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
Notice: Accepted version subject to English editing.	

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], and Jingyi Jessica Li[3*]

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

Abstract:

For high-dimensional sparse linear models, how to construct confidence intervals for coefficients remains a difficult question. The main reason is the complicated limiting distributions of common estimators such as the lasso. Several confidence interval construction methods have been developed, and bootstrap lasso+ols is notable for its simple technicality, good interpretability, and comparable performance with other more complicated methods. However, bootstrap lasso+ols depends on the beta-min assumption, a theoretic criterion that is often violated in practice. In this paper, we introduce a new method called bootstrap lasso+partial ridge to relax this assumption. Lasso+partial ridge is a two-stage estimator: first using lasso to select features and subsequently using partial ridge 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 violating the beta-min assumption. For such coefficients, compared to bootstrap lasso+ols, confidence intervals constructed by bootstrap lasso+partial ridge have on average 50% larger coverage probabilities. Bootstrap lasso+partial ridge also has on average 35% shorter confidence in-

terval lengths than the de-sparsified lasso methods, regardless of whether linear models are misspecified. Additionally, we provide theoretical guarantees of 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

There has been rapid growth in the production and needs to analyze high dimensional data in a variety of fields including information technology, astronomy, neuroscience and bioinformatics, to name just a few. 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, methodology, and algorithms have been developed to tackle high-dimensional data problems under certain sparsity constraints, e.g., the number of nonzero linear model coefficients s is much smaller than the sample size n. Regularization is required to perform sparse estimation under this regime. For example, the lasso (Tibshirani, 1996) uses l_1 regularization to perform model selection and parameter estimation simultaneously in high dimensional sparse linear regression. Previous work has focused on the recovery of a sparse parameter vector (denoted by $\beta^0 \in R^p$) based on common criteria such as: (i) model selection consistency; (ii) l_q estimation error $||\hat{\beta} - \beta^0||_q$, where $\hat{\beta}$ is an estimate

of β^0 and q typically equals 1 or 2; (iii) 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 question at the frontier of high-dimensional statistical research is how to perform statistical inference, i.e., 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 prediction. However, inference is difficult for high-dimensional model parameters, because the limiting distribution of common estimators, e.g., the lasso, is complicated and hard to compute in high-dimensions. Facing this challenge, here we develop a novel and practical inference procedure called bootstrap lasso+partial ridge, which is based on three canonical methods: the bootstrap, the lasso and the ridge. Before presenting our method, we first briefly review the existing high-dimensional inference methods in the next two paragraphs.

There is a growing statistical literature that tackles high-dimensional inference problems. Existing methods belong to several categories, including the sample splitting based methods, the bootstrap based methods, the de-sparsified lasso methods, the post-selection inference methods, and the knockoff filter. In particular, Wasserman and Roeder proposed a sample splitting method (Wasser-

man & Roeder, 2009), which splits n data points into two halves, with the first half to be used for model selection (say by the lasso) and the second half for constructing 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 the scenarios with p going 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 which are similar to the irrepresentible condition and the beta-min condition (Beta-min condition means the minimum absolute value of nonzero regression coefficients is much larger than $n^{-1/2}$) that 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 and showed consistency under the irrepresentible condition, the beta-min condition and other regularity conditions (Liu & Yu, 2013) (lasso+ols means using lasso to select a model and next using ols, ordinary least squares, to refit the coefficients in the selected

model). A main issue with these methods is that they all require the rather restrictive beta-min condition, which should be better relaxed in high-dimensional inference if possible.

