Submitted to the Annals of Statistics

RIDGE REGRESSION REVISITED: DEBIASING, THRESHOLDING AND BOOTSTRAP

BY YUNYI ZHANG [1] AND DIMITRIS N. POLITIS [2]

[1]*Department of Mathematics, Univ. of California–San Diego, yuz334@ucsd.edu*

[2]*Department of Mathematics and Halicioglu Data Science Institute, Univ. of California–San Diego, dpolitis@ucsd.edu*

The success of the Lasso in the era of high-dimensional data can be attributed to its conducting an implicit model selection, i.e., zeroing out regression coefficients that are not significant. By contrast, classical ridge regression can not reveal a potential sparsity of parameters, and may also introduce a large bias under the high-dimensional setting. Nevertheless, recent work on the Lasso involves debiasing and thresholding, the latter in order to further enhance the model selection. As a consequence, ridge regression may be worth another look since –after debiasing and thresholding– it may offer some advantages over the Lasso, e.g., it can be easily computed using a closed-form expression. In this paper, we define a debiased and thresholded ridge regression method, and prove a consistency result and a Gaussian approximation theorem. We further introduce a wild bootstrap algorithm to construct confidence regions and perform hypothesis testing for a linear combination of parameters. In addition to estimation, we consider the problem of prediction, and present a novel, hybrid bootstrap algorithm tailored for prediction intervals. Extensive numerical simulations further show that the debiased and thresholded ridge regression has favorable finite sample performance and may be preferable in some settings.

1. Introduction. Linear regression is a fundamental topic in statistical inference. The classical setting assumes the dimension of parameters in a linear model is constant. However, in the modern era, observations may have a comparable or even larger dimension than the number of samples. To perform a consistent estimation with high-dimensional data, statisticians often assume the underlying parameters are sparse (i.e., the parameter vector contains lots of zeros), and proceed with statistical inference based on this assumption.

The success of the Lasso in the setting of high-dimensional data can be attributed to its conducting an implicit model selection, i.e., zeroing out regression coefficients that are not significant; see Tibshirani (1996). More recent work includes: Meinshausen and Bühlmann (2006), Meinshausen and Yu (2009), and van de Geer (2008) for the Lasso estimator's (model-selection) consistency and applications; Chatterjee and Lahiri (2010, 2011), Zhang and Cheng (2017), and Dezeure, Bühlmann and Zhang (2017) for confidence interval construction and hypothesis testing; and Javanmard and Montanari (2018), Fan and Li (2001), and Chen and Zhou (2020) for improvements of the Lasso estimator. Although the Lasso has the desirable property of zeroing out some regression coefficients, van de Geer, Bühlmann and Zhou (2011) proposed to further *threshold* the estimated coefficients, leading to a sparser fitted model. Furthermore, Bühlmann and van de Geer (2011), and Dezeure, Bühlmann and Zhang (2017), proposed to *debias* the Lasso in constructing confidence intervals; see van de Geer (2019) and Javanmard and Javadi (2019) for recent works on debiased Lasso.

MSC2020 subject classifications: 62J05, 62F40.

Keywords and phrases: Gaussian approximation, high-dimensional data, Lasso, prediction, regression, resampling.

An alternative approach providing consistent estimators for a high dimensional linear model is the so-called *post-selection inference*. It first applies Lasso to select influential parameters, then fits an ordinary least squares regression on the selected parameters; see Lee et al. (2016), Liu and Yu (2013), and Tibshirani et al. (2018). We refer to Bühlmann and van de Geer (2011) for a comprehensive overview of the Lasso method for high dimensional data.

Ridge regression is a classical method, and its estimator has a closed-form expression, making statistical inference easier than Lasso. However, there is relatively little research on the ridge regression under the high-dimensional setting. Shao and Deng (2012) proposed a threshold ridge regression method and proved its consistency. Dai et al. (2018) introduced a broken adaptive ridge estimator to approximate L_0 penalized regression. Dobriban and Wager (2018) derived the limit of high dimensional ridge regression's expected predictive risk. Bühlmann (2013) used Lasso to correct the bias in a ridge regression estimator, while Lopes (2014) applied a residual-based bootstrap to construct confidence intervals.

Three issues have prevented ridge regression from being suitable for a high dimensional linear model:

1. *The ridge regression cannot preserve/recover sparsity.* Typically, a ridge regression estimator of the parameter vector will not contain any zeros, even though the parameters may be sparse.

2. *Bias in the ridge regression estimator can be large.* To illustrate this, suppose the parameter of interest is $a^T\beta$ in a linear model $y = X\beta + \epsilon$; here, the dimension $p < n$ (the sample size), X has rank p, and a is a known vector. The ridge estimator is $a^T\widetilde{\theta}^\star$ with $\widetilde{\theta}^\star = (X^TX + \rho_n I_p)^{-1}X^Ty$, for some $\rho_n > 0$, with I_p denoting the p-dimensional identity matrix. Performing a thin singular value decomposition $X = P\Lambda Q^T$ (as in Theorem 7.3.2 in Horn and Johnson (2013)), and assuming the error vector ϵ consists of independent identically distributed (i.i.d.) components, the bias and the standard deviation can be calculated (and controlled) as follows:

(1)

$$\mathbf{E}a^T\widetilde{\theta}^\star - a^T\beta = -\rho_n a^T Q(\Lambda^2 + \rho_n I_p)^{-1}Q^T\beta \text{ which implies } |\mathbf{E}a^T\widetilde{\theta}^\star - a^T\beta| \le \frac{\rho_n\|a\|_2 \times \|\beta\|_2}{\lambda_p^2 + \rho_n}$$

$$\text{and } \sqrt{Var(a^T\widetilde{\theta}^\star)} = \sqrt{Var(\epsilon_1) \times a^TQ(\Lambda^2 + \rho_n I_p)^{-2}\Lambda^2 Q^T a} \le \frac{\sqrt{Var(\epsilon_1)} \times \|a\|_2}{\lambda_p}.$$

In the above, λ_p is the smallest singular value of X, and $\|\cdot\|_2$ is the Euclidean norm of a vector. If $\|\beta\|_2$ does not have a bounded order, the bias may tend to infinity. Another critical problem is that the absolute value of the bias can be significantly larger than the standard deviation, which makes constructing confidence intervals difficult.

3. *When the dimension of parameters is larger than the sample size, ridge regression estimates the projection of parameters on the linear space spanned by rows of X* (Shao and Deng (2012)). The projection (which can now be considered to be the 'parameters' of the linear model) is not sparse, bringing extra burdens for statistical inference.

The third issue comes from the nature of ridge regression, and it is not necessarily bad; our section 6 provides an example to illustrate this. The first two issues can be solved by *thresholding and debiasing* respectively, yielding an *improved* ridge regression that will be the focus of this paper. If the Lasso is in need of thresholding and debiasing –as van de Geer, Bühlmann and Zhou (2011), Dezeure, Bühlmann and Zhang (2017), and Bühlmann and van de Geer (2011) seem to suggest– then it loses some of its attractiveness, in which case (improved) ridge regression may be worth another look. If (improved) ridge regression turns out to have comparable performance to threshold Lasso, then the former would be preferable since it can be easily computed using a closed-form expression. Indeed, numerical simulations in section 6 indicate that improved ridge regression has favorable finite-sample

performance, and has a further advantage over the Lasso: it is *robust* against a non-optimal choice of the hyperparameters.

Apart from point estimation using improved ridge regression, this paper presents a Gaussian approximation theorem for the improved ridge regression estimator. Applying this result, we propose a wild bootstrap algorithm to construct a confidence region for $\gamma = M\beta$ with M a known matrix and/or test the null hypothesis $\gamma = \gamma_0$ with γ_0 a known vector, versus the alternative hypothesis $\gamma \neq \gamma_0$. The wild bootstrap was developed in the 1980s by Wu (1986) and Liu (1988); its applicability to high-dimensional problems was recognized early on by Mammen (1993). Here we will use the wild bootstrap in its Gaussian residuals version that has been found useful in high-dimensional regression; see Chernozhukov, Chetverikov and Kato (2013). Estimating and testing γ are important problems in econometrics, e.g., Dolado and Lütkepohl (1996), Sun (2011, 2013), and Gonçalves and Vogelsang (2011). Besides, estimating γ directly contributes to prediction, which is an important topic in modern age statistics.

Finally, we consider statistical prediction based on the improved ridge regression estimator for a high-dimensional linear model. For a regression problem, quantifying a predictor's accuracy can be as important as predicting accurately. To do that, it is useful to be able to construct a prediction interval to accompany the point prediction; this is usually done by some form of bootstrap; see Stine (1985) for a classical result, and Politis (2015) for a comprehensive treatment of both model-based and model-free prediction intervals in regression. As an alternative to the bootstrap, conformal prediction may be a tool to yield prediction intervals; see e.g. Romano, Patterson and Candès (2019) and Romano, Sesia and Candès (2020). In our point of view, however, the bootstrap is preferable as it captures the underlying variability of estimated quantities; Section 5 in what follows gives the details.

The remainder of this paper is organized as follows: Section 2 introduces frequently used notations and assumptions. Section 3 presents the consistency result and the Gaussian approximation theorem for the improved ridge regression estimator. Section 4 constructs a confidence region for $\gamma = M\beta$, and tests the null hypothesis $\gamma = \gamma_0$ versus the alternative hypothesis $\gamma \neq \gamma_0$ via a bootstrap algorithm. Section 5 constructs bootstrap prediction intervals in our ridge regression setting using a novel, hybrid resampling procedure. Finally, Section 6 provides extensive simulations to illustrate the finite sample performance, while Section 7 gives some concluding remarks; technical proofs are deferred to the supplement Zhang and Politis (2021a).

2. Preliminaries. Our work focuses on the fixed design linear model

$$y = X\beta + \epsilon \tag{2}$$

where the (unknown) parameter vector β is p-dimensional, and the $n \times p$ fixed (nonrandom) design matrix X is assumed to have rank r. The error vector ϵ has mean zero and satisfies assumptions to be specified later.

Define the known matrix of linear combination coefficients as $M = (m_{ij})_{i=1,2,...,p_1, j=1,2,...,p}$ so that M has p_1 rows. The linear combination of interest are $\gamma = (\gamma_1, ..., \gamma_{p_1})^T = M\beta$.

Perform a thin singular value decomposition $X = P\Lambda Q^T$ as in Theorem 7.3.2 in Horn and Johnson (2013); here, P and Q respectively is $n \times r$ and $p \times r$ orthonormal matrices that satisfy $P^TP = Q^TQ = I_r$, where I_r denotes the $r \times r$ identity matrix. Furthermore, $\Lambda = diag(\lambda_1, ..., \lambda_r)$, and $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_r > 0$ are positive singular values of X.

Denote $Q_\perp$ as the $p \times (p-r)$ orthonormal complement of Q; then we have

$$Q_\perp^T Q_\perp = I_{p-r},\ Q^T Q_\perp = 0,\ \text{and}\ QQ^T + Q_\perp Q_\perp^T = I_p; \tag{3}$$

in the above, 0 is the $r \times (p-r)$ matrix having all elements 0. Define $\zeta = Q^T\beta$ and $\theta = (\theta_1, ..., \theta_p)^T = Q\zeta = QQ^T\beta$, then $X\beta = X\theta$, $\theta^T\theta = \zeta^T Q^T Q\zeta = \zeta^T\zeta$. According to Shao and Deng (2012), the ridge regression estimates θ rather than β.

