UNIVERSITY OF CALIFORNIA

Los Angeles

Joint and Post-Selection Confidence Sets for High-Dimensional Regression

A dissertation submitted in partial satisfaction

of the requirements for the degree

Doctor of Philosophy in Statistics

by

Kun Zhou

2020

ABSTRACT OF THE DISSERTATION

Joint and Post-Selection Confidence Sets for High-Dimensional Regression

by

Kun Zhou
Doctor of Philosophy in Statistics
University of California, Los Angeles, 2020
Professor Qing Zhou, Chair

Construction of confidence sets is an important topic in statistical inference. In this dissertation, we propose an adaptive method to construct honest confidence sets for the regression mean vector and a framework to construct confidence sets after model selection. The whole dissertation is divided into two parts.

The issue of honesty in constructing confidence sets arises in nonparametric regression. While the 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. After combining the confidence sets for the two groups, our resulting 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, 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.

Apart from the construction of joint confidence sets, the construction of confidence sets after model selection is essentially a different and more challenging problem, as the sampling distributions are restricted to irregular subsets, which increases the difficulty in maintaining the confidence level. To address this problem, we develop a new framework, which contains Bayesian interpretation and constructs credible sets conditioning on active sets of lasso estimates. This framework provides flexible choices of the prior distributions serving as regularizers for the credible sets. Our preliminary research demonstrates that certain credible sets are proved to be confidence sets in the frequentist framework, yet the size of credible sets and the adaption of their diameters should be further studied. Lastly, we seek the possibility to generalize this framework into a large amount of generalized linear models and into confidence sets conditioning on block lasso estimates.

The dissertation of Kun Zhou is approved.

Yingnian Wu

Hongquan Xu

Jingyi Li

Qing Zhou, Committee Chair

University of California, Los Angeles

2020

This dissertation work is specially dedicated to my mother
whose words of encouragement inspire me all the time.
Thank you for bringing me to the world.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I would like to express my sincere gratitude to my supervisor Qing Zhou for his consistent support and guidance throughout my research. I benefit from every meeting, discussion and question and am grateful for his patience, motivation, enthusiasm, and immense knowledge.

VITA

2011–2015 B.S., Department of Mathematics, Zhejiang University

2015–2020 Teaching Assistant and Graduate Student Researcher, Department of Statistics, University of California, Los Angeles

CHAPTER 1

Introduction

Statistical inference is one of the most important fields in the current research. Particularly, high-dimensional inference has been receiving significant attention, since there is a growing demand for methodologies and theories when data is insufficient compared to the number of parameters in models. For example, in biology, especially genetics, researchers want to screen out a small group of genes associated with one specific trait among millions of genes, while the number of subjects is limited. More examples include finance, social networking, online advertising and the list goes on. To broaden the scope of high-dimensional inference, this dissertation aims at providing a novel adaptive method to construct confidence sets for the mean vector in regression models and a new framework of constructing confidence sets after model selection. Before formally presenting our ideas, we review three adaptive methods for regression models, which correspond to three topics on statistical inference: the inference on the mean vector, the inference on the coefficient vector and the inference on the individual coefficient. Besides, we review Stein estimation (Li, 1989), on which our method is based, and a simulation-based method for post-selection inference.

The Stein estimation and the adaptive method for the mean vector are discussed in Section 1.1. The rest of two adaptive methods are discussed in Section 1.2 and Section 1.3, respectively. The method for post-selection inference is discussed in Section 1.4. Lastly, we describe the outline of this dissertation in Section 1.5.

1.1 Inference on the mean vector

Our problem is directly 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; Robins and van der Vaart, 2006; Cai and Low, 2006). Beran and Dümbgen (1998) mentioned that the problem of recovering a signal from observation of the signal plus noise may be formulated as inference of the mean vector, which justifies the practical importance of inference on the mean vector.

Li (1989) provided a fundamental study for this problem. The author considered the nonparametric statistical model

$$y = \mu + \varepsilon,$$

where $y \in \mathbb{R}^n$ is the observed vector, $\mu \in \mathbb{R}^n$ is the unknown mean vector and $\varepsilon \sim \mathcal{N}_n(0, \sigma^2 \mathbf{I}_n)$. Their aim is to construct an asymptotic confidence set $\widehat{C} = \widehat{C}_n$ of small diameter for μ, which achieves honesty in the sense that, for any significance level $\alpha \in (0, 1)$,

$$\liminf_{n \to 0} \inf_{\mu \in \mathbb{R}^n} \mathbb{P}\{\mu \in \widehat{C}\} \geq 1 - \alpha. \tag{1.1}$$

Note that (1.1) means that the confidence set should achieve the nominal significance level for any $\mu \in \mathbb{R}^n$. A naive confidence set can be constructed by $\{\mu \in \mathbb{R}^n : \frac{1}{n}\|\mu - y\|^2 \leq \frac{\sigma^2}{n}\chi^2_{n,\alpha}\}$, where $\chi^2_{n,\alpha}$ denotes the $(1-\alpha)$-quantile of χ^2-distribution with n degrees of freedom. It is easy to verify such a naive confidence set satisfies honesty property. However, the normalized radius $\frac{\sigma^2}{n}\chi^2_{n,\alpha}$ is of the order of 1, indicating the diameter never converges, and thus it is of limited interest. The author further proved that any honest confidence set satisfying (1.1) cannot have a diameter converging at a rate faster than $n^{-1/4}$.

The achievability of this optimal rate $n^{-1/4}$ is demonstrated by a simplified Stein estimate and the associated unbiased risk estimate. 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

$$\hat{\mu}(y;\tilde{\mu}) = y - \frac{\sigma^2 \operatorname{tr}(R_n)}{\|R_n y\|^2} R_n y, \tag{1.2}$$

$$\hat{L}(y;\tilde{\mu}) = 1 - \frac{\sigma^2 \left(\operatorname{tr}(R_n)\right)^2}{n\|R_n y\|^2}, \tag{1.3}$$

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 $c_s(\alpha) > 0$ such that*

$$\liminf_{n\to\infty} \inf_{\mu\in\mathbb{R}^n} \mathbb{P}_\mu \left\{ \left| \sigma^2\hat{L} - n^{-1}\|\hat{\mu} - \mu\|^2 \right| \le c_s(\alpha)\sigma^2 n^{-1/2} \right\} \ge 1 - \alpha, \tag{1.4}$$

where $\hat{\mu}$ and $\hat{L}$ are defined in (1.2) *and* (1.3)*.*

By Lemma 1, we let $\hat{\mu}$ be the center and $\sigma^2\hat{L} + c_s(\alpha)n^{-1/2}$ be the radius to construct a confidence set in the form of $\{\mu : \|\mu - \hat{\mu}\|^2 \le \sigma^2\hat{L} + c_s(\alpha)n^{-1/2}\}$. It follows from (1.4) that such a confidence set satisfies (1.1). The Stein method can achieve the optimal rate of $n^{-1/4}$ if $\tilde{\mu}$ is close to μ in the sense of ℓ_2-norm. That is, y is shrunk to the subspace that μ lies in. Baraud (2004) proposed another method with multiple hypotheses, which increases the chance of adaption at the optimal rate.

Here, we introduce an adaptive method (Robins and van der Vaart, 2006), which constructs an honest confidence set for a Hilbert space-valued parameter including the mean vector in $\mathbb{R}^n$. The authors provided five different models as examples — a model with regular parameters, a finite sequence model, an infinite sequence model, a density estimator and random regression — to illustrate the wide application of their method. Particularly, under a finite sequence model, the observation is a vector following the n-dimensional normal distribution with a mean vector $\theta = \theta^{(n)} = (\theta_1, \theta_2, \ldots, \theta_n)^\mathsf{T}$ and a covariance matrix $\frac{\sigma^2}{n}\mathbf{I}_n$. The variance σ^2 is known and the parameter θ belongs to a subset Θ of $\mathbb{R}^n$, where Θ is possibly equal to $\mathbb{R}^n$. They justified that the confidence set by their method is honest over the parameter space Θ and its diameter adapts to a subset of Θ to achieve the optimal diameter rate $n^{-1/4}$.

Their method is based on sample splitting. Suppose an initial estimator $\hat{\theta} = \hat{\theta}^{(n)}$ independent of y is given. Their confidence set is connected with an estimator $R_n = R_n(\hat{\theta}, y)$ of the squared norm $\|\theta - \hat{\theta}\|^2$ such that

$$\lim_{n\to\infty} \inf_{\theta\in\Theta} \mathbb{P}_\theta \left\{ R_n(\hat{\theta}) - \|\theta - \hat{\theta}\|^2 \geq -z_\alpha \hat{\tau}_{n,\theta} | \hat{\theta} \right\} \geq 1 - \alpha, \tag{1.5}$$

where $\hat{\tau}_{n,\theta}$ is a "scale estimator" and z_α is a quantile. Following from (1.5), one can derive an honest confidence set

$$\widehat{C}_n = \left\{ \theta \in \Theta : \|\theta - \hat{\theta}\| \leq \sqrt{z_\alpha \hat{\tau}_{n,\theta} + R_n(\theta)} \right\}. \tag{1.6}$$

Note that $\widehat{C}$ may not be a ball as the right hand side of (1.6) is also a function of $\hat{\theta}$. Tne inequality (1.5) demonstrates that the confidence set (1.6) essentially depends on a good estimator R_n to achieve the adaptive diameter, while being honesty over Θ. The authors used the construction by Laurent (1996, 1997) for R_n. This construction is based on estimating the squared norm of the projection of $\theta - \hat{\theta}$ by $\|\Pi_k\theta - \Pi_k\hat{\theta}\|^2$, where $\Pi_k\theta = (\theta_1, \ldots, \theta_k, 0, \ldots, 0)$, and minimizing the total effect of the resulting bias and the variance of the estimator. The bias can bounded above by a multiple of

$$B_k^2 := \sup_{\theta\in\Theta} \|\theta - \Pi_k\theta\|^2. \tag{1.7}$$

Note that when $\Theta = \mathbb{R}^n$, $B_k^2 = \infty$ for any $k < n$. The variance is of the order of

$$\hat{\tau}_{k,n,\theta} := \frac{2\sigma^4 k}{n^2} + \frac{4\sigma^2 \|\Pi_k\theta - \Pi_k\hat{\theta}\|^2}{n}. \tag{1.8}$$

Let $|\widehat{C}|$ denote the diameter of the confidence set. Combining (1.7) and (1.8) together, they showed that $\widehat{C}$ is honest over Θ with its diameter

$$|\widehat{C}| = O_p(\frac{\sigma k^{1/4}}{\sqrt{n}} + B_k + \|\theta - \hat{\theta}\|), \tag{1.9}$$

where k can be chosen by an optimal value to minimize the order. For any R_n which is of the same order as $\|\Pi_k\theta - \Pi_k\hat{\theta}\|^2$, the result in (1.9) is still valid. One can see from (1.9) that $\|\theta - \hat{\theta}\|$ is the key to improve the diameter. If $\hat{\theta}$ performs well for a subset of Θ, $k^{1/4}/\sqrt{n}$ and B_k dominate in (1.9). When $\Theta = \mathbb{R}^n$, $|\widehat{C}|$ in (1.9) is reduced to $|\widehat{C}| = O_p(\sigma n^{-1/4} + \|\theta - \hat{\theta}\|)$ verifying $n^{-1/4}$ is the optimal rate for the honest confidence set over $\mathbb{R}^n$.

Further, under the aforementioned finite sequence model, they derived the estimator of $\|\theta - \hat{\theta}\|^2$,

$$R_{k,n}(\theta) = \sum_{i=1}^{k}(X_i - \hat{\theta}_i)^2 - \frac{k\sigma^2}{n}, \tag{1.10}$$

and the associated scale estimator,

$$\hat{\tau}^2_{k,n,\theta} = \frac{2k\sigma^4}{n^2} + \frac{4\sigma^2}{n}\sum_{i=1}^{k}(\theta_i - \hat{\theta}_i)^2. \tag{1.11}$$

With such a choice, z_α in (1.5) is the $(1-\alpha)$-quantile of the standard normal distribution. Later, we will compare our method against this adaptive method by Robins and van der Vaart (2006).

1.2 Inference on the coefficient vector

Consider a linear model

$$y = X\beta + \varepsilon, \tag{1.12}$$

where $y \in \mathbb{R}^n$, $X \in \mathbb{R}^{n\times p}$, $\beta \in \mathbb{R}^p$ and $\varepsilon \sim \mathcal{N}(0, \sigma^2\mathbf{I}_n)$. The inference on β is of great interest for various purposes: model selection, point estimation, hypothesis testing, etc. Carpentier (2015) proposed a method to construct a confidence set for β, which adapts to the unknown sparsity, under the assumption of a separation between large and small coefficients. Recently, 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 for any significance level $\alpha \in (0, 1)$, and vice versa; however, low dimension ($p < n$) is a vital condition in their proof, making it impossible to generalize their idea to the high-dimensional problem that we are studying. Besides, Cai and Guo (2020) considered semi-supervised inference for explained variance, i.e., $\beta^\mathsf{T}\Sigma\beta$, where Σ is the covariance matrix of a random design matrix, and then applied it to construct a confidence set for β. We present the important finding in Nickl and van de Geer (2013) in the rest of this section. In Section 1.1, we have seen that as the space of $\mu = X\beta$ is $\mathbb{R}^n$, the confidence set for μ cannot

maintain the honesty with $|\widehat{C}| = o(n^{-1/4})$. Nickl and van de Geer (2013) showed that this conclusion is also valid to the confidence set for β and the adaption to sparsity at a faster rate $o(n^{-1/4})$ can only happen when a small region is removed from the parameter space of β.

Denote by $B_0(k) = \{\beta \in \mathbb{R}^p : \|\beta\|_0 \leq k\}$ the subset of $\mathbb{R}^p$, where the number of nonzero coefficients is no more than k. The authors studied a sparse linear model (1.12) in the sense that $\beta \in B_0(k_1)$ and $k_1 = k_1(a_1) \sim p^{1-a_1}$ for $0 < a_1 < 1$. Here, $a_1 \in (0, \frac{1}{2}]$ is considered as a moderately sparse case and $a_1 \in (\frac{1}{2}, 1)$ is considered as a highly sparse case. Under proper conditions and $k_1 = o(n/\log p)$, they proved that there exists a confidence set $\widehat{C}$ that is honesty over $B_0(k_1)$ in the sense of

$$\liminf_{n\to\infty} \inf_{\beta \in B_0(k_1)} \mathbb{P}_\beta \left\{\beta \in \widehat{C}\right\} \geq 1 - \alpha$$

with any given significance level. Moreover, for any $k \leq k_1$ and $\beta \in B_0(k)$, its diameter satisfies

$$|\widehat{C}|^2 = O_p(\frac{k}{n} \log p + n^{-1/2}). \tag{1.13}$$

In high-dimensional inference, the equation (1.13) indicates that the construction of confidence sets is essentially a different problem from the estimation of risk bound, where a sparse adaptive estimator $\hat{\beta} = \hat{\beta}(X, y)$ can satisfy

$$\|\hat{\beta} - \beta\|^2 \lesssim \frac{k}{n} \log p$$

up to a multiplicative constant in high probability.

Their next interest is what kind of critical region could be removed from $B_0(k_1)$ to encourage adaption to sparsity at any rate $o(n^{-1/4})$. Let $B_0(k_0)$ be a subset of $B_0(k_1)$ with $k_0 \sim p^{1-a_0}$. Clearly, $k_0 < k_1$ and $a_0 > a_1$. Then, remove those $\beta \in B_0(k_1)$ that are close in ℓ_2-distance to $B_0(k_0)$ to get

$$\tilde{B}_0(k_1, \rho) = \{\beta \in B_0(k_1) : \|\beta - B_0(k_0)\| \geq \rho\},$$

where $\rho = \rho_{n,p}$ is a separation sequence and $\|\beta - B\| := \inf_{b\in B} \|\beta - b\|$ for any $B \subseteq \mathbb{R}^p$. The new model is studied over $\beta \in B(\rho) := B_0(k_0) \cup \tilde{B}_0(k_1, \rho)$. The goal is to find a confidence

set $\widehat{C}_{n,p}$ for β satisfying a weaker honesty property

$$\liminf_{n\to\infty} \inf_{\beta\in B(\rho_{n,p})} \mathbb{P}_\beta \left\{ \beta \in \widehat{C}_{n,p} \right\} \geq 1-\alpha \tag{1.14}$$

for any $\alpha \in (0,1)$. Besides, its diameter satisfies that, for some constant $L > 0$ and $\alpha' \in (0,1)$,

$$\limsup_{n\to 0} \sup_{\beta\in B_0(k_0)} \mathbb{P}_\beta \left\{ |\widehat{C}_{n,p}|^2 > L\frac{k_0}{n}\log p \right\} \leq \alpha' \tag{1.15}$$

and

$$\limsup_{n\to 0} \sup_{\beta\in \tilde{B}_0(k_1,\rho_{n,p})} \mathbb{P}_\beta \left\{ |\widehat{C}_{n,p}|^2 > L\frac{k_1}{n}\log p \right\} \leq \alpha'. \tag{1.16}$$

Assuming certain conditions as well as $k_0 = o(\sqrt{n}/\log p)$ and $k_1 = o(n/\log p)$, they proved that a confidence set satisfying (1.14), (1.15) and (1.16) exists if and only if $\rho_{n,p} \gtrsim n^{-1/4}$, where $\gtrsim$ denotes greater than up to a multiplicative constant. In a moderately sparse case that $0 < a_1 \leq 1/2 < a_0 \leq 1$, $\rho_{n,p}$ attains the rate of $n^{-1/4}$. On the other hand, in a highly spare case that $\frac{1}{2} < a_1 < a_0 \leq 1$, the rate of $\rho_{n,p}$ can be potentially relaxed from $n^{-1/4}$ to $\min\{\frac{k_1}{n}\log p, n^{-1/4}\}$. This conclusion affirms that (1.13) cannot be improved if the confidence sets are honest over all $B_0(k_1)$. One may question if the ρ-separation condition can be avoided at the cost of a mild penalty added to the adaption rate in (1.15). The authors remarked that such a penalty does not alter the necessity of $\rho_{n,p} \gtrsim n^{-1/4}$ by their proof.

1.3 Inference on the individual coefficient

Under the linear model (1.12), the construction of confidence sets for $\mu = X\beta$ or β 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 (Zhang and Zhang, 2014; van de Geer et al., 2014; Javanmard and Montanari, 2014). Although simultaneous inference methods have been proposed based on bootstrapping de-biased lasso estimates (Zhang and Cheng,

2017; Dezeure et al., 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. We introduce another adaptive method proposed by Cai and Guo (2017), which is based on the convergence rates of the minimax expected length for confidence intervals.

Specifically, the authors considered constructing a confidence interval for a linear functional $T(\beta) = \xi^{\mathsf{T}}\beta$, where $\xi \in \mathbb{R}^p$ is named as the loading vector. Let k denote the sparsity level, i.e., $\|\beta\|_0 \leq k$. Based on the sparsity of ξ, they define a sparse loading regime where only a part of ξ_i is nonzero, say $\|\xi\|_0 \lesssim k$, and a dense loading regime where $|\xi| \gg k$. A typical example for sparse loading regime is $T(\beta) = \beta_i$ and a typical example for dense loading regime is $T(\beta) = \sum_{i=1}^p \beta_i$. We first introduce their "minimax expected length" framework. They assume the Gaussian design that all rows of $X_{i\cdot} \overset{i.i.d.}{\sim} \mathcal{N}_p(0, \Sigma)$, and Σ and σ are both unknown. Denote as $\theta = (\beta, \Sigma^{-1}, \sigma)$ the tuple of all parameters. Given the significance level $\alpha \in (0, 1)$, a parameter space $\Theta \subseteq \mathbb{R}^p$ of β and a linear functional $T(\beta)$, let $\mathcal{I}_\alpha(\Theta, T)$ be the set of all honesty confidence intervals for $T(\beta)$ over Θ, namely,

$$\mathcal{I}_\alpha(\Theta, T) = \left\{ CI_\alpha(T, Z) = [l(Z), u(Z)] : \inf_{\theta \in \Theta} \mathbb{P}_\theta\{l(Z) \leq T(\beta) \leq u(Z)\} \geq 1 - \alpha \right\},$$

where $Z = (X, y)$ is the observed data, $l(Z)$ is the lower bound and $u(Z)$ is the upper bound. For any honesty confidence interval $CI_\alpha(T, Z) \in \mathcal{I}_\alpha(\Theta, T)$, define the maximum expected length over a parameter space Θ as

$$L(CI_\alpha(T, Z), \Theta, T) = \sup_{\theta \in \Theta} \mathbb{E}_\theta L(CI_\alpha(T, Z)),$$

where $L(CI_\alpha(T, Z)) = u(Z) - l(Z)$ is the length of that confidence interval. Given two parameter spaces $\Theta_1 \subseteq \Theta$, define

$$L^*_\alpha(\Theta_1, \Theta, T) := \inf_{CI_\alpha(T,Z) \in \mathcal{I}_\alpha(\Theta, T)} L(CI_\alpha(T, Z), \Theta_1, T).$$

Essentially, $\mathcal{I}_\alpha(\Theta, T)$ is the infimum of the maximum length of confidence intervals over the subspace Θ_1, when these confidence intervals are honest over Θ with α significance level. If $\Theta_1 = \Theta$, $L^*_\alpha(\Theta, T) = L^*_\alpha(\Theta, \Theta, T)$ is exactly the minimax expected length of honest confidence intervals over Θ. A confidence interval $CI_\alpha(T, Z)$, which is honest over Θ and adapts to Θ_1,

should satisfy

$$L(CI_\alpha(T,Z),\Theta_1,T) \asymp L^*_\alpha(\Theta_1,T), \qquad L(CI_\alpha(T,Z),\Theta,T) \asymp L^*_\alpha(\Theta,T).$$

In other words, the length of this $CI_\alpha(T,Z)$ should be of the order of the optimal length simultaneously over Θ_1 and Θ, while $CI_\alpha(T,Z)$ maintains the honesty over Θ. As a consequence, if $L^*_\alpha(\Theta_1,\Theta,T) \gg L^*_\alpha(\Theta_1,T)$, the sparse adaption to Θ_1 from Θ is unfeasible.

Let k and k_1 respectively denote the sparsity of Θ and Θ_1. For the sparse loading regime, the authors proved that with the condition $k_1 < k \leq \min\{p^\gamma, \frac{n}{\log p}\}$ for $\gamma \in (0,1)$,

$$L^*_\alpha(\Theta,T) \asymp \|\xi\|_2 \left(\frac{1}{\sqrt{n}} + k\frac{\log p}{n}\right), \tag{1.17}$$

$$L^*_\alpha(\Theta_1,\Theta,T) \geq c_1\|\xi\|_2 \left(\frac{1}{\sqrt{n}} + k\frac{\log p}{n}\right)\sigma, \tag{1.18}$$

where $c_1 > 0$ is a constant. It can bee seen from (1.17) that the length of the honest confidence interval cannot adapt at $o(n^{-1/2})$. The inequality (1.18) indicates that when $\frac{\sqrt{n}}{\log p} \ll k \lesssim \frac{n}{\log p}$ and $k_1 \ll k$, there does not exist an honest confidence interval that adapts to Θ_1, since

$$L^*_\alpha(\Theta_1,\Theta,T) \asymp L^*_\alpha(\Theta,T) \asymp \|\xi\|_2 k\frac{\log p}{n} \gg L^*_\alpha(\Theta_1,T).$$

Therefore, the adaption can only be achieved in the ultra-spare case $k \lesssim \frac{\sqrt{n}}{\log p}$, while the optimal rate $n^{-1/2}$ does not depend on sparsity in this case. For the dense loading regime, they proved that

$$L^*_\alpha(\Theta,T) \asymp \|\xi\|_\infty k\sqrt{\frac{\log p}{n}}, \tag{1.19}$$

$$L^*_\alpha(\Theta_1,\Theta,T) \geq c_1\|\xi\|_\infty k\sqrt{\frac{\log p}{n}}\sigma, \tag{1.20}$$

where $c_1 > 0$ is a constant. It follows from (1.19) and (1.20) that

$$L^*_\alpha(\Theta_1,\Theta,T) \gtrsim \|\xi\|_\infty k\sqrt{\frac{\log p}{n}} \gg L^*_\alpha(\Theta_1,T),$$

which means that the the adaption in the dense loading regime is impossible. Lastly, they studied whether the prior knowledge $\Sigma = \mathbf{I}_n$ and $\sigma = \sigma_0$ can improve the result. For the

sparse loading regime, the minimax expected length is improved and becomes

$$L_\alpha^*(\Theta, T) \asymp \frac{\|\xi\|_2}{\sqrt{n}}$$

so that an adaptive honest confidence interval for $T(\beta)$ is possible over Θ with $k = O(\frac{n}{\log p})$, while for the dense loading regime, the prior knowledge cannot improve the minimax expected length.

1.4 Post-selection inference

Post-selection inference is a topic different from the previous topics and attracts increasing interest in recent years. (Berk et al., 2013; Lee et al., 2016; Tibshirani et al., 2016; Tian and Taylor, 2017; Taylor and Tibshirani, 2018; Liu et al., 2018).

Min and Zhou (2019) developed a novel method to construct a confidence set after lasso variable selection. Their method takes advantages of the closed-form sampling density conditioning on a lasso active set (Zhou, 2014) and a randomization step. Together with a carefully developed Markov chian Monte Carlo (MCMC) algorithm, they empirically showed that the confidence set constructed by their method can achieve the desired coverage rate with its diameter substantially smaller than other state-of-the-art methods. We summarize their work in the rest of the section. Consider (1.12) with a fixed design matrix X. The parameter of interest is

$$\nu := X_A^+ \mu_0 = \underset{\beta \in \mathbb{R}^{|A|}}{\operatorname{argmin}} \|\mu_0 - X_A \beta\|^2, \tag{1.21}$$

where $\mu_0 = X\beta$, namely, the projection of μ_0 onto the column space spanned by X_A. Inference on ν becomes more challenging when the selection of A is also based on the same data set (X, y), because the distribution of ν essentially conditions on the model selection event. In their method, the selection step is achieved by the lasso estimator, namely, $\mathcal{A} = \operatorname{supp}(\hat{\beta})$.

The primary goal is to construct a marginal confidence interval for ν_j conditioning on $\mathcal{A} = A$ with a desired coverage. One possible direction is to construct a confidence interval from $[\{X_A^+(y^* - \tilde{\mu})\}_j | \mathcal{A}(y^*) = A]$ where $\tilde{\mu}$ is an estimate of μ and y^* denotes a sample drawn

from an estimated distribution of y (e.g., $y^* \sim \mathcal{N}_n(\tilde{\mu}, \sigma^2\mathbf{I}_n)$). However, it is unrecommended due to the lack of theories to support that $[X_A^+(y^* - \tilde{\mu})|\mathcal{A}(y^*) = A]$ is a consistent estimator for $[X_A^+(y - \mu_0)|\mathcal{A}(y) = A]$. Therefore, the coverage rates of this raw method could drop with a poor choice of $\tilde{\mu}$. We can theoretically develop a confidence interval for ν_j without knowing the true value of μ_0. Let a set $C \subseteq \mathbb{R}^n$ satisfy $\mu_0 \in C$ and $q_{j,\alpha}(\mu)$ be the α-quantile of $[\{X_A^+(y^* - \mu)\}_j|\mathcal{A}(y^*) = A]$ where $y^* \sim \mathcal{N}_n(\mu, \sigma^2\mathbf{I}_n)$. For any $\alpha < 1/2$, define

$$q^*_{j,\alpha}(C) = \min_{\mu \in C} q_{j,\alpha}(\mu), \qquad\qquad q^*_{j,1-\alpha}(C) = \max_{\mu \in C} q_{j,1-\alpha}(\mu).$$

If $\widehat{C}$ is a $(1 - \alpha/2)$ confidence set for μ_0, then

$$\mathbb{P}\left\{\nu_j \in [\hat{\nu}_j - q^*_{j,\alpha/4}(\widehat{C}), \hat{\nu}_j - q^*_{j,1-\alpha/4}(\widehat{C})]\middle|\mathcal{A}(y) = A\right\} \geq 1 - \alpha, \tag{1.22}$$

where $\hat{\nu}_j = [X_A^+ y]_j$ is the center of this confidence interval. Determining its length by the worst scenarios over all $\mu \in \widehat{C}$, the confidence interval $[\hat{\nu}_j - q^*_{j,\alpha/4}(\widehat{C}), \hat{\nu}_j - q^*_{j,1-\alpha/4}(\widehat{C})]$ maintains the significance level. Inspired by this conservative method, they proposed a three-step simulation-based algorithm to trade off incorporating more variation than single estimate $\tilde{\mu}$ versus controlling the interval length:

1. Draw $\tilde{u}^{(k)}$ uniformly from a $(1 - \alpha)$ confidence set $\widehat{C}$ for μ_0. Denote the uniform distribution over $\widehat{C}$ as $\mathcal{U}(\widehat{C})$. When n is large, one can sample from the boundary of $\widehat{C}$, $\mathcal{U}(\partial\widehat{C})$, as most points in $\widehat{C}$ are close to the boundary.

2. For each $\tilde{u}^{(k)}$, draw $\{y^*_{k,i}\}_i$ from $[y^*|\mathcal{A}(y^*) = A]$, the density of which is derived by estimator augmentation (Zhou, 2014) with a point estimate $\tilde{u}^{(k)}$ in place of the true μ_0.

3. Construct a confidence interval for ν_j based on the quantiles of $\{X_A^+(y^*_{k,i} - \hat{\mu})_j\}$, where $\hat{\mu}$ is some estimate of μ_0.

The randomization of the plug-in $\tilde{\mu}^{(k)}$ has a Bayesian interpretation. Regard $\mathcal{U}(\widehat{C})$ as a posterior distribution for μ_0 and let the density be $p(\mu_0|y)$. Then samples are drawn from the posterior distribution of y^* conditioning on A, i.e.,

$$p(y^*|\mathcal{A}(y^*) = A, y) = \int p(y^*|\mathcal{A}(y^*) = A, \mu_0)p(\mu_0|y)d\mu_0.$$

Further, the algorithm can be generalized to learn the joint distribution

$$\left[\|H(X_A^+ y^* - X_A^+ \hat{\mu})\|_\delta \middle| \mathcal{A}(y^*) = A\right], \tag{1.23}$$

where $H \in \mathbb{R}^{m\times|A|}$, $m \leq |A|$ and $\|.\|_\delta$ is the ℓ_δ-norm, and then a $(1-\alpha)$ confidence set for $H\nu$ can be

$$\{\eta \in \mathbb{R}^m : \|\eta - H\hat{\nu}\|_\delta \leq q\}$$

where q is the $(1-\alpha)$-quantile of the distribution in (1.23). One typical choice is to let $H = \mathbf{I}_{m\times m}$ and $\delta = 2$ to do the inference on ν in (1.21).

1.5 Outline and overview

In this dissertation, we propose an adaptive method to construct a confidence set and a new framework for post-selection inference. Remaining chapters of this dissertation are structured as follows:

- Chapter 2 develops our two-step Stein method in details, including its theoretical properties and algorithmic implementation.
- Chapter 3 presents three alternative methods to construct confidence sets. A various amount of simulations and real-data analysis are conducted to illustrate the effectiveness of our two-step Stein method.
- Chapter 4 establishes a new framework for constructing confidence sets after model selection and includes a preliminary study of its theoretical properties.
- Chapter 5 concludes this dissertation with further discussions and future work.

Notations used throughout the dissertation are defined here. 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 $A \subseteq [p]$. 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. Similarly, define $M_{BA} = (M_{ij})_{i\in B, j\in A}$ and $M_{B\cdot} = (M_{ij})_{i\in B}$. 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 $a, b \in \mathbb{R}$.

CHAPTER 2

A Two-Step Stein Method

Consider high-dimensional linear regression

$$y = X\beta + \varepsilon, \tag{2.1}$$

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 et al. (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 $1-\alpha$,

$$\liminf_{n\to\infty} \inf_{\beta\in\mathbb{R}^p} \mathbb{P}_\beta \left\{ X\beta \in \widehat{C} \right\} \geq 1-\alpha, \tag{2.2}$$

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 the sparsity and the strength of β. In practical applications, sparsity assumptions are very hard to verify, and for many data sets they are at most a good approximation. The first property guarantees that our confidence sets reach the nominal coverage without imposing any sparsity assumption, while the second property allows us to leverage sparse estimation when β is indeed sparse.

Throughout the chapter, we always assume model (2.1) with $\varepsilon \sim \mathcal{N}_n(0, \sigma^2\mathbf{I}_n)$ unless otherwise noted. The remainder of this chapter is organized as follows. We introduce more detailed background, demonstrate our motivation and formulate the problem in Section 2.1. Section 2.2 develops our two-step Stein method. Section 2.3 studies the size of the confidence sets by our method. We provide a data-driven selection of the candidate set in Section 2.4 and develop the implementable algorithm in Section 2.5. Section 2.6 considers theoretical

properties when estimated error variance is plugged in. Finally, all proofs are included in Section 2.7.

2.1 Introduction

Despite notable advances of many developed methods in Section 1.1, lack of numerical support casts doubt on the merit of borrowing these nonparametric regression methods directly for sparse regression. Taking the adaptive method based on sample splitting in Robins and van der Vaart (2006) in Section 1.1 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 (2.1) for $p > n$ is a sparse estimator, such as the lasso (Tibshirani, 1996) or ℓ_0-penalized least-squares estimator. With high probability, the prediction loss of the lasso estimator typically satisfies

$$\frac{1}{n}\|X\hat{\beta} - X\beta\|^2 \leq c\frac{s\log p}{n} \tag{2.3}$$

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 et al. (2009). Under this choice, the diameter $|\widehat{C}_a|$ is of the order

$$|\widehat{C}_a| = O_p\left(n^{-1/4} + \sqrt{s\log p/n}\right) \tag{2.4}$$

for all $\beta \in \mathscr{B}(s)$. For a precise statement, see Theorem 8 below. 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 side of (2.4) approaches ∞ as $p > n \to \infty$ and thus offers little insight into the performance of the confidence set. The upper bound (2.3) 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) in Section 1.2 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 any $s \leq k_1$, the diameter is on the same order as that in (2.4) for any $\beta \in \mathscr{B}(s)$. Compared to the unrestricted honesty in (2.2) 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 $p > n$.

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 (2.4) 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 (2.1) with $p \geq n$, the diameter of an honest confidence set for μ, in the sense of (2.2), cannot adapt at any rate $o(n^{-1/4})$. This is in sharp contrast to error bounds for a sparse estimator, such as that in (2.3), 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 (2.3) 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 (2.1). 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 towards 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 signal strength and sparsity. 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 (2.4), 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, which shrinks to the optimal rate $n^{-1/4}$ when the signal sparsity $s = O(\sqrt{n})$, as opposed to $s = O(\sqrt{n}/\log p)$ in (2.4). Third, in addition to proving the optimal rates like many existing works, we made a lot of efforts in approximating all involved constants in our method, making it practical in real data analysis. 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. We demonstrate with extensive numerical results that our method can construct much smaller confidence sets than other competing methods, including the adaptive method (Robins and van der Vaart, 2006) discussed above and oracle approaches making use of the *true* sparsity of β (the oracle). These results highlight the practical usefulness of our method.

2.2 Method of construction

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 (2.2). 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.

Now, consider the linear model (2.1) and let $\mu = X\beta$. Given a pre-constructed 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 data set 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

$$\widehat{C} = \left\{ \mu \in \mathbb{R}^n : \frac{\|P_A \mu - \hat{\mu}_A\|^2}{n r_A^2} + \frac{\|P_A^\perp \mu - \hat{\mu}_\perp\|^2}{n r_\perp^2} \le 1 \right\}. \tag{2.5}$$

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 the major and minor axes, respectively. 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

$$\|\hat{\mu}_A - \mu_A\|^2 = \|P_A \varepsilon\|^2 \mid A \sim \sigma^2 \chi_k^2. \tag{2.6}$$

Thus, we choose

$$r_A^2 = c_1 \tilde{r}_A^2 = c_1 \sigma^2 \chi_{k,\alpha/2}^2 / n, \tag{2.7}$$

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

$$\mathbb{P}\left\{ \frac{\|P_A \mu - \hat{\mu}_A\|^2}{n r_A^2} \le 1/c_1 \right\} = 1 - \alpha/2. \tag{2.8}$$

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 $y_\perp$ towards 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 (1.2) and (1.3), we obtain

$$\hat{\mu}_\perp = \hat{\mu}(y_\perp; 0) = (1 - B) y_\perp, \tag{2.9}$$

$$\hat{L} = \hat{L}(y_\perp; 0) = (1 - B), \tag{2.10}$$

where the shrinkage factor

$$B = (n-k)\sigma^2 / \|y_\perp\|^2. \tag{2.11}$$

It then follows from Lemma 1 that

$$\liminf_{(n-k)\to\infty} \inf_{\beta \in \mathbb{R}^p} \mathbb{P}\left\{ \left| \sigma^2 \hat{L} - (n-k)^{-1} \|\hat{\mu}_\perp - \mu_\perp\|^2 \right| \le c_s(\alpha) \sigma^2 (n-k)^{-1/2} \right\} \ge 1 - \alpha, \tag{2.12}$$

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

$$r_\perp^2 = c_2 \tilde{r}_\perp^2 = c_2 \frac{n-k}{n} \sigma^2 \left\{ \hat{L} + c_s(\alpha/2)(n-k)^{-1/2} \right\}, \tag{2.13}$$

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

$$\liminf_{(n-k)\to\infty} \inf_{\beta \in \mathbb{R}^p} \mathbb{P} \left\{ \frac{\|\mu_\perp - \hat{\mu}_\perp\|^2}{n r_\perp^2} \le 1/c_2 \right\} \ge 1 - \alpha/2. \tag{2.14}$$

In practical implementation, we estimate the constant $c_s(\alpha)$ in (2.12) by simulation, which will be discussed in Section 2.5.

If $1/c_1 + 1/c_2 = 1$, the confidence set (2.5) made up from (2.8) and (2.14) 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 $\widehat{C}$ (2.5) constructed by the two-step Stein method is honest in the sense of (2.2). Furthermore, the squared diameter of $\widehat{C}$ has expectation*

$$\mathbb{E}|\widehat{C}|^2 = 4\sigma^2 \max \left\{ c_1 \frac{\chi^2_{k,\alpha/2}}{n}, c_2 \frac{n-k}{n} \left(1 - \mathbb{E} \frac{n-k}{\chi^2_{n-k}(\rho)} + c_s(\alpha/2)(n-k)^{-1/2} \right) \right\}, \tag{2.15}$$

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 (2.2) 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

$$\mathbb{E}|\widehat{C}|^2 \le 4\sigma^2 \max \left\{ c_1 \frac{\chi^2_{k,\alpha/2}}{n}, c_2 \frac{n-k}{n} \left(1 + c_s(\alpha/2)(n-k)^{-1/2} \right) \right\} \tag{2.16}$$

holds for all $\beta \in \mathbb{R}^p$. In particular, when β is dense, the diameter will be comparable to that of the naive χ^2 region. As corroborated with the numerical results in Section 3.2.4, this protects our method from inferior performance when sparsity assumptions are violated, making it robust to different data sets. Next, we will show that our confidence set is adaptive: When β is indeed sparse with separable strong and weak signals, 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 set $c_1 = c_2 = 2$ in this section 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 $n \to \infty$. Lemma 2 specifies conditions for the diameter of $\widehat{C}$ to converge at the optimal rate $n^{-1/4}$.

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

$$|\widehat{C}| = O_p\left\{(n^{-1/4} + \|\beta_{A^c}\|) \vee \sqrt{k/n}\right\} \tag{2.17}$$

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 model selection method, while all other signals are weak such that $\|\beta_{A^c}\| = O(n^{-1/4})$. Then we

can 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 (2.4) which requires $s \log p = O(\sqrt{n})$, we have potentially relaxed the sparsity assumption on β to attain the optimal rate $n^{-1/4}$ by imposing a mild condition on the decay rate of the weak signals $\|\beta_{A^c}\|$. In the worst case, if all signals are weak signals, which are of the order of $\sqrt{\log p/n}$, the rate of $|\widehat{C}|$ in (2.17) is reduced to (2.4), the same rate derived by Robins and van der Vaart (2006).

Now we discuss a few methods to find A so that our confidence set can adapt to the sparsity and signal strength of β. We split the whole data set 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'):

$$\hat{\beta} = \hat{\beta}(y', X'; \lambda) := \operatorname*{argmin}_{\beta \in \mathbb{R}^p} \left[\frac{1}{2n} \|y' - X'\beta\|^2 + \lambda \|\beta\|_1 \right], \tag{2.18}$$

where λ is a tuning parameter. Then choose

$$A = \{j : \hat{\beta}_j \neq 0\}, \tag{2.19}$$

that is, we define strong signals by the support of the lasso. This choice of A is justified by the following corollary. 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.

Corollary 3. *Suppose that X and X' satisfy $\mathrm{SRC}(s^*, c_*, c^*)$, where $0 < c_* < c^*$ are constants. Let the confidence set $\widehat{C}$ (2.5) be constructed by the two-step Stein method with A chosen by (2.19) 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*

$$|\widehat{C}| = O_p\left(n^{-1/4} + \sqrt{s \log p/n}\right). \tag{2.20}$$

If in addition $\|\beta_{A_0 \setminus S_0}\| = O(n^{-1/4})$, then

$$|\widehat{C}| = O_p\left(n^{-1/4} \vee \sqrt{s/n}\right). \tag{2.21}$$

The rate of $|\widehat{C}|$ in (2.20) does not depend on any assumption on signal strength, and it is identical to (2.4). However, our method can achieve a faster rate (2.21) 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):

$$\rho(t; \lambda, \gamma) = \int_0^{|t|} \left(1 - \frac{u}{\gamma\lambda}\right)_+ du = \begin{cases} |t| - t^2/(2\gamma\lambda) & \text{if } |t| \le \gamma\lambda \\ \gamma\lambda/2 & \text{if } |t| > \gamma\lambda \end{cases}, \tag{2.22}$$

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

$$\hat{\beta}^{\mathrm{mcp}}_{\lambda,\gamma} = \hat{\beta}^{\mathrm{mcp}}_{\lambda,\gamma}(y', X') := \operatorname*{argmin}_{\beta \in \mathbb{R}^p} \left[\frac{1}{2n} \|y' - X'\beta\|^2 + \lambda \sum_{j=1}^{p} \rho(|\beta_j|; \lambda, \gamma) \right]. \tag{2.23}$$

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 (2.21) under the same SRC assumption but a weaker beta-min condition than Corollary 3.

Corollary 4. *Suppose that X and X' satisfy* $\mathrm{SRC}(s^*, c_*, c^*)$*, 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*

$$|\widehat{C}| = O_p\left(n^{-1/4} \vee \sqrt{s/n}\right). \tag{2.24}$$

Remark 2. Compared to (2.4) for confidence sets centering at a sparse estimator, the diameter of our method in (2.21) and (2.24) 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 (2.4). 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 (2.4) is obtained by adaption to sparsity only (cf. Theorem 8). 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 (2.24). 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 data set, 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 data set.

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, i.e., $M_n \to \infty$. In what follows, we show that under mild conditions, (2.8) and (2.14) 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 $k = \mathrm{rank}(X_A)$ 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 probability of $\widehat{C}_m$.

For strong signals, we apply the following concentration inequality to show (2.8) holds uniformly:

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

$$\mathbb{P}\left\{\sqrt{n}\left|1-\frac{1}{n}\chi^2_n\right| \geq \delta\right\} \leq 2\exp\left(-\frac{\delta^2}{4}\right). \tag{2.25}$$

This lemma with a union bound implies

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

Then choosing

$$r_A^2 = c_1\tilde{r}_A^2 = \frac{c_1\sigma^2}{n}\left[k + 2\sqrt{k\log(4M/\alpha)}\right] \tag{2.26}$$

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 (2.14) uniformly over $\mathcal{H}$ via the following result:

Lemma 4. *Suppose all components of ε in (2.1), $\varepsilon_i, i = 1,\ldots,n$, have mean 0, common second, forth 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*

$$\begin{aligned}&\mathbb{P}\left\{\sup_{A\in\mathcal{H}}\sqrt{n-k}\left|\sigma^2\hat{L} - (n-k)^{-1}\|\hat{\mu}_\perp - \mu_\perp\|^2\right| \geq \sigma^2\delta\right\}\\ &\quad\leq \mathbb{P}\left\{\sup_{A\in\mathcal{H}}\sqrt{n-k}\left|\sigma^2 - \frac{1}{n-k}\|P_A^\perp\varepsilon\|^2\right| \geq \sigma^2\frac{\delta}{2}\right\} + D\sum_{A\in\mathcal{H}}\frac{1}{(n-k)^2} + D\frac{M}{\delta^4}.\end{aligned} \tag{2.27}$$

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 (2.27) 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

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

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

$$r_\perp^2 = c_2\tilde{r}_\perp^2 = c_2\frac{n-k}{n}\sigma^2\left\{\hat{L} + c_m(\alpha/2)\frac{M^{1/4}}{\sqrt{n-k}}\right\} \tag{2.28}$$

for any $A \in \mathcal{H}$, so that the upper bound in (2.27) 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 (2.26) and (2.28), 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 $\mathcal{H}$, i.e.,*

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

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

Remark 4. The increment of r_A^2 in (2.26), $2\sqrt{k\log(4M/\alpha)}/n$, 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 the k here is the size of the selected candidate set A_{m^*}. Compared with (2.13), the factor $M^{1/4}/\sqrt{n-k}$ in (2.28), 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 (2.18) calculated from (X', y') to obtain

$$A_m = \{j \in [p] : |\hat{\beta}_j| > \tau_m\}, \tag{2.29}$$

for a sequence of threshold values $\tau_m = a_m\lambda$, e.g. $a_m \in [0, 4]$. 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, i.e., $\{\mu \in \mathbb{R}^n : \|\mu - \hat{\mu}_\perp\|^2 \le n r_\perp^2\}$.

2.5 Algorithm and implementation

We implement our method with a sequence of candidate sets A_m defined by (2.29). Given the data set, σ^2, λ in (2.18) and threshold values $\{a_m\lambda\}_{1\le m\le M}$, this section describes some technique details in our algorithm to construct the confidence set (2.5) by the two-step Stein method.

Data splitting. We split the original data set into (X', y') and (X, y). Apply lasso on (X', y') to get $\hat{\beta}$ in (2.18) with the tuning parameter λ. Threshold $\hat{\beta}$ by $\tau_m = a_m\lambda$ for $m = 1, \ldots, M$ in (2.29) to define candidate sets A_m. Note that A_m, $m = 1, \ldots, M$, are independent of (X, y).

Choice of c_1 and c_2. When $A \neq \varnothing$, we consider two criteria to choose the constants c_1 in (2.7) and c_2 in (2.13). 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

$$\begin{aligned} &\min_{c_1, c_2} \left\{ k \log(\sqrt{c_1}\tilde{r}_A) + (n - k) \log(\sqrt{c_2}\tilde{r}_\perp) \right\}, \\ &\text{subject to } 1/c_1 + 1/c_2 = 1 \text{ and } 1 < c_1, c_2 \le E, \end{aligned} \tag{2.30}$$

where $\tilde{r}_A$ and $\tilde{r}_\perp$ are defined in (2.7) and (2.13) and $E > 2$ is a pre-determined upper bound. It is easy to obtain the solution

$$c_1 = \frac{E}{E-1} \vee \left(\frac{n}{k} \wedge E \right), \qquad c_2 = \frac{E}{E-1} \vee \left(\frac{n}{n-k} \wedge E \right). \tag{2.31}$$

For all numerical results in this dissertation, 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 $|\widehat{C}|$

$$\min_{c_1,c_2} \max\{r_A, r_\perp\}, \text{ subject to } 1/c_1 + 1/c_2 = 1, \tag{2.32}$$

which yields the solution

$$c_1 = (\tilde{r}_A^2 + \tilde{r}_\perp^2)/\tilde{r}_A^2, \qquad c_2 = (\tilde{r}_A^2 + \tilde{r}_\perp^2)/\tilde{r}_\perp^2. \tag{2.33}$$

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. Since r_A and $r_\perp$ are less than $(\tilde{r}_A + \tilde{r}_\perp)$, all theoretical justifications in Section 2.3 hold.

Computation of $c_s(\alpha)$. For any candidate set A, the radius $r_\perp$ (2.13) depends on the constant $c_s(\alpha)$, 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 $c_s(\alpha)$: 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

$$\check{\mu}_j = \left(1 - \frac{n\sigma^2}{\|\check{Y}_j\|^2}\right) \check{Y}_j \quad \text{and} \quad \check{L}_j = \left(1 - \frac{n\sigma^2}{\|\check{Y}_j\|^2}\right)_+. \tag{2.34}$$

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

$$\frac{\sqrt{n}}{\sigma^2} \left| \sigma^2 \check{L}_j - n^{-1}\|\check{\mu}_j\|^2 \right|, \quad j = 1, \ldots, N, \tag{2.35}$$

is a consistent estimator of $c_s(\alpha)$ as long as $\|\mu_\perp\| = o(\sqrt{n})$, which is the case under the assumptions of Corollary 3. Expression (2.35) can be written as a function of a χ^2_n random variable, which simplifies its simulation.

Clearly, the estimate of $c_s(\alpha)$ 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,

i.e., 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 (2.7) and (2.13) for each $A \in \mathcal{H}$.

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

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 (2.9)

compute c_1 and c_2 according to one of the two criteria

compute r_A and $r_\perp$ by (2.7) and (2.13)

construct $\widehat{C}_m$ in the form of (2.5)

end for

find m^* by minimizing the volume or the diameter of $\widehat{C}_m$ over m

Remark 5. In the calculation of $r_\perp$ and $c_s(\alpha)$, we use truncated Stein estimation for $\hat{\mu} = (1 - B)_+ y_\perp$ in (2.9) and $\hat{L} = (1 - B)_+$ in (2.10) as well as for $\check{\mu}_j$ and $\check{L}_j$ in (2.34). 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 (2.7) and (2.13) to construct the confidence set $\widehat{C}$ (2.5). 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 we may assume 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 at 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* (1.2)

and $\hat{L}(y,0)$ *in* (1.3) *with* σ^2 *replaced by* $\hat{\sigma}^2$. *For any* $\alpha \in (0,1)$ *and any sequence* $\hat{\sigma}^2 = \hat{\sigma}_n^2$ *satisfying* $|\hat{\sigma}_n^2 - \sigma^2| \le M_1/\sqrt{n}$ *when* n *is large, there exists a constant* $c_s'(\alpha) > 0$ *(depending on* M_1*) such that*

$$\limsup_{n\to\infty} \sup_{\mu\in\mathbb{R}^n} \mathbb{P}\left\{\left|\hat{\sigma}^2\tilde{L} - n^{-1}\|\tilde{\mu}-\mu\|^2\right| \ge c_s'(\alpha)\hat{\sigma}^2 n^{-1/2}\right\} \le \alpha. \tag{2.36}$$

Theorem 6. *Suppose all assumptions in Theorem 1 hold and in addition that* $k = o(n)$. *Let* $\hat{\sigma}^2 = \hat{\sigma}_n^2$ *be a sequence satisfying* $|\hat{\sigma}_n^2 - \sigma^2| \le M_1/\sqrt{n}$ *when* n *is large. Let* r_A *be computed as in* (2.7) *with* $\hat{\sigma}^2$ *in place of* σ^2 *and* $r_\perp$ *be computed as in* (2.13) *with* $\hat{\sigma}^2$ *and* $c'(\alpha)$ *in place of* σ^2 *and* $c_s(\alpha)$. *Then the confidence set* $\widehat{C}$ (2.5) *is honest.*

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* $\hat{\sigma}^2 = \hat{\sigma}_n^2$ *be a sequence satisfying* $|\hat{\sigma}_n^2 - \sigma^2| \le M_1/\sqrt{n}$ *when* n *is large. Let* r_A *and* $r_\perp$ *be computed with* $\hat{\sigma}^2$ *as in Theorem 6. Then*

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

which are exactly the same rates in Lemma 2.

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 $\hat{\sigma}^2 - \sigma = O(1/\sqrt{n})$. 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 certain conditions. Given the design matrix X', define a compatibility factor (van de Geer and Bühlmann, 2009) as $\kappa(\xi, T)$ with $\xi > 1$ and $T \subseteq [p]$. Suppose the infimum $\kappa^*(\xi) = \inf_{|T|\le s} \kappa(\xi, T) > 0$ exists and $s \log p \ll \sqrt{n}$. Then their Theorem 2 demonstrates that for any s-sparse β, $\hat{\sigma}^2$ estimated by scaled lasso is the $\sqrt{n}$-consistent estimator and the central limit theorem about $\hat{\sigma}$ holds, i.e., $n^{1/2}(\hat{\sigma}/\sigma - 1) \xrightarrow{d} \mathcal{N}(0, 1/2)$. This conclusion

together with Theorem 6 directly justifies that our method maintains the desired honesty and achieves the adaptive radii in probability converging to 1 when the even splitting and scaled lasso are applied. Lastly, 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 6. Note that $c_s(\alpha)$ is invariant to the value of the true σ^2. Even if we plug $\hat{\sigma}^2$ in the simulation of $c_s(\alpha)$ discussed in Section 2.5, we will still estimate the $c_s(\alpha)$ associated with the true σ^2 instead of $c'_s(\alpha)$. However, the empirical study in Section 3.2.5 shows that using so estimated $c_s(\alpha)$ 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 $c'_s(\alpha)$ from $c_s(\alpha)$. In particular, if $\hat{\sigma}^2$ is estimated by scaled lasso, we propose an efficient method to approximate $c'_s(\alpha)$. See the Supplementary Material for more details.

2.7 Proofs

Proof of Lemma 2. By the law of large number, we have

$$\frac{\chi^2_{k,\alpha} - k}{\sqrt{2k}} = o(1) + \Phi^{-1}(\alpha) \Rightarrow \chi^2_{k,\alpha} = k + o(\sqrt{2k}) + \sqrt{2k}\Phi^{-1}(\alpha) \asymp k, \tag{2.37}$$

where Φ^{-1} is the inverse of the cumulative distribution function of $\mathcal{N}(0,1)$. It follows from (2.7) and (2.37) that

$$r_A^2 = c_1 \cdot \sigma^2 \chi^2_{k,\alpha}/n \asymp k/n. \tag{2.38}$$

Let $\varepsilon_\perp = P_A^\perp \varepsilon$. Under the normality assumption of ε, we have

$$\begin{aligned} 1/B &= \frac{\|y_\perp\|^2}{(n-k)\sigma^2} = \frac{\|\varepsilon_\perp\|^2 + 2\langle \mu_\perp, \varepsilon_\perp\rangle + \|\mu_\perp\|^2}{(n-k)\sigma^2} \\ &= 1 + O_p\left(\frac{1}{\sqrt{n-k}}\right) + O_p\left(\frac{\|\mu_\perp\|}{n-k}\right) + \frac{\|\mu_\perp\|^2}{(n-k)\sigma^2}. \end{aligned}$$

It follows, by noting $\|\mu_\perp\| = o(\sqrt{n-k})$, that

$$\hat{L} = 1 - B = O_p\left(\frac{1}{\sqrt{n-k}}\right) + O_p\left(\frac{\|\mu_\perp\|^2}{n-k}\right). \tag{2.39}$$

By plugging (2.39) in (2.13), we obtain

$$r_\perp^2 = c_2 \cdot \sigma^2 \frac{n-k}{n} \left\{ O_p\left(\frac{1}{\sqrt{n-k}}\right) + O_p\left(\frac{\|\mu_\perp\|^2}{n-k}\right) + c_s(\alpha/2)\frac{1}{\sqrt{n-k}} \right\}$$
$$= O_p\left(\frac{\sqrt{n-k}}{n}\right) + O_p\left(\frac{\|\mu_\perp\|^2}{n}\right). \tag{2.40}$$

If $k = O_p(\sqrt{n})$ and $\|\mu_\perp\| = O(n^{1/4})$, it follows from (2.38) and (2.40) that $|\widehat{C}| \asymp_p n^{-1/4}$. □

Proof of Theorem 2. Under sparse Riesz condition, letting $G = A^c \cap \mathrm{supp}(\beta)$, we have

$$\|\mu_\perp\| = \|P_A^\perp X_{A^c}\beta_{A^c}\| = \|P_A^\perp X_G \beta_G\| \le c^*\sqrt{n}\|\beta_G\| = c^*\sqrt{n}\|\beta_{A^c}\|,$$

which, together with $k = o(n)$ and $\|\beta_{A^c}\| = o(1)$, implies $\|\mu_\perp\| = o(\sqrt{n}) = o(\sqrt{n-k})$. Thus, by Lemma 2, $r_\perp^2 = O_p(n^{-1/2} + \|\beta_{A^c}\|^2)$ and the rest of the proof is straightforward. □

Proof of Corollary 3. Under the choice of λ in this corollary and the assumption that $s \le (s^*-1)/(2+4c^*/c_*)$, Theorem 1 and Theorem 3 in Zhang and Huang (2008) imply that, for any $\epsilon > 0$, there exists N such that when $n > N$,

$$\mathbb{P}\left\{ |A| \le M_1^* s \text{ and } \|\hat{\beta} - \beta\| \le M_2^* \sigma \sqrt{(s\log p)/n} \right\} > 1 - \epsilon, \tag{2.41}$$

where M_1^* and M_2^* are two constants depending on c_0, c_* and c^*. It follows from (2.41) that

$$k \le |A| = O_p(s) = o_p(n), \quad \|\hat{\beta} - \beta\| = O_p\left(\sqrt{s\log p/n}\right).$$

Thus, we have

$$\|\beta_{A^c}\| \le \|\hat{\beta} - \beta\| = O_p\left(\sqrt{s\log p/n}\right) = o_p(1). \tag{2.42}$$

Now, all the conditions in Theorem 2 are satisfied, leading to (2.20). Further, (2.42) implies that $S_0 \subset A$ and thus $\|\beta_{A^c}\| = \|\beta_{A^c \cap A_0}\| \le \|\beta_{A_0 \setminus S_0}\| = O(n^{-1/4})$ with probability at least $1-\epsilon$. Consequently, (2.21) follows from (2.17). □

Proof of Corollary 4. If $\mathbb{P}(A = A_0) \to 1$, then the rate of $|\widehat{C}|$ in (2.24) follows immediately from (2.17) in Theorem 2. Thus, it remains to show that $\hat{\beta}^{\mathrm{mcp}}_{\lambda_n,\gamma_n} = \hat{\beta}^{\mathrm{mcp}}_{\lambda,\gamma}(y', X')$ (2.23) is model selection consistent by verifying the conditions of the following corollary, which is a simplified version of Corollary 4.2 in Huang et al. (2012).

Corollary 7. *Let* $\lambda_{\min}$ *be the smallest eigenvalue of* $(X'_{A_0})^{\mathsf{T}} X'_{A_0}/n$, $\tau_n = \sigma\sqrt{2\log s/(n\lambda_{\min})}$ *and* $\lambda^* = 2\sigma\sqrt{2c^*\log(p-s)/n}$. *Suppose that* X' *satisfies* $\mathrm{SRC}(s^*, c_*, c^*)$, *where* $0 < c_* < c^*$ *are constants and* $s^* \geq (c^*/c_* + 1/2)s$. *If a sequence of* (λ_n, γ_n) *satisfies* $\inf_{A_0} |\beta_j| \geq \gamma_n\lambda_n + a_n\tau_n$ *with* $a_n \to \infty$, $\lambda_n \geq a_n\lambda^*$, $n\lambda_n^2/(4c^*) > \sigma^2$ *and* $\gamma_n \geq c_*^{-1}\sqrt{4 + c_*/c^*}$, *then* $\mathbb{P}\{\mathrm{supp}(\hat{\beta}^{\mathrm{mcp}}_{\lambda_n,\gamma_n}) = A_0\} \to 1$.

Under the SRC assumption $\lambda_{\min}$ is bounded from below by $c_* > 0$. It follows from $\tau_n = O(\sqrt{\log s/n})$, $\lambda^* = O(\sqrt{\log p/n})$ and $\lambda_n \gg \sqrt{\log p/n}$ that there exists $a_n \to \infty$ such that $\lambda_n \geq a_n(\lambda^* \vee \tau_n)$. Then we have the following: $\inf_{A_0} |\beta_j| \geq (\gamma_n + 1)\lambda_n \geq \gamma_n\lambda_n + a_n\tau_n$, $\lambda_n \geq a_n\lambda^*$, and $n\lambda_n^2/(4c^*) \gg \log p > \sigma^2$ when n is sufficiently large. Thus all the conditions in Corollary 7 are satisfied under the assumptions of Corollary 4. This completes the proof.

Technically, we did not invoke the assumption $s\log p = o(n)$ in the proof. But it is required for the sparse Riesz condition to hold (e.g. for Gaussian designs). □

Proof of Lemma 3. We have the following inequalities for any positive x and degree of freedom of n from Lemma 1 in Laurent and Massart (2000):

$$\mathbb{P}\left\{\chi_n^2 - n \geq 2\sqrt{n}\sqrt{x} + 2x\right\} \leq e^{-x}, \tag{2.43}$$

$$\mathbb{P}\left\{\chi_n^2 - n \leq -2\sqrt{n}\sqrt{x}\right\} \leq e^{-x}. \tag{2.44}$$

The solutions of $2\sqrt{n}\sqrt{x_1} + 2x_1 = \sqrt{n}\delta$ and $2\sqrt{n}\sqrt{x_2} = \sqrt{n}\delta$ are plugged in (2.43) and (2.44) to obtain

$$\mathbb{P}\left\{\frac{\chi_n^2}{n} - 1 \geq \sqrt{n}\delta\right\} \leq \exp\left\{-\frac{(\sqrt{1 + 2\delta/\sqrt{n}} - 1)^2}{4}n\right\},$$
$$\mathbb{P}\left\{\frac{\chi_n^2}{n} - 1 \leq -\sqrt{n}\delta\right\} \leq \exp\left\{-\frac{\delta^2}{4}\right\},$$

so that

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

To finish the proof, we will show that

$$f(n) = \left(\sqrt{1 + 2\delta/\sqrt{n}} - 1\right)^2 n \tag{2.45}$$

is bounded by δ^2 for any n. Replacing $\sqrt{1+2\delta/\sqrt{n}}$ with its Taylor expansion $1+\delta/\sqrt{n}+O(\delta^2/n)$ in (2.45), we get $f(n)=\delta^2+O(n^{-1/2})\to\delta^2$, as $n\to\infty$. If $f(n)$ is monotonically increasing in n, then δ^2 is a tight upper bound of $f(n)$ for all n. Lastly, to prove the monotonicity, it suffices to show the derivative

$$f'(n)=2+\delta/\sqrt{n}-\frac{2+3\delta/\sqrt{n}}{\sqrt{1+2\delta/\sqrt{n}}}\geq 0,$$

which can be verified easily. Now the proof is completed. □

Proof of Lemma 4. Let

$$\begin{aligned}Q(A)&=\mathbb{E}\|P_A^\perp y\|^2=\mathbb{E}\|P_A^\perp(\mu+\varepsilon)\|^2\\&=\|P_A^\perp\mu\|^2+\mathrm{tr}(P_A^\perp)\sigma^2=\|P_A^\perp\mu\|^2+(n-k)\sigma^2.\end{aligned}$$

A few steps of derivation shows that

$$\begin{aligned}&\sigma^2\hat{L}-(n-k)^{-1}\|\hat{\mu}_\perp-\mu_\perp\|^2\\&\quad=\sigma^2-\frac{\sigma^4(n-k)}{\|P_A^\perp y\|^2}-\frac{1}{n-k}\left\|\left(1-\frac{(n-k)\sigma^2}{\|P_A^\perp y\|^2}\right)P_A^\perp y-P_A^\perp\mu\right\|^2\\&\quad=\sigma^2-\frac{1}{n-k}\|P_A^\perp\varepsilon\|^2+\frac{2\sigma^2}{\|P_A^\perp y\|^2}\left(\langle\varepsilon,P_A^\perp\mu\rangle+\|P_A^\perp\varepsilon\|^2-\sigma^2(n-k)\right).\end{aligned}\tag{2.46}$$

It follows from (2.46) that

$$\begin{aligned}&\mathbb{P}\left\{\sup_{A\in\mathcal{H}}\sqrt{n-k}\left|\sigma^2\hat{L}-(n-k)^{-1}\|\hat{\mu}_\perp-\mu_\perp\|^2\right|\geq\sigma^2\delta\right\}\\&\quad\leq\mathbb{P}\left\{\sup_{A\in\mathcal{H}}\sqrt{n-k}\left|\sigma^2-\frac{1}{n-k}\|P_A^\perp\varepsilon\|^2\right|\geq\sigma^2\delta/2\right\}\\&\qquad+\mathbb{P}\left\{\sup_{A\in\mathcal{H}}\sqrt{n-k}\left|\frac{2\sigma^2}{\|P_A^\perp y\|^2}\left(\langle\varepsilon,P_A^\perp\mu\rangle+\|P_A^\perp\varepsilon\|^2-\sigma^2(n-k)\right)\right|\geq\sigma^2\delta/2\right\},\end{aligned}$$

where the second probability on the right hand side is bounded by

$$\begin{aligned}&\sum_{A\in\mathcal{H}}\mathbb{P}\left\{\left|\frac{2\sigma^2}{\|P_A^\perp y\|^2}\left(\langle\varepsilon,P_A^\perp\mu\rangle+\|P_A^\perp\varepsilon\|^2-\sigma^2(n-k)\right)\right|\geq\frac{\sigma^2\delta}{2\sqrt{n-k}}\right\}\\&\quad\leq\sum_{A\in\mathcal{H}}\left[\mathbb{P}\left\{\|P_A^\perp y\|^2\leq\frac{1}{2}Q(A)\right\}\right.\\&\qquad+\mathbb{P}\left\{2\left|\|P_A^\perp\varepsilon\|^2-(n-k)\sigma^2\right|\geq\frac{\delta Q(A)}{2^3\sqrt{n-k}}\right\}\\&\qquad\left.+\mathbb{P}\left\{2\left|\langle\varepsilon,P_A^\perp\mu\rangle\right|\geq\frac{\delta Q(A)}{2^3\sqrt{n-k}}\right\}\right].\end{aligned}\tag{2.47}$$

To prove the theorem, it suffices to show that all three probabilities in (2.47) can be bounded by either $D/(n-k)^2$ or D/δ^4 for some constant $D > 0$. Before that, we introduce the following three inequalities derived from Theorem 2 in Whittle (1960):

$$\mathbb{E}\left(\|P_A^\perp y\|^2 - Q(A)\right)^4 \leq D_1\left[\sigma^4(n-k)^2 + \|P_A^\perp \mu\|^4\right], \tag{2.48}$$

$$\mathbb{E}\left(\|P_A^\perp \varepsilon\|^2 - (n-k)\sigma^2\right)^4 \leq D_1\sigma^4(n-k)^2, \tag{2.49}$$

$$\mathbb{E}\left(\langle \varepsilon, P_A^\perp \mu\rangle\right)^4 \leq D_1\|P_A^\perp \mu\|^4, \tag{2.50}$$

for some constant D_1 depending on the moments of ε_i. In our case, D_1 only depends on the upper bound d of the eighth moment. The first term of (2.47) can be bounded by

$$\begin{aligned}
&\mathbb{P}\left\{\|P_A^\perp y\|^2 \leq \frac{1}{2}Q(A)\right\} \leq \mathbