiased selective p-value $p_{\text{SI}}(H|S, y)$ for regions H and S with smooth boundary surfaces:

$$p_{\text{SI}}(H|S, y) := \frac{\bar{\Phi}(\varphi_H(-1|\theta_H))}{\bar{\Phi}(\varphi_H(-1|\theta_H) + \varphi_S(0|\theta_S))},$$

where φ_S is the model for the normalized bootstrap z-value related to the selective region S, and θ_S is the parameter of the model φ_S.

Theorem 1 (Theorem 4.3 in Terada and Shimodaira 2017) *The boundary surfaces ∂H and ∂S are assumed to be sufficiently smooth and nearly parallel in the sense that the first derivatives of h and s differ only $O(n^{-1})$. Then, the selective p-value $p_{\text{SI}}(H|S, y)$ has the second-order accuracy*:

$$\forall \mu \in \partial H; \frac{P(p_{\text{SI}}(H|S, Y) < \alpha \mid \mu)}{P(Y \in S \mid \mu)} = \alpha + O(n^{-1}).$$

The detailed calculation of $p_{\text{SI}}(H|S, y)$ is provided as Algorithm 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. Since the boundary surface ∂S of the selection region is generally not smooth in the selective inference after feature selection, the theory of nearly flat surfaces is used to derive the properties of the proposed method described in Section 4.3. Roughly speaking, both theories essentially assume that the boundary surfaces ∂H and ∂S are more or less flat and parallel to each other, at least locally around the data point. In analogy with the local alternative framework, we consider that given surfaces are on the way to approaching mutually parallel flat surfaces.

4 Selective inference after feature selection via multiscale bootstrap

4.1 Model space for regression analysis

Here, we describe the setting of the selective inference after feature 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 $Z = (Z_1, \dots, Z_n)$ drawn from the multivariate Gaussian distribution:

$$Z \sim N_n(\xi, \tau^2 I_n),$$

where $\xi \in \mathbb{R}^n$ is an unknown parameter, I_n is the n-dimensional identity matrix, and τ^2 is assumed to be known. We will denote by $z \in \mathbb{R}^n$ the observed value of Z. Let $X = (x_1, \dots, x_p) = (x_{ij})_{n \times p}$ be a non-random full rank matrix whose columns represent the features. Note that the error variance τ^2 can be estimated if ξ is modeled as

a function of features $x_1, \dots, x_p \in \mathbb{R}^n$. Assuming a specific feature selection method, such as Lasso and MCP, is applied to $(X, z) \in \mathbb{R}^{n \times p} \times \mathbb{R}^n$, let $\hat{M} \subseteq \{1, \dots, p\}$ be the set of selected features, and $\hat{s}_j \in \{+, -\}$ be the sign of the estimated coefficient $\hat{\beta}_j$ of the feature $j \in \hat{M}$.

First, note that the general selective inference approach described as Algorithm 1 cannot be directly applied to regression analysis. Since we need to change the sample size n' of bootstrap samples in the usual multiscale bootstrap, it is assumed that the hypothesis and selective regions can be represented as specific regions, which are independent of n', in the model space. For the selective inference in regression analysis, however, the shape of the selective region inevitably depends on n' because it is the dimensionality of the model space as explained below.

We recall that it is assumed that $Z \sim N_n(\xi, \tau^2 I_n)$ and that the selection event can be represented as the region of the space of Z. Then, it is realized that the normalized space of

$$Y := Z/\tau, \quad \mu := \xi/\tau \tag{3}$$

can be considered as the model space described in Section 3 with $m + 1 = n$. Thus, the selective region for multiscale bootstrap inevitably depends on n'. Another choice of model space is given by

$$Y := \tau^{-1} B Z, \quad \mu := \tau^{-1} B \xi,$$

where $B = (X^T X)^{-1/2} X^T \in \mathbb{R}^{p \times n}$. The selective region represented in this model space also depends on n' because feature selection algorithms take account of sample size. Although the latter model space is preferable for the asymptotic theory because it has the fixed dimensionality $m + 1 = p$, we use the former model (3) below for easy illustration.

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, 10000.
2: For each n', perform bootstrap resampling to generate Y^* for B times and compute $\alpha_{\sigma^2}(H|y) = C_H/B$ and $\alpha_{\sigma^2}(S|y) = C_S/B$ by counting the frequencies $C_H = \#\{Y^* \in H\}$ and $C_S = \#\{Y^* \in S\}$. We may actually work on $\mathcal{X}_{n'}^*$ instead of Y^*. Compute $\psi_{\sigma^2}(H|y) = \sigma\bar{\Phi}^{-1}(\alpha_{\sigma^2}(H|y))$ and $\psi_{\sigma^2}(S|y) = \sigma\bar{\Phi}^{-1}(\alpha_{\sigma^2}(S|y))$.
3: Estimate parameters $\theta_H(y)$ and $\theta_S(y)$ by fitting models $\psi_{\sigma^2}(H|y) = \varphi_H(\sigma^2|\theta_H)$ and $\psi_{\sigma^2}(S|y) = \varphi_S(\sigma^2|\theta_S)$, respectively. The parameter estimates are denoted as $\hat{\theta}_H(y)$ and $\hat{\theta}_S(y)$. If we have several candidate models, apply above to each and choose the best model based on AIC value.
4: Approximately unbiased p-values of non-selective inference (p_{AU}) and of selective inference (p_{SI}) are computed by $p_{\mathrm{AU}}(H|y) = \bar{\Phi}(z_H)$ and $p_{\mathrm{SI}}(H|S, y) = \bar{\Phi}(z_H)/\bar{\Phi}(z_H + z_S)$, where $z_H = \varphi_H(-1|\hat{\theta}_H(y))$ and $z_S = \varphi_S(0|\hat{\theta}_S(y))$.

4.2 Appropriate selection event

Recently, Liu et al. (2018) suggests the use of the selection event $\{j \in \hat{M}\}$ for a specified $j \in \{1, \dots, p\}$, which increases the statistical power and thus leading to shorter confidence intervals. This is explained by the monotonicity of the selective error provided in Fithian et al. (2014), as mentioned in Section 2. The event $\{\hat{M} = M\}$ for a specified $M \subset \{1, \dots, p\}$ is over-conditioning and reducing the statistical power because the other features $M \setminus \{j\}$ are not relevant for the null hypothesis $H_0 : \beta_j = 0$ of the feature j in the full-model view.

Here, we actually consider testing feature j with its sign. More precisely, whenever the feature j is selected and $\hat{\beta}_j > 0$ (or < 0), the hypothesis $H_0 : \beta_j \leq 0$ (or ≥ 0) is tested. The minimal selection event is then $\{j \in \hat{M}, \hat{s}_j = s_j\}$ where $s_j \in \{+, -\}$. Hence, the main goal of our selective inference is to compute the unbiased selective p-value $p_j(y)$, which satisfies

$$\mathbb{P}(p_j(Y) < \alpha \mid j \in \hat{M} \text{ and } \hat{s}_j = s_j) = \alpha \qquad (4)$$

for any μ with $\beta_j = 0$.