Define $\theta_\perp = Q_\perp Q_\perp^T\beta$, so $\beta = \theta + \theta_\perp$. If the design matrix X has rank $p \le n$, then $Q_\perp$ does not exist. In this situation, we define $\theta_\perp = 0$, the p dimensional vector with all elements 0. For a threshold b_n, define the set $\mathcal{N}_{b_n} = \{i \,||\theta_i| > b_n\}$. After selecting a suitable b_n, define

$$c_{ik} = \sum_{j\in\mathcal{N}_{b_n}} m_{ij}q_{jk},\ \forall\ i = 1, 2, ..., p_1,\ k = 1, 2, ..., r,\ \text{and}\ \mathcal{M} = \{i \mid \sum_{k=1}^{r} c_{ik}^2 > 0\} \tag{4}$$

Define $\tau_i,\ i = 1, 2, ..., p_1$ as

$$\tau_i = \sqrt{\sum_{k=1}^{r} c_{ik}^2 \left(\frac{\lambda_k}{\lambda_k^2+\rho_n} + \frac{\rho_n\lambda_k}{(\lambda_k^2+\rho_n)^2}\right)^2 + \frac{1}{n}} \tag{5}$$

In section 3 and (B.14) to (B.16)(in the supplement Zhang and Politis (2021a)), we show that the estimation error $\widehat{\gamma} - \gamma$ (see (17) for the definition of $\widehat{\gamma}$) asymptotically can be approximated by the random vector

$$\left(\sum_{l=1}^{n}\sum_{k=1}^{r} c_{1k}p_{lk}\left(\frac{\lambda_k}{\lambda_k^2+\rho_n} + \frac{\rho_n\lambda_k}{(\lambda_k^2+\rho_n)^2}\right)\epsilon_l, ..., \sum_{l=1}^{n}\sum_{k=1}^{r} c_{p_1k}p_{lk}\left(\frac{\lambda_k}{\lambda_k^2+\rho_n} + \frac{\rho_n\lambda_k}{(\lambda_k^2+\rho_n)^2}\right)\epsilon_l\right)^T \tag{6}$$

here $P = (p_{lk})_{l=1,...,n,k=1,...,r}$ and $\epsilon = (\epsilon_1, ..., \epsilon_n)^T$. Moreover, if we assume that $\epsilon_i, i = 1, ..., n$ are i.i.d. with mean 0 and variance 1, then

$$Var\left(\sum_{l=1}^{n}\sum_{k=1}^{r} c_{ik}p_{lk}\left(\frac{\lambda_k}{\lambda_k^2+\rho_n} + \frac{\rho_n\lambda_k}{(\lambda_k^2+\rho_n)^2}\right)\epsilon_l\right)$$

$$= \sum_{l=1}^{n}\left(\sum_{k=1}^{r} c_{ik}p_{lk}\left(\frac{\lambda_k}{\lambda_k^2+\rho_n} + \frac{\rho_n\lambda_k}{(\lambda_k^2+\rho_n)^2}\right)\right)^2 = \sum_{k=1}^{r} c_{ik}^2\left(\frac{\lambda_k}{\lambda_k^2+\rho_n} + \frac{\rho_n\lambda_k}{(\lambda_k^2+\rho_n)^2}\right)^2 \tag{7}$$

In section 3, we will estimate τ_i by $\widehat{\tau}_i$(defined in (25)) and use $\widehat{\tau}_i$ to normalize the estimation error. The extra $1/n$ in (5) is introduced to assure $\tau_i > 0$.

We will use the standard order notations $O(\cdot)$, $o(\cdot)$, $O_p(\cdot)$, and $o_p(\cdot)$. For two numerical sequences $a_n, b_n, n = 1, 2, ...$, we say $a_n = O(b_n)$ if $\exists$ a constant $C > 0$ such that $|a_n| \le C|b_n|$ for all n, and $a_n = o(b_n)$ if $\lim_{n\to\infty}\frac{a_n}{b_n} = 0$. For two random variable sequences X_n, Y_n, we say $X_n = O_p(Y_n)$ if for any $0 < \epsilon < 1$, $\exists$ a constant $C_\epsilon > 0$ such that $\sup_n Prob(|X_n| \ge C_\epsilon|Y_n|) \le \epsilon$; and $X_n = o_p(Y_n)$ if $\frac{X_n}{Y_n} \to_p 0$; see e.g. Definition 1.9 and Chapter 1.5.1 of Shao (2003). *All order notations and convergences in this paper will be understood to hold as the sample size* $n \to \infty$. For a vector $a = (a_1, ..., a_n)^T$ and a fixed number $q \ge 1$, define $\|a\|_q = (\sum_{i=1}^n |a_i|^q)^{1/q}$. For a finite set A, $|A|$ denotes the number of elements in A. Notations $\exists$ and $\forall$ denote "there exists" and "for all" respectively. $Prob^*(\cdot)$ and $\mathbf{E}^*\cdot$ respectively represent probability and expectation in the *"bootstrap world"*, i.e., they are the conditional probability $Prob(\cdot|y)$ and the conditional expectation $\mathbf{E}(\cdot|y)$.

Suppose $H(x)$ is a cumulative distribution function and $0 < \alpha < 1$; then the $1-\alpha$ quantile of H is defined as

$$c_{1-\alpha} = \inf\{x \in \mathbf{R} | H(x) \ge 1 - \alpha\}. \tag{8}$$

In particular, given some order statistics $X_1 \le X_2 \le ... \le X_B$, the $1-\alpha$ sample quantile $C_{1-\alpha}$ is defined as

$$C_{1-\alpha} = X_{i_*} \text{ such that } i_* = \min\left\{ i \;\Big|\; \frac{1}{B}\sum_{j=1}^{B} \mathbf{1}_{X_j \le X_i} \ge 1-\alpha \right\}. \tag{9}$$

Other notations will be defined before being used. Without being explicitly specified, the convergence results in this paper assume the sample size $n \to \infty$.

The high dimensionality in this paper comes from two aspects: the number of parameters p may increase with the sample size n, and (for statistical inference/hypothesis testing) the number of simultaneous linear combinations p_1 and $|\mathcal{M}|$ can also increase with n.

Our work adopts the following assumptions:

Assumptions

1. Assume a fixed design, i.e., the design matrix X is deterministic. Also assume that there exists constants $c_\lambda, C_\lambda > 0$, $0 < \eta \le 1/2$, such that the positive singular values of X satisfy

$$C_\lambda n^{1/2} \ge \lambda_1 \ge \lambda_2 \ge ... \ge \lambda_r \ge c_\lambda n^{\eta}. \tag{10}$$

Furthermore, the Euclidean norm of θ is assumed to satisfy $\|\theta\|_2 = \sqrt{\sum_{i=1}^{p} \theta_i^2} = O(n^{\alpha_\theta})$ with $0 < \alpha_\theta < 3\eta$.

2. The ridge parameter satisfies $\rho_n = O(n^{2\eta-\delta})$ with a positive constant δ such that $\frac{\eta+\alpha_\theta}{2} < \delta < 2\eta$

3. The errors $\epsilon = (\epsilon_1, ..., \epsilon_n)^T$ driving regression (2) are assumed to be i.i.d., with $\mathbf{E}\epsilon_1 = 0$, and $\mathbf{E}|\epsilon_1|^m < \infty$ for some $m > 4$.

4. The dimension of the parameter vector β satisfies $p = O(n^{\alpha_p})$ for some constant $\alpha_p \in [0, m\eta)$ where m, η are as defined in Assumptions 1–3. Furthermore, the threshold b_n is chosen as $b_n = C_b \times n^{-\nu_b}$ with constants $C_b, \nu_b > 0$ and $\nu_b + \frac{\alpha_p}{m} - \eta < 0$. We assume $\exists$ a constant $0 < c_b < 1$ such that $\max_{i \notin \mathcal{N}_{b_n}} |\theta_i| \le c_b \times b_n$, and $\min_{i \in \mathcal{N}_{b_n}} |\theta_i| \ge \frac{b_n}{c_b}$.

The intuitive meaning of assumption 4 is that the θ_is that are not being truncated should be significantly larger than the θ_i being truncated.

5. $\mathcal{M}$ (defined in (4)) is not empty and $|\mathcal{M}| = O(n^{\alpha_\mathcal{M}})$ with $\alpha_\mathcal{M} < m\eta$ where m, η are as defined in Assumptions 1–3. Besides, assume $\exists$ constants $c_\mathcal{M}, C_\mathcal{M}$ such that $0 < c_\mathcal{M} < \sum_{k=1}^{r} c_{ik}^2 \le C_\mathcal{M}$ for all $i \in \mathcal{M}$. Also assume

$$\max_{i=1,2,...,p_1} \Big| \sum_{j \notin \mathcal{N}_{b_n}} m_{ij}\theta_j \Big| = o\left(\frac{1}{\sqrt{n\log(n)}}\right) \text{ and } \max_{i=1,2,...,p_1} \Big| \sum_{j=1}^{p} m_{ij}\theta_{\perp,j} \Big| = o\left(\frac{1}{\sqrt{n\log(n)}}\right) \tag{11}$$

We assume (11) to maintain the sparsity of θ and assure that the projection bias $\beta - \theta$ is negligible compared to the stochastic errors. It allows an inexact sparsity, i.e., some θ_i may not equal 0 even if $i \notin \mathcal{N}_{b_n}$. Theoretical results for other linear regression estimators (e.g., Lasso) need an exact sparse assumption ($\theta_i = 0$ for all $i \notin \mathcal{N}_{b_n}$), see Zhao and Yu (2006) and Basu and Michailidis (2015) for a further introduction.

6. $\exists$ a constant α_σ satisfying $\eta \ge \alpha_\sigma > 0$ such that

$$n^{-\nu_b} \sum_{j \notin \mathcal{N}_{b_n}} |\theta_j| = O(n^{-\alpha_\sigma}), \quad \frac{\sqrt{|\mathcal{N}_{b_n}|}}{n^\eta} = O(n^{-\alpha_\sigma}) \tag{12}$$

7. $|\mathcal{M}| \le r$, the matrix $T = (c_{ik})_{i \in \mathcal{M}, k=1,2,...,r}$ has rank $|\mathcal{M}|$, and one of the two following conditions holds true:

7.1.

$$\max_{i\in\mathcal{M},l=1,2,...,n} |\frac{1}{\tau_i}\times\sum_{k=1}^{r} c_{ik}p_{lk}\left(\frac{\lambda_k}{\lambda_k^2+\rho_n}+\frac{\rho_n\lambda_k}{(\lambda_k^2+\rho_n)^2}\right)| \tag{13}$$
$$= o(\min(n^{(\alpha_\sigma-1)/2}\times\log^{-3/2}(n),\ n^{-1/3}\times\log^{-3/2}(n)))$$

7.2. $\alpha_\sigma < 1/2$ and

$$|\mathcal{M}| = o(n^{\alpha_\sigma}\times\log^{-3}(n))$$

$$\max_{i\in\mathcal{M},l=1,2,...,n} |\frac{1}{\tau_i}\times\sum_{k=1}^{r} c_{ik}p_{lk}\left(\frac{\lambda_k}{\lambda_k^2+\rho_n}+\frac{\rho_n\lambda_k}{(\lambda_k^2+\rho_n)^2}\right)| = O(n^{-\alpha_\sigma}\times\log^{-3/2}(n)) \tag{14}$$

According to (6), the normalized estimation error $\frac{\widehat{\gamma}_i-\gamma_i}{\widehat{\tau}_i}$ asymptotically will be approximated by $\sum_{l=1}^{n}\left(\frac{1}{\tau_i}\sum_{k=1}^{r} c_{ik}p_{lk}\left(\frac{\lambda_k}{\lambda_k^2+\rho_n}+\frac{\rho_n\lambda_k}{(\lambda_k^2+\rho_n)^2}\right)\right)\epsilon_l$. Therefore, the intuitive meaning of assumption 7 is that all terms $\frac{1}{\tau_i}\sum_{k=1}^{r} c_{ik}p_{lk}\left(\frac{\lambda_k}{\lambda_k^2+\rho_n}+\frac{\rho_n\lambda_k}{(\lambda_k^2+\rho_n)^2}\right)\epsilon_l$ in the summation are negligible and the number of simultaneous linear combinations $|\mathcal{M}|$ cannot be too large.

Recall that the paper at hand focuses on fixed design regression, i.e., no randomness involves in the design matrix X. However, all results of this paper still hold true in the case of random design after conditioning on X, as long as X can be assumed independent of the error vector ϵ. In this case, to intrepret the results we would need to replace $Prob(\cdot)$ by $Prob(\cdot|X)$, $\mathbf{E}\cdot$ by $\mathbf{E}\cdot|X$, $Prob^*(\cdot)$ by $Prob(\cdot|X,y)$ and $\mathbf{E}^*\cdot$ by $\mathbf{E}\cdot|X,y$.

REMARK 1. *We do not require that the design matrix has rank* $\min(n,p)$ *or that* $p<n$. *However, when these conditions are not satisfied, the sparsity of* θ*, i.e., assumption 5 and 6, can be violated. Section 6 uses a numerical simulation to illustrate this problem.*

Example 1 below provides an instance in which assumption 1 is satisfied.

