Statistica Sinica Preprint No: SS-2018-0131	
Title	A Bootstrap Lasso + Partial Ridge Method to Construct Confidence Intervals for Parameters in High-dimensional Sparse Linear Models
Manuscript ID	SS-2018-0131
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202018.0131
Complete List of Authors	Hanzhong Liu Xin Xu and Jingyi Jessica Li
Corresponding Author	Jingyi Jessica Li
E-mail	jli@stat.ucla.edu

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0131

Statistica Sinica

A Bootstrap Lasso + Partial Ridge Method to Construct Confidence Intervals for Parameters in High-dimensional Sparse Linear Models

Hanzhong Liu[1], Xin Xu[2,3], and Jingyi Jessica Li[3*]

[1] *Tsinghua Universty,* [2] *Yale University,* [3] *UCLA*

Abstract:

Constructing confidence intervals for the coefficients of high-dimensional sparse linear models remains a challenge, mainly because of the complicated limiting distributions of the widely used estimators, such as the lasso. Several methods have been developed for constructing such intervals. Bootstrap lasso+ols is notable for its technical simplicity, good interpretability, and performance that is comparable with that of other more complicated methods. However, bootstrap lasso+ols depends on the beta-min assumption, a theoretic criterion that is often violated in practice. Thus, we introduce a new method, called bootstrap lasso+partial ridge, to relax this assumption. Lasso+partial ridge is a two-stage estimator. First, the lasso is used to select features. Then, the partial ridge is used to refit the coefficients. Simulation results show that bootstrap lasso+partial ridge outperforms bootstrap lasso+ols when there exist small, but nonzero coefficients, a common situation that violates the beta-min assumption. For such coefficients, the confidence intervals constructed using bootstrap lasso+partial ridge have, on average, 50% larger coverage probabilities than those of bootstrap lasso+ols. Bootstrap lasso+partial ridge also has, on average, 35%

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0131

shorter confidence interval lengths than those of the de-sparsified lasso methods, regardless of whether the linear models are misspecified. Additionally, we provide theoretical guarantees for bootstrap lasso+partial ridge under appropriate conditions, and implement it in the R package "HDCI."

Key words and phrases: Bootstrap, Confidence interval, High-dimensional inference, Lasso+partial ridge, Model selection consistency.

1. Introduction

The proliferation of high-dimensional data in fields such as information technology, astronomy, neuroscience, and bioinformatics has necessitated new analysis methods. Data are high dimensional if the number of predictors p is comparable to, or much larger than, the sample size n. Over the past two decades, statistical and machine learning theory, methodologies, and algorithms have been developed to tackle high-dimensional data problems under certain sparsity constraints, such as the number of nonzero linear model coefficients s being much smaller than the sample size n. Regularization is required to perform a sparse estimation under this regime. For example, the lasso (Tibshirani, 1996) uses l_1 regularization to perform model selection and parameter estimation simultaneously in a high-dimensional sparse linear regression. Previous works have focused on recovering a sparse parameter vector (denoted by $\beta^0 \in \mathbb{R}^p$), based on criteria such as (i) model selection consistency, (ii) the l_q estimation

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0131

error $||\hat{\beta} - \beta^0||_q$, where $\hat{\beta}$ is an estimate of β^0 and q is typically equal to one or two, and (iii) the prediction error $||X\hat{\beta} - X\beta^0||_2$, with X as the design matrix. The book (Bühlmann & van de Geer, 2011) and the review paper (Fan & Lv, 2010) give a thorough summary of the recent advances in high-dimensional statistics.

An important research question in high-dimensional statistics is how to perform statistical inference, that is, constructing confidence intervals and hypothesis tests for individual coefficients in linear models. Inference is crucial when the purpose of statistical modeling is to understand scientific principles beyond those of prediction. However, inference is difficult for high-dimensional model parameters, because the limiting distributions of the widely used estimators, such as the lasso, are complicated and difficult to compute in high dimensions. To address this challenge, we develop a novel and practical inference procedure called bootstrap lasso+partial ridge (LPR), which is based on three canonical methods: the bootstrap, lasso, and ridge. Before presenting our method, we briefly review several existing high-dimensional inference methods.

There is a growing body of statistical literature on high-dimensional inference problems. Existing methods are divided into several categories, including the sample-splitting-based methods, bootstrap-based methods, de-sparsified lasso methods, post-selection inference methods, and knockoff filter. In partic-

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0131

ular, Wasserman and Roeder proposed a sample-splitting method (Wasserman & Roeder, 2009) that splits n data points into two halves. The first half is used for model selection (say, by the lasso), and the second half is used to construct confidence intervals or p-values for the parameters in the selected model. For a fixed dimension p, Minnier et al. developed a perturbation resampling-based method to approximate the distribution of penalized regression estimates under a general class of loss functions (Minnier et al., 2009). Chatterjee and Lahiri proposed a modified residual bootstrap lasso method (Chatterjee & Lahiri, 2011), which is consistent in estimating the limiting distribution of a modified lasso estimator. For scenarios in which p goes to infinity at a polynomial rate of n, Chatterjee and Lahiri showed that a residual bootstrap adaptive lasso estimator can consistently estimate the limiting distribution of the adaptive lasso estimator under several intricate conditions (Chatterjee & Lahiri, 2013). Two of these conditions are similar to the irrepresentible condition and the beta-min condition (the beta-min condition means that the minimum absolute value of the nonzero regression coefficients is much larger than $n^{-1/2}$), which together guarantee the model selection consistency of the lasso. Liu and Yu proposed another residual bootstrap method based on a two-stage estimator (lasso+ols), showing its consistency under the irrepresentible condition, beta-min condition, and other regularity conditions (Liu & Yu, 2013). Here, lasso+ols denotes using the lasso method

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0131

to select a model, and then using the ordinary least squares (OLS) method to refit the coefficients in the selected model. However, a common issue with these methods is that they all require the rather restrictive beta-min condition, which should be relaxed in high-dimensional inference, if possible.

The de-sparsified lasso, proposed by Zhang & Zhang (2014), and later investigated by van de Geer et al. (2014), Javanmard & Montanari (2014), is another type of method. This method aims to remove the biases of the lasso estimates and produce an asymptotically normal estimate for each parameter. Specifically, we refer to the popular de-sparsified lasso methods developed by Zhang & Zhang (2014) and Javanmard & Montanari (2014) as LDPE and JM, respectively. These methods do not rely on the beta-min condition, but do require that we estimate the precision matrix of predictors using the graphical lasso method (Zhang & Zhang, 2014; van de Geer et al., 2014), or some other convex optimization procedure (Javanmard & Montanari, 2014). There are two main issues with these methods. First, they rely heavily on the sparse linear model assumption and, thus, may exhibit poor performance for misspecified models. Second, the computational costs of these methods are quite high. For example, constructing confidence intervals for all entries of β^0 requires solving $(p+1)$ separate quadratic optimization problems. Despite these drawbacks, the methods can serve as a theoretically proven benchmark for high-dimensional inference.

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0131

Other new tools include the post-selection inference methods (Berk, 2013; Lee et al., 2015), knockoff filter (Barber & Candès, 2015), covariance test (Lockhart et al., 2014), group-bound confidence intervals (Meinshausen, 2015), bootstrapping ridge regression (Lopes, 2014), and ridge projection and bias correction (Bühlmann, 2013), among others; see Dezeure et al. (2014) for a comprehensive review of high-dimensional inference methods.