4.3 Computing selective *p*-values by multiscale bootstrap

In this section, we describe our proposed method. We develop a new algorithm to compute the approximately unbiased selective p-value, which approximately satisfies the equation (4). We will update the computation of $\psi_{\sigma^2}(H|y)$ and $\psi_{\sigma^2}(S|y)$ in Algorithm 1 to obtain Algorithm 2.

For the feature selection via Lasso, the selection event $\{j \in \hat{M}, \hat{s}_j = s_j\}$ $(j \in \{1, \dots, p\}, s_j \in \{+, -\})$ 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 Lasso and the corresponding region in the response vector space when $n = 2$. In contrast, for more complicated feature selection methods such as MCP and SCAD, the region S of the selective event $\{j \in \hat{M}, \hat{s}_j = s_j\}$ will become complicated, and the explicit shape of the selective region S may not be obtained easily. The right panel of Figure 2 shows the relationship between the selected model by MCP and the corresponding region in the response vector space when $n = 2$. We had to numerically evaluate which features are selected at each point since no explicit representation of the selection event is available. Although Lee et al. (2016) and Tibshirani et al. (2016) consider exact selective inference for Lasso, it is difficult to consider exact selective inference for these complicated feature selection methods.

As shown in Figure 2, the selective region S which represents the selection event $\{j \in \hat{M}, \hat{s}_j = s_j\}$ could be complicated and has generally non-smooth boundary surfaces. In contrast, for $\eta \in \mathbb{R}$, the hypotheses $H_0 : \beta_j \lesseqgtr \eta$, namely the two cases of $\beta_j \leq \eta$ and $\beta_j \geq \eta$, can be represented as the following regions in the space of Y:
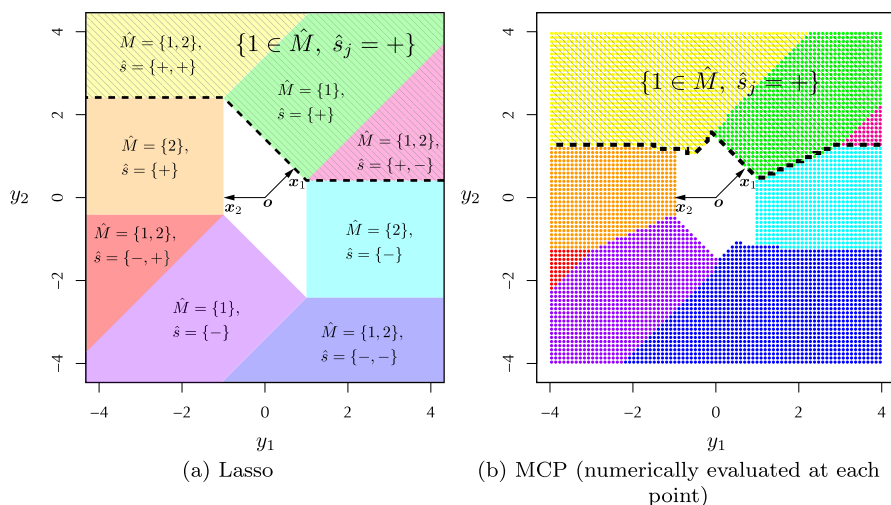
Fig. 2 Selective regions of feature selection ($n = 2$). The shaded region is corresponding to the event $\{1 \in \hat{M}, \hat{s}_1 = +\}$) which is a union of regions for the selection events of the form $\{\hat{M} = M, \hat{s}_M = s_M\}$. (a) For Lasso, each region is represented as a polyhedron. (b) For MCP, each region has a more complicated shape. Our selective inference method works easily in both cases

$$H = \{z/\tau \in \mathbb{R}^n \mid a_j^T z \lessgtr \eta\}, \quad A = [a_1, \ldots, a_p]^T = (X^T X)^{-1} X^T.$$

Since the hypothesis region H has a flat boundary surface with mean curvature $\gamma(H|y) = 0$, we can easily obtain the expression of $\psi_{\sigma^2}(H|y)$ without multiscale bootstrap. In particular, for $H_0 : \beta_j \lessgtr 0$,

$$\psi_{\sigma^2}(H|y) = v(H|y) = \pm\tau^{-1} a_j^T z / \|a_j\|_2, \tag{5}$$

where $v(H|y)$ is the signed distance from y to the hypothesis region H. Thus, once we obtain the expression of $\psi_{\sigma^2}(S|y)$, we can compute the selective p-value $p_{\mathrm{SI}}(H|S, y)$.

Next, we consider how to obtain the expression of $\psi_{\sigma^2}(S|y)$ for the selective region S by multiscale bootstrap; the bootstrap probability of S is computed at several scales via an adaptation of Step 2 of Algorithm 1. In the usual setting of multiscale bootstrap, changing n' in the data space corresponds to changing σ in the model space. Thus, based on bootstrap probabilities with several n', we can estimate the expression of $\psi_{\sigma^2}(S|y)$. However, as described in Section 4.1, this framework is not applicable here. Fortunately, since we can access the model space, we directly change the scale σ in the model space and compute the bootstrap probabilities with several scales. With the normality of response Z, the parametric bootstrap method, i.e., sampling directly from $N_n(y, \sigma^2 I_n)$, could be applied to the computation of bootstrap probabilities $\alpha_{\sigma^2}(S|y)$ at several scales $\sigma > 0$. To relax the Gaussian assumption, here we consider the resampling of residuals with scale change. More formally, we resample the scaled residuals to compute $\alpha_{\sigma^2}(S|y)$ at several $\sigma > 0$ as follows. Let

$\hat{\beta}^{(LS)} = (X^T X)^{-1} X^T z$ be the least-squares estimator based on the full-model. Write $\hat{e} := z - X\hat{\beta}^{(LS)}$ 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 $\hat{\epsilon}_i = \hat{e}_i / \sqrt{1 - h_i}$. To compute the bootstrap probability $\alpha_{\sigma^2}(S|y)$ at $\sigma > 0$, we use the following bootstrap sample:

$$z_\sigma^* = X\hat{\beta}^{(LS)} + \sigma\hat{\epsilon}^*, \tag{6}$$

where $\hat{\epsilon}^* = (\hat{\epsilon}_1^*, \ldots, \hat{\epsilon}_n^*)^T$ is a bootstrap sample with size n from $(\hat{\epsilon}_1, \ldots, \hat{\epsilon}_n)$. For each $\sigma > 0$, we generate z_σ^* for B times, and apply a particular model selection procedure to them for computing $\alpha_{\sigma^2}(S|y) = C_S/B$ by counting the frequency of the selective event $\{j \in \hat{M}, \hat{s}_j = s_j\}$.

