Journal of the American Statistical Association

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/uasa20

Honest Confidence Sets for High-Dimensional Regression by Projection and Shrinkage

Kun Zhou, Ker-Chau Li & Qing Zhou

To cite this article: Kun Zhou, Ker-Chau Li & Qing Zhou (2021): Honest Confidence Sets for High-Dimensional Regression by Projection and Shrinkage, Journal of the American Statistical Association, DOI: 10.1080/01621459.2021.1938581

To link to this article: https://doi.org/10.1080/01621459.2021.1938581

JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2021, VOL. 00, NO. 0, 1–20: Theory and Methods
https://doi.org/10.1080/01621459.2021.1938581

Honest Confidence Sets for High-Dimensional Regression by Projection and Shrinkage

Kun Zhou[a], Ker-Chau Li[a,b], and Qing Zhou[a]

[a]Department of Statistics, University of California, Los Angeles, CA; [b]Institute of Statistical Science, Academia Sinica, Nangang, Taiwan

ABSTRACT

The issue of honesty in constructing confidence sets arises in nonparametric regression. While optimal rate in nonparametric estimation can be achieved and utilized to construct sharp confidence sets, severe degradation of confidence level often happens after estimating the degree of smoothness. Similarly, for high-dimensional regression, oracle inequalities for sparse estimators could be utilized to construct sharp confidence sets. Yet, the degree of sparsity itself is unknown and needs to be estimated, which causes the honesty problem. To resolve this issue, we develop a novel method to construct honest confidence sets for sparse high-dimensional linear regression. The key idea in our construction is to separate signals into a strong and a weak group, and then construct confidence sets for each group separately. This is achieved by a projection and shrinkage approach, the latter implemented via Stein estimation and the associated Stein unbiased risk estimate. Our confidence set is honest over the full parameter space without any sparsity constraints, while its size adapts to the optimal rate of $n^{-1/4}$ when the true parameter is indeed sparse. Moreover, under some form of a separation assumption between the strong and weak signals, the diameter of our confidence set can achieve a faster rate than existing methods. Through extensive numerical comparisons on both simulated and real data, we demonstrate that our method outperforms other competitors with big margins for finite samples, including oracle methods built upon the true sparsity of the underlying model.

ARTICLE HISTORY
Received May 2019
Accepted May 2021

KEYWORDS
Adaptive confidence set;
High-dimensional inference;
Sparse linear regression;
Stein estimate

1. Introduction

Consider high-dimensional linear regression

$$y = X\beta + \varepsilon, \tag{1}$$

where $y \in \mathbb{R}^n$, $X = [X_1|\cdots|X_p] \in \mathbb{R}^{n\times p}$, $\beta \in \mathbb{R}^p$, $\varepsilon \sim \mathcal{N}_n(0, \sigma^2\mathbf{I}_n)$ and $p > n$. While there is a rich body of research on parameter estimation under this model concerning signal sparsity (e.g., Bickel, Ritov, and Tsybakov 2009; Zhang and Huang 2008; Negahban et al. 2012), how to construct confidence sets remains elusive. In this work, we focus on confidence sets for the mean $\mu = X\beta$ with the following two properties: First, the confidence set $\widehat{C}$ is (asymptotically) honest over all possible parameters. That is, for a given confidence level $1-\alpha$,

$$\liminf_{n\to\infty} \inf_{\beta\in\mathbb{R}^p} \mathbb{P}_\beta \left\{X\beta \in \widehat{C}\right\} \geq 1-\alpha, \tag{2}$$

where $\mathbb{P}_\beta$ is taken with respect to the distribution of $y \sim \mathcal{N}_n(X\beta, \sigma^2\mathbf{I}_n)$, regarding X as fixed. Second, the diameter of $\widehat{C}$ is able to adapt to, respectively, the sparsity and the strength of β. In practical applications, sparsity assumptions are very hard to verify, and for many datasets they are at most a good approximation. The first property guarantees that our confidence sets reach the nominal coverage probability without imposing any sparsity assumption, while the second property allows us to leverage sparse estimation when β is indeed sparse. Building confidence sets for the mean of a multivariate Gaussian distribution is one of the most classical problems in statistical inference. Under a regression setting, it arises naturally when one is interested in making simultaneous inference for the mean responses among a group of individuals. As another example, the problem of recovering signals from noisy observations may be formulated as inference of a mean vector as well (Beran and Dümbgen 1998).

Our problem is related to the construction of confidence sets in nonparametric regression, for which a line of work has laid down important theoretic foundations and provided methods of construction (Li 1989; Beran and Dümbgen 1998; Baraud 2004; Cai and Low 2006; Robins and van der Vaart 2006). Despite such notable advances, lack of numerical support casts doubt on the merit of borrowing these nonparametric regression methods directly for sparse regression. Taking the adaptive method in Robins and van der Vaart (2006) as an example, an honest confidence set for μ can be constructed as $\widehat{C}_a = \{\mu \in \mathbb{R}^n : n^{-1/2}\|\mu - X\hat{\beta}\| \leq r_n\}$, where $X\hat{\beta}$ is an initial estimate independent of y, and its (normalized) diameter $|\widehat{C}_a| := 2r_n = O_p(n^{-1/4} + n^{-1/2}\|X\hat{\beta} - X\beta\|)$. A common choice for $\hat{\beta}$ under model (1) for $p > n$ is a sparse estimator, such as the lasso (Tibshirani 1996) or ℓ_0-penalized least-square estimator. With high probability, the prediction loss of the lasso estimator typically satisfies

$$\frac{1}{n}\|X\hat{\beta} - X\beta\|^2 \leq c\frac{s\log p}{n} \tag{3}$$

CONTACT Qing Zhou zhou@stat.ucla.edu Department of Statistics, University of California, Los Angeles, CA 90095.
Supplementary materials for this article are available online. Please go to www.tandfonline.com/r/JASA.

for some $c > 0$, uniformly for all $\beta \in \mathscr{B}(s) := \{v \in \mathbb{R}^p : \|v\|_0 \leq s\}$; see, for example, Bickel, Ritov, and Tsybakov (2009). Under this choice, the diameter $|\widehat{C}_a|$ is of the order

$$|\widehat{C}_a| = O_p\left(n^{-1/4} + \sqrt{s \log p/n}\right) \tag{4}$$

for all $\beta \in \mathscr{B}(s)$. For a precise statement, see Theorem 7. This method has nice theoretical properties when $s = o(n/\log p)$. But even for moderately sparse signals with $s/n \to \delta \in (0, 1)$, the bound on the right-hand side of Equation (4) approaches ∞ as $p > n \to \infty$ and thus offers little insight into the performance of the confidence set. The upper bound (3) also critically depends on the regularization parameter used for the initial estimate $\hat{\beta}$. In fact, our numerical results show that, for finite samples with $(s, n, p) = (10, 200, 800)$, this confidence set can be worse than a naive χ^2 region $\{\mu : \|y - \mu\|^2 \leq \sigma^2 \chi^2_{n,\alpha}\}$, where $\chi^2_{n,\alpha}$ denotes the $1 - \alpha$ quantile of the χ^2 distribution with n degrees of freedom. A similar issue occurs in the related but different problem of constructing confidence sets for β. Nickl and van de Geer (2013) have shown that one can construct a confidence set for β that is honest over $\mathscr{B}(k_1)$ for $k_1 = o(n/\log p)$, and for $s \leq k_1$, the diameter is on the same order as that in Equation (4) for any $\beta \in \mathscr{B}(s)$. Compared to the unrestricted honesty in Equation (2) over the entire space $\mathbb{R}^p$, the restriction on the honesty region to $\mathscr{B}(k_1)$ also reflects the challenge faced in the construction of confidence sets when $p > n$. Carpentier (2015) further extended the result of Nickl and van de Geer (2013) to construct confidence sets that are adaptive to multiple sparsity levels, by imposing margins between subspaces with different sparsity. Ewald and Schneider (2018) provided an exact formula to compute a lower bound of the coverage rate of a confidence set centered at the lasso, over the entire parameter space; however, low dimension ($p < n$) is a vital condition in their proof, making it difficult to generalize their idea to the high-dimensional problem that we are studying.

The construction of confidence sets is fundamentally different from the problem of inferring error bounds for a sparse estimator (Nickl and van de Geer 2013). It is seen from Equation (4) that no matter how sparse the true β is, the diameter of $\widehat{C}_a$ cannot converge at a rate faster than $n^{-1/4}$. Indeed, results in Li (1989) imply that, for the linear model (1) with $p \geq n$, the diameter of an honest confidence set for μ, in the sense of Equation (2), cannot adapt at any rate $o(n^{-1/4})$. The rate of $n^{-1/4}$ has also been observed in hypothesis testing and accuracy assessment for high-dimensional regression (Arias-Castro, Candès, and Plan 2011; Cai and Guo 2018; Ingster, Tsybakov, and Verzelen 2010; Verzelen 2012). This is in sharp contrast to error bounds for a sparse point estimator, such as that in Equation (3), which can decay at a much faster rate when β is sufficiently sparse. It is not desired to construct confidence sets directly from error bounds like Equation (3) even we only require honesty for $\beta \in \mathscr{B}(k_1)$ with a given $k_1 = o(n/\log p)$, because its diameter, on the order of $\sqrt{k_1 \log p/n}$, cannot adapt to any sparser $\beta \in \mathscr{B}(s)$ for $s < k_1$.

Motivated by these challenges, we propose a new two-step method to construct a confidence set for $\mu = X\beta$, allowing the dimension $p \gg n$ in Equation (1). The basic idea of our method is to estimate the radius of the confidence set separately for strong and weak signals defined by the magnitude of $|\beta_j|$. Using a sparse estimate, such as the lasso, one can recover the set A of large $|\beta_j|$ accurately and expect a small radius for a confidence ball for μ_A, the projection of μ onto the subspace spanned by $X_j, j \in A$. By construction, $(\mu - \mu_A)$ is composed of weak signals. Thus, in the second step, we shrink our estimate of this part toward zero by Stein's method and construct a confidence set with Stein's unbiased risk estimate (Stein 1981). Combining the inferential advantages of sparse estimators and Stein estimators, our method overcomes many of the aforementioned difficulties. First, our confidence set is honest for all $\beta \in \mathbb{R}^p$, and its diameter is well under control for all possible values of β including the dense case. Second, by using elastic radii our confidence set, an ellipsoid in general, can adapt to both sparsity and signal strength. The radius for strong signals adapts to the sparsity of the underlying model via sparse estimation or model selection, while the radius for weak signals adapts according to the degree of shrinkage of the Stein estimate. Without any signal strength assumption, the diameter of our confidence set is $O_p(n^{-1/4} + \sqrt{s \log p/n})$, the same as Equation (4), for $\beta \in \mathscr{B}(s)$. It may further reduce to $O_p(n^{-1/4} + \sqrt{s/n})$ under an assumption on the separability between the strong and the weak signals. Third, we provide a data-driven selection of the set A from multiple candidates, which protects our method from a bad choice and thus makes it very robust. To maximize the practical significance of our method, we have developed efficient algorithms to approximate all constants involved in our theory. We demonstrate with extensive numerical results on both simulated and real-world datasets that our method can construct much smaller confidence sets than the adaptive method (Robins and van der Vaart 2006) discussed above and oracle approaches making use of the *true* sparsity of β (the oracle).

Note that the construction of confidence sets for $\mu = X\beta$ is different in nature from the construction of confidence intervals for an individual β_j or a low-dimensional projection of β. For the latter, the optimal rate of an interval length can be $n^{-1/2}$ when β is sufficiently sparse (Schneider 2016; Cai and Guo 2017), such as the intervals constructed by de-biased lasso methods (Javanmard and Montanari 2014; van de Geer et al. 2014; Zhang and Zhang 2014). Li (2020) further developed a bias-corrected de-biased lasso with bootstrap, of which the sample complexity for asymptotic normality is improved from $n \gg (s \log p)^2$ as in the above work to $n \gg \max\{s \log p, (\tilde{s} \log p)^2\}$, where $\tilde{s}$ is the number of weak signals of $O(\sqrt{\log p/n})$. The idea of separating strong and weak signals is in a similar spirit to our method. Although simultaneous inference methods have been proposed based on bootstrapping de-biased lasso estimates (Zhang and Cheng 2017; Dezeure, Bühlmann, and Zhang 2017), these methods are shown to achieve the desired coverage only for extremely sparse β such that $\|\beta\|_0 = o(\sqrt{n/(\log p)^3})$, which severely limits their practical application.

The remainder of this article is organized as follows: Section 2 develops our two-step Stein method in details, including its theoretical properties and algorithmic implementation. To demonstrate the advantage of our method, we develop in Section 3 a few competing methods making use of the lasso prediction or the oracle of the true sparsity. Extensive numerical comparisons are provided in Sections 4 and 5 to show the

superior performance of our two-step Stein method, relative to the competitors, in a variety of simulation settings, including when β is quite dense, and in a real data analysis. The paper is concluded in Section 6 with discussions on some limitations of and potential improvements for our method. Proofs of all theoretical results are deferred to the supplementary materials.

Throughout the article, we always assume model (1) with $\varepsilon \sim \mathcal{N}_n(0, \sigma^2 \mathbf{I}_n)$ unless otherwise noted. We denote by $\mathbb{P}_\beta$ the distribution of $[y \mid X]$ and $\mathbb{E}_\beta$ the corresponding expectations, where the subscript β may be dropped when its meaning is clear from the context. Denote by $[p]$ the index set $\{1, \ldots, p\}$ and by $|A|$ the size of a set A. Write $a_n = \Omega(b_n)$ if $b_n = O(a_n)$ and $a_n \asymp b_n$ if $a_n = O(b_n)$ and $b_n = O(a_n)$. We use $\Omega_p(.)$ and $\asymp_p$ if the above statements hold in probability. For a vector $v = (v_j)_{1:m}$, let $v_A = (v_j)_{j \in A}$ be the restriction of v to the components in A. For a matrix $M = [M_1 \mid \ldots \mid M_m]$, where M_j is the jth column, denote by $M_A = (M_j)_{j \in A}$ the submatrix consisting of columns in A. We use $\|v\|_q$ to denote the ℓ_q norm, $q \in [0, \infty]$, of a vector v, and $\|v\|$ for the Euclidean norm (ℓ_2 norm). For $a, b \in \mathbb{R}^n$, $\langle a, b \rangle := a^\mathsf{T} b$ is the inner product. Define $a \vee b := \max\{a, b\}$ and $a \wedge b := \min\{a, b\}$ for $a, b \in \mathbb{R}$.

2. Two-Step Stein method

Dividing β into strong and weak signals, our method constructs a confidence set $\widehat{C}(y)$ with an ellipsoid shape for $X\beta$ that is honest as defined in (2). Note that under a high-dimensional asymptotic framework, all variables $X = X(n)$, $y = y(n)$, $\beta = \beta(n)$ and $s = s_n$ depend on n as $p = p_n \gg n \to \infty$, while $X(n)$ is regarded as a fixed design matrix for each n. We often suppress the dependence on n to simplify the notation.

2.1. Preliminaries on Stein Estimation

We will use a simplified Stein estimate (Li 1989) to construct the confidence set for weak signals. For a linear estimate $\tilde{\mu} = T_n y$, where $y \sim \mathcal{N}_n(\mu, \sigma^2 \mathbf{I}_n)$ and $T_n \in \mathbb{R}^{n \times n}$, let $R_n = \mathbf{I}_n - T_n$, and define

$$\hat{\mu}(y; \tilde{\mu}) = y - \frac{\sigma^2 \mathrm{tr}(R_n)}{\|R_n y\|^2} R_n y, \tag{5}$$

$$\hat{L}(y; \tilde{\mu}) = 1 - \frac{\sigma^2 \left(\mathrm{tr}(R_n)\right)^2}{n \|R_n y\|^2}, \tag{6}$$

where $\hat{\mu}(y; \tilde{\mu})$ is the Stein estimate associated with the initial estimate $\tilde{\mu}$ and $\sigma^2 \hat{L}(y; \tilde{\mu})$ is the Stein unbiased risk estimate (SURE). Li (1989) proved the uniform consistency of $\hat{L}$.

Lemma 1 (Theorem 3.1 in Li 1989). Assume that $y \sim \mathcal{N}_n(\mu, \sigma^2 \mathbf{I}_n)$. For any $\alpha \in (0, 1)$, there exists a constant $c_{\mathrm{st}}(\alpha) > 0$ such that

$$\liminf_{n \to \infty} \inf_{\mu \in \mathbb{R}^n} \mathbb{P}_\mu \left\{ \left| \sigma^2 \hat{L} - n^{-1} \|\hat{\mu} - \mu\|^2 \right| \le c_{\mathrm{st}}(\alpha) \sigma^2 n^{-1/2} \right\} \ge 1 - \alpha, \tag{7}$$

where $\hat{\mu}$ and $\hat{L}$ are defined in Equations (5) and (6).