EXAMPLE 1. *Suppose* $n>p$ *and* $\lim_{n\to\infty} p/n = c\in(0,1)$. *Choose* $X=(x_{ij})_{i=1,...,n,j=1,...,p}$ *such that the* x_{ij} *are a realization of i.i.d. random variables with mean* 0*, variance* 1*, and finite fourth order moment. According to Bai and Yin (1993), the smallest eigenvalue of* $\frac{1}{n}X^TX$ *would then converge to* $(1-\sqrt{c})^2$ *almost surely as* $n\to\infty$. *So the smallest singular value of* X *(which is the smallest eigenvalue of the square root of* X^TX*) is greater than* $\frac{1-\sqrt{c}}{2}\sqrt{n}$ *for sufficiently large* n*, almost surely. On the other hand, the largest eigenvalue of* $\frac{1}{n}X^TX$ *converges to* $(1+\sqrt{c})^2$ *as* $n\to\infty$. *Hence, the largest singular value of* X *also has order* $O(\sqrt{n})$ *almost surely.*

3. Consistency and the Gaussian approximation theorem. Throughout, we will use the notations developed in section 2. For a chosen ridge parameter $\rho_n>0$, define the classical ridge regression estimator $\widetilde{\theta}^\star$ and the de-biased estimator $\widetilde{\theta}$ as

$$\widetilde{\theta}^\star = (X^TX+\rho_nI_p)^{-1}X^Ty$$
$$\widetilde{\theta} = (\widetilde{\theta}_1,...,\widetilde{\theta}_p)^T = \widetilde{\theta}^\star+\rho_n\times Q(\Lambda^2+\rho_nI_r)^{-1}Q^T\widetilde{\theta}^\star \tag{15}$$

Then we have

$$\widetilde{\theta}-\theta = -\rho_n^2Q(\Lambda^2+\rho_nI_r)^{-2}\zeta + Q\left((\Lambda^2+\rho_nI_r)^{-1}\Lambda+\rho_n(\Lambda^2+\rho_nI_r)^{-2}\Lambda\right)P^T\epsilon \tag{16}$$

Similar to $\mathcal{N}_{b_n}$, define the set $\widehat{\mathcal{N}}_{b_n}$, the estimator $\widehat{\theta} = (\widehat{\theta}_1, ..., \widehat{\theta}_p)^T$ and $\widehat{\gamma}$ as

$$\widehat{\mathcal{N}}_{b_n} = \left\{ i \,\middle|\, |\widetilde{\theta}_i| > b_n \right\}, \quad \widehat{\theta}_i = \widetilde{\theta}_i \times \mathbf{1}_{i \in \widehat{\mathcal{N}}_{b_n}}, \quad \widehat{\gamma} = M\widehat{\theta} \tag{17}$$

Then, $\widehat{\theta}$ and $\widehat{\gamma}$ constitute the improved, i.e., debiased and thresholded, ridge regression estimator for the parameter vector θ and $\gamma = M\beta$ respectively. Apart from parameter estimation, we need to estimate the error variance $\sigma^2 = \mathbf{E}\epsilon_1^2$. The estimator for σ^2 is

$$\widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \sum_{j=1}^{p} x_{ij}\widehat{\theta}_j)^2 \tag{18}$$

Here $X = (x_{ij})_{i=1,...,n, j=1,...,p}$.

REMARK 2. *According to (16), the estimation error $\widetilde{\theta} - \theta$ is decomposed into a bias term $-\rho_n^2 Q(\Lambda^2 + \rho_n I_r)^{-2}\zeta$ and a variance term $Q\left((\Lambda^2 + \rho_n I_r)^{-1}\Lambda + \rho_n(\Lambda^2 + \rho_n I_r)^{-2}\Lambda\right)P^T\epsilon$. For $\|\rho_n^2 Q(\Lambda^2 + \rho_n I_r)^{-2}\zeta\|_2 \leq \frac{\rho_n^2\|\beta\|_2}{(\lambda_r^2+\rho_n)^2}$. In order to control the bias term, $\|\beta\|_2$ cannot be too large (which is achievable if β is sparse); in addition, ρ_n/λ_r^2 must be small.*

We can now explain why debiasing helps decrease the estimation error; we will use the notation of section 2. According to (1), for a fixed vector $a \in \mathbf{R}^p$,

$$\begin{aligned} a^T\widetilde{\theta}^\star - a^T\beta &= a^T\widetilde{\theta}^\star - a^T\theta - a^T Q_\perp Q_\perp^T\beta \\ &= a^T Q(\Lambda^2 + \rho_n I_r)^{-1}\Lambda P^T\epsilon - \rho_n a^T Q(\Lambda^2 + \rho_n I_r)^{-1}\zeta - a^T Q_\perp Q_\perp^T\beta \end{aligned} \tag{19}$$

Assume $a^T Q_\perp Q_\perp^T\beta = 0$. Then the bias term of $a^T\widetilde{\theta}^\star$ will be $-\rho_n a^T Q(\Lambda^2 + \rho_n I_r)^{-1}\zeta$. We can estimate this by $-\rho_n a^T Q(\Lambda^2 + \rho_n I_r)^{-1}Q^T\widetilde{\theta}^\star$ and subtract the estimated bias from $a^T\widetilde{\theta}^\star$, yielding the debiased estimator.

Compared to (16), the debiased estimator $\widetilde{\theta}$ changes the bias term from $-\rho_n a^T Q(\Lambda^2 + \rho_n I_r)^{-1}\zeta$ (having order $O\left(\frac{\rho_n\|a\|_2\times\|\beta\|_2}{\lambda_r^2+\rho_n}\right)$) to $-\rho_n^2 Q(\Lambda^2+\rho_n I_r)^{-2}\zeta$ (having order $O\left(\frac{\rho_n^2\|a\|_2\times\|\beta\|_2}{(\lambda_r^2+\rho_n)^2}\right)$). At the same time, $\widetilde{\theta}$ will enlarge the variance from $Var(\epsilon_1) \times a^T Q(\Lambda^2 + \rho_n I_r)^{-2}\Lambda^2 Q^T a$ to

$$Var(\epsilon_1) \times a^T Q\left((\Lambda^2 + \rho_n I_r)^{-1}\Lambda + \rho_n(\Lambda^2 + \rho_n I_r)^{-2}\Lambda\right)^2 Q^T a. \tag{20}$$

Assume $\rho_n/\lambda_r^2 = o(1)$; then, $\widetilde{\theta}$'s variance enlargement is asymptotically negligible but its decrease in bias is significant.

Even when $\rho_n > \lambda_r^2$, numerical simulations in figure 1 show that debiasing still may help decrease the estimation error.

REMARK 3 (Further discussion on the debiased estimator). *Apart from our work, there are other procedures that help decrease the bias of an estimator. For example, Bühlmann (2013) proposed a bias-corrected ridge regression estimator, and Zhang and Zhang (2014) considered correcting bias for a general linear regression estimator. However, the purpose of our work and those procedures are different. The bias-corrected ridge regression estimator focuses on eliminating $Q_\perp Q_\perp^T\beta$(i.e., the projection bias in Bühlmann (2013)). Therefore, if $p < n$ and X has rank p, then the bias-corrected ridge regression estimator equals the classical ridge regression estimator $\widetilde{\theta}^\star$. Our work does not focus on the projection bias but wants to diminish the estimation bias $-\rho_n Q(\Lambda^2 + \rho_n I_r)^{-1}\zeta$. Thus, even if $p < n$ and X has rank p, the debiased estimator $\widetilde{\theta}$ is still different from $\widetilde{\theta}^\star$(which is demonstrated in figure 1).*

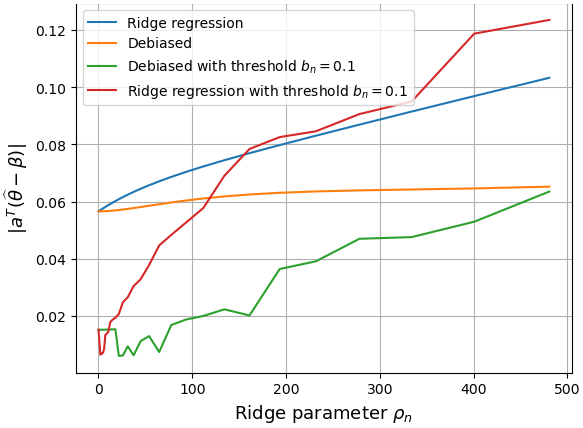


FIG 1. *Estimation errors of the ridge regression estimator* $a^T\widetilde{\theta}^\star$*, the debiased estimator(Debiased)* $a^T\widetilde{\theta}$*, the debiased and threshold ridge regression estimator(Debiased with threshold* $b_n = 0.1$*)* $a^T\widehat{\theta}$ *and the threshold ridge regression estimator(Ridge regression with threshold* $b_n = 0.1$ *as in section 4 in Shao and Deng (2012)) with respect to different* ρ_n*. The threshold* b_n *is chosen to be* 0.1*,* a *is a fixed linear combination vector with* $\|a\|_2 = 1$*, and* $\lambda_r = 12.684$*.*

THEOREM 1. *(i). Suppose assumptions 1 to 5 hold true. Then*

$$Prob\left(\widehat{\mathcal{N}}_{b_n} \neq \mathcal{N}_{b_n}\right) = O(n^{\alpha_p + m\nu_b - m\eta}) \tag{21}$$

$\mathcal{N}_{b_n}$ *is defined in section 2. In other words, the variable selection consistency holds true asymptotically. Besides,*

$$\max_{i=1,2,\ldots,p_1} |\widehat{\gamma}_i - \gamma_i| = O_p(|\mathcal{M}|^{1/m} \times n^{-\eta}) \tag{22}$$

where $\gamma_i, i = 1, \ldots, p_1$ *are defined in section 2.*
(ii). Suppose assumptions 1 to 6 hold true. Then

$$|\widehat{\sigma}^2 - \sigma^2| = O_p(n^{-\alpha_\sigma}). \tag{23}$$

An advantage of using $\widehat{\theta}$ is that it can be computed by a closed-form formula, making it simpler to practically calculate as well as derive its theoretical guarantees. As an example, define $\widehat{\theta}$'s prediction loss $\frac{1}{n}\|X\widehat{\theta} - X\theta\|_2^2 = \frac{1}{n}\|X\widehat{\theta} - X\beta\|_2^2$. If $\widehat{\mathcal{N}}_{b_n} = \mathcal{N}_{b_n}$, then from (B.8) and (B.9) in the supplement Zhang and Politis (2021a) we have

$$\begin{aligned} \frac{1}{n}\|X\widehat{\theta} - X\theta\|_2^2 \leq \frac{2}{n}\sum_{i=1}^{n}\left(\sum_{j\in\mathcal{N}_{b_n}} x_{ij}(\widetilde{\theta}_j - \theta_j)\right)^2 + \frac{2}{n}\sum_{i=1}^{n}\left(\sum_{j\notin\mathcal{N}_{b_n}} x_{ij}\theta_j\right)^2 \\ \Rightarrow \frac{1}{n}\|X\widehat{\theta} - X\theta\|_2^2 = O_p\left(n^{-\alpha_\sigma}\right). \end{aligned} \tag{24}$$

On the other hand, the prediction loss of other estimators (e.g., Lasso) can be hard to derive. Dalalyan, Hebiri and Lederer (2017), Bickel, Ritov and Tsybakov (2009) and Sun and Zhang (2012) provided oracle inequalities for the Lasso estimator. However, those inequalities depend on terms that are hard to bound. Numerical experiments in section 6 show that $\widehat{\theta}$ has comparable performance with complex estimators like the threshold Lasso or post-selection estimators. In this case, it is beneficial to choose an estimator that has clear theoretical guarantees.