Algorithm 2 Approximately unbiased selective p-value for the selected feature $j \in \hat{M}$ with $\hat{s} = s_j$

1: Specify several $\sigma > 0$ values. Set the number of bootstrap replicates B, say, 10000.
2: Compute the adjusted residuals $\hat{\epsilon}$.
3: For each σ, perform bootstrap resampling to generate z_σ^* by (6) for B times and compute $\alpha_{\sigma^2}(S|y) = C_S/B$ by counting the frequency $C_S = \#\{j \in \hat{M}_\sigma^* \text{ and } \hat{s}_{j,\sigma}^* = s_j\}$ where $\hat{M}_\sigma^*$ and $\hat{s}_{j,\sigma}^*$ are the set of selected features and the estimated sign of feature j by applying the specific algorithm to (X, z_σ^*), respectively. Compute $\psi_{\sigma^2}(S|y) = \sigma\bar{\Phi}^{-1}(\alpha_{\sigma^2}(S|y))$.
4: Estimate parameters $\theta_S(y)$ by fitting model $\psi_{\sigma^2}(S|y) = \varphi_S(\sigma^2|\theta_S)$.
5: Compute the selective p-value by

$$p_{\text{SI}}(H|S, y) = \frac{\bar{\Phi}(z_H)}{\bar{\Phi}(z_H + z_S)},$$

where $z_H = s_j(a_j^T z/\|a_j\|_2)/\tau$, $z_S = \varphi_S(0|\hat{\theta}_S(y))$, and $\hat{\theta}_S(y)$ is the estimated value of $\theta_S(y)$.

With the updated computation of $\psi_{\sigma^2}(H|y)$ and $\psi_{\sigma^2}(S|y)$ for regression analysis, we may use Step 3 and Step 4 of Algorithm 1 to compute the selective p-value. This becomes Algorithm 2 for computing an approximately unbiased selective p-value for the selected feature $j \in \hat{M}$. It is worth noting that Algorithm 2 can be applied to almost any feature selection methods, including MCP and SCAD, in addition to Lasso. Note also that the multiscale bootstrap is not very sensitive to the choice of the scales. For B, several thousand replications are enough in practice. The computational cost is the same order as the classical bootstrap method. Also note that, in actual implementation of Algorithm 2, the Steps 1 to 3 are shared by all the features $j \in \hat{M}$. Thus, this algorithm works even for large p such as $p > 20$.

Now, we provide the theoretical justification of the proposed algorithm. Since the boundary surface of the selective region is generally non-smooth as shown in Figure 2, we consider the asymptotic theory of nearly flat surfaces. In the model space, we take the coordinate system $(u, v) \in \mathbb{R}^{n-1} \times \mathbb{R}$ such that the hypothesis region H can be written by $\{(u, v) \mid v \leq 0\}$. Using this coordinate system, let us denote the selective region by $S = \{(u, v) \mid v_s - s(u) \leq v\}$ at least locally in a neighborhood of y,

where $v_s \in \mathbb{R}$ and s is a function from $\mathbb{R}^{n-1}$ to $\mathbb{R}$ which represents the boundary surface of the selective region. Here, the L^1-norm and L^∞-norm of function s are defined as $\|s\|_1 = \int |s(u)|\, du$ and $\|s\|_\infty = \sup_u |s(u)|$, respectively. Let $\lambda = \|s\|_\infty$ be the magnitude of the boundary surface ∂S of the selective region. Even in regression analysis, we assume that the selection event can be written as $S = \{(u,v) \mid v_s - s(u) \le v\}$ and the magnitude of the boundary surface ∂S is relatively small at least around the data point y in the model space. That is, we will consider the asymptotic theory in which $\lambda \to 0$. In the same way as the local alternative framework, this assumption can be interpreted as that the given surface is on the way to approach the flat surface which is parallel to ∂H. In this paper, this assumption is called nearly flatness of the boundary surface. In the asymptotic theory of nearly flat surfaces, the proposed p-value $p_{\mathrm{SI}}(H|S, y)$ has the second-order accuracy.

Theorem 2 *Let us denote the selective region as* $S = \{(u, v) \mid v \ge v_s - s(u)\}$. *Let* $\tilde{s}$ *be the Fourier transform of the function* s. *Suppose that the L^1-norms* $\|s\|_1$ *and* $\|\tilde{s}\|_1$ *are bounded. Let us assume that* $\lambda = \|s\|_\infty$ *is sufficiently small. Then, the selective p-value described in Algorithm 2 has the second-order accuracy*:

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

Theorem 2 can be considered as a special case of Theorem 5.3 in Terada and Shimodaira (2017). In the current situation, we can directly obtain the signed distance from y to the boundary surface ∂H. Thus, the proof of Theorem 2 is much simpler than Theorem 5.3 in Terada and Shimodaira (2017). The proof is given in Appendix B. This theorem provides a theoretical justification for the proposed p-value when the magnitude of the boundary surface is small, at least locally around the data point. The assumption requires that the boundary surface ∂S of the selection event is nearly flat and its limiting hyperplane is parallel to the surface ∂H of the hypothesis region, at least locally around the data point. Through the numerical simulations in the next section, it seems that this assumption is reasonably satisfied in practice. Of course, when the two surfaces ∂H and ∂S are not very parallel to each other, the proposed p-value has a non-negligible bias. Solving this problem is an important future work of this research.

Remark 1 In contrast with the p-value $p_{\mathrm{SI}}(H|S, y)$ in Algorithm 2, we also propose a simple selective p-value based on the classical bootstrap probability $\alpha_1(S|y)$. We replace $z_S = \varphi_S(0|\hat{\theta}_S(y))$ in Step 5 by $z'_S = \Phi^{-1}(\alpha_1(S|y))$ to define

$$p_{\mathrm{SI\text{-}BP}}(H|S, y) := \frac{\bar{\Phi}(z_H)}{\bar{\Phi}(z_H + z'_S)}.$$

It is also computed with $z'_S = \psi_1(S|y) \approx \varphi_S(1|\hat{\theta}_S(y))$. If the boundary surface of S is flat, $z'_S = z_S$ and thus $p_{\mathrm{SI\text{-}BP}} = p_{\mathrm{SI}}$. Under the assumption of Theorem 2, we can obtain

$$\forall \mu \in \partial H; \ \frac{P(p_{\text{SI-BP}}(H|S,Y) < \alpha \mid \mu)}{P(Y \in S \mid \mu)} = \alpha + O(\lambda),$$

indicating that $p_{\text{SI-BP}}$ has a larger bias than p_{SI}. This result can be proved in much the same way as Theorem 2.