Besides the above sample splitting or bootstrap based methods, the de-sparsified lasso, which was first proposed by Zhang & Zhang (2014) and later investigated by van de Geer et al. (2014), Javanmard & Montanari (2014), is another type of methods. They aim to remove the biases of the lasso estimates and produce an asymptotically normal estimate for each individual parameter. Specifically, for the two typical de-sparsified lasso methods developed by Zhang & Zhang (2014) and Javanmard & Montanari (2014), we will refer to them as LDPE and JM. These methods do not rely on the beta-min condition but on the other hand require estimating the precision matrix of predictors using the graphical lasso (Zhang & Zhang, 2014; van de Geer et al., 2014) or another convex optimization procedure (Javanmard & Montanari, 2014). There are two main issues with these methods. First, these methods reply heavily on the sparse linear model assumption and thus may have 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, they can serve as a theoretically proven benchmark for high-dimensional infer-

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

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

geous to adopt a different second-step procedure after the lasso to replace the ols. Ideally this procedure should put no penalty on the selected coefficients by the lasso to reduce the bias and have a small but nonzero l_2 penalty on the unselected coefficients to recover the small but nonzero ones. Here, we name this estimator as lasso+partial ridge (LPR), which will be consistently used throughout the reminder of this paper. An independent work of Gao et al. (2017) who also proposed a post selection ridge estimator that is similar to our lasso+partial ridge estimator. The difference is that they aimed to improve the prediction performance and thus added a thresholding step to achieve this goal. Another work by Chernozhukov, Hansen & Liao (2017) proposed a penalization-based estimation strategy called Lava to deal with the "sparse + dense" coefficients. However, their goal is also to improve prediction performance instead of inference.

In this paper, we propose a new inference procedure called bootstrap lasso+partial ridge (bootstrap LPR) to improve over the bootstrap lasso+ols method. The problem setting is to construct confidence intervals for individual regression coefficients $\beta_j^0, j = 1, \ldots, p$ in a high-dimensional linear regression model where β^0 is weakly sparse (Negahban et al., 2009), i.e., 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 if we order the absolute coefficients from the largest

to the smallest, there exits a big drop like a cliff that divides the coefficients into tow groups. Obviously, the cliff-weak-sparsity is a weaker assumption than the hard (or exact) sparsity (β^0 has at most s ($s \ll n$) nonzero elements) and the beta-min condition.

Inference for small coefficients was also investigated by Shi & Qu (2017), who proposed a two-step inference procedure to identify weak signals (small coefficients). Their method is designed for the orthogonal design matrix and based on a combination of the asymptotic normality of a biased-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 $p \ll n$, while our method based on the bootstrap can be used when $p \gg n$.

After submitting this paper, it was brought to our attention two interesting and thought-provoking papers (Dezeure et al., 2017; Zhang & Cheng, 2017), which combined the bootstrap and the de-sparsified lasso method to deal with non-Gaussian and heteroscedastic errors. We refer this method as the bootstrap version of LDPE (BLDPE), and we add it to the method comparison in our simulation and real data studies.

Our contributions 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 to construct confidence interval for each individual 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: (1) Compared with bootstrap lasso+ols, bootstrap LPR improves the coverage probabilities of 95% confidence intervals by about 50% on average for small but nonzero regression coefficients, at the price of 15% heavier computational burden for $n = 200,\ p = 500$; (2) Compared with two de-sparsified lasso methods, LDPE and JM, bootstrap LPR has comparably good coverage probabilities for large and small regression coefficients, and in some cases outperforms their methods by producing conference intervals with more than 50% shorter interval lengths on average; (3) bootstrap LPR is more than 30% faster than these two de-sparsified lasso methods and is robust to model misspecification. We also demonstrate the performance of bootstrap LPR on two real data sets: fMRI (functional magnetic resonance imaging) data and neuroblastoma gene expression data.

Third, we extend 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 the "large" elements of β^0 while shrinking all the "small" elements to zero.

Fourth, we develop an R package "HDCI" to implement the bootstrap lasso, the bootstrap lasso+ols and our new 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 criteria such as stability selection (Meinshausen & Bühlmann, 2010), SCAD (Fan & Li, 2001), Dantzig selector (Candès & Tao, 2007) and post-double selection (Belloni et al., 2014). The post-double selection is promising because it does not require the beta-min condition. If we replace the lasso by this method, the resulting confidence intervals may achieve better coverages for medium-size coefficients. This is an interesting research direction that worth further investigation, because the methodology, computation and theory will be different from our current work in many aspects.