Define $\widehat{\tau}_i,\ i=1,2,...,p_1$ and $H(x), x\in\mathbf{R}$ as

$$\widehat{\tau}_i=\sqrt{\sum_{k=1}^{r}\left(\sum_{j\in\widehat{\mathcal{N}}_{b_n}} m_{ij}q_{jk}\right)^2\times\left(\frac{\lambda_k}{\lambda_k^2+\rho_n}+\frac{\rho_n\lambda_k}{(\lambda_k^2+\rho_n)^2}\right)^2+\frac{1}{n}} \tag{25}$$

$$H(x)=Prob\left(\max_{i\in\mathcal{M}}\frac{1}{\tau_i}|\sum_{k=1}^{r}c_{ik}\left(\frac{\lambda_k}{\lambda_k^2+\rho_n}+\frac{\rho_n\lambda_k}{(\lambda_k^2+\rho_n)^2}\right)\xi_k|\le x\right)$$

Here $\xi_k,\ k=1,2,...,r$ are independent normal random variables with mean 0 and variance $\sigma^2=\mathbf{E}\epsilon_1^2$. $|\mathcal{M}|$ (defined in (4)) and p_1 may grow as the sample size increases. In this case, the estimator $\max_{i=1,2,...,p_1}\frac{|\widehat{\gamma}_i-\gamma_i|}{\widehat{\tau}_i}$ does not have an asymptotic distribution. However, the cumulative distribution function of $\max_{i=1,2,...,p_1}\frac{|\widehat{\gamma}_i-\gamma_i|}{\widehat{\tau}_i}$ still can be approximated by $H(x)$ (whose expression changes as the sample size increases as well). Define $c_{1-\alpha}$ as the $1-\alpha$ quantile of H; theorem 2 implies that the set

$$\left\{\gamma=(\gamma_1,...,\gamma_{p_1})\Big|\max_{i=1,...,p_1}\frac{|\widehat{\gamma}_i-\gamma_i|}{\widehat{\tau}_i}\le c_{1-\alpha}\right\} \tag{26}$$

is an asymptotically valid $(1-\alpha)\times 100\%$ confidence region for the parameter of interest γ.

THEOREM 2. *Suppose assumptions 1 to 7 hold true. Then*

$$\lim_{n\to\infty}\sup_{x\ge 0}|Prob\left(\max_{i=1,2,...,p_1}\frac{|\widehat{\gamma}_i-\gamma_i|}{\widehat{\tau}_i}\le x\right)-H(x)|=0 \tag{27}$$

where $\gamma_i, i=1,...,p_1$ are defined in section 2.

Gaussian approximation theorems like theorem 2 are useful tools not only in linear models but also in other high dimensional statistics; e.g., Chernozhukov, Chetverikov and Kato (2013) and Zhang and Wu (2017).

4. Bootstrap inference and hypothesis testing. An obstacle for constructing a practical confidence region or testing a hypothesis via theorem 2 are the unknown $\mathcal{M}$, $\mathcal{N}_{b_n}$, and σ. Besides, H is too complicated to have a closed-form formula. Fortunately, statisticians can simulate normal random variables on a computer, so they may use Monte-Carlo simulations to find the $1-\alpha$ quantile of H. Based on this idea, this section develops a wild bootstrap algorithm similar to Mammen (1993) and Chernozhukov, Chetverikov and Kato (2013) for the following tasks: constructing the confidence region for the parameter of interest $\gamma=M\beta$; and testing the null hypothesis $\gamma=\gamma_0$ (for a known γ_0) versus the alternative hypothesis $\gamma\neq\gamma_0$. Similar to Zhang and Cheng (2017), Chernozhukov, Chetverikov and Kato (2013), and Zhang and Wu (2017), we use the maximum statistic $\max_{i=1,2,...,p_1}\frac{|\widehat{\gamma}_i-\gamma_i|}{\widehat{\tau}_i}$ to construct a simultaneous confidence region.

ALGORITHM 1 (Wild bootstrap inference and hypothesis testing). ***Input:*** *Design matrix X, dependent variables $y=X\beta+\epsilon$, linear combination matrix M, ridge parameter ρ_n, threshold b_n, nominal coverage probability $1-\alpha$, number of bootstrap replicates B*
Additional input for testing: $\gamma_0=(\gamma_{0,1},...,\gamma_{0,p_1})^T$
1. Calculate $\widehat{\theta}$, $\widehat{\gamma}=(\widehat{\gamma}_1,...,\widehat{\gamma}_{p_1})^T$ defined in (17), $\widehat{\tau}_i,\ i=1,2,...,p_1$ defined in (25), and $\widehat{\sigma}$ defined in (18).

2. Generate i.i.d. errors $\epsilon^* = (\epsilon_1^*, ..., \epsilon_n^*)^T$ *with* $\epsilon_i^*, i = 1, ..., n$ *having normal distribution with mean* 0 *and variance* $\widehat{\sigma}^2$*, then calculate* $y^* = X\widehat{\theta} + \epsilon^*$ *and* $\widehat{\theta}_\perp = Q_\perp Q_\perp^T \widehat{\theta}$ *(*$Q_\perp$ *is defined in section 2).*
3. Calculate $\widetilde{\theta}^{\star *} = (X^T X + \rho_n I_p)^{-1} X^T y^*$ *and* $\widetilde{\theta}^* = (\widetilde{\theta}_1^*, ..., \widetilde{\theta}_p^*)^T = \widetilde{\theta}^{\star *} + \rho_n \times Q(\Lambda^2 + \rho_n I_r)^{-1} Q^T \widetilde{\theta}^{\star *} + \widehat{\theta}_\perp$.
4. Calculate $\widehat{\mathcal{N}}_{b_n}^* = \left\{ i \middle| |\widetilde{\theta}_i^*| > b_n \right\}$ *and* $\widehat{\theta}^* = (\widehat{\theta}_1^*, ..., \widehat{\theta}_p^*)^T$ *with* $\widehat{\theta}_i^* = \widetilde{\theta}_i^* \times \mathbf{1}_{i \in \widehat{\mathcal{N}}_{b_n}^*}$ *for* $i = 1, 2, ..., p$.
5. Calculate $\widehat{\gamma}^* = M\widehat{\theta}^*$, $\widehat{\tau}_i^*, i = 1, 2, ..., p_1$, *and* E_b^* *such that*

(28)
$$\widehat{\tau}_i^* = \sqrt{\sum_{k=1}^{r} \left(\sum_{j \in \widehat{\mathcal{N}}_{b_n}^*} m_{ij} q_{jk} \right)^2 \times \left(\frac{\lambda_k}{\lambda_k^2 + \rho_n} + \frac{\rho_n \lambda_k}{(\lambda_k^2 + \rho_n)^2} \right)^2 + \frac{1}{n}}, \quad E_b^* = \max_{i=1,2,...,p_1} \frac{|\widehat{\gamma}_i^* - \widehat{\gamma}_i|}{\widehat{\tau}_i^*}$$

6.a (For constructing a confidence region) Repeat steps 2 to 5 for B *times to generate* E_b^*, $b = 1, 2, ..., B$*; then calculate the* $1 - \alpha$ *sample quantile* $C_{1-\alpha}^*$ *of* E_b^**. The* $1 - \alpha$ *confidence region for the parameter of interest* $\gamma = M\beta$ *is given by the set*

(29)
$$\left\{ \gamma = (\gamma_1, ..., \gamma_{p_1})^T \middle| \max_{i=1,2,...,p_1} \frac{|\widehat{\gamma}_i - \gamma_i|}{\widehat{\tau}_i} \leq C_{1-\alpha}^* \right\}$$

6.b (For hypothesis testing) Repeat steps 2 to 5 for B *times to generate* E_b^*, $b = 1, 2, ..., B$*; then calculate the* $1 - \alpha$ *sample quantile* $C_{1-\alpha}^*$ *of* E_b^**. Reject the null hypothesis* $\gamma = \gamma_0$ *when*

(30)
$$\max_{i=1,2,...,p_1} \frac{|\widehat{\gamma}_i - \gamma_{0,i}|}{\widehat{\tau}_i} > C_{1-\alpha}^*.$$

As in section 2, if X *has rank* $p \leq n$*, we define* $\widehat{\theta}_\perp = 0$*, the* p *dimensional vector with all elements* 0.

According to theorem 1.2.1 in Politis, Romano and Wolf (1999), the consistency of algorithm 1 –either for asymptotic validity of confidence regions or consistency of the hypothesis test– is ensured if

(31)
$$Prob\left(\max_{i=1,2,...,p_1} \frac{|\widehat{\gamma}_i - \gamma_i|}{\widehat{\tau}_i} \leq c_{1-\alpha}^* \right) \to 1 - \alpha$$

where $c_{1-\alpha}^*$ is the $1-\alpha$ quantile of the conditional distribution $Prob^*\left(\max_{i=1,2,...,p_1} \frac{|\widehat{\gamma}_i^* - \widehat{\gamma}_i|}{\widehat{\tau}_i^*} \leq x \right)$; we prove this in theorem 3 below.

THEOREM 3. *Suppose assumptions 1 to 7 hold true. Then*

(32)
$$\sup_{x \geq 0} \left| Prob^* \left(\max_{i=1,2,...,p_1} \frac{|\widehat{\gamma}_i^* - \widehat{\gamma}_i|}{\widehat{\tau}_i^*} \leq x \right) - H(x) \right| = o_P(1).$$

In addition, for any given $0 < \alpha < 1$*, (31) holds true.*

Theorem 3 has two implications. On the one hand, the confidence region introduced in step 6.a of algorithm 1 is asymptotically valid, i.e., its coverage tends to $1 - \alpha$. On the other hand, consider the hypothesis test of step 6.b of algorithm 1; Theorem 3 implies that, if the null hypothesis is true, then the probability for incorrectly rejecting the null hypothesis is asymptotically α, i.e., the test is consistent.

5. Bootstrap interval prediction. Given our data from the linear model $y = X\beta + \epsilon$, consider a new $p_1 \times p$ regressor matrix X_f, i.e., a collection of regressor (column) vectors that happen to be of interest; as with X itself, X_f is assumed given, i.e., deterministic. The prediction problem involves (a) finding a predictor for the *future* (still unobserved) vector $y_f = X_f\beta + \epsilon_f$, and (b) finding a $1-\alpha$ prediction region $A \subset R^{p_1}$ so that $Prob(y_f \in A) \to 1-\alpha$ as the (original) sample size $n \to \infty$. Here $\epsilon_f = (\epsilon_{f,1}, ..., \epsilon_{f,p_1})^T$ are i.i.d. errors with the same marginal distribution as ϵ_1, and ϵ_f is independent with ϵ.

Finding a good predictor based on different criteria is a big topic. For example, Greenshtein and Ritov (2004) applied Lasso in constructing predictors and their predictor's mean square error is minimal asymptotically. We construct an intuitive predictor based on the following idea: if β were known, the predictor of y_f that is optimal with respect to total mean squared error is $X_f\beta$; since β is typically unknown, we can estimate it by $\widehat{\theta}$ as in (17), yielding the practical predictor $\widehat{y}_f = X_f\widehat{\theta}$. In what follows, we would like to derive a $1-\alpha$ prediction region for y_f based on the intuitive predictor $\widehat{y}_f$.

We adopt definition 2.4.1 of Politis (2015), and define a consistent prediction region in terms of conditional coverage as follows.

DEFINITION 1 (Consistent prediction region). *A set $\Gamma = \Gamma(X, y, X_f)$ is called a $1-\alpha$ consistent prediction region for the future observation $y_f = X_f\beta + \epsilon_f$ if*

$$Prob\,(y_f \in \Gamma | y) \to_p 1-\alpha \ \ as\ n \to \infty. \tag{33}$$

Note that the convergence in (33) is "in probability" since $Prob\,(y_f \in \Gamma|y)$ is a function of y, and therefore random; see also Lei and Wasserman (2014) for more on the notion of conditional validity.

Other authors, including Stine (1985), Romano, Patterson and Candès (2019), and Chernozhukov, Wüthrich and Zhu (2019), considered another definition of prediction interval consistency focusing on unconditional coverage, i.e., insisting that

$$Prob(y_f \in \Gamma) \to 1-\alpha. \tag{34}$$

However, the conditional coverage of definition 1 is a stronger property. To see why, define the random variables $U_n = Prob\,(y_f \in \Gamma|y)$, noting that y has dimension n. Then, the boundedness of U_n can be invoked to show that if $U_n \to_p 1-\alpha$, then $\mathbf{E}U_n \to 1-\alpha$ as well. Hence, (33) implies (34); see Zhang and Politis (2021b) for a further discussion on conditional vs. unconditional coverage.