According to the results of simulation studies in an independent assessment (Dezeure et al., 2014), the bootstrap lasso+ols method produces confidence intervals with coverage probabilities and lengths that are comparable with those of other existing methods when the beta-min condition holds. Bootstrap lasso+ols is built on three canonical statistical techniques (i.e., the bootstrap, lasso, and OLS), all of which are well known to a broad audience and, hence, easily accessible to data scientists. However, as mentioned, the main drawback of bootstrap lasso+ols is the rather restrictive beta-min condition, which results in poor coverage probabilities for the confidence intervals of small, but nonzero coefficients (e.g., 95% confidence intervals with coverage probabilities lower than 50%). This is because these small coefficients are seldom selected by the lasso and, hence, are not refitted by the OLS, resulting in coefficient estimates of zero in most bootstrap runs. Therefore, the confidence intervals produced by bootstrap lasso+ols have lengths and coverage probabilities that are close to zero. Intu-

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0131

itively, it seems advantageous to adopt a different second-step procedure after the lasso to replace the OLS. Ideally, this procedure should not place a penalty on the coefficients selected by the lasso, in order to reduce the bias. However, it should place a small, but nonzero l_2 penalty on the unselected coefficients in order to recover them. We call this the LPR estimator. An independent work by Gao et al. (2017) proposed a post-selection ridge estimator similar to our LPR estimator. However, their aim was to improve the prediction performance, and they achieved it by adding a thresholding step. Chernozhukov, Hansen & Liao (2017) proposed a penalization-based estimation strategy called Lava to deal with "sparse + dense" coefficients. However, they also focused on improving the prediction performance rather than the quality of the inference.

In this paper, we propose a new inference procedure called bootstrap LPR as an improvement over the bootstrap lasso+ols method. The problem setting is to construct confidence intervals for individual regression coefficients β_j^0, for $j = 1, \ldots, p$, in a high-dimensional linear regression model, where β^0 is weakly sparse (Negahban et al., 2009). That is, its elements can be divided into two groups: "large" coefficients, with absolute values $\gg n^{-\frac{1}{2}}$, and "small" coefficients, with absolute values $\ll n^{-\frac{1}{2}}$. We define this type of sparsity as the *cliff-weak-sparsity*, which means that if we order the absolute coefficients from the largest to the smallest, there exits a cliff-like drop that divides the coeffi-

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0131

cients into two groups. Obviously, cliff-weak-sparsity is a weaker assumption than hard (or exact) sparsity (β^0 has at most s ($s \ll n$) nonzero elements) and the beta-min condition.

Inference for small coefficients has been investigated by Shi & Qu (2017), who proposed a two-step inference procedure to identify weak signals (small coefficients). Their method is designed for an orthogonal design matrix, and is based on a combination of the asymptotic normality of a bias-corrected adaptive lasso estimator (for large coefficients) and the least squares estimator (for small coefficients) instead of the bootstrap. However, their method performs well only when $p \ll n$, whereas our method, based on the bootstrap, can be used when $p \gg n$.

Dezeure et al. (2017) and Zhang & Cheng (2017) combined the bootstrap and de-sparsified lasso methods to deal with non-Gaussian and heteroscedastic errors. We refer to this method as the bootstrap version of LDPE (BLDPE), and we include it in the method comparison in our simulation and real-data studies.

Our contributions to the literature are summarized as follows.

First, our proposed bootstrap LPR method relaxes the beta-min condition required by the bootstrap lasso+ols method. We provide conditions under which the bootstrap LPR method can consistently estimate the distribution of the LPR estimator and, therefore, is valid for constructing a confidence interval for each

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0131

coefficient.

Second, we conduct comprehensive simulation studies to evaluate the finite-sample performance of the bootstrap LPR method for both sparse linear models and misspecified models. Our main findings are as follows. (1) Compared with bootstrap lasso+ols, bootstrap LPR improves the coverage probabilities of the 95% confidence intervals by about 50%, on average, for small nonzero regression coefficients. However, this improvement incurs a 15% heavier computational burden for $n = 200,\ p = 500$. (2) Compared with the two de-sparsified lasso methods, LDPE and JM, bootstrap LPR produces good coverage probabilities for large and small regression coefficients. In some cases, it even outperforms these methods by producing confidence intervals with lengths that are more than 50% shorter, on average. (3) Bootstrap LPR is more than 30% faster than the two de-sparsified lasso methods, and is robust to model misspecification. We also demonstrate the performance of bootstrap LPR on two real data sets: functional magnetic resonance imaging (fMRI) data, and neuroblastoma gene expression data.

Third, we extend the model selection consistency of the lasso from the hard sparsity case (Zhao & Yu, 2006; Wainwright, 2009) to a more general *cliff-weak-sparsity* case. Under the irrepresentable condition and other reasonable conditions, we show that the lasso can correctly select all "large" elements of β^0,

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0131

while shrinking all "small" elements to zero.

Fourth, we develop an R package "HDCI" to implement the bootstrap lasso, bootstrap lasso+ols, and the proposed bootstrap LPR methods. This package makes these methods easily accessible to practitioners.

Fifth, our method is not limited to using the lasso in the selection stage, but can be extended to any other model selection criterion, such as the stability selection (Meinshausen & Bühlmann, 2010), the smoothly clipped absolute deviation (SCAD) estimator (Fan & Li, 2001), the Dantzig selector (Candès & Tao, 2007), and the post-double selection (Belloni et al., 2014), which is promising because it does not require the beta-min condition. If we replace the lasso by the post-double selection method, the resulting confidence intervals may achieve better coverages for medium-sized coefficients. This is an interesting research direction that is worth further investigation, because the methodology, computation, and theory will differ from those of our current work in many respects.

The remainder of this paper proceeds as follows. In Section 2, we define the LPR estimator and introduce the residual bootstrap LPR (rBLPR) and the paired bootstrap LPR (pBLPR) methods. In Section 3, we investigate the theoretical properties of the proposed method. In Section 4, we conduct comprehensive simulation studies to compare the finite-sample performance of rBLPR, pBLPR, bootstrap lasso+ols, and three de-sparsified lasso methods (LDPE, JM,

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0131

and BLDPE). In Sections 5 and 6, we present two real-data case studies. Section 7 concludes the paper. All relevant proofs, algorithms, and simulation details can be found in the online Supplementary Material.

2. Framework and definitions

2.1 Overview and background

In this section, we begin by introducing high-dimensional sparse linear models. We next define cliff-weak-sparsity and the LPR estimator. Finally, we propose two bootstrap procedures (residual bootstrap and paired bootstrap), based on the LPR estimator, to construct confidence intervals for individual regression coefficients.

We assume the data are generated from the following linear model:

$$Y = X\beta^0 + \epsilon, \tag{2.1}$$

where $\epsilon = (\epsilon_1, \ldots, \epsilon_n)^{\mathrm{T}}$ is a vector of independent and identically distributed (i.i.d.) random error variables, with mean 0 and variance σ^2, $Y = (y_1, \ldots, y_n)^{\mathrm{T}} \in \mathbb{R}^n$ is an n-dimensional response vector, and $X = (x_1^{\mathrm{T}}, \ldots, x_n^{\mathrm{T}})^{\mathrm{T}} = (X_1, \ldots, X_p) \in \mathbb{R}^{n\times p}$ is a deterministic or random design matrix. Without loss of generality, we assume that every predictor is centered, that is, $\sum_{i=1}^n x_{ij}/n = 0$, for $j = 1, \ldots, p$, and there is no intercept term in the linear model. Denoting

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0131

$\beta^0 \in \mathbb{R}^p$ as a vector of coefficients, we assume that β^0 satisfies cliff-weak-sparsity.