Our paper is organized as follows: In Section 2, we define the lasso+partial ridge (LPR) estimator and introduce the residual bootstrap LPR (rBLPR) and the paired bootstrap 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

and BLDPE). In Sections 5 and 6, we present two real data case studies. Section 7 is the conclusion and future work. The relevant proofs, algorithms, and simulation details can be found in the Supplementary Material.

2. Framework and definitions

2.1 Overview and background

In this section, we begin with an introduction to the basic background of high-dimensional sparse linear models. We next define the cliff-weak-sparsity and the lasso+partial ridge (LPR) estimator. Finally, we propose two bootstrap procedures (the residual bootstrap and the paired bootstrap) based on the LPR estimator to construct confidence intervals for individual regression coefficients.

Assuming data are generated from a 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.) error random 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, i.e., $\sum_{i=1}^n x_{ij}/n = 0, j = 1, \ldots, p$, and there is no intercept term in the linear model. Denoting $\beta^0 \in \mathbb{R}^p$ as a vector of

coefficients, we assume that β^0 satisfies the cliff-weak-sparsity.

Definition 1 (Cliff-weak-sparsity)**.** β^0 satisfies the 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 $n^{-1/2}$.

In this paper, we are interested in constructing confidence intervals of each individual coefficient $\beta_j^0, j = 1, \ldots, p$. 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, which 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 controling the amount of regularization applied to the estimate. The λ_1 generally 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 work relies on two restrictive assumptions: hard sparsity and beta-min condition. In

order to relax these two often unrealistic assumptions, we propose a lasso+partial ridge (LPR) estimator and two bootstrap procedures on it, i.e., residual bootstrap LPR (rBLPR) and paired bootstrap LPR (pBLPR) in the following subsections.

2.2 Lasso+partial ridge (LPR) estimator

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

In *cliff-weak-sparsity* case, existing bootstrap methods, e.g., bootstrap lasso+ols, give very poor coverage probabilities for the small but nonzero regression coefficients because they are seldom selected by the lasso and hence a large fraction of the bootstrap lasso+ols estimates for them are 0, producing zero-length and non-coverage confidence intervals like $[0, 0]$. To fix this problem, we need to increase the variance of our estimates for small coefficients whose corresponding predictors are missed by the lasso. This is the motivation for the lasso+partial ridge (LPR) estimator proposed in this paper.

The LPR estimator is a two-stage estimator that adopts the lasso to select predictors and then refits the coefficients by the partial ridge, which is defined to minimize the empirical l_2 loss with no penalty on the selected predictors but an l_2

penalty on the unselected predictors, so as to reduce 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 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 selected predictors by the lasso, then we define the LPR estimator as:

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

The λ_2 is a tuning parameter generally depending 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 $\lambda_2 = 1/n$ in this paper with the understanding that further research should be done on the selection of λ_2.

In the next two subsections, we will separately discuss two commonly used bootstrap procedures on the LPR estimator and explain how to use them to construct confidence intervals for each individual coefficient.

2.3 Residual bootstrap lasso+partial ridge (rBLPR)

For a deterministic design matrix X in a linear regression model, the residual bootstrap is a standard method for constructing confidence intervals. In this

subsection, we will introduce the residual bootstrap LPR procedure.

We first need to appropriately define residuals so that their empirical distribution can well approximate the true distribution of the error ϵ_i's. In high-dimensional linear regression, there are different ways to obtain residuals; for example, we may calculate the residuals from different estimation methods such as the lasso, the lasso+ols and the LPR. Simulation suggests the residuals obtained from the lasso+ols approximate the true distribution of the error ϵ_i's the best and hence will be adopted in this paper (Note that, when the beta-min condition is not satisfied, the lasso+ols could fail to correctly select all the nonzero coefficients, i.e., it is not consistent for model selection, but its prediction performance could still be good, i.e., smaller mean-squared-error than 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: $\hat{\epsilon} = (\hat{\epsilon}_1, \ldots, \hat{\epsilon}_n)^{\mathrm{T}} = Y - X\hat{\beta}_{\text{lasso+ols}}$. Consider the centered residuals at the mean $\{\hat{\epsilon}_i - \tilde{\epsilon},\ i = 1, \ldots, n\}$, where $\tilde{\epsilon} = \sum_{i=1}^n \hat{\epsilon}_i/n$. For the residual bootstrap, one obtains $\epsilon^* = (\epsilon_1^*, \ldots, \epsilon_n^*)^{\mathrm{T}}$ by resampling with replacement from the centered residuals $\{\hat{\epsilon}_i - \tilde{\epsilon},\ i = 1, \ldots, n\}$, and constructs the residual bootstrap ("rboot") version of Y:

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

Then, based on the residual bootstrap sample (X, Y^*_{rboot}), one 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 $\hat{S}^*_{\text{rBlasso}} = \{j \in \{1, \ldots, p\} : (\hat{\beta}^*_{\text{rBlasso}})_j \neq 0\}$ and also the residual bootstrap LPR (rBLPR) estimator $\hat{\beta}^*_{\text{rBLPR}}$ as in (2.7) in the same way as in equation (2.3) except that replacing $Y, \hat{S}$ by $Y^*_{\text{rboot}}, \hat{S}^*_{\text{rBlasso}}$ 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 whole procedure.

2.4 Paired bootstrap lasso+partial ridge (pBLPR)

In this subsection, we will introduce the paired bootstrap LPR (pBLPR) procedure. Paired bootstrap is another bootstrap procedure widely used for the inference in linear models. In this procedure, one generates a resample $\{(x^*_i, y^*_i), i = 1, \ldots, n\}$ from the empirical distribution of $\{(x_i, y_i), i = 1, \ldots, n\}$ and then computes 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 selected predictors by the paired bootstrap lasso and define the paired bootstrap LPR (pBLPR) estimator by

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

The paired bootstrap LPR 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 residual bootstrap LPR (rBLPR). In particular, we first show that, under the cliff-weak-sparsity and other reasonable conditions, the lasso has model selection consistency in the sense that it can correctly identify all the large components of β^0 while shrinking all the small ones to zeros; see Theorem 1. Second and more interestingly, we show in Theorem 2 that, under one more condition, the residual bootstrap lasso

estimator can also achieve the same kind of model selection consistency. Based on these properties, we finally provide 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 $u \in R^p$; 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.

In (Zhao & Yu, 2006; Wainwright, 2009), the authors 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, the beta-min condition and the hard sparsity.

Definition 2 (Zhao & Yu (2006))**.** If an estimator $\hat{\beta}$ is equal in sign with 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 1, negative entries to -1 and zero entries to zero.

In this paper, we will extend this result to the *cliff-weak-sparsity* case. Without loss of generality, we assume $\beta^0 = (\beta^0_1, \ldots, \beta^0_s, \beta^0_{s+1}, \ldots, \beta^0_p)$ with $\beta^0_j \gg n^{-1/2}$ for $j = 1, \ldots, s$ and $\beta^0_j \ll n^{-1/2}$ 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. We let $C = X^{\mathrm{T}}X/n$, which can be expressed in a 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 eigenvalue of a matrix A. To obtain model selection consistency, we require the following assumptions:

Condition 1. ϵ_i's are i.i.d. subgaussian random variables.

Condition 2. The predictors are standardized, i.e.,

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

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

Conditions 1 and 2 are fairly standard in sparse linear regression literature; see for example (Zhao & Yu, 2006; Huang et al., 2008; Huang, Ma & Zhang, 2008). Theorems 1, 2, 3 hold if we replace Condition 2 by a bounded second moment condition. However, to simplify our argument, we use Condition 2. Condition 3 ensures that the smallest eigenvalue of C_{11} is bounded away from 0 so 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 1 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 1. 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 $M > 0$ so that

