Electronic Journal of Statistics
Vol. 7 (2013) 3124–3169
ISSN: 1935-7524
DOI: 10.1214/14-EJS875

Asymptotic properties of Lasso+mLS and Lasso+Ridge in sparse high-dimensional linear regression

Hanzhong Liu

School of Mathematical Sciences
Peking University, Beijing 100871, P.R. China
e-mail: lhz2009@pku.edu.cn

and

Bin Yu

Department of Statistics
Department of Electrical Engineering and Computer Sciences
University of California, Berkeley, CA 94720, USA
e-mail: binyu@stat.berkeley.edu

Abstract: We study the asymptotic properties of Lasso+mLS and Lasso+Ridge under the sparse high-dimensional linear regression model: Lasso selecting predictors and then modified Least Squares (mLS) or Ridge estimating their coefficients. First, we propose a valid inference procedure for parameter estimation based on parametric residual bootstrap after Lasso+mLS and Lasso+Ridge. Second, we derive the asymptotic unbiasedness of Lasso+mLS and Lasso+Ridge. More specifically, we show that their biases decay at an exponential rate and they can achieve the oracle convergence rate of s/n (where s is the number of nonzero regression coefficients and n is the sample size) for mean squared error (MSE). Third, we show that Lasso+mLS and Lasso+Ridge are asymptotically normal. They have an oracle property in the sense that they can select the true predictors with probability converging to 1 and the estimates of nonzero parameters have the same asymptotic normal distribution that they would have if the zero parameters were known in advance. In fact, our analysis is not limited to adopting Lasso in the selection stage, but is applicable to any other model selection criteria with exponentially decay rates of the probability of selecting wrong models.

MSC 2010 subject classifications: Primary 62F12, 62F40; secondary 62J07.
Keywords and phrases: Lasso, irrepresentable condition, Lasso+mLS and Lasso+Ridge, sparsity, asymptotic unbiasedness, asymptotic normality, residual bootstrap.

Received June 2013.

1. Introduction

Consider the sparse linear regression model

$$Y = X\beta^* + \epsilon, \tag{1}$$

where $\epsilon = (\epsilon_1, \ldots, \epsilon_n)^T$ is a vector of independent and identically distributed (i.i.d.) random variables with mean 0 and variance σ^2. $Y = (y_1, \ldots, y_n)^T \in \mathbb{R}^n$ is the response vector, and $X \in \mathbb{R}^{n \times p}$ is the design matrix which is deterministic. $\beta^* \in \mathbb{R}^p$ is the vector of model coefficients with at most s ($s < n$) non-zero components. We consider the high-dimensional setting which allows p and s to grow with n (p can be comparable to or larger than n). Note that, in here and what follows, Y, X, and β^* are all indexed by the sample size n, but we omit the index whenever this does not cause confusion.

In sparse linear regression models, an active line of research focuses on the recovery of sparse vector β^* by a popular l_1 regularization method called Lasso [47]. The Lasso has been studied under at least three common criteria: (i) model selection criteria, meaning the correct recovery of the support set $S = \{j \in \{1, 2, \ldots, p\} : \beta_j^* \neq 0\}$ of the model coefficients β^*; (ii) l_q estimation errors $||\hat{\beta} - \beta^*||_q^q$, especially l_2 and l_1, where $\hat{\beta}$ is the estimate of β^*; and (iii) prediction error $||X\hat{\beta} - X\beta||_2^2$.

The Lasso estimator is defined by

$$\hat{\beta}(\lambda_n) = \underset{\beta}{argmin} \left\{ ||Y - X\beta||_2^2 + \lambda_n ||\beta||_1 \right\}, \tag{2}$$

where $\lambda_n \geq 0$ is the tuning parameter which controls the amount of regularization applied to the estimate. Setting $\lambda_n = 0$ reverses the Lasso problem to Ordinary Least Squares (OLS) which minimizes the unregularized empirical loss.

Replacing ℓ_1 penalty by ℓ_2 penalty in (2) gives the Ridge estimator [28]:

$$\hat{\beta}_{Ridge}(\lambda_n) = \underset{\beta}{argmin} \left\{ ||Y - X\beta||_2^2 + \lambda_n ||\beta||_2^2 \right\}, \tag{3}$$

The Lasso estimator has two nice properties, namely, (i) it generates sparse models by means of ℓ_1 regularization and (ii) it is also computationally feasible (see [43, 19, 24]). The asymptotic behavior of Lasso-type estimators has been studied by [32] for fixed p and β^* as $n \to \infty$. In particular, they have shown that under some regularity conditions on the design, $\lambda_n = o(n)$ is sufficient for consistency in the sense that $\hat{\beta}(\lambda_n) \to_p \beta^*$, and λ_n should grow more slowly (i.e. $\lambda_n = O(\sqrt{n})$) for asymptotic normality of the Lasso estimator. On the model selection consistency front, [36] proposed the neighborhood stability condition which is equivalent to the Irrepresentable condition [25, 48, 54, 51] to prove the Lasso consistency for Gaussian graphical model selection. [54] showed that the Irrepresentable condition is almost necessary (for fixed p) and sufficient for the Lasso to select the true model both in the classical fixed p setting and in the high-dimensional setting. [52] considered a weaker sparsity assumption, meaning that the regression coefficients outside an ideal model are small but not necessarily zero (the sum of their absolute values is of the order $O(s\lambda_n/n)$), and

imposed the sparse Riesz condition to prove the rate-consistent in terms of the sparsity, bias and the norm of missing large coefficients. [51] further established precise conditions on the scalings of (n, p, s) that are necessary and sufficient for sparsity pattern recovery using the Lasso. In addition, thresholded Lasso and Dantzig estimators were introduced in [31] and the authors proved their model selection consistency under less restrictive conditions on the decay rates of the nonzero regression coefficients. Other related work includes [17, 55, 10, 1]. All the aforementioned papers imposed suitable mutual incoherence conditions on the design. On the l_2 estimation error front, the Lasso has been shown under a weaker restricted eigenvalue condition to achieve l_2 convergence rate of $(s \log p)/n$ [49, 9, 39, 41, 42], which is the minimax optimal rate [45]. Other work focuses on the convergence rates of $||X\hat{\beta}(\lambda_n) - X\beta^*||_2^2$ and $||\hat{\beta}(\lambda_n) - \beta^*||_1$, see [27, 50, 11] for example.

However, even if p is fixed and the Irrepresentable Condition is satisfied, there does not exist a tuning parameter λ_n which can lead to both variable selection consistency and asymptotic normality [21, 55]. More importantly, for the case of $p \gg n$, statistical inference for the Lasso estimator with theoretical guarantees is still an insufficiently explored area.

The bootstrap is very useful for inference. For fixed p, [40] developed a perturbation resampling-based method to approximate the distribution of a general class of penalized regression estimates. [13] proposed a modified residual bootstrapping Lasso method that is consistent in estimating the limiting distribution of the Lasso estimator. [53] developed a low-dimensional projection (LDP) approach to constructing confidence intervals. Though LDP works for $(s \log p)/\sqrt{n} \to 0$, it has nothing to do with the idea of resampling and bootstrap. Does the bootstrap provide a valid approximation in the case of $p \gg n$? In this paper, we will give an affirmative answer to this question based on residual bootstrap after two post-Lasso estimators: Lasso+mLS and Lasso+Ridge. Our method provides consistent estimate of the limiting distribution of Lasso+mLS (or Lasso+Ridge) even if p grows at an exponential rate in n.

Post-Lasso estimator is a special case of two stage estimators: (1) selection stage: one selects predictors using the Lasso; and (2) estimation stage: modified Least Square (mLS) or Ridge, is applied to estimate the coefficients of the selected predictors. Our estimator is referred to as Lasso+mLS or Lasso+Ridge. Lasso+mLS is very close to Lasso+OLS [3], which uses Ordinary Least Squares (OLS) in the second stage. Several authors have previously considered two stage estimators to improve the performance of the Lasso, such as the Lars-OLS hybrid [19], adaptive Lasso [55], relaxed Lasso [37], and marginal bridge estimator [29], to name just a few.

Our contributions are summarized as follows:

1. We propose a valid inference procedure for parameter estimation based on parametric residual bootstrap after two post-Lasso estimators: Lasso+mLS and Lasso+Ridge. More specifically, we show that the Mallows distance between the distributions of the bootstrap estimator and the Lasso+mLS (or Lasso+Ridge) estimator converges to 0 in probability.

2. Under the Irrepresentable condition and other regularity conditions, we derive the asymptotic unbiasedness of Lasso+mLS and Lasso+Ridge. We show that their biases decay at an exponential rate and that they can achieve the oracle convergence rate of s/n for mean squared error $E||\tilde{\beta} - \beta^*||_2^2$ where $\tilde{\beta}$ is either Lasso+mLS or Lasso+Ridge.
3. We prove the asymptotic normality of Lasso+mLS and Lasso+Ridge. As we show in Theorem 3 and Corollary 2, these two post-Lasso estimators display an oracle property that the Lasso does not have: they can select the true predictors with probability converging to 1 and the estimates of nonzero parameters have the same asymptotic normal distribution that they would have if the zero parameters were known in advance.
4. Our analysis is not limited to adopting the Lasso in the selection stage, but is applicable to any other model selection criteria with exponentially decay rates of the probability of selecting wrong models, for example, stability selection [38], SCAD [21, 34] and Dantzig selector [12, 9, 26].

Our key assumptions for the validity of residual bootstrap after Lasso+mLS or Lasso+Ridge are the Irrepresentable condition and that s goes to infinity slower than $\sqrt{n}$. The Irrepresentable condition can be weakened by the sparse eigenvalue condition if we adopt stability selection [38] to enhance the selection performance of the Lasso. Without considering model selection, [8] showed that residual bootstrap OLS fails if p^2/n does not tend to 0. Therefore, the condition $s^2/n \to 0$ cannot be weakened. Our conditions on the scalings (n, p, s) are not the sharpest but have been previously used in the literature [54] and make our convergence rate more explicit. In addition, we require a gap of size $n^{\frac{c_3}{2}}$ ($c_3 \in (0, 1]$ is a constant) between the decay rate of β^* and $n^{-\frac{1}{2}}$ which prevents the estimation from being dominated by the noise terms. [22] proposed a similar constraint $min_{1\le i\le s}|\beta_i^*| \ge \frac{M}{n^\kappa}$, $0 \le \kappa < \frac{1}{2}$ to show model selection consistency of the Sure Independent Screening. This assumption is weaker than $min_{1\le i\le s}|\beta_i^*| \ge M$, which was assumed by [29] in connection with the asymptotic properties of the Bridge estimator. However, as mentioned by [33], inference results based on post-model selection methods can be misleading when this kind of "beta min" condition fails. Therefore, the proposed inference procedure should be used in practice only when there is believed to be a gap between the decay rate of the nonzero elements of β^* and the $n^{-1/2}$. Finally, we need some regularity conditions (conditions (a)–(c) in Section 2) which are standard in sparse high-dimensional linear regression literature [54, 29, 30].

After we had obtained our bootstrap results in Theorem 4 and Corollary 3, our attention was brought to an independent result in [14] where a variant of the Irrepresentable condition was used to prove the second-order correctness of the residual bootstrap applied to a suitable studentized version of the adaptive Lasso estimator. However, the main results of [14] are valid only for linear combinations of the adaptive Lasso estimator while our results hold for the joint distribution of Lasso+mLS (or Lasso+Ridge). The distance between distributions used in [14] and the proof there are also different from ours. Specifically, [14] adopted the total variation distance and used the Edgeworth expansion in the proof

while we study the Mallows distance and our proof is direct. In addition, [14] allows p to grow only at polynomial rates in n while we allow p to grow at an exponential rate in n.

In our paper, before stating the bootstrap result, we first derive the asymptotic unbiasedness and asymptotic normality of Lasso+mLS and Lasso+Ridge. It is well known that $(s \log p)/n$ is the minimax optimal rate for the Lasso under the restricted eigenvalue condition. Since we assume the stronger Irrepresentable condition and some conditions on scaling (n, p, s), we are able to attain a better rate of s/n for MSE which indicates that one can avoid the feature selection penalty of $\log p$ by combining the Lasso and Least Squares or Ridge. We should mention that previous work [3] has obtained l_2 convergence rate ($||\tilde{\beta}_{Lasso+OLS} - \beta^*||_2^2 = O_p(s/n)$) of Lasso+OLS estimator under weaker conditions. However, their results hold in probability and it is not clear whether Lasso+OLS can achieve the oracle convergence rate of $O(s/n)$ in L_2-expectation, i.e., whether $E||\tilde{\beta} - \beta^*||_2^2 = O(s/n)$ holds, which we need to prove the validity of residual bootstrap. On the asymptotic normality front, the authors in [4, 5] also adopted the OLS after model selection and derived the asymptotic normality for inference on the effect of a treatment variable on a scalar outcome in the presence of very many controls. However, they studied a partial linear model which is different from ours and the l_1 regularization was imposed on the effect of the control variables without on the effect of the treatment variable.

Notation. For any vector $a = (a_1, \ldots, a_m)^T$, we denote $||a||_2^2 = \sum_{i=1}^m a_i^2$, $||a||_1 = \sum_{i=1}^m |a_i|$, and $\|a\|_\infty = \max_{i=1,\ldots,m} |a_i|$. For a vector $\beta \in R^p$ and a set $S \subset \{1, \ldots, p\}$, denote S^c the complementary set of S and let $\beta_S = \{\beta_j : j \in S\}$. Given an n by p matrix X, write $x_i^T \in R^p, i = 1, \ldots, n$ and $X_j \in R^n, j = 1, \ldots, p$ the i-th row and the j-th column of X respectively, where x_i^T is the transpose of x_i. For a given $m \times m$ matrix A, let $\Lambda_{min}(A)$ and $\Lambda_{max}(A)$ denote the smallest and largest eigenvalues of A respectively. Write $tr(A)$ the trace of A which is the sum of the diagonal entries of A.

The rest of the paper is organized as follows: in Section 2, we define modified Least Squares and Ridge after model selection and study their asymptotic properties. In Section 3, we apply these general properties to the special cases of Lasso+mLS and Lasso+Ridge and then derive their asymptotic unbiasedness, asymptotic normality and the approximation property of residual bootstrap. Similar asymptotic properties of modified Least Squares and Ridge after stability selection are obtained in Section 4. Simulation examples are given in Section 5. We conclude in Section 6. The proofs can be found in the Appendix.

2. Asymptotic properties of modified least squares and ridge after model selection

In this section, we begin with a precise definition of the modified Least Squares or Ridge after model selection, and then study their asymptotic properties, including asymptotic unbiasedness, asymptotic normality and the validity of residual bootstrap.

2.1. Definitions and assumptions

Modified Least Squares (or Ridge) after model selection refers to a special type of two stage estimators. In the first stage, one uses certain model selection methods to select predictors. For example, let $\hat{\beta}$ be the Lasso estimator defined in (2), one gets a set of selected predictors $\hat{S} = \{j \in \{1, 2, \ldots, p\} : \ \hat{\beta}_j \neq 0\}$. Again, $\hat{\beta}$ and $\hat{S}$ are dependent on λ_n, but we omit the dependence whenever this does not cause confusion.

In the second stage, a low-dimensional estimation method is applied to the selected predictors in $\hat{S}$. For example, one can adopt OLS and then form the OLS after model selection (denoted by Select+OLS):

$$\tilde{\beta}_{Select+OLS} = \underset{\beta:\ \beta_j=0,\ j\in\hat{S}^c}{argmin} ||Y - X\beta||_2^2. \tag{4}$$

The solution of (4) is $\tilde{\beta}_{Select+OLS,\hat{S}} = (X_{\hat{S}}^T X_{\hat{S}})^{-1} X_{\hat{S}}^T Y$ if $X_{\hat{S}}^T X_{\hat{S}}$ is invertible. When $X_{\hat{S}}^T X_{\hat{S}}$ is not invertible, the solution of (4) is not unique. In this case one can use the generalized inverse. However, the generalized inverse is not stable when the smallest nonzero eigenvalue of $X_{\hat{S}}^T X_{\hat{S}}$ approximately equals 0, which may result in poor performance. We propose a modified Least Squares method in the second stage and form our Select+mLS estimator. Let $d = |\hat{S}|$ and write $\frac{1}{\sqrt{n}} X_{\hat{S}}$ in its singular value decomposition (SVD) form

$$\frac{1}{\sqrt{n}} X_{\hat{S}} = UDV^T \tag{5}$$

where U is an $n \times n$ orthogonal matrix, D is an $n \times d$ diagonal matrix with singular values $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d$ on the diagonal, and V^T (the transpose of V) is a $d \times d$ orthogonal matrix. By simple algebraic operations, OLS based on $(X_{\hat{S}}, Y)$ has the following form:

$$\tilde{\beta}_{OLS} = \frac{1}{\sqrt{n}} V D^{-1} U^T Y \tag{6}$$

where D^{-1} is a $d \times n$ diagonal matrix with diagonal entries λ_1^{-1}, $\lambda_2^{-1}, \ldots, \lambda_d^{-1}$. If one or more of the singular values are 0, one can utilize generalized inverse which just takes $\lambda_k^{-1} = 0$ for all zero-valued λ_k.

We propose a hard thresholding on the singular values, that is, shrinking those singular values smaller than τ_n ($\tau_n > 0$) to zero. Then define a modified Least Square estimator $\tilde{\beta}_{mLS}(\tau_n)$ in the same form of (6) except that we take $\lambda_k^{-1} = 0$ for all $\lambda_k < \tau_n$. This estimator is similar to principal components regression [35]. Note that if $0 \leq \tau_n^2 \leq \Lambda_{min}(\frac{1}{n} X_{\hat{S}}^T X_{\hat{S}})$, $\tilde{\beta}_{mLS}(\tau_n)$ is the same as $\tilde{\beta}_{OLS}$, that is,

$$\tilde{\beta}_{mLS}(\tau_n) = \tilde{\beta}_{OLS} = (X_{\hat{S}}^T X_{\hat{S}})^{-1} X_{\hat{S}}^T Y, \quad if \quad 0 \leq \tau_n^2 \leq \Lambda_{min}\left(\frac{1}{n} X_{\hat{S}}^T X_{\hat{S}}\right).$$

Our final modified Least Squares after model selection (Select+mLS) is defined by:

$$\tilde{\beta}_{Select+mLS,\hat{S}}(\tau_n) = \tilde{\beta}_{mLS}(\tau_n)\ , \quad \tilde{\beta}_{Select+mLS,\hat{S}^c}(\tau_n) = 0. \tag{7}$$

One can also use Ridge in the second stage and form the Ridge after model selection (Select+Ridge):

$$\tilde{\beta}_{Select+Ridge}(\mu_n) = \underset{\beta:\ \beta_j=0,\ j\in\hat{S}^c}{argmin} ||Y - X\beta||_2^2 + \mu_n||\beta||_2^2 \tag{8}$$

where $\mu_n \geq 0$ is a smoothing parameter. (8) is equivalent to

$$\tilde{\beta}_{Select+Ridge,\hat{S}}(\mu_n) = (X_{\hat{S}}^T X_{\hat{S}} + \mu_n I)^{-1} X_{\hat{S}}^T Y, \quad \tilde{\beta}_{Select+Ridge,\hat{S}^c}(\mu_n) = 0. \tag{9}$$

When the Lasso is used in the selection stage, we refer to our final estimators as Lasso+mLS and Lasso+Ridge respectively. τ_n and μ_n are tuning parameters. In our theorems and simulation, $\tau_n \propto \frac{1}{n}$ and $\mu_n \propto \frac{1}{n}$ can get good estimation and prediction performance. For the sake of notational simplicity, we omit the dependence of estimators on λ_n and τ_n or μ_n whenever this does not cause confusion.