Definition 1 (Cliff-weak-sparsity)**.** β^0 satisfies cliff-weak-sparsity if its elements can be divided into two groups. The first group has s $(s \ll n)$ large elements, with absolute values much larger than $n^{-1/2}$, and the second group contains $p-s$ small elements, with absolute values much smaller than $n^{-1/2}$.

We are interested in constructing a confidence interval for each coefficient β_j^0, for $j = 1, \ldots, p$. We consider the high-dimensional setting where both p and s grow with n. Here, and in what follows, Y, X, and β^0 are all indexed by n, but we omit the index n whenever this does not cause confusion.

The lasso estimator (Tibshirani, 1996) is a useful tool for enforcing sparsity when estimating high-dimensional parameters. The estimator is defined as follows:

$$\hat{\beta}_{\text{lasso}} = \arg\min_{\beta} \left\{ ||Y - X\beta||_2^2/(2n) + \lambda_1 ||\beta||_1 \right\}, \tag{2.2}$$

where $\lambda_1 \geq 0$ is the tuning parameter controlling the amount of regularization applied to the estimate. In general, λ_1 depends on n, but we omit this dependence in the notation, for simplicity. The limiting distribution of the lasso is complicated (Knight & Fu, 2000), and the usual residual bootstrap lasso fails to construct valid confidence intervals (Chatterjee & Lahiri, 2010). Various modifications have been proposed to form a valid inference procedure, but these rely on

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0131

two restrictive assumptions: hard sparsity, and the beta-min condition. In order to relax these two often unrealistic assumptions, we propose the LPR estimator, with two associated bootstrap procedures (the rBLPR and pBLPR).

2.2 The LPR estimator

In this subsection, we first describe the rationale of the LPR estimator and then formally define it. We argue that the LPR estimator is useful for weakly sparse linear models, the coefficients of which have many small, but nonzero elements decaying at a certain rate, satisfying *cliff-weak-sparsity*.

In case of *cliff-weak-sparsity*, existing bootstrap methods, such as bootstrap lasso+ols, give very poor coverage probabilities for the small, but nonzero regression coefficients because they are seldom selected by the lasso. Hence, a large fraction of the bootstrap lasso+ols estimates are zero, producing zero-length confidence intervals $[0, 0]$ that do not cover the corresponding non-zero coefficients. To fix this problem, we need to increase the variance of our estimates for small coefficients of predictors that are missed by the lasso. This is the motivation for the LPR estimator proposed in this paper.

The LPR estimator is a two-stage estimator. It adopts the lasso to select the predictors, and it then refits the coefficients using the partial ridge. The latter is defined to minimize the empirical l_2 loss with no penalty on the selected

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0131

predictors, but with an l_2 penalty on the unselected predictors. This reduces the bias of the coefficient estimates of the selected predictors, while increasing the variance of the coefficient estimates of the unselected predictors. The l_2 penalty (as used in a ridge regression (Hoerl & Kennard, 1970)) is used because it regularizes the coefficient estimates without imposing sparsity. Formally, let $S = \{j \in \{1, \ldots, p\} : \beta_j^0 \neq 0\}$ be the support set of β^0, and let $\hat{S} = \{j \in \{1, \ldots, p\} : (\hat{\beta}_{\text{lasso}})_j \neq 0\}$ be the set of predictors selected by the lasso. Then, we define the LPR estimator as

$$\hat{\beta}_{\text{LPR}} = \arg\min_{\beta} \left\{ \frac{1}{2n} ||Y - X\beta||_2^2 + \frac{\lambda_2}{2} \sum_{j \notin \hat{S}} \beta_j^2 \right\}. \tag{2.3}$$

Here, λ_2 is a tuning parameter that, in general, depends on n, but we omit the dependence in the notation, for simplicity. Our simulations in Section 4 show that fixing λ_2 at $O(1/n)$ works quite well for a range of error variance levels. For the sake of simplicity, we set $\lambda_2 = 1/n$, with the understanding that further research should be done on the selection of λ_2.

In the next two subsections, we will discuss two commonly used bootstrap procedures for the LPR estimator, and we will explain how to use them to construct a confidence interval for each coefficient, respectively.

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0131

2.3 The rBLPR method

For a deterministic design matrix X in a linear regression model, the residual bootstrap is a standard method used to construct confidence intervals. In this subsection, we introduce the rBLPR procedure.

We first need to appropriately define residuals so that their empirical distribution can well approximate the true distribution of the error, ϵ_i. In a high-dimensional linear regression, there are different ways to obtain residuals. For example, we can calculate the residuals using estimation methods such as the lasso, lasso+ols, and LPR. Simulations suggest that the residuals obtained from the lasso+ols approximate the true distribution of ϵ_i the best and, hence, are adopted in this study. Note that, when the beta-min condition does not hold, lasso+ols would fail to select all nonzero coefficients correctly. That is, lasso+ols is not consistent for model selection, but its prediction performance could still be good (i.e., it has a smaller mean squared error than that of the lasso). Let $\hat{\beta}_{\text{lasso+ols}}$ denote the lasso+ols estimator,

$$\hat{\beta}_{\text{lasso+ols}} = \underset{\beta:\ \beta_{\hat{S}^c}=0}{\arg\min} \left\{ \frac{1}{2n} ||Y - X\beta||_2^2 \right\}, \text{ where, } \beta_{\hat{S}^c} = \{\beta_j : j \notin \hat{S}\}. \quad (2.4)$$

The residual vector is defined as $\hat{\epsilon} = (\hat{\epsilon}_1, \ldots, \hat{\epsilon}_n)^{\mathrm{T}} = Y - X\hat{\beta}_{\text{lasso+ols}}$. We consider the centered residuals $\{\hat{\epsilon}_i - \tilde{\epsilon},\ i = 1, \ldots, n\}$, where $\tilde{\epsilon} = \sum_{i=1}^{n} \hat{\epsilon}_i/n$. For the residual bootstrap, we obtain $\epsilon^* = (\epsilon_1^*, \ldots, \epsilon_n^*)^{\mathrm{T}}$ by resampling with

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0131

replacement from the centered residuals $\{\hat{\epsilon}_i - \tilde{\epsilon},\ i = 1, \ldots, n\}$, and then we construct the residual bootstrap ("rboot") version of Y:

$$Y^*_{\text{rboot}} = X\hat{\beta}_{\text{lasso+ols}} + \epsilon^*. \tag{2.5}$$

Then, based on the residual bootstrap sample (X, Y^*_{rboot}), we can compute the residual bootstrap lasso (rBlasso) estimator $\hat{\beta}^*_{\text{rBlasso}}$, as in (2.6) (replacing Y in equation (2.2) by Y^*_{rboot}), and its selected predictor set $\hat{S}^*_{\text{rBlasso}} = \{j \in \{1, \ldots, p\} : (\hat{\beta}^*_{\text{rBlasso}})_j \neq 0\}$. We can also compute the rBLPR estimator $\hat{\beta}^*_{\text{rBLPR}}$, as in (2.7), in the same way as in equation (2.3), except that we replace Y and $\hat{S}$ by Y^*_{rboot} and $\hat{S}^*_{\text{rBlasso}}$, respectively:

$$\hat{\beta}^*_{\text{rBlasso}} = \arg\min_{\beta} \left\{ \frac{1}{2n} ||Y^*_{\text{rboot}} - X\beta||_2^2 + \lambda_1 ||\beta||_1 \right\}, \tag{2.6}$$

$$\hat{\beta}^*_{\text{rBLPR}} = \arg\min_{\beta} \left\{ \frac{1}{2n} ||Y^*_{\text{rboot}} - X\beta||_2^2 + \frac{\lambda_2}{2} \sum_{j \notin \hat{S}^*_{\text{rBlasso}}} \beta_j^2 \right\}. \tag{2.7}$$

If the conditional distribution (given ϵ) of $T^*_n = \sqrt{n}(\hat{\beta}^*_{\text{rBLPR}} - \hat{\beta}_{\text{lasso+ols}})$ from the bootstrap is a good approximation of the distribution of $T_n = \sqrt{n}(\hat{\beta}_{\text{LPR}} - \beta^0)$, then we can use the residual bootstrap to construct asymptotically valid confidence intervals; see Algorithm S1 for the complete procedure.