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

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

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

Condition 4 implies both the number of larger components of β^0 (i.e., s) and the number of predictors (i.e., p) diverge with 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 work (Zhao & Yu, 2006). Condition 6 is the cliff-weak-sparsity assumption on β^0, and it 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 statement of Condition 6 on $\min_{1\le i\le s} |\beta_i^0|$ and the first half statement of Condition 7 on λ_1 are the same as those used in paper (Zhao & Yu, 2006) to show the sign consistency of lasso. Condition 8 is a technical assumption suggesting that the projection of small effects i.e., $X_{S^c}\beta^0_{S^c}$, onto the linear subspace spanned by the predictors corresponding to the large coefficients, i.e., the predictors in S, decays at a certain rate. In the Supplementary Material, we will present examples where this condition holds. Conditions 1 - 5 and 7 are also assumed in paper (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 1 at an exponential rate. This is a natural generalization of the sign consistency of the lasso from the hard sparsity to the cliff-weak-sparsity. We adopt the analytical techniques in paper (Zhao & Yu, 2006) with necessary modifications to account for the cliff-weak-sparsity. The proof details are provided in the Supplementary Material.

3.3 Weak convergence of the residual bootstrap lasso+partial ridge

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

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

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

Condition 9 is stronger than Condition 4 by requiring $0 < c_1 < 1/2$. Without considering model selection, Bickel & Freedman (1983) showed that residual bootstrap ols fails if p^2/n does not tend to 0. Therefore, Condition 9 cannot be easily weakened. 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 in 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, i.e. $|x_{ij}| \leq M,\ i = 1, \ldots, n,\ j = 1, \ldots, s$. This condition is also assumed in (Huang et al., 2008) to obtain asymptotic normality of the bridge estimator. The second part in 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 the 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 has sign consistency, i.e.,*

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

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 lasso+partial ridge (LPR) method will use the partial ridge regression to recover those small but nonzero ones.

With Theorems 1 and 2 and under the following Condition 11, we can show that the residual bootstrap LPR (rBLPR) procedure can consistently estimate

the distribution of $\hat{\beta}_{\text{LPR}}$ and thus construct asymptotically valid confidence intervals for 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 $u \in R^p$ 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

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

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, and thus the conditional distribution of ϵ_i^* is a valid approximation of the distribution of ϵ_i. The other two statements are a Linderberg type condition and a technical condition to obtain asymptotic normality.

Remark 5. For orthogonal design matrix, i.e., $(1/n)X^{\mathrm{T}}X = I$, which implies that there are no correlations between predictors and that $p \le n$, $\sigma_1^2 = \sigma^2$ and Condition 11 reduces to the following much simpler form: $\max_{1\le k\le n}|u^{\mathrm{T}}X_k| = o(\sqrt{n})$. When $u = e_j$, a basis vector whose jth element equals 1 and other elements equal 0, this condition is equivalent to $\max_{1\le k\le n}|x_{kj}| = o(\sqrt{n})$, which is not a strong condition that is expected to hold in many practical situation.

The conclusion is still true when the correlation between two covariates satisfies $\text{cor}(x_i, x_j) = \rho^{|i-j|}$ 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 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 by the bootstrap. Based on the estimated conditional distribution of $\sqrt{n}u^{\mathrm{T}}(\hat{\beta}^*_{\text{rBLPR}} - \hat{\beta}_{\text{lasso+ols}})$, we 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 could also show 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 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 can work as well as the rBLPR method. We leave the theoretical analysis of the pBLPR to future research.

4. Simulation studies

We carry out 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 coverage probabilities and confidence interval lengths. More details of the simulation studies and results are included in the Supplementary Material, but the main conclusions are summarized as follows.

(1) Setting $\lambda_2 = O(1/n)$ 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, while 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), while, rBLPR and pBLPR do not perform well when SNRs are low, i.e., no greater than 1. This is mainly because the lasso cannot correctly select all the important predictors. The rBLPR and pBLPR produce much better confidence intervals when SNRs are high, i.e., larger than 5: with comparable coverage probabilities, its interval lengths are 50% shorter than LDPE and JM on average.