Consider the prediction error $y_f - X_f\widehat{\theta} = \epsilon_f - X_f(\widehat{\theta} - \beta)$. If we can put bounds on the prediction error that are valid with conditional probability $1-\alpha$ (asymptotically), then a consistent prediction region ensues. Note that the prediction error has two parts: ϵ_f and $-X_f(\widehat{\theta} - \beta)$. Although the latter may be asymptotically negligible, it is important in practice to not approximate it by zero as it would yield finite-sample undercoverage; see e.g. Ch. 3 of Politis (2015) for an extensive discussion.

Theorem 2 indicates that the asymptotically negligible estimation error can be approximated by normal random variables. On the other hand, the non-negligible error ϵ_f may not have a normal distribution; so in order to approximate the distribution of $\epsilon_f - X_f(\widehat{\theta} - \beta)$, we need to estimate the errors' marginal distribution as well.

This section requires some additional assumptions.

Additional assumptions

8. The cumulative distribution function of errors $F(x) = Prob\,(\epsilon_1 \leq x)$ is continuous
9. The number of regressors of interest is bounded, i.e., $p_1 = O(1)$

Since F is increasing and bounded, if $F(x)$ is continuous, then F is uniformly continuous on $\mathbf{R}$. this property is useful in the proof of lemma 1.

LEMMA 1. *Suppose assumption 1 to 6 and 8 hold true. Define the residuals* $\widehat{\epsilon}' = (\widehat{\epsilon}'_1, ..., \widehat{\epsilon}'_n)^T = y - X\widehat{\theta}$, *as well as the centered residuals* $\widehat{\epsilon} = (\widehat{\epsilon}_1, ..., \widehat{\epsilon}_n)^T$ *with* $\widehat{\epsilon}_i = \widehat{\epsilon}'_i - \frac{1}{n}\sum_{i=1}^n \widehat{\epsilon}'_i$. *If we let* $\widehat{F}(x) = \frac{1}{n}\sum_{i=1}^n \mathbf{1}_{\widehat{\epsilon}_i \leq x}$, *then*

$$\sup_{x \in \mathbf{R}} |\widehat{F}(x) - F(x)| \to_p 0 \text{ as } n \to \infty. \tag{35}$$

We emphasize that the dimension of parameters p in lemma 1 can grow to infinity as long as assumption 4 is satisfied. Furthermore, the validity of lemma 1 –as well as that of theorem 4 that follows– does not require assumption 7.

We will resample the centered residuals $\widehat{\epsilon}_i, i = 1, 2, ..., n$ (in other words, generate random variables with distribution $\widehat{F}$) in algorithm 2. Lemma 1 will ensure that the centered residuals can capture the distribution of the non-negligible errors.

For a high dimensional linear model, lemma 1 is not an obvious result; see Mammen (1996) for a detailed explanation. Lemma 1 is the foundation for a new resampling procedure as follows; this is a hybrid bootstrap as it combines the residual-based bootstrap to replicate the new error ϵ_f with the normal approximation to the estimation error $-X_f(\widehat{\theta} - \beta)$.

ALGORITHM 2 (Hybrid bootstrap for prediction region). ***Input:*** *Design matrix* X, *dependent variables* $y = X\beta + \epsilon$, *a new* $p_1 \times p$ *linear combination matrix* X_f, *ridge parameter* ρ_n, *threshold* b_n, *nominal coverage probability* $0 < 1 - \alpha < 1$, *the number of bootstrap replicates* B

1. Calculate $\widehat{\theta}$ *defined in* (17), $\widehat{\sigma}$ *defined in* (18), $\widehat{\epsilon}$ *defined in lemma 1,* $\widehat{y}_f = (\widehat{y}_{f,1}, ..., \widehat{y}_{f,p_1})^T = X_f\widehat{\theta}$, *and* $\widehat{\theta}_\perp = Q_\perp Q_\perp^T \widehat{\theta}$.

2. Generate i.i.d. errors $\epsilon^* = (\epsilon_1^*, ..., \epsilon_n^*)^T$ *with* $\epsilon_i^*, i = 1, ..., n$ *having normal distribution with mean* 0 *and variance* $\widehat{\sigma}^2$. *Then generate i.i.d. errors* $\epsilon_f^* = (\epsilon_{f,1}^*, ..., \epsilon_{f,p_1}^*)^T$ *with* $\epsilon_{f,i}^*, i = 1, ..., p_1$ *having cumulative distribution function* $\widehat{F}$ *defined in lemma 1. Calculate* $y^* = X\widehat{\theta} + \epsilon^*$.

3. Calculate $\widetilde{\theta}^{**} = (X^TX + \rho_n I_p)^{-1}X^Ty^*$ *and* $\widetilde{\theta}^* = \widetilde{\theta}^{**} + \rho_n \times Q(\Lambda^2 + \rho_n I_r)^{-1}Q^T\widetilde{\theta}^{**} + \widehat{\theta}_\perp$. *Then derive* $\widehat{\mathcal{N}}_{b_n}^* = \left\{i \middle| |\widetilde{\theta}_i^*| > b_n\right\}$, $\widehat{\theta}^* = (\widehat{\theta}_1^*, ..., \widehat{\theta}_p^*)^T$ *with* $\widehat{\theta}_i^* = \widetilde{\theta}_i^* \times \mathbf{1}_{i \in \widehat{\mathcal{N}}_{b_n}^*}$ *for* $i = 1, 2, ..., p$.

4. Calculate $y_f^* = (y_{f,1}^*, ..., y_{f,p_1}^*)^T = X_f\widehat{\theta} + \epsilon_f^*$ *and* $\widehat{y}_f^* = (\widehat{y}_{f,1}^*, ..., \widehat{y}_{f,p_1}^*)^T = X_f\widehat{\theta}^*$. *Define* $E_b^* = \max_{i=1,2,...,p_1} |y_{f,i}^* - \widehat{y}_{f,i}^*|$.

5. Repeat steps 2 to 4 for B *times, and generate* E_b^*, $b = 1, 2, ..., B$. *Calculate the* $1 - \alpha$ *sample quantile* $C_{1-\alpha}^*$ *of* E_b^*. *Then, the* $1 - \alpha$ *prediction region for* $y_f = X_f\beta + \epsilon_f$ *is given by*

$$\left\{y_f = (y_{f,1}, ..., y_{f,p_1})^T \middle| \max_{i=1,2,...,p_1} |y_{f,i} - \widehat{y}_{f,i}| \leq C_{1-\alpha}^*\right\}. \tag{36}$$

If the design matrix X *has rank* p, *then* $\widehat{\theta}_\perp$ *is defined to be* 0.

Similar to section 4, here we define $c^*_{1-\alpha}$ as the $1-\alpha$ quantile of the conditional distribution $Prob^*\left(\max_{i=1,...,p_1}|y^*_{f,i}-\widehat{y}^*_{f,i}|\le x\right)$, which can be approximated by $C^*_{1-\alpha}$ by letting $B\to\infty$. Theorem 4 below proves $Prob\left(\max_{i=1,2,...,p_1}|y_{f,i}-\widehat{y}_{f,i}|\le c^*_{1-\alpha}\right)\to 1-\alpha$ as the sample size $n\to\infty$, which justifies the consistency of the prediction region (36).

THEOREM 4. *Suppose assumptions 1 to 6 and 8 to 9 hold true(here consider* $M=(m_{ij})_{i=1,...,p_1,j=1,...,p}$ *in assumption 5 as* X_f*). Then*

(37)
$$\sup_{x\ge 0}|Prob^*\left(\max_{i=1,2,...,p_1}|y^*_{f,i}-\widehat{y}^*_{f,i}|\le x\right)-Prob^*\left(\max_{i=1,2,...,p_1}|y_{f,i}-\widehat{y}_{f,i}|\le x\right)|=o_p(1).$$

For any fixed $0<\alpha<1$*, it follows that*

(38)
$$Prob^*\left(\max_{i=1,2,...,p_1}|y_{f,i}-\widehat{y}_{f,i}|\le c^*_{1-\alpha}\right)\to_p 1-\alpha \text{ as } n\to\infty.$$

Note that the bootstrap probability $Prob^*(\cdot)$ is probability conditional on the data y, thus justifying the notion of conditional validity of our definition 1.

A version of the algorithm 2 can be constructed where the residual-based bootstrap part is conducted by resampling from the empirical distribution of the (centered) predictive, i.e., leave-one-out, residuals instead of the fitted residuals $\widehat{\epsilon}_i$; see Ch. 3 of Politis (2015) for a discussion.

6. Numerical Simulations. Define $k_n=\sqrt{n\log(n)}$ and the following four terms

(39)
$$\mathcal{K}_1=\max_{i=1,2,...,p_1}k_n|\sum_{j\notin\mathcal{N}_{b_n}}m_{ij}\theta_j|,\quad \mathcal{K}_2=\max_{i=1,2,...,p_1}k_n|\sum_{j=1}^{r}m_{ij}\theta_{\perp,j}|,$$
$$\mathcal{K}_3=b_n\sum_{j\notin\mathcal{N}_{b_n}}|\theta_j|,\quad \mathcal{K}_4=\frac{\sqrt{|\mathcal{N}_{b_n}|}}{\lambda_r};$$

see section 2 for the meaning of notations in the above. Assumptions 5 and 6 imply that these terms converge to 0 as the sample size $n\to\infty$. Indeed, if one of the $\mathcal{K}_i$ is large, the debiased and threshold ridge regression estimator may have a large bias, which affects the performance of the bootstrap algorithms.

In this section, we generate the design matrix X, the linear combination matrix M, and the parameters β through the following strategies:

Design matrix X: define $X=[x_1,...,x_n]^T$ with $x_i=(x_{i1},...,x_{ip})^T\in\mathbf{R}^p, i=1,...,n$. Generate $x_1,x_2,\ldots$ as i.i.d. normal random vectors with mean 0 and covariance matrix $\Sigma\in\mathbf{R}^{p\times p}$. We choose Σ with diagonal elements equal to 2.0 and off-diagonal elements equal to 0.5.

(40)
$$M=\begin{bmatrix} m_{11} & m_{12} & ... & m_{1\tau} & m_{1\tau+1} & ... & m_{1p}\\ m_{21} & m_{22} & ... & m_{2\tau} & m_{2\tau+1} & ... & m_{2p}\\ \vdots & \vdots & ... & \vdots & \vdots & ... & \vdots\\ m_{|\mathcal{M}|1} & m_{|\mathcal{M}|2} & ... & m_{|\mathcal{M}|\tau} & m_{|\mathcal{M}|\tau+1} & ... & m_{|\mathcal{M}|p}\\ 0 & 0 & ... & 0 & m_{|\mathcal{M}|+1\tau+1} & ... & m_{|\mathcal{M}|+1p}\\ \vdots & \vdots & ... & \vdots & \vdots & ... & \vdots\\ 0 & 0 & ... & 0 & m_{p_1\tau+1} & ... & m_{p_1p}\end{bmatrix}$$

M and β when $p < n$: choose $\tau = 50$ in (40). Generate $m_{ij}^{'}, i = 1, 2, ..., |\mathcal{M}|,\ j = 1, 2, ..., \tau$ as i.i.d. normal with mean 0.5 and variance 1.0, and generate $m_{ij}^{'}, i = 1, 2, ..., p_1, j = \tau + 1, ..., p$ as i.i.d. normal with mean 1.0 and variance 4.0. Use $m_{ij} = 2.0 \times m_{ij}^{'} / \sqrt{\sum_{j=1}^{\tau} m_{ij}^{'2}}$ for $i = 1, 2, ..., |\mathcal{M}|,\ j = 1, 2, ..., \tau$; $m_{ij} = 4.0 \times m_{ij}^{'} / \sqrt{\sum_{j=\tau+1}^{p} m_{ij}^{'2}}$ for $i = 1, 2, ..., |\mathcal{M}|,\ j = \tau + 1, ..., p$; and $m_{ij} = 6.0 \times m_{ij}^{'} / \sqrt{\sum_{j=\tau+1}^{p} m_{ij}^{'2}}$ for $i = |\mathcal{M}| + 1, ..., p_1, j = \tau + 1, ..., p$. Choose $\beta = (\beta_1, ..., \beta_p)^T$ with $\beta_i = 2.0, i = 1, 2, 3,\ \beta_i = -2.0, i = 4, 5, 6,\ \beta_i = 1.0, i = 7, 8, 9,\ \beta_i = -1.0, i = 10, 11, 12,\ \beta_i = 0.01, i = 13, 14, 15, 16$, and 0 otherwise.
M and β when $p > n$: choose $\tau = 6$ in (40). Generate $m_{ij}^{'}, i = 1, 2, ..., |\mathcal{M}|,\ j = 1, 2, ..., \tau$ as i.i.d. normal with mean 0.5 and variance 1.0, and generate $m_{ij}^{'}, i = 1, 2, ..., p_1,\ j = \tau + 1, ..., p$ as i.i.d. normal with mean 1.0 and variance 4.0. Use $m_{ij} = 2.0 \times m_{ij}^{'} / \sqrt{\sum_{j=1}^{\tau} m_{ij}^{'2}}$ for $i = 1, 2, ..., |\mathcal{M}|,\ j = 1, 2, ..., \tau$; and $m_{ij} = m_{ij}^{'} / \sqrt{\sum_{j=\tau+1}^{p} m_{ij}^{'2}}$ for $i = 1, 2, ..., p_1,\ j = \tau + 1, ..., p$. Choose $\beta_i = 1.0, i = 1, 2, 3,\ \beta_i = -1.0, i = 4, 5, 6$, and 0 otherwise. When $p > n$, β may not be identifiable (Shao and Deng (2012)), and β may not equal θ(defined in section 2) despite $X\beta = X\theta$. We consider both situations and evaluate the performance of proposed methods on the linear model $y = X\beta + \epsilon$ and $y = X\theta + \epsilon$. We fix X and M in each simulation.