To state our main theorems, we need the following assumptions. Without loss of generality, assume $\beta^* = (\beta_1^*, \ldots, \beta_s^*, \beta_{s+1}^*, \ldots, \beta_p^*)$ with $\beta_j^* \neq 0$ for $j = 1, \ldots, s$ and $\beta_j^* = 0$ for $j = s+1, \ldots, p$. Let $S = \{1, \ldots, s\}$ and $\beta_S^* = (\beta_1^*, \ldots, \beta_s^*)$. Now write X_S and X_{S^c} as the first s and the last $p-s$ columns of X respectively and let $C = \frac{1}{n}X^T X$ which can be expressed in a block-wise form as follows:

$$C = \begin{pmatrix} C_{11} & C_{12} \\ C_{21} & C_{22} \end{pmatrix} \tag{10}$$

where $C_{11} = \frac{1}{n}X_S^T X_S$, $C_{12} = \frac{1}{n}X_S^T X_{S^c}$, $C_{21} = \frac{1}{n}X_{S^c}^T X_S$ and $C_{22} = \frac{1}{n}X_{S^c}^T X_{S^c}$.

Assumption (a). *ϵ_i are i.i.d. gaussian random variables with mean 0 and variance σ^2.*

Assumption (b). *Suppose that the predictors are standardized, i.e.*

$$\frac{1}{n}\sum_{i=1}^n x_{ij} = 0 \text{ and } \frac{1}{n}\sum_{i=1}^n x_{ij}^2 = 1,\ j = 1, \ldots, p. \tag{11}$$

Assumption (c). *There exists an constant $\Lambda_{min} > 0$ such that*

$$\Lambda_{min}(C_{11}) \geq \Lambda_{min}. \tag{12}$$

Assumption (d). *The model is high-dimensional and sparse, i.e. there exists $0 \leq c_1 < 1$ and $0 < c_2 < 1 - c_1$ such that*

$$s = s_n = O(n^{c_1}),\ p = p_n = O(e^{n^{c_2}}). \tag{13}$$

Assumption (e).[1] *$\tau_n \propto \frac{1}{n}$ and $\mu_n \propto \frac{1}{n}$.*

[1] In fact, Theorem 3 (asymptotic normality) and Theorem 4 (bootstrap) are valid for any $\mu_n \to 0$ with rate neither faster than $e^{-n^{c_2}/4}$ nor slower than $\frac{1}{n}$.

The gaussian assumption (a) is fairly standard in the literature. Assumption (c) ensures the smallest eigenvalue of C_{11} is bounded away from 0 so that its inverse behaves well. (a)–(c) are typical assumptions in sparse linear regression literature, see for example [54, 29, 30]. Assumption (d) means that the number of relevant predictors s is allowed to diverge but much slower than n, and that the number of predictors p can grow faster than n (up to exponentially fast), which is standard in almost all high-dimensional inference literature. Though this assumption is stronger than the typical one $\frac{s\log p}{n}\to 0$, it has been used previously [54].

In the following subsections, we will show that both Select+mLS and Select+Ridge have good asymptotic properties if the probability of selecting wrong models $P(\hat{S}\neq S)$ decays fast, say $o(e^{-n^{\kappa}})$ (where $\kappa>0$ is a constant).

2.2. Bias and MSE of Select+mLS and Select+Ridge

In a high-dimensional setting, bias is not the only consideration of estimates because of the bias and variance trade-off. Regularization has been a popular technique for model fitting which results in a biased estimator but decreases the mean squared error (MSE) dramatically. However, if two estimators have the same MSE, we prefer the unbiased one. The following Theorems 1 and 2 provide general bounds for the bias and MSE of the Select+mLS and Select+Ridge.

Theorem 1. *Suppose that Gaussian assumption (a) is satisfied and* $\tau_n^2\le \Lambda_{min}(C_{11})$*, then the bias and the MSE of* $\tilde{\beta}_{Select+mLS}(\tau_n)$ *satisfy*

$$\begin{aligned} &||E\tilde{\beta}_{Select+mLS}(\tau_n)-\beta^*||_2^2 \\ \le\ & 2P(\hat{S}\neq S)\left\{\frac{\sigma^2}{n}tr(C_{11}^{-1})+||\beta^*||_2^2+\frac{1}{\tau_n^2}\frac{1}{n}||X\beta^*||_2^2+\frac{1}{\tau_n^2}\sigma^2\right\}, \qquad (14)\\ &E||\tilde{\beta}_{Select+mLS}(\tau_n)-\beta^*||_2^2 \\ \le\ & \frac{\sigma^2}{n}tr(C_{11}^{-1})+8\sqrt{P(\hat{S}\neq S)}\left\{||\beta^*||_2^2+\frac{1}{\tau_n^2}\frac{1}{n}||X\beta^*||_2^2+\frac{1}{\tau_n^2}\sigma^2\right\}. \qquad (15)\end{aligned}$$

Remark 2.1. Let $\beta_S^{OLS}=(X_S^TX_S)^{-1}X_S^TY=\beta_S^*+(X_S^TX_S)^{-1}X_S^T\epsilon$ be the oracle OLS estimator. It is easy to see that $E||\beta_S^{OLS}-\beta^*||_2^2=\frac{\sigma^2}{n}tr(C_{11}^{-1})$. The first term on the right hand side of (15) corresponds to the oracle convergence rate. The second term is related to model selection accuracy. In the case of the Lasso, $P(\hat{S}\neq S)$ can decay at an exponential rate. Hence the MSE is completely determined by the first term $\frac{\sigma^2}{n}tr(C_{11}^{-1})$, which can not be improved.

Remark 2.2. From theorem 1, one can easily get an upper bound for prediction mean squared error $E\{\frac{1}{n}||X\tilde{\beta}-X\beta^*||_2^2\}$ since

$$\frac{1}{n}||X\tilde{\beta}-X\beta^*||_2^2\le\Lambda_{max}\left(\frac{1}{n}X^TX\right)||\tilde{\beta}-\beta^*||_2^2.$$

Similarly, for $\tilde{\beta}_{Select+Ridge}(\mu_n)$, we have:

Theorem 2. *Suppose that Gaussian assumption (a) is satisfied, then the bias and the MSE of* $\tilde{\beta}_{Select+Ridge}(\mu_n)$ *satisfy*

$$\begin{aligned} & ||E\tilde{\beta}_{Select+Ridge}(\mu_n) - \beta^*||_2^2 \\ \leq \; & \frac{2\mu_n^2}{n^2\Lambda_{min}^2}||\beta^*||_2^2 + 2P(\hat{S} \neq S)\frac{n}{\mu_n}\left\{\frac{1}{n}||X\beta^*||_2^2 + \sigma^2\right\}, \qquad (16) \\ & E||\tilde{\beta}_{Select+Ridge}(\mu_n) - \beta^*||_2^2 \\ \leq \; & \frac{\sigma^2}{n}tr\left\{\left(C_{11} + \frac{\mu_n}{n}I\right)^{-2}C_{11}\right\} + \frac{\mu_n^2}{n^2\Lambda_{min}^2}||\beta^*||_2^2 \\ & + 2\sqrt{P(\hat{S} \neq S)}\left\{||\beta^*||_2^2 + \frac{n}{\mu_n}\left\{\frac{1}{n}||X\beta^*||_2^2 + \sigma^2\right\}\right\}. \qquad (17) \end{aligned}$$

Theorems 1 and 2 indicate that as long as the probability of selecting wrong models $P(\hat{S} \neq S)$ decays fast, say at an exponential rate, Select+mLS and Select+Ridge are asymptotically unbiased and their MSEs decay at the oracle rate. We have known that under the Irrepresentable condition and other regularity conditions, the probability of the Lasso selecting wrong models satisfies $P(\hat{S} \neq S) = o(e^{-n^{c_2}})$ [54]. Then applying the above theorems to Lasso+mLS and Lasso+Ridge as special cases, we can easily derive their convergence rates of bias and MSE, see Section 3 for more details.

2.3. Asymptotic normality of Select+mLS and Select+Ridge

In this section, we show asymptotic normality of Select+mLS and Select+Ridge. Let $\hat{\Psi}$ and Ψ be the distribution functions of $\sqrt{n}(\tilde{\beta}_S - \beta_S^*)$ and $N(0, \sigma^2 C_{11}^{-1})$ respectively, where $\tilde{\beta}$ can be any one of the two post model selection estimators: Select+mLS and Select+Ridge. Let $\hat{S}$ be the selected predictor set,

Theorem 3. *Suppose that assumptions (a)–(e) are satisfied and that the model selection procedure is consistent, i.e.,* $P(\hat{S} \neq S) = o(1)$*, then Select+mLS and Select+Ridge are asymptotically normal*[2]*, that is,*

$$\sup_{t \in R^s} |\hat{\Psi}(t) - \Psi(t)| \to 0, \; as \; n \to \infty. \qquad (18)$$

This theorem states that model selection consistency in the first stage implies the asymptotic normality of the second stage estimators: Select+mLS and Select+Ridge. The proof of this theorem can be found in Appendix C.

2.4. Residual bootstrap after Select+mLS and Select+Ridge

To make reliable scientific discoveries, we need to establish valid inference procedures including constructing confidence regions and testing for the parameter

[2]For Select+Ridge, we need assumption (g) proposed in the next subsection to hold. Due to the restricted space, we don't state it separately.

estimation. Although we have derived the asymptotic normality of Select+mLS and Select+Ridge, it is difficult to use in practice because the noise variance σ^2 is not known and hard to estimate in a high-dimensional setting. The bootstrap is a popular alternative in this case. A summary of bootstrap methods in linear and generalized linear penalized regression models for fixed p can be found in [46]. We will consider the $p \gg n$ case by proposing a new inference procedure: residual bootstrap after two stage estimators. Our method allows p to grow at an exponential rate in n.

In the context of a regression model, residual bootstrap is a standard method to bootstrap when the design matrix X is deterministic [18, 23, 32]. Let $\tilde{\beta}$ denote Select+mLS or Select+Ridge, the residual vector is given by:

$$\hat{\epsilon} = (\hat{\epsilon}_1, \ldots, \hat{\epsilon}_n)^T = Y - X\tilde{\beta}. \tag{19}$$

Consider the centered residuals at the mean $\{\hat{\epsilon}_i - \hat{\mu},\ i = 1, \ldots, n\}$, where $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} \hat{\epsilon}_i$. For residual bootstrap, one obtains $\epsilon^* = (\epsilon_1^*, \ldots, \epsilon_n^*)^T$ by resampling with replacement from the centered residuals $\{\hat{\epsilon}_i - \hat{\mu},\ i = 1, \ldots, n\}$, and formulates Y^* as follows

$$Y^* = X\tilde{\beta} + \epsilon^*. \tag{20}$$

Then one can define the selected predictor set $\hat{S}^*$ and Select+mLS or Select+Ridge $\tilde{\beta}^*$ based on the bootstrap sample (X, Y^*). For the bootstrap to be valid, one needs to verify that the conditional distribution of $T_n^* = \sqrt{n}(\tilde{\beta}^* - \tilde{\beta})$ given ϵ, which can be computed directly from the data, approximates the distribution of $T_n = \sqrt{n}(\tilde{\beta} - \beta^*)$. The difference between two distributions can be characterized by Mallows metric.

Definition 1. The Mallows metric d, relative to the Euclidean norm $||\cdot||$, of two distributions F and $\tilde{F}$ is the infimum of $(E||Z - W||_2^2)^{\frac{1}{2}}$ over all pairs of random vectors Z and W, where Z has distribution F and W has distribution $\tilde{F}$. That is,

$$d(F, \tilde{F}) = \inf_{Z \sim F, W \sim \tilde{F}} (E||Z - W||_2^2)^{\frac{1}{2}}. \tag{21}$$

By Lemma 8.1 in [7], the infimum can be attained. To proceed, denote G_n and G_n^* the distribution of T_n and the conditional distribution of T_n^* respectively. Let P^* denote the conditional probability given the error variables $\{\epsilon_i, i = 1, \ldots, n\}$. To show the validity of residual bootstrap after Select+mLS or Select+Ridge, we need more conditions:

Assumption (dd). *Suppose that* $s^2/n \to 0,\ as\ n \to \infty$.

Assumption (f). *Suppose that both the probability of selecting wrong models based on the original data* (X, Y) *and that based on the resample* (X, Y^*) *decay at an exponential rate, i.e.* $P(\hat{S} \neq S) = o(e^{-n^{c_2}})$ *and* $P^*(\hat{S}^* \neq \hat{S}) = o_p(e^{-n^{c_2}})$.

Assumption (g). *Suppose that*

$$\frac{1}{n}||X\beta^*||_2^2 = O(n). \tag{22}$$

Assumption (h). *Suppose that* $max_{1\le i\le n}\sum_{j=1}^{s} x_{ij}^2 = o(n^{\frac{1}{2}})$.

Assumption (dd) is stronger than assumption (d) because it requires that s grows slower than $\sqrt{n}$. Without considering model selection, [8] showed that residual bootstrap OLS fails if p^2/n does not tend to 0. Therefore, the assumption (dd) cannot be weakened. As we shown in the next section, assumption (f) is satisfied if the Irrepresentable condition and some regularity conditions hold. Assumption (g) is a technical assumption which makes the convergence rates in Theorems 1 and 2 more clear. If we suppose $\Lambda_{max}(C_{11}) = \Lambda_{max}(\frac{1}{n}X_S^T X_S) \le \Lambda_{max} < \infty$ where Λ_{max} is a constant, assumption (g) is equivalent to $||\beta^*||_2^2 = O(n)$ because

$$\Lambda_{min}||\beta^*||_2^2 \le \frac{1}{n}||X\beta^*||_2^2 = \frac{1}{n}||X_S\beta_S^*||_2^2 \le \Lambda_{max}\left(\frac{1}{n}X_S^T X_S\right)||\beta^*||_2^2 \le \Lambda_{max}||\beta^*||_2^2.$$

Since β^* has only $s \ll n$ nonzero components, this assumption is not very restrictive. Obviously, it is satisfied when the maximum of β_j^* is upper bounded by a constant. Assumption (h) is not very restrictive either because the number of terms in the sum is $s \ll n^{\frac{1}{2}}$ and it clearly holds when all the predictors corresponding to the nonzero coefficients are bounded by a constant M, i.e. $|x_{ij}| \le M,\ i = 1,\dots,n,\ j = 1,\dots,s$. [29] also assumed this condition to show the asymptotic normality of Bridge estimator.

Let "$\to_p$" denote convergence in probability,

Theorem 4. *Suppose that assumptions (a)–(h) and (dd) are satisfied, then* $d(G_n, G_n^*)$ *converges in probability to zero, i.e.,*

$$d(G_n, G_n^*) \to_p 0. \tag{23}$$

Theorem 4 states that residual bootstrap after Select+mLS or Select+Ridge gives a valid approximation to the distribution of T_n if the probabilities of selecting wrong models $P(\hat{S} \ne S)$ and $P^*(\hat{S}^* \ne \hat{S})$ decay at an exponential rate and the number of true predictors s grows slower than $\sqrt{n}$. The proof of this theorem can be found in Appendix C.

In the next section, we will apply the above Theorems 1, 2, 3 and 4 to two special cases: Lasso+mLS and Lasso+Ridge.

3. Asymptotic properties of Lasso+mLS and Lasso+Ridge

As we shown in Section 2, Select+mLS and Select+Ridge have attractive asymptotic properties (see Theorem 1, 2, 3 and 4) if the probability of selecting wrong models decays exponentially fast. Applying these theorems to Lasso+mLS and Lasso+Ridge, we can easily attain their convergence rates of bias and MSE, asymptotic normality and the validity of residual bootstrap. To proceed, we will first give a brief overview of the assumptions used to get model selection consistency of the Lasso investigated by [54]. Let $sign(\cdot)$ map positive entries to 1, negative entries to -1 and zero to zero.

Definition 2 (**Irrepresentable condition** [25, 48, 36, 54, 51])**.** There exists a positive constant vector η, such that

$$|C_{21}C_{11}^{-1}sign(\beta_S^*)| \leq \mathbf{1} - \eta \tag{24}$$

where $\mathbf{1}$ is a $p-s$ by 1 vector with entries 1 and the inequality holds element-wise.

Assumption (i). *There exist constant* $c_1 + c_2 < c_3 \leq 1$ *and* $M > 0$ *so that*

$$n^{\frac{1-c_3}{2}} \min_{1\leq i\leq s} |\beta_i^*| \geq M. \tag{25}$$

Assumption (j). *Suppose that the tuning parameter* λ_n *in the definition of the Lasso satisfies* $\lambda_n \propto n^{\frac{1+c_4}{2}}$ *with* $c_2 < c_4 < c_3 - c_1$.

The Irrepresentable Condition is a key assumption on the design that can be weakened by e.g. using stability selection instead of the Lasso. [38] showed that stability selection can achieve the same convergence rate of the probability of selecting wrong models as the Lasso does but under a less restrictive sparse eigenvalue condition. We will give a brief overview of stability selection combined with randomized Lasso in sparse linear regression in the Appendix B and discuss the asymptotic properties of two stage estimators based on stability selection in the next section. Assumption (i) requires a gap of size $n^{\frac{c_3}{2}}$ between the decay rate of β^* and $n^{-\frac{1}{2}}$ thus preventing the estimation from being dominated by the noise terms. It is weaker than $min_{1\leq i\leq s}|\beta_i^*| \geq M$, which was assumed by [29] who studied the asymptotic properties of the Bridge estimator. [22] also proposed a similar constraint $min_{1\leq i\leq s}|\beta_i^*| \geq \frac{M}{n^\kappa}$, $0 \leq \kappa < \frac{1}{2}$ to show model selection consistency of the Sure Independent Screening.

Now, we are ready to state our main results. Let $\tilde{\beta}$ and $\tilde{\beta}^*$ denote the Lasso+mLS (or Lasso+Ridge) based on the original data (X, Y) and that based on the resample (X, Y^*) respectively, then we can define $\hat{\Psi}$, Ψ, T_n, T_n^*, G_n and G_n^* as Section 2 does.

Firstly, combining Theorem 1 and the model selection property of the Lasso (see Lemma 2 in Appendix A), we can attain the following corollary:

Corollary 1. *Suppose that assumptions (a)–(d), (g), (i), (j) and the Irrepresentable condition* (24) *are satisfied, if* $\tau_n \propto \frac{1}{n}$ *and* $\mu_n \propto e^{-n^{c_2}/4}$*, then the bias and the MSE of Lasso+mLS and Lasso+Ridge estimators* $\tilde{\beta}$ *satisfy*

$$||E\tilde{\beta} - \beta^*||_2^2 = o(e^{-n^{c_2}/2}) \to 0, \ as \ n \to \infty, \tag{26}$$

$$E||\tilde{\beta} - \beta^*||_2^2 = O\left(\frac{\sigma^2}{\Lambda_{min}}\frac{s}{n}\right). \tag{27}$$

Proof. We only consider the Lasso+mLS since the proof for Lasso+Ridge is similar. By Lemma 2 in Appendix A, we have

$$P(\hat{S} \neq S) \leq P(sign(\hat{\beta}(\lambda_n)) \neq sign(\beta^*)) \leq o(e^{-n^{c_2}}).$$

From assumption (c) (12), we know that $\Lambda_{min}(C_{11}) \geq \Lambda_{min} > 0$. And since C_{11} is an s by s matrix, we have

$$\frac{\sigma^2}{n} tr(C_{11}^{-1}) \leq \frac{\sigma^2 s}{n} \Lambda_{min}^{-1}.$$

Under condition (g) or equation (22), we have

$$||\beta^*||_2^2 \leq \Lambda_{min}^{-1} \frac{1}{n} ||X\beta^*||_2^2 = O(n).$$

The corollary is obtained directly from Theorem 1. □

Corollary 1 indicates that Lasso+mLS and Lasso+Ridge are asymptotically unbiased. In particular, their biases decay at an exponential rate and their MSEs achieve the oracle convergence rate of $\frac{s}{n}$ which is much faster than that of the Lasso. (Under the restricted eigenvalue condition, the Lasso achieves convergence rate of $\frac{s \log p}{n}$ for MSE, see for example [45, 52]).