5 Numerical experiments

Here, we show some numerical experiments to demonstrate the usefulness of our method. For Lasso, the exact unbiased selective test conditioned on $j \in \hat{M}$ and $\hat{s}_j = s_j$ ($j \in \{1, \ldots, p\}$, $s_j \in \{+, -\}$) can be constructed (Lee et al. 2016; Liu et al. 2018). At first, we will show that our selective p-value with multiscale bootstrap (i.e., p_{SI}) approximates the exact selective p-value for Lasso. Here, Lasso is defined as $\min_{\beta \in \mathbb{R}^p} \|z - X\beta\|_2^2/2 + \sum_{j=1}^p \rho|\beta_j|$. Set $(n, p) = (50, 25)$ and $\beta = (2, 2, 2, 2, 2, 0, \ldots, 0)^T \in \mathbb{R}^p$; $\beta_j = 0$ for the features $j = 6, \ldots, 25$. The elements x_{ij} of the input matrix X were independently generated from the standard normal distribution $N(0, 1)$. Then, the response $z \in \mathbb{R}^n$ was generated as $z = X\beta + \epsilon$, where ϵ was generated from the n-dimensional standard normal distribution $N_n(0_n, I_n)$. In our algorithm, we used $\sigma^2 = 0.5, 0.6, \ldots, 1.5$ as the scales and $B = 10^4$ as the number of bootstrap replicates. Here, we note that the choice of scales has only little effect on the stability of the result. The experiment about this point can be found in Appendix C. We simulated 2000 independent datasets. We set the significance level $\alpha = 5\%$. For computing false positive rates accurately, the variables with zero coefficients need to be selected several hundred times. Here, we used Lasso with the penalty parameter $\rho = 10$ as the feature selection method. Each variable with zero coefficient was selected approximately 250 times out of 2000 datasets in this experiment. In each dataset, we performed the classical t-test, the selective (one-sided) test conditioned on $\hat{M} = M$ and $\hat{s}_M = s_M$ for $M \subset \{1, \ldots, p\}$, $s_M \in \{+, -\}^{|M|}$ (Lee et al. 2016; only for Lasso), the selective (two-sided) test conditioned on $j \in \hat{M}$ for $j \in \{1, \ldots, p\}$ (Liu et al. 2018; only for Lasso) and our approximately unbiased (one-sided) test with multiscale bootstrap for each selected feature. We count how many times, say N_j, the feature j is selected. For each test, we also count how many times (say R_j) the null hypothesis $H_0 : \beta_j \lesseqgtr \eta$ is rejected, and the selective rejection probability is estimated by R_j/N_j.

The panel (a) of Figure 3 shows the selective rejection probabilities of each feature for Lasso, where the four test methods are compared. In this plot, we can see that the selective rejection probabilities of our test with p_{SI} for the features 6 to 25 with $\beta_j = 0$ are around 5%. Thus, it is shown that our multiscale bootstrap method approximately satisfies the unbiasedness in the sense of the equation (4). We can also see that the classical inference does not provide a valid inference after feature selection; the classical t-test gives more false positives than expected from the specified α level. Moreover, instead of the Lasso penalty $\rho|\beta_j|$, we also used the following
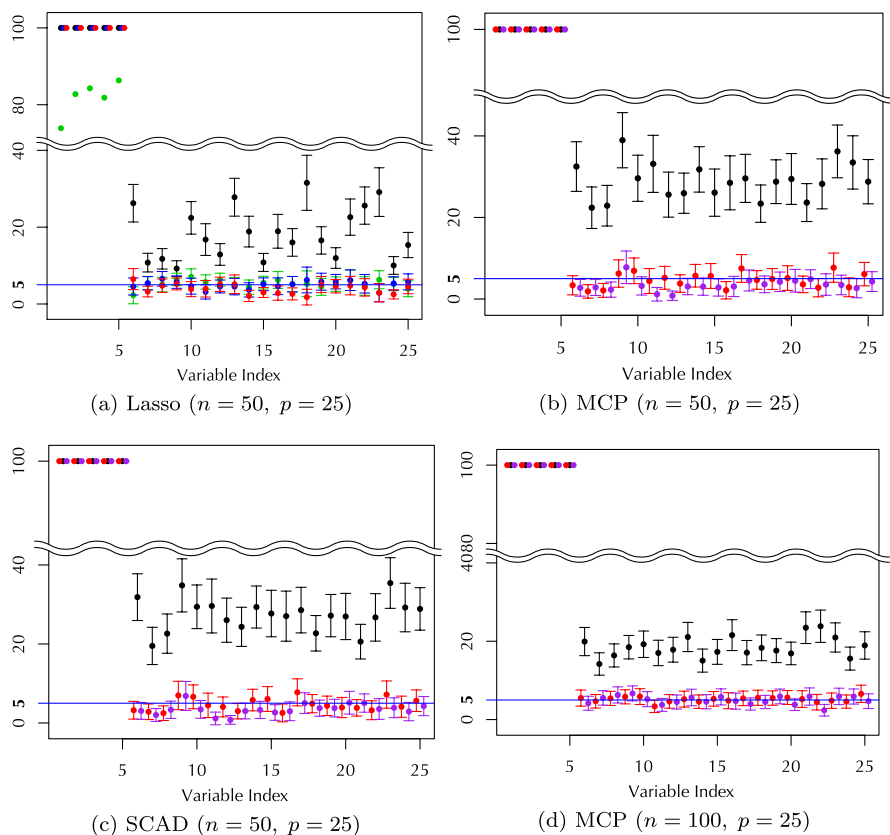
Fig. 3 Selective rejection probabilities (in percent) with two standard deviations for each feature. Black: the classical *t*-test, Green: the exact selective test conditioned on $\hat{M} = M$ and $\hat{s}_M = s_M$ (Lee et al. 2016; only for Lasso), Blue: the exact selective test conditioned on $\{j \in \hat{M}\}$ (Liu et al. 2018; only for Lasso), Red: our approximately unbiased test with the multiscale bootstrap conditioned on $j \in \hat{M}$ and $\hat{s}_j = s_j$ (i.e., p_{SI}), Purple (for MCP and SCAD): our approximately unbiased test with the classical bootstrap conditioned on $j \in \hat{M}$ and $\hat{s}_j = s_j$ (i.e., $p_{\text{SI-BP}}$)

MCP and SCAD penalties with the tuning parameter $(\rho, \gamma) = (10, 3.7)$ as the feature selection methods:

$$\text{MCP}(\beta_j | \rho, \gamma) = \begin{cases} \rho|\beta_j| - \beta_j^2/(2\gamma) & (|\beta_j| \leq \gamma\rho) \\ \gamma\rho^2/2 & (|\beta_j| > \gamma\rho) \end{cases},$$

$$\text{SCAD}(\beta_j | \rho, \gamma) = \begin{cases} \rho|\beta_j| & (|\beta_j| \leq \rho) \\ \dfrac{2\gamma\rho|\beta_j| - \beta_j^2 - \rho^2}{2(\gamma-1)} & (\rho < |\beta_j| < \gamma\rho) \cdot \\ \rho^2(\gamma+1)/2 & (|\beta_j| \geq \gamma\rho) \end{cases}$$

For non-convex penalties, we have local minimum and multiple global minimum issues. In general, we assume that the selection event can be represented as the fixed set in the data space. Thus, both issues have unexpected effects in selective inference. To overcome these issues, we used the fixed initial values for MCP and SCAD[1]. In this experiment, we used the R package `plus` (Zhang and Melnik 2012) for MCP and SCAD. The algorithm of this package generates a piecewise linear path of coefficients, starting with zero coefficients for infinity penalty.