The different regression algorithms considered are the debiased and threshold ridge regression(Deb Thr), ridge regression, Lasso, threshold ridge regression (Thr Ridge), threshold Lasso (Thr Lasso), and the post-selection algorithms, i.e., Lasso + OLS (Post OLS), and Lasso + Ridge (Post Ridge). We consider 6 cases for simulation involving a different p/n ratio, and Normal vs. Laplace (2-sided exponential) errors; we present detailed information about each simulation case in table 1, compare the performance of different regression algorithms in Figure 2 to 4 and Table 2, and record the performance of bootstrap algorithms on estimation/hypothesis testing and interval-prediction in Table 3 and Figure 5. The optimal ridge parameter ρ_n and threshold b_n are chosen by 5-fold cross validation. To adapt to assumption 9, we choose X_f as the first 100 lines of M for prediction.

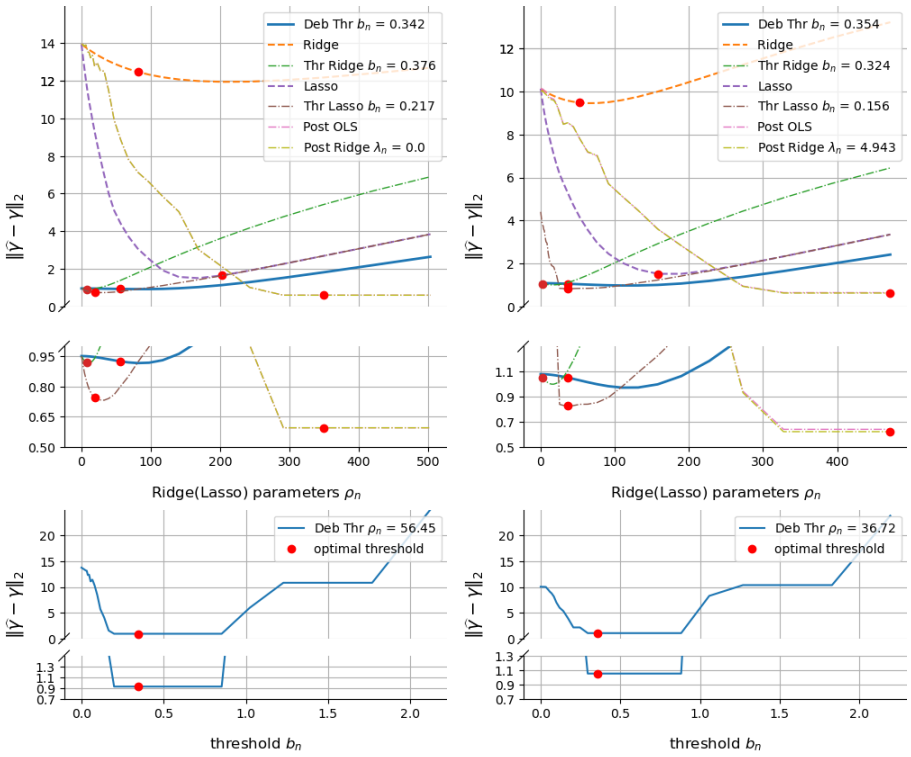
TABLE 1

Information about X, M and ϵ in each simulation case. For the normal distribution we choose variance 4, for the Laplace distribution we choose the scale $\sqrt{2}$. By doing this, the variance of residuals is 4. When $p > n$, $\beta \neq \theta$. ***The left(right) side of the slashes represent $\mathcal{K}_2$ calculated by the linear model $y = X\beta + \epsilon$ ($y = X\theta + \epsilon$).*** *The difference between β and θ does not change other terms in case 5 and 6.*

Case	n	p	Residual	p_1	$\lvert\mathcal{M}\rvert$	λ_r	ρ_n	b_n	$\mathcal{K}_1$	$\mathcal{K}_2$	$\mathcal{K}_3$	$\mathcal{K}_4$
1	1000	500	Normal	800	300	12.978	56.453	0.343	1.370	0.0	0.013	1.712
2	1000	500	Laplace	800	300	12.561	36.728	0.354	1.636	0.0	0.014	1.769
3	1000	650	Laplace	800	300	8.226	56.432	0.396	1.553	0.0	0.016	3.085
4	1000	500	Laplace	800	700	12.847	55.317	0.346	1.510	0.0	0.014	1.730
5	1000	1500	Normal	800	300	9.766	1.201	0.228	6.938	129 / 0.0	8.214	3.962
6	1000	1500	Laplace	800	300	9.766	1.201	0.228	6.938	129 / 0.0	8.214	3.962

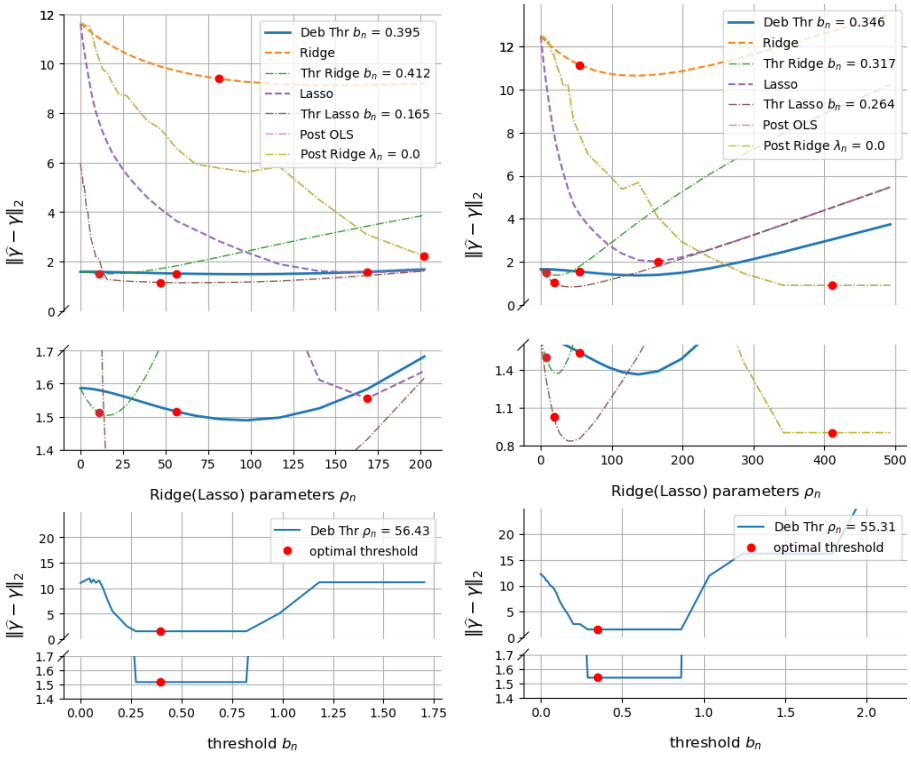
Case 5 and 6 consider both the linear model $y = X\beta + \epsilon$ and $y = X\theta + \epsilon$, here $\beta \neq \theta = QQ^T\beta$. The difference in β and θ affects the value of $\mathcal{K}_2$(but does not affect others), so we have two values in table 1.

Figure 2 plots the Euclidean norm $\|\widehat{\gamma} - \gamma\|_2$, with $\widehat{\gamma}$ defined in (17), and γ defined in Section 2, for various linear regression methods. When the underlying linear model is sparse, thresholding decreases the ridge regression estimator's error(from around 10 to around 2 in our experiment). However, the performance of the threshold ridge regression method is