Secondly, combining Theorem 3 and Lemma 2, we can easily derive the asymptotic normality of Lasso+mLS and Lasso+Ridge.

Corollary 2. *Suppose conditions (a)–(e), (i), (j) and the Irrepresentable Condition (24) are satisfied, then Lasso+mLS and Lasso+Ridge are asymptotically normal*[3]*. That is,*

$$\sup_{t \in R^s} |\hat{\Psi}(t) - \Psi(t)| \to 0, \ as \ n \to \infty. \tag{28}$$

Remark 3.1. For fixed p and fixed β^*, [32] showed that for $\lambda_n = o(\sqrt{n})$, the Lasso estimator is also asymptotically normal $N(0, \sigma^2 C^{-1})$ under conditions $\frac{1}{n} X^T X \to C$ and $\frac{1}{n} max_{1 \leq i \leq n} x_i^T x_i \to 0$. Then the asymptotic covariance matrix of the rescaled and centered Lasso estimator $\sqrt{n}(\hat{\beta}_S - \beta_S^*)$ is σ^2 multiplied by

$$C_{11}^{-1} + C_{11}^{-1} C_{12} (C_{22} - C_{21} C_{11}^{-1} C_{12}) C_{21} C_{11}^{-1} \quad (\succeq C_{11}^{-1})$$

where matrix $A \succeq B$ means $A - B$ is positive definite. Therefore, Lasso+mLS and Lasso+Ridge have smaller asymptotic covariance matrix (which is $\sigma^2 C_{11}^{-1}$) compared with the Lasso and hence reduce estimation uncertainty. Moreover, as pointed out by [21], one cannot find a λ_n such that the Lasso estimator is model selection consistent and asymptotically normal ($\sqrt{n}-$consistency) simultaneously. In this sense, Lasso+mLS and Lasso+Ridge improve the performance of the Lasso.

Lastly, we verify the validity of residual bootstrap after Lasso+mLS and Lasso+Ridge. We would like to begin with showing an interesting result that the Lasso estimator $\hat{\beta}^*$ based on the resample (X, Y^*) also has model selection consistency.

$$\hat{\beta}^* = \underset{\beta}{argmin} \left\{ ||Y^* - X\tilde{\beta}||^2 + \lambda_n ||\beta||_1 \right\}. \tag{29}$$

[3]For Lasso+Ridge, we need assumption (g) to hold. Due to the restricted space, we don't state it separately.

Recall that $\hat{S} = \{j \in, \{1, 2, \ldots, p\} : \tilde{\beta}_j \neq 0\}$ and $\hat{S}^* = \{j \in \{1, 2, \ldots, p\} : \hat{\beta}^*_j \neq 0\}$ are the sets of selected predictors by $\tilde{\beta}$ and $\hat{\beta}^*$ respectively.

Lemma 1. *Suppose conditions (a)–(e), (g)–(j), (dd) and the Irrepresentable Condition* (24) *are satisfied, then the following holds,*

$$P^*(\hat{S}^* \neq \hat{S}) = o_p(e^{-n^{c_2}}).$$

Now, applying Theorem 4, Lemma 1 and Lemma 2, we can state our main result:

Corollary 3. *Suppose that conditions (a)–(e), (g)–(j), (dd) and the Irrepresentable Condition* (24) *are satisfied, then residual bootstrap after Lasso+mLS or Lasso+Ridge is consistent in the sense that*

$$d(G_n, G^*_n) \to_p 0. \tag{30}$$

Remark 3.2. In practice, if the Irrepresentable Condition does not hold, one needs to try other model selection methods, e.g., Bolasso [2] or stability selection. As stated before, as long as the probabilities of selecting wrong models $P(\hat{S} \neq S)$ and $P^*(\hat{S}^* \neq \hat{S})$ decay at an exponential rate $o(e^{-n^{c_2}})$, residual bootstrap after two stage estimator gives valid approximation.

Corollary 3 indicates that residual bootstrap after Lasso+mLS or Lasso+Ridge gives a valid approximation to the distribution of T_n and then can be used to construct confidence intervals and test for parameter estimation. The proof is straightforward and we omit it.

4. Asymptotic properties of modified least squares and ridge after stability selection

If the Irrepresentable Condition is violated, the Lasso cannot correctly select the true model. In this case, one needs to apply other model selection criteria instead of the Lasso. Stability selection is one popular method among many others. [38] showed that it can achieve the same convergence rate of the probability of selecting wrong models but under a less restrictive sparse eigenvalue condition. More details are given in the Appendix B.

Let $\tilde{\beta}$ and $\tilde{\beta}^*$ denote the modified Least Squares (or Ridge) after stability selection (SS+mLS or SS+Ridge) based on the original data (X, Y) and that based on the resample (X, Y^*) respectively, then we can define $\hat{\Psi}$, Ψ, T_n, T^*_n, G_n and G^*_n as Section 2 does. Applying the asymptotic properties of Select+mLS and Select+Ridge in Section 2, we can derive without any proofs the following corollaries parallel to Corollary 1, 2, 3.

Corollary 4. *Assume the conditions in Lemma 3 in the Appendix B and conditions (b), (c) and (g) in the previous sections are satisfied, if* $\tau_n \propto \frac{1}{n}$ *and* $\mu_n \propto e^{-n^{c_2}/4}$*, then the bias and the MSE of the SS+mLS and SS+Ridge* $\tilde{\beta}$ *satisfy*

$$||E\tilde{\beta} - \beta^*||_2^2 = o(e^{-n^{c_2}/2}) \to 0, \ n \to \infty, \tag{31}$$

$$E||\tilde{\beta} - \beta^*||_2^2 = O\left(\frac{\sigma^2}{\Lambda_{min}}\frac{s}{n}\right). \tag{32}$$

Corollary 5. *Assume the conditions in Lemma 3 in the Appendix B and conditions (b)–(e), (g), (h) and (dd) in the previous sections are satisfied, then the SS+mLS and SS+Ridge $\tilde{\beta}$ are asymptotically normal*[4]*, that is,*

$$\sup_{t \in R^s} |\hat{\Psi}(t) - \Psi(t)| \to 0, \text{ as } n \to \infty. \tag{33}$$

Corollary 6. *Assume the conditions in Lemma 3 in the Appendix B and conditions (b), (c), (e), (g), (h) and (dd) in the previous sections are satisfied, then residual bootstrap after SS+mLS or SS+Ridge is consistent in the sense that*

$$d(G_n, G_n^*) \to_p 0. \tag{34}$$

5. Simulation

In this section we carry out simulation studies to evaluate the finite sample performance of Lasso+mLS and Lasso+Ridge. We have also constructed simulations for SS+mLS and SS+Ridge (modified Least Squares or Ridge after stability selection). Though in theory stability selection achieves the same convergence rate of the probability of selecting wrong models under weaker conditions compared with the Lasso, their finite sample performance is similar to the Lasso unless the signal to noise ratio is very high. In our simulation, Lasso+mLS and Lasso+Ridge work well and perform similarly with SS+mLS and SS+Ridge regardless of whether the Irrepresentable condition holds or not. Therefore we only present here the results for Lasso+mLS and Lasso+Ridge.

In the following simulations, we also compared the performance of the Lasso+mLS (with $\tau_n = 1/n$) with that of the Lasso+OLS and found that their finite sample results are almost the same. This is true because the smallest singular value of the matrix $X_{\hat{S}}$ containing the predictors selected by Lasso is bounded well from 0 in all examples which makes the hard thresholding step not necessary. We will omit the results of Lasso+OLS for the sake of brevity.

5.1. Comparison of Bias^2, MSE and PMSE

This subsection compares the bias^2 ($||E\tilde{\beta} - \beta^*||_2^2$), MSE and prediction mean squared error (PMSE) of the Lasso+mLS and Lasso+Ridge with those of the Lasso. We use R package "glmnet" to compute the Lasso solution. As part of the simulation, we fix $\tau_n = 1/n$ and $\mu_n = 1/n$, so the only tuning parameter for all the three methods is λ_n which can be chosen by a 5-fold cross-validation.

Our simulated data are drawn from the following model

$$y = x^T\beta^* + \epsilon, \ \epsilon \sim N(0, \sigma^2).$$

[4]For SS+Ridge, we need assumption (g) to hold. Due to the restricted space, we don't state it separately.

TABLE 1
Example settings

Example	n	p	s	$\Sigma_{ij}, i \neq j$	β^*
1	200	500	10	0	case (1)
2	400	500	10	0	case (1)
3	200	500	10	$0.5^{\lvert i-j\rvert}$	case (1)
4	400	500	10	$0.5^{\lvert i-j\rvert}$	case (1)
5	200	500	10	0	case (2)
6	400	500	10	0	case (2)
7	200	500	10	$0.5^{\lvert i-j\rvert}$	case (2)
8	400	500	10	$0.5^{\lvert i-j\rvert}$	case (2)

We set $p = 500$, $s = 10$ and $\sigma = 1$. The predictor vector x is generated from a multivariate normal distribution $N(0, \Sigma)$. The value of x is generated once and then kept fixed. We consider two different Toeplitz covariance matrices Σ which control the correlation among the predictors: (1) $\Sigma = I$ and (2) $\Sigma_{ij} = \rho^{|i-j|}$ where $\rho = 0.5$. For the true parameter β^*, the first $s = 10$ elements are nonzero with two different patterns of sign:

$$\begin{aligned} &\textit{case } (1): \quad \beta^*_{1-10} = \{1.5, 1.5, 1.5, 1.5, 1.5, 0.75, 0.75, 0.75, 0.75, 0.75\}, \\ &\textit{case } (2): \quad \beta^*_{1-10} = \{1.5, 1.5, -1.5, -1.5, 1.5, 0.75, -0.75, 0.75, -0.75, -0.75\}. \end{aligned}$$

The remaining $p - s = 490$ elements of β^* are zero. Table 1 summaries eight different example settings.

After X was generated, we examined the Irrepresentable condition and found that it holds in examples 1–6 and is violated in examples 7–8. In order to evaluate the prediction performance, we generate an independent testing data set of size 500 and compute the PMSE. Summary statistics are calculated based on 100 replications (keeping X fixed) and showed in Table 2 and Figure 1.

We see that Lasso+mLS and Lasso+Ridge perform almost the same. They not only dramatically decrease the bias^2 of the Lasso by more than 90% but also can reduce the variance (in fact, their variances are 20%–55% smaller than that of the Lasso in all examples except in example 7 where their variances are 45% larger), therefore they improve the MSE and PMSE by 40%–80% and 5%–25% respectively. These benefits occur regardless of whether the Irrepresentable condition holds or not, which indicates that Lasso+mLS and Lasso+Ridge dominate the performance of the Lasso in terms of estimation.

5.2. Finite sample distribution

This section evaluates the finite sample distribution of the scaled and centered Lasso+mLS estimator $T_n = \sqrt{n}(\hat{\beta} - \beta^*)$. Lasso+Ridge behaves similarly, so we omit it.

We show in Figure 2 the histograms and Normal Q-Q Plots of the scaled and centered Lasso and Lasso+mLS estimators based on 1000 replications. We only present the results for individual coefficients $T_{n,j}$ in example 1 with $j = 1, 6, 11$

TABLE 2
Comparison of Lasso, Lasso+mLS and Lasso+Ridge in terms of bias^2, MSE and PMSE

Example		Lasso	Lasso+mLS	Lasso+Ridge
1	bias^2	0.266	**0.003**	0.005
	MSE	0.42(0.11)	0.08(0.05)	**0.07(0.04)**
	PMSE	1.45(0.15)	**1.08(0.08)**	1.08(0.08)
2	bias^2	0.081	**0.001**	0.001
	MSE	0.13(0.04)	0.03(0.03)	**0.03(0.01)**
	PMSE	1.14(0.08)	**1.04(0.07)**	1.04(0.07)
3	bias^2	0.062	**0.001**	0.001
	MSE	0.18(0.06)	**0.09(0.05)**	0.09(0.05)
	PMSE	1.24(0.11)	1.07(0.08)	**1.07(0.07)**
4	bias^2	0.02	**0.001**	0.001
	MSE	0.08(0.03)	**0.04(0.02)**	0.04(0.02)
	PMSE	1.09(0.07)	**1.04(0.07)**	1.04(0.07)
5	bias^2	0.212	**0.001**	0.002
	MSE	0.35(0.09)	0.06(0.04)	**0.06(0.03)**
	PMSE	1.36(0.14)	1.08(0.08)	**1.08(0.07)**
6	bias^2	0.097	0.0002	**0.0001**
	MSE	0.15(0.04)	0.03(0.02)	**0.03(0.01)**
	PMSE	1.16(0.08)	**1.03(0.06)**	1.03(0.06)
7	bias^2	0.494	**0.053**	0.079
	MSE	0.72(0.19)	**0.39(0.31)**	0.44(0.38)
	PMSE	1.48(0.16)	**1.3(0.18)**	1.32(0.2)
8	bias^2	0.343	**0.004**	0.007
	MSE	0.45(0.11)	**0.07(0.04)**	0.08(0.08)
	PMSE	1.26(0.09)	**1.05(0.08)**	1.05(0.08)

$*$ The numbers in parentheses are the corresponding standard deviations.

corresponding to the largest, the medium sized and the zero-valued coefficients respectively. Other coefficients in example 1 and the coefficients in examples 2–8 behave similarly (except example 7 where the Irrepresentable condition does not hold). We can see that the finite sample distribution of the Lasso+mLS highly coincides with the asymptotically normal distribution which verifies the claims in Theorem 3 and Corollary 2. Although the finite sample distributions of the Lasso estimator for the largest and the medium sized coefficients also seem to somewhat resemble normality, the centers shift away from 0.

We should mention that Lasso+mLS suffers the same issue proposed in [44] which studied the distribution of the adaptive Lasso estimator, that is, the finite sample distribution can be highly non-normal when there is not a gap between the decay rate of the nonzero β^* and the order $n^{-1/2}$.

5.3. Confidence intervals and coverage probabilities

In this subsection, we study the finite sample performance of residual bootstrap Lasso+mLS and Lasso+Ridge using the examples from Table 1. Since Lasso+mLS and Lasso+Ridge behave similarly, we only show the results of Lasso+mLS for the sake of brevity.

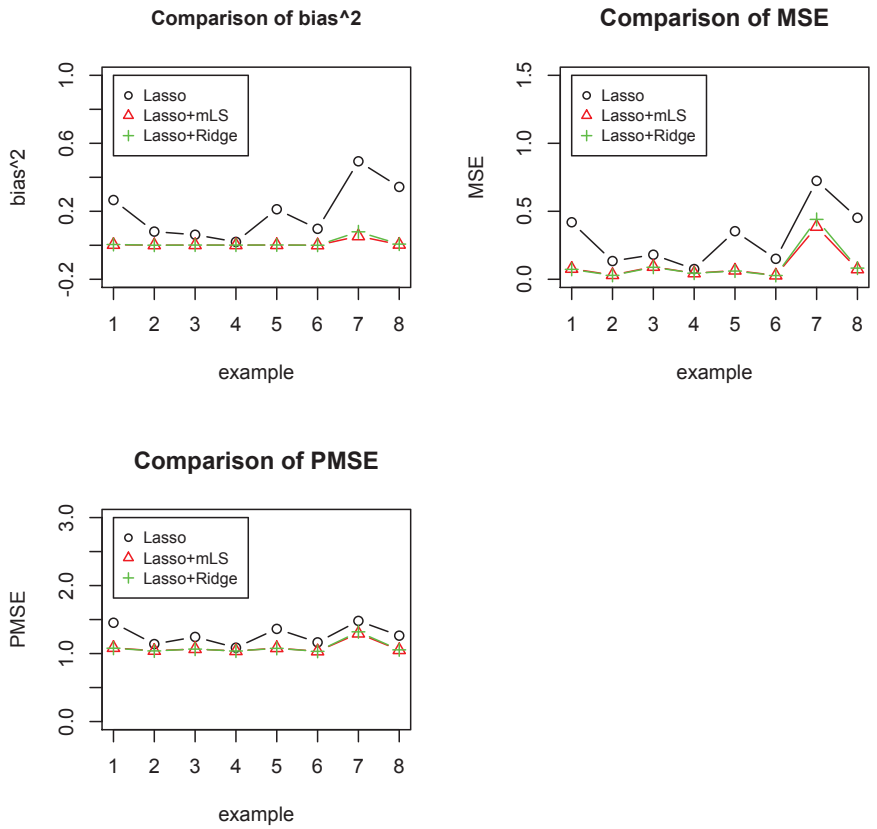

FIG 1. *Comparisons of bias*^{2}*, MSE and PMSE. Three methods are considered: Lasso (black circle), Lasso+mLS (red triangle) and Lasso+Ridge (green "+"). Lasso+mLS and Lasso+Ridge behave similarly, both dominate the performance of the Lasso.*

For each data set (X, Y), we generated 500 bootstrap samples (X, Y^*) by residual bootstrap and then computed the Lasso and Lasso+mLS based on each bootstrap sample (X, Y^*), both are denoted by $\hat{\beta}^*_{(b)}, b = 1, \ldots, 500$. Let $\hat{t}_{\alpha/2}$ and $\hat{t}_{1-\alpha/2}$ be the $\alpha/2$ and $1-\alpha/2$ quantiles of the empirical distribution of $\hat{\beta}^*$. Two approaches are considered to construct $1-\alpha$ confidence intervals for each individual parameter $\beta^*_j, j = 1, \ldots, p$: (1) percentile confidence intervals defined by $[\hat{t}_{\alpha/2}, \hat{t}_{1-\alpha/2}]$; and (2) basic confidence intervals defined by $[2\hat{\beta} - \hat{t}_{1-\alpha/2}, 2\hat{\beta} - \hat{t}_{\alpha/2}]$ where $\hat{\beta}$ is the Lasso estimator for residual bootstrap Lasso or Lasso+mLS for residual bootstrap Lasso+mLS, see [20, 15]. This procedure is repeated 100 times and then an estimate of the coverage probability is obtained.

