Statistica Sinica **32** (2022), 915-937
doi:https://doi.org/10.5705/ss.202019.0283

PROJECTION-BASED INFERENCE FOR HIGH-DIMENSIONAL LINEAR MODELS

Sangyoon Yi and Xianyang Zhang

Texas A&M University

Abstract: We develop a new method to estimate the projection direction in the debiased Lasso estimator. The basic idea is to decompose the overall bias into two terms, corresponding to strong and weak signals, respectively. We propose estimating the projection direction by balancing the squared biases associated with the strong and weak signals and the variance of the projection-based estimator. A standard quadratic programming solver can solve the resulting optimization problem efficiently. We show theoretically that the unknown set of strong signals can be estimated consistently, and that the projection-based estimator enjoys asymptotic normality under suitable assumptions. A slight modification of our procedure leads to an estimator with a potentially smaller order of bias than that of the original debiased Lasso. We further generalize our method to conduct an inference for a sparse linear combination of the regression coefficients. Numerical studies demonstrate the advantage of the proposed approach in terms of coverage accuracy over several existing alternatives.

Key words and phrases: Confidence interval, high-dimensional linear models, Lasso, quadratic programming.

1. Introduction

Uncertainty quantification after model selection is an active field of research in statistics. The problem is challenging because the Lasso-type estimator does not admit a tractable asymptotic limit, owing to its noncontinuity at zero. Standard bootstrap and subsampling techniques cannot capture such noncontinuity and thus fail for the Lasso estimator, even in a low-dimensional regime. Several attempts have been made to tackle this challenge. For example, (Multi-) sample splitting and subsequent statistical inference procedures are developed in Wasserman and Roeder (2009) and Meinshausen, Meier and Bühlmann (2009). Meinshausen and Bühlmann (2010) proposed the so-called stability selection method, based on subsampling in combination with selection algorithms. Chatterjee and Lahiri (2011, 2013) considered bootstrap methods that provide valid approxima-

Corresponding author: Sangyoon Yi, Department of Statistics, Texas A&M University, College Station, TX 77843, USA. E-mail: syi@stat.tamu.edu.

tions to the limiting distributions of the Lasso and adaptive Lasso estimators.

For statistical inference after model selection, Berk et al. (2013) developed a post-selection inference procedure by reducing the problem to one of simultaneous inference. Lockhart et al. (2014) constructed a statistic from the Lasso solution path, and showed that it converges to a standard exponential distribution. To account for the selection effects, Lee et al. (2016) developed an exact post-selection inference procedure by characterizing the distribution of a post-selection estimator conditioned on the selection event. By leveraging the same core statistical framework, Tibshirani et al. (2016) proposed a general scheme to derive post-selection hypothesis tests at any step of forward-stepwise and least-angle regressions as well as any step along the Lasso regularization path. Barber and Candès (2015) proposed an inferential procedure by adding knockoff variables to create certain symmetry among the original variables and their knockoff copies. By exploring this symmetry, they showed that the method provides finite-sample false discovery rate control. The knockoff procedure is extended to the high-dimensional linear model in Barber and Candès (2019), and to settings in which the conditional distribution of the response is completely unknown in Candès et al. (2018).

More closely related to the current work, Zhang and Zhang (2014) introduced the idea of regularized projection, which is further explored and extended in van de Geer et al. (2014) and Javanmard and Montanari (2014). The common idea is to find a projection direction designed to remove the bias term in the Lasso estimator. The resulting debiased Lasso estimator, which is no longer sparse, is shown to admit an asymptotic normal limit. To find the projection direction, the nodewise Lasso regression of Meinshausen and Bühlmann (2006) was adopted in both Zhang and Zhang (2014) and van de Geer et al. (2014), whereas Javanmard and Montanari (2014) considered a convex optimization problem to approximate the precision matrix of the design. Zhang and Cheng (2017) and Dezeure, Bühlmann and Zhang (2017) proposed boostrap-assisted procedures for simultaneous inferences based on the debiased Lasso estimators. Belloni, Chernozhukov and Hansen (2014) developed a two-stage procedure with the so-called post-double-selection as the first stage and a least squares estimation as the second stage. Ning and Liu (2017) proposed a decorrelated score test in a likelihood based framework. Zhu and Bradic (2018a,b) developed projection-based methods that are robust to the lack of sparsity in the model parameter. More recent advances along this line include Neykov et al. (2018) and Chang et al. (2020). Focusing on the theoretical aspects of the debiased Lasso, Javanmard and Montanari (2018) studied its optimal sample size and Cai and Guo (2017) showed that the debiased estimator achieves the minimax rate. Although the methodology and theory for the debiased Lasso estimator are elegant, its empirical performance could be undesirable. For instance, the average coverage rate for active variables could be far lower than the nominal levels in a finite sample (see, e.g., van de Geer et al. (2014)).

A natural question to ask is whether there exist alternative projection directions that can improve the finite-sample performance in the original debiased Lasso estimator. In this paper, we propose a new method for estimating the projection direction and construct a novel bias-reducing projection (BRP) estimator that is designed to further reduce the bias of the original debiased Lasso estimator. In contrast to the nodewise Lasso adopted in both Zhang and Zhang (2014) and van de Geer et al. (2014), we propose a direct approach to estimate the projection direction. Our method is related to the procedure in Javanmard and Montanari (2014), but differs in the following respects. (i) We formulate a different objective function that appropriately balances the squared bias and the variance of the BRP estimator. (ii) We decompose the bias term into two parts based on a preliminary estimate of the signal strength: one associated with the strong signals, and the other related to the weak signals and noise. (iii) We develop new methods to estimate the set of strong signals and to select the tuning parameters involved in the objective function.

Our approach relies crucially on the following observation in a finite sample: the bias term associated with the strong signals contributes more to the overall bias. Motivated by this fact, we estimate the projection direction by minimizing an objective function that assigns different weights to the squared bias terms associated with the strong and weak signals. The set of strong signals is unknown, but can be estimated consistently based on a preliminary debiased Lasso estimator. The resulting optimization problem can be cast as a quadratic programming problem that can be solved efficiently using a standard quadratic programming solver. We use a residual bootstrap to estimate the coverage probabilities associated with different choices of weights. Then, we select the one that delivers the shortest interval width, while ensuring that the bootstrap estimate of the coverage probability is close to the nominal level.

We show theoretically that the unknown set of strong signals can be estimated consistently using a surrogate set based on a preliminary projection-based Lasso estimator, where the projection direction is obtained using a novel formulation. The BRP estimator is shown to enjoy asymptotic normality under suitable assumptions. As one of the main contributions, we prove that a slight modification of our BRP estimator leads to an estimator with a potentially smaller order of bias than that of the original debiased Lasso. We further generalize our BRP estimator to conduct a statistical inference for a sparse linear combination of the regression coefficients, under suitable assumptions on a loading vector. We demonstrate the usefulness of the proposed approach by comparing it with several current approaches using simulations.

The rest of the paper is organized as follows. We introduce the projection-based estimator and develop a new formulation to find the projection direction in Section 2. We propose a method to estimate the set of strong signals and show its consistency in Section 3.1. We establish the asymptotic normality of the BRP estimator in Section 3.2. In Section 3.3, we propose a modified BRP estimator that could result in a potentially smaller order of bias than that of the original debiased Lasso. Section 4 generalizes the method to conduct an inference for a linear combination of the regression coefficients. We develop a bootstrap-assisted procedure for choosing the tuning parameters in Section 5. Section 6 presents some numerical results. Section 7 concludes the paper. All technical details and additional numerical results are gathered in Supplementary Material.

Throughout this paper, we use the following notation. For a matrix $\mathbf{A} \in \mathbb{R}^{d\times d}$ and two sets $I, J \subseteq [d] := \{1, 2, \ldots, d\}$, denote by $\mathbf{A}_{I,J}$ ($\mathbf{A}_{-I,-J}$) the submatrix of $\mathbf{A}$ with (without) the rows in I and columns in J. Write $\mathbf{A}_{[d],-I} = \mathbf{A}_{-I}$. Similarly, for a vector $a \in \mathbb{R}^q$, write a_I (a_{-I}) as the subvector of a with (without) the components in I. Let $\|a\|_q$, with $0 \le q \le \infty$, be the l_q-norm of a, and write $\|a\| = \|a\|_2$. For two sets $\mathcal{S}_1, \mathcal{S}_2$, let $\mathcal{S}_1 \setminus \mathcal{S}_2$ be the set of elements in $\mathcal{S}_1$, but not in $\mathcal{S}_2$. Denote $|\mathcal{S}_1|$ as the cardinality of $\mathcal{S}_1$. For a square matrix $\mathbf{A}$, let $\lambda_{\max}(\mathbf{A})$ and $\lambda_{\min}(\mathbf{A})$ be its largest and smallest eigenvalues, respectively. Define $\|\mathbf{A}\| = \|\mathbf{A}\|_{\text{op}} = \sup_{a\in\mathcal{S}^{d-1}} \|\mathbf{A}a\|$ as the operator norm of $\mathbf{A}$, where $\mathcal{S}^{d-1}$ is the unit sphere in $\mathbb{R}^d$. The sub-gaussian norm of a random variable X, which we denote by $\|X\|_{\psi_2}$, is defined as $\|X\|_{\psi_2} = \sup_{q\ge1} q^{-1/2}(E|X|^q)^{1/q}$. For a random vector $X \in \mathbb{R}^d$, its sub-gaussian norm is defined as $\|X\|_{\psi_2} = \sup_{a\in\mathcal{S}^{d-1}} \|a^\top X\|_{\psi_2}$. The sub-exponential norm of a random variable X, which we denote by $\|X\|_{\psi_1}$, is defined as $\|X\|_{\psi_1} = \sup_{q\ge1} q^{-1}(E|X|^q)^{1/q}$. For a random vector $X \in \mathbb{R}^d$, its sub-exponential norm is defined as $\|X\|_{\psi_1} = \sup_{a\in\mathcal{S}^{d-1}} \|a^\top X\|_{\psi_1}$. Let $(\mathcal{M}, \rho)$ be a metric space, and let $\varepsilon > 0$. A subset $\mathcal{N}_\varepsilon$ of $\mathcal{M}$ is called an ε-net of $\mathcal{M}$ if every point $x \in \mathcal{M}$ can be approximated within ε by some point $y \in \mathcal{N}_\varepsilon$, that is, $\rho(x, y) \le \varepsilon$. The minimal cardinality of an ε-net of $\mathcal{M}$ is called the covering number of $\mathcal{M}$.

2. Projection-Based Estimator

To illustrate the idea, we focus on the high-dimensional linear model such that

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

where $Y = (y_1, \ldots, y_n)^\top \in \mathbb{R}^{n\times 1}$ is the response vector, $\mathbf{X} = (X_1, \ldots, X_p) \in \mathbb{R}^{n\times p}$ is the design matrix, $\beta = (\beta_1, \ldots, \beta_p)^\top \in \mathbb{R}^{p\times 1}$ is the vector of unknown regression coefficients with $\|\beta\|_0 = s_0$, and $\epsilon = (\epsilon_1, \ldots, \epsilon_n)^\top$ is a vector of independent errors with common variance σ^2.

2.1. Motivation

Suppose we are interested in conducting an inference for a single regression coefficient β_j, for $1 \le j \le p$. We first rewrite model (2.1) as

$$\eta_j := Y - \mathbf{X}_{-j}\beta_{-j} = X_j\beta_j + \epsilon. \tag{2.2}$$

If the value of η_j is known, the problem reduces to an inference about β_j in a simple linear regression model. Because η_j is not directly observable, a natural idea is to replace η_j with a suitable estimator, defined as

$$\hat{\eta}_j = Y - \mathbf{X}_{-j}\hat{\beta}_{-j} = X_j\beta_j + \epsilon + \mathbf{X}_{-j}(\beta_{-j} - \hat{\beta}_{-j}), \tag{2.3}$$

where $\hat{\beta}$ is a preliminary estimator for β. Here, (2.3) approximates (2.2), with the extra term $\mathbf{X}_{-j}(\beta_{-j} - \hat{\beta}_{-j})$ due to the estimation effect, by replacing β_{-j} with $\hat{\beta}_{-j}$. In this paper, we focus on the Lasso estimator given by

$$\hat{\beta} = \underset{\tilde{\beta}\in\mathbb{R}^p}{\operatorname{argmin}} \left\{ \frac{1}{2n}\|Y - \mathbf{X}\tilde{\beta}\|^2 + \lambda\|\tilde{\beta}\|_1 \right\},$$

the properties of which are now well understood (see, e.g., Bühlmann and van de Geer (2011); Hastie, Tibshirani and Wainwright (2015)). We also try the alternative Lasso formulation without penalizing β_j in our numerical studies, and find that it does not improve the finite-sample performance. Now, given a projection vector $v_j = (v_{j,1}, \ldots, v_{j,n})^\top \in \mathbb{R}^{n\times 1}$ such that $v_j^\top X_j = n$, we define the projection-based estimator for β_j as

$$\tilde{\beta}_j(v_j) := \frac{1}{n}v_j^\top\hat{\eta}_j = \beta_j + \frac{1}{n}v_j^\top\epsilon + R(v_j, \beta_{-j}), \tag{2.4}$$

where $R(v_j, \beta_{-j}) = n^{-1}v_j^\top\mathbf{X}_{-j}(\beta_{-j} - \hat{\beta}_{-j})$ is the bias term caused by the estimation effect. Here, (2.4) implies that

$$\sqrt{n}(\tilde{\beta}_j(v_j) - \beta_j) = \frac{1}{\sqrt{n}} v_j^\top \epsilon + \sqrt{n} R(v_j, \beta_{-j}).$$

To ensure that $\tilde{\beta}_j(v_j)$ has an asymptotically tractable limiting distribution, we require the bias term $\sqrt{n}R(v_j, \beta_{-j})$ to be dominated by the leading term $n^{-1/2} v_j^\top \epsilon$, which converges to a normal limit under suitable assumptions. In other words, the bias term $\sqrt{n}R(v_j, \beta_{-j})$ controls the normality of $\tilde{\beta}_j(v_j)$. A practical challenge here is that the bias $\sqrt{n}R(v_j, \beta_{-j})$ cannot be estimated directly from the data. Thus, it is common in the literature to replace $|\sqrt{n}R(v_j, \beta_{-j})|$ with a conservative estimator using the $l_1 - l_\infty$ bound, that is,

$$\|\sqrt{n}(\beta_{-j} - \hat{\beta}_{-j})\|_1 \|n^{-1} v_j^\top \mathbf{X}_{-j}\|_\infty, \tag{2.5}$$

as in Zhang and Zhang (2014), van de Geer et al. (2014), and Javanmard and Montanari (2014). Note that the variance of $n^{-1/2} v_j^\top \epsilon$ is equal to $\sigma^2 n^{-1} \|v_j\|^2$. To achieve efficiency, we also try to minimize $\sigma^2 n^{-1} \|v_j\|^2$, given that the bias $\sqrt{n}R(v_j, \beta_{-j})$ is properly controlled. Because the first term in (2.5) is independent of v_j, we can seek a projection direction to minimize a linear combination of $\|n^{-1} v_j^\top \mathbf{X}_{-j}\|_\infty^2$ and the variance $\sigma^2 n^{-1} \|v_j\|^2$. However, the $l_1 - l_\infty$ bound on the whole bias term could be conservative, because it does not take into account the specific form of the bias term. Note that the bias term can be written as

$$\begin{aligned}
\sqrt{n}R(v_j, \beta_{-j}) &= \frac{1}{\sqrt{n}} \sum_{k \neq j} v_j^\top X_k (\beta_k - \hat{\beta}_k) \\
&= \frac{1}{\sqrt{n}} \sum_{k \in \mathcal{S}_j^{(1)}(\nu)} v_j^\top X_k (\beta_k - \hat{\beta}_k) + \frac{1}{\sqrt{n}} \sum_{k \in \mathcal{S}_j^{(2)}(\nu)} v_j^\top X_k (\beta_k - \hat{\beta}_k) \\
&= \sqrt{n} R_{(1)}(v_j, \beta_{-j}) + \sqrt{n} R_{(2)}(v_j, \beta_{-j}),
\end{aligned} \tag{2.6}$$

where $\mathcal{S}_j^{(1)}(\nu) := \mathcal{S}(\nu) \setminus \{j\}$ and $\mathcal{S}_j^{(2)}(\nu) := \mathcal{S}(\nu)^\complement \setminus \{j\}$ denote the index sets (except j) associated with the strong and weak signals, respectively, for $\mathcal{S}(\nu) := \{k : |\beta_k| \geq \nu\}$, and $R_{(1)}(v_j, \beta_{-j})$ and $R_{(2)}(v_j, \beta_{-j})$ are defined accordingly. Here, ν is a threshold that separates the coefficients into two-groups, namely, the group with strong signals, and the group with weak signals or zero signal. For example, one can set $\nu = c_0 \sqrt{\log(p)/n}$ for some large enough constant c_0, which is the minimax rate for support recovery.

The formulation (2.6), using the decomposition associated with the signal strengths, can be motivated empirically. Specifically, it generally provides a smaller bias than the one without such a decomposition using the simulated data. Figure 4 illustrates one such case, where we compare the biases for projection vectors calculated based on two different methods: one solves (2.8) using the estimated set of strong signals, as in Section 3.1 (denoted by "With Decomposition"); the other solves the same problem, but with $\mathcal{A}_j^{(1)} = \emptyset$ (denoted by "Without Decomposition"). It can be seen that "With Decomposition" shows a smaller bias than "Without Decomposition." Similar results were observed in various simulation settings.

2.2. A new projection direction

In this subsection, we propose a novel formulation to find the projection direction. When $|\mathcal{S}_j^{(1)}(\nu)| \leq n$, we have the freedom to choose v_j to make the term $\|n^{-1}v_j^\top \mathbf{X}_{\mathcal{S}_j^{(1)}(\nu)}\|_\infty$ arbitrarily small. In fact, we can always choose v_j such that it is orthogonal to all X_k, with $k \in \mathcal{S}_j^{(1)}(\nu)$. The basic idea here is to find a projection direction v_j such that it is "more orthogonal" to the space spanned by $\{X_k\}_{k\in\mathcal{S}_j^{(1)}(\nu)}$ as compared to the space spanned by $\{X_k\}_{k\in\mathcal{S}_j^{(2)}(\nu)}$. With this intuition in mind, and the goal of balancing the squared bias with the variance, we formulate the following optimization problem:

$$\begin{aligned}&\min_{v_j}\left(\gamma_1 \max_{k\in S_j^{(1)}(\nu)} |n^{-1}v_j^\top X_k|^2 + \gamma_2 \max_{k\in S_j^{(2)}(\nu)} |n^{-1}v_j^\top X_k|^2 + \sigma^2 n^{-1}\|v_j\|^2\right),\\ &\text{s.t.} \quad v_j^\top X_j = n, \qquad (2.7)\end{aligned}$$

where $\gamma_1, \gamma_2 > 0$ are tuning parameters that control the trade-off between the squared bias and the variance. The term $\gamma_1 \max_{k\in S_j^{(1)}(\nu)} |n^{-1}v_j^\top X_k|^2$ ($\gamma_2 \max_{k\in S_j^{(2)}(\nu)} |n^{-1}v_j^\top X_k|^2$) corresponds to the $l_1 - l_\infty$ bound for $R_{(1)}^2$ ($R_{(2)}^2$). By introducing two ancillary variables u_{j1}, u_{j2}, problem (2.7) can be cast as the following quadratic programming problem:

$$\begin{aligned}&\min_{u_{j1},u_{j2},v_j} (\gamma_1 u_{j1}^2 + \gamma_2 u_{j2}^2 + \sigma^2 n^{-1}\|v_j\|^2),\\ &\text{s.t.} \quad v_j^\top X_j = n,\\ &- u_{j1} \leq n^{-1}v_j^\top X_k \leq u_{j1}, \quad k \in \mathcal{S}_j^{(1)}(\nu),\\ &- u_{j2} \leq n^{-1}v_j^\top X_k \leq u_{j2}, \quad k \in \mathcal{S}_j^{(2)}(\nu),\end{aligned}$$

which can be solved efficiently using existing quadratic programming solvers.

In general, the set $\mathcal{S}_j^{(1)}(\nu)$ is unknown, and needs to be replaced by a surrogate set $\mathcal{A}_j^{(1)}$, with $|\mathcal{A}_j^{(1)}| \leq n$. In Section 3.1, we describe a method for selecting $\mathcal{A}_j^{(1)}$ based on preliminary projection-based estimators. We show that $\mathcal{A}_j^{(1)}$ converges asymptotically to a nonrandom limit, that is,

$$P\left(\mathcal{A}_j^{(1)} = \mathcal{B}_j^{(1)}\right) \to 1,$$

for a nonrandom subset $\mathcal{B}_j^{(1)}$ of $[p]$. Note that $\mathcal{B}_j^{(1)}$ does not need to agree with $\mathcal{S}_j^{(1)}(\nu)$ for our procedure to be valid. To ensure that the remainder term is negligible, the theoretical analysis in Section 3.2 suggests that γ_1 and γ_2 should both be of order $O\left(\sigma^2 \log p/n\right)$. Combining the above discussions, we now state the optimization problem for obtaining the optimal projection direction:

$$\begin{aligned}
&\min_{u_{j1},u_{j2},v_j} \left(C_1 \frac{n}{\log p} u_{j1}^2 + C_2 \frac{n}{\log p} u_{j2}^2 + n^{-1}\|v_j\|^2\right), \\
&\text{s.t. } \ v_j^\top X_j = n, \\
&- u_{j1} \leq n^{-1} v_j^\top X_k \leq u_{j1}, \quad k \in \mathcal{A}_j^{(1)}, \\
&- u_{j2} \leq n^{-1} v_j^\top X_k \leq u_{j2}, \quad k \in \mathcal{A}_j^{(2)},
\end{aligned} \tag{2.8}$$

where $\mathcal{A}_j^{(2)} := (\mathcal{A}_j^{(1)})^{\complement} \setminus \{j\}$, and $C_1, C_2 > 0$ are tuning parameters, the choice of which is discussed in Section 5.

Remark 1. A related method is the refitted Lasso by Liu and Yu (2013). The idea is to refit the model selected by the Lasso, and to conduct an inference based on the refitted least squares estimator. Such an estimator fits into the framework of projection-based estimators. To see this, let $\hat{S}$ be the set of active variables selected by the Lasso, and note that $\hat{\beta}_k = 0$, for $k \notin \hat{S}$. For each $j \in \hat{S}$, let $\hat{w}_j$ be the projection of X_j onto the orthogonal space of $\mathbf{X}_{\hat{S}\setminus\{j\}}$. Then, the refitted least squares estimator is given by $\hat{w}_j^\top (Y - X_{-j}\hat{\beta}_{-j})/(\hat{w}_j^\top X_j)$. It is easy to see that the bias for this estimator is proportional to $\sum_{k\notin\hat{S}} \hat{w}_j^\top X_k \beta_k$, which disappears when the selected model contains all significant variables. However, when the model selection consistency fails, this procedure is no longer valid owing to the nonnegligible bias.

3. Methodology

3.1. Surrogate set

We describe a procedure to estimate the set of strong signals based on a preliminary projection-based estimator. Note that the estimator here is different from the original debiased Lasso because it is based on the novel formulation (2.8). Specifically, for some $\tau > 0$, we define our estimate for the set of strong signals as

$$\mathcal{A}(\tau) := \left\{ l : |T_l| > \sqrt{\tau \log p} \right\} \quad \text{where} \quad T_l = \frac{\sqrt{n}\tilde{\beta}_l(\hat{v}_l)}{\hat{\sigma} n^{-1/2} ||\hat{v}_l||}, \tag{3.1}$$

where $\hat{\sigma}$ is an estimator of the noise level σ, and $\tilde{\beta}_l(\hat{v}_l)$ is a projection-based estimator, with $\hat{v}_l$ being the solution to the following optimization problem:

$$\begin{aligned} &\min_{u_l, v_l} \left(C_0 \frac{n}{\log p} u_l^2 + n^{-1} \|v_l\|^2 \right), \\ &\text{s.t.} \quad v_l^\top X_l = n, \\ &- u_l \le n^{-1} v_l^\top X_k \le u_l, \quad k \ne l. \end{aligned} \tag{3.2}$$

In practice, both C_0 and τ need to be chosen appropriately. The details for the selection are discussed in Section S1. Note that (3.2) is a special case of (2.8) when we have no knowledge about the set of strong signals, that is, $\mathcal{A}_l^{(1)} = \emptyset$. We define the surrogate sets as

$$\mathcal{A}_j^{(1)}(\tau) := \mathcal{A}(\tau) \setminus \{j\}, \quad \mathcal{A}_j^{(2)}(\tau) := \mathcal{A}(\tau)^{\complement} \setminus \{j\}. \tag{3.3}$$

Throughout the paper, we consider the variance estimator

$$\hat{\sigma}^2 = \frac{1}{n} \|Y - \mathbf{X}\hat{\beta}\|^2, \tag{3.4}$$

which appears to outperform the estimator $\|Y - \mathbf{X}\hat{\beta}\|^2 / (n - \|\hat{\beta}\|_0)$ studied in Reid, Tibshirani and Friedman (2016); see Figure 22 in Supplementary Material for a comparison. Before presenting the main result of this subsection, we introduce some assumptions.

Assumption 1. *There exist a set $\mathcal{B} \subseteq [p] = \{1, 2, \ldots, p\}$ and $0 \le d_0 < d_1$ such that*

$$\begin{aligned} \max_{l \in \mathcal{B}^{\complement}} \frac{|\sqrt{n}\beta_l|}{\sigma} &\le \sqrt{d_0 \log p}, \\ \min_{l \in \mathcal{B}} \frac{|\sqrt{n}\beta_l|}{\sigma} &\ge \sqrt{d_1 \log p}. \end{aligned}$$

Assumption 2. *The error ϵ is a mean-zero sub-gaussian random vector with the sub-gaussian norm κ_ϵ.*

Assumption 3. *The preliminary estimator satisfies*

$$\sqrt{n}\|\hat{\beta}-\beta\|_1 = O_p\Big(s_0\sqrt{\log(p)}\Big).$$

Assumption 4. *The variance estimator* $\hat{\sigma}^2$ *is consistent in the sense that* $\hat{\sigma}/\sigma \xrightarrow{p} 1$.

Assumption 5. *Suppose the design matrix* $\mathbf{X} \in \mathbb{R}^{n\times p}$ *has independent and identically distributed (i.i.d.) rows with a zero population mean and covariance matrix* $\mathbf{\Sigma} = (\Sigma_{i,j})_{i,j=1}^{p}$. *Assume that*

1. $\max_j \Sigma_{j,j} < \infty$;
2. $\lambda_{\min}(\mathbf{\Sigma}) \geq \Lambda_{\min} > 0$;
3. *The rows of* $\mathbf{X}$ *are sub-gaussian with the sub-gaussian norm* $\kappa < \infty$.

Assumption 6. n, p, *and* s_0 *satisfy the rate condition* $s_0 \log p/\sqrt{n} = o(1)$.

Assumption 1 allows the strengths of strong and weak signals to be the same order, and thus is much weaker than the “beta-min” condition, which requires the weak signals to be of smaller order. Assumptions 3 and 4 are satisfied for the Lasso estimator and the variance estimator $\hat{\sigma}$ in (3.4) under suitable regularity conditions [Bühlmann and van de Geer (2011)]. Assumptions 2 and 5 require the error and design to be sub-gaussian. Similar assumptions have been made in van de Geer et al. (2014). Like Javanmard and Montanari (2014), the validity of our method does not rely on the sparsity of the precision matrix of the design, which is required in the nodewise Lasso regression for the original debiased Lasso. In view of Cai and Guo (2017), the rate condition in Assumption 6 cannot be relaxed without extra information. Zhu and Bradic (2018a,b) proposed testing procedures in high-dimensional linear models that impose much weaker restrictions on model sparsity or the loading vector representing the hypothesis. However, their methods require certain auxiliary sparse models, which are not needed for our procedure.

Define $\mathbf{\Sigma}_{j\backslash -j} = \mathbf{\Sigma}_{j,j} - \mathbf{\Sigma}_{j,-j}\mathbf{\Sigma}_{-j,-j}^{-1}\mathbf{\Sigma}_{-j,j}$ and $\kappa_{0j} = 2(1+\sqrt{\Lambda_{\min}^{-1}\mathbf{\Sigma}_{j,j}})\kappa^2$, for $1 \leq j \leq p$. The following proposition shows that the surrogate set $\mathcal{A}_j^{(1)}(\tau)$ with a properly chosen τ converges to $\mathcal{B}\setminus\{j\}$.

Proposition 1. *Define* $\mathcal{A}_j^{(1)}(\tau)$ *and* $\mathcal{A}_j^{(2)}(\tau)$ *as in* (3.3), *and let* $\hat{v}_l$ *be the solution to* (3.2) *for* $l \neq j$. *Suppose* d_0, d_1, *and* τ *satisfy*

$$\frac{\sigma^2}{32e\kappa_\epsilon^2}\left(\sqrt{\tau} - \sqrt{d_0 \max_l \mathbf{\Sigma}_{l,l}}\right)^2 > 1,$$

and $\sqrt{d_1/M} - \sqrt{\tau} > 0$*, where*

$$M = \left(\min_{1\le l\le p} \mathbf{\Sigma}_{l\backslash -l}\right)^2 \left(2C_0 \left(\min_{1\le l\le p} \frac{1}{8e^2}\frac{1}{(\kappa_{0l})^2}\right)^{-1} + \max_{1\le l\le p} \mathbf{\Sigma}_{l\backslash -l}\right).$$

Then under Assumptions 1–6*, we have*

$$\mathbb{P}\left(\max_{l\in\mathcal{B}_j^{(2)}} |T_l| \le \sqrt{\tau \log p}\right) \to 1,$$
$$\mathbb{P}\left(\min_{l\in\mathcal{B}_j^{(1)}} |T_l| > \sqrt{\tau \log p}\right) \to 1,$$

where $\mathcal{B}_j^{(1)} := \mathcal{B} \setminus \{j\}$ *and* $\mathcal{B}_j^{(2)} := (\mathcal{B}_j^{(1)})^{\complement} \setminus \{j\}$*. As a consequence,* $\mathbb{P}(\mathcal{A}_j^{(1)}(\tau) = \mathcal{B}_j^{(1)}) \to 1$*.*

Remark 2. As shown in Proposition 1, the surrogate set in (3.3) has an asymptotic (nonrandom) limit, which implies that the projection direction obtained in (2.8) is asymptotically independent of the random error ϵ. This fact is useful in the proof of Theorem 1 later. To ensure the independence between the projection direction and the random error, we can also employ a sample splitting strategy. That is, we split the samples into two subsamples, estimate the set of strong signals based on the first subsample, and construct the projection-based estimator based on another subsample. Because we use all samples in building the projection-based estimator, our method is more efficient than the sample splitting strategy.

Remark 3. When $d_0 = 0$, $\mathcal{B}$ coincides with the support of β. Proposition 1 suggests that one can consistently recover the support of β by thresholding the projection-based estimator.

3.2. BRP estimator

In this subsection, we introduce the BRP estimator and study its asymptotic behavior. Let $\tilde{v}_j$ be the solution to (2.8) based the surrogate sets in (3.3). Then, the BRP estimator $\tilde{\beta}_j(\tilde{v}_j)$ is defined as

$$\tilde{\beta}_j(\tilde{v}_j) = \frac{1}{n}\tilde{v}_j^\top \hat{\eta}_j = \frac{1}{n}\tilde{v}_j^\top (Y - \mathbf{X}_{-j}\hat{\beta}_{-j}).$$

In the following, we introduce two asymptotic results, depending on whether the surrogate set is estimated from the same data set used to find the projection direction. We first state the following theorem on the asymptotic normality when the surrogate set is estimated using (3.3).