2.4 The pBLPR method

In this subsection, we introduce the pBLPR procedure. Paired bootstraps are

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0131

widely used for the inference in linear models. In this procedure, we generate a bootstrap sample $\{(x_i^*, y_i^*), i = 1, \ldots, n\}$ from the empirical distribution of $\{(x_i, y_i), i = 1, \ldots, n\}$, and then we compute the paired bootstrap lasso (pBlasso) estimator

$$\hat{\beta}^*_{\text{pBlasso}} = \arg\min_{\beta} \left\{ \frac{1}{2n} ||Y^*_{\text{pboot}} - X^*_{\text{pboot}}\beta||_2^2 + \lambda_1 ||\beta||_1 \right\}, \tag{2.8}$$

where $Y^*_{\text{pboot}} = (y_1^*, \ldots, y_n^*)^{\text{T}}$ and $X^*_{\text{pboot}} = ((x_1^*)^{\text{T}}, \ldots, (x_n^*)^{\text{T}})^{\text{T}}$ denote the paired bootstrap samples. Let $\hat{S}^*_{\text{pBlasso}} = \{j \in \{1, \ldots, p\} : (\hat{\beta}^*_{\text{pBlasso}})_j \neq 0\}$ be the set of predictors selected by the paired bootstrap lasso. We define the pBLPR estimator as

$$\hat{\beta}^*_{\text{pBLPR}} = \arg\min_{\beta} \left\{ \frac{1}{2n} ||Y^*_{\text{pboot}} - X^*_{\text{pboot}}\beta||_2^2 + \frac{\lambda_2}{2} \sum_{j \notin \hat{S}^*_{\text{pBlasso}}} \beta_j^2 \right\}. \tag{2.9}$$

The pBLPR procedure for constructing confidence intervals is summarized in Algorithm S2.

3. Theoretical results

3.1 Overview

In this section, we investigate the theoretical properties of the rBLPR method. In particular, we first show that, under cliff-weak-sparsity and other reasonable conditions, the lasso exhibits model selection consistency, in the sense that it

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0131

correctly identifies all large components of β^0, while shrinking all small components to zero; see Theorem 1. Second, and more interestingly, we show in Theorem 2 that, under one further condition, the residual bootstrap lasso estimator achieves the same kind of model selection consistency. Based on these properties, we provide the conditions under which the limiting distribution of $\sqrt{n}u^{\mathrm{T}}T_n^* = \sqrt{n}u^{\mathrm{T}}(\hat{\beta}^*_{\text{rBLPR}} - \hat{\beta}_{\text{lasso+ols}})$, conditional on ϵ, is the same as the limiting distribution (unconditional) of $\sqrt{n}u^{\mathrm{T}}T_n = \sqrt{n}u^{\mathrm{T}}(\hat{\beta}_{\text{LPR}} - \beta^0)$, for a general class of $u \in \mathbb{R}^p$; see Theorem 3.

3.2 Model selection consistency of the lasso under *cliff-weak-sparsity*

In this subsection, we extend the model selection consistency of the lasso from the hard sparsity case to the more general *cliff-weak-sparsity* case, where β^0 has many small but nonzero elements.

Zhao & Yu (2006) and Wainwright (2009) showed that the lasso is sign-consistent (i.e., $\text{pr}(\text{sign}(\hat{\beta}_{\text{lasso}}) = \text{sign}(\beta^0)) \to 1$ as $n \to \infty$, which implies model selection consistency) under appropriate conditions, including the irrepresentable condition, beta-min condition, and hard sparsity.

Definition 2 (Zhao & Yu (2006))**.** If an estimator $\hat{\beta}$ is equal in sign to the true β^0, we write $\hat{\beta} =_s \beta^0$, which is equivalent to $\text{sign}(\hat{\beta}) = \text{sign}(\beta^0)$, where $\text{sign}(\cdot)$ maps positive entries to one, negative entries to -1, and zero entries to zero.

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0131

We extend this result to the *cliff-weak-sparsity* case. Without loss of generality, we assume $\beta^0 = (\beta_1^0, \ldots, \beta_s^0, \beta_{s+1}^0, \ldots, \beta_p^0)$, with $\beta_j^0 \gg n^{-1/2}$ for $j = 1, \ldots, s$, and $\beta_j^0 \ll n^{-1/2}$ for $j = s+1, \ldots, p$. Let $S = \{1, \ldots, s\}$ and $\beta_S^0 = (\beta_1^0, \ldots, \beta_s^0)$. Assuming the columns of X are ordered in accordance with the components of β^0, we write X_S and X_{S^c} as the first s and the last $p - s$ columns of X, respectively. Let $C = X^{\mathrm{T}}X/n$, which can be expressed in block-wise form, with four blocks, $C_{11} = X_S^{\mathrm{T}}X_S/n$, $C_{12} = X_S^{\mathrm{T}}X_{S^c}/n$, $C_{21} = X_{S^c}^{\mathrm{T}}X_S/n$, and $C_{22} = X_{S^c}^{\mathrm{T}}X_{S^c}/n$. Let $\Lambda_{\min}(A)$ and $\Lambda_{\max}(A)$ denote the smallest and largest eigenvalues of a matrix A. To obtain model selection consistency, we require the following assumptions:

Condition 1. ϵ_i are i.i.d. sub-Gaussian random variables.

Condition 2. The predictors are standardized, that is,

$$\frac{1}{n}\sum_{i=1}^{n} x_{ij} = 0, \qquad \frac{1}{n}\sum_{i=1}^{n} x_{ij}^2 = 1, \quad j = 1, \ldots, p.$$

Condition 3. There exists a constant $\Lambda > 0$, such that $\Lambda_{\min}(C_{11}) \geq \Lambda$.

Conditions 1 and 2 are fairly standard in the sparse linear regression literature; see, for example, (Zhao & Yu, 2006; Huang et al., 2008; Huang, Ma & Zhang, 2008). Theorems 1, 2, and 3 hold if we replace Condition 2 with a bounded second-moment condition. However, to simplify our argument, we use

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0131

Condition 2. Condition 3 ensures that the smallest eigenvalue of C_{11} is bounded away from zero, such that its inverse behaves well.

Condition 4. There exist constants $0 < c_1 < 1$ and $0 < c_2 < 1 - c_1$, such that

$$s = s_n = O(n^{c_1}) \ , \ p = p_n = O(e^{n^{c_2}}). \tag{3.1}$$

Condition 5 (Irrepresentable condition (Zhao & Yu, 2006))**.** There exists a constant vector η with entries in $(0, 1]$, such that $|C_{21}C_{11}^{-1}\text{sign}(\beta_S^0)| \leq \mathbf{1} - \eta$, where $\mathbf{1}$ is a $(p - s) \times 1$ vector with entries equal to one, and the inequality holds, element-wise.