Our multiscale bootstrap method works well not only for Lasso but also for more complicated feature selection methods such as MCP and SCAD. The panels (b) and (c) of Figure 3 show the selective rejection probabilities in the cases of MCP and SCAD, respectively. Here, we note that the existing selective inference methods (Green, Blue) cannot be applied to MCP and SCAD. Each feature with zero coefficient was selected approximately 250 times by MCP and SCAD in this experiment. In this setting, whereas no exact unbiased selective inference is proposed, the selective rejection probabilities of our test for the features 6 to 25 are around 5%. In general, we can get a more accurate result as the sample size increases. In fact, the panel (d) of Figure 3 is the result of the same experiment about MCP with larger sample size $n = 100$. Compared with the case of $n = 50$ of the panel (b), we can see the more accurate result (i.e., less variations) in the case of $n = 100$.

Moreover, we also computed the simpler selective p-value $p_{\text{SI-BP}}$ based on the classical bootstrap in both settings with MCP and SCAD. For MCP with $n = 50$, the selective rejection probabilities of p_{SI} and $p_{\text{SI-BP}}$ under the null hypotheses are 4.62% and 3.45%, respectively. For SCAD with $n = 50$, the selective rejection probabilities of p_{SI} and $p_{\text{SI-BP}}$ under the null hypotheses are 4.62% and 3.52%, respectively. These results show that the multiscale bootstrap method is more accurate than the classical bootstrap method in accordance with theory. In the setting with $n = 100$, the selective rejection probabilities of p_{SI} and $p_{\text{SI-BP}}$ under the null hypotheses are 5.16% and 5.02%, respectively. As the sample size increases, both p_{SI} and $p_{\text{SI-BP}}$ provide almost unbiased results. It could be because the boundary surface of the selection event is almost flat, at least locally around the data point, in a larger sample case.

In addition, set $\beta = (\theta, \theta, \theta, \theta, \theta, 0, \dots, 0)^T \in \mathbb{R}^p$ and $(n, p) = (50, 10)$. We compare the true positive rates (TPRs; i.e., statistical powers) and the false-positive rates (FPRs; i.e., type-I errors) of these tests with changes of θ in the case of Lasso. Here, TPR is defined by the proportion of selected nonzero features that are correctly identified, and FPR is defined by the proportion of selected zero-features that are incorrectly detected. Figure 4 shows that both the proposed method via multiscale bootstrap and the exact selective test conditioned on $j \in \hat{M}$ (Liu et al. 2018) not only have desirable high TPRs but also control FPRs at the significance level $\alpha = 5\%$. For the non-selective t-test, the FPR is not controlled, whereas the highest TPR is attained. The t-test (the black line) does not control the false positive rate. Thus it is not valid. Focusing on the TPR of the *over-conditioning* selective test (Lee et al. 2016), we can see that the unnecessarily restrictive selection event $\{\hat{M} = M, \hat{s}_M = s_M\}$ leads to the lower statistical power.

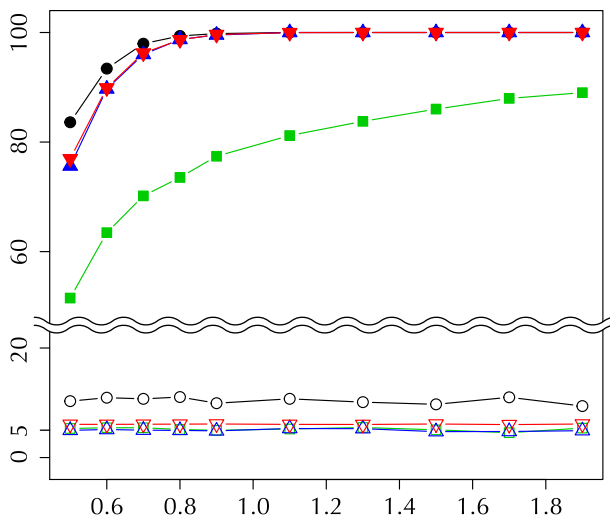[1] Another approach is extending the data space to include the space of initial values.

Fig. 4 True positive rates (upper) and false positive rates (lower) of four tests. Colors correspond to the tests as shown in the panel (a) of Figure 3. The horizontal axis represents the strength θ of the weights for the nonzero features

Next, we deal with the prostate cancer data (Stamey et al. 1989), which is available in the R package `ElemStatLearn` (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 feature has a mean zero and unit variance.

The main purpose is to provide the selective confidence intervals (CIs) for the coefficients of the 6 selected features by Lasso with the penalty $\rho = 5$. Here, we also set $\alpha = 5\%$. We computed four types of confidence intervals with confidence level $1 - \alpha$ as shown in Figure 5: the non-selective CI $[L_j^{(a)}, U_j^{(a)}]$ using t-distribution, the selective CI $[L_j^{(b)}, U_j^{(b)}]$ conditioned on $\{\hat{M} = M, \hat{s}_M = s_M\}$ (Lee et al. 2016), the selective CI $[L_j^{(c)}, U_j^{(c)}]$ conditioned on $\{j \in \hat{M}, \hat{s}_j = s_j\}$ (Liu et al. 2018), and the approximate selective CI $[L_j^{(d)}, U_j^{(d)}]$ based on our approximately unbiased p-values p_{SI} based on the multiscale bootstrap. We note that the first three CIs satisfy the following equations, respectively, for $j \in \{1, \dots, p\}$, $s_j \in \{+, -\}$, $M \subseteq \{1, \dots, p\}$, $s_M \in \{+, -\}^{|M|}$:
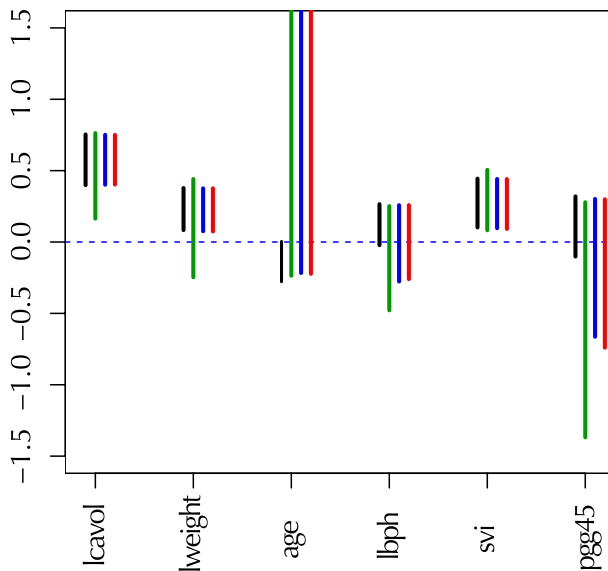
Fig. 5 Various confidence intervals for the coefficients of the selected features. (Black) the non-selective CIs $[L_j^{(a)}, U_j^{(a)}]$. (Green) the selective CIs $[L_j^{(b)}, U_j^{(b)}]$. (Blue) the selective CIs $[L_j^{(c)}, U_j^{(c)}]$. (Red) our approximate selective CIs $[L_j^{(d)}, U_j^{(d)}]$ with the multiscale bootstrap

$$P(\beta_j \in [L_j^{(a)}, U_j^{(a)}]) = 1 - \alpha, \; P(\beta_j \in [L_j^{(b)}, U_j^{(b)}] \mid \hat{M} = M, \hat{s}_M = s_M) = 1 - \alpha,$$

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

From the plot, we can see that our selective CIs $[L_j^{(d)}, U_j^{(d)}]$ approximates the exact selective CIs $[L_j^{(c)}, U_j^{(c)}]$ very well. Moreover, the over-conditioning of the selection event $\{\hat{M} = M, \hat{s}_M = s_M\}$ made CIs $[L_j^{(b)}, U_j^{(b)}]$ wider than $[L_j^{(c)}, U_j^{(c)}]$, and this indicates that the less restrictive selection event $\{j \in \hat{M}, \hat{s}_j = s_j\}$ is preferable.