We found that basic confidence intervals based on residual bootstrap Lasso provide more accurate coverage probabilities while having the same length as

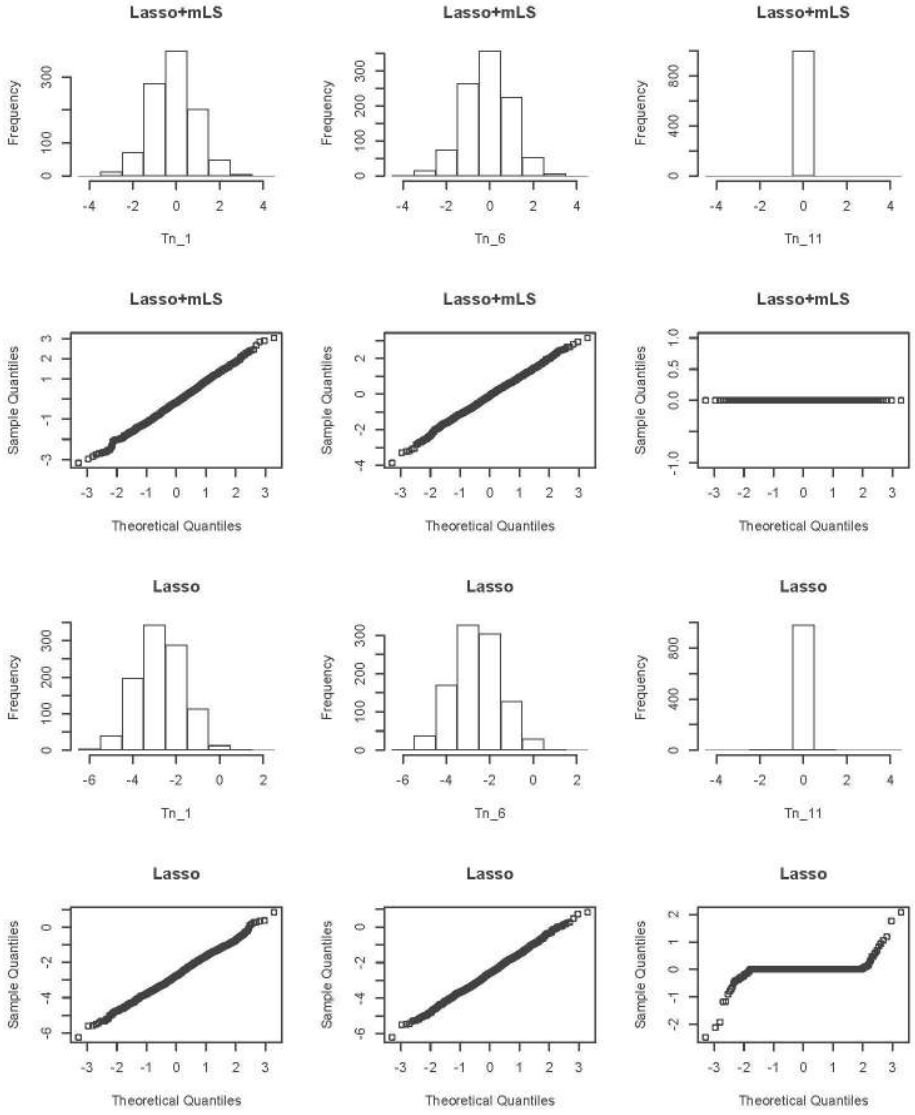

FIG 2. *Histograms and Normal Q-Q Plots of the scaled and centered Lasso and Lasso+mLS estimators in example 1 for* $T_{n,j} = \sqrt{n}(\hat{\beta}_j - \beta_j^*), j = 1, 6, 11$.

percentile confidence intervals (the basic 90% confidence intervals can achieve coverage probabilities larger than 80% while the percentile confidence intervals are too biased that their coverage probabilities can be lower than 20% for the nonzero-valued parameters, see Figure 3). The distribution of residual bootstrap Lasso is far away from being centered at the true value (Figure 4) which makes

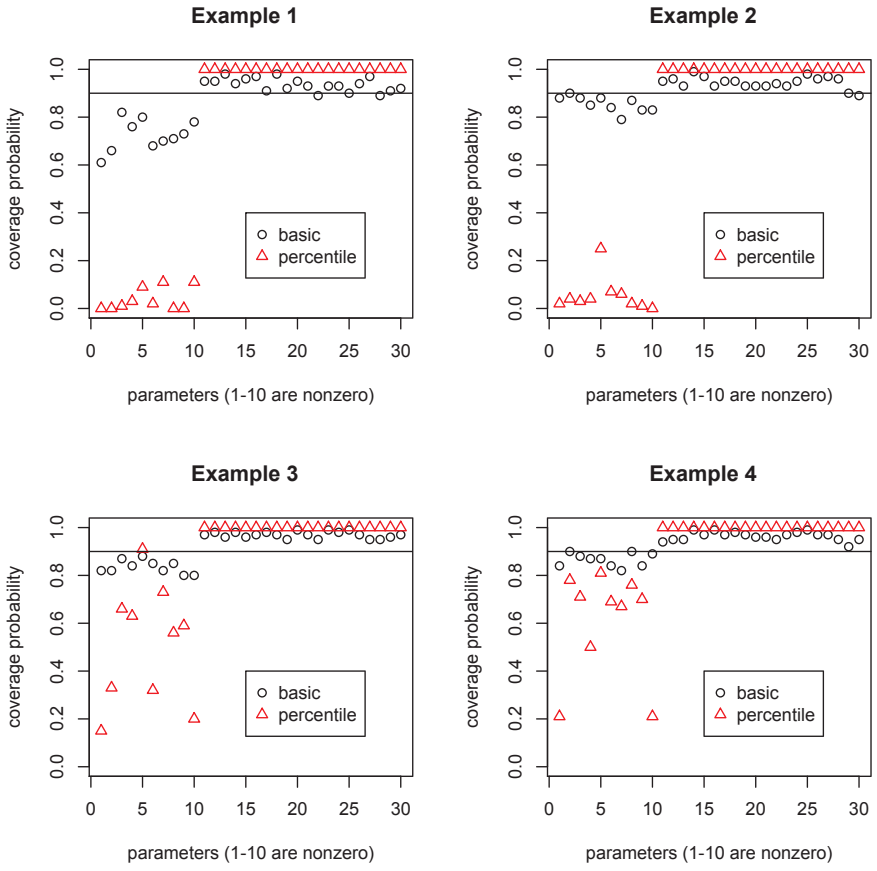

FIG 3. *Comparison of two confidence intervals construction approaches: basic (black circle) and percentile (red triangle). Coverage probabilities of* 90% *confidence intervals for each* $\beta_j^*, j = 1, \ldots, s, s+1, \ldots, s+20$ *based on residual bootstrap Lasso are shown in this figure. We only show the results for examples 1–4 since the results for examples 5–8 are similar. For a better view, the coverage probabilities for only* 20 *zero-valued parameters are present and those for the remaining* $p - s - 20$ *zero-valued parameters are similar to the* 20 *presented and therefore are omitted. Basic confidence intervals provide much more accurate coverage probabilities than percentile confidence intervals.*

the percentile confidence intervals fail. Similar phenomenon happens for paired bootstrap Lasso method. Therefore, we suggest using the basic confidence intervals in practice with high-dimensional data. In what follows, our confidence intervals are all basic.

Figure 5 shows the coverage probabilities of 90% confidence intervals based on residual bootstrap Lasso+mLS and residual bootstrap Lasso. In this figure, only 20 zero-valued parameters are presented for the sake of brevity. The coverage probabilities for the remaining $p - s - 20$ zero-valued parameters are

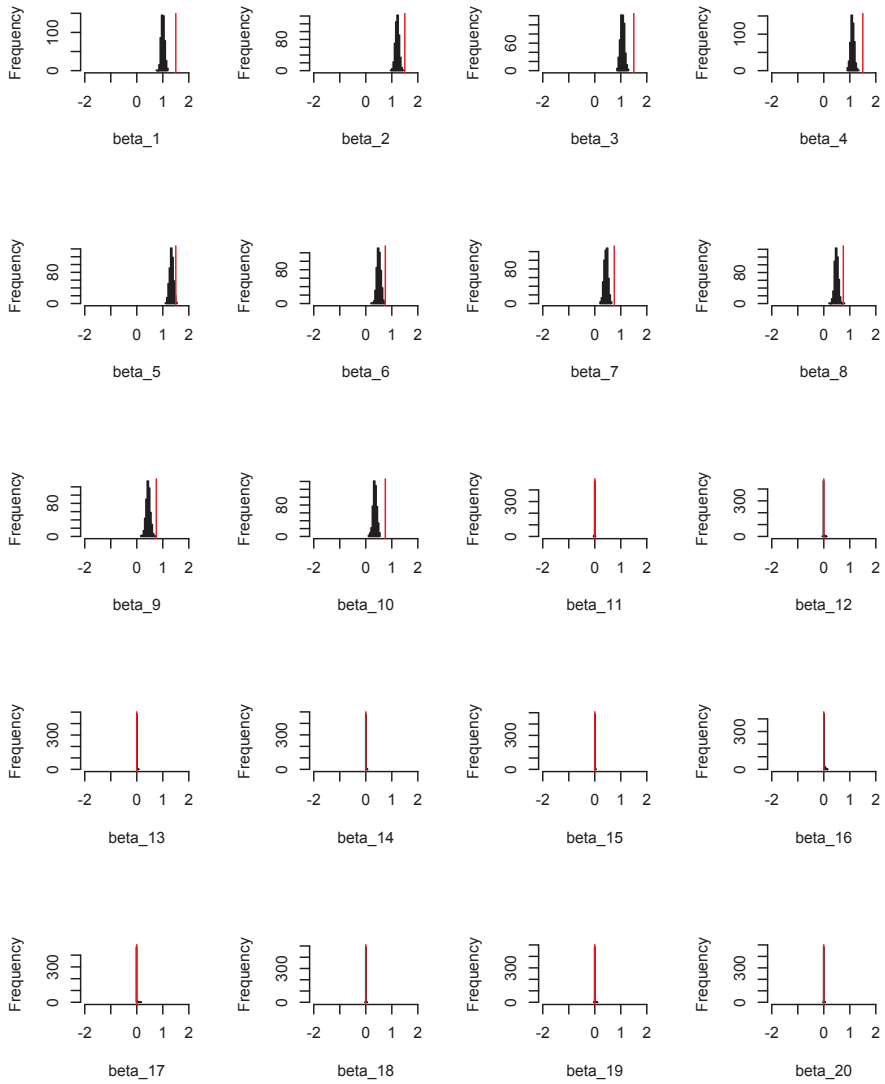

FIG 4. *Histograms of the distribution of residual bootstrap Lasso with the true value plotted as a red vertical line. This figure only show the results for the first* 20 *parameters in example 1. Other parameters and other examples behave similarly. The large bias of the residual bootstrap Lasso makes percentile confidence intervals fail.*

similar to the 20 presented and therefore are omitted. In addition, we average the coverage probabilities and interval lengths over nonzero-valued parameters part and zero-valued parameters part respectively (see Tables 3 and 4). From

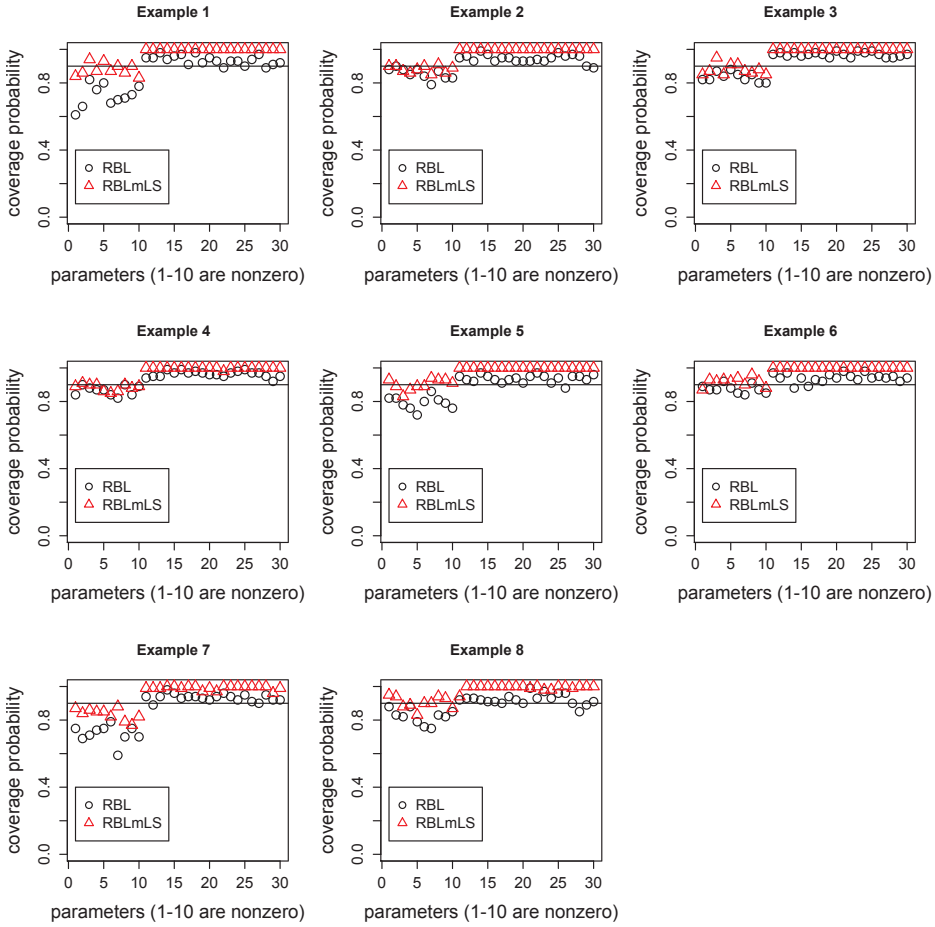

FIG 5. *Coverage probabilities of* 90% *confidence intervals for each* $\beta_j^*, j = 1, \ldots, s, s+1, \ldots, s+20$ *based on two methods: residual bootstrap Lasso+mLS (RBLmLS, red triangle) and residual bootstrap Lasso (RBL, black circle). For a better view, only* 20 *zero-valued parameters are present. The results for the remaining* $p - s - 20$ *zero-valued parameters are similar to the* 20 *presented and therefore are omitted. Residual bootstrap Lasso+mLS provides more accurate coverage probabilities (*7% *in average closer to the preassigned level* (90%) *for nonzero* β_j^* *and* 5% *closer to* 1 *for zero* β_j^**).*

Figure 5 and Tables 3 and 4, we can see that residual bootstrap Lasso+mLS gives accurate coverage probabilities (approximately 88% for nonzero β_j^* and 1 for zero β_j^*) when the Irrepresentable condition holds (see examples 1–6), which verifies Corollary 3. Note that, for zero-valued parameters, residual bootstrap Lasso+mLS produces confidence intervals with coverage probabilities close to 1 and very short lengths (approximately 0, see Figure 6 and Table 4), reflecting the oracle properties in Corollary 2. By contrast, residual bootstrap Lasso cannot

TABLE 3
Mean coverage probability of residual bootstrap Lasso (RBL), residual bootstrap Lasso+mLS (RBLmLS) and paired bootstrap Lasso (PBL)

Example		1	2	3	4	5	6	7	8
nonzero β_j^*	RBL	0.725	0.855	0.835	0.865	0.792	0.875	0.717	0.821
	RBLmLS	0.880	0.882	0.880	0.884	0.901	0.917	0.835	0.903
	PBL	0.772	0.634	0.833	0.816	0.801	0.609	0.866	0.512
zero β_j^*	RBL	0.937	0.947	0.965	0.968	0.940	0.947	0.931	0.926
	RBLmLS	0.999	1.000	1.000	1.000	1.000	1.000	0.991	0.997
	PBL	0.990	0.996	0.997	0.999	0.991	0.997	0.987	0.992

TABLE 4
Mean interval length of residual bootstrap Lasso (RBL), residual bootstrap Lasso+mLS (RBLmLS) and paired bootstrap Lasso (PBL)

Example		1	2	3	4	5	6	7	8
nonzero β_j^*	RBL	0.209	0.158	0.278	0.202	0.222	0.165	0.266	0.203
	RBLmLS	0.248	0.178	0.322	0.241	0.254	0.204	0.350	0.290
	PBL	0.288	0.170	0.307	0.207	0.293	0.176	0.412	0.238
zero β_j^*	RBL	0.011	0.004	0.002	0.001	0.009	0.004	0.017	0.014
	RBLmLS	0.001	0.000	0.000	0.000	0.000	0.000	0.009	0.002
	PBL	0.024	0.016	0.012	0.010	0.022	0.016	0.031	0.025

provide accurate coverage probabilities unless n is large enough (the coverage probability for nonzero β_j^* is around 75% for $n = 200$ and is around 85% for $n = 400$). Even when the Irrepresentable condition doesn't hold (see examples 7–8), residual bootstrap Lasso+mLS can also provide reasonable coverage probabilities (83.5%, 90.3% for nonzero β_j^*, and 99.1%, 99.7% for zero β_j^*) while residual bootstrap Lasso does not (71.7%, 82.1% for nonzero β_j^* and 93.1%, 92.6% for zero β_j^*). Compared with residual bootstrap Lasso, the coverage probability of residual bootstrap Lasso+mLS is about 7% in average closer to the preassigned level (90%) for nonzero β_j^* and 5% closer to 1 for zero β_j^*. Even though residual bootstrap Lasso has shorter (17% in average) interval lengths for the nonzero β_j^*, it loses accuracy in coverage. Overall, residual bootstrap Lasso+mLS is better than residual bootstrap Lasso. Moreover, when n increases, the performance of both methods become better.

In practice, many people prefer to perform paired bootstrap Lasso (resampling from the pairs $(x_i, y_i), i = 1, \ldots, n$ instead of from the residual) even when it makes sense to think of the design matrix as fixed. Therefore, we give some comparisons of residual bootstrap Lasso+mLS and paired bootstrap Lasso. Figure 7 shows the coverage probabilities v.s. average interval lengths based on residual bootstrap Lasso+mLS and paired bootstrap Lasso for different examples. Again, we average the coverage probabilities and interval lengths over nonzero-valued parameters part and zero-valued parameters part respectively and show them in Table 3 and Table 4. We can see that residual bootstrap Lasso+mLS provides more accurate coverage probabilities (0.5% closer to 1 for zero β_j^* and 14% in average closer to the preassigned level for nonzero β_j^*) with more than 90% shorter (for zero β_j^*) or at least comparable (for nonzero β_j^*)

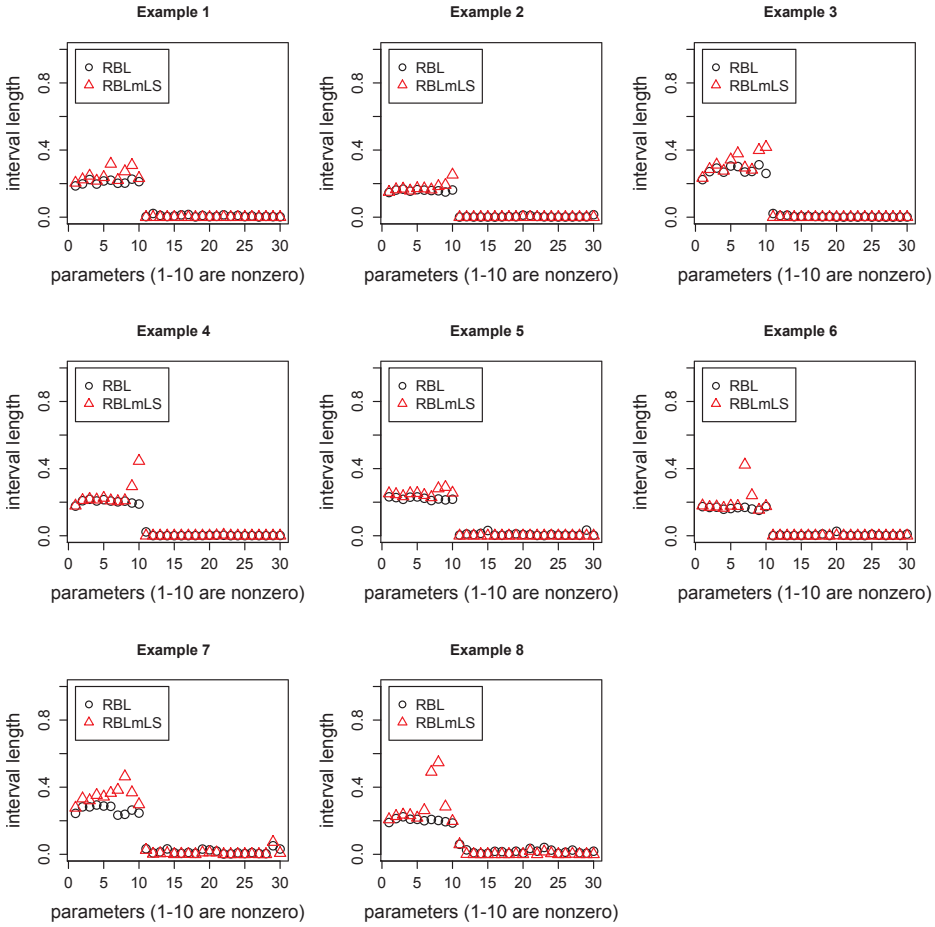

FIG 6. *Lengths of* 90% *confidence intervals for each* $\beta_j^*, j = 1, \ldots, s, s+1, \ldots, s+20$ *based on two methods: residual bootstrap Lasso+mLS (RBLmLS, red triangle) and residual bootstrap Lasso (RBL, black circle). For a better view, only* 20 *zero-valued parameters are present. The interval lengths for the remaining* $p - s - 20$ *zero-valued parameters are similar to the* 20 *presented and therefore are omitted.*

interval lengths compared with paired bootstrap Lasso. Based on our simulations, we conclude that residual bootstrap Lasso+mLS and residual bootstrap Lasso+Ridge are better choices for constructing confidence intervals.