(a) Case 1

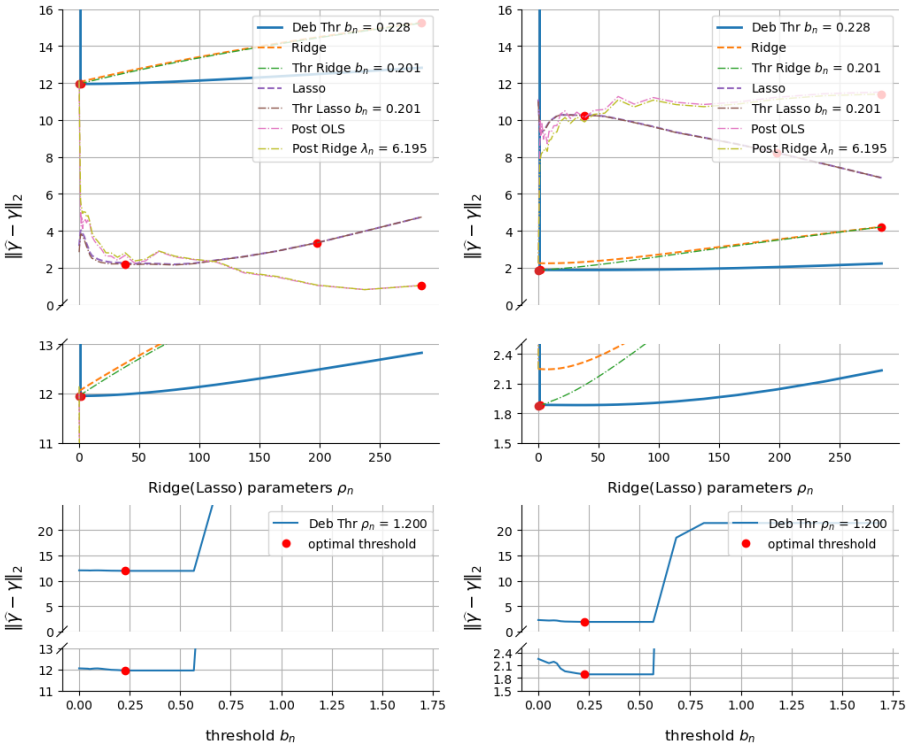
(b) Case 2



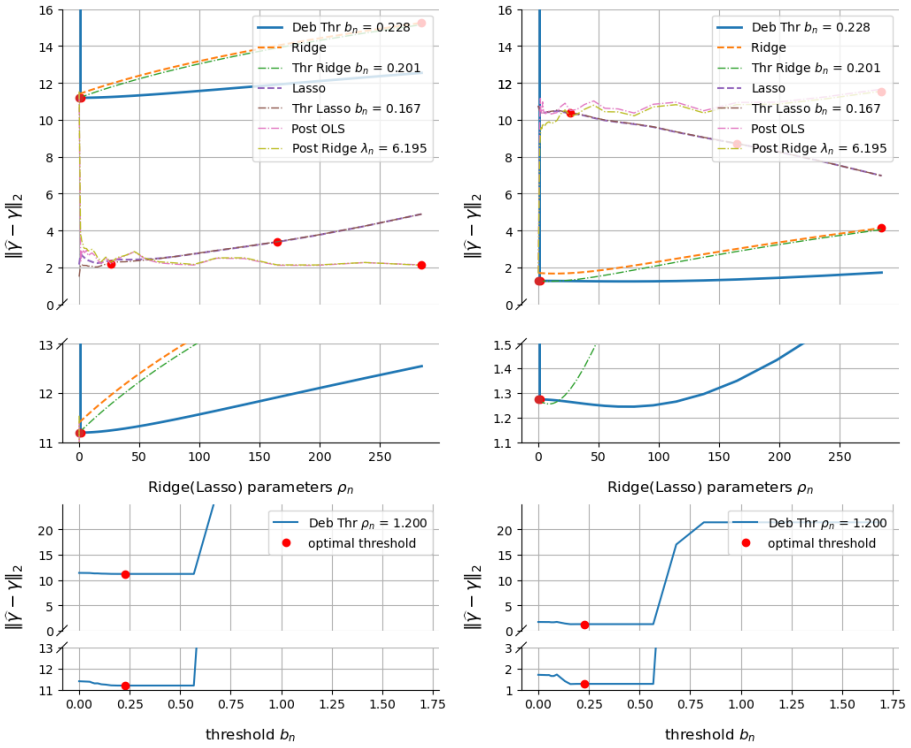
(c) Case 3

(d) Case 4

FIG 2. *Estimation performance of various linear regression methods over case 1 to 4. 'Deb' abbreviates 'Debiased', 'Thr' abbreviates 'Threshold', 'Post' abbreviates 'Post-selection', and 'OLS' abbreviates 'ordinary least square'. Red dots represent the ridge/Lasso parameters ρ_n and the thresholds b_n selected by 5-fold cross validation. The vertical axis represents the Euclidean norm of $\widehat{\gamma} - \gamma$ where $\widehat{\gamma}$ is defined in (17) and γ is defined in Section 2. The little graphs in the middle of each of the four graphs show a zoomed-in part of the graph above it. The little graphs below each of the four graphs show the estimation performance of the debiased and threshold ridge regression method with respect to different thresholds.*

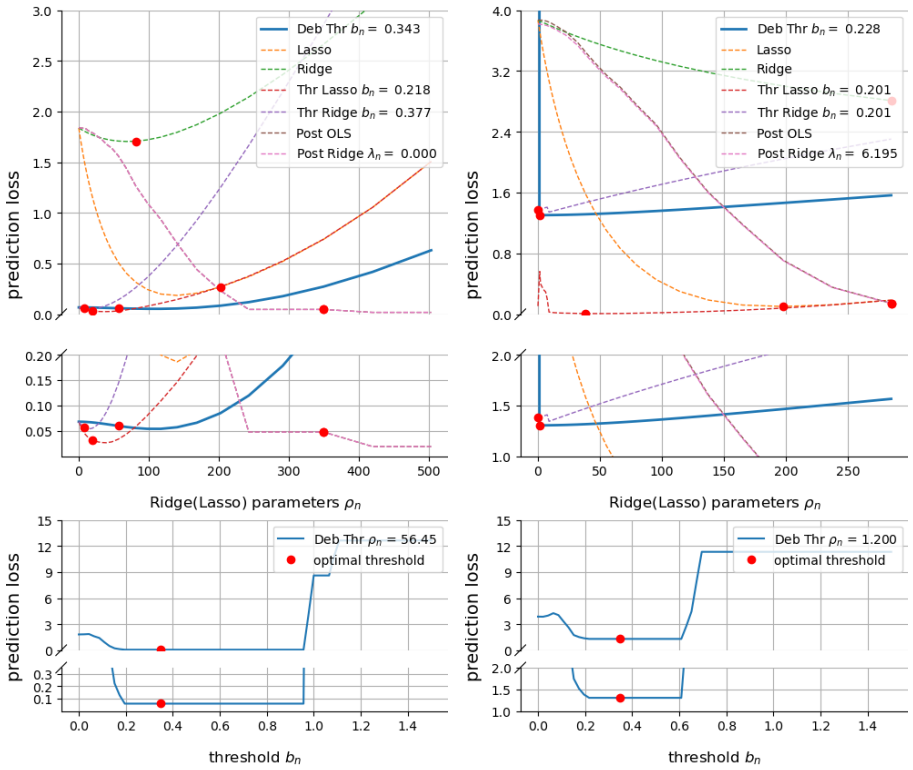


(a) Case 5(use β) (b) Case 5(use θ)



(c) Case 6(use β) (d) Case 6(use θ)

FIG 3. *Estimation performance of various linear regression methods over case 5 to 6. The meaning of symbols coincides with figure 2.*



(a) Case 1 (b) Case 5

FIG 4. *Prediction loss(see section 3) of various linear regression methods. The meaning of symbols coincides with figure 2.*

sensitive to the ridge parameter ρ_n, i.e., $\|\widehat{\gamma} - \gamma\|_2$ can be significantly larger than its minimum despite ρ_n is close to the minimizer of $\|\widehat{\gamma} - \gamma\|_2$.

In reality, cross validation does not necessarily guarantee selection of the optimal ρ_n and b_n, so it is risky to use the threshold ridge regression method. Debiasing helps decrease the ridge regression estimator's error; more importantly, it is robust to changes in the choice of ρ_n. Even if a cross validation selects a sub-optimal ρ_n, the error of the debiased and threshold ridge regression estimator does not surge, and the estimator's performance does not notably deteriorate. On the other hand, in Figure 2 and 3, $\|\widehat{\gamma} - \gamma\|_2$ reaches its minimum and does not increase for a wide range of b_n. For example, in case 3 the cross validation chooses $b_n = 0.395$, but any value between 0.30 and 0.75 can be the optimal threshold(the threshold having the smallest $\|\widehat{\gamma} - \gamma\|_2$). Since there is a wide interval of thresholds b_n that have small $\|\widehat{\gamma} - \gamma\|_2$, the regression algorithm has robustness regarding the choice of b_n. Because of these good properties, we consider the debiased and threshold ridge regression as a practical method to handle real-life data.

Thresholding also helps improve the performance of Lasso, especially when the Lasso parameter is small. However, when the Lasso parameter becomes large, Lasso method already recovers the underlying sparsity of the linear model, and thresholding becomes unnecessary (but large Lasso parameters tend to introduce large bias).

When the dimension of parameters p is greater than the sample size n, both parameters β and θ(see section 2) could be considered as the 'parameters' for the linear model. Lasso methods estimate linear combinations of β, while ridge regression methods estimate linear combinations of θ. Under this situation, the difference between β and θ is the main factor for

the estimators' error. In reality, statisticians cannot distinguish between β and θ based on data. So they need to design which parameters to estimate a priori and select a suitable regression method (e.g., Lasso, ridge regression, or their variations) reflecting their preferences.

The optimal prediction loss of the debiased and threshold ridge regression method is comparable to the threshold Lasso and the post-selection algorithms. Furthermore, the prediction loss is robust to changes in the choice of ρ_n, b_n. When $p > n$, the sparsity of θ is violated and a large bias is introduced to the estimator(see table 1). As a result the prediction loss will be enlarged.

As a summary of Figure 2 to 4, apart from having a closed-form formula, the debiased and threshold ridge regression has the smallest estimation error and prediction loss among all ridge regression variations, and has comparable performance to the threshold Lasso. Furthermore, it is not overly sensitive on changes in the ridge parameter ρ_n as well as the threshold b_n. Therefore, even when a sub-optimal ρ_n or b_n are selected, the performance of the debiased and threshold ridge regression is not severely affected. When $p > n$, this method (and other ridge regression methods) considers θ rather than β to be the parameter of the linear model. So, in this case, ridge regression methods are suitable if the underlying linear model is indeed $y = X\theta + \epsilon$(in other words, the projection does not have effect on the parameters of the linear model).

Table 2 demonstrates the model selection performance of various linear regression algorithms. Following Fithian, Sun and Taylor (2017), we evaluate the algorithms through the frequency of model misspecification(i.e., $\widehat{\mathcal{N}}_{b_n} \neq \mathcal{N}_{b_n}$), $P(\widehat{\mathcal{N}}_{b_n} \neq \mathcal{N}_{b_n})$; the average size of model misspecification $|\widehat{\mathcal{N}}_{b_n} \Delta \mathcal{N}_{b_n}|$ (here Δ denotes the symmetric difference, i.e. $A \Delta B = (A - B) \cup (B - A)$); and the average false discovery rate $|\widehat{\mathcal{N}}_{b_n} - \mathcal{N}_{b_n}| / \max(|\widehat{\mathcal{N}}_{b_n}|, 1)$. Notice that Lasso and the ridge regression do not have thresholds. For these algorithms we say $i \in \widehat{\mathcal{N}}_{b_n}$ if the estimated parameter $|\widehat{\beta}_i| > 0.001$. When the sparsity assumption is not violated, the debiased and threshold ridge regression can perfectly recover the model sparsity, and thresholding is also an essential tool that improves Lasso's model selection performance. On the other hand, if the sparsity assumption is violated, then $|\theta_i|$ can be close to b_n even if $i \notin \mathcal{N}_{b_n}$. Despite the stochastic errors are still small, the summation of θ_i and the stochastic error can exceed b_n, which results in selecting a false model.

TABLE 2

Model selection performance of various linear regression methods over case 1 and 5. The hyper-parameters are chosen by 5-fold cross validation. The overscore represents calculating the sample mean among 1000 simulations.

Case	Algorithm	$P(\widehat{\mathcal{N}}_{b_n} \neq \mathcal{N}_{b_n})$	$\overline{\lvert\widehat{\mathcal{N}}_{b_n} \Delta \mathcal{N}_{b_n}\rvert}$	$\overline{\text{False discovery rate}}$
1	Deb Thr	0.0	0.0	0.0
	Lasso	1.0	9.674	0.463
	Ridge	1.0	240.53	0.967
	Thr Lasso	0.009	0.009	0.001
	Thr Ridge	0.0	0.0	0.0
5(use θ)	Deb Thr	0.132	0.140	0.034
	Lasso	1.0	761.59	0.308
	Ridge	1.0	452.69	0.283
	Thr Lasso	0.004	0.004	0.001
	Thr Ridge	0.508	0.729	0.157

Table 3 records the average errors of the proposed statistics $\widehat{\gamma}$(defined in (17)), $\widehat{\sigma}^2$(defined in (18)), and the coverage probability of the confidence region (29) as well as the coverage

probability of the prediction region (36), in 1000 numerical simulations. We also record the frequency of model misspecification $P(\widehat{\mathcal{N}}_{b_n} \neq \mathcal{N}_{b_n})$. When the sample size n is greater than the dimension of parameters p, thresholding is likely to recover the sparsity of the parameters. In all these cases, i.e., Case 1–4, our confidence intervals achieve near-perfect coverage. The slight under-coverage in prediction intervals is a well-known phenomenon; see e.g. Ch. 3.7 of Politis (2015).

However, in cases 5 and 6 where $p > n$, θ is not necessarily sparse, and model misspecification may happen. Notably, $\widehat{\gamma}$'s error in estimating linear combinations of θ does not surge even when $p > n$. However, the difference between β and θ introduces a large bias to $\widehat{\gamma}$. Besides, when $p > n$, assumption 6 can be violated. Correspondingly the variance estimator $\widehat{\sigma}^2$ may have a large error. The difference between β and θ invalidates the confidence region (29). For prediction region (36), this problem still exists. However, the prediction region catches non-negligible errors apart from the asymptotically negligible errors and it is wider than the confidence region. Consequently, as long as the absolute values of difference are small, the prediction interval's performance will not be severely affected.

TABLE 3

Frequency of model misspecification; average errors of $\widehat{\gamma}$ and $\widehat{\sigma}^2$; and the coverage probability for the confidence region (29) and the prediction region (36). The nominal coverage probability is $1-\alpha = 95\%$. The overscore represents calculating the sample mean among 1000 simulations. We choose the number of bootstrap replicates $B = 500$.

Estimation and Confidence region construction					Prediction
Case #	$P(\widehat{\mathcal{N}}_{b_n} \neq \mathcal{N}_{b_n})$	$\overline{\max_{i=1,2,...,p_1} \lvert\widehat{\gamma}_i - \gamma_i\rvert}$	$\overline{\lvert\widehat{\sigma}^2 - \sigma^2\rvert}$	coverage	coverage
1	0.0	0.185	0.144	95.4%	91.5%
2	0.0	0.183	0.228	93.6%	90.4%
3	0.0	0.209	0.232	95.9%	92.6%
4	0.0	0.191	0.224	95.3%	90.6%
5(use β)	0.129	1.578	1.341	0.0%	97.2%
5(use θ)	0.122	0.258	1.354	97.6%	98.2%
6(use β)	0.126	1.579	1.342	0.0%	94.6%
6(use θ)	0.137	0.258	1.364	97.3%	92.8%

Figure 5 plots the power curve of the hypothesis test of $\gamma = \gamma_0$ versus $\gamma \neq \gamma_0$; here, we use $\gamma_0 = \gamma + \delta \times (1, 1, ..., 1)^T$ and $\delta > 0$.

7. Conclusion. The paper at hand proposes an improved, i.e., debiased and thresholded, ridge regression method that recovers the sparsity of parameters and avoids introducing a large bias. Besides, it derives a consistency result and the Gaussian approximation theorem for the improved ridge estimator. An asymptotically valid confidence region for $\gamma = M\beta$ and a hypothesis test of $\gamma = \gamma_0$ are also constructed based on a wild bootstrap algorithm. In addition, a novel, hybrid resampling procedure was proposed that can be used to perform interval prediction based on the improved ridge regression. When the dimension of parameters p is larger than the sample size n, the proposed method estimates linear combinations of $\theta = QQ^T\beta$ instead of linear combinations of β. If the underlying parameter is indeed β and the projection bias $\theta - \beta$ is not negligible, then the proposed methods may fail to provide a consistent result.

Numerical simulations indicate that improved ridge regression has comparable performance to the threshold Lasso while having at least two major advantages: (a) Ridge regression is easily computed using a closed-form expression, and (b) it appears to be quite robust against a non-optimal choice of the ridge parameter ρ_n as well as the threshold b_n. Therefore,

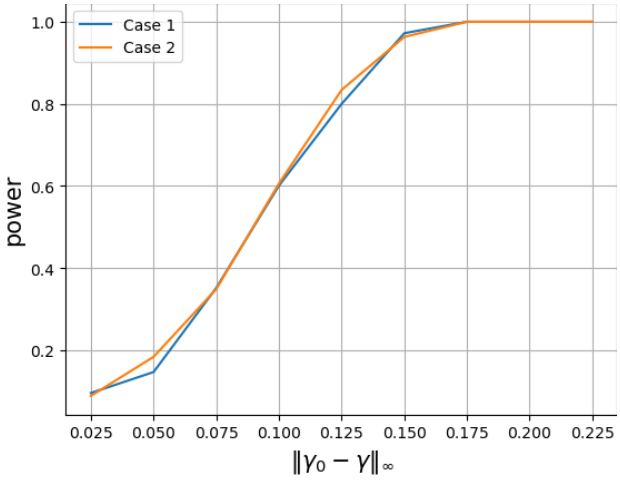


FIG 5. *Power of the test for cases 1 and 2; the x-axis represents* $\max_{i=1,\ldots,p_1} |\gamma_{0,i} - \gamma_i|$. *Nominal size for the test is* 5%*; see algorithm 1 for the meaning of notations.*

ridge regression may be found useful again in applied work using high-dimensional data as long as practitioners make sure to include debiasing and thresholding.

Acknowledgments. The authors are grateful to the anonymous referees for their valuable suggestions that significantly improve the content of this article.

Funding. This research was partially supported by NSF Grant DMS 19-14556.

SUPPLEMENTARY MATERIAL

Proofs

(DOI: 10.1214/[provided by typesetter]) Proofs of the aforementioned theorems will be presented in the Supplementary Material Zhang and Politis (2021a)

REFERENCES

BAI, Z. D. and YIN, Y. Q. (1993). Limit of the Smallest Eigenvalue of a Large Dimensional Sample Covariance Matrix. *Ann. Probab.* **21** 1275 – 1294.

BASU, S. and MICHAILIDIS, G. (2015). Regularized estimation in sparse high-dimensional time series models. *Ann. Statist.* **43** 1535 – 1567.

BICKEL, P. J., RITOV, Y. and TSYBAKOV, A. B. (2009). Simultaneous analysis of Lasso and Dantzig selector. *Ann. Statist.* **37** 1705 – 1732.

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

BÜHLMANN, P. and VAN DE GEER, S. (2011). *Statistics for high-dimensional data*. *Springer Series in Statistics*. Springer, Heidelberg. MR2807761

CHATTERJEE, A. and LAHIRI, S. N. (2010). Asymptotic properties of the residual bootstrap for Lasso estimators. *Proc. Amer. Math. Soc.* **138** 4497–4509.

CHATTERJEE, A. and LAHIRI, S. N. (2011). Bootstrapping Lasso Estimators. *J. Amer. Statist. Assoc.* **106** 608–625.

CHEN, X. and ZHOU, W.-X. (2020). Robust inference via multiplier bootstrap. *Ann. Statist.* **48** 1665–1691.

CHERNOZHUKOV, V., CHETVERIKOV, D. and KATO, K. (2013). Gaussian approximations and multiplier bootstrap for maxima of sums of high-dimensional random vectors. *Ann. Statist.* **41** 2786–2819.

CHERNOZHUKOV, V., WÜTHRICH, K. and ZHU, Y. (2019). Distributional conformal prediction. Preprint. arXiv:1909.07889.

DAI, L., CHEN, K., SUN, Z., LIU, Z. and LI, G. (2018). Broken adaptive ridge regression and its asymptotic properties. *J. Multivariate Anal.* **168** 334 - 351.

DALALYAN, A. S., HEBIRI, M. and LEDERER, J. (2017). On the prediction performance of the Lasso. *Bernoulli* **23** 552 – 581.

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

DOBRIBAN, E. and WAGER, S. (2018). High-dimensional asymptotics of prediction: Ridge regression and classification. *Ann. Statist.* **46** 247–279.

DOLADO, J. J. and LÜTKEPOHL, H. (1996). Making Wald tests work for cointegrated VAR systems. *Econometric Rev.* **15** 369-386.

FAN, J. and LI, R. (2001). Variable Selection via Nonconcave Penalized Likelihood and its Oracle Properties. *J. Amer. Statist. Assoc.* **96** 1348-1360.

FITHIAN, W., SUN, D. and TAYLOR, J. (2017). Optimal Inference After Model Selection. Preprint. arXiv:1410.2597.

GONÇALVES, S. and VOGELSANG, T. J. (2011). Block bootstrap HAC robust tests: the sophistication of the naive bootstrap. *Econometric Theory* **27** 745 - 791.

GREENSHTEIN, E. and RITOV, Y. (2004). Persistence in high-dimensional linear predictor selection and the virtue of overparametrization. *Bernoulli* **10** 971 – 988.

HORN, R. A. and JOHNSON, C. R. (2013). *Matrix analysis*, Second ed. Cambridge University Press, Cambridge. MR2978290

JAVANMARD, A. and JAVADI, H. (2019). False discovery rate control via debiased lasso. *Electron. J. Statist.* **13** 1212 – 1253.

JAVANMARD, A. and MONTANARI, A. (2018). Debiasing the lasso: Optimal sample size for Gaussian designs. *Ann. Statist.* **46** 2593–2622.

LEE, J. D., SUN, D. L., SUN, Y. and TAYLOR, J. E. (2016). Exact post-selection inference, with application to the lasso. *Ann. Statist.* **44** 907–927.

LEI, J. and WASSERMAN, L. (2014). Distribution-free prediction bands for non-parametric regression. *J. Roy. Statist. Soc. Ser. B* **76** 71–96.

LIU, R. Y. (1988). Bootstrap Procedures under some Non-I.I.D. Models. *Ann. Statist.* **16** 1696 – 1708.

LIU, H. and YU, B. (2013). Asymptotic properties of Lasso+mLS and Lasso+Ridge in sparse high-dimensional linear regression. *Electron. J. Statist.* **7** 3124–3169.

LOPES, M. (2014). A Residual Bootstrap for High-Dimensional Regression with Near Low-Rank Designs. In *Advances in Neural Information Processing Systems 27* 3239–3247.

MAMMEN, E. (1993). Bootstrap and Wild Bootstrap for High Dimensional Linear Models. *Ann. Statist.* **21** 255 – 285.

MAMMEN, E. (1996). Empirical process of residuals for high-dimensional linear models. *Ann. Statist.* **24** 307 – 335.

MEINSHAUSEN, N. and BÜHLMANN, P. (2006). High-dimensional graphs and variable selection with the Lasso. *Ann. Statist.* **34** 1436–1462.

MEINSHAUSEN, N. and YU, B. (2009). Lasso-type recovery of sparse representations for high-dimensional data. *Ann. Statist.* **37** 246–270.

POLITIS, D. N. (2015). *Model-free prediction and regression*. *Frontiers in Probability and the Statistical Sciences*. Springer, Cham. MR3442999

POLITIS, D. N., ROMANO, J. P. and WOLF, M. (1999). *Subsampling*. *Springer Series in Statistics*. Springer-Verlag, New York. MR1707286

ROMANO, Y., PATTERSON, E. and CANDÈS, E. (2019). Conformalized Quantile Regression. In *Advances in Neural Information Processing Systems* **32** 3543–3553. Curran Associates, Inc.

ROMANO, Y., SESIA, M. and CANDÈS, E. (2020). Deep Knockoffs. *J. Amer. Statist. Assoc.* **115** 1861-1872.

SHAO, J. (2003). *Mathematical statistics*, second ed. *Springer Texts in Statistics*. Springer-Verlag, New York. MR2002723

SHAO, J. and DENG, X. (2012). Estimation in high-dimensional linear models with deterministic design matrices. *Ann. Statist.* **40** 812–831.

STINE, R. A. (1985). Bootstrap Prediction Intervals for Regression. *J. Amer. Statist. Assoc.* **80** 1026-1031.

SUN, Y. (2011). Robust trend inference with series variance estimator and testing-optimal smoothing parameter. *J. Econometrics* **164** 345 - 366.

SUN, Y. (2013). A heteroskedasticity and autocorrelation robust F test using an orthonormal series variance estimator. *Econom. J.* **16** 1-26.

SUN, T. and ZHANG, C.-H. (2012). Scaled sparse linear regression. *Biometrika* **99** 879-898.

TIBSHIRANI, R. (1996). Regression Shrinkage and Selection Via the Lasso. *J. Roy. Statist. Soc. Ser. B* **58** 267-288.

TIBSHIRANI, R. J., RINALDO, A., TIBSHIRANI, R. and WASSERMAN, L. (2018). Uniform asymptotic inference and the bootstrap after model selection. *Ann. Statist.* **46** 1255–1287.

VAN DE GEER, S. A. (2008). High-dimensional generalized linear models and the lasso. *Ann. Statist.* **36** 614–645.

VAN DE GEER, S. (2019). On the asymptotic variance of the debiased Lasso. *Electron. J. Statist.* **13** 2970 – 3008.

VAN DE GEER, S., BÜHLMANN, P. and ZHOU, S. (2011). The adaptive and the thresholded Lasso for potentially misspecified models (and a lower bound for the Lasso). *Electron. J. Statist.* **5** 688 – 749.

WU, C. F. J. (1986). Jackknife, Bootstrap and Other Resampling Methods in Regression Analysis. *Ann. Statist.* **14** 1261 – 1295.

ZHANG, X. and CHENG, G. (2017). Simultaneous Inference for High-Dimensional Linear Models. *J. Amer. Statist. Assoc.* **112** 757-768.

ZHANG, Y. and POLITIS, D. N. (2021a). Supplement to "Ridge Regression Revisited: Debiasing, Thresholding and Bootstrap".

ZHANG, Y. and POLITIS, D. N. (2021b). Bootstrap prediction intervals with asymptotic conditional validity and unconditional guarantees. Preprint. arXiv:2005.09145.

ZHANG, D. and WU, W. B. (2017). Gaussian approximation for high dimensional time series. *Ann. Statist.* **45** 1895–1919.

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

ZHAO, P. and YU, B. (2006). On Model Selection Consistency of Lasso. *J. Mach. Learn. Res.* **7** 2541 - 2563.