6 Discussion

A new multiscale bootstrap method (Algorithm 2) is proposed to compute approximately unbiased selective *p*-values and confidence intervals for regression coefficients after feature selection. The new method is useful in particular for complicated feature selection algorithms such as MCP and SCAD, while existing methods are only available for simpler feature selection algorithms such as Lasso. The new method also computes shorter confidence intervals than most existing methods by minimally conditioning on each selected feature instead of over-conditioning on all selected features.
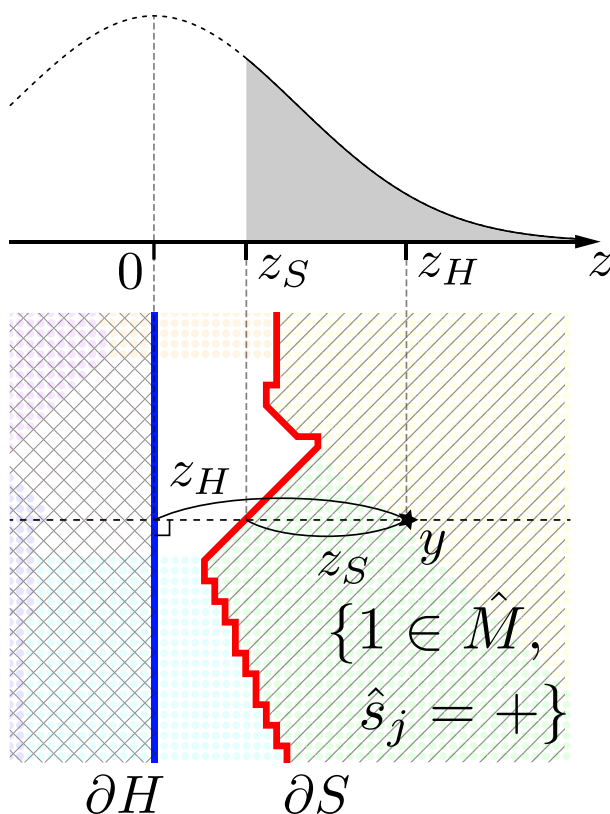
Fig. 6 The exact selective inference for the hypothesis region with the flat boundary

The proposed method is closely related to the exact selective inference such as Lee et al. (2016) and Liu et al. (2018). Here, in addition to the Gaussian assumption, we assume that the boundary surface of the hypothesis region is flat. Let us consider the line passing through the point y and perpendicular to the boundary of H. By setting the projection point $\hat{\mu}(y) \in \partial H$ of y as the origin, we can consider the one-dimensional coordinate system z on the line. If we know the distance from y to ∂H as well as intervals representing the intersection of the line and the selective region S, we can perform the exact selective inference. For example, in Figure 6, the distance is z_H and the interval is $[z_S, \infty)$. As with the polyhedral lemma, the following p-value provides the exact selective inference:

$$p(y) = \frac{\bar{\Phi}(z_H)}{\bar{\Phi}(z_H + z_S)}.$$

The explicit forms of the intervals can be obtained for the Lasso case, but it may not be possible for more complicated cases. In the proposed method, we estimate the geometric quantities z_H and z_S indirectly via the multiscale bootstrap. Alternating to this approach, we can use the grid search on the one-dimensional coordinate system

z to obtain the intervals. In association with this approach, Duy and Takeuchi (2021) proposes a parametric programming-based method that can perform the exact selective inference for the Lasso very efficiently without conditioning on signs.

Our method is applicable to the case of $p > n$, provided that $(X^T X)^+ X^T \mu$ is the target of inference, where A^+ denotes the pseudo-inverse matrix of A. However, in this case, it is difficult to estimate the error variance tau or the residuals reasonably well. Thus, the selective inference in the case of $p > n$ is important future work. In theory, we may consider the application of the proposed framework for the submodel setting. However, the selection event $\{\hat{M} = M\}$ is often too small, i.e., the bootstrap probability becomes very small for the selection event, and thus the proposed framework does not work well. Combining with the randomized response method by Tian and Taylor (2018), we may propose an appropriate selective inference based on the multiscale bootstrap method for the submodel setting.

In this paper, the multiscale bootstrap is used only for the selective region, because the hypothesis region with a flat boundary surface (i.e., $\beta_j = 0$) is easily expressed in the model space. However, Algorithm 1 is valid even for a general hypothesis region with a curved surface. Therefore, we may extend our method for, say, nonlinear regression or multiple comparisons of regression coefficients in future work.

Appendix

A Asymptotic theory for nearly flat surfaces

In Section 3, 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 λ, becomes small, that is, any boundary surfaces approach flat surfaces at least locally in a neighborhood. The artificial parameter λ is introduced, and consider the situation of $\lambda \to 0$ instead of $n \to \infty$. More precisely, suppose that the L^1-norms of function h and its Fourier transform $\tilde{h}$, i.e., $\|h\|_1$ and $\|\tilde{h}\|_1$, are bounded and that the L^∞-norm $\|h\|_\infty$ of h has the same order

as λ. Here, the function satisfying these properties is called nearly flat. Then, we consider the asymptotic theory as $\lambda \to 0$. Note that λ in this theory is corresponding to $1/\sqrt{n}$ in the classical large sample theory. Here, we assume that the hypothesis and selective regions are defined as follows, respectively:

$$H = \{(u, v) \mid v \leq -h(u), u \in \mathbb{R}^m, v \in \mathbb{R}\}, \text{ and } S = \{(u, v) \mid v_s - s(u) < v, u \in \mathbb{R}^m, v \in \mathbb{R}\},$$

where h and s are nearly flat functions, and $v_s \in \mathbb{R}$.