Theorem 1. *Denote by* $\tilde{v}_j$ *the solution to* (2.8)*, with* $\mathcal{A}_j^{(1)}(\tau)$ *and* $\mathcal{A}_j^{(2)}(\tau)$ *in* (3.3)*. Suppose the assumptions in Proposition* 1 *hold, and further assume that, for some* $\delta > 0$,

$$\|\tilde{v}_j\|_{2+\delta} = o_{a.s.}(\|\tilde{v}_j\|). \tag{3.5}$$

Then, we have

$$\frac{\sqrt{n}\left(\tilde{\beta}_j(\tilde{v}_j) - \beta_j\right)}{\hat{\sigma} n^{-1/2}\|\tilde{v}_j\|} \xrightarrow{d} N(0,1). \tag{3.6}$$

Thus, an asymptotic $100(1-\alpha)\%$ *confidence interval for* β_j *is given by*

$$CI(1-\alpha) = \left\{ b \in \mathbb{R} : \left| \frac{\sqrt{n}(\tilde{\beta}_j(\tilde{v}_j) - b)}{\hat{\sigma} n^{-1/2}\|\tilde{v}_j\|} \right| \le z_{1-\alpha/2} \right\}, \tag{3.7}$$

where $z_{1-\alpha/2}$ *is the* $1-\alpha/2$ *quantile of* $N(0,1)$.

(3.5) is a Lyapunov-type condition, which implies the central limit theorem. This type of assumption on the projection direction is also imposed in Dezeure, Bühlmann and Zhang (2017). It can be dropped under the Gaussian assumption on the errors. If the surrogate set is chosen based on prior knowledge or estimated from an independent data set (e.g., based on sample splitting), then Assumptions 1-2 can be relaxed and we have the following result.

Corollary 1. *Suppose the surrogate set* $\mathcal{A}_j^{(1)}$ *is independent of the data. Under Assumptions* 3–6 *and assuming that for some* $\delta > 0$, $E[|\epsilon_i|^{2+\delta}] < \infty$ *and* $\|\tilde{v}_j\|_{2+\delta} = o_{a.s.}(\|\tilde{v}_j\|)$*, then* (3.6) *still holds.*

3.3. Modified BRP estimator

We introduce a modified bias-reducing projection (MBRP) estimator that is motivated by Proposition 1 and the refitted Lasso. This new estimator leads to a potentially smaller order of bias compared to that of the original debiased Lasso estimator under suitable assumptions, as shown in Proposition 2. Thus, it is expected to provide better empirical coverage probability; see Section 6. To motivate the MBRP estimator, we note that the bias associated with the BRP estimator based on some estimator $\check{\beta}$ for β can be written as

$$\sqrt{n}R(v_j, \beta_{-j}) = \frac{1}{\sqrt{n}} \sum_{k \ne j} v_j^\top X_k(\beta_k - \check{\beta}_k)$$

$$= \frac{1}{\sqrt{n}} \sum_{k \in \mathcal{B}_j^{(1)}} v_j^\top X_k(\beta_k - \check{\beta}_k) + \frac{1}{\sqrt{n}} \sum_{k \in \mathcal{B}_j^{(2)}} v_j^\top X_k(\beta_k - \check{\beta}_k),$$

where $\mathcal{B}_j^{(1)}, \mathcal{B}_j^{(2)}$ are as in Proposition 1. When $|\mathcal{B}_j^{(1)}| \leq n$, we can always require v_j to be exactly orthogonal to $\mathbf{X}_{\mathcal{B}_j^{(1)}}$. Therefore, the bias associated with the set of strong signals becomes zero. Thus, it suffices to control the bias term associated with $\mathcal{B}_j^{(2)}$ by properly choosing v_j and $\check{\beta}$, as clarified below.

To find the projection direction for the MBRP estimator, we consider the optimization problem:

$$\begin{aligned} &\min_{u_{j2}, v_j} \left(C_2 \frac{n}{\log p} u_{j2}^2 + n^{-1} \|v_j\|^2 \right), \\ &\text{s.t. } \; v_j^\top X_j = n, \\ &n^{-1} v_j^\top X_k = 0, \quad k \in \mathcal{A}_j^{(1)}, \\ &- u_{j2} \leq n^{-1} v_j^\top X_k \leq u_{j2}, \quad k \in \mathcal{A}_j^{(2)}. \end{aligned} \tag{3.8}$$

In contrast to (2.8), we require the projection direction to be orthogonal to the column space of $\mathbf{X}_{\mathcal{A}_j^{(1)}}$ in (3.8). Instead of using the Lasso estimator $\hat{\beta}$, we adopt the refitted least squares estimator $\check{\beta}$ as our preliminary estimator; that is,

$$\check{\beta}_{\mathcal{A}_j^{(1)}} = \operatorname*{argmin}_{\tilde{\beta}} \frac{1}{2n} \|Y - \mathbf{X}_{\mathcal{A}_j^{(1)}} \tilde{\beta}\|^2, \quad \check{\beta}_{\mathcal{A}_j^{(2)}} = 0. \tag{3.9}$$

The MBRP estimator is then defined as

$$\tilde{\beta}_j(\bar{v}_j) = \frac{1}{n} \bar{v}_j^\top (Y - \mathbf{X}_{-j} \check{\beta}_{-j}) = \beta_j + \frac{1}{n} \bar{v}_j^\top \epsilon + R(\bar{v}_j, \beta_{-j}), \tag{3.10}$$

where $R(\bar{v}_j, \beta_{-j}) = n^{-1} \bar{v}_j^\top \mathbf{X}_{-j}(\beta_{-j} - \check{\beta}_{-j})$, and $\bar{v}_j$ is the solution to problem (3.8). The MBRP estimator can be viewed as an intermediate estimator between the refitted Lasso and the BRP estimator based on (2.8). Here, (3.8) is a variant of (2.8) seeking a projection direction that is exactly orthogonal to the column space of $\mathbf{X}_{\mathcal{A}_j^{(1)}}$. In contrast, the modified procedure uses the refitted estimator for β, as the refitted Lasso does, as noted in Remark 1.

We argue that the bias term $\sqrt{n} R(\bar{v}_j, \beta_{-j})$ that controls normality could have a potentially smaller order than that of the original debiased Lasso estimator.

Proposition 2. *Denote by $\bar{v}_j$ the solution to* (3.8), *with $\mathcal{A}_j^{(1)}(\tau)$ and $\mathcal{A}_j^{(2)}(\tau)$ defined in* (3.3). *Let $\check{\beta}$ be the refitted least squares estimator in* (3.9). *Conditional on the event $\{\mathcal{A}_j^{(2)} = \mathcal{B}_j^{(2)}\}$, we have*

$$|\sqrt{n}R(\bar{v}_j, \beta_{-j})| \leq O_p\left(\sqrt{d_0}\|\beta_{\mathcal{B}_j^{(2)}}\|_0 \frac{\log p}{\sqrt{n}}\right), \tag{3.11}$$

under Assumptions 1 and 5. If we further assume that

$$\sqrt{d_0}\|\beta_{\mathcal{B}_j^{(2)}}\|_0 = o(s_0), \tag{3.12}$$

the bias $\sqrt{n}R(\bar{v}_j, \beta_{-j})$ *is asymptotically negligible with smaller order than that of the original debiased Lasso, given by* $O_p(s_0 \log p/\sqrt{n})$.

In particular, (3.12) holds if $d_0 = o(1)$ and $d_1 = O(1)$; that is, the strength of the weak signals is of smaller order than that of the strong signals. It is more stringent than Assumption 1, where the magnitudes of the set of strong signals and weak signals are allowed to be of the same order. However, note that Proposition 2 is not necessary for the asymptotic normality in Corollary 2 to be achieved. The following result shows the asymptotic normality of (3.10), which can be proved using similar arguments to those for Theorem 1.

Corollary 2. *Under the assumptions in Theorem* 1*, we have*

$$\frac{\sqrt{n}\left(\tilde{\beta}_j(\bar{v}_j) - \beta_j\right)}{\hat{\sigma}n^{-1/2}\|\bar{v}_j\|} \xrightarrow{d} N(0,1),$$

where $\tilde{\beta}_j(\bar{v}_j)$ *is defined in* (3.10) *and* $\bar{v}_j$ *is the solution to* (3.8).

4. Inference on a Sparse Linear Combination of Parameters

In some applications, one may be interested in conducting an inference on $a^\top\beta$ for a (sparse) loading vector $a = (a_1, \ldots, a_p)^\top \in \mathbb{R}^p$, with $\|a\|_0 = s \ll n$. Denote by $S = S(a) = \{1 \leq j \leq p : a_j \neq 0\}$ the support set of a. Our method can be generalized to construct an estimator and conduct an inference for $a^\top\beta = a_S^\top\beta_S$. Recall that $\hat{\beta}$ is the preliminary estimator of β. Define

$$\eta_S = Y - \mathbf{X}_{-S}\beta_{-S} = \mathbf{X}_S\beta_S + \epsilon$$

and

$$\hat{\eta}_S = Y - \mathbf{X}_{-S}\hat{\beta}_{-S} = \mathbf{X}_S\beta_S + \epsilon + \mathbf{X}_{-S}(\beta_{-S} - \hat{\beta}_{-S}).$$

We construct an estimator for $a^\top\beta$ in the form of $n^{-1}v_a^\top\hat{\eta}_S$, where $v_a = (v_{a,1}, \ldots, v_{a,n})^\top$ is a projection direction such that $n^{-1}v_a^\top\hat{\eta}_S$ has a tractable asymptotic limit. Note that

$$n^{-1}v_a^\top\hat{\eta}_S = n^{-1}v_a^\top\mathbf{X}_S\beta_S + n^{-1}v_a^\top\epsilon + n^{-1}v_a^\top\mathbf{X}_{-S}(\beta_{-S} - \hat{\beta}_{-S})$$

$$= a_S^\top \beta_S + (n^{-1} v_a^\top \mathbf{X}_S - a_S^\top)\beta_S + n^{-1} v_a^\top \epsilon + n^{-1} v_a^\top \mathbf{X}_{-S}(\beta_{-S} - \hat{\beta}_{-S}).$$

Under the equality constraint that $n^{-1} v_a^\top \mathbf{X}_S - a_S^\top = 0$, and by rearranging the above terms, we have

$$\sqrt{n}(n^{-1} v_a^\top \hat{\eta}_S - a_S^\top \beta_S) = n^{-1/2} v_a^\top \epsilon + \sqrt{n} R(v_a, \beta_{-S}), \tag{4.1}$$

where $R(v_a, \beta_{-S}) = n^{-1} v_a^\top \mathbf{X}_{-S}(\beta_{-S} - \hat{\beta}_{-S})$. Similarly to (2.6), the bias term can be decomposed into two parts, corresponding to different strengths of the signals. Let $\mathcal{A}_S^{(1)}$ be the surrogate set for the set of strong signals (excluding the elements in S), which can be obtained in a similar way to that described in Section 3.1. Following the derivations in Section 2, we formulate the following optimization problem to find v_a:

$$\begin{aligned} &\min_{u_{a1}, u_{a2}, v_a} \left(C_1 \frac{n}{\log p} u_{a1}^2 + C_2 \frac{n}{\log p} u_{a2}^2 + n^{-1} \|v_a\|^2 \right), \\ &\text{s.t.} \;\; v_a^\top X_S = n a_S^\top, \\ &- u_{a1} \le n^{-1} v_a^\top X_k \le u_{a1}, \quad k \in \mathcal{A}_S^{(1)}, \\ &- u_{a2} \le n^{-1} v_a^\top X_k \le u_{a2}, \quad k \in \mathcal{A}_S^{(2)}, \end{aligned} \tag{4.2}$$

where $\mathcal{A}_S^{(2)} := (\mathcal{A}_S^{(1)} \cup S)^\complement$. Denote by $(\tilde{u}_{a1}, \tilde{u}_{a2}, \tilde{v}_a)$ the solution to (4.2). Our estimator for $a^\top \beta$ is thus given by $n^{-1} \tilde{v}_a^\top \hat{\eta}_S$, the asymptotic normality of which is established in the following theorem.

Theorem 2. *With $\|a\|_0 = s \ll n$, suppose the assumptions in Proposition 1 hold and $\|\tilde{v}_a\|_{2+\delta} = o_{a.s.}(\|\tilde{v}_a\|)$, for some $\delta > 0$. Then, we have*

$$\frac{\sqrt{n}\left(n^{-1} \tilde{v}_a^\top \hat{\eta}_S - a^\top \beta\right)}{\hat{\sigma} n^{-1/2} \|\tilde{v}_a\|} \xrightarrow{d} N(0,1). \tag{4.3}$$

Thus, an asymptotic $100(1-\alpha)\%$ confidence interval for $a^\top \beta$ is given by

$$CI(1-\alpha) = \left\{ b \in \mathbb{R} : \left| \frac{\sqrt{n}\left(n^{-1} \tilde{v}_a^\top \hat{\eta}_S - b\right)}{\hat{\sigma} n^{-1/2} \|\tilde{v}_a\|} \right| \le z_{1-\alpha/2} \right\},$$

where $z_{1-\alpha/2}$ is the $1-\alpha/2$ quantile of $N(0,1)$.

We mention some existing works for inferences on linear combinations of β. When the sparsity level s_0 is known, Cai and Guo (2017) obtained the minimax expected length of the confidence intervals for $a^\top \beta$ in both the sparse and dense loading regions. They further showed that without knowledge of s_0, a rate-optimal adaptation in the sparse loading regime is only possible under Assumption 6, and in the dense loading regime, an adaptation to s_0 is impossible. In Zhu and Bradic (2018b), the authors propose a test for a linear hypothesis that does not impose a restriction on the model sparsity or the loading vector representing the hypothesis. Nevertheless, compared with our method, the method of Zhu and Bradic (2018b) requires an additional sparse model to account for the dependence between the so-called synthesized feature and the stabilized feature.

Parallel to Corollary 1, if the surrogate set is estimated based on prior information or an independent data set, Assumptions 1–2 can be dropped, and the asymptotic normality can be established as follows.

Corollary 3. *Suppose the surrogate set $\mathcal{A}_j^{(1)}$ is independent of the data. Under Assumptions 3–6 and further assuming that for some $\delta > 0$, $E[|\epsilon_i|^{2+\delta}] < \infty$ and $\|\tilde{v}_a\|_{2+\delta} = o_{a.s.}(\|\tilde{v}_a\|)$, (4.3) still holds.*

5. Selecting the Tuning Parameters

The bootstrap for the debiased Lasso is studied in Zhang and Cheng (2017) and Dezeure, Bühlmann and Zhang (2017) to approximate the sampling distribution of the debiased Lasso estimator. Here, we propose a bootstrap-assisted approach for choosing the tuning parameters in (2.8), (3.2), and (3.8). Specifically, the residual bootstrap is used to obtain the empirical coverage rate and its standard error for selecting the optimal tuning parameters. We focus our discussion on (2.8), and remark that the procedure is applicable to (3.2) and (3.8) as well. Let

$$\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)^\top = Y - \mathbf{X}\hat{\beta},$$

and let $\bar{\varepsilon}_i = \varepsilon_i - n^{-1}\sum_{j=1}^n \varepsilon_j$ be the centered residual, where $\hat{\beta}$ denotes the cross-validated Lasso estimator. Given a sequence of tuning parameters $\{(c_{1,j,(k)}, c_{2,j,(k)})\}_{k=1}^K$, we first calculate $\tilde{v}_j\left(c_{1,j,(k)}, c_{2,j,(k)}\right)$, which is the solution to (2.8) given $\left(c_{1,j,(k)}, c_{2,j,(k)}\right)$. Note that the projection direction $\tilde{v}_j$ only needs to be calculated once for each pair of tuning parameters. Given $\{\tilde{v}_j\left(c_{1,j,(k)}, c_{2,j,(k)}\right)\}_{k=1}^K$, we do the following.

1. To generate the bth bootstrap sample, we sample n residuals with replacement from $\{\bar{\varepsilon}_i\}_{i=1}^n$, and denote the corresponding samples by $\varepsilon_b^* = (\varepsilon_{b,1}^*, \ldots, \varepsilon_{b,n}^*)^\top$. Then, generate Y_b^* such that $Y_b^* = \mathbf{X}\hat{\beta} + \varepsilon_b^*$.

2. With (X, Y_b^*), calculate the cross-validated Lasso estimator $\hat{\beta}_b^*$ and the projection-based estimator

$$\tilde{\beta}_j(\tilde{v}_j(c_{1,j,(k)}, c_{2,j,(k)})) = \frac{\tilde{v}_j(c_{1,j,(k)}, c_{2,j,(k)})^\top (Y_b^* - \mathbf{X}_{-j}\hat{\beta}^*_{b,-j})}{n},$$

where $\hat{\beta}^*_{b,-j}$ denotes $\hat{\beta}^*_b$ without the jth component. We then calculate the $100(1-\alpha)\%$ confidence interval $\mathrm{CI}^*_{b,j,(k)}$ using (3.7). For each j, calculate $I(\hat{\beta}_j \in \mathrm{CI}^*_{b,j,(k)})$, which is one if $\hat{\beta}_j$ is covered by $\mathrm{CI}^*_{b,j,(k)}$, and zero otherwise. In addition, calculate the length of $\mathrm{CI}^*_{b,j,(k)}$, and denote it as $\mathrm{Len}^*_{b,j,(k)}$.

3. Repeat the above steps for B bootstrap samples. We choose the tuning parameters for β_j as

$$\begin{aligned}(c^*_{1,j,(k)}, c^*_{2,j,(k)}) &= \operatorname*{argmin}_k \mathrm{AvgLen}_{j,(k)} \\ \text{s.t. } \widehat{\mathrm{Cover}}_{j,(k)} &+ \mathrm{SE}(\widehat{\mathrm{Cover}}_{j,(k)}) \geq 1-\alpha,\end{aligned}$$

where $\mathrm{AvgLen}_{j,(k)} = B^{-1}\sum_{b=1}^{B} \mathrm{Len}^*_{b,j,(k)}$ and

$$\widehat{\mathrm{Cover}}_{j,(k)} = \frac{\sum_{b=1}^{B} I(\hat{\beta}_j \in \mathrm{CI}^*_{b,j,(k)})}{B},$$

$$\mathrm{SE}(\widehat{\mathrm{Cover}}_{j,(k)}) = \sqrt{\frac{\widehat{\mathrm{Cover}}_{j,(k)}(1-\widehat{\mathrm{Cover}}_{j,(k)})}{B}}.$$

In other words, the optimal pair of tuning parameters is selected, where the minimum average interval length among all pairs with an empirical coverage rate that increased by one standard error is at least the nominal level $1-\alpha$.

6. Numerical Results

6.1. Confidence interval for a single regression coefficient

We conduct simulations to evaluate the finite-sample performance of the proposed BRP and MBRP estimators. We use the R package `quadprog` to solve the quadratic programming problems in our methods, and the R package `doMC` with five cores for parallel computation. All remaining implementation details are as described in Section S1. For comparison, we implement the debiased Lasso of van de Geer et al. (2014) (denoted by DB), using the R package `hdi`, and the method of Javanmard and Montanari (2014) (denoted by JM), using the code posted on the authors' website. We encounter some numerical issues when implementing JM's code for the equicorrelation covariance structure of $\mathbf{X}$ in (ii). Therefore, we report only the results of JM for Toeplitz covariance structure of $\mathbf{X}$. In addition, we present the results of the double selection approach of

Belloni, Chernozhukov and Hansen (2014) (denoted by BCH), using the R package `hdm`. Owing to the high computational cost of BCH in the case of the equicorrelation covariance, we report only the result for the active set. We also implement the method of Zhu and Bradic (2018b) (denoted by "ZB" and "ZB2"). The only difference between ZB and ZB2 is in the choice of the constant c in the tuning parameter $\eta = \sqrt{c(\log p)/n}$ in (12) of their paper. In ZB, we set $c = 2$, as suggested by the authors, while in ZB2, we let $c = 10^{-3}$.

In (2.1), the rows of $\mathbf{X}$ are considered to be i.i.d. realizations from $N(0, \mathbf{\Sigma})$ with $\Sigma_{jj} = 1$ under two scenarios: (i) $\Sigma_{j,k} = 0.9^{|j-k|}$ (denoted as Tp); (ii) $\Sigma_{j,k} = 0.8$, for all $j \neq k$ (denoted as Eq). To generate β, we consider the following two cases:

Case 1: $\beta_j \overset{i.i.d.}{\sim} U(0, 4)$ with $s_0 = 3, 5, 10, 15$.

Case 2: Half of the nonzero β_j are independently generated from $U(0, 0.5)$, and the rest are generated from $U(2.5, 3)$, with $s_0 = 4, 8, 12, 16$.

The errors are independently generated from (a) the standard normal distribution, (b) the Studentized $t(4)$ distribution, that is, $t(4)/\sqrt{2}$, and (c) the centralized and Studentized Gamma(4,1) distribution, that is, $(\text{Gamma}(4, 1) - 4)/2$. The simulation results for (b) and (c) are summarized in Supplementary Material. To save space, we include only the results for BCH, ZB, and ZB2 for case (a). Throughout the simulations, we set $n = 100$, $p = 500$, and the nominal level $1 - \alpha = 0.95$. All simulation results are based on 100 independent simulation runs.

We summarize the empirical coverage probabilities, corresponding confidence interval lengths, and the absolute value of the overall normalized, bias defined as

$$\text{Bias} = \frac{|\sqrt{n} R(v_j, \beta_{-j})|}{\sqrt{\hat{\sigma}^2 n^{-1} \|v_j\|^2}}, \tag{6.1}$$

for both the active set and the inactive set in Figures 5–8. The R code of Javanmard and Montanari (2014) makes a finite-sample adjustment. To avoid an unfair comparison, we do not include their method in the bias comparison. Because inverting the test statistic in Zhu and Bradic (2018b) does not provide a closed form of confidence interval, the interval lengths of ZB and ZB2 are calculated numerically by using the bisection-type method. To avoid the computational burden therein, we calculate only the lengths of five confidence intervals of ZB and ZB2 for the inactive set in each simulation run.

We observe that (i) BRP and MBRP provide more accurate coverage, in general, for the active set in comparison to DB and JM. The coverage probability for the active set based on DB can be significantly lower than the nominal level. While BCH shows a similar or slightly higher coverage rate than BRP for the Toeplitz covariance structure, its coverage rate is lower than the nominal level in the equicorrelation case. (ii) The interval length of BCH is, in general, similar or wider than the lengths of BRP and MBRP, which are wider than that of DB for the active set. ZB and ZB2 tend to provide wider confidence intervals than the other methods do. (iii) For the equicorrelation covariance structure and $s_0 \geq 10$, ZB2 delivers the most accurate coverage rate, followed by MBRP. In contrast, the other methods significantly undercover in these cases. (iv) The better coverage of the active set for our method is closely related to the smaller bias. Interestingly, the coverage rate for the inactive set seems not to be sensitive to the bias. (v) The computation time of our method is between those of DB and ZB, as shown in Table 1. (vi) The bias associated with the active set tends to be larger than that with the inactive set, especially in the case of the Toeplitz covariance. Overall, BRP seems to reduce the bias associated with the active and inactive sets in this case. (vii) The coverage rate for the inactive set is usually close to or above the nominal level for all methods, except for ZB. According to our extensive simulations, the over-coverage is partly caused by the overestimation of the noise level, as illustrated in Figure 22 in Supplementary Material. Overall, our proposed method appears to outperform DB, JM, BCH, and ZB in terms of coverage accuracy.

Figures 9–10 plot the bias and length of BRP and MBRP against C_2 selected using the procedure in Section 5. Note that for BRP, the interval width increases, in general, while the bias decreases with C_2. The pattern is less obvious for MBRP, with most of the values of C_2 concentrated around the lower end of the grid points in (S1.1).

6.2. Confidence interval for a sparse linear combination of regression coefficients

In this subsection, we investigate the finite-sample performance of the method in Section 4. We consider the case where a linear contrast for two coefficients is of interest. We set the true regression coefficient $\beta = (b_1, b_1, b_2, b_3, 0, \dots, 0)^\top$, where b_1, b_2, b_3 are drawn independently from $U(0, 4)$. Depending on a, we consider the following two cases:

1. Contrast 1: $a = (1, -1, 0, \dots, 0)^\top$ and $a^\top \beta = b_1 - b_1 = 0$;
2. Contrast 2: $a = (0, 0, 1, -1, , 0, \dots, 0)^\top$ and $a^\top \beta = b_2 - b_3 \neq 0$.

We adopt the same procedures as before for choosing the surrogate set and the tuning parameters, but the results are based on 300 independent simulation runs. The configuration for ϵ is the same as in the previous subsection. The results for the t-distributed and gamma errors are presented in Supplementary Material.

Figure 11 shows the empirical coverage rates, corresponding confidence interval widths, and bias for each contrast. For the Toeplitz covariance structure, BRP and MBRP provide closer coverage rates to the nominal level, but with wider interval lengths than those of DB. In particular, MBRP delivers the smallest bias. Thus, the better coverage for our method is again closely related to the smaller bias in the finite sample. For the equicorrelation covariance structure, the coverage rates of all the methods are close to the nominal level. We also note that ZB2 provides satisfactory coverage probabilities, while ZB significantly undercovers in the case of the Toeplitz covariance structure. Similarly to the case for a single regression coefficient, the lengths of ZB and ZB2 are, in general, wider than those of the other methods.

6.3. Real-data analysis

As a real data-application, we consider a data set of riboflavin (vitamin B_2) production by Bacillus subtilis. The data set is available in the R package `hdi`, and has also been analyzed in van de Geer et al. (2014) and Javanmard and Montanari (2014). It contains $n = 71$ observations of $p = 4{,}088$ covariates of gene expressions and a response of riboflavin production. We model the data using (2.1), and consider the following multiple hypothesis tests for the significance of each gene:

$$H_{j,0} : \beta_j = 0 \quad \text{for} \quad j = 1, \ldots, 4{,}088.$$

We use Theorem 1 and Corollary 2 to calculate the p-values based on BRP and MBRP, respectively. The Holm procedure is adopted for the multiplicity adjustment with a 5% significance level. Neither of our methods finds any significant predictors, which is also the case for DB; JM identifies two significant genes, YXLD-at and YXLE-at.

7. Conclusion

We have proposed a new method for finding the projection direction in the debiased Lasso estimator, and demonstrated its advantage over the original debiased Lasso estimator of van de Geer et al. (2014) and the method of Javanmard and Montanari (2014). The main contributions of this study are summarized below:

- We propose a new formulation to estimate the projection direction by properly balancing the biases associated with the strong and weak signals.

- We show that the set of strong signals can be estimated consistently, and establish the asymptotic normality of the proposed estimator.
- We propose a modified estimator that can lead to a smaller order of bias than that of the original debiased Lasso, both theoretically and empirically.
- We generalize our idea to conduct an inference for a sparse linear combination of the regression coefficients.

We expect that our method can be extended to other settings, such as the generalized linear models, Cox proportional hazards model, and nonparametric additive models.

Supplementary Material

The online Supplementary Material provides the appendix for the main paper, technical details, and additional numerical results.

Acknowledgments

This research was supported in part by NSF grants DMS-1607320 and DMS-1811747.

References

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

Barber, R. F. and Candès, E. J. (2019). A knockoff filter for high-dimensional selective inference. *The Annals of Statistics* **47**, 2504–2537.

Belloni, A., Chernozhukov, V. and Hansen, C. (2014). Inference on treatment effects after selection among high-dimensional controls. *The Review of Economic Studies* **81**, 608–650.

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

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

Cai, T. T. and Guo, Z. (2017). Confidence intervals for high-dimensional linear regression: Minimax rates and adaptivity. *The Annals of Statistics* **45**, 615–646.

Candès, E., Fan, Y., Janson, L. and Lv, J. (2018). Panning for gold: 'model-X' knockoffs for high dimensional controlled variable selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **80**, 551–577.

Chang, J., Chen, S. X., Tang, C. Y. and Wu, T. T. (2020). High-dimensional empirical likelihood inference. *Biometrika* **108**, 127–147.

Chatterjee, A. and Lahiri, S. N. (2011). Bootstrapping Lasso estimators. *Journal of the American Statistical Association* **106**, 608–625.

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

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

Hastie, T., Tibshirani, R. and Wainwright, M. (2015). *Statistical Learning with Sparsity: The Lasso and Generalizations*. CRC press, Florida.

Javanmard, A. and Montanari, A. (2014). Confidence intervals and hypothesis testing for high-dimensional regression. *Journal of Machine Learning Research* **15**, 2869–2909

Javanmard, A. and Montanari, A. (2018). Debiasing the Lasso: Optimal sample size for Gaussian designs. *The Annals of Statistics* **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. *The Annals of Statistics* **44**, 907–927.

Liu, H. and Yu, B. (2013). Asymptotic properties of Lasso+mLS and Lasso+Ridge in sparse high-dimensional linear regression. *Electronic Journal of Statistics* **7**, 3124–3169.

Lockhart, R., Taylor, J., Tibshirani, R. and Tibshirani, R. (2014). A significance test for the Lasso. *The Annals of Statistics* **42**, 413–468.

Meinshausen, N. and Bühlmann, P. (2006). High-dimensional graphs and variable selection with the Lasso. *The Annals of Statistics* **34**, 1436–1462.

Meinshausen, N., Meier, L. and Bühlmann, P. (2009). P-values for high-dimensional regression. *Journal of the American Statistical Association* **104**, 1671–1681.

Meinshausen, N. and Bühlmann, P. (2010). Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **72**, 417–473.

Neykov, M., Ning, Y., Liu, J. S. and Liu, H. (2018). A unified theory of confidence regions and testing for high-dimensional estimating equations. *Statistical Science* **33**, 427–443.

Ning, Y. and Liu, H. (2017). A general theory of hypothesis tests and confidence regions for sparse high dimensional models. *The Annals of Statistics* **45**, 158–195.

Reid, S., Tibshirani, R. and Friedman, J. (2016). A study of error variance estimation in Lasso regression. *Statistica Sinica* **26**, 35–67.

Tibshirani, R. J., Taylor, J., Lockhart, R. and Tibshirani, R. (2016). Exact post-selection inference for sequential regression procedures. *Journal of the American Statistical Association* **111**, 600–620.

van de Geer, S., Bühlmann, P., Ritov, Y. and Dezeure, R. (2014). On asymptotically optimal confidence regions and tests for high dimensional models. *The Annals of Statistics* **42**, 1166–1202.

Wasserman, L. and Roeder, K. (2009). High dimensional variable selection. *The Annals of Statistics* **37**, 2178–2201.

Zhang, C. H. and Zhang, S. S. (2014). Confidence intervals for low-dimensional parameters with high-dimensional data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **76**, 217–242.

Zhang, X. and Cheng, G. (2017). Simultaneous inference for high-dimensional linear models. *Journal of the American Statistical Association*, **112**, 757–768.

Zhu, Y. and Bradic, J. (2018a). Significance testing in non-sparse high-dimensional linear models. *Electronic Journal of Statistics* **12**, 3312–3364.

Zhu, Y. and Bradic, J. (2018b). Linear hypothesis testing in dense high-dimensional linear models. *Journal of the American Statistical Association* **113**, 1583–1600.

Sangyoon Yi

Department of Statistics, Texas A&M University, College Station, TX 77843, USA.

E-mail: syi@stat.tamu.edu

Xianyang Zhang

Department of Statistics, Texas A&M University, College Station, TX 77843, USA.

E-mail: zhangxiany@stat.tamu.edu

(Received August 2019; accepted September 2020)