Remark 1. The irrepresentable condition is implied by the slightly stronger condition, $|C_{21}C_{11}^{-1}| \leq \mathbf{1} - \eta$. This condition basically imposes a regularization constraint on the regression coefficients of the unimportant covariates (with small coefficients) on the important covariates (with large coefficients): the absolute value of any unimportant covariate's regression coefficient, represented by the important covariates, is strictly smaller than one. This condition can be weakened if we use other model selection criteria, such as stability selection.

Condition 6. There exist constants $c_1 + c_2 < c_3 \leq 1$ and $M > 0$, such that

$$n^{\frac{1-c_3}{2}} \min_{1 \leq i \leq s} |\beta_i^0| \geq M; \quad n^{\frac{1+c_1}{2}} \max_{s < j \leq p} |\beta_j^0| \leq M. \tag{3.2}$$

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0131

Condition 7. There exists a constant c_4 $(c_2 < c_4 < c_3 - c_1)$, such that the tuning parameter λ_1 in the definition of the lasso in equation (2.2) satisfies $\lambda_1 \propto n^{(c_4-1)/2}$. Based on empirical evidence from the simulation results (see subsection 4.2), we assume the tuning parameter $\lambda_2 \propto n^{-1}$.

Condition 8. Let c_4 be the constant defined in Condition 7, and suppose that

$$||\sqrt{n}C_{11}^{-1}C_{12}\beta^0_{S^c}||_\infty = O(1);\ ||\sqrt{n}(C_{21}C_{11}^{-1}C_{12} - C_{22})\beta^0_{S^c}||_\infty = o(n^{\frac{c_4}{2}}). \quad (3.3)$$

Condition 4 implies that both the number of larger components of β^0 (i.e., s) and the number of predictors (i.e., p) diverge with the sample size n. In particular, s is allowed to diverge much more slowly than n, and p can grow much faster than n (up to exponentially fast), which is standard in almost all of the high-dimensional inference literature. Although this assumption is stronger than the typical one $(s \log p)/n \to 0$, it has been used in previous works (Zhao & Yu, 2006). Condition 6 is the cliff-weak-sparsity assumption on β^0, which allows the existence of small, but nonzero coefficients, and is thus weaker than the hard sparsity and beta-min conditions. Conditions 1–5, the first half of the statement of Condition 6 on $\min_{1\le i\le s} |\beta_i^0|$, and the first half of the statement of Condition 7 on λ_1 are the same as those used in (Zhao & Yu, 2006) to show the sign-consistency of the lasso. Condition 8 is a technical assumption stating that the projection of small effects (i.e., $X_{S^c}\beta^0_{S^c}$) onto the linear subspace spanned

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0131

by the predictors corresponding to the large coefficients (i.e., the predictors in S) decays at a certain rate. In the Supplementary Material, we present examples where this condition holds. Conditions 1–5 and 7 are also assumed in (Liu & Yu, 2013) to show the validity of residual bootstrap lasso+ols.

An interesting fact is that both the lasso and the residual bootstrap lasso are model selection consistent under cliff-weak-sparsity and appropriate conditions.

Theorem 1. *Under Conditions 1 – 8, we have*

$$\text{pr}\left((\hat{\beta}_{\text{lasso}})_S =_s \beta^0_S,\ (\hat{\beta}_{\text{lasso}})_{S^c} = \mathbf{0}\right) = 1 - o(e^{-n^{c_2}}) \to 1 \ \text{ as } \ n \to \infty.$$

Remark 2. Theorem 1 shows that, under suitable conditions, the probability that the lasso correctly identifies the large coefficients of β^0, while shrinking the small ones to zero, goes to one at an exponential rate. This is a natural generalization of the sign consistency of the lasso from hard sparsity to cliff-weak-sparsity. We adopt the analytical techniques in (Zhao & Yu, 2006), with necessary modifications to account for the cliff-weak-sparsity. The proof is provided in the Supplementary Material.

3.3 Weak convergence of the rBLPR method

Condition 9. The number of large coefficients s satisfies $s^2/n \to 0$.

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0131

Condition 10. There exists a constant $D > 0$, such that

$$\max_{1\le i\le n} ||x_{i,S}||_2^2 = o(\sqrt{n});\ \max_{1\le i\le n} |x_{i,S^c}^{\mathrm{T}}\beta_{S^c}^0| < D,\ \text{where } x_{i,S} = (x_{i1},\ldots,x_{is})^{\mathrm{T}}.$$

Condition 9 is stronger than Condition 4 because it requires $0 < c_1 < 1/2$. Without considering model selection, Bickel & Freedman (1983) showed that a residual bootstrap OLS fails if p^2/n does not tend to zero. Therefore, Condition 9 cannot be weakened easily. This condition is weaker than $(s\log p)/\sqrt{n} \to 0$, as required by the de-sparsified lasso (Zhang & Zhang, 2014; van de Geer et al., 2014; Javanmard & Montanari, 2014). The first part of Condition 10 is not very restrictive, because the length of the vector $x_{i,S}$ is $s \ll \sqrt{n}$, and it holds, for example, when the predictors corresponding to the large coefficients are bounded by a constant M; that is, $|x_{ij}| \le M$, for $i = 1,\ldots,n,\ j = 1,\ldots,s$. This condition is also assumed in (Huang et al., 2008) to obtain the asymptotic normality of the bridge estimator. The second part of Condition 10 assumes that the small effects, $\{x_{i,S^c}^{\mathrm{T}}\beta_{S^c}^0, i = 1,\ldots,n\}$, are bounded from above by a constant.

Theorem 2 shows that the residual bootstrap lasso estimator also has sign-consistency under cliff-weak-sparsity and other appropriate conditions. The proof of this theorem is given in the Supplementary Material.

Theorem 2. *Under Conditions 1 – 10 , the residual bootstrap lasso estimator*

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0131

has sign-consistency; that is,

$$\text{pr}\left((\hat{\beta}^*_{\text{rBlasso}})_S =_s \beta^0_S,\ (\hat{\beta}^*_{\text{rBlasso}})_{S^c} = \mathbf{0} \mid \epsilon\right) = 1 - o_p(e^{-n^{c_2}}).$$

Remark 3. By Theorem 2, the residual bootstrap lasso correctly identifies the large coefficients and shrinks the small ones to zero, with probability approaching one. The proposed bootstrap LPR method uses the partial ridge regression to recover these small, but nonzero coefficients.

Using Theorems 1 and 2 and Condition 11, we can show that the rBLPR procedure can consistently estimate the distribution of $\hat{\beta}_{\text{LPR}}$ and, thus, construct asymptotically valid confidence intervals for the regression coefficients β^0.

Let I be a $(p-s)\times(p-s)$ identity matrix, and denote the matrix C_{λ_2} as

$$C_{\lambda_2} = \begin{pmatrix} C_{11} & C_{12} \\ C_{21} & C_{22} + \lambda_2 I \end{pmatrix}. \tag{3.4}$$

Condition 11. Let $u \in \mathbb{R}^p$ be a fixed vector, with $||u||_2 = 1$. Assume $\sigma_1^2 = \lim_{n\to\infty}\left(u^{\mathrm{T}} C_{\lambda_2}^{-1} C (C_{\lambda_2}^{-1})^{\mathrm{T}} u\right)\sigma^2 < \infty$ and

$$\max\left\{\left(\beta^0_{S^c}\right)^{\mathrm{T}} C_{22}\left(\beta^0_{S^c}\right),\ \max_{1\le k\le n}\frac{\left|u^{\mathrm{T}} C_{\lambda_2}^{-1} x_k\right|}{\sqrt{n}},\ \frac{u^{\mathrm{T}} C_{\lambda_2}^{-1}\left(\mathbf{0}^{\mathrm{T}}, (\beta^{\mathbf{0}}_{\mathbf{S^c}})^{\mathrm{T}}\right)^{\mathrm{T}}}{\sqrt{n}}\right\} = o(1).$$

Remark 4. The first statement $\left(\beta^0_{S^c}\right)^{\mathrm{T}} C_{22}\left(\beta^0_{S^c}\right) = o(1)$ is used to guarantee that the conditional variance of ϵ_i^*, given ϵ, converges to σ^2, the variance of ϵ_i. Hence, the conditional distribution of ϵ_i^* is a valid approximation to the distribution of

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0131

ϵ_i. The other two statements are a Linderberg-type condition and a technical condition, which are used to obtain the asymptotic normality.