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

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

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

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

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

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

where $E_{\sigma^2}[\cdot|u]$ is the expectation related to U^* and $\mathcal{F}^{-1}$ 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:

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

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

$$\mathcal{E}_{\sigma^2}^{-1} h(u) = \mathcal{E}_{-\sigma^2} h(u) = \mathcal{F}^{-1}\left[e^{\sigma^2\|\omega\|^2/2}\tilde{h}(\omega)\right](u).$$

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

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

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

For the smooth h, it follows that $\mathcal{E}_{\sigma^2} h(u) = \sum_{j=0}^{\infty} \sigma^{2j}\theta_j(u)$. That is, letting $\theta_{H,0} = v + \theta_0(u)$ and $\theta_{H,j} = \theta_j(u)$ ($j \geq 1$), $\varphi_H(\sigma^2|\theta_H)$ can be modeled as $\theta_{H,0} + \theta_{H,1}\sigma^2 + \theta_{H,2}(\sigma^2)^2 + \cdots$. Thus, using a polynomial regression with predictor σ^2, we can compute the p-value $p_{\mathrm{AU}}(H|y) = \bar{\Phi}(\theta_{H,0} - \theta_{H,1} + \theta_{H,2} - \cdots)$ by formally letting $\sigma^2 = -1$. In contrast, for a cone-shaped region H, it is shown that $\mathcal{E}_{\sigma^2} h(u) = \sum_{j=0}^{\infty} \sigma^{1-j}\theta_j(u)$. Since we have $\beta_j(u) = O(\|u\|^j)$ as $\|u\| \to 0$, focusing on first two terms, we obtain

$$\mathcal{E}_{\sigma^2}h(u) \approx \theta_0(u)\sqrt{\sigma^2} + \theta_1(u).$$

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 $\theta_{H,0} = v + \theta_1(u)$ and $\theta_{H,1} = \theta_0(u)$, and then the normalized bootstrap z-value $\psi_{\sigma^2}(H|y)$ can be approximated by the model $\theta_{H,0} + \theta_{H,1}\sigma$; note the predictor is $\sigma = \sqrt{\sigma^2}$ instead of σ^2. Here, we also denote by $\varphi_H(\sigma^2|\theta_H)$ the model which approximates the normalized bootstrap z-value $\psi_{\sigma^2}(H|y)$. For fixed $\sigma_0^2 > 0$, let $\varphi_{H,k}(\sigma^2|\theta_H, \sigma_0^2)$ be the truncated Taylor expansion of $\varphi_H(\sigma^2|\theta_H)$ with the first k terms at σ_0^2:

$$\varphi_{H,k}(\sigma^2|\theta_H, \sigma_0^2) = \sum_{j=0}^{k-1} \frac{(\sigma^2 - \sigma_0^2)^j}{j!} \frac{\partial^j \varphi_H(\sigma^2|\theta_H)}{\partial(\sigma^2)^j} \bigg|_{\sigma^2 = \sigma_0^2}. \tag{7}$$

We can always use the above formula for extrapolating $\psi_{\sigma^2}(H|y)$ to $\sigma^2 \le 0$. Therefore, Algorithm 1 is updated to Algorithm 3. In practice, Algorithm 3 with $k = 2$ can be simply implemented as Algorithm 1 with the linear model $\theta_{H,0} + \theta_{H,1}\sigma^2$ and a narrow range of $\sigma^2 = n/n'$ values around σ_0^2.

Algorithm 3 Computing approximately unbiased p-values for general regions H and S (applicable to non-smooth boundary surfaces)

1: Same as Algorithm 1
2: Same as Algorithm 1
3: Same as Algorithm 1
4: Approximately unbiased p-values of non-selective inference (p_{AU}) and of selective inference (p_{SI}) are computed as follows. Specify $k \in \mathbb{N}$, $\sigma_0^2, \sigma_{-1}^2 > 0$ (e.g., $k = 3$ and $\sigma_{-1}^2 = \sigma_0^2 = 1$). Compute p-values by $p_{\text{AU},k}(H|y) = \bar{\Phi}(z_{H,k})$ and $p_{\text{SI},k}(H|S,y) = \bar{\Phi}(z_{H,k})/\bar{\Phi}(z_{H,k} + z_{S,k})$, where $z_{H,k} = \varphi_{H,k}(-1|\hat{\theta}_H(y), \sigma_{-1}^2)$ and $z_{S,k} = \varphi_{S,k}(0|\hat{\theta}_S(y), \sigma_0^2)$ computed by formula (7).

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

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

Under some regularity conditions, Shimodaira ([2008]) proves that

$$\lim_{k \to \infty} P(p_{\text{AU},k}(H|Y) < \alpha \mid \mu) = \alpha + O(\lambda^2)$$

at each $\mu \in \partial H$. Moreover, for general selective inference with possibly non-smooth boundary surfaces, Terada and Shimodaira ([2017]) proposes the following selective p-value:

$$p_{\mathrm{SI},k}(H|S,y) = \frac{\bar{\Phi}(\varphi_{H,k}(-1|\theta_H, \sigma_{-1}^2))}{\bar{\Phi}(\varphi_{H,k}(-1|\theta_H, \sigma_{-1}^2) + \varphi_{S,k}(0|\theta_S, \sigma_0^2))},$$

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

$$\lim_{k \to \infty} \frac{P(p_{\mathrm{SI},k}(H|S,Y) < \alpha \mid \mu)}{P(Y \in S \mid \mu)} = \alpha + O(\lambda^2)$$

at each $\mu \in \partial H$.

B Proof of Theorem 2

Proof First, we show that, for given $\alpha \in (0, 1)$, there exists a nearly flat function r such that

$$\forall \mu \in \partial H; \; \frac{P(Y \in R \mid \mu)}{P(Y \in S \mid \mu)} = \alpha + O(\lambda^2), \tag{8}$$

where $R = \{(u, v) \mid v > v_r - r(u)\}$ and $v_r = \bar{\Phi}^{-1}(\alpha \bar{\Phi}(v_s))$. By Lemma 5.1 in Terada and Shimodaira (2017) or equivalently eq. (5.3) in Shimodaira (2008), we have

$$P(Y \in S \mid \mu) = 1 - \Phi(v_s - \mathcal{E}_1 s(\theta)) + O(\lambda^2)$$
$$= 1 - \{\Phi(v_s) - \phi(v_s)\mathcal{E}_1 s(\theta)\} + O(\lambda^2)$$

for $\mu = (\theta, 0) \in \partial H$. Let us temporarily assume that r is nearly flat. Then, we also have

$$P(Y \in R \mid \mu) = 1 - \{\Phi(v_r) - \phi(v_r)\mathcal{E}_1 r(\theta)\} + O(\lambda^2).$$

From Eq. (8), it follows that

$$\bar{\Phi}(v_r) + \phi(v_r)\mathcal{E}_1 r(\theta) + O(\lambda^2) = \alpha\big[\bar{\Phi}(v_s) + \phi(v_s)\mathcal{E}_1 s(\theta) + O(\lambda^2)\big].$$

Since $\bar{\Phi}(v_r) = \alpha\bar{\Phi}(v_s)$, we have $\mathcal{E}_1 r(\theta) = \alpha C \mathcal{E}_1 s(\theta) + O(\lambda^2)$, where $C = \phi(v_s)/\phi(v_r)$. Thus, applying the inverse operator $\mathcal{E}_{-1}$ to both sides, we obtain $r(u) = \alpha C s(u) + O(\lambda^2)$. Since s is nearly flat, r should be nearly flat. Similarly, we can show that the above r actually satisfies Eq. (8). Combining $\phi(v_r)r(u) = \alpha\phi(v_s)s(u) + O(\lambda^2)$ with $\bar{\Phi}(v_r) = \alpha\bar{\Phi}(v_s)$, we obtain $\bar{\Phi}(v_r) + \phi(v_r)r(u) = \alpha\{\bar{\Phi}(v_s) + \phi(v_s)s(u)\} + O(\lambda^2)$. By Taylor's theorem, we deduce that

$$\bar{\Phi}(v_r - r(u)) = \alpha\bar{\Phi}(v_s - s(u)) + O(\lambda^2). \tag{9}$$

Now, we consider the rejection region R based on p_{SI}, that is, $R = \{(u, v) \mid p_{\mathrm{SI}}(H|S, y) < \alpha\}$. By Lemma 5.1 in Terada and Shimodaira
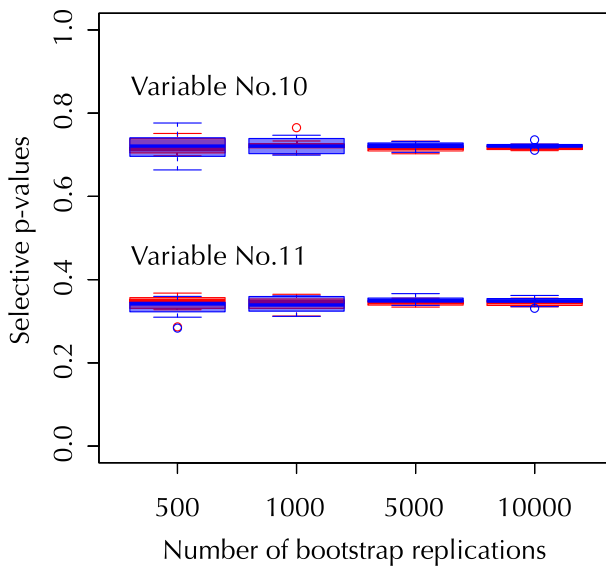
Fig. 7 Stability of multiscale bootstrap with the choice of the scales and the number of replications

(2017) or equivalently eq. (5.3) in Shimodaira (2008), for $y = (u, v)$, we have $\psi_{\sigma^2}(S|y) = -v + v_s - \mathcal{E}_{\sigma^2} s(u) + O(\lambda^2)$. Since $\mathcal{E}_0 s(u) = s(u)$, we have $z_S = \psi_0(S|y) = -v + v_s - s(u) + O(\lambda^2)$. Let us recall that $\psi_{\sigma^2}(H|y) = v(H|y) = v$ for $H = \{(u', v') \mid u' \in \mathbb{R}^m, v' \leq 0\}$, and so $z_H = v$. Thus, we obtain

$$p_{\text{SI}}(H|S, y) = \frac{\bar{\Phi}(z_H)}{\bar{\Phi}(z_H + z_S)} = \frac{\bar{\Phi}(v)}{\bar{\Phi}(v_s - s(u))} + O(\lambda^2).$$

By substituting $v = v_r - r(u)$ above, it follows from Eq. (9) that $p_{\text{SI}}(H|S, y) = \alpha + O(\lambda^2)$ for $y = (u, v_r - r(u)) \in \partial R$. This finishes the proof. $\square$

C Choice of the tuning parameters in multiscale bootstrap

The multiscale bootstrap is not very sensitive to the choice of the scales. For confirming this fact, we additionally performed experiments with two settings of scales, as shown in Figure 7. We choose ℓ (= 5 and 10) scales from the interval between 0.1 and 2 with equal spaces in the log-scale. We also changed the number of replications B (= 500, 1000, 5000, 10000) in multiscale bootstrap. The other parameters are the same as the experiment in Section 5. For a simulated data, we computed selective p-values under the various settings ($B = 500, 1000, 5000, 10000; \ell = 5, 10$) of multiscale bootstrap. Under each setting, we computed the selective p-value 10 times for two selected features No. 10 and No. 11 whose true coefficients are zero.

In Figure 7, the red box plots correspond to the setting $\ell = 5$ and the blue ones to the setting $\ell = 10$. The averages of 10 p-values are almost the same among all settings, and thus the p-values are not sensitive to ℓ. On the other hand, the variance of the p-values decreases as B increases. Several thousand replications are enough in practice.

Declaration

Conflict of interest The authors declare no conflicts of interest associated with this manuscript.

References

Berk, R., Brown, L., Buja, A., Zhang, K., 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.

Duy, V. N. L., Takeuchi, I. (2021). Parametric programming approach for more powerful and general lasso selective inference. *Proceedings of The 24th international conference on artificial intelligence and statistics* (*AISTATS2021*), pp 901–909.

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

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* 93:13, 429–13, 434.

Fan, J., 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., 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*. https://CRAN.R-project.org/package=ElemStatLearn, r package.

Lee, J. D., Sun, D. L., Sun, Y., 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. *Annals of Mathematical Statistics, 23,* 541–552.

Lehmann, E. L. (1999). *Elements of large-sample theory*. New York: Springer.

Lim, J., Yamada, M., Jitkrittum, W., Terada, Y., Matsui, S., Shimodaira, H. (2020). More powerful selective kernel tests for feature selection. *Proceedings of the 23rd International conference on artificial intelligence and statistics* (*AISTATS 2020*), pp 820–830.

Liu, K., Markovic, J., Tibshirani, R. (2018). More powerful post-selection inference, with application to the lasso. arXiv:1801.09037.

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., Terada, Y. (2019). Selective inference for testing trees and edges in phylogenetics. *Frontiers in Ecology and Evolution, 7,* 174.

Slim, L., Chatelain, C., Azencott, C. A., Vert, J. P. (2019). kernelPSI: A post-selection inference framework for nonlinear variable selection. *Proceedings of the 36th International conference on machine learning* (*ICML 2019*), pp 5857–5865.

Stamey, T., Kabalin, J., Johnstone, I., Freiha, F., Redwine, E., 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.

Suzumura, S., Nakagawa, K., Umezu, Y., Tsuda, K., Takeuchi, I. (2017). Selective inference for sparse high-order interaction models. *Proceedings of the 17th International conference on machine learning* (*ICML 2017*), pp 3338–3347.

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

Tian, X., Taylor, J. (2018). Selective inference with a randomized response. *Annals of Statistics, 46,* 679–710.

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., 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., Wasserman, L. (2018). Uniform asymptotic inference and the bootstrap after model selection. *Annals of Statistics, 46,* 1255–1287.

Yang, F., Foygel Barber, R., Jain, P., Lafferty, J. (2016). Selective inference for group-sparse linear models. *Advances in Neural Information Processing Systems, 29,* 2469–2477.

Zhang, C. H. (2010). Nearly unbiased variable selection under minimax concave penalty. *Annals of Statistics, 38,* 894–942.

Zhang, C. H., Melnik, O. (2012). *plus*: *Penalized linear unbiased selection*. https://cran.r-project.org/package=plus, r package.