2.2. Method of Construction

Now, consider the linear model (1) and let $\mu = X\beta$. Given a preconstructed candidate set $A = A_n \subseteq [p]$, independent of (X, y), define

$$\mu_A = P_A \mu, \qquad \mu_\perp = P_A^\perp \mu = (\mathbf{I}_n - P_A)\mu,$$

where P_A is the orthogonal projection from $\mathbb{R}^n$ onto $\mathrm{span}(X_A)$ and $P_A^\perp$ is the projection to the orthogonal complement. A good candidate set A is supposed to include all strong signals, say $A = \{j : |\beta_j| > \tau\}$. With such a choice, $\|\mu_\perp\|$ will be small. Typically, we split our dataset into two halves, (X, y) and (X', y'), and apply a model selection method on (X', y') to construct the set A. See Section 2.3 for more detailed discussion.

We estimate μ_A and $\mu_\perp$, respectively, by $\hat{\mu}_A$ and $\hat{\mu}_\perp$, compute radii r_A and $r_\perp$, and construct a $(1 - \alpha)$ confidence set $\widehat{C}$ for μ in the form of

$$\widehat{C} = \left\{ \mu \in \mathbb{R}^n : \frac{\|P_A \mu - \hat{\mu}_A\|^2}{n r_A^2} + \frac{\|P_A^\perp \mu - \hat{\mu}_\perp\|^2}{n r_\perp^2} \le 1 \right\}. \tag{8}$$

Note that $\widehat{C}$ is an ellipsoid in $\mathbb{R}^n$, where $r_A = r_A(\alpha)$ and $r_\perp = r_\perp(\alpha)$ correspond to two radii. Our method consists of a projection and a shrinkage step:

Step 1: Projection. Let $\hat{\mu}_A = P_A y$ and $k = \mathrm{rank}(X_A) \le |A|$. Since A is independent of (y, X), we have

$$\|\hat{\mu}_A - \mu_A\|^2 = \|P_A \varepsilon\|^2 \mid A \sim \sigma^2 \chi_k^2. \tag{9}$$

Thus, we choose

$$r_A^2 = c_1 \tilde{r}_A^2 = c_1 \sigma^2 \chi_{k, \alpha/2}^2 / n, \tag{10}$$

where $\chi_{k, \alpha/2}^2$ is the $(1 - \alpha/2)$ quantile of the χ_k^2 distribution and $c_1 > 1$ is a constant, so that

$$\mathbb{P}\left\{ \frac{\|P_A \mu - \hat{\mu}_A\|^2}{n r_A^2} \le 1/c_1 \right\} = 1 - \alpha/2. \tag{11}$$

Step 2: Shrinkage. Let $y_\perp = P_A^\perp y$. As mentioned above, under a good choice of A that contains strong signals, $\|\mu_\perp\|$ is expected to be small. Therefore, we shrink $y_\perp$ toward zero via Stein estimation to construct $\hat{\mu}_\perp$. Note that $y_\perp$ is in an $(n - k)$-dimensional subspace of $\mathbb{R}^n$. Letting $\tilde{\mu} = 0$ and $R_n = P_A^\perp$ in Equations (5) and (6), we obtain

$$\hat{\mu}_\perp = \hat{\mu}(y_\perp; 0) = (1 - B) y_\perp, \tag{12}$$

$$\hat{L} = \hat{L}(y_\perp; 0) = 1 - B, \tag{13}$$

where the shrinkage factor

$$B = (n - k)\sigma^2 / \|y_\perp\|^2. \tag{14}$$

It then follows from Lemma 1 that

$$\liminf_{(n-k) \to \infty} \inf_{\beta \in \mathbb{R}^p} \mathbb{P}\left\{ \left| \sigma^2 \hat{L} - (n - k)^{-1} \|\hat{\mu}_\perp - \mu_\perp\|^2 \right| \le c_{\mathrm{st}}(\alpha) \sigma^2 (n - k)^{-1/2} \right\} \ge 1 - \alpha, \tag{15}$$

for any sequence of $A = A_n$ as long as $(n - k) \to \infty$. Therefore, if we choose

$$r_\perp^2 = c_2 \tilde{r}_\perp^2 = c_2 \frac{n - k}{n} \sigma^2 \left\{ \hat{L} + c_{\mathrm{st}}(\alpha/2)(n - k)^{-1/2} \right\}, \tag{16}$$

where $c_2 > 1$ is a constant, we have

$$\liminf_{(n-k)\to\infty} \inf_{\beta\in\mathbb{R}^p} \mathbb{P}\left\{\frac{\|\mu_\perp - \hat{\mu}_\perp\|^2}{nr_\perp^2} \le 1/c_2\right\} \ge 1-\alpha/2. \quad (17)$$

In practical implementation, we estimate the constant $c_{\mathrm{st}}(\alpha)$ in Equation (15) by simulation, which will be discussed in Section 2.5.

If $1/c_1 + 1/c_2 = 1$, confidence set (8) made up from Equations (11) and (17) is honest and the expectation of its (normalized) diameter $|\widehat{C}| := 2(r_A \vee r_\perp)$ can be calculated explicitly for all $\beta \in \mathbb{R}^p$:

Theorem 1. Assume $1/c_1 + 1/c_2 = 1$, A is independent of (y, X) with $\mathrm{rank}(X_A) = k$, and $(n-k) \to \infty$ as $n \to \infty$. Then the confidence set $\widehat{C}$ (8) constructed by the two-step Stein method is honest in the sense of Equation (2). Furthermore, the squared diameter of $\widehat{C}$ has expectation

$$\mathbb{E}|\widehat{C}|^2 = 4\sigma^2 \max\left\{c_1\frac{\chi^2_{k,\alpha/2}}{n}, c_2\frac{n-k}{n} \times \left(1 - \mathbb{E}\frac{n-k}{\chi^2_{n-k}(\rho)} + c_{\mathrm{st}}(\alpha/2)(n-k)^{-1/2}\right)\right\}, \quad (18)$$

where $\chi^2_{n-k}(\rho)$ follows a noncentral χ^2 distribution with $n-k$ degrees of freedom and noncentrality parameter $\rho = \|\mu_\perp\|^2/\sigma^2$.

In the above result, we did not impose any assumptions on A except $(n-k) \to \infty$, which allows many choices of A. Our confidence set $\widehat{C}$ is honest as in Equation (2) and its diameter is under control for all $\beta \in \mathbb{R}^p$. Since $\mathbb{E}[1/\chi^2_{n-k}(\rho)] > 0$, a uniform but very loose upper bound

$$\mathbb{E}|\widehat{C}|^2 \le 4\sigma^2 \max\left\{c_1\frac{\chi^2_{k,\alpha/2}}{n}, c_2\frac{n-k}{n} \times \left(1 + c_{\mathrm{st}}(\alpha/2)(n-k)^{-1/2}\right)\right\} \quad (19)$$

holds for all $\beta \in \mathbb{R}^p$. In particular, when β is dense, the diameter will be comparable to that of a naive χ^2 region. As corroborated with the numerical results in Section 4.4, this protects our method from inferior performance when sparsity assumptions are violated, making it robust to different datasets. Note that our choice of the confidence level of $1-\alpha/2$ for both the strong and the weak signal parts in Equations (11) and (17) is out of convenience. There might be an optimal choice of the two confidence levels that minimizes the diameter $|\widehat{C}|$. But we do not expect a faster convergence rate.

Next, we will show that our confidence set is adaptive: When β is indeed sparse, the radii r_A and $r_\perp$ will adapt to the optimal rate with a proper choice of A that contains strong signals.

2.3. Adaptation of the Diameter

To simplify our analysis, we first set $c_1 = c_2 = 2$ so that they can be ignored when calculating the convergence rates of r_A and $r_\perp$. These rates do not change as long as c_1 and c_2 stay as constants when $n \to \infty$. We will discuss choices of c_1 and c_2 near the end of this subsection.

Lemma 2 gives the rates of r_A and $r_\perp$, and specifies conditions for the diameter of $\widehat{C}$ to converge at the optimal rate $n^{-1/4}$.

Lemma 2. Suppose that A is independent of (y, X), $k = \mathrm{rank}(X_A)$, and $\|\mu_\perp\| = o(\sqrt{n-k})$. Then

$$r_A^2 \asymp_p k/n, \quad r_\perp^2 = O_p\left(\frac{\sqrt{n-k}}{n} + \frac{\|\mu_\perp\|^2}{n}\right).$$

Therefore, if $k = O(\sqrt{n})$ and $\|\mu_\perp\| = O(n^{1/4})$, then the diameter of $\widehat{C}$

$$|\widehat{C}| = 2(r_A \vee r_\perp) \asymp_p n^{-1/4}.$$

The ℓ_2 norm of the weak signals $\|\mu_\perp\|$ can be bounded by $\|\beta_{A^c}\|$ under the sparse Riesz condition on X and a sparsity assumption on β. A design matrix X satisfies the sparse Riesz condition (Zhang and Huang 2008) with rank s^* and spectrum bounds $0 < c_* < c^* < \infty$, denoted by $\mathrm{SRC}(s^*, c_*, c^*)$, if

$$c_* \le \frac{\|X_A v\|^2}{n\|v\|^2} \le c^*, \quad \text{for all } A \text{ with } |A| = s^* \text{ and all nonzero } v \in \mathbb{R}^{s^*}.$$

Under our asymptotic framework, s^*, c^* and c_* are allowed to depend on n.

Theorem 2. Suppose A is independent of (y, X), $k = \mathrm{rank}(X_A)$, and X satisfies $\mathrm{SRC}(s^*, c_*, c^*)$ with $s^* \ge |\mathrm{supp}(\beta) \cap A^c|$. If $\limsup_n c^* < \infty$, $k = o(n)$ and $\|\beta_{A^c}\| = o(1)$, then

$$|\widehat{C}| = O_p\left\{(n^{-1/4} + \|\beta_{A^c}\|) \vee \sqrt{k/n}\right\} \quad (20)$$

for the two-step Stein method. In particular, $|\widehat{C}| \asymp_p n^{-1/4}$ if $k = O(\sqrt{n})$ and $\|\beta_{A^c}\| = O(n^{-1/4})$.

Remark 1. Let us take a closer look at the conditions in this theorem for $|\widehat{C}| \asymp_p n^{-1/4}$. Suppose that β has $O(\sqrt{n})$ strong coefficients that can be reliably detected by a sparse estimator, while all other signals are weak such that $\|\beta_{A^c}\| = O(n^{-1/4})$. Then we may have $k \le |A| = O(\sqrt{n})$ with high probability. This shows that the sparsity $s = \|\beta\|_0$ is allowed to be $O(\sqrt{n})$. The only additional constraint on s comes from the assumption $\mathrm{SRC}(s^*, c_*, c^*)$ with $s^* \ge s$, which holds for Gaussian designs if $s \log p = o(n)$ (Zhang and Huang 2008). Compared to Equation (4) which requires $s \log p = O(\sqrt{n})$, the sparsity assumption on β to attain the optimal rate $n^{-1/4}$ for our method will be relaxed if the weak signals $\|\beta_{A^c}\| \ll \sqrt{s \log p/n}$.

Now we discuss a few methods to find A so that our confidence sets can adapt to the sparsity and the signal strength of β, respectively. We split the whole dataset into (X, y) and (X', y'), with respective sample sizes n and n', so that they are independent. Henceforth, we assume an even partition with $n' = n$, which simplifies the notation and is commonly used in practice, unless otherwise noted. The first method is to apply lasso on (X', y'):

$$\hat{\beta} = \hat{\beta}(y', X'; \lambda) := \mathrm{argmin}_{\beta\in\mathbb{R}^p}\left[\frac{1}{2n}\|y' - X'\beta\|^2 + \lambda\|\beta\|_1\right], \quad (21)$$

where λ is a tuning parameter. Then, we define the set of strong signals by the support of $\hat{\beta}$:

$$A = \{j : \hat{\beta}_j \ne 0\}. \quad (22)$$

Under this choice, $\hat{\beta}_{A^c} = 0$ by definition and thus $\|\beta_{A^c}\|$ in Equation (20) can be bounded by

$$\|\beta_{A^c}\| \le \|\hat{\beta} - \beta\| = O_p(\sqrt{s \log p/n}),$$

where the ℓ_2 error bound of the lasso is valid without any beta-min condition (Zhang and Huang 2008; Bickel, Ritov, and Tsybakov 2009). This leads to the same rate (4) for $|\widehat{C}|$, which adapts to the sparsity s without any assumption on signal strength. This is the first conclusion of the following corollary:

Corollary 3. Suppose that X and X' satisfy SRC(s^*, c_*, c^*), where $0 < c_* < c^*$ are constants. Let the confidence set $\widehat{C}$ (8) be constructed by the two-step Stein method with A chosen by Equation (22) and $\lambda = c_0\sigma\sqrt{c^* \log p/n}, c_0 > 2\sqrt{2}$. Assume $s \le (s^* - 1)/(2 + 4c^*/c_*)$ and $s \log p = o(n)$. Then for any $\beta \in \mathscr{B}(s)$ we have

$$|\widehat{C}| = O_p\left(n^{-1/4} + \sqrt{s \log p/n}\right). \tag{23}$$

Let $A_0 = \mathrm{supp}(\beta)$ and $S_0 = \{j \in A_0 : |\beta_j| \ge K\sqrt{s \log p/n}\}$ for a sufficiently large K. If in addition $\|\beta_{A_0 \setminus S_0}\| = O(n^{-1/4})$, then

$$|\widehat{C}| = O_p\left(n^{-1/4} \vee \sqrt{s/n}\right). \tag{24}$$

The second conclusion of the above corollary shows that our method can achieve a faster rate (24) if $\|\beta_{A_0 \setminus S_0}\| = O(n^{-1/4})$. Together with the definition of S_0, this essentially imposes a separability assumption between the strong and the weak signals when $s \log p \gg \sqrt{n}$. To weaken the beta-min condition on strong signals in S_0, we may apply a better model selection method to define A, such as using the minimax concave penalty (MCP) (Zhang 2010):

$$\begin{aligned}\rho(t; \lambda, \gamma) &= \int_0^{|t|} \left(1 - \frac{u}{\gamma\lambda}\right)_+ du \\ &= \begin{cases} |t| - t^2/(2\gamma\lambda) & \text{if } |t| \le \gamma\lambda \\ \gamma\lambda/2 & \text{if } |t| > \gamma\lambda \end{cases}, \end{aligned} \tag{25}$$

for $\gamma > 1$. Accordingly, a regularized least-square estimate is defined by

$$\begin{aligned}\hat{\beta}^{\mathrm{mcp}}_{\lambda,\gamma} &= \hat{\beta}^{\mathrm{mcp}}_{\lambda,\gamma}(y', X') \\ &:= \mathrm{argmin}_{\beta \in \mathbb{R}^p}\left[\frac{1}{2n}\|y' - X'\beta\|^2 + \lambda \sum_{j=1}^{p} \rho(|\beta_j|; \lambda, \gamma)\right]. \end{aligned} \tag{26}$$

Suppose we choose $A = \mathrm{supp}(\hat{\beta}^{\mathrm{mcp}}_{\lambda,\gamma})$ in our two-step Stein method. The model selection consistency of $\hat{\beta}^{\mathrm{mcp}}_{\lambda,\gamma}$ makes it possible for $|\widehat{C}|$ to adapt at the rate (24) under the same SRC assumption but a weaker beta-min condition than that in the definition of S_0.

Corollary 4. Suppose that X and X' satisfy SRC(s^*, c_*, c^*), where $0 < c_* < c^*$ are constants, $s^* \ge (c^*/c_* + 1/2)s$, and $s \log p = o(n)$. Choose a sequence of (λ_n, γ_n) satisfying $\lambda_n \gg \sqrt{\log p/n}$ and $\gamma_n \ge c_*^{-1}\sqrt{4 + c_*/c^*}$. If $\beta \in \mathscr{B}(s)$ and $\inf_{A_0} |\beta_j| \ge (\gamma_n + 1)\lambda_n$, then $\mathbb{P}\{\mathrm{supp}(\hat{\beta}^{\mathrm{mcp}}_{\lambda_n,\gamma_n}) = A_0\} \to 1$, and consequently the $\widehat{C}$ constructed by the two-step Stein method with $A = \mathrm{supp}(\hat{\beta}^{\mathrm{mcp}}_{\lambda_n,\gamma_n})$ has diameter

$$|\widehat{C}| = O_p\left(n^{-1/4} \vee \sqrt{s/n}\right). \tag{27}$$

Remark 2. Compared to (4) for confidence sets centering at a sparse estimator, the diameter of our method in Equations (24) and (27) converges faster by a factor of $(\log p)^{1/2}$ when $s = \Omega(\sqrt{n})$. Accordingly, our method achieves the optimal rate when $s = O(\sqrt{n})$ instead of $s = O(\sqrt{n}/\log p)$ as for Equation (4). Under a high-dimensional setting with $p \gg n$, say $p = \exp(n^a)$ for $a \in (0, 1/2)$, this improvement in rate can be very substantial, which is supported by our numerical results. The faster rate of our method is made possible by its adaption to *both* signal strength and sparsity, while the rate of Equation (4) is obtained by adaption to sparsity only (cf. Theorem 7). We emphasize that our method achieves the adaptive rates in the above results, while being uniformly honest over the entire $\mathbb{R}^p$ (Theorem 1). One could construct a confidence set with diameter $O_p(\sqrt{s/n})$ using only the covariates selected by a consistent model selection method, which would be faster than the rate (27). However, such a confidence set is *not* honest over $\mathbb{R}^p$, because it cannot reach the nominal coverage rate for those β that do not satisfy the required beta-min condition for model selection consistency. Our method overcomes this difficulty with the shrinkage step, based on the uniform consistency of the SURE (Lemma 1).