Remark 5. For an orthogonal design matrix (i.e., $(1/n)X^{\mathrm{T}}X = I$), in which there are no correlations between predictors, $p \leq n$, and $\sigma_1^2 = \sigma^2$. Then Condition 11 reduces to the following, much simpler form: $\max_{1\leq k\leq n} |u^{\mathrm{T}}x_k| = o(\sqrt{n})$. When $u = e_j$, a basis vector with the jth element equal to one and other elements equal to zero, this condition is equivalent to $\max_{1\leq k\leq n} |x_{kj}| = o(\sqrt{n})$, which is not a strong condition, and is expected to hold in many practical situations. The conclusion is still true when the correlation between two predictors satisfies $\mathrm{cor}(X_i, x_j) = \rho^{|i-j|}$, with $\rho < 1/5$ (see Section S3 for more detail).

Theorem 3. *Under conditions 1 – 11, we have*

$$\sqrt{n}u^{\mathrm{T}}(\hat{\beta}_{\text{LPR}} - \beta^0) = U + o_p(1); \quad \sqrt{n}u^{\mathrm{T}}(\hat{\beta}^*_{\text{rBLPR}} - \hat{\beta}_{\text{lasso+ols}}) = U^* + o_p(1).$$

Both U and $(U^* \mid \epsilon)$ converge in distribution to the normal distribution $N(0, \sigma_1^2)$.

Remark 6. Theorem 3 shows that, under appropriate conditions, the limiting distribution of $\sqrt{n}u^{\mathrm{T}}(\hat{\beta}^*_{\text{rBLPR}} - \hat{\beta}_{\text{lasso+ols}})$, conditional on ϵ, is the same as the limiting distribution (unconditional) of $\sqrt{n}u^{\mathrm{T}}(\hat{\beta}_{\text{LPR}} - \beta^0)$. Thus, the unknown distribution of $\sqrt{n}u^{\mathrm{T}}(\hat{\beta}_{\text{LPR}} - \beta^0)$ can be approximated by the conditional distribution of $\sqrt{n}u^{\mathrm{T}}(\hat{\beta}^*_{\text{rBLPR}} - \hat{\beta}_{\text{lasso+ols}})$, which can be estimated using the bootstrap. Based on the estimated conditional distribution of $\sqrt{n}u^{\mathrm{T}}(\hat{\beta}^*_{\text{rBLPR}} - \hat{\beta}_{\text{lasso+ols}})$, we

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0131

can construct asymptotically valid confidence intervals for the linear combination $u^{\mathrm{T}}\beta^0$. Specifically, by setting $u = e_j$, we can construct an asymptotically valid confidence interval for an individual coefficient β_j^*.

We can also show the model selection consistency of the paired bootstrap lasso estimator (similar to Theorem 2). However, even in the orthogonal design matrix case, the design matrix X^* of the paired bootstrap samples is no longer orthogonal, making the components of the pBLPR estimates, $(\hat{\beta}^*_{\text{pBLPR}})_S$ and $(\hat{\beta}^*_{\text{pBLPR}})_{S^c}$, dependent on each other and, thus, have complicated forms. Hence, it becomes difficult to verify the convergence property of the pBLPR estimator using techniques similar to those used to prove Theorem 3 for the rBLPR estimator. Our simulation studies in the following section indicate that the pBLPR method works as well as the rBLPR method. We leave the theoretical analysis of the pBLPR method to future research.

4. Simulation studies

We perform simulation studies to evaluate the finite-sample performance of two bootstrap LPR methods, rBLPR and pBLPR. We compare our method with the bootstrap lasso+ols method and three de-sparsified lasso methods (LDPE, JM, and BLDPE) in terms of their coverage probabilities and confidence interval lengths. Additional information about the simulation studies and results is

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0131

provided in the Supplementary Material. The main conclusions are summarized as follows:

(1) $\lambda_2 = O(1/n)$ works well for a wide range of noise levels.

(2) pBLPR is slightly better than rBLPR, in most cases.

(3) Under the setting of normal design matrices, bootstrap lasso+ols has the shortest confidence interval lengths, with good coverage probabilities for large coefficients. However, for small, but nonzero coefficients, rBLPR and pBLPR have the shortest confidence interval lengths, with good coverage probabilities.

(4) LDPE and JM are more robust to low signal-to-noise ratios (SNRs), whereas rBLPR and pBLPR do not perform well when the SNRs are low, that is, no greater than one. This is mainly because the lasso cannot select all of the important predictors correctly. The rBLPR and pBLPR methods produce much better confidence intervals when the SNRs are high, that is, larger than five: with comparable coverage probabilities, their interval lengths are 50% shorter than those of LDPE and JM, on average.

(5) With regard to the point estimates of the linear model coefficients, the LPR estimator has smaller biases for most coefficients than those of LDPE and JM. However, its standard deviations are larger than those of LDPE and JM for large coefficients, and are smaller for small coefficients. Overall, its root mean squared errors (RMSEs) are 60% smaller than those of LDPE, but 42% larger

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0131

than those of JM.

(6) When the predictors are generated from a Student's t distribution with two degrees of freedom, the methods all fail to produce valid confidence intervals. New statistical techniques are needed for inference in this case.

(7) Our rBLPR and pBLPR methods are robust to model misspecification, and the confidence intervals constructed using our methods are more than 50% shorter, on average, than those produced by LDPE and JM.

(8) BLDPE has the best coverage probabilities of the considered methods. Its confidence interval lengths are close to the better ones of LDPE and JM, but are still longer than those of pBLPR and rBLPR.

5. Real-data case study 1: fMRI data

In this section, we demonstrate the performance of our method on a real fMRI data set and compare its performance with that of two de-sparsified methods. The fMRI data were provided by the Gallant Lab at UC Berkeley (Kay et al., 2008). The fMRI measured blood oxygen level-dependent activity at 1331 discretized 3D brain volumes ($2 \times 2 \times 2.5$ millimeters): cube-like units called voxels. We use a sub-dataset focusing on the responses in the ninth voxel, located in the brain region responsible for visual functions. A single human subject was shown pictures of everyday objects, such as trees, stars, and so on. Each

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0131

picture was a 128 pixel by 128 pixel grayscale image, reduced to a vector of length 10921, as follows: (1) use a Gabor transform of the gray image to generate local contrast energy features Z_j; and (2) take the nonlinear transformation $X_j = \log(1 + \sqrt{Z_j})$, for $j = 1, \ldots, 10921$. Training and validation data sets were collected during the experiment. There were 1750 natural images in the training data, consisting of a design matrix of dimensions 1750×10921. The validation data set contained responses to 120 natural images (we do not use the validation data in this study).

After reading the training data set into R, we calculate the variance of each feature (column) in X, and delete those columns with variances $\leq 1e^{-4}$. Then, we have a matrix of dimension 1750×9076. We further reduce the dimension of the design matrix using correlation screening, that is, sorting the correlations (Pearson correlation between every feature in X and the response Y) in decreasing order of absolute value, and then selecting the top 500 features with the largest correlations. We use the lasso+ols estimate of the feature coefficients, based on the 1750×500 design matrix, as the pseudo-true parameter values, denoted by β^0. We randomly choose a subset of $n = 200$ rows to create a high-dimensional simulation setting, and then generate Y from a linear regression model $y_i = x_i^{\mathrm{T}}\beta^0 + \epsilon_i$. We set $B = 1000$ for the number of replications in the bootstrap, and compare the performance of the pBLPR method with that of

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0131

LDPE and JM.

Based on the sub-dataset with $n = 200$ and $p = 500$, we evaluate the performance of pBLPR, LDPE, and JM in their construction of confidence intervals. The 95% confidence intervals constructed by these three methods cover 95.8%, 97%, and 99.6%, respectively, of the 500 components of β^0. All three methods cover more than 95% of the pseudo-true values and, thus, have satisfactory performance in terms of coverage. In terms of interval lengths, however, our pBLPR method produces much shorter confidence intervals than those of the other two methods for most of the coefficients, especially the small ones. Figure S15 shows the confidence interval lengths of 100 coefficients (44 nonzero coefficients in β^0 and 56 randomly chosen zero coefficients) produced by the three methods. The satisfactory coverage and much shorter lengths of the confidence intervals produced by pBLPR, demonstrate that it outperforms LDPE and JM, overall, in this real-data case study.

6. Real-data case study 2: neuroblastoma gene expression data

In this section, we apply our proposed pBLPR and rBLPR methods and three de-sparsified lasso methods (LDPE, JM, and BLDPE), to a data set containing $43,827$ gene expression measurements from the Illumina RNA sequencing of 498 neuroblastoma samples. More details about this data set can be found in the

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0131

Supplementary Material.

Constructing gene-gene regulatory relationships is of primary interest for this data set. In this section, we apply five methods (pBLPR, rBLPR, LDPE, JM, and BLDPE) to identify the significant genes that affect the expression of a gene called *CAMTA1*, which is known to be neuroblastoma-related and is observed to be highly correlated with the risk of neuroblastoma. Given our lack of knowledge on the complex regulatory relationships between genes, the linear model is almost certainly a misspecified model. However, this case study serves as a reasonable real-data example to demonstrate the ability of our pBLPR and rBLPR methods and the three de-sparsified lasso methods (LDPE, JM, and BLDPE) to identify significant predictors in a misspecified linear model.

The results show that LDPE and its bootstrap version, BLDPE, find the most significant genes; pBLPR and rBLPR find 91 and 26 significant genes, respectively; JM finds only one significant gene. The functions related to natural and regulated cell deaths (e.g., apoptosis and autophagy), which are key processes used to prevent cancer, are only enriched in the significant genes found by pBLPR or rBLPR, but not in those found by any of the de-sparsified lasso methods. On the other hand, only general functions, such as basic processes in cells, are enriched in the significant genes found by each de-sparsified lasso method, but not by our methods. This suggests that pBLPR and rBLPR find

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0131

significant features that are more reasonable and interpretable, based on domain knowledge, implying that pBLPR and rBLPR are robust to model misspecification. The detailed analysis results are provided in the Supplementary Material and additional Supplementary File.

7. Conclusion and future work

Assigning p-values and constructing confidence intervals for parameters in high-dimensional sparse linear models are challenging tasks. The bootstrap, as a standard inference tool, has been shown to be useful in addressing this problem. However, previous works that extended the bootstrap technique to high-dimensional models relied on two key assumptions: hard sparsity and the beta-min condition. The beta-min condition is rather restrictive. In order to relax it, we propose two new bootstrap procedures based on a new two-stage estimator, called lasso+partial ridge. Our methods improve the performance of the bootstrap lasso+ols method proposed in (Liu & Yu, 2013) when there exist a group of small, but nonzero regression coefficients. We conduct extensive simulation studies to compare our methods with three de-sparsified methods (LDPE, JM, and the bootstrap version of LDPE (BLDPE)). We find that our methods yield comparable coverage probabilities, but shorter (on average) intervals, and are more robust to misspecified models than the other methods are under many sce-

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0131

narios. We apply our methods to an fMRI data set, finding that it gives reasonable coverage probabilities and shorter interval lengths than those of LDPE, JM, and BLDPE. In a second real-data application, we applied our methods to identify genes that have significant effects on predicting a cancer gene's expression levels in a (likely) misspecified linear model. Compared with three de-sparsified lasso methods, our methods find genes that are biologically more reasonable and interpretable, suggesting that our methods are robust to model misspecification in certain applications, despite the lack of rigorous theoretical analysis in this work. Future work is needed to investigate the robustness of various inference methods to different types of model misspecification, from both theoretical and empirical perspectives.

A disadvantage of our method is that its resulting inference is not uniformly valid over the class of sparse models, owing to the cliff-weak-sparsity assumption. It is possible that our methods are uniformly valid for some pseudo-true parameter, that is, the parameters of the nearest model that satisfies cliff-weak-sparsity; we leave this to future work. Moreover, compared with uniformly valid inference procedures such as the de-sparsified lasso methods, our empirical studies show that our methods are more likely to identify small, but nonzero coefficients, owing to the shorter confidence interval lengths returned by our methods. In many real-world applications, the covariates (or features) with small effects

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0131

are not negligible, but may be important. For example, in genomic applications, where complex gene-gene regulatory relationships are of primary interest, researchers searching for regulators of a target gene are not only interested in the genes with large effects, but also in other genes with small effects. This is because many small effects have been discovered to play important functional roles in biological mechanisms. In this application, our methods provide a means to identify genes with small effects. However, note that subsequently experiments are still required to validate the identified genes. Furthermore, when an individual coefficient is too small, no method can successfully identify it; then, a statistical procedure should instead aim to detect the joint significance of a set of covariates.

Overall, the bootstrap lasso+ols method has the shortest confidence interval lengths, with good coverage probabilities, for large coefficients. However, for small, but nonzero coefficients, the bootstrap LPR method (rBLPR and pBLPR) has the shortest confidence interval lengths, with good coverage probabilities. Therefore, if practitioners focus on the confidence intervals for large coefficients, we recommend the bootstrap lasso+ols method; however, if they are also interested in identifying small, but significant coefficients in a possibly misspecified linear model, we recommend our bootstrap LPR methods. Nevertheless, note that the confidence intervals of the coefficients, with magnitudes

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0131

of order $1/\sqrt{n}$, may be invalid. If practitioners' major concern is the coverage probabilities of confidence intervals, they should use the de-sparsified lasso methods, which are uniformly valid over the class of sparse models. Moreover, from an application perspective, our bootstrap LPR methods have the advantages of being technically simple, interpretable, and easy to implement and parallelize.

Finally, multiple testing is another important task in hypothesis testing, and is closely related to confidence interval construction. Several procedures, such as the Bonferroni correction, Benjamini–Hochberg procedure and FDR control, have been proposed to correct multiple testing in low-dimensional settings. However, these procedures are based on accurate estimations of the p-values of each test, where small p-values can only be obtained using large numbers of bootstrap runs (e.g., a p-value of 0.001 requires at least 1000 runs), thus imposing too much computational complexity. We leave the correction for multiple testing in high-dimensional models as future work.

Supplementary Material

The online Supplementary Material includes proofs, algorithms, and simulation results. An additional Supplementary File contains the detailed Gene Ontology analysis results for real-data case study 2.