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

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

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

(8) BLDPE has the best coverage probabilities among the considered methods. Its confidence interval lengths are close to the better ones of LDPE and JM, but they 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 (functional Magnetic Resonance Imaging) data set and compare its performance with the two de-sparsified methods. The fMRI data were provided by the Gallant Lab at UC Berkeley Kay et al. (2008). The fMRI recorded measurements of 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-data set 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, etc. Every picture was a 128 pixel by 128 pixel gray scale image, which was reduced to a vector of length 10921 through the following procedure: (1) using Gabor transform of the gray image to generate local contrast energy features Z_j, and (2) taking non-linear transformation $X_j = \log(1 + \sqrt{Z_j}), j = 1, \ldots, 10921$. 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. And the validation data set contained responses to 120 natural images (We will not use the validation data in this paper).

After reading the training data set into R, we calculate the variance of each feature (column) in X and delete those columns whose variances are $\leq 1e^{-4}$.

Then we have a matrix of dimension 1750×9076. We further reduce the dimension of the design matrix by correlation screening, i.e., sorting the correlations (Pearson correlation between every feature in X and the response Y) in an decreasing order of absolute values and selecting the top 500 features with the largest correlation. We use the lasso+ols estimate of 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 rows with size $n = 200$ 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 LDPE and JM.

Based on the sub-data set 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% of the 500 components of β^0, respectively. All the 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 the other two methods do for most of the coefficients, especially the small ones. As shown in Figure S15, we illustrate the confidence interval lengths of 100 coefficients (44

nonzero coefficients in β^0 and 56 randomly chosen zero coefficients) produced by the three methods. Seeing the satisfactory coverage and much shorter lengths of the confidence intervals produced by pBLPR, we demonstrate that pBLPR has overall better performance than LDPE and JM 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, as well as the three de-sparsified lasso methods LDPE, and BLDPE, to a data set contained $43,827$ gene expression measurements from Illumina RNA sequencing of 498 neuroblastoma samples. More details about this data set could be found in the Supplemental Material.

Constructing gene-gene regulatory relationships are of primary interests 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 particular gene named *CAMTA1* that is known to be neuroblastoma-related and is observed to be highly correlated with the risk of neuroblastoma. Given our knowledge on the complex regulatory relationships among genes, the linear model is almost certainly a misspecified model. However, this case study would serve as a reasonable real data example to demonstrate the capacity of our pBLPR and rBLPR methods and the three de-sparsified lasso methods

LDPE, JM and BLDPE in identifying 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 only finds 1 significant genes. The functions related to natural and regulated cell deaths (e.g., apoptosis and autophagy), which are the key processes to prevent cancer occurrences, are only enriched in the significant genes found by pBLPR or rBLPR but not by any de-sparisied lasso methods. On the other hand, only general functions, such as basic processes in cells, are enriched in the significant genes found only by a de-sparsified lasso method but not by our methods. This suggests that pBLPR and rBLPR find 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 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 useful for tackling this problem. However, previous work that extended bootstrap technique to high-dimensional models

rely on two key assumptions: hard sparsity and the beta-min condition. The beta-min condition is rather restrictive and in order to relax it, we propose two new bootstrap procedures based on a new two-stage estimator called lasso+partial ridge. Our methods dramatically improve the performance of the bootstrap lasso+ols method proposed in (Liu & Yu, 2013) in the scenarios 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 give comparable coverage probabilities but shorter (on average) intervals and are robust to misspecified models than the other methods under many scenarios. We apply our methods to an fMRI data set and find that it gives reasonable coverage probabilities and shorter interval lengths than LDPE, JM and BLDPE. In another real data application, we apply our methods to find genes that have significant effects on predicting a cancer gene's expression levels in a likely misspecified linear model. Compared with the three de-sparsified lasso methods, our methods find genes that are more biologically reasonable and interpretable, suggesting that our methods can be 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 specification, from both theoretical and empirical perspectives.