Remark 3. For an uneven partition of the whole dataset, the conclusions of Corollaries 3 and 4 still hold as long as both $n' \asymp n \to \infty$. However, it is a common and reasonable choice to have $n = n'$, since (X', y') and (X, y) can be swapped to construct a confidence set for $X'\beta$, making full use of the whole dataset.

Carpentier (2015) developed algorithms to construct a confidence set for β that is honest and adaptive for multiple sparsity levels $\{1, \ldots, \bar{s}\}$ over a restricted parameter space. Approximate sparse vectors are considered by Carpentier, but here we focus on exact sparse vectors $\mathscr{B}(k)$ for easy comparison. For two sparsity levels $k_1 < k_2$ and some $\delta \in (0, 1)$, define a set

$$\widetilde{\mathscr{B}}(k_2, k_1) := \left\{u \in \mathscr{B}(k_2) : \inf_{v \in \mathscr{B}(k_1)} \|u - v\| \ge \delta\right\} \tag{28}$$

by enforcing a margin from the set $\mathscr{B}(k_1)$ of k_1-sparse vectors. The restricted parameter space is defined as

$$\mathcal{P} := \mathscr{B}(1) \bigcup \left[\bigcup_{k=2}^{\bar{s}} \widetilde{\mathscr{B}}(k, k-1)\right].$$

A confidence set $\widehat{B}$ of Carpentier (2015) is honest with level $1 - \delta$ for $\beta \in \mathcal{P}$, while its diameter is on the order of $(s \log(p/\delta)/n)^{1/2}$ with probability $\ge 1 - \delta$ for any $\beta \in \mathscr{B}(s) \cap \mathcal{P}$ and $s \le \bar{s}$. This is a very interesting theoretical result. However, it is difficult to verify in practice whether β lies in the restricted parameter space $\mathcal{P}$. For $\beta \notin \mathcal{P}$, the confidence set $\widehat{B}$ may not achieve the desired coverage probability. On the contrary, the coverage of our confidence set $\widehat{C}$ is guaranteed for the full parameter space without any restriction. This is a critical result for practical

applications, since it is hard, if not impossible, to verify such a margin condition as in Equation (28) or even a sparsity assumption. The price is that our confidence set can only adapt to a rate no faster than $n^{-1/4}$. Note that for Carpentier's method to achieve a faster rate than $n^{-1/4}$, it is necessary that $s\log(p/\delta) \ll \sqrt{n}$. For high-dimensional data with a relatively small sample size n, this is again a severe assumption on the sparsity level s. Furthermore, Carpentier (2015) did not consider adaptation to signal strength, although the margin condition implies certain minimum signal strength: For any $u \in \widetilde{\mathscr{B}}(k_2, k_1)$, we have $\sum_{j>k_1} u_{(j)}^2 \geq \delta^2$, where $u_{(j)}$ is in descending order in terms of $|u_j|$.

Choice of c_1 and c_2. When $A \neq \varnothing$, we consider two criteria to choose the constants c_1 in (10) and c_2 in (16). The first criterion is to minimize the log-volume of $\widehat{C}$, namely,

$$\log V(\widehat{C}) = k\log(r_A) + (n-k)\log(r_\perp)$$

up to an additive constant, which becomes a constrained optimization problem

$$\min_{c_1,c_2} \left\{k\log(\sqrt{c_1}\tilde{r}_A) + (n-k)\log(\sqrt{c_2}\tilde{r}_\perp)\right\}, \quad \text{subject to } 1/c_1 + 1/c_2 = 1 \text{ and } 1 < c_1, c_2 \leq E, \tag{29}$$

where $\tilde{r}_A$ and $\tilde{r}_\perp$ are defined in Equations (10) and (16) and $E > 2$ is a predetermined upper bound. It is easy to obtain the solution

$$c_1 = \frac{E}{E-1} \vee \left(\frac{n}{k} \wedge E\right), \qquad c_2 = \frac{E}{E-1} \vee \left(\frac{n}{n-k} \wedge E\right). \tag{30}$$

For all numerical results in this article, we use $E = 10$. Without the constraint $c_1, c_2 \leq E$, the minimizer would be $(c_1, c_2) = (n/k, n/(n-k))$ so that under the conditions of Corollary 3, $r_A = \sqrt{n/k}\tilde{r}_A \asymp_p 1$ and thus the diameter $|\widehat{C}|$ would not converge to 0. Therefore, a finite upper bound E must be imposed. The second criterion is to minimize the diameter $|\widehat{C}|$

$$\min_{c_1,c_2} \max\{r_A, r_\perp\}, \text{ subject to } 1/c_1 + 1/c_2 = 1, \tag{31}$$

which yields the solution

$$c_1 = (\tilde{r}_A^2 + \tilde{r}_\perp^2)/\tilde{r}_A^2, \qquad c_2 = (\tilde{r}_A^2 + \tilde{r}_\perp^2)/\tilde{r}_\perp^2. \tag{32}$$

As a result, we have $r_A = r_\perp = (\tilde{r}_A^2 + \tilde{r}_\perp^2)^{1/2}$ and the confidence set reduces to a ball. Note that by definition $\tilde{r}_A$ (10) and $\tilde{r}_\perp$ (16) are independent of the constants c_1, c_2. Their rates are given by the rates of r_A and $r_\perp$ in Lemma 2 (established under the assumption that $c_1 = c_2 = 2$). Consequently, under this criterion,

$$r_A^2 = r_\perp^2 = \tilde{r}_A^2 + \tilde{r}_\perp^2 = O_p\left(\frac{k}{n} + \frac{\sqrt{n-k}}{n} + \frac{\|\mu_\perp\|^2}{n}\right).$$

Then it is easy to verify that all the results in this subsection hold for the second criterion as well.

2.4. Multiple Candidate Sets

It is common to have multiple choices for the candidate set A in our two-step Stein method. Let

$$\mathcal{H} = \{A_m \subseteq [p],\ m = 1, \ldots, M_n\}$$

be a collection of candidate sets. We can apply the two-step Stein method to construct $M = M_n$ confidence sets for μ, denoted by $\widehat{C}_m$, and then choose an optimal set $\widehat{C}_{m^*}$ by certain criterion such as minimizing the volume or the diameter. Furthermore, the cardinality of $\mathcal{H}$ may be unbounded as n increases, that is, $M_n \to \infty$. In what follows, we show that under mild conditions, Equations (11) and (17) hold uniformly for all $A \in \mathcal{H}$ after modifying r_A and $r_\perp$ accordingly, which implies $\widehat{C}_{m^*}$ is asymptotically honest.

Put $k = \text{rank}(X_A)$ for $A \in \mathcal{H}$ and $k_{\max} = \max_{A\in\mathcal{H}} k$. Intuitively, the cardinality of $\mathcal{H}$ (i.e., M) and the maximum size of A in $\mathcal{H}$ (i.e., $k_{\max}$) determine the radii and the coverage probabilities of $\{\widehat{C}_m\}$.

For strong signals, we apply the following concentration inequality to show Equation (11) holds uniformly:

Lemma 3. Suppose χ_n^2 follows a χ^2 distribution with n degrees of freedom. Then for any $\delta > 0$,

$$\mathbb{P}\left\{\sqrt{n}\left|1 - \frac{1}{n}\chi_n^2\right| \geq \delta\right\} \leq 2\exp\left(-\frac{\delta^2}{4}\right). \tag{33}$$

This lemma with a union bound implies

$$\mathbb{P}\left\{\sup_{A\in\mathcal{H}} \sqrt{k}\left|\frac{\chi_k^2}{k} - 1\right| \geq \delta\right\} \leq \sum_{A\in\mathcal{H}} \mathbb{P}\left\{\sqrt{k}\left|\frac{\chi_k^2}{k} - 1\right| \geq \delta\right\} \leq 2M\exp\left(-\frac{\delta^2}{4}\right).$$

Then choosing

$$r_A^2 = c_1\tilde{r}_A^2 = \frac{c_1\sigma^2}{n}\left[k + 2\sqrt{k\log(4M/\alpha)}\right] \tag{34}$$

as the radius for strong signals, we have

$$\mathbb{P}\left\{\sup_{A\in\mathcal{H}} \frac{\|P_A\mu - \hat{\mu}_A\|^2}{nr_A^2} \leq 1/c_1\right\} \geq 1 - \alpha/2.$$

For weak signals, we establish Equation (17) uniformly over $\mathcal{H}$ via the following result:

Lemma 4. Suppose all components of ε in (1), $\varepsilon_i, i = 1, \ldots, n$, have mean 0, common second, fourth, and sixth moments and their eighth moments are bounded by some constant d. For any $\delta > 0$ there exists a positive number D depending on d such that

$$\begin{aligned}\mathbb{P}\left\{\sup_{A\in\mathcal{H}} \sqrt{n-k}\left|\sigma^2\hat{L} - (n-k)^{-1}\|\hat{\mu}_\perp - \mu_\perp\|^2\right| \geq \sigma^2\delta\right\} &\leq \mathbb{P}\left\{\sup_{A\in\mathcal{H}} \sqrt{n-k}\left|\sigma^2 - \frac{1}{n-k}\|P_A^\perp\varepsilon\|^2\right| \geq \sigma^2\frac{\delta}{2}\right\} \\ &+ D\sum_{A\in\mathcal{H}} \frac{1}{(n-k)^2} + D\frac{M}{\delta^4}.\end{aligned} \tag{35}$$

The proof of Lemma 4 mainly follows the ideas in Li (1985). In our model with $\varepsilon \sim \mathcal{N}_n(0, \sigma^2 \mathbf{I}_n)$, the first term on the right-hand side of Equation (35) simplifies to

$$\mathbb{P}\left\{\sup_{A\in\mathcal{H}} \sqrt{n-k}\left|\sigma^2 - \frac{1}{n-k}\|P_A^{\perp}\varepsilon\|^2\right| \geq \sigma^2\frac{\delta}{2}\right\} \leq 2M\exp\left(-\frac{\delta^2}{16}\right)$$

via Lemma 3. Assume that the cardinality of $\mathcal{H}$ and the maximum size of $A \in \mathcal{H}$ satisfy $M \ll (n-k_{\max})^2$. To achieve the desired coverage for weak signals, it is sufficient to pick δ such that $\delta^2 = \Omega(\log M)$ and $\delta^4 = \Omega(M)$. Therefore, we can set

$$\delta = c(\alpha/2)M^{1/4} \gg (\log M)^{1/2}$$

for some constant $c(\alpha/2) > 0$, and the corresponding radius

$$r_\perp^2 = c_2\tilde{r}_\perp^2 = c_2\frac{n-k}{n}\sigma^2\left\{\hat{L} + c(\alpha/2)\frac{M^{1/4}}{\sqrt{n-k}}\right\} \tag{36}$$

for any $A \in \mathcal{H}$, so that the upper bound in Equation (35) is $\leq \alpha/2$. Now we generalize Theorem 1 to establish asymptotic honesty uniformly over $\mathcal{H}$:

Theorem 5. Given $\mathcal{H}$, construct confidence sets $\widehat{C}_m, m = 1,\ldots,M$, with r_A and $r_\perp$ as in Equations (34) and (36), respectively, for $A = A_m$. Suppose $\lim_{n\to\infty} M/(n-k_{\max})^2 = 0$, $1/c_1 + 1/c_2 = 1$, and each A_m is independent of (X, y). Then the confidence sets $\widehat{C}_m$ are uniformly honest over $\mathcal{H}$, that is,

$$\liminf_{n\to\infty}\inf_{\beta\in\mathbb{R}^p}\mathbb{P}\left[\bigcap_m \{X\beta \in \widehat{C}_m\}\right] \geq 1-\alpha.$$

Consequently, $\widehat{C}_{m^*}$ chosen by any criterion is asymptotically honest.

Remark 4. The increment of r_A^2 in Equation (34), $2\sqrt{k\log(4M/\alpha)}/n$, reflects the cost for achieving uniform honesty over $\mathcal{H}$. But this factor will not cause a slower rate for r_A if $\log M = O_p(k^*)$, where k^* is the size of the selected candidate set A_{m^*}. Compared with Equation (16), the factor $M^{1/4}/\sqrt{n-k}$ in Equation (36), also the cost for uniform honesty, will in general lead to slower convergence of $r_\perp$. However, this is a worthwhile price to protect our method from an improper candidate set A that does not satisfy the assumptions in Theorem 2. For example, if the candidate set A misses some strong signals, we may end up with $\hat{L} \asymp_p 1$ and the radius of weak signals $r_\perp$ will not converge to 0 at all. Such bad choices of A will be excluded if $\widehat{C}_{m^*}$ is chosen by minimizing its volume over $\mathcal{H}$. In this sense, our method provides a data-driven selection of an optimal candidate set.

To construct $\mathcal{H}$, we threshold the lasso $\hat{\beta}$ in Equation (21) calculated from (X', y') to obtain

$$A_m = \{j \in [p] : |\hat{\beta}_j| > \tau_m\}, \tag{37}$$

for a sequence of threshold values $\tau_m = a_m\lambda$, for example, $a_m \in [0, 4]$. It is possible for two different τ_m to define the same A, which will be counted once in $\mathcal{H}$. By setting $\tau_m = 0$ for some m, $A = \text{supp}(\hat{\beta})$ will be included in $\mathcal{H}$, though it may not be selected as the optimal $\widehat{C}_{m^*}$. In the proof of Corollary 3, we have shown $\|\hat{\beta}\|_0 = O_p(\sqrt{n})$, and therefore both M and $k_{\max}$ are $O_p(\sqrt{n})$, which means $M \ll (n-k_{\max})^2$ with high probability. As a result, we can guarantee uniform honesty over all $\widehat{C}_m$. Other choices of $\mathcal{H}$ are possible, such as stepwise variable selection with BIC. It is possible that $A = \varnothing$ for a large value of τ_m. In this special case, $r_A = 0$, so the confidence set reduces to a ball, that is, $\{\mu \in \mathbb{R}^n : \|\mu - \hat{\mu}_\perp\|^2 \leq nr_\perp^2\}$.

Remark 5. The multiple candidate sets $\{A_m\}$ define a family of linear subspaces for the projection step. This idea has some connection to the work in Baraud (2004). Given a finite family of linear subspaces $\{S_m\}$ of $\mathbb{R}^n$, Baraud (2004) first tested the hypothesis that the mean vector $\mu \in S_m$ for each m, and calculate a corresponding radius ρ_m. Let $\hat{m}$ index the subspace with the minimum ρ_m among all S_m for which the hypothesis is accepted. Put $\hat{\mu} = P_{S_{\hat{m}}}y$ and $\hat{\rho} = \rho_{\hat{m}}$. A confidence ball is then constructed with center $\hat{\mu}$ and radius $\hat{\rho}$. At a conceptual level, our method is different in a few aspects. First, we have a shrinkage step using Stein estimator to handle the residual after projection. This step will be especially helpful, compared to Baraud's method, if none of the proper subspaces contains the mean vector. Second, Baraud's method was not designed to exploit any potential separation between strong and weak signals, which is one of the key contributions of our approach. More technically, the coverage probability of Baraud's method is guaranteed for any finite sample size n, while ours is asymptotic in nature. Baraud (2004) established an upper bound for the radius $\hat{\rho}$ if the true mean μ is in some subspace S_m. Our general result on the diameter of $\widehat{C}$ does not restrict μ to any subspace, except that $\|\beta_{A^c}\|$ is small (c.f. Theorem 2). An interesting future direction for our work is to develop a similar test-based procedure to select a good subspace for the projection step of the two-step Stein method, such that the volume or diameter of the constructed confidence set is minimized.

2.5. Algorithm and Implementation

We implement our method with a sequence of candidate sets A_m defined by (37). Given the dataset, σ^2, λ in Equation (21) and threshold values $\{a_m\lambda\}_{1\leq m\leq M}$, this section describes some technique details in our algorithm to construct the confidence set (8) by the two-step Stein method.