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0131

Acknowledgments

The authors would like to thank the Gallant Lab at UC Berkeley for providing the fMRI data, Simon Walter (UC Berkeley) and Dr. Chad Hazlett (UCLA) for their edits and suggestions that have helped clarify the text, and Prof. Bin Yu at UC Berkeley for her helpful discussions and comments that have helped improve the quality of the paper. Dr. Hanzhong Liu's research is partially supported by NSF grants DMS-1613002, DMS-1228246, AFOSR grant FA9550-14-1-0016, and the National Natural Science Foundation of China 11701316. Dr. Jingyi Jessica Li's research is supported by the Hellman Fellowship, the PhRMA Foundation Research Starter Grant in Informatics, the Sloan Research Fellowship, the Johnson & Johnson WiSTEM2D Award, NIH/NIGMS grant R01GM120507, and NSF grants DMS-1613338 and DBI-1846216.

References

Barber, R. F. and Candès, E. J. (2015). Controlling the false discovery rate via knockoffs. *Ann. Statist.* **43**, 2055–2085.

Belloni, A., Chernozhukov, V., and Hansen C. (2014). Inference on Treatment Effects after Selection among High-Dimensional Controls. *Review of Economic Studies* **81**, 608–650.

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

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0131

Bickel, P. J. and Freedman, D. A. (1983). Bootstrapping regression models with many parameters. In *Festschrift for Erich L. Lehmann* (P. Bickel, K. Doksum, and J. Hodges, Jr., eds.) 28–48. Wadsworth, Belmont, Calif.

Bühlmann, P. (2013). Statistical significance in high-dimensional linear models. *Bernoulli* **19**, 1212–1242.

Bühlmann, P. and van de Geer, S. (2011). *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Springer.

Candès, E. J. and Tao, T. (2007). The Dantzig selector: Statistical estimation when p is much larger than n. *Ann. Statist.* **35**, 2312–2351.

Chatterjee, A. and Lahiri, S. N. (2010). Asymptotic Properties of the Residual bootstrap for lasso Estimators. *P. Am. Math. Soc.* **138**, 4497–4509.

Chatterjee, A. and Lahiri, S. N. (2011). Bootstrapping lasso estimators. *J. Am. Statist. Assoc.* **106**, 608–625.

Chatterjee, A. and Lahiri, S. N. (2013). Rates of convergence of the adaptive lasso estimators to the oracle distribution and higher order refinements by the bootstrap. *Ann. Statist.* **41**, 1232–1259.

Chernozhukov, V., Hansen C. and Liao, Y. (2017). A lava attack on the recovery of sums of dense and sparse signals. *Ann. Statist.* **45**, 39–76.

Dezeure, R., Bühlmann, P., Meier, L. and Meinshausen, N. (2014). High-dimensional Inference: Confidence intervals, p-values and R-Software hdi. *Stat. Sci.* **30**, 533–558.

Dezeure, R., Bühlmann, P. and Zhang, C-H. (2017). High-dimensional simultaneous inference with the bootstrap. *Test* **26**, 685–719.

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0131

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Am. Statist. Assoc.* **96**, 1348–1360.

Fan, J. and Lv, J. (2010). A selective overview of variable selection in high dimensional feature space. *Stat. Sinica* **20**, 101–148.

Gao, X., Ahmed, S. E. and Feng, Y. (2017). Post selection shrinkage estimation for high-dimensional data analysis. *Appl. Stoch. Model Bus.* **33**, 97–120.

Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics* **12**, 55–67.

Huang, J., Horowitz, J. and Ma, S. (2008). Asymptotic properties of bridge estimators in sparse high-dimensional regression models. *Ann. Statist.* **36**, 587–613.

Huang, J., Ma, S. and Zhang C.-H. (2008). Adaptive lasso for sparse high-dimensional regression models. *Stat. Sinica* **18**, 1603–1618.

Javanmard, A. and Montanari, A. (2014). Confidence Intervals and Hypothesis Testing for High-Dimensional Regression. *J. Mach. Learn. Res.* **15**, 2869–2909.

Kay, K. N., Naselaris, T., Prenger, R. J. and Gallant, J. L. (2008). Identifying natural images from human brain activity. *Nature* **452**, 352–355.

Knight, K. and Fu, W. J. (2000). Asymptotics for lasso-type estimators. *Ann. Statist.* **28**, 1356–1378.

Lee, J. D., Sun, D. L., Sun, Y., and Taylor, J. E. (2015). Exact post-selection inference, with application to the lasso. *arXiv*: 1311.6238.

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0131

Lockhart, R., Taylor, J., Tibshirani, R. J. and Tibshirani, R. (2014). A significance test for the lasso. *Ann. Statist.* **42**, 413–468.

Lopes, M. (2014). Residual bootstrap for High-Dimensional Regression with Near Low-Rank Designs. *NIPS* **15**, 3239–3247.

Liu, H. and Yu, B. (2013). Asymptotic properties of lasso+mLS and lasso+Ridge in sparse high-dimensional linear regression. *Electron. J. Stat.* **7**, 3124–3169.

Meinshausen, N. and Bühlmann, P. (2006). High dimensional graphs and variable selection with the lasso. *Ann. Statist.* **34**, 1436–1462.

Meinshausen, N. and Bühlmann, P. (2010). Stability selection. *J. R. Statist. Soc.* B **72**, 417–473.

Meinshausen, N. (2015). Group-bound: confidence intervals for groups of variables in sparse high-dimensional regression without assumptions on the design. *J. R. Statist. Soc.* B **77**, 923–945.

Minnier, J., Tian, L. and Cai, T. (2009). A perturbation method for inference on regularized regression estimates. *J. Am. Statist. Assoc.* **106**, 1371–1382.

Negahban, S., Ravikumar, P., Wainwright M. J. and Yu, B. (2009). A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers. *Stat. Sci.* **28**, 538–557.

Shi, P. and Qu, A. (2017). Weak Signal Identification and Inference in Penalized Model Selection. *Ann. Statist.* **45**, 1214–1253.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. R. Statist. Soc.* B **58**, 267–288.

van de Geer, S., Bühlmann, P., Ritov, Y. and Dezeure, R. (2014). On asymptotically optimal confidence

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0131

regions and tests for high-dimensional models. *J. R. Statist. Soc.* B **42**, 1166–1202.

Wainwright, M. J. (2009). Sharp thresholds for high-dimensional and noisy sparsity recovery using-constrained quadratic programming (lasso). *Information Theory, IEEE* **55**, 2183–2202..

Wasserman, L. and Roeder, K. (2009). Weak Signal Identification and Inference in Penalized Model Selection. *Ann. Statist.* **45**, 1214–1253.

Zhang, X., and Cheng, G. (2017). Simultaneous inference for high-dimensional linear models. *J. Am. Statist. Assoc.* B **112**, 757–768.

Zhang, C-H., and Zhang, S. S. (2014). Confidence intervals for low dimensional parameters in high dimensional linear models. *J. R. Statist. Soc.* B **76**, 217–242.

Zhao, P. and Yu, B. (2006). On Model Selection Consistency of lasso. *J. Mach. Learn. Res.* **7**, 2541–2563.

Hanzhong Liu

Center for Statistical Science, Department of Industrial Engineering, Tsinghua University, Beijing, China

E-mail: lhz2016@tsinghua.edu.cn

Xin Xu

Department of Statistics, Yale University, New Haven, Connecticut, U.S.A. E-mail: xin.xu@yale.edu

Jingyi Jessica Li

Department of Statistics, University of California, Los Angeles, California, U.S.A.

* To whom correspondence should be addressed: E-mail: jli@stat.ucla.edu