6. Conclusion

We have derived for the first time the asymptotic properties of Lasso+mLS and Lasso+Ridge in sparse high-dimensional linear regression models where $p \gg n$.

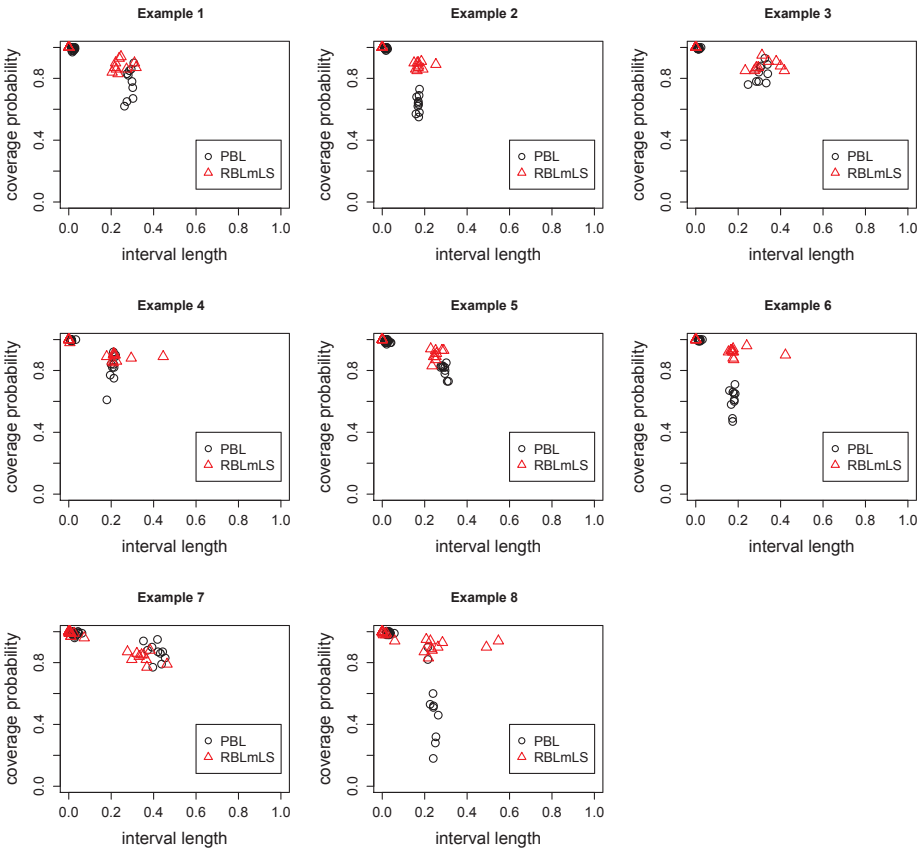

FIG 7. *Coverage probabilities v.s. average interval lengths for* 90% *confidence intervals based on two methods: paired bootstrap Lasso (PBL, black circle) and residual bootstrap Lasso+mLS (RBLmLS, red triangle). The latter is better since it provides more accurate coverage probabilities (*0.5% *closer to* 1 *for zero* β_j^* *and* 14% *in average closer to the preassigned level (*90%*) for nonzero* β_j^**) but with* 90% *shorter (for zero* β_j^**) or at least comparable (for nonzero* β_j^**) interval lengths.*

Under the Irrepresentable condition and other common conditions on scaling (n, s, p), we showed that both Lasso+mLS and Lasso+Ridge are asymptotically unbiased and they achieve oracle convergence rate of $\frac{s}{n}$ for MSE which improves the performance of the Lasso. In addition, Lasso+mLS and Lasso+Ridge estimators have an oracle property in the sense that they can select the true predictors with probability converging to 1 and the estimates of nonzero parameters have the same asymptotic normal distribution that they would have if the zero parameters were known in advance.

We then proposed residual bootstrap after Lasso+mLS and Lasso+Ridge methods and showed that they give valid approximations to the distributions

of Lasso+mLS and Lasso+Ridge, respectively, provided that the probability of the Lasso selecting wrong models decays at an exponential rate and the number of true predictors s goes to infinity slower than $\sqrt{n}$. In fact, our analysis is not limited to adopting Lasso in the selection stage, but is applicable to any other model selection criteria with exponentially decay rates of the probability of selecting wrong models, for example, stability selection, SCAD and Dantzig selector.

Lastly, we presented simulation results assessing the finite sample performance of Lasso+mLS and Lasso+Ridge and observe that they not only dramatically decrease the bias^2 of the Lasso by more than 90% but also reduce the MSE and PMSE by $40\% - 80\%$ and $5\% - 25\%$ respectively. Further, we constructed 90% confidence interval based on our residual bootstrap Lasso+mLS (or Lasso+Ridge) and examined the coverage accuracy. We found that our method resulted in coverage probability approximately 88% for nonzero β_j^* and 1 for zero β_j^*, which is much more accurate (approximately 7% closer to the desired level) than bootstrap Lasso method.

Acknowledgements

Hanzhong Liu would like to thank the China Scholarship Council (CSC) for financial support and the Statistics Department at UC Berkeley for hosting his visit during which this work was finished. This research is also supported in part by NSF grants SES-0835531 (CDI), DMS-1107000, DMS-1228246, ARO grant W911NF-11-1-0114, and the Center for Science of Information (CSoI), an US NSF Science and Technology Center, under grant agreement CCF-0939370. The authors are very grateful to Taesup Moon, Siqi Wu, Jyothsna Sainath, Adam Bloniarz and the referees for many insightful comments and helpful suggestions that lead to a substantially improved manuscript.

Appendix A: Model selection consistency of the Lasso

Lemma 2 (Zhao and Yu (2006))**.** *Under conditions (a)–(d), (i), (j) and the Irrepresentable Condition* (24)*, the Lasso has strong sign consistency. That is,*

$$P(sign(\hat{\beta}(\lambda_n)) = sign(\beta^*)) \geq 1 - o(e^{-n^{c_2}}) \to 1 \ as \ n \to \infty.$$

Remark A.1. In fact, looking carefully through the proof of Lemma 2 in [54], the Gaussian assumption (a) can be relaxed by a subgaussian assumption. That is, assume that there exists constants C, $c > 0$ so

$$P(|\epsilon_i| \geq t) \leq Ce^{-ct^2}, \ \forall t \geq 0.$$

This result tells us that using the Lasso we can allow p to grow faster than n (up to exponentially fast) while the probability of correct model selection still converges to 1 fast.

Appendix B: Stability selection

Here we will give a brief introduction of stability selection combined with randomized Lasso in sparse linear regression.

The randomized Lasso is a generalization of the Lasso, which penalizes the absolute value $|\beta_k|$ of every component with a penalty randomly chosen in the range $[\lambda, \lambda/\alpha]$ for $\alpha \in (0, 1]$. Let W_k be i.i.d. random variables in $[\alpha, 1]$ for $k = 1, \ldots, p$. The randomized Lasso estimator $\hat{\beta}^{\lambda,W}$ is defined by

$$\hat{\beta}^{\lambda,W} = \underset{\beta}{argmin} \left\{ ||Y - X\beta||^2 + \lambda \sum_{k=1}^{p} \frac{|\beta_k|}{W_k} \right\} \tag{35}$$

where $\lambda \in R^+$ is the regularization parameter. [38] proposed an appropriate distribution of the weights W_k: $W_k = \alpha$ with probability $p_w \in (0, 1)$ and $W_k = 1$ otherwise.

For any given regularization parameter $\lambda \in \Lambda \subseteq R^+$, denote the selected predictor set based on the samples $I \subset \{1, \ldots, n\}$ as $\hat{S}^{\lambda,W}(I) = \{k : \hat{\beta}_k^{\lambda,W} \neq 0\}$.

Definition 3 (Meinshausen and Buhlmann (2010))**.** Let I be a random subsample of $\{1, \ldots, n\}$ of size $\lfloor n/2 \rfloor$, drawn with replacement. For every set $K \subseteq \{1, \ldots, p\}$, the probability of being in the selected set $\hat{S}^{\lambda,W}(I)$ is

$$\hat{\Pi}_K^{\lambda} = P^*\{K \subseteq \hat{S}^{\lambda,W}(I)\} \tag{36}$$

where the probability P^* is with respect to both the random subsampling and the randomness of the weights W_k.

With stability selection, we subsample the data many times and then choose the predictors with a high selection probability.

Definition 4 (Meinshausen and Buhlmann (2010))**.** For a cut-off π_{thr} with $0 < \pi_{thr} < 1$ and a set of regularization parameters Λ, the set of stable variables is defined as

$$\hat{S}^{stable} = \{k : \underset{\lambda \in \Lambda}{max}\, \hat{\Pi}_k^{\lambda} \geq \pi_{thr}\}. \tag{37}$$

For stability selection with randomized Lasso, one can obtain selection consistency by just assuming sparse eigenvalues, a condition that is much weaker than that of the Irrepresentability. Sparse eigenvalues condition essentially requires that the minimum and maximum eigenvalues, for a selection of order s predictors, are bounded away from 0 and ∞ respectively.

Definition 5 (Meinshausen and Buhlmann (2010))**.** For any $K \subseteq \{1, \ldots, p\}$, let X_K be the restriction of X to columns in K. The minimal sparse eigenvalue ϕ_{min} is defined for $k \leq p$ as

$$\phi_{min}(k) = \underset{a \in R^{\lceil k \rceil}, K \subseteq \{1,\ldots,p\} : |K| \leq \lceil k \rceil}{inf} \left\{ \frac{||X_K a||}{||a||} \right\} \tag{38}$$

and analogously for the maximal sparse eigenvalue ϕ_{max}.

Sparse eigenvalues assumption. *There are some $C > 1$ and some $\kappa \geq 9$ such that*

$$\frac{\phi_{max}(Cs^2)}{\phi^{3/2}_{min}(Cs^2)} < \sqrt{C}/\kappa.$$

Under this assumption, we can state the selection consistency result obtained directly from Theorem 2 in [38].

Lemma 3 (Meinshausen and Buhlmann (2010))**.** *For the randomized Lasso, let α be given by $\alpha^2 = \nu\phi_{min}(m)/m$, for any $\nu \in ((7/\kappa)^2, 1/\sqrt{2})$, and $m = Cs^2$. Let $c_2 < c < 1$ and $\lambda_{min} = 2\sigma\{\sqrt{2C}s + 1\}n^{(c-1)/2}$. Assume that $p = O(e^{n^{c_2}}) > 10$ and $s \geq 7$ and that the gaussian assumption (a) and the sparse eigenvalues assumption are satisfied. For any $\lambda \geq \lambda_{min}$, if $min_{1\leq i\leq s}|\beta^*_i| \geq (Cs)^{3/2}(0.3\lambda)$, then there is some $\delta \in (0,1)$ such that, for all $\pi_{thr} \geq 1-\delta$, stability selection with randomized Lasso satisfies*

$$P(\hat{S}^{stable}_\lambda = S) \geq 1 - o(e^{-n^{c_2}})$$

where $\hat{S}^{stable}_\lambda = \{k : \Pi^\lambda_k \geq \pi_{thr}\}$.

Appendix C: Technical details

Proof of Theorem 1. Denote $\tilde{\beta}$ the Select+mLS. Conditioned on $\{\hat{S} = S\}$, for $\tau_n^2 \leq \Lambda_{min}(C_{11})$, we have

$$\tilde{\beta}_S = (X_S^T X_S)^{-1} X_S^T Y = \beta^*_S + (X_S^T X_S)^{-1} X_S^T \epsilon.$$

Combine with triangle inequality,

$$\begin{aligned}
||E\tilde{\beta} - \beta^*||_2 &\leq ||E\tilde{\beta}\mathbb{I}_{\hat{S}=S} - \beta^*||_2 + ||E\tilde{\beta}\mathbb{I}_{\hat{S}\neq S}||_2 \\
&= ||E\{(X_S^T X_S)^{-1} X_S^T Y \mathbb{I}_{\hat{S}=S}\} - \beta^*||_2 + ||E\tilde{\beta}\mathbb{I}_{\hat{S}\neq S}||_2 \\
&\leq ||E\{(X_S^T X_S)^{-1} X_S^T Y - \beta^*_S\}||_2 + ||E\{(X_S^T X_S)^{-1} X_S^T Y \mathbb{I}_{\hat{S}\neq S}\}||_2 \\
&\quad + ||E\tilde{\beta}\mathbb{I}_{\hat{S}\neq S}||_2 \\
&= ||E\{(X_S^T X_S)^{-1} X_S^T Y \mathbb{I}_{\hat{S}\neq S}\}||_2 + ||E\tilde{\beta}\mathbb{I}_{\hat{S}\neq S}||_2
\end{aligned}$$

where $\mathbb{I}_A$ is the indicater function. The last equality holds since

$$E\{(X_S^T X_S)^{-1} X_S^T Y\} = \beta^*_S + (X_S^T X_S)^{-1} X_S^T E\epsilon = \beta^*_S.$$

By Cauchy-Schwarz inequality,

$$||E\{(X_S^T X_S)^{-1} X_S^T Y \mathbb{I}_{\hat{S}\neq S}\}||_2^2 \leq E||\{(X_S^T X_S)^{-1} X_S^T Y\}||_2^2 P(\hat{S} \neq S),$$

$$||E\tilde{\beta}\mathbb{I}_{\hat{S}\neq S}||_2^2 \leq E||\tilde{\beta}||_2^2 P(\hat{S} \neq S).$$

So we need to control $E||\{(X_S^T X_S)^{-1} X_S^T Y\}||_2^2$ and $E||\tilde{\beta}||_2^2$ respectively.

$$E||\{(X_S^T X_S)^{-1} X_S^T Y\}||_2^2 = E||\beta^*_S + (X_S^T X_S)^{-1} X_S^T \epsilon||_2^2$$

$$\begin{aligned} &= ||\beta_S^*||_2^2 + E||(X_S^T X_S)^{-1} X_S^T \epsilon||_2^2 \\ &= ||\beta^*||_2^2 + \sigma^2 tr(X_S^T X_S)^{-1} \\ &= ||\beta^*||_2^2 + \frac{\sigma^2}{n} tr(C_{11}^{-1}). \end{aligned}$$

By definition (7), we have

$$||\tilde{\beta}||_2^2 = \left|\left| \frac{1}{\sqrt{n}} V \tilde{D} U^T Y \right|\right|_2^2 = \frac{1}{n} Y^T U \tilde{D}^T \tilde{D} U^T Y \leq \frac{1}{n} \tau_n^{-2} ||Y||_2^2, \tag{39}$$

where $\tilde{D}$ is a $d \times n$ diagonal matrix with diagonal entries $\lambda_1^{-1}, \lambda_2^{-1}, \ldots, \lambda_d^{-1}$ (Note that we take $\lambda_k^{-1} = 0$ for all $\lambda_k < \tau_n$). The last equality holds since the largest singular value of $\tilde{D}$ is no more than τ_n^{-1} and U is an orthogonal matrix. Moreover,

$$E||Y||_2^2 = E||X\beta^* + \epsilon||_2^2 = ||X\beta^*||_2^2 + n\sigma^2. \tag{40}$$

Combining the above results, we obtain (14).

Next, we prove the second part of Theorem 1.

$$\begin{aligned} E||\tilde{\beta} - \beta^*||_2^2 &= E||\tilde{\beta} - \beta^*||_2^2 \mathbb{I}_{\hat{S}=S} + E||\tilde{\beta} - \beta^*||_2^2 \mathbb{I}_{\hat{S}\neq S} \\ &= E||\{(X_S^T X_S)^{-1} X_S^T \epsilon\}||_2^2 \mathbb{I}_{\hat{S}=S} + E||\tilde{\beta} - \beta^*||_2^2 \mathbb{I}_{\hat{S}\neq S} \\ &\leq E||\{(X_S^T X_S)^{-1} X_S^T \epsilon\}||_2^2 + 2(E||\tilde{\beta}||_2^2 \mathbb{I}_{\hat{S}\neq S} + E||\beta^*||_2^2 \mathbb{I}_{\hat{S}\neq S}) \\ &\leq \frac{\sigma^2}{n} tr(C_{11}^{-1}) + 2\sqrt{P(\hat{S} \neq S)}\left(\sqrt{E||\tilde{\beta}||_2^4} + ||\beta^*||_2^2\right). \end{aligned} \tag{41}$$

The last inequality holds for Cauchy-Schwarz inequality. Since $Y = X\beta^* + \epsilon$, we have

$$||Y||_2^4 = (||Y||_2^2)^2 \leq 4(||X\beta^*||_2^2 + ||\epsilon||_2^2)^2 \leq 8(||X\beta^*||_2^4 + ||\epsilon||_2^4).$$

Because $\epsilon_i \sim N(0, \sigma^2)$, *i.i.d.*, we get $E(\epsilon_i)^4 = 3\sigma^4$ and

$$E||\epsilon||_2^4 = E\left(\sum_{i=1}^n \epsilon_i^2\right)^2 = \sum_{i=1}^n E\epsilon_i^4 + 2\sum_{i<j} E\epsilon_i^2 \epsilon_j^2 = (n^2 + 2n)\sigma^4, \tag{42}$$

hence when $n \geq 2$

$$E||Y||_2^4 \leq 8(||X\beta^*||_2^4 + E||\epsilon||_2^4) \leq 16(||X\beta^*||_2^4 + n^2\sigma^4).$$

Connect (39), then

$$E||\tilde{\beta}||_2^4 \leq \frac{1}{n^2} \tau_n^{-4} E||Y||_2^4 \leq 16\tau_n^{-4}\left(\frac{1}{n^2}||X\beta^*||_2^4 + \sigma^4\right). \tag{43}$$

Taking (43) back to (41) gives the result. □

Proof of Theorem 2. Denote $\tilde{\beta}$ the Select+Ridge, we have

$$\tilde{\beta}_{\hat{S}} = (X_{\hat{S}}^T X_{\hat{S}} + \mu_n I)^{-1} X_{\hat{S}}^T Y. \tag{44}$$

Applying SVD decomposition of $\frac{1}{\sqrt{n}}X_{\hat{S}}$ in (5), it is easy to obtain

$$\tilde{\beta}_{\hat{S}} = VD_1D^TU^TY \tag{45}$$

where D_1 is a diagonal matrix with diagonal entries $\frac{\sqrt{n}}{n\lambda_1^2+\mu_n}, \ldots, \frac{\sqrt{n}}{n\lambda_d^2+\mu_n}$. Since V is an orthogonal matrix,

$$||\tilde{\beta}_{\hat{S}}||_2^2 = Y^TUDD_1^TV^TVD_1D^TU^TY = Y^TUD_2U^TY$$

where

$$D_2 = diag\left\{\frac{n\lambda_1^2}{(n\lambda_1^2+\mu_n)^2}, \ldots, \frac{n\lambda_d^2}{(n\lambda_d^2+\mu_n)^2}\right\}.$$

Therefore,

$$||\tilde{\beta}_{\hat{S}}||_2^2 \leq \Lambda_{max}(D_2)Y^TUU^TY = \Lambda_{max}(D_2)||Y||_2^2 \leq \frac{1}{4\mu_n}||Y||_2^2,$$

the last inequality is due to $\frac{n\lambda_i^2}{(n\lambda_i^2+\mu_n)^2} \leq \frac{1}{4\mu_n},\ i = 1, \ldots, d$. Then,

$$E||\tilde{\beta}_{\hat{S}}||_2^2 \leq \frac{1}{4\mu_n}E||Y||_2^2 = \frac{1}{4\mu_n}(||X\beta^*||_2^2 + n\sigma^2),$$

$$E||\tilde{\beta}_{\hat{S}}||_2^4 \leq \frac{1}{16\mu_n^2}E||Y||_2^4 = \frac{1}{16\mu_n^2}16(||X\beta^*||_2^4 + n^2\sigma^4).$$

Combine Cauchy-Schwarz inequality, we have

$$||E\tilde{\beta}_{\hat{S}}\mathbb{I}_{\hat{S}\neq S}||_2^2 \leq P(\hat{S} \neq S)\frac{n}{4\mu_n}\left(\frac{1}{n}||X\beta^*||_2^2 + \sigma^2\right),$$

$$||E\tilde{\beta}_S\mathbb{I}_{\hat{S}\neq S}||_2^2 \leq P(\hat{S} \neq S)\frac{n}{4\mu_n}\left(\frac{1}{n}||X\beta^*||_2^2 + \sigma^2\right).$$

Moreover,

$$\begin{aligned}\tilde{\beta}_S - \beta_S^* &= VD_1D^TU^TY - \beta_S^* = (VD_1D^TU^TX_S - I)\beta_S^* + VD_1D^TU^T\epsilon \\ &= (\sqrt{n}VD_1D^TU^TUDV^T - I)\beta_S^* + VD_1D^TU^T\epsilon \\ &= VD_3V^T\beta_S^* + VD_1D^TU^T\epsilon\end{aligned}$$

where

$$D_3 = diag\left\{\frac{-\mu_n}{(n\lambda_1^2+\mu_n)}, \ldots, \frac{-\mu_n}{(n\lambda_s^2+\mu_n)}\right\},$$

and using $(\frac{\mu_n}{n\lambda_i^2+\mu_n})^2 \leq \frac{\mu_n^2}{n^2\Lambda_{min}^2},\ i = 1, \ldots, s$, we have

$$||E\tilde{\beta}_S - \beta_S^*||_2^2 = ||VD_3V^T\beta_S^*||_2^2 \leq \frac{\mu_n^2}{n^2\Lambda_{min}^2}||\beta^*||_2^2.$$

Therefore,

$$\begin{aligned} ||E\tilde{\beta} - \beta^*||_2 &\leq ||E\tilde{\beta}\mathbb{I}_{\hat{S}=S} - \beta^*||_2 + ||E\tilde{\beta}\mathbb{I}_{\hat{S}\neq S}||_2 \\ &\leq ||E\tilde{\beta}_S - \beta^*_S||_2 + ||E\tilde{\beta}_S\mathbb{I}_{\hat{S}\neq S}||_2 + ||E\tilde{\beta}_{\hat{S}}\mathbb{I}_{\hat{S}\neq S}||_2 \\ &\leq \frac{\mu_n}{n\Lambda_{min}}||\beta^*||_2 + 2\sqrt{P(\hat{S}\neq S)\frac{n}{4\mu_n}\left(\frac{1}{n}||X\beta^*||_2^2 + \sigma^2\right)}, \end{aligned}$$

which proves the first part of Theorem 2. For the second part, we have

$$\begin{aligned} E||\tilde{\beta}_S - \beta^*_S||_2^2 &= ||VD_3V^T\beta^*_S||_2^2 + E||VD_1D^TU^T\epsilon||_2^2 \\ &\leq \frac{\mu_n^2}{n^2\Lambda_{min}^2}||\beta^*||_2^2 + \sigma^2 tr(VD_1D^TU^TUDD_1V^T) \\ &= \frac{\mu_n^2}{n^2\Lambda_{min}^2}||\beta^*||_2^2 + \sigma^2\sum_{i=1}^{s}\frac{n\lambda_i^2}{(n\lambda_i^2+\mu_n)^2}. \end{aligned}$$

Algebraic operation yields

$$\sigma^2\sum_{i=1}^{s}\frac{n\lambda_i^2}{(n\lambda_i^2+\mu_n)^2} = \frac{\sigma^2}{n}tr\left\{\left(C_{11}+\frac{\mu_n}{n}I\right)^{-2}C_{11}\right\},$$

then,

$$E||\tilde{\beta}_S - \beta^*_S||_2^2 \leq \frac{\sigma^2}{n}tr\left\{\left(C_{11}+\frac{\mu_n}{n}I\right)^{-2}C_{11}\right\} + \frac{\mu_n^2}{n^2\Lambda_{min}^2}||\beta^*||_2^2.$$

On the other hand,

$$\begin{aligned} E||\tilde{\beta} - \beta^*||_2^2\mathbb{I}_{\hat{S}\neq S} &\leq 2\sqrt{P(\hat{S}\neq S)}\left(\sqrt{E||\tilde{\beta}||_2^4} + ||\beta^*||_2^2\right) \\ &\leq 2\sqrt{P(\hat{S}\neq S)}\left\{||\beta^*||_2^2 + \frac{n}{\mu_n}\left[\frac{1}{n}||X\beta^*||_2^2 + \sigma^2\right]\right\}. \end{aligned}$$

Applying the same trick as (41), we obtain the result. □

Proof of Theorem 3. We prove the results for Select+mLS and Select+Ridge respectively.

(1) Select+mLS: as stated in the proof of Theorem 1, conditioned on $\{\hat{S}=S\}$, when n is large enough, $\tau_n^2 \propto \frac{1}{n^2} \leq \Lambda_{min} \leq \Lambda_{min}(C_{11})$, then Select+mLS $\tilde{\beta}$ satisfies:

$$\tilde{\beta}_S = (X_S^TX_S)^{-1}X_S^TY = \beta^*_S + (X_S^TX_S)^{-1}X_S^T\epsilon.$$

Because $\epsilon_i \sim N(0,\sigma^2)$, *i.i.d.*, then $\sqrt{n}(X_S^TX_S)^{-1}X_S^T\epsilon \sim N(0,\sigma^2n(X_S^TX_S)^{-1})$, which is $N(0,\sigma^2C_{11}^{-1})$ since $C_{11} = \frac{1}{n}X_S^TX_S$. Therefore,

$$\begin{aligned} \hat{\Psi}(t) &= P(\sqrt{n}(\tilde{\beta}_S - \beta^*_S) \leq t) \\ &= P(\sqrt{n}(\tilde{\beta}_S - \beta^*_S) \leq t, \hat{S}=S) + P(\sqrt{n}(\tilde{\beta}_S - \beta^*_S) \leq t,\ \hat{S}\neq S) \end{aligned}$$

$$\begin{aligned} &= P(\sqrt{n}(X_S^T X_S)^{-1} X_S^T \epsilon \le t,\ \hat{S} = S) + P(\sqrt{n}(\tilde{\beta}_S - \beta_S^*) \le t,\ \hat{S} \ne S) \\ &= \Psi(t) - P(\sqrt{n}(X_S^T X_S)^{-1} X_S^T \epsilon \le t,\ \hat{S} \ne S) \\ &\quad + P(\sqrt{n}(\tilde{\beta}_S - \beta_S^*) \le t,\ \hat{S} \ne S). \end{aligned}$$

Then

$$\begin{aligned} &\sup_{t \in R^s} |\hat{\Psi}(t) - \Psi(t)| \\ \le\ &\sup_{t \in R^s} \{P(\sqrt{n}(X_S^T X_S)^{-1} X_S^T \epsilon \le t,\ \hat{S} \ne S) + P(\sqrt{n}(\tilde{\beta}_S - \beta_S^*) \le t,\ \hat{S} \ne S)\} \\ \le\ &2P(\hat{S} \ne S) \to 0,\ as\ n \to \infty. \end{aligned}$$

(2) Select+Ridge: again, conditioned on $\{\hat{S} = S\}$, the Select+Ridge and Select+mLS are respectively

$$\tilde{\beta}_{Select+Ridge,S} = V D_1 D^T U^T Y,$$

$$\tilde{\beta}_{Select+mLS,S} = (X_S^T X_S)^{-1} X_S^T Y = \frac{1}{\sqrt{n}} V (D^T D)^{-1} D^T U^T Y.$$

By simple calculation, the difference between these two estimators is

$$\begin{aligned} &||\sqrt{n}(\tilde{\beta}_{Select+Ridge,S} - \tilde{\beta}_{Select+mLS,S})||_2^2 \\ =\ &n \cdot \left|\left| V diag\left\{ \frac{-\sqrt{n}\mu_n}{n\lambda_1^2(n\lambda_1^2 + \mu_n)}, \ldots, \frac{-\sqrt{n}\mu_n}{n\lambda_s^2(n\lambda_s^2 + \mu_n)} \right\} D^T U^T Y \right|\right|_2^2 \\ \le\ &\max_{1 \le i \le s} \left\{ \frac{n\mu_n^2}{n\lambda_i^2(n\lambda_i^2 + \mu_n)^2} \right\} ||Y||_2^2 \\ \le\ &\frac{\mu_n^2}{n^2 \Lambda_{min}^3} ||Y||_2^2 = O_p(\mu_n^2) \end{aligned} \tag{46}$$

where the last equality comes from assumption (g) and $E||Y||_2^2 = ||X\beta^*||_2^2 + n\sigma^2 = O(n^2)$. Therefore, Select+Ridge has the same asymptotic distribution as Select+mLS, which completes the proof. □

In the following, we will prove the validity of residual bootstrap after Select+mLS. The proof for residual bootstrap after Select+Ridge is omitted since the techniques are almost the same.

Firstly, we re-characterize the conditional distribution of bootstrap error terms $\epsilon_i^*, i = 1, \ldots, n$ as follows:

Lemma 4. *Suppose that assumptions (a)–(e), (h) and (dd) are satisfied and that $P(\hat{S} \ne S) \to 0$, then with probability converging to 1, $\epsilon_i^*, i = 1, \ldots, n$ are conditionally i.i.d. subgaussian random variables. That is, there exists constant C^*, $c^* > 0$ such that*

$$P^*(|\epsilon_i^*| \ge t) \le C^* e^{-c^* t^2},\ \forall t \ge 0 \tag{47}$$

holds in probability.

Proof. Note that $P^*(|\epsilon_i^*| \geq t) = \frac{1}{n}\sum_{i=1}^n \mathbb{I}_{|\hat{\epsilon}_i - \hat{\mu}| \geq t}$, hence (47) is equivalent to

$$\sup_{t \geq 0} \left\{ \frac{1}{n} \sum_{i=1}^n e^{c^* t^2} \mathbb{I}_{|\hat{\epsilon}_i - \hat{\mu}| \geq t} \right\} \leq C^*. \tag{48}$$

We know that

$$\begin{aligned} \hat{\epsilon}_i - \hat{\mu} &= y_i - x_i^T \tilde{\beta} - (\overline{y} - \overline{x}^T \tilde{\beta}) \\ &= x_i^T \beta^* + \epsilon_i - x_i^T \tilde{\beta} - (\overline{x}^T \beta^* + \overline{\epsilon} - \overline{x}^T \tilde{\beta}) \\ &= x_i^T (\beta^* - \tilde{\beta}) + \epsilon_i - \overline{\epsilon} \end{aligned}$$

where x_i^T is the i-th row of X, $\overline{y} = \frac{1}{n}\sum_{i=1}^n y_i$,$\overline{\epsilon} = \frac{1}{n}\sum_{i=1}^n \epsilon_i$ and $\overline{x} = \frac{1}{n}\sum_{i=1}^n x_i = 0$. It is easy to see that $sup_{t\geq 0}\{\frac{1}{n}\sum_{i=1}^n e^{c^* t^2}\mathbb{I}_{|\hat{\epsilon}_i - \hat{\mu}| \geq t}\}$ can be bounded by

$$\frac{1}{n}\sum_{i=1}^n \left\{ \sup_{t\geq 0}\{e^{c^* t^2}\mathbb{I}_{|x_i^T(\beta^* - \tilde{\beta})| \geq t/3}\} + \sup_{t \geq 0}\{e^{c^* t^2}\mathbb{I}_{|\overline{\epsilon}| \geq t/3}\} + \sup_{t\geq 0}\{e^{c^* t^2}\mathbb{I}_{|\epsilon_i| \geq t/3}\} \right\}. \tag{49}$$

For the first term in (49), let $x_{i,S} = (x_{i1}, \ldots, x_{is})^T$,

$$\begin{aligned} & P(\max_{1\leq i\leq n} |x_i^T(\beta^* - \tilde{\beta})| \geq 1/3) \\ = \; & P(\max_{1\leq i\leq n} |x_i^T(\beta^* - \tilde{\beta})| \geq 1/3, \hat{S} = S) \\ & + P(\max_{1\leq i\leq n} |x_i^T(\beta^* - \tilde{\beta})| \geq 1/3, \hat{S} \neq S) \\ \leq \; & P(\max_{1\leq i\leq n} |x_{i,S}^T (X_S^T X_S)^{-1} X_S^T \epsilon| \geq 1/3) + P(\hat{S} \neq S). \end{aligned} \tag{50}$$

By Markov inequality,

$$P(\max_{1\leq i\leq n} |x_{i,S}^T (X_S^T X_S)^{-1} X_S^T \epsilon| \geq 1/3) \leq 9E \max_{1\leq i\leq n} |x_{i,S}^T (X_S^T X_S)^{-1} X_S^T \epsilon|^2.$$

Note that

$$\max_{1\leq i\leq n} |x_{i,S}^T (X_S^T X_S)^{-1} X_S^T \epsilon|^2 \leq \max_{1\leq i\leq n} ||x_{i,S}||^2 \cdot ||(X_S^T X_S)^{-1} X_S^T \epsilon||^2,$$

therefore

$$P(\max_{1\leq i\leq n} |x_{i,S}^T (X_S^T X_S)^{-1} X_S^T \epsilon| \geq 1/3) \leq 9 \max_{1\leq i\leq n} ||x_{i,S}||^2 \cdot E||(X_S^T X_S)^{-1} X_S^T \epsilon||^2.$$

Because $\epsilon \sim N(0, \sigma^2 I)$, hence $(X_S^T X_S)^{-1} X_S^T \epsilon \sim N(0, \sigma^2 (X_S^T X_S)^{-1})$. Then,

$$E||(X_S^T X_S)^{-1} X_S^T \epsilon||^2 = \sigma^2 tr(X_S^T X_S)^{-1} = \sigma^2 \frac{1}{n} tr(C_{11}^{-1}) \leq \frac{\sigma^2}{\Lambda_{min}} \frac{s}{n},$$

hence

$$P(\max_{1\leq i\leq n} |x_{i,S}^T (X_S^T X_S)^{-1} X_S^T \epsilon| \geq 1/3) \leq 9 \frac{\sigma^2}{\Lambda_{min}} \frac{s}{n} \max_{1\leq i\leq n} ||x_{i,S}||^2 \to 0$$

where "$\to 0$" comes from the assumptions $s^2/n \to 0$ and $max_{1\le i\le n}||x_{i,S}||^2 = max_{1\le i\le n}\sum_{j=1}^{s} x_{ij}^2 = o(n^{\frac{1}{2}})$. Connect with $P(\hat{S} \neq S) \to 0$, we have

$$P(\max_{1\le i\le n} |x_i^T(\beta^* - \tilde{\beta})| \ge 1/3) \to 0,$$

hence

$$P\left(\frac{1}{n}\sum_{i=1}^{n} sup_{t\ge 1}\{e^{c^*t^2}\mathbb{I}_{|x_i^T(\beta^*-\tilde{\beta})|\ge t/3}\} \le e^{c^*}\right) \ge P(\max_{1\le i\le n} |x_i^T(\beta^* - \tilde{\beta})| < 1/3) \to 1.$$

The inequality holds since it is easy to show that

$$\left\{\frac{1}{n}\sum_{i=1}^{n} sup_{t\ge 1}\{e^{c^*t^2}\mathbb{I}_{|x_i^T(\beta^*-\tilde{\beta})|\ge t/3}\} \le e^{c^*}\right\} \supseteq \{\max_{1\le i\le n} |x_i^T(\beta^* - \tilde{\beta})| < 1/3\}.$$

It is clear that

$$\frac{1}{n}\sum_{i=1}^{n} \sup_{0\le t\le 1}\{e^{c^*t^2}\mathbb{I}_{|x_i^T(\beta^*-\tilde{\beta})|\ge t/3}\} \le e^{c^*},$$

therefore, with probability going to 1, we have

$$\begin{aligned} & \frac{1}{n}\sum_{i=1}^{n} sup_{t\ge 0}\{e^{c^*t^2}\mathbb{I}_{|x_i^T(\beta^*-\tilde{\beta})|\ge t/3}\} \\ = \quad & \max\left(\frac{1}{n}\sum_{i=1}^{n} \sup_{0\le t\le 1}\{e^{c^*t^2}\mathbb{I}_{|x_i^T(\beta^*-\tilde{\beta})|\ge t/3}\},\ \frac{1}{n}\sum_{i=1}^{n} sup_{t\ge 1}\{e^{c^*t^2}\mathbb{I}_{|x_i^T(\beta^*-\tilde{\beta})|\ge t/3}\}\right) \\ \le \quad & e^{c^*}. \end{aligned} \tag{51}$$

For the second term in (49), by strong law of large numbers, we have $\overline{\epsilon} \to 0,\ a.s.$, then

$$P(|\overline{\epsilon}| \ge 1/3) \to 0.$$

It is easy to show that

$$\{sup_{t\ge 1}\{e^{c^*t^2}\mathbb{I}_{|\overline{\epsilon}|\ge t/3}\} \supseteq \{|\overline{\epsilon}| < 1/3\}.$$

Hence

$$P(sup_{t\ge 1}\{e^{c^*t^2}\mathbb{I}_{|\overline{\epsilon}|\ge t/3}\} \le e^{c^*}) \ge P(|\overline{\epsilon}| < 1/3) \to 1. \tag{52}$$

Using the same trick as (51), we have

$$\frac{1}{n}\sum_{i=1}^{n} sup_{t\ge 0}\{e^{c^*t^2}\mathbb{I}_{|\overline{\epsilon}|\ge t/3}\} = \max(\sup_{0\le t\le 1}\{e^{c^*t^2}\mathbb{I}_{|\overline{\epsilon}|\ge t/3}\},\ sup_{t\ge 1}\{e^{c^*t^2}\mathbb{I}_{|\overline{\epsilon}|\ge t/3}\}) \le e^{c^*} \tag{53}$$

holds in probability.

For the third term in (49), we will show that if $c^* = \frac{1}{36\sigma^2}$

$$E \sup_{t\geq 0}\{e^{c^*t^2}\mathbb{I}_{|\epsilon_1|\geq t/3}\} = \int_0^\infty P(\sup_{t\geq 0}\{e^{c^*t^2}\mathbb{I}_{|\epsilon_1|\geq t/3}\} > u)d_u < \infty. \tag{54}$$

The above integral can be divided into two parts $[0, e^{9\sigma^2c^*}]$ and $[e^{9\sigma^2c^*}, \infty]$. And the first part is bounded using $P(sup_{t\geq 0}\{e^{c^*t^2}\mathbb{I}_{|\epsilon_1|\geq t/3}\} > u) \leq 1$ that

$$\int_0^{e^{9\sigma^2c^*}} P(\sup_{t\geq 0}\{e^{c^*t^2}\mathbb{I}_{|\epsilon_1|\geq t/3}\} > u)d_u \leq e^{9\sigma^2c^*}. \tag{55}$$

For the second part, we have

$$\begin{aligned}
& P(\sup_{t\geq 0}\{e^{c^*t^2}\mathbb{I}_{|\epsilon_1|\geq t/3}\} > u) \\
= \;& P\left(\sup_{t\geq 0}\{e^{c^*t^2}\mathbb{I}_{|\epsilon_1|\geq t/3}\} > u, |\epsilon_1| < \sqrt{\frac{\log u}{c^*}}/3\right) \\
& + P\left(\sup_{t\geq 0}\{e^{c^*t^2}\mathbb{I}_{|\epsilon_1|\geq t/3}\} > u, |\epsilon_1| \geq \sqrt{\frac{\log u}{c^*}}/3\right) \\
\leq \;& P\left(\sup_{t\geq 0}\{e^{c^*t^2}\mathbb{I}_{|\epsilon_1|\geq t/3}\} > u, |\epsilon_1| < \sqrt{\frac{\log u}{c^*}}/3\right) + P\left(|\epsilon_1| \geq \sqrt{\frac{\log u}{c^*}}/3\right).
\end{aligned}$$

On $\{|\epsilon_1| < \sqrt{\frac{\log u}{c^*}}/3\}$, we have

$$e^{c^*t^2}\mathbb{I}_{|\epsilon_1|\geq t/3}\begin{cases} \leq u & if\ t \leq \sqrt{\frac{\log u}{c^*}} \\ 0 & otherwise \end{cases} \tag{56}$$

Hence on $\{|\epsilon_1| < \sqrt{\frac{\log u}{c^*}}/3\}$,

$$\sup_{t\geq 0}\{e^{c^*t^2}\mathbb{I}_{|\epsilon_1|\geq t/3}\} \leq u,$$

then

$$P\left(\sup_{t\geq 0}\{e^{c^*t^2}\mathbb{I}_{|\epsilon_1|\geq t/3}\} > u, |\epsilon_1| < \sqrt{\frac{\log u}{c^*}}/3\right) = 0.$$

Therefore,

$$P(\sup_{t\geq 0}\{e^{c^*t^2}\mathbb{I}_{|\epsilon_1|\geq t/3}\} > u) \leq P\left(|\epsilon_1| \geq \sqrt{\frac{\log u}{c^*}}/3\right) \leq 2e^{-\frac{1}{2\sigma^2}\frac{\log u}{9c^*}} = \frac{2}{u^2} \tag{57}$$

where the second inequality comes from the Gaussian tail bound,

$$P(|\epsilon_1| \geq t) \leq 2e^{-\frac{t^2}{2\sigma^2}} \ \forall t \geq \sigma.$$

Therefore

$$\int_{e^{9\sigma^2 c^*}}^{\infty} P(\sup_{t\geq 0}\{e^{c^* t^2}\mathbb{I}_{|\epsilon_1|\geq t/3}\} > u)d_u \leq \int_{e^{9\sigma^2 c^*}}^{\infty} \frac{2}{u^2} d_u < \infty. \tag{58}$$

Combine (55) and (58), we obtain (54).

Again, by strong law of large numbers,

$$\frac{1}{n}\sum_{i=1}^{n} \sup_{t\geq 0}\{e^{c^* t^2}\mathbb{I}_{|\epsilon_i|\geq t/3}\} \to E \sup_{t\geq 0}\{e^{c^* t^2}\mathbb{I}_{|\epsilon_i|\geq t/3}\} \ a.s.,$$

hence,

$$\frac{1}{n}\sum_{i=1}^{n} \sup_{t\geq 0}\{e^{c^* t^2}\mathbb{I}_{|\epsilon_i|\geq t/3}\} \leq 2E \sup_{t\geq 0}\{e^{c^* t^2}\mathbb{I}_{|\epsilon_i|\geq t/3}\} \ \textit{holds in probability.}$$

If we chose $C^* = (2e^{c^*} + 2E\, sup_{t\geq 0}\{e^{c^* t^2}\mathbb{I}_{|\epsilon_i|\geq t/3}\})$ and $c^* = \frac{1}{36\sigma^2}$, then (47) holds in probability, which verifies the claim. □

Secondly, we need the following Lemma 5, which provides upper bounds of mean squared error (MSE) of Select+mLS $\tilde{\beta}^*$ based on the resample (X, Y^*). Let $\sigma_*^2 = C^*/c^*$.

Lemma 5. *Under assumptions (a)–(h) and (dd), the following hold*

$$E^*||\tilde{\beta}^* - \tilde{\beta}||_2^2 \mathbb{I}_{\hat{S}^*=S} = O_p\left(\frac{\sigma_*^2}{n} tr(C_{11}^{-1})\right),$$

$$E^*||\tilde{\beta}^* - \tilde{\beta}||_2^2 \mathbb{I}_{\hat{S}^*\neq S} = o_p(e^{-n^{c_2}/4})$$

where E^* denotes the conditional expectation given the error variables $\{\epsilon_i, i = 1, \ldots, n\}$.

Proof. By Lemma 4, we have shown that with probability going to 1, ϵ_i^* are i.i.d. subgaussian variables, i.e.,

$$P^*(|\epsilon_i^*| \geq t) \leq C^* e^{-c^* t^2}, \ \forall t \geq 0.$$

Simple calculation yields $E^*(\epsilon_1^*)^2 \leq \sigma_*^2$ and $E^*(\epsilon_1^*)^4 \leq \sigma_*^4$. Conditioned on $\{\hat{S} = S\}$, we have

$$||\tilde{\beta}^* - \tilde{\beta}||_2^2 \mathbb{I}_{\hat{S}^*=S} = ||(X_S^T X_S)^{-1} X_S^T Y^* - \tilde{\beta}||_2^2 \mathbb{I}_{\hat{S}^*=S} = ||(X_S^T X_S)^{-1} X_S^T \epsilon^*||_2^2 \mathbb{I}_{\hat{S}^*=S}.$$

Take expectation, then

$$E^*||\tilde{\beta}^* - \tilde{\beta}||_2^2 \mathbb{I}_{\hat{S}^*=S} \leq E^*||(X_S^T X_S)^{-1} X_S^T \epsilon^*||_2^2 \leq \frac{\sigma_*^2}{n} tr(C_{11}^{-1}).$$

Since $P(\hat{S} = S) \to 1$, we obtain

$$E^*||\tilde{\beta}^* - \tilde{\beta}||_2^2 \mathbb{I}_{\hat{S}^* = S} = O_p\left(\frac{\sigma_*^2}{n} tr(C_{11}^{-1})\right).$$

For the second part, we also conditioned on $\{\hat{S} = S\}$. Using the same procedure as proving Theorem 1, we can get

$$E^*||\tilde{\beta}^* - \tilde{\beta}||_2^2 \mathbb{I}_{\hat{S}^* \neq S} \leq 8\sqrt{P^*(\hat{S}^* \neq \hat{S})}\left\{||\tilde{\beta}||_2^2 + \frac{1}{\tau_n^2}\frac{1}{n}||X\tilde{\beta}||_2^2 + \frac{1}{\tau_n^2}\sigma_*^2\right\}. \quad (59)$$

By Theorem 1,

$$E||\tilde{\beta} - \beta^*||_2^2 \leq \frac{\sigma^2}{n} tr(C_{11}^{-1}) + 8\sqrt{P(\hat{S} \neq S)}\left\{||\beta^*||_2^2 + \frac{1}{\tau_n^2}\frac{1}{n}||X\beta^*||_2^2 + \frac{1}{\tau_n^2}\sigma^2\right\}. \quad (60)$$

As shown in Corollary 1, we have

$$E||\tilde{\beta} - \beta^*||_2^2 = O\left(\frac{\sigma^2}{\Lambda_{min}}\frac{s}{n}\right),$$

then $||\tilde{\beta} - \beta^*||_2^2 \to_p 0$ and therefore $||\tilde{\beta}||_2^2 = ||\beta^*||_2^2 + o_p(1)$. Moreover,

$$\frac{1}{n}||X\tilde{\beta} - X\beta^*||_2^2 = \frac{1}{n}||X_S(X_S^T X_S)^{-1} X_S^T \epsilon||_2^2 \to_p 0 \quad (61)$$

where "$\to_p 0$" comes from

$$E\frac{1}{n}||X_S(X_S^T X_S)^{-1} X_S^T \epsilon||_2^2 = \frac{s}{n}\sigma^2 \to 0,$$

therefore $\frac{1}{n}||X\tilde{\beta}||_2^2 = \frac{1}{n}||X\beta^*||_2^2 + o_p(1)$. Taking these results back to (59) and combining assumption (f), we have

$$\begin{aligned} E^*||\tilde{\beta}^* - \tilde{\beta}||_2^2 \mathbb{I}_{\hat{S}^* \neq \hat{S}} &\leq 8o_p(e^{-n^{c_2}/2})\left\{||\beta^*||_2^2 + \frac{1}{\tau_n^2}\frac{1}{n}||X\beta^*||_2^2 + \frac{1}{\tau_n^2}\sigma_*^2 + o_p(1)\right\} \\ &= o_p(e^{-n^{c_2}/4}) \end{aligned}$$

where the last equality holds since we suppose that $\tau_n \propto \frac{1}{n}$ and that

$$\frac{1}{n}||X\beta^*||_2^2 = O(n).$$

□

Finally, we can prove Theorem 4 now.

Proof of Theorem 4. Using the same notations as [7], let F be the true distribution of ε_i; let F_n be the empirical distribution of $\epsilon_1, \ldots, \epsilon_n$; let $\tilde{F}_n$ be the empirical distribution of the residuals $\hat{\epsilon}_1, \ldots, \hat{\epsilon}_n$; and let $\hat{F}_n$ be $\tilde{F}_n$ centered at its mean $\hat{\mu}$. We first show that the Mallows metric of G_n and G_n^* can be bounded by $\sqrt{s} \cdot d(F, \hat{F}_n)$.

Lemma 6. *Suppose that conditions (a)–(h) and (dd) are satisfied, then*

$$d^2(G_n, G_n^*) \le 4tr\left\{\left(\frac{1}{n}X_S^T X_S\right)^{-1}\right\} d^2(F, \hat{F}_n) + o_p(1) \le \frac{4s}{\Lambda_{min}} d^2(F, \hat{F}_n) + o_p(1).$$

Proof. By assumption (f), there exists a set A_n be such that $P(A_n) \to 1$ and for every $\omega \in A_n$,

$$P^*(\hat{S}^* \neq \hat{S}) = o(e^{-n^{c_2}}).$$

Fix $\omega \in A_n \bigcap \{\hat{S} = S\}$. By definition of Mallows metric, we have

$$\begin{aligned} d^2(G_n, G_n^*) &= \inf_{\epsilon \sim F, \epsilon^* \sim \hat{F}_n} E||T_n - T_n^*||_2^2 \\ &= \inf_{\epsilon \sim F, \epsilon^* \sim \hat{F}_n} E||\sqrt{n}(\tilde{\beta} - \beta^*) - \sqrt{n}(\tilde{\beta}^*(\omega) - \tilde{\beta}(\omega))||_2^2. \end{aligned}$$

By Lemma 8.1 in [7], the infimum in Mallows metric can be obtained. Then we can choose pairs $\{\epsilon_i \sim F, \epsilon_i^* \sim \hat{F}_n, i = 1, \ldots, n\}$ which are independent and $E(\epsilon_i - \epsilon_i^*)^2 = d^2(F, \hat{F}_n)$. Now, Let $A = \{\hat{S} = S\}$ and $A^* = \{\hat{S}^* = S\}$, straightforward computation and triangle inequality yield

$$\begin{aligned} & E||\sqrt{n}(\tilde{\beta} - \beta^*) - \sqrt{n}(\tilde{\beta}^*(\omega) - \tilde{\beta}(\omega))||_2^2 \\ = \; & E||\sqrt{n}(\tilde{\beta} - \beta^*)\mathbb{I}_A + \sqrt{n}(\tilde{\beta} - \beta^*)\mathbb{I}_{A^c} - \sqrt{n}(\tilde{\beta}^*(\omega) - \tilde{\beta}(\omega))\mathbb{I}_{A^*} \\ & - \sqrt{n}(\beta^*(\omega) - \tilde{\beta}(\omega))\mathbb{I}_{(A^*)^c}||_2^2 \\ \le \; & 2E||\sqrt{n}(\tilde{\beta} - \beta^*)\mathbb{I}_A - \sqrt{n}(\tilde{\beta}^*(\omega) - \tilde{\beta}(\omega))\mathbb{I}_{A^*}||_2^2 + 4E||\sqrt{n}(\tilde{\beta} - \beta^*)\mathbb{I}_{A^c}||_2^2 \\ & + 4E^*||\sqrt{n}(\tilde{\beta}^*(\omega) - \tilde{\beta}(\omega))\mathbb{I}_{(A^*)^c}||_2^2 \\ = \; & 2E||\sqrt{n}(X_S^T X_S)^{-1} X_S^T \epsilon \mathbb{I}_A - \sqrt{n}(X_S^T X_S)^{-1} X_S^T \epsilon^* \mathbb{I}_{A^*}||_2^2 \\ & + 4E||\sqrt{n}(\tilde{\beta} - \beta^*)\mathbb{I}_{A^c}||_2^2 + 4E^*||\sqrt{n}(\tilde{\beta}^*(\omega) - \tilde{\beta}(\omega))\mathbb{I}_{(A^*)^c}||_2^2 \end{aligned}$$

where E^* is the conditional expectation over ϵ^* given ω.

From the proof of Theorem 1, we have

$$E||\sqrt{n}(\tilde{\beta} - \beta^*)\mathbb{I}_{A^c}||_2^2 \le 8n\sqrt{P(\hat{S} \neq S)}\left\{||\beta^*||_2^2 + \frac{1}{\tau_n^2}\frac{1}{n}||X\beta^*||_2^2 + \frac{1}{\tau_n^2}\sigma^2\right\} \to 0.$$

By Lemma 5, we have

$$E^*||\sqrt{n}\beta^*(\omega) - \tilde{\beta}(\omega))\mathbb{I}_{(A^*)^c}||_2^2 = o_p(1),$$

therefore

$$\begin{aligned} & E||\sqrt{n}(\tilde{\beta} - \beta^*) - \sqrt{n}(\tilde{\beta}^*(\omega) - \tilde{\beta}(\omega))||_2^2 \\ \le \; & 2E||\sqrt{n}(X_S^T X_S)^{-1} X_S^T \epsilon \mathbb{I}_A - \sqrt{n}(X_S^T X_S)^{-1} X_S^T \epsilon^* \mathbb{I}_{A^*}||_2^2 + o_p(1) \\ \le \; & 4E||\sqrt{n}(X_S^T X_S)^{-1} X_S^T \epsilon - \sqrt{n}(X_S^T X_S)^{-1} X_S^T \epsilon^*||_2^2 \\ & + 4E||\sqrt{n}(X_S^T X_S)^{-1} X_S^T \epsilon \mathbb{I}_{A^c} - \sqrt{n}(X_S^T X_S)^{-1} X_S^T \epsilon^* \mathbb{I}_{(A^*)^c}||_2^2 + o_p(1) \\ \le \; & 4E||\sqrt{n}(X_S^T X_S)^{-1} X_S^T \epsilon - \sqrt{n}(X_S^T X_S)^{-1} X_S^T \epsilon^*||_2^2 \end{aligned}$$

$$+ 8E||\sqrt{n}(X_S^TX_S)^{-1}X_S^T\epsilon\mathbb{I}_{A^c}||_2^2 + 8E^*||\sqrt{n}(X_S^TX_S)^{-1}X_S^T\epsilon^*\mathbb{I}_{(A^*)^c}||_2^2 + o_p(1). \tag{62}$$

Next, we will bound the first three parts in the last inequality respectively. By straightforward computation, we have

$$\begin{aligned} & E||\sqrt{n}(X_S^TX_S)^{-1}X_S^T\epsilon - \sqrt{n}(X_S^TX_S)^{-1}X_S^T\epsilon^*||_2^2 \\ = \;& Etr\left\{(\epsilon-\epsilon^*)^T(\sqrt{n}(X_S^TX_S)^{-1}X_S^T)^T(\sqrt{n}(X_S^TX_S)^{-1}X_S^T)(\epsilon-\epsilon^*)\right\} \\ = \;& tr\left\{(nX_S(X_S^TX_S)^{-2}X_S^TE(\epsilon-\epsilon^*)(\epsilon-\epsilon^*)^T)\right\} \\ = \;& d^2(F,\hat{F}_n)tr\left\{nX_S(X_S^TX_S)^{-2}X_S^T\right\} \\ = \;& d^2(F,\hat{F}_n)tr\left\{\left(\frac{1}{n}X_S^TX_S\right)^{-1}\right\}. \end{aligned} \tag{63}$$

The penultimate equality is because $E(\epsilon-\epsilon^*)(\epsilon-\epsilon^*)^T = d^2(F,\hat{F}_n)I$. Then we only need to show that

$$E||\sqrt{n}(X_S^TX_S)^{-1}X_S^T\epsilon\mathbb{I}_{A^c}||_2^2 = o(1), \tag{64}$$

$$E^*||\sqrt{n}(X_S^TX_S)^{-1}X_S^T\epsilon^*\mathbb{I}_{(A^*)^c}||_2^2 = o_p(1). \tag{65}$$

It is easy to see that

$$\begin{aligned} ||\sqrt{n}(X_S^TX_S)^{-1}X_S^T\epsilon||_2^2 &= n\epsilon^TX_S(X_S^TX_S)^{-2}X_S^T\epsilon \\ &\le \Lambda_{max}(nX_S(X_S^TX_S)^{-2}X_S^T)||\epsilon||_2^2 \\ &\le \Lambda_{max}\left(\left(\frac{1}{n}X_S^TX_S\right)^{-1}\right)||\epsilon||_2^2 \\ &\le \Lambda_{min}^{-1}||\epsilon||_2^2. \end{aligned}$$

By Cauchy-Schwarz inequality,

$$\begin{aligned} & E||\sqrt{n}(X_S^TX_S)^{-1}X_S^T\epsilon\mathbb{I}_{A^c}||_2^2 \\ \le \;& \sqrt{E||\sqrt{n}(X_S^TX_S)^{-1}X_S^T\epsilon||_2^4E\mathbb{I}_{A^c}} \\ \le \;& \sqrt{\Lambda_{min}^{-2}E||\epsilon||_2^4P(A^c)} \end{aligned}$$

In the proof of Theorem 1, we have shown that $E||\epsilon||_2^4 = O(n^2)$ (see (42)) and connect with $P(A^c) = P(\hat{S}\neq S) = o(e^{-n^{c_2}})$, we obtain (64). The proof of (65) is the same as (64), so we omit it.

Combine (63), (64), (65) and (62), we have

$$E||\sqrt{n}(\tilde{\beta}-\beta^*) - \sqrt{n}(\tilde{\beta}^*(\omega)-\tilde{\beta}(\omega))||_2^2 \le 4d^2(F,\hat{F}_n)tr\left\{\left(\frac{1}{n}X_S^TX_S\right)^{-1}\right\} + o_p(1).$$

Therefore, we can obtain Lemma 6 by

$$d^2(G_n,G_n^*) = \inf_{\epsilon\sim F,\epsilon^*\sim\hat{F}_n} E||\sqrt{n}(\tilde{\beta}-\beta^*) - \sqrt{n}(\tilde{\beta}^*(\omega)-\tilde{\beta}(\omega))||_2^2$$

$$\begin{aligned} &\leq E||\sqrt{n}(\tilde{\beta}-\beta^*)-\sqrt{n}(\tilde{\beta}^*(\omega)-\tilde{\beta}(\omega))||_2^2 \\ &\leq 4d^2(F,\hat{F}_n)tr\left\{\left(\frac{1}{n}X_S^TX_S\right)^{-1}\right\}+o_p(1) \\ &\leq \frac{4s}{\Lambda_{min}}d^2(F,\hat{F}_n)+o_p(1) \end{aligned}$$

where the last inequality holds because $(\frac{1}{n}X_S^TX_S)^{-1}$ is a s by s matrix and $\Lambda_{max}((\frac{1}{n}X_S^TX_S)^{-1}) \leq \Lambda_{min}^{-1}$. □

In order to prove Theorem 4, we only have to show that

$$\frac{s}{\Lambda_{min}}d^2(F,\hat{F}_n)=o_p(1).$$

Lemma 7. *Suppose that assumptions (a)–(c), (e) and (dd) are satisfied and that* $P(\hat{S}\neq S)\to 0$*, then*

$$sd^2(\tilde{F}_n,F_n)=o_p(1).$$

Proof. By definition,

$$d^2(\tilde{F}_n,F_n)\leq\frac{1}{n}\sum_{i=1}^n(\hat{\epsilon}_i-\epsilon_i)^2=\frac{1}{n}||\hat{\epsilon}-\epsilon||_2^2.$$

Since $\hat{\epsilon}=Y-X\tilde{\beta}$ and $\epsilon=Y-X\beta^*$, we have

$$\hat{\epsilon}-\epsilon=X(\beta^*-\tilde{\beta}).$$

Conditioned on $\{\hat{S}=S\}$,

$$\frac{s}{n}||\hat{\epsilon}-\epsilon||_2^2=\frac{s}{n}||X_S(X_S^TX_S)^{-1}X_S^T\epsilon||_2^2\to_p 0$$

because

$$E\frac{s}{n}||X_S(X_S^TX_S)^{-1}X_S^T\epsilon||_2^2=\frac{s}{n}tr\left\{X_S(X_S^TX_S)^{-1}X_S^T\right\}\sigma^2=\frac{s^2}{n}\sigma^2\to 0.$$

Therefore,

$$sd^2(\tilde{F}_n,F_n)=o_p(1).$$ □

Lemma 8. *Suppose that assumptions (a)–(c), (e) and (dd) are satisfied and that* $P(\hat{S}\neq S)\to 0$*, then*

$$sd^2(\hat{F}_n,F_n)=o_p(1).$$

Proof. Application of Lemma 8.8 in [7], shows that for random variables U and V with finite second moment,

$$d^2(U,V)=d^2(U-EU,V-EV)+||EU-EV||_2^2.$$

Therefore, if let $\overline{F_n}$ be the empirical distribution of $\epsilon_1, \ldots, \epsilon_n$ centered at its mean $\overline{\epsilon} = \frac{1}{n}\sum_{i=1}^n \epsilon_i$, we have

$$d^2(\hat{F}_n, F_n) = d^2(\hat{F}_n, \overline{F_n}) + \left(\frac{1}{n}\sum_{i=1}^n \epsilon_i\right)^2,$$

$$d^2(\tilde{F}_n, F_n) = d^2(\hat{F}_n, \overline{F_n}) + \left(\frac{1}{n}\sum_{i=1}^n \epsilon_i - \frac{1}{n}\sum_{i=1}^n \hat{\epsilon}_i\right)^2.$$

Connecting the above two equalities,

$$\begin{aligned} d^2(\hat{F}_n, F_n) &= d^2(\tilde{F}_n, F_n) + \left(\frac{1}{n}\sum_{i=1}^n \epsilon_i\right)^2 - \left(\frac{1}{n}\sum_{i=1}^n \{\hat{\epsilon}_i - \epsilon_i\}\right)^2 \\ &\le d^2(\tilde{F}_n, F_n) + \left(\frac{1}{n}\sum_{i=1}^n \epsilon_i\right)^2. \end{aligned}$$

Now use the fact $sE(\frac{1}{n}\sum_{i=1}^n \epsilon_i)^2 = \frac{s\sigma^2}{n} \to 0$ and the previous Lemma 7, we obtain Lemma 8. □

Since d is a metric,

$$\frac{1}{2}d^2(F, \hat{F}_n) \le d^2(F, F_n) + d^2(F_n, \hat{F}_n). \tag{66}$$

We still need to bound $d^2(F, F_n)$. Denote ϕ and Φ the density and distribution functions of standard normal distribution $N(0,1)$ respectively. To control $d^2(F, F_n)$, we use the following result obtained by [16]: let $\to_\omega$ denote weak convergence,

Lemma 9 (del Barrio et al. (2000)). *Let $\epsilon_i, i = 1, \ldots, n$ be a sequence of i.i.d. normal random variables with mean 0 and variance σ^2. Then*

$$n\left(\frac{d^2(F, F_n)}{\sigma^2} - a_n\right) \to_\omega -\frac{3}{2} + \sum_{j=3}^{\infty} \frac{Z_j^2 - 1}{j}$$

where $\{Z_j, j = 3, \ldots, \infty\}$ are a sequence of independent $N(0,1)$ random variables and

$$a_n = \frac{1}{n}\int_{\frac{1}{n+1}}^{\frac{n}{n+1}} \frac{t(1-t)}{[\phi(\Phi^{-1}(t))]^2} dt.$$

In fact we have (see [6])

$$\int_{\frac{1}{n}}^{1-\frac{1}{n}} \frac{t(1-t)}{[\phi(\Phi^{-1}(t))]^2} d_t = \log\log n + \log 2 + \gamma + o(1)$$

where $\gamma = \lim_{k\to\infty}(\sum_{i=1}^n j^{-1} - \log k)$ is Euler's constant. Then, we have

$$d^2(F, F_n) = O_p\left(\frac{1}{n}\right) + \sigma^2\frac{\log\log n}{n} = o_p\left(\frac{1}{s}\right),$$

which together with Lemma 6, Lemma 8 and inequality (66) complete the proof. □

Proof of Lemma 1. Let $\hat{F}_n$ be the empirical distribution of the centered residuals $\hat{\epsilon}_1 - \hat{\mu}, \ldots, \hat{\epsilon}_n - \hat{\mu}$. For the stared data (X, Y^*), we have

$$Y^* = X\tilde{\beta} + \epsilon^*,\ \epsilon_i^* \sim \hat{F}_n\ i.i.d.$$

We only need to verify: the stared version (replacing β^* and ϵ by $\tilde{\beta}$ and ϵ^* respectively) of conditions (a)–(d), (i), (j) and the Irrepresentable Condition (24) hold in probability. Given $\{\hat{S} = S\}$, these conditions have the following forms:

(*Irrepresentable Condition**): there exists a positive constant vector η, such that

$$|C_{21}C_{11}^{-1} sign(\tilde{\beta}_S)| \leq \mathbf{1} - \eta. \tag{67}$$

(a^*)[5]: ϵ_i^* are i.i.d. subgaussian random variables. That is, there exists constant C^*, $c^* > 0$ such that

$$P^*(|\epsilon_i^*| \geq t) \leq C^* e^{-c^* t^2},\ \forall t \geq 0.$$

(b^*): Suppose that the predictors are standardized, i.e.

$$\sum_{i=1}^{n} x_{ij} = 0\ and\ \frac{1}{n}\sum_{i=1}^{n} x_{ij}^2 = 1,\ j = 1, \ldots, p. \tag{68}$$

(c^*): there exists an constant $\Lambda_{min} > 0$ such that

$$\Lambda_{min}(C_{11}) \geq \Lambda_{min}. \tag{69}$$

(d^*): let $s^* = |\hat{S}|$, there exists $0 \leq c_1 < 1$ and $0 < c_2 < 1 - c_1$

$$s^* = s_n^* = O(n^{c_1}),\ p = p_n = O(e^{n^{c_2}}). \tag{70}$$

(i^*): there exists constant $c_1 + c_2 < c_3 \leq 1$ and $M > 0$ so that

$$n^{\frac{1-c_3}{2}} \min_{1 \leq i \leq s} |\tilde{\beta}_i| \geq M. \tag{71}$$

(j^*): $\lambda_n \propto n^{\frac{1+c_4}{2}}$ with $c_2 < c_4 < c_3 - c_1$.

Clearly, conditions (b^*), (c^*) and (j^*) hold because they are not relative to $\tilde{\beta}$ and ϵ^*. By Lemma 4, condition (a*) holds. (d^*) is satisfied since $s^* = |\hat{S}| = s$. By Corollary 2 (asymptotic normality), $\tilde{\beta}$ is $\sqrt{n}$-consistent, then condition (i^*) holds in probability. The sign-consistency of $\tilde{\beta}$ (Lemma 2) ensures condition (*Irrepresentable Condition**). □

[5] By Remark 2.1, subgaussian assumption of ϵ_i^* ensures model selection consistency of Lasso

References

[1] Adel, J. and Andrea, M. (2013). Model selection for high-dimensional regression under the generalized irrepresentability condition. http://arxiv.org/abs/1305.0355.

[2] Bach, F. (2008). Bolasso: Model consistent Lasso estimation through the bootstrap. In *Proc. 25th Int. Conf. Machine Learning*, 33–40.

[3] Belloni, A. and Chernozhukov, V. (2009). Least squares after model selection in high-dimensional sparse models. *Bernoulli* **19**, 521–547.

[4] Belloni, A., Chernozhukov, V. and Hansen, C. (2011). Inference for high-dimensional sparse econometric models. http://arxiv.org/abs/1201.0220.

[5] Belloni, A., Chernozhukov, V. and Hansen, C. (2011). Inference on treatment effects after selection amongst high-dimensional controls. http://arxiv.org/abs/1201.0224.

[6] Bickel, P. J. and van Zwet, W. R. (1978). Asymptotic expansions for the power of distribution free tests in the two-sample problem. *Annals of Statistics* **6**, 937–1004. MR0499567 (80j:62043)

[7] Bickel, P. J. and Freedman, D. A. (1981). Some asymptotic theory for the bootstrap. *Annals of Statistics* **9**, 1196–1217. MR0630103

[8] 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.

[9] Bickel, P. J., Ritov, Y. and Tsybakov A. (2009). Simultaneous analysis of Lasso and Dantzig selector. *Annals of Statistics* **37**, 1705–1732. MR2533469 (2010j:62118)

[10] Bunea, F. (2008). Honest variable selection in linear and logistic regression models via l_1 and $l_1 + l_2$ penalization. *Electronic Journal of Statistics* **2**, 1153–1194. MR2461898 (2010b:62143)

[11] Bunea, F., Tsybakov A. and Wegkamp, M. (2006). Sparsity oracle inequalities for the Lasso. *Electronic Journal of Statistics* **1**, 169–194. MR2312149 (2008h:62101)

[12] Candes, E. and Tao, T. (2007). The Dantzig selector: Statistical estimation when p is much larger than n. *Annals of Statistics* **35**, 2312–2351. MR2382651

[13] Chatterjee, A. and Lahiri, S. N. (2011). Bootstrapping Lasso estimators. *Journal of the American Statistical Association* **106**, 608–625. MR2847974 (2012i:62199)

[14] Chatterjee, A. and Lahiri, S. N. (2012). Rates of convergence of the adaptive Lasso estimators to the oracle distribution and higher order refinements by the bootstrap. *Annals of Statistics* (to appear).

[15] Davison, A. C. and Hinkley, D. V. (1997). *Bootstrap methods and their application*, Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press. MR1700749

[16] del Barrio, E., Cuesta-Albertos, J. and Matran, C. (2000). Contributions of empirical and quantile processes to the asymptotic theory of goodness-of-fit tests. *Test* **9**, 1–96. MR1790430 (2001h:62026)
[17] Donoho, D., Elad, M. and Temlyakov, V. (2006). Stable recovery of sparse overcomplete representations in the presence of noise. *IEEE Transactions on Information Theory* **52**, 6–18. MR2237332
[18] Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *Annals of Statistics* **7**, 1–26. MR0515681 (80b:62021)
[19] Efron, B., Hastie, T. and Tibshirani, R. (2004). Least angle regression. *Annals of Statistics* **32**, 407–499. MR2060166
[20] Efron, B. and Tibshirani, R. (1993). *An Introduction to the Bootstrap*. Boca Raton, FL: Chapman & Hall/CRC.
[21] Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association* **96**, 1348–1360. MR1946581 (2003k:62160)
[22] Fan, J. and Lv, J. (2008). Sure independence screening for ultra-high dimensional feature space. *Journal of the Royal Statistical Society, Series B* **70**, 849–911.
[23] Freedman, D. A. (1981). Bootstrapping regression models. *Annals of Statistics* **9**, 1218–1228. MR0630104
[24] Friedman, J., Hastie, T., Holfing, H., and Tibshirani, R. (2007). Pathwise coordinate optimization. *The Annals of Applied Statistics* **1**, 302–332. MR2415737
[25] Fuchs, J. J. (2005). Recovery of exact sparse representations in the presence of noise. *IEEE Transactions on Information Theory* **51**, 3601–3608.
[26] Gai, Y., Zhu, L. and Lin, L. (2013). Model selection consistency of Dantzig selector. *Statistica Sinica* **23**, 615–634.
[27] Greenshtein, E. and Ritov, Y. (2004). Persistence in high-dimensional linear predictor selection and the virtue of overparametrization. *Bernoulli* **10**, 971–988. MR2108039
[28] Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics* **12**, 55–67.
[29] Huang, J., Horowitz, J. and Ma, S. (2008). Asymptotic properties of bridge estimators in sparse high-dimensional regression models. *Annals of Statistics* **36**, 587–613. MR2396808 (2009g:62094)
[30] Huang, J., Ma, S. and Zhang, C.-H. (2008). Adaptive Lasso for sparse high-dimensional regression models. *Statistica Sinica* **18**, 1603–1618. MR2469326 (2010a:62214)
[31] Karim, L. (2008). Sup-norm convergence rate and sign concentration property of Lasso and Dantzig estimators. *Electronic Journal of Statistics* **2**, 90–102. MR2386087 (2009a:62287)
[32] Knight, K. and Fu, W. J. (2000). Asymptotics for Lasso-type estimators. *Annals of Statistics* **28**, 1356–1378. MR1805787 (2002a:62099)
[33] Leeb, H. and Pötscher, B. M. (2005). Model selection and inference: Facts and fiction. *Econometric Theory* **21**, 21–59. MR2153856

[34] Lv, J. and Fan, Y. (2009). A unified approach to model selection and sparse recovery using regularized least squares. *Annals of Statistics* **37**, 3498–3528. MR2549567 (2010m:62219)

[35] Massy, W. F. (1965). Principal components regression in exploratory statistical research. *Journal of the American Statistical Association* **60**, 234–256.

[36] Meinshausen, N. and Buhlmann, P. (2006). High dimensional graphs and variable selection with the Lasso. *Annals of Statistics* **34**, 1436–1462. MR2278363 (2008b:62044)

[37] Meinshausen, N. (2007). Relaxed Lasso. *Computational Statistics and Data Analysis* **52**, 374–393. MR2409990

[38] Meinshausen, N. and Buhlmann, P. (2010). Stability selection. *Journal of the Royal Statistical Society: Series B* **72**, 417–473. MR2758523

[39] Meinshausen, N. and Yu, B. (2009). Lasso-type recovery of sparse representations for high-dimensional data. *Annals of Statistics* **37**, 246–270. MR2488351 (2010e:62176)

[40] Minnier, J., Tian, L. and Cai, T. (2009). A perturbation method for inference on regularized regression estimates. *Journal of the American Statistical Association* **106**(496), 1371–1382. MR2896842

[41] Negahban, S., Ravikumar, P., Wainwright, M. J., and Yu, B. (2009). A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers. In *Advances in Neural Information Processing Systems* **22**, 1348–1356.

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

[43] Osborne, M. R., Presnell, B. and Turlach, B. A. (2000). On the Lasso and its dual. *Journal of Computational and Graphical Statistics* **9**, 319–337. MR1822089

[44] Pötscher, B. M. and Schneider, U. (2009). On the distribution of the adaptive LASSO estimator. *Journal of Statistical Planning and Inference* **139**, 2775–2790. MR2523666 (2010h:62084)

[45] Raskutti, G., Wainwright, M. J. and Yu, B. (2011). Minimax rates of estimation for high-dimensional linear regression over l_q-balls. *IEEE Transactions on Information Theory* **57**, 6976–6994. MR2882274 (2012k:62204)

[46] Sartori, S. (2011). *Penalized regression: Bootstrap confidence intervals and variable selection for high dimensional data sets*. PhD thesis, Universit degli Studi di Milano, 2011. Available online: `http://air.unimi.it/bitstream/2434/153099/6/phd_unimi_R07738.pdf`.

[47] Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society: Series B* **58**, 267–288. MR1379242

[48] Tropp, J. (2004). Greed is good: Algorithmic results for sparse approximation. *IEEE Transactions on Information Theory* **50**, 2231–2242. MR2097044

[49] Van de Geer, S. (2007). The deterministic Lasso. *Proc. of Joint Statistical Meeting*

[50] Van de Geer, S. (2007). High-dimensional generalized linear models and the Lasso. *Annals of Statistics* **36**, 614–645. MR2396809 (2009h:62048)

[51] Wainwright, M. (2009). Sharp thresholds for noisy and high-dimensional recovery of sparsity using l_1-constrained quadratic programming (Lasso). *IEEE Transactions on Information Theory* **55**, 2183–2202.

[52] Zhang, C.-H. and Huang J. (2008). The sparsity and bias of the Lasso selection in high-dimensional linear regression. *Annals of Statistics* **36**, 1567–1594. MR2435448 (2010h:62204)

[53] Zhang, C.-H. and Zhang, S. S. (2011). Confidence intervals for low-dimensional parameters in high-dimensional linear models. arxiv.org/abs/1110.2563.

[54] Zhao, P. and Yu, B. (2006). On model selection consistency of Lasso. *The Journal of Machine Learning Research* **7**, 2541–2563. MR2274449

[55] Zou, H. (2006). The adaptive Lasso and its oracle properties. *Journal of the American Statistical Association* **101**, 1418–1429. MR2279469 (2008d:62024)

[56] Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B* **67**, 301–320. MR2137327