Data splitting. We split the original dataset into (X', y') and (X, y). Apply lasso on (X', y') to get $\hat{\beta}$ in (21) with the tuning parameter λ. Threshold $\hat{\beta}$ by $\tau_m = a_m\lambda$ for $m = 1,\ldots,M$ in (37) to define candidate sets A_m. Note that A_m, $m = 1,\ldots,M$, are independent of (X, y).

Computation of $c_{st}(\alpha)$. For any candidate set A, the radius $r_\perp$ (16) depends on the constant $c_{\text{st}}(\alpha)$, which is essentially the quantile of the deviation between $\sigma^2\hat{L}$ and the loss of the Stein estimator $\hat{\mu}_\perp$. We use the following simulation procedure to estimate $c_{\text{st}}(\alpha)$. First draw $\check{Y}_j \sim \mathcal{N}_n(0, \sigma^2\mathbf{I}_n)$ for $j = 1, 2, \ldots, N$. For each j, compute

$$\check{\mu}_j = \left(1 - \frac{n\sigma^2}{\|\check{Y}_j\|^2}\right)_+ \check{Y}_j \quad \text{and} \quad \check{L}_j = \left(1 - \frac{n\sigma^2}{\|\check{Y}_j\|^2}\right)_+. \tag{38}$$

Algorithm 1 Two-step Stein method

for $m = 1, \ldots, M$ **do**
  $A = A_m$
  compute $\hat{\mu}_A = P_A y$ and $\hat{\mu}_\perp$ by (12)
  compute c_1 and c_2 according to (29) or (31)
  compute r_A and $r_\perp$ by (10) and (16)
  construct $\widehat{C}_m$ in the form of (8)
end for
find m^* by minimizing the volume or the diameter of $\widehat{C}_m$ over m

Then the $(1-\alpha)$ quantile of the empirical distribution of

$$\frac{\sqrt{n}}{\sigma^2}\left|\sigma^2 \check{L}_j - n^{-1}\|\check{\mu}_j\|^2\right|, \quad j = 1, \ldots, N, \tag{39}$$

is a consistent estimator of $c_{\mathrm{st}}(\alpha)$ as long as $\|\mu_\perp\| = o(\sqrt{n})$, which is the case under the assumptions of Corollary 3. Expression (39) can be written as a function of a χ^2_n random variable, which simplifies its simulation.

Clearly, the estimate of $c_{\mathrm{st}}(\alpha)$ does not depend on A and is used for any candidate set $A \in \mathcal{H}$ in our implementation. Moreover, we find the multiple set adjustments on the radii, that is, the factors of $(\log M)^{1/2}$ and $M^{1/4}$, are usually negligible given a reasonable sample size, say $n \geq 100$. Therefore, we simply use the radii r_A and $r_\perp$ in Equations (10) and (16) for each $A \in \mathcal{H}$.

Algorithm 1 summarizes the two-step Stein method with multiple candidate sets A_m.

Remark 6. In the calculation of $r_\perp$ and $c_{\mathrm{st}}(\alpha)$, we use truncated SURE for $\hat{L} = (1-B)_+$ in (13) and similarly for $\check{L}_j$ in (38). Such a truncated rule has been used for the James-Stein estimator (Efron and Morris 1973) and does not affect the asymptotic validity of our method.

2.6. Estimated Noise Variance

In practice, the noise variance σ^2 is usually unknown. Consequently, an estimated variance $\hat{\sigma}^2$ will be used in (10) and (16) to construct the confidence set $\widehat{C}$ (8). Similar to the candidate set A, we use sample splitting to estimate $\hat{\sigma}^2 = \hat{\sigma}^2(y', X')$ from (X', y') so that $\hat{\sigma}^2$ is independent of (X, y). Under a suitable convergence rate of $\hat{\sigma}^2$, we establish that $\widehat{C}$ is honest and its diameter adapts to the same rate as that in Theorem 2.

Our first step is to generalize Lemma 1 with $\hat{\sigma}^2$ in place of the true error variance σ^2, based on which we show that $\widehat{C}$ is honest over the whole parameter space $\beta \in \mathbb{R}^p$.

Lemma 5. Assume that $y \sim \mathcal{N}_n(\mu, \sigma^2 \mathbf{I}_n)$. Let $\tilde{\mu}$ and $\tilde{L}$ be the Stein estimate $\hat{\mu}(y; 0)$ in (5) and $\hat{L}(y; 0)$ in (6) with σ^2 replaced by $\hat{\sigma}^2$. For any $\alpha \in (0, 1)$ and any sequence $\hat{\sigma}^2 = \hat{\sigma}^2_n$ satisfying $|\hat{\sigma}^2_n - \sigma^2| \leq M_1/\sqrt{n}$ when n is large, there exists a constant $c'_{\mathrm{st}}(\alpha) > 0$ (depending on M_1) such that

$$\limsup_{n\to\infty} \sup_{\mu\in\mathbb{R}^n} \mathbb{P}\left\{\left|\hat{\sigma}^2\tilde{L} - n^{-1}\|\tilde{\mu} - \mu\|^2\right| \geq c'_{\mathrm{st}}(\alpha)\hat{\sigma}^2 n^{-1/2}\right\} \leq \alpha. \tag{40}$$

Theorem 6. Suppose all assumptions in Theorem 1 hold and in addition that $k = o(n)$. Let $\hat{\sigma}^2 = \hat{\sigma}^2_n$ be a sequence satisfying $|\hat{\sigma}^2_n - \sigma^2| \leq M_1/\sqrt{n}$ when n is large. Let r_A be computed as in (10) with $\hat{\sigma}^2$ in place of σ^2 and $r_\perp$ be computed as in (16) with $\hat{\sigma}^2$ and $c'_{\mathrm{st}}(\alpha)$ in place of σ^2 and $c_{\mathrm{st}}(\alpha)$. Then the confidence set $\widehat{C}$ (8) is honest in the sense of (2).

The key assumption in the above theorem on $\hat{\sigma}^2$ is its $\sqrt{n}$-consistency, under which the next lemma shows that the radii of the strong and weak signals, r_A and $r_\perp$, computed with $\hat{\sigma}^2$ converge at the same rates as in Lemma 2.

Lemma 6. Suppose all assumptions in Lemma 2 hold. Let $\hat{\sigma}^2 = \hat{\sigma}^2_n$ be a sequence satisfying $|\hat{\sigma}^2_n - \sigma^2| \leq M_1/\sqrt{n}$ when n is large. If r_A and $r_\perp$ are computed with $\hat{\sigma}^2$ as in Theorem 6, then

$$r_A^2 \asymp_p k/n, \quad r_\perp^2 = O_p\left(\frac{\sqrt{n-k}}{n} + \frac{\|\mu_\perp\|^2}{n}\right).$$

It follows from Lemma 6 that Theorem 2 holds when $\hat{\sigma}^2$ is used in place of σ^2. As discussed in Remark 3, we split the whole data into two equal halves with sample sizes $n = n'$. In the above results, we have assumed that $\hat{\sigma}^2 - \sigma^2 = O(1/\sqrt{n})$. Consequently, if $\hat{\sigma}^2$ is $\sqrt{n}$-consistent, then all nice properties of our method are reserved with probability approaching one. The scaled lasso (Sun and Zhang 2012) provides one way to construct a $\sqrt{n}$-consistent estimator. Under a similar SRC condition and a sparse scaling $s \log p \ll \sqrt{n}$, Theorem 2 in Sun and Zhang (2012) implies that for any $\beta \in \mathscr{B}(s)$, the $\hat{\sigma}^2$ estimated by scaled lasso is $\sqrt{n}$-consistent and a central limit theorem holds, $n^{1/2}(\hat{\sigma}/\sigma - 1) \xrightarrow{d} \mathcal{N}(0, 1/2)$. Finally, we emphasize that $\hat{\sigma}^2$ and the candidate set A can be estimated by different methods, as long as the estimators satisfy their respective conditions with high probability.

Remark 7. Note that $c_{\mathrm{st}}(\alpha)$ is invariant to the value of the true σ^2. Even if we plug $\hat{\sigma}^2$ in the simulation of $c_{\mathrm{st}}(\alpha)$ discussed in Section 2.5, we will still estimate the $c_{\mathrm{st}}(\alpha)$ associated with the true σ^2 instead of $c'_{\mathrm{st}}(\alpha)$. However, the empirical study in Section 4.5 shows that using so estimated $c_{\mathrm{st}}(\alpha)$ with $\hat{\sigma}^2$ does not lead to any decrease in coverage. On the other hand, the proof of Lemma 5 provides a conservative way to theoretically compute $c'_{\mathrm{st}}(\alpha)$ from $c_{\mathrm{st}}(\alpha)$. In particular, if $\hat{\sigma}^2$ is estimated by scaled lasso, we propose an efficient method to approximate $c_{\mathrm{st}}{}'(\alpha)$. See the supplementary materials for more details.

2.7. Main Contributions of Our Work

Here, we briefly summarize the key contributions of our method. By dealing with strong and weak signals separately, our work combines sparse regression techniques with Stein estimation to build an honest and adaptive confidence set in high-dimensional regression. Corollaries 3 and 4 provide theoretical guarantees for the use of popular sparse regression methods, lasso and MCP, in our two-step method. In contrast to many existing works in this area which focus primarily on theoretical aspects, we also make a lot of efforts in practical implementation by approximating all involved constants in our method, such as the computation of $c_{\mathrm{st}}(\alpha)$ in Section 2.5. To

broaden its application, our honesty result in Theorem 1 is almost assumption-free, without restricting to a sparse setting as in Nickl and van de Geer (2013); Carpentier (2015). The numerical results will show that our method works well even if β is dense (Section 4.4) or the relation between y and X is potentially nonlinear (Section 5).

Moreover, the confidence sets by our method can adapt to both sparsity and signal strength. When the signal strength is separable to certain degrees, the diameter of our confidence sets achieves $|\widehat{C}| = O_p(n^{-1/4} \vee \sqrt{s/n})$ (Corollaries 3 and 4) with candidate set A defined by different sparse regression methods. Theorem 2 provides a general statement relating the adaptive rate to the subspace for projection. As far as we know, such theoretical results have not been established for high-dimensional inference. We have also proposed a data-driven way to choose an optimal candidate set among multiple choices, making our method more applicable in practice. Theoretical guarantees (Theorem 5) are established for this data-driven selection.

3. Competing Methods

To illustrate the effectiveness of our two-step Stein method, we first present three alternative procedures that can be derived by extending ideas from construction of nonparametric regression confidence sets in conjunction with lasso estimation. Since all of them make use of the lasso, we review an error bound for lasso prediction due to Bickel, Ritov, and Tsybakov (2009).

3.1. Lasso Prediction Error

Given X, y and $\lambda > 0$, consider the lasso estimator $\hat{\beta} = \hat{\beta}(y, X; \lambda)$ defined as in (21). Let $\omega(X) = \max_j(\|X_j\|^2/n)$. Error bounds of lasso prediction have been established under the restricted eigenvalue assumption (Bickel, Ritov, and Tsybakov 2009). For $S \subseteq [p]$ and $c_0 > 0$, define the cone

$$\mathscr{C}(S, c_0) := \left\{ \delta \in \mathbb{R}^p : \sum_{j \in S^c} |\delta_j| \le c_0 \sum_{j \in S} |\delta_j| \right\}. \tag{41}$$

We say the design matrix X satisfies RE(s, c_0), for $s \in [p]$ and $c_0 > 0$, if

$$\kappa(s, c_0; X) := \min_{|S| \le s} \min_{\delta \ne 0} \left\{ \frac{\|X\delta\|}{\sqrt{n}\|\delta_S\|} : \delta \in \mathscr{C}(S, c_0) \right\} > 0. \tag{42}$$

Lemma 7 (Theorem 7.2 in Bickel, Ritov, and Tsybakov 2009). Let $n \ge 1$ and $p \ge 2$. Suppose that $\|\beta\|_0 \le s$ and X satisfies Assumption RE$(s, 3)$. Choose $\lambda = K\sigma\sqrt{\log(p)/n}$ for $K > 2\sqrt{2}$. Then we have

$$\mathbb{P}\left\{ \|X(\hat{\beta} - \beta)\|^2 \le \frac{16K^2\sigma^2\omega(X)}{\kappa^2(s, 3; X)} s \log p \right\} \ge 1 - p^{1 - K^2/8}. \tag{43}$$

Remark 8. The original theorem in Bickel, Ritov, and Tsybakov (2009) assumes that all the diagonal elements of the Gram matrix $X^{\mathsf{T}}X/n$ are 1 for simplicity, while we remove this assumption by including the term $\omega(X)$.

3.2. Another Adaptive Method

Here, we develop another adaptive method following the procedure in (Robins and van der Vaart 2006, sec. 3), which constructs a confidence set for μ from $y \sim \mathcal{N}_n(\mu, \sigma^2\mathbf{I}_n)$ via sample splitting. Applied to the linear model (1), the method can be described as follows. Split the original dataset into (X', y') and (X, y), of which the former is used to obtain an initial lasso estimate $\hat{\beta} = \hat{\beta}(y', X'; \lambda)$ (21), and the latter is used to compute two quantities

$$R_n = \frac{1}{n}\|y - X\hat{\beta}\|^2 - \sigma^2, \qquad \hat{\tau}_n^2 = \frac{2\sigma^4}{n} + \frac{4\sigma^2}{n^2}\|X\beta - X\hat{\beta}\|^2, \tag{44}$$

where R_n is an estimate of the loss $\|X\beta - X\hat{\beta}\|^2/n$. Then, a confidence ball for $\mu = X\beta$ is constructed in the form of

$$\widehat{C}_a = \left\{ \mu \in \mathbb{R}^n : \frac{R_n - n^{-1}\|\mu - X\hat{\beta}\|^2}{\hat{\tau}_n} \ge -z_\alpha \right\}, \tag{45}$$

where z_α is the $(1-\alpha)$ quantile of the standard normal distribution. Note that $\hat{\tau}_n$ in Equation (45) contains the term $\|\mu - X\hat{\beta}\|$ as well so an explicit form of the confidence ball is

$$\left\{ \mu \in \mathbb{R}^n : \frac{1}{n}\|\mu - X\hat{\beta}\|^2 \le r_a^2 = R_n + O\left(\sqrt{(R_n + 1)/n}\right) \right\},$$

where r_a is the radius.

To establish the convergence rate of the diameter of $\widehat{C}_a$, we need an assumption, similar to RE(s, c_0), on the restricted maximum eigenvalue of $X^{\mathsf{T}}X/n$ over the cone $\mathscr{C}(S, c_0)$ (41). For $s \in [p]$ and $c_0 > 0$, let

$$\zeta(s, c_0; X) := \max_{|S| \le s} \max_{\delta \ne 0} \left\{ \frac{\|X\delta\|}{\sqrt{n}\|\delta_S\|} : \delta \in \mathscr{C}(S, c_0) \right\}.$$

Theorem 7. The $(1 - \alpha)$ confidence set $\widehat{C}_a$ (45) is honest for all $\beta \in \mathbb{R}^p$. Suppose $s \log p = o(n)$, the sequence $X = X(n)$ satisfies

$$\liminf_{n\to\infty} \kappa(2s, 3; X) = \kappa > 0, \quad \limsup_{n\to\infty} \zeta(s, 3; X) = \zeta < \infty,$$
$$\limsup_{n\to\infty} \omega(X) = \omega < \infty,$$

and so does the sequence $X' = X'(n)$. Then with a proper choice of $\lambda \asymp \sqrt{\log p/n}$, for any $\beta \in \mathscr{B}(s)$ the diameter

$$|\widehat{C}_a| = O_p\left(n^{-1/4} + \sqrt{s \log p/n}\right). \tag{46}$$

These properties have been informally discussed in the introduction (Section 1). Although $\widehat{C}_a$ is also honest over the entire parameter space, the upper bound on its diameter critically depends on the sparsity of β. The scaling $s \log p = o(n)$ is the minimum requirement for the lasso to be consistent in estimating μ or β. In general, this scaling is also needed for the RE assumption to hold with $\liminf_n \kappa(2s, 3; X) > 0$ (Negahban et al. 2012) and for the upper bound on $|\widehat{C}_a|$ to be informative. This is different from the universal bound (19) on $\mathbb{E}|\widehat{C}|^2$ for the two-step method. The diameter $|\widehat{C}_a|$ adapts to the optimal rate for sufficiently sparse β as $s \log p = O(\sqrt{n})$; see Remark 2 for related discussion. Our numerical results in Section 4.4 demonstrate that $|\widehat{C}_a|$ can be substantially larger than the diameter of our two-step Stein method when β is not sparse.

3.3. An Oracle Lasso Method

We calculate the lasso $\hat{\beta} = \hat{\beta}(y, X; \lambda)$ from the whole dataset without sample splitting, which we denote by (X, y) in this subsection.