A disadvantage of our methods is that the resulting inference is not uniformly valid over the class of sparse models, due to the cliff-weak-sparsity assumption. It is possible that our methods are uniformly valid for some 'pseudo-true' parameters, i.e., parameters in the nearest model that satisfies the cliff-weak-sparsity, and we leave this direction to future work. Moreover, compared with the uniformly valid inference procedures such as the de-sparsified lasso methods, we observe from our empirical studies that our methods are more likely to identify small but nonzero coefficients due to the shorter confidence interval lengths returned by our methods. In many real-world applications, the covariates (or features) with small effects are not negligible but may be important. For example, in genomic applications where complex gene-gene regulatory relationships are of primary interests, researchers searching for regulators of a target gene are not only interested in the genes with large effects but also the other genes with small effects, 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, but we note that subsequent experiments are still required to validate the identified genes. Besides, in the case where 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, while for small but nonzero coefficients, the bootstrap LPR methods (rBLPR and pBLPR) have the shortest confidence interval lengths with good coverage probabilities. Therefore, if practitioners are only concerned with the confidence intervals of large coefficients, we recommend the bootstrap lasso+ols method; if practitioners are also interested in identifying small but significant coefficients in a possibly misspecified linear model, we recommend our bootstrap LPR methods but meanwhile warn practitioners that the confidence intervals of the coefficients, whose magnitudes are of the order $1/\sqrt{n}$, may be invalid; otherwise, 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 for implementation and parallelization.

Finally, multiple testing is another important task in hypothesis testing, which is closely related to confidence interval construction. Several procedures such as the Bonferroni correction, the Benjamini-Hochberg procedure and the FDR control have been proposed to correct multiple testing in low-dimensional settings. However, these procedures are based on accurate estimation of p-values of each

single test, where small p-values can only be obtained by 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 to future work.

Supplementary Material

Supplementary Material includes proofs, algorithms, and simulation results. Another Supplementary File contains the detailed Gene Ontology analysis results of real data application 2.

Acknowledgements

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, NIH/NIGMS grant R01GM120507, and NSF grant NSF grant DMS-1613338.

References

BARBER R. F. & CANDÈS, E. J. (2015). Controlling the false discovery rate via knockoffs. *Ann. Statist.* **43**, 2055–2085.

BELLONI A., CHERNOZHUKOV V., & 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. & ZHAO L. (2013). Valid post-selection inference. *Ann. Statist.* **41**, 802–837.

BICKEL, P. J. & 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. & VAN DE GEER, S. (2011). *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Springer.

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

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

CHATTERJEE, A. & LAHIRI, S. N. (2011). Bootstrapping lasso estimators. *J. Am. Statist. Assoc.* **106**,

608–625.

CHATTERJEE, A. & 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. & 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. & MEINSHAUSEN, N. (2014). High-dimensional Inference: Confidence intervals, *p*-values and R-Software hdi. *Stat. Sci.* **30**, 533–558.

DEZEURE, R., BÜHLMANN, P. & ZHANG, C-H. (2017). High-dimensional simultaneous inference with the bootstrap. *Test* **26**, 685–719.

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

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

GAO, X., AHMED, S. E. & FENG, Y. (2017). Post selection shrinkage estimation for high-dimensional data analysis. *Appl. Stoch. Model Bus.* **33**, 97–120.

HOERL, A. E. & KENNARD, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics* **12**, 55–67.

HUANG, J., HOROWITZ, J. & MA, S. (2008). Asymptotic properties of bridge estimators in sparse high-dimensional regression models. *Ann. Statist.* **36**, 587–613.

HUANG, J., MA, S. & ZHANG C.-H. (2008). Adaptive lasso for sparse high-dimensional regression models. *Stat. Sinica* **18**, 1603–1618.

JAVANMARD, A. & 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. & GALLANT, J. L. (2008). Identifying natural images from human brain activity. *Nature* **452**, 352–355.

KNIGHT, K. & 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.

LOCKHART, R., TAYLOR, J., TIBSHIRANI, R. J. & 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. & 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. & BÜHLMANN, P. (2006). High dimensional graphs and variable selection with the lasso. *Ann. Statist.* **34**, 1436–1462.

MEINSHAUSEN, N. & 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. & 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. & YU, B. (2009). A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers. *Stat. Sci.* **28**, 538–557.

SHI, P. & 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. & DEZEURE, R. (2014). On asymptotically optimal confidence 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. & 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. & 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