Assuming the true sparsity $s_\beta = \|\beta\|_0$ is known (the oracle), a $(1-\alpha)$ confidence ball for $X\beta$ is constructed as

$$\left\{\mu \in \mathbb{R}^n : \frac{1}{n}\|\mu - X\hat{\beta}\|^2 \le c_o(\alpha)\sigma^2 \frac{s_\beta \log p}{n} := r_o^2\right\},$$

where $c_o(\alpha)$ is a constant depending on the design matrix X and the tuning parameter λ. We estimate $c_o(\alpha)$ by a similar procedure to be described in Section 3.4 for a two-step lasso method. Although there are sharper upper bounds, for example, $O(s_\beta \log(p/s_\beta)/n)$, for lasso prediction error (e.g., Hastie, Tibshirani, and Wainwright 2015, chap. 11), our choice of λ is tuned to achieve the desired coverage rate in our numerical results and thus the corresponding r_o is already optimized in this sense.

It should be pointed out that the oracle lasso is *not* implementable in practice since the true sparsity s_β is unknown. In theory, it can build a confidence set with a diameter on the order of $(s_\beta \log p/n)^{1/2}$, potentially faster than the rate $n^{-1/4}$, however, the constant $c_o(\alpha)$ can be large and difficult to approximate. Indeed, in comparison with the oracle lasso, our method often constructs confidence sets with a smaller volume even under highly sparse settings, which highlights the practical usefulness of our two-step method.

3.4. A Two-Step Lasso Method

To appreciate the advantage of using Stein estimates in the shrinkage step of our construction, we compare our method with a two-step lasso method, in which we replace the Stein estimate by the lasso to build a confidence set for $\mu_\perp$, the mean for weak signals. Consider the two-step method in Section 2.2 with a given candidate set A. Let $k = \text{rank}(X_A)$ and further assume A contains strong signals only, that is, $A \subseteq \text{supp}(\beta)$. We use the same method to find $\hat{\mu}_A$ and r_A (10) in the projection step. Like the oracle lasso, we assume the true sparsity $s_\beta = \|\beta\|_0$ is given and construct a confidence set for $\mu_\perp$ based on the error bound for lasso prediction.

Apply lasso on $(P_A^\perp X, y_\perp) = (P_A^\perp X, P_A^\perp y)$ with a tuning parameter

$$\lambda_2 = K\sigma\sqrt{\log(p-k)/n}, \qquad K > 2\sqrt{2}, \tag{47}$$

to find the estimate

$$\tilde{\beta} = \tilde{\beta}(\lambda_2) = \text{argmin}_{\beta\in\mathbb{R}^p}\left[\frac{1}{2n}\|y_\perp - P_A^\perp X\beta\|^2 + \lambda_2\|\beta\|_1\right]. \tag{48}$$

It is natural to estimate the center $\mu_\perp = P_A^\perp \mu$ by the lasso prediction $\hat{\mu}_\perp = P_A^\perp X\tilde{\beta}$. As a corollary of Lemma 7, we find an error bound for $\|\hat{\mu}_\perp - \mu_\perp\|^2$:

Corollary 8. Let $n \ge 1$ and $p \ge 2$. Suppose that $\|\beta\|_0 \le s$ and Assumption RE$(s, 3)$ holds for X. Choose λ_2 as in Equation (47). Then for any fixed $A \subseteq \text{supp}(\beta)$ with $k = \text{rank}(X_A) < s$, we have

$$\mathbb{P}\left\{\|P_A^\perp X(\tilde{\beta} - \beta)\|^2 \le \frac{16K^2\sigma^2\omega(X)}{\kappa^2(s,3;X)}(s-k)\log(p-k)\right\} \ge 1 - (p-k)^{1-K^2/8}. \tag{49}$$

Accordingly, the radius for weak signals is chosen as

$$r_\perp^2 = c_2\tilde{r}_\perp^2 = c_2 c_l(\alpha/2)\sigma^2\frac{(s_\beta - k)\log(p-k)}{n}, \tag{50}$$

where $c_l(\alpha/2) = c_l(\alpha/2; P_A^\perp X)$ is a constant. Last, we combine $(\hat{\mu}_\perp, r_\perp)$ with $(\hat{\mu}_A, r_A)$ as in Equation (8) to define the confidence set $\widehat{C}$.

Again we use sample splitting to define the candidate set A by thresholding the lasso estimate $\hat{\beta}(y', X'; \lambda)$ in Equation (21) with a threshold value $\tau = \Omega_p(\|\hat{\beta} - \beta\|_\infty)$ so that $\mathbb{P}\left(A \subseteq \text{supp}(\beta)\right) \to 1$, satisfying the assumption in Corollary 8. Upper bounds on $\|\hat{\beta} - \beta\|_\infty$ are available under certain conditions; see, for example, (Hastie, Tibshirani, and Wainwright 2015, theor. 11.3).

Remark 9. Suppose β is sufficiently sparse so that $s_\beta \log p \ll \sqrt{n}$. Then, it follows that both r_A and $r_\perp$ of the two-step lasso converge faster than the rate of $n^{-1/4}$. This is not surprising and shows the advantage of the oracle knowledge of the true sparsity s_β. Of course, in practice we do not know s_β and therefore, this two-step lasso method, like the oracle lasso, is not implementable for real problems. The numerical comparisons in the next section will show that our two-step Stein method, which does not use the true sparsity in its construction, is more appealing than the two-step lasso: Its adaptation to the underlying sparsity is comparable to the two-step lasso, while its coverage turns out to be much more robust.

We follow the same procedure as the two-step Stein method to implement the two-step lasso method with multiple candidate sets $A_m, m = 1, \ldots, M$ — threshold $\hat{\beta}(y', X'; \lambda)$ with a sequence of threshold values to construct A_m (37) and then choose the confidence set with the minimum volume or diameter. The main difference lies in how to approximate $c_l(\alpha)$ in (50), which is done by the following approach.

We first use $b = \max_{i\in[p]}(X_i'^{\mathsf{T}} y')/\|X_i'\|^2$ as a rough upper bound for $\|\beta\|_\infty$. For $j = 1, 2, \ldots, N$, we draw an s_β-sparse vector, $\gamma_j \in \mathbb{R}^p$, of which the nonzero components follow $\mathcal{U}(-b, b)$. Then we sample $Y_j^* \sim \mathcal{N}_n(X\gamma_j, \sigma^2\mathbf{I}_n)$ and calculate lasso estimate $\hat{\gamma}_j(\lambda) = \hat{\beta}(Y_j^*, X; \lambda)$ as in Equation (21) with the tuning parameter λ for all j. Let $c_j = \|X(\hat{\gamma}_j(\lambda) - \gamma_j)\|^2/(\sigma^2 s_\beta \log p)$. For a large N, $c_l(\alpha)$ can be approximated by the $(1-\alpha)$ quantile of $\{c_j\}$. Here, $\lambda = \nu \cdot K\sigma^2\sqrt{\log p/n}$, where $\nu \le 1$ is a predetermined constant. This choice is slightly smaller than the theoretical value in Lemma 7, but gives a stable estimate of $c_l(\alpha)$ with the desired coverage. As we calculate b with (X', y') in the above, our estimate of $c_l(\alpha)$ is independent of the response y. It is possible that a candidate set A_m defined by Equation (37) may contain s or more predictors. In this case, we will only include the largest $s-1$ predictors in terms of their absolute lasso coefficients, as Corollary 8 requires $|A_m| < s$.

4. Numerical Results

We will first compare our method with the above competing methods when β is sparse relative to the sample size, that is, s/n is small, and then consider the more challenging settings in which s is comparable to n. We will also examine the performance of these methods with an estimated error variance and their robustness when key assumptions for the error distribution are violated.

4.1. Simulation Setup

The rows of X and X', both of size $n \times p$, are independently drawn from $\mathcal{N}_p(0, \Sigma)$ and the columns are normalized to have an identical ℓ_2 norm. We use three designs for Σ as in Dezeure et al. (2015):

$$\begin{aligned} \text{Toeplitz:} &\quad \Sigma_{i,j} = 0.5^{|i-j|}, \\ \text{Exp.decay:} &\quad (\Sigma^{-1})_{i,j} = 0.4^{|i-j|}, \\ \text{Equi.corr:} &\quad \Sigma_{i,j} = 0.8 \text{ for all } i \neq j, \Sigma_{i,i} = 1 \text{ for all } i. \end{aligned}$$

The support of β is randomly chosen and its s nonzero components are generated in two ways:

1. They are drawn independently from a uniform distribution $\mathcal{U}(-b, b)$.
2. Half of the nonzero components follow $\mathcal{U}(-b, b)$ while the other half following $\mathcal{U}(-0.2, 0.2)$, so there are two signal strengths under this setting.

Last, y and y' are drawn from $\mathcal{N}_n(X\beta, \sigma^2\mathbf{I}_n)$ and $\mathcal{N}_n(X'\beta, \sigma^2\mathbf{I}_n)$, respectively. In our results, we chose $n = n' = 200$, $p = 800$, $\sigma^2 = 1$ and $s = 10$, and b took 10 values evenly spaced between $(0, 1)$ and $(1, 5)$. In total, we had 60 simulation settings, each including one design for Σ, one way of generating β, and one value for b. Under each setting, 100 datasets were generated independently, so that the total number of datasets used in this simulation study was 6000.

The confidence level $1 - \alpha$ was set to 0.95. The threshold values $\{a_m\}$ in Equation (37) were evenly spaced from 0 to 4 with a step of 0.05. All the competing methods use lasso in some of the steps, and the tuning parameter λ was chosen by three approaches: 1) the minimum theoretical value in Bickel, Ritov, and Tsybakov (2009), $\lambda_{\text{val}} = 2\sqrt{2}\sigma\sqrt{\log p/n}$, 2) cross-validation λ_{cv}, and 3) one standard error rule λ_{1se}. For the one standard error rule, we choose the largest λ whose test error in cross-validation is within one standard error of the error for λ_{cv}. Since it is time-consuming to approximate $c_o(\alpha) = c_o(\alpha; X, \lambda)$ for the oracle lasso when λ is chosen by a data-dependent way, we set $c_o(\alpha; X, \lambda_{\text{cv}}) = \eta_1 c_o(\alpha; X, \lambda_{\text{val}})$ and $c_o(\alpha; X, \lambda_{\text{1se}}) = \eta_2 c_o(\alpha; X, \lambda_{\text{val}})$, where the factors η_k were chosen such that the overall coverage rate across datasets simulated with $b > 0.3$ was around the desired level.

Unlike the adaptive method in Section 3.2 and our two-step methods, the oracle lasso method does not require sample splitting. Consequently, a confidence set is constructed based on the whole dataset including both (X, Y) and (X', Y') for a fair comparison. We compare the geometric average radius $\bar{r} = (r_A^{|A|} r_\perp^{n-|A|})^{1/n}$ of our two-step methods with r_a of the adaptive method and r_o of the oracle lasso. This is equivalent to comparing the volumes of the confidence sets.

4.2. Results on the Two-Step Stein Method

In this subsection, we compare the two-step Stein method with the adaptive method and the oracle lasso. The constants c_1 and c_2 of our method were chosen by minimizing the volume in Equation (29) with upper bound $E = 10$.

Figure 1 compares the geometric average radius $\bar{r}$ among the three methods against the signal strength b under the first way of drawing β. Each point in a panel was computed by averaging $\bar{r}$ from 100 datasets under a particular simulation setting. It is seen from the figure that $\bar{r}$ by our method was dramatically smaller than the other two methods for almost every setting. This suggests that the volumes of our confidence sets were orders of magnitude smaller than the other two methods, as the ratio of the radii will be raised to the power of $n = 200$ for comparing volumes. When X was drawn from the equal correlation (Equi.corr) design, $\bar{r}$ of the oracle lasso and the adaptive methods kept increasing as b increased, while $\bar{r}$ by our method became stable after $b > 2$. Overall, the equal correlation design was more challenging than the other two designs, for which our method outperformed the other two methods with the largest margin. Unlike the other two methods, our method was less sensitive to the choices of λ and the designs of X. Essentially, r_A and $r_\perp$ by our method are determined by the candidate set A. Even if a different λ is used, our method can choose adaptively an optimal A close to $\text{supp}(\beta)$, showing the advantage of using multiple candidate sets.

In a similar way, Figure 2 plots $\bar{r}$ against b in the second scenario of drawing β. When b is large (e.g, $b \geq 1$), the β contains a mixture of weak and strong signals. Again, we see that $\bar{r}$ of our method was smaller than the other two competitors for most settings. The average radius by our method often decreased as $b > 1$, which shows that our method can properly distinguish strong signals and weak signals.

The coverage rates, each computed from 100 datasets, for each of the three ways of choosing λ are summarized in Figure 3. We pooled the results from three types of design matrices together in the figure, because the coverage rates distributed similarly across them. The coverage rates of our method matched the desired 95% confidence level very well, with coverage rate > 0.9 for 96% of the cases. This result is particularly satisfactory for a quite small sample size of $n = 200$. The adaptive method also showed a good coverage, but slightly more conservative than the desired level. The oracle lasso had the most variable coverage rate across different settings when λ was selected in a data-dependent way (λ_{cv} or λ_{1se}). In fact, its coverage could drop below 0.5 for these two cases (not shown in the figure). This shows the difficulty in practice to construct stable confidence sets using error bounds like Equation (43) even with a known sparsity. Together with the results in Figures 1 and 2, this comparison demonstrates the advantage of the proposed two-step Stein method: It builds much smaller confidence sets, while closely matching the desired confidence level. In particular, our confidence sets were uniformly smaller than those by the adaptive method (Section 3.2) for all simulation settings and all choices of λ.

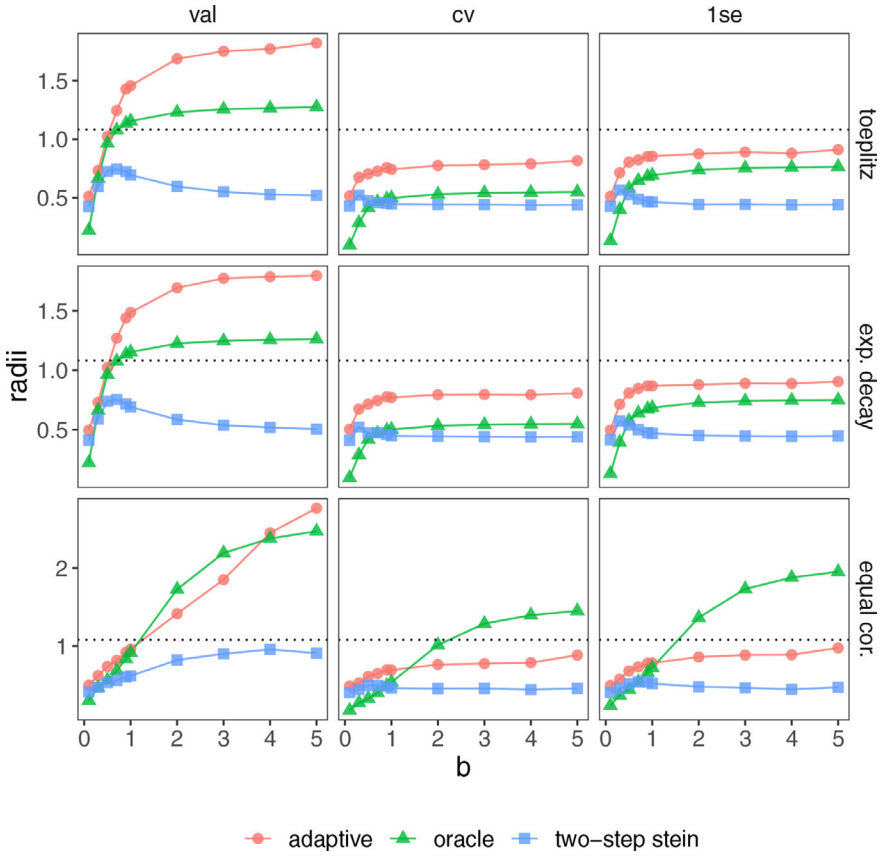

Figure 1. Geometric average radius against b under the first way of generating β. Each panel reports the results for one type of design (row) and one way of choosing λ (column), where the dashed line indicates the naive χ^2 radius.

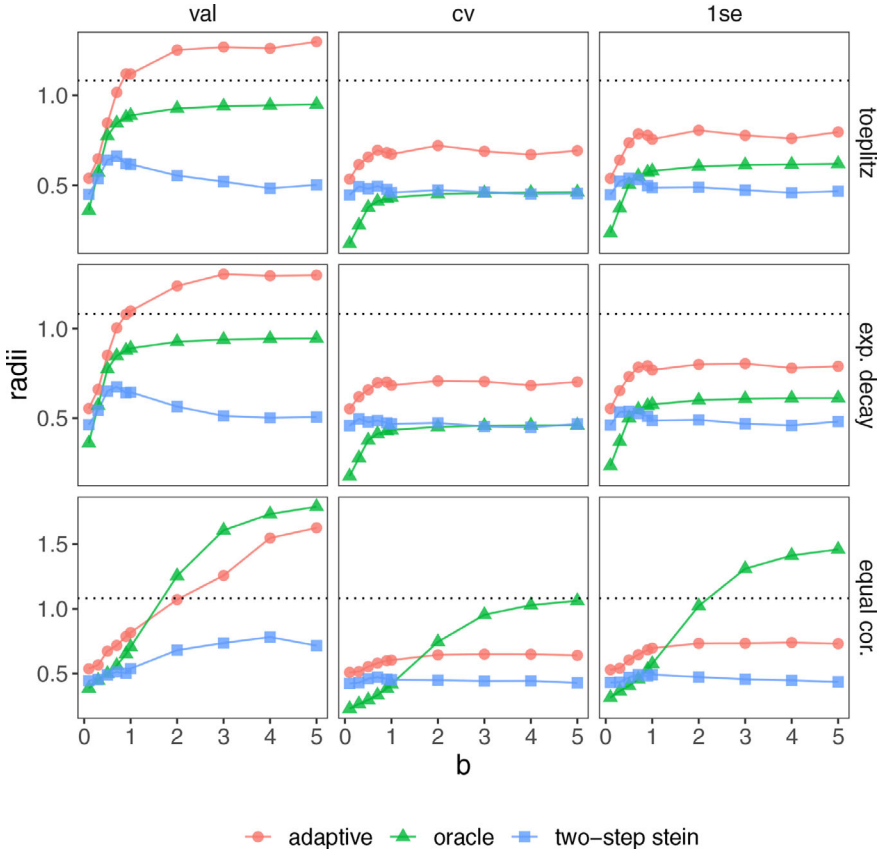

Figure 2. Average radius $\bar{r}$ against b in the second scenario of generating β.

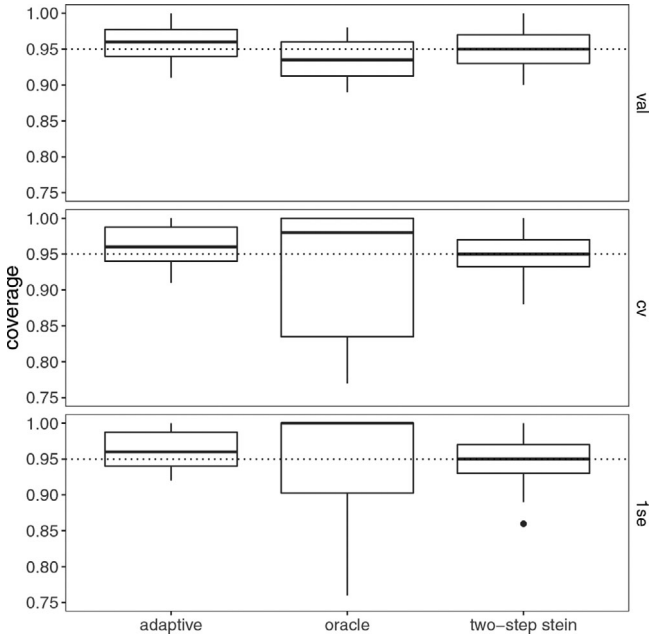

Figure 3. Boxplots of coverage rates for each choice of λ, pooling data from three designs. The dashed lines indicate the desired confidence level of 95%.

4.3. Comparison With the Two-Step Lasso Method

We discussed in Section 2.3 two ways to choose c_1 and c_2, that is, by minimizing the volume or by minimizing the diameter of the confidence set for our proposed two-step framework. Here we compare the two-step Stein method and the two-step lasso, each with the two ways to choose the constants. The two-step Stein method by minimizing the volume (abbreviated as TSV) is the same method used in the previous comparison. Similarly, we use the short-hand TSD, TLV, and TLD for the two-step Stein method by minimizing diameter, the two-step lasso method by minimizing volume and by minimizing diameter, respectively. The true sparsity $s = 10$ was given to the two-step lasso methods. Only the first scenario of generating β was considered in this comparison, since most results in the second scenario were similar. Figure 4 shows the plots of radius against b by the four methods under different settings, while Figure 5 reports the distribution of the coverage rates. The two-step lasso methods apply the lasso twice, one to generate candidate sets A_m and the other to compute $\hat{\mu}_\perp$ and $r_\perp$ for weak signals. To clarify, the three ways of choosing λ in these figures refer to the step to generate candidate sets A_m, while λ_2 in Equation (48) was set to $\nu K\sigma^2\sqrt{\log(p-|A|)/(n-|A|)}$, where $\nu = 0.5$ in our simulation.

We make the following observations from the two figures. First, the two-step Stein methods showed a substantially more satisfactory coverage than the two-step lasso methods. The coverage was close to 0.95 for both TSV and TSD, while the coverage rates of TLV and TLD had a much larger variance and were especially poor when λ was chosen via cross-validation. The confidence sets by the two-step lasso methods had a slightly smaller average radius than the two-step Stein methods for the Toeplitz and the exponential decay designs. However, given their low and unstable coverage rates, this does not imply the two-step lasso methods constructed better confidence sets. Recall that $|\widehat{C}| = O_p(n^{-1/4} \vee \sqrt{s/n})$ for the two-step Stein methods and $|\widehat{C}| = O_p(\sqrt{s\log p/n})$ for the two-step lasso methods. The signals were very sparse in our simulation, with $s = 10$ much smaller than p, favorable for the two-step lasso methods. Even so, we find the two-step Stein methods very competitive, noting that the radii of both TSV and TSD were actually comparable or slightly smaller than the two-step lasso methods for the equal correlation designs, in which the predictors were highly correlated. This comparison demonstrates that the two-step Stein method is more appealing in practice, as it does not require any prior knowledge about the underlying sparsity but gives a better and more stable coverage. Second, both ways of choosing the constants c_1 and c_2 worked well for the two-step Stein method. On the contrary, it is seen from Figure 5 that the coverage rate of TLV was significantly lower than that of TLD in the bottom two panels. Lastly, between using λ_{cv} and λ_{1se} in the lasso for defining candidate sets A_m, we recommend the latter, as it tends to give comparable radii but a better coverage, especially for the two-step lasso.

We also compared the performance between the oracle lasso method and TLD, both constructing confidence sets based on the lasso prediction (43) with a known sparsity. The coverage rates of the two methods were quite comparable as reported in Figures 3 and 5. The geometric average radius of the oracle lasso method (Figure 1) was 2 to 5 times that of TLD (Figure 4). The difference was especially significant when the signal strength was high (large b). This comparison confirms that, by separating strong and weak signals, our two-step framework can greatly improve the efficiency of the constructed confidence sets.

4.4. Dense Signal Settings

We have shown the advantages of our two-step Stein method in the last two subsections under sparse settings. Recall that the dimension of our data was $(n, p) = (200, 800)$ with sparsity $s = 10$ for β in the previous comparisons. The goal of this subsection is to illustrate the stable performance of our method when the true signal is dense. As such, we changed the sparsity to $s = 100$ for the first way of generating β and $s = 200$ for the second way of generating β. We focused on the equal correlation design, which was the most difficult one among the three designs. With the same set of values for the signal strength b, we had 20 distinct parameter settings for data generation in this comparison, and again we simulated 100 datasets under each setting. The tuning parameter λ was selected as λ_{1se} for all the results here.

Figure 6 compares the geometric average $\bar{r}$ against b and the coverage among the adaptive method, the oracle lasso and our two-step Stein method. In all the scenarios reported in panels (a) and (b), our method outperformed the other two methods with very big margins in terms of the volume of a confidence set. For $b > 1$, the radius of our method approached the naive χ^2 radius $(\chi^2_{n,\alpha}/n)^{1/2}$ as suggested by Theorem 1, while the radii of the oracle lasso and the adaptive methods kept increasing to a level much greater than the naive χ^2 radius. This shows that the two competing methods failed to construct acceptable confidence sets when the signal was dense. Since the sparsity level s is comparable to n for the datasets here, the upper bounds for the diameters of these two methods, $|\widehat{C}_o| = O_p(\sqrt{s\log p/n})$

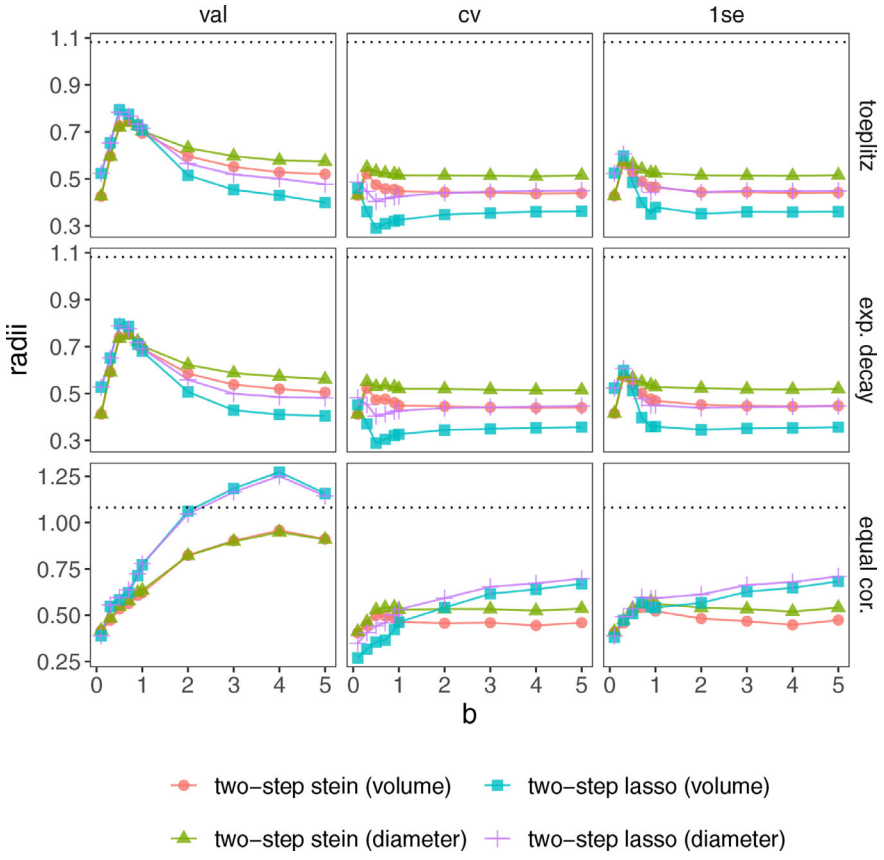

Figure 4. Average radius $\bar{r}$ against b in the first scenario of generating β.

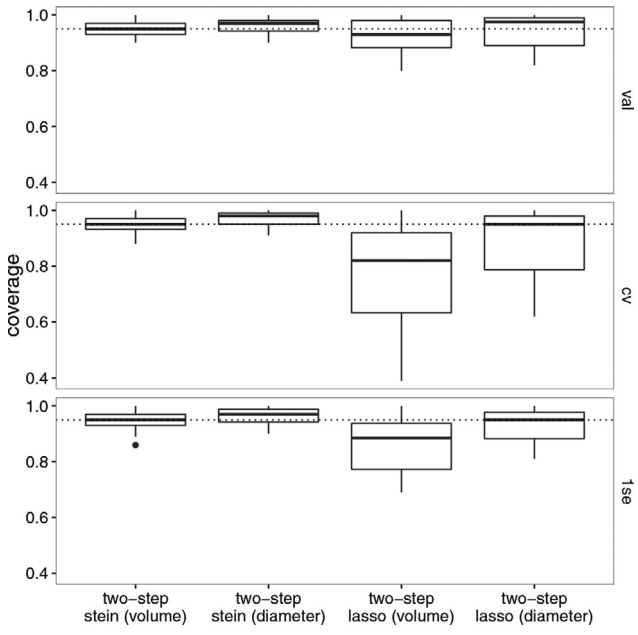

Figure 5. Boxplots of coverage rates for each choice of λ. The dashed lines indicate the desired confidence level of 95%.

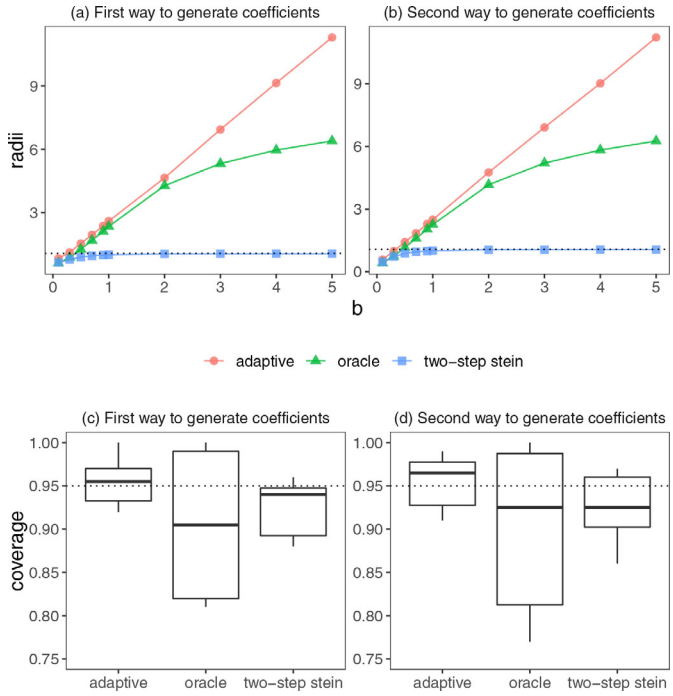

Figure 6. Comparison results under dense signal settings. (a) and (b) Geometric average radius against b. (c) and (d) Boxplots of the coverage rates.

and $|\widehat{C}_a| = O_p(n^{-1/4} + \sqrt{s \log p/n})$, are no longer useful or even valid. It is seen from Figure 6(c) and (d) that the coverage rates of the two-step Stein method were much better than the oracle lasso, but slightly lower than the adaptive method. Nevertheless, our confidence sets still maintained a minimum coverage of 0.9 in most cases, which is quite satisfactory given the way smaller diameters than the adaptive method.

To understand the behavior of our method in this dense signal setting, we examined the number of variables selected as

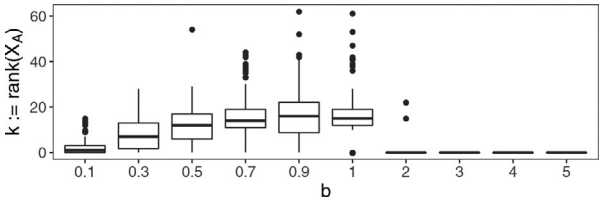

Figure 7. The boxplot of k across datasets for each value of b

strong signals in the set A, that is, $k = |A|$. Figure 7 displays the boxplot of k across 100 datasets for each value of b under the first way to generate β. When $b \leq 1$, our two-step method still chose a nonempty candidate set, but k dropped to 0 for $b \geq 2$, that is, $A = \varnothing$. Note that the radius of our method will be close to the naive χ^2 radius when $k = n$ or $k = 0$; see (18) in Theorem 1. When the signal strength $b \leq 1$, some small nonzero coefficients are close to zero so β is effectively quite sparse, in which case the lasso can select a good subset A of strong signals. On the contrary, when b is large, the lasso will not be able to select a majority of the strong signals, leaving $\|\mu_\perp\| = \|P_A^\perp \mu\|$ too big. In this setting, our method automatically adjusts its "optimal" choice to $A = \varnothing$, constructing a confidence set centered at the Stein estimate $\hat{\mu}(y; 0)$ with radius estimated via the SURE.

4.5. Estimated Error Variance

We further examine the performance of our method using a plug-in $\hat{\sigma}^2$ instead of the true variance σ^2. Recall that we split our sample into (X', y') and (X, y). First, an estimated variance $\hat{\sigma}^2 = \hat{\sigma}^2(X', y')$ was calculated by ordinary least-square regression of y' onto $X'_{A'}$, where A' is the set of variables selected by the scaled lasso (Sun and Zhang 2012, 2013). Although the scaled lasso provides a consistent estimator for σ^2, it sometimes yielded extremely large $\hat{\sigma}^2$, which led to inaccurate inference by all the methods. In contrast, the least-square estimate after the scaled lasso selection gave a much more stable value. To simplify the comparison, we only used a single candidate set $A = \text{supp}(\hat{\beta})$ in this comparison, where $\hat{\beta}$ is the lasso estimate with λ chosen by the three approaches in Section 4.1. In particular, $\hat{\sigma}^2$ was used in place of σ^2 to calculate the theoretical value λ_{val}. We input the same $\hat{\sigma}^2$ to the adaptive and the oracle lasso methods.

For brevity, we only present results on the datasets simulated under the first way of generating β as in Section 4.1. The average radii and coverage rates are reported in Figures 8 and 9, respectively. It is seen from Figure 8 that the trend of $\bar{r}$ against the signal strength b is quite similar to Figure 1 for all three methods. Our two-step Stein method constructed smaller confidence sets than the other two methods for most settings, except for the equal correlation design under which the $\bar{r}$ of our method was quite comparable to that of the adaptive method when λ was selected by cross-validation or the one standard error rule. As shown in Figure 9, the overall coverage of the adaptive method and our method was around or above the desired level of 95% for most settings. In particular, the coverage rates of our method were slightly higher than the adaptive methods when using λ_{cv} or $\lambda_{1\text{se}}$, two practical ways of choosing the lasso tuning parameters. There are some outliers in the boxplots, representing low coverage rates for some datasets generated under the equal correlation design—the most difficult design due to high correlation among the predictors. Using λ_{val}, the adaptive method and our method yielded almost an equal number of outliers, while using λ_{cv} or $\lambda_{1\text{se}}$ our method had fewer outliers.

As expected, the coverage rates here in Figure 9 are somewhat lower than those reported in Figure 3 assuming σ^2 is known. Among those datasets for which either our method or the adaptive method failed to cover the true β, the $\hat{\sigma}^2$ for more than 60% of them was either < 0.8 or > 1.2 (recall $\sigma^2 = 1$), suggesting that the lower coverage was mostly caused by the inaccuracy of $\hat{\sigma}^2$. On the other hand, the pattern of $\bar{r}$ of our method under the Toeplitz and the exponential designs is very similar between Figure 1 for known σ^2 and Figure 8 here, while the $\bar{r}$ of the adaptive method increased slightly when $\hat{\sigma}^2$ was plugged in. Under the equal correlation design, the $\bar{r}$ of our method also increased but not faster than the adaptive method.

4.6. Normality and Homogeneity Assumptions

Our method is developed under normality and homogeneity assumptions that the error vector $\varepsilon \sim \mathcal{N}_n(0, \sigma^2 \mathbf{I})$, which may not hold in practice. In this section, we test the robustness of our two-step Stein method when the above assumptions are violated in comparison with the adaptive method. To this end, we designed the following four simulation settings. Let t_d denote the t-distribution with d degrees of freedom. In the first setting, all components of ε were independently drawn from t_4 with a scale parameter σ, while in the second setting from t_7. These two settings were designed to test the robustness against the violation of normality, and the next two settings against the homogeneity assumption. Let μ_α be the α-percentile of the components $\mu_i, i \in [n]$ of the mean vector $\mu = X\beta$. We drew $\varepsilon_i \sim \mathcal{N}(0, \sigma_i^2)$ independently for $i = 1, \ldots, n$, where

$$\sigma_i = \sigma + 4\sigma(\mu_i - \mu_{0.05})_+/(\mu_{0.95} - \mu_{0.05})$$

in the third setting and

$$\sigma_i = \sigma + 9\sigma\{(\mu_i - \mu_{0.05})_+/(\mu_{0.95} - \mu_{0.05})\}^2$$

in the fourth setting. These two models were motivated by the observation that the variance of ε_i usually increases with μ_i. In particular, σ_i increases quadratically with μ_i in the fourth setting, severely against the homogeneity error assumption. We only tested the Toeplitz design in this study, while using the same choices of the other parameters in data generation as in Section 4.1. The lasso tuning parameter λ for both methods was selected by the one standard error rule, and $\hat{\sigma}^2$ was estimated in the same way as in Section 4.5.

The average radii and coverage rates of the constructed confidence sets are summarized in Figure 10. It is comforting to see that the coverage rates of both methods across all settings were above or close to the nominal level of 95%, with only mild drop compared to their coverage rates under iid normal errors (lower panel of Figure 9). This observation shows that both methods are quite robust against possible violation of error assumptions. On the other hand, the average radius of our two-step Stein method was uniformly smaller than that of the adaptive method (top panels of Figure 10) in all the four settings, demonstrating

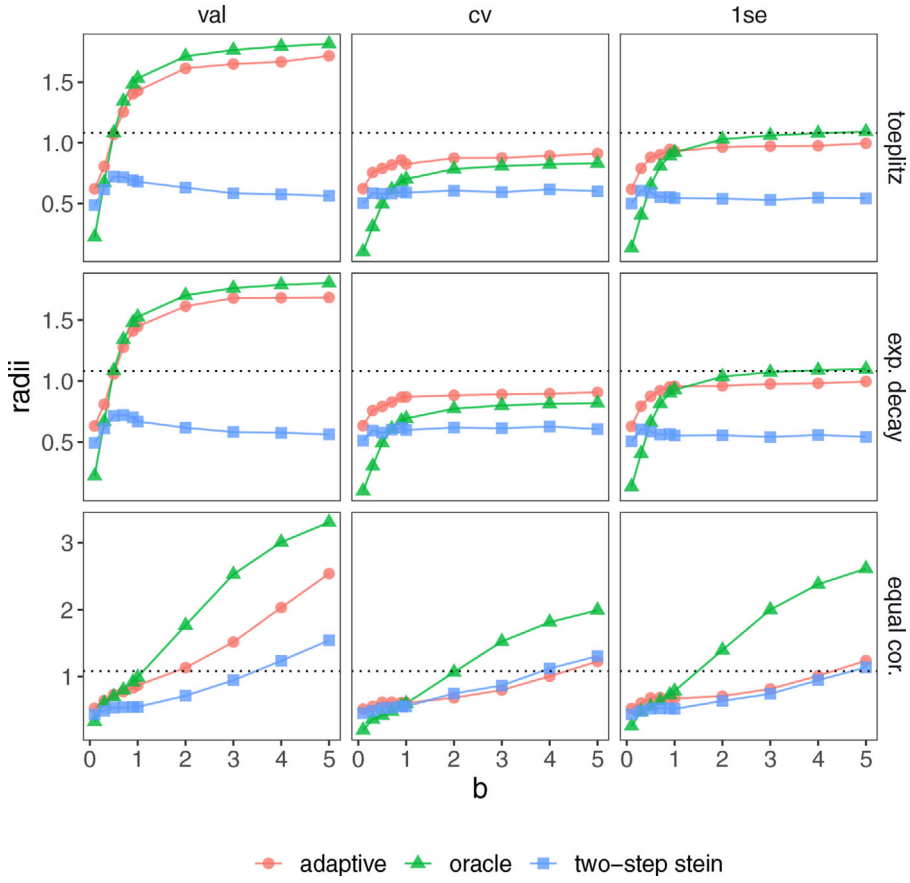

Figure 8. Average radius $\bar{r}$ against b with estimated error variance $\hat{\sigma}^2$.

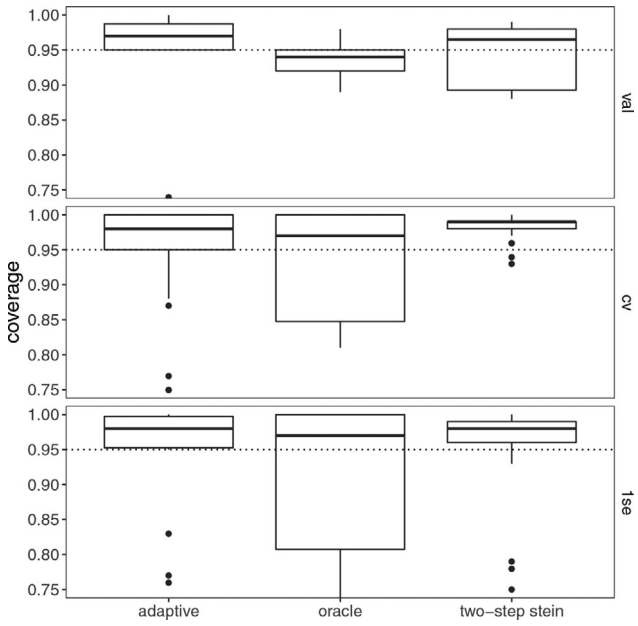

Figure 9. Boxplots of coverage rates for each choice of λ with estimated $\hat{\sigma}^2$. The dashed lines indicate the desired confidence level of 95%. Outliers below 0.75 are truncated.

the higher relative efficiency of our confidence sets when model assumptions are not satisfied, or even severely violated.

As shown in Figure 10, as b increased, the average radius of the adaptive method approached or exceeded the estimated naive χ^2 radius, $\hat{\sigma}(\chi^2_{n,\alpha}/n)^{1/2}$, under iid normal errors, where $\hat{\sigma}^2$ is the estimated error variance. This trend suggests that the adaptive method could be too conservative when the model assumptions are violated, with diameter not necessarily converging to 0. In contrast, the average radius of our two-step Stein was stable and uniformly $< \hat{\sigma}$ for all values of b.

For our method, the shrinkage factor $B = (n-k)\hat{\sigma}^2/\|y_\perp\|^2$ defined in (14) plays a vital role against heterogeneity. Note that the left-hand side of the inequality (15) is essentially determined by B. Even the error variances are different, $\|y_\perp\|^2/\sqrt{n-k}$ still follows approximately a normal distribution when $n-k$ is large, similar to the case with homogeneous errors. Consequently, the distribution of B does not change that much and the inequality (15) still holds in spite of error heterogeneity, which guarantees good coverage for our method.

5. Real Data Analysis

In this section, we apply the two-step Stein method on the riboflavin dataset compiled by Bühlmann, Kalisch, and Meier (2014) to demonstrate its practical significance. This dataset contains a real-valued response variable y, which is the logarithm of the riboflavin production rate, and the expression levels in log-scale of $p = 4088$ genes as covariates. There are $n = 71$ individuals in total so that the design matrix X is 71×4088. Unlike van de Geer et al. (2014) and Dezeure et al. (2015) that aim at gene selection, we focus on joint inference about the mean riboflavin production rates for a

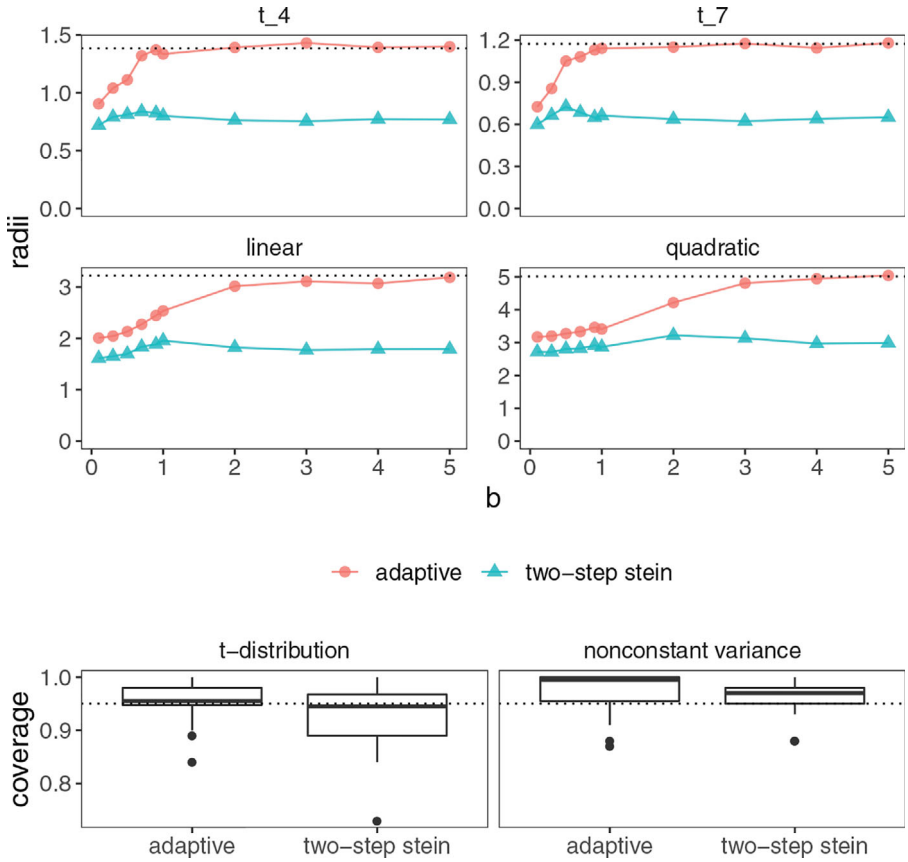

Figure 10. (Upper) Average radius $\bar{r}$ against b and (lower) boxplots of coverage rates of all settings under t-distributions or heterogeneous error variance. The dashed lines in the four top panels indicate the average naive χ^2 radius. The dashed lines in the boxplots indicate the nominal coverage level of 95%.

group of individuals, which is also a scientifically significant problem.

Before our analysis, the columns of X was normalized to have an identical ℓ_2 norm and y was centered to have zero mean. Again, we split (X, y) into two subsamples. One of them was used to calculate an initial lasso estimate $\hat{\beta}$ (21) for the adaptive method and a single candidate set $A = \text{supp}(\hat{\beta})$ for our method, as well as an estimated variance $\hat{\sigma}^2$. The tuning parameter for the lasso estimate $\hat{\beta}$ was chosen by the one standard error rule, while $\hat{\sigma}^2$ was calculated by least-square regression after scaled lasso selection, the same procedure used in Section 4.5. The other subsample was used to construct a confidence set. In our analysis, the n individuals were partitioned into two subsamples by their gene expression clustering pattern. Define the distance between two individuals by $1 - |\rho|$, where ρ is the correlation coefficient between their gene expression vectors. The hierarchical clustering dendrogram on the n gene expression vectors is shown in Figure 11, from which we see a clear separation into two clusters. It makes sense to infer the riboflavin production rates simultaneously for individuals in the same cluster, due to the strong correlation among their gene expression profiles. We also swapped the two subsamples to build two confidence sets, one for each subsample.

We applied our method and the adaptive method to this dataset to construct 95% confidence sets. The results are summarized in Table 1. One sees that the radius $\bar{r}$ of the adaptive method was substantially greater than the $\bar{r}$ of our method. Considering $\|y\|/\sqrt{n} = 0.914$, the confidence sets constructed

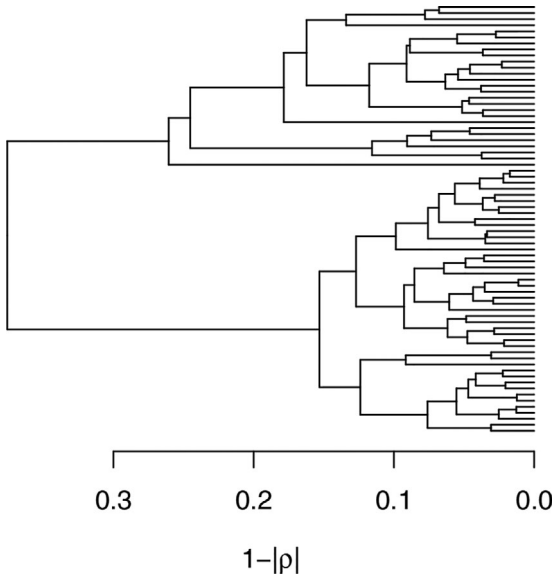

Figure 11. Hierarchical clustering of gene expression vectors among all individuals.

by our method achieved substantial reduction in the uncertainty in μ, especially given the small sample sizes $n \leq 44$ after sample splitting and the large number $p > 4000$ of covariates. To appreciate how much smaller the confidence sets of our method were, consider a sequence of vectors $v^{(m)} \in \mathbb{R}^n$ for $m = 0, 1, \ldots, n$. The ith coordinate of $v^{(m)}$ is given by $v_i^{(m)} = y_i 1(|y_i| > t_m)$, where t_m is a threshold value and $1(\cdot)$ the

Table 1. Confidence sets constructed on the riboflavin dataset.

		Cluster 1	Cluster 2
	Cluster size	27	44
Adaptive	$\bar{r}$	0.641	0.943
Two-step Stein	$\bar{r}$	0.375	0.337
	r_s	0.409	0.175
	$r_\perp$	0.354	0.348

indicator function. We chose an increasing sequence, $t_0 = 0$ and $t_m = |y_{(m)}|$ for $m = 1, \ldots, n$, where $y_{(m)}$ is ordered so that $|y_{(1)}| < \cdots < |y_{(n)}|$. In particular, $v^{(0)} = y$ is the observed response vector and $v^{(n)} = 0$ is the origin. The confidence set for cluster 2 ($n = 44$, Table 1) by the adaptive method contains all $v^{(m)}$, including $v^{(n)} = 0$. In contrast, our confidence set contains only the first 24 of them, that is, $v^{(0)}, \ldots, v^{(23)}$. Loosely speaking, the adaptive method allows the mean of the individual with the largest absolute response value $y_{(n)}$ to be zero, suggesting that the responses of all the 44 individuals in cluster 2 could be the same (as y had been centered). Our method suggests this is not the case. Ranking the individuals by the absolute values of their responses, the first 23 individuals, that is, those with responses $y_{(1)}, \ldots, y_{(23)}$, behave rather differently from the other 21 individuals.

The sharp reduction in size of the confidence sets in favor of our method is clearly observed in Table 1. But since the true riboflavin production rate for each individual is unknown, one may doubt if the reduction in the volume of a confidence set might result from compromising the coverage probability. To address this concern, we conducted two simulation studies based on this dataset to assess the coverage probability. In the first simulation, we generated a coefficient vector $\beta' \in \mathbb{R}^p$ by randomly choosing s nonzero components uniformly in $(-b\tilde{\sigma}, b\tilde{\sigma})$, and then drew $y' = X\beta' + \varepsilon'$, $\varepsilon' \sim \mathcal{N}_n(0, \tilde{\sigma}^2 \mathbf{I}_n)$, where $\tilde{\sigma} = 0.320$ was estimated from (X, y). We computed confidence sets for the two clusters given (X, y'), and verified whether the confidence sets covered the true mean $X\beta'$. The whole process was repeated 100 times for each $b \in \{0.1, 0.3, 0.5, 0.7, 0.9, 1\}$. We chose $b \le 1$ since the maximum magnitude of the estimated β from the original data was close to $\tilde{\sigma}$. The sparsity $s = 14$ was chosen to match the support size of the scaled lasso solution. Note that a total of 600 datasets were generated, each under a random β' with different support and coefficients. In the second study, we perturbed y to simulate $y^* \sim \mathcal{N}_n(y, \tilde{\sigma}^2 \mathbf{I}_n)$, and then applied each of the two competing methods on the perturbed data (X, y^*) to construct confidence sets and checked whether they covered the original response vector y. Although y is the mean of y^*, the relation between y and the predictors X is noisy and could be nonlinear, which makes this test more challenging. The whole process, starting from the simulation of y^*, was repeated 400 times.

The average results of both simulations are summarized in Table 2. Considering the extremely high dimension and small sample sizes ($p > 4000, n \le 44$), the coverage probabilities of both methods ($\ge$ 90%) are quite satisfactory. Consistent with the results for the original data (Table 1), our method achieved much smaller average radius $\bar{r}$ across all cases than the adaptive method. Moreover, the $\bar{r}$ of the adaptive method was even greater than the radius of the naive χ^2 set for the more challenging perturbation datasets, making it not practically useful, while the $\bar{r}$ of our method was still smaller than the naive radius. This is a very encouraging result given the noisy and potentially nonlinear relationship between y and X, as we mentioned above.

Table 2. Confidence sets for datasets simulated by linear models or via perturbation based on the riboflavin data.

		Cluster 1	Cluster 2
	Cluster size	27	44
	χ^2 Radius	0.541	0.424
Adaptive	$\bar{r}$	0.425	0.438
(linear model)	Coverage	0.983	0.975
	$\bar{r}$	0.297	0.259
Two-step Stein	r_s	0.413	0.311
(linear model)	$r_\perp$	0.293	0.255
	Coverage	0.975	0.905
Adaptive	$\bar{r}$	0.745	0.985
(perturbation)	Coverage	0.898	0.942
	$\bar{r}$	0.530	0.411
Two-step Stein	r_s	0.631	0.482
(perturbation)	$r_\perp$	0.478	0.404
	Coverage	0.960	0.933

Lastly, it is worth reiterating that our confidence set makes *simultaneous inference* on all μ_i, $i = 1, \ldots, n$. As the sample size n becomes large, the diameter of the set will shrink to zero at certain rate (e.g., $n^{-1/4}$). This is particularly useful when we wish to control family-wise error rate over a large number of individual tests (n large). On the contrary, if we apply Bonferroni correction on n individual inferences, each on a single μ_i, the power can be much lower than our approach. This highlights another aspect of the practical significance of our inference method.

6. Discussion

For high-dimensional regression, oracle inequalities for sparse estimators cannot be directly utilized to construct honest and adaptive confidence sets due to the unknown signal sparsity. To overcome this difficulty, we have developed a two-step Stein method, via projection and shrinkage, to construct confidence sets for $\mu = X\beta$ by separating signals into a strong group and a weak group. Not only is honesty achieved over the full parameter space $\mathbb{R}^p$, but also our confidence sets can adapt to the sparsity and the strength of β. We also implemented an adaptive way to choose a proper subspace for the projection step among multiple candidate sets, which protects our method from a poor separation between strong and weak signals. Our two-step Stein method showed very satisfactory performance in extensive numeric comparisons, outperforming other competing methods under various parameter settings.

The focus of this work is on the confidence set for $\mu = X\beta$. Although related, it is different from the problem of inference on β. In general, it is difficult to infer a confidence set for β from the confidence set for $X\beta$ without any constraint on X and β, because X does not have a full column rank under the high-dimensional setting. However, if we know that $\|\beta\|_0 \le s$, then a confidence set $\widehat{C}$ for μ can be converted into a confidence set for β as $\widehat{B} := \{\beta \in \mathscr{B}(s) : X\beta \in \widehat{C}\}$, which is the union of s-dimensional subspaces intersecting $\widehat{C}$. It is interesting

future work to study the convergence rate of $|\widehat{B}|$ and related computational issues, such as how to draw β from $\widehat{B}$. On the other hand, if X satisfies SRC(s, c_*, c^*), then

$$c^*\|\beta\|^2 \geq \|X\beta\|^2/n, \quad \forall\ \beta \in \mathscr{B}(s).$$

A hypothesis test about the mean $X\beta$ can be carried out by using the confidence set $\widehat{C}$ to obtain a lower bound on $\|X\beta\|$, which carries over to a lower bound on $\|\beta\|$ with the above inequality and thus can be used to perform a test about β. See Nickl and van de Geer (2013) for a related discussion. A recent work of Cai and Guo (2020) develops methods to construct confidence intervals for $\beta^\mathsf{T}\Sigma\beta$, where Σ is the covariance matrix of the covariates in a random design. For a fixed design, $\beta^\mathsf{T}\Sigma\beta = \|X\beta\|^2/n$ is a function of the mean vector $\mu = X\beta$, and thus it is interesting to explore connections between inferences on μ and on $\beta^\mathsf{T}\Sigma\beta$. Although asymptotic coverage guarantees have been established, our method may have a lower coverage rate than the nominal level for finite samples with an estimate of σ^2, as reported in the numerical results (e.g., Figure 9). Recall that the shrinkage factor B (14) also depends on the estimated σ^2, making our method more sensitive to inaccurate $\hat{\sigma}^2$. Incorporating a robust estimate of the noise variance is an important future direction to improve our method.

We have also demonstrated that our method works well even when the underlying β is dense, for example, $\|\beta\|_0 \asymp n$, which is important for practical applications. See Bradic, Fan, and Zhu (2018) for recent theoretical results on high-dimensional inference for non-sparse β. Besides linear regression models, an abundance of literature has contributed to the construction of confidence sets in functional space (Hoffman and Lepski 2002; Juditsky and Lambert-Lacroix 2003; Genovese and Wasserman 2005; Bull and Nickl 2013). It remains an open and interesting question how to apply the idea of separating strong and weak signals to this problem. Another future direction is to incorporate the confidence set $\widehat{C}$ with the method of estimator augmentation (Zhou 2014; Zhou and Min 2017) for lasso-based inference. Estimator augmentation can be used to simulate from the sampling distribution of the lasso without solving the lasso problem repeatedly, provided a point estimate of $\mu = X\beta$. Given $\widehat{C}$, one may randomize the point estimate of μ by sampling from the confidence set, which has been shown to improve the inferential performance of estimator augmentation (Min and Zhou 2019).

Finally, we briefly comment on predictive inference in high-dimensional linear regression, which is related to this work. As a recent example, Lei et al. (2018) proposed a general framework for distribution-free prediction inference. Assuming that ε follows a zero-mean distribution, their method constructs prediction bands for new responses as well as in-sample prediction intervals with any estimator $\hat{\mu}(x)$ for the mean response value given covariates x. They showed that the constructed prediction bands are close to oracle bands constructed with a known error distribution. In general, in-sample prediction bands can be built upon a confidence set for the mean vector $X\beta$, together with the error distribution. For example, under a normal error assumption, one may construct in-sample prediction bands for all observations $(x_i, y_i), i \in [n]$ as $\{\mu + v : \mu \in \widehat{C}, v \in B(r)\}$, where $\widehat{C}$ is a confidence set for μ and $B(r)$ is a ball with center 0 and radius r. The radius r depends on the error variance σ^2 and the confidence level for the prediction bands. A novel contribution of the method in Lei et al. (2018) is to study the distribution of $\{|y_i - \hat{\mu}(x_i)| : i \in [n]\}$ and make use of order statistics to get rid of restrictive assumptions on the error distribution. It is an interesting future direction to investigate how our method can take advantage of such order statistics to minimize assumptions on the error distribution.

Acknowledgments

The authors thank to the editor, the AE and two referees for their helpful and constructive comments which significantly improved the article.

Funding

This work was partially supported by National Science Foundation under award numbers IIS-1546098, DMS-1952929, and DMS-1513622.

Supplementary Materials

Proofs and additional technical results: Proofs and several auxiliary lemmas are provided in the supplementary materials.

References

Arias-Castro, E., Candès, E. J., and Plan, Y. (2011), "Global Testing Under Sparse Alternatives: ANOVA, Multiple Comparisons and the Higher Criticism," *Annals of Statistics*, 39, 2533–2556. [2]

Baraud, Y. (2004), "Confidence Balls in Gaussian Regression," *Annals of Statistics*, 32, 528–551. [1,7]

Beran, R., and Dümbgen, L. (1998), "Modulation of Estimators and Confidence Sets," *Annals of Statistics*, 26, 1826–1856. [1]

Bühlmann, P., Kalisch, M., and Meier, L. (2014), "High-Dimensional Statistics with a View Toward Applications in Biology," *Annual Review of Statistics and Its Application*, 1, 255–278. [16]

Bickel, P. J., Ritov, Y., and Tsybakov, A. B. (2009), "Simultaneous Analysis of Lasso and Dantzig Selector," *Annals of Statistics*, 37, 1705–1732. [1,2,5,9,11]

Bradic, J., Fan, J., and Zhu, Y. (2018), "Testability of High-Dimensional Linear Models With Non-Sparse Structures," arXiv:1802.09117. [19]

Bull, A. D., and Nickl, R. (2013), "Adaptive Confidence Sets in L^2," *Probability Theory and Related Fields*, 156, 889–919. [19]

Cai, T. T., and Guo, Z. (2017), "Confidence Intervals for High-Dimensional Linear Regression: Minimax Rates and Adaptivity," *Annals of Statistics*, 45, 615–646. [2]

——— (2018), "Accuracy Assessment for High-Dimensional Linear Regression," *Annals of Statistics* , 46, 1807–1836. [2]

——— (2020), "Semisupervised Inference for Explained Variance in High Dimensional Linear Regression and its Applications," *Journal of the Royal Statistical Society*, Series B. [19]

Cai, T. T., and Low, M. G. (2006), "Adaptive Confidence Balls," *Annals of Statistics* , 34, 202–228. [1]

Carpentier, A. (2015), "Implementable Confidence Sets in High Dimensional Regression," in *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics*, Vol. 38 of *Proceedings of Machine Learning Research*,eds. G. Lebanon and S. V. N. Vishwanathan, San Diego, CA: PMLR, pp. 120–128. [2,5,6,9]

Dezeure, R., Bühlmann, P., Meier, L., and Meinshausen, N. (2015), "High-Dimensional Inference: Confidence Intervals, p-Values and R-Software Hdi," *Statistical Science*, 30, 533–558. [11,16]

Dezeure, R., Bühlmann, P., and Zhang, C.-H. (2017), "High-Dimensional Simultaneous Inference With the Bootstrap," *TEST*, 26, 685–719. [2]

Efron, B., and Morris, C. (1973), "Stein's Estimation Rule and its Competitors–An Empirical Bayes Approach," *Journal of the American Statistical Association*, 68, 117–130. [8]

Ewald, K., and Schneider, U. (2018), "Uniformly Valid Confidence Sets Based on the Lasso," *Electronic Journal of Statistics*, 12, 1358–1387. [2]

Genovese, C. R., and Wasserman, L. (2005), "Confidence Sets for Nonparametric Wavelet Regression," *Annals of Statistics* , 33, 698–729. [19]

Hastie, T., Tibshirani, R., and Wainwright, M. (2015), *Statistical Learning with Sparsity: The Lasso and Generalizations*, New York: Chapman & Hall/CRC. [10]

Hoffman, M., and Lepski, O. (2002), "Random Rates in Anisotropic Regression" (with discussion), *Annals of Statistics*, 30, 325–396. [19]

Ingster, Y. I., Tsybakov, A. B., and Verzelen, N. (2010), "Detection Boundary in Sparse Regression," *Electronic Journal of Statistics*, 4, 1476–1526. [2]

Javanmard, A. and Montanari, A. (2014), "Confidence Intervals and Hypothesis Testing for High-Dimensional Regression," *Journal of Machine Learning Research*, 15, 2869–2909. [2]

Juditsky, A., and Lambert-Lacroix, S. (2003), "Nonparametric Confidence Set Estimation," *Mathematical Methods of Statistics*, 12, 410–428. [19]

Lei, J., G'Sell, M., Rinaldo, A., Tibshirani, R. J., and Wasserman, L. (2018), "Distribution-Free Predictive Inference for Regression," *Journal of the American Statistical Association*, 113, 1094–1111. [19]

Li, K.-C. (1985), "From Stein's Unbiased Risk Estimates to the Method of Generalized Cross Validation," *Annals of Statistics*, 13, 1352–1377. [7]

——— (1989), "Honest Confidence Regions for Nonparametric Regression," *Annals of Statistics* , 17, 1001–1008. [1,2,3]

Li, S. (2020), "Debiasing the Debiased Lasso With Bootstrap," *Electronic Journal of Statistics*, 14, 2298–2337. [2]

Min, S., and Zhou, Q. (2019), "Constructing Confidence Sets After Lasso Selection by Randomized Estimator Augmentation," arXiv: 1904.08018. [19]

Negahban, S. N., Ravikumar, P., Wainwright, M. J., and Yu, B. (2012), "A Unified Framework for High-Dimensional Analysis of M-Estimators With Decomposable Regularizers," *Statistical Science*, 27, 538–557. [1,9]

Nickl, R., and van de Geer, S. (2013), "Confidence Sets in Sparse Regression," *Annals of Statistics* , 41, 2852–2876. [2,9,19]

Robins, J., and van der Vaart, A. (2006), "Adaptive Nonparametric Confidence Sets," *Annals of Statistics* , 34, 229–253. [1,2,9]

Schneider, U. (2016), "Confidence Sets Based on Thresholding Estimators in High-Dimensional Gaussian Regression Models," *Econometric Reviews*, 35, 1412–1455. [2]

Stein, C. M. (1981), "Estimation of the Mean of a Multivariate Normal Distribution," *The Annals of Statistics*, 9, 1135–1151. [2]

Sun, T., and Zhang, C.-H. (2012), "Scaled Sparse Linear Regression," *Biometrika*, 99, 879–898. [8,15]

——— (2013), "Sparse Matrix Inversion with Scaled Lasso," *Journal of Machine Learning Research*, 14, 3385–3418. [15]

Tibshirani, R. (1996), "Regression Shrinkage and Selection via the Lasso," *Journal of the Royal Statistical Society*, Series B, 58, 267–288. [1]

van de Geer, S., Bühlmann, P., Ritov, Y., and Dezeure, R. (2014), "On Asymptotically Optimal Confidence Regions and Tests for High-Dimensional Models," *Annals of Statistics*, 42, 1166–1202. [2,16]

Verzelen, N. (2012), "Minimax Risks for Sparse Regressions: Ultra-High Dimensional Phenomenons," *Electronic Journal of Statistics*, 6, 38–90. [2]

Zhang, C.-H. (2010), "Nearly Unbiased Variable Selection Under Minimax Concave Penalty," *Annals of Statistics* , 38, 894–942. [5]

Zhang, C.-H., and Huang, J. (2008), "The Sparsity and Bias of the Lasso Selection in High-Dimensional Linear Regression," *Annals of Statistics*, 36, 1567–1594. [1,4,5]

Zhang, C.-H., and Zhang, S. S. (2014), "Confidence Intervals for Low Dimensional Parameters in High Dimensional Linear Models," *Journal of the Royal Statistical Society*, Series B, 76, 217–242. [2]

Zhang, X., and Cheng, G. (2017), "Simultaneous Inference for High-Dimensional Linear Models," *Journal of the American Statistical Association*, 112, 757–768. [2]

Zhou, Q. (2014), "Monte Carlo Simulation for Lasso-Type Problems by Estimator Augmentation," *Journal of the American Statistical Association*, 109, 1495–1516. [19]

Zhou, Q., and Min, S. (2017), "Estimator Augmentation With Applications in High-Dimensional Group Inference," *Electronic Journal of Statistics*, 11, 3039–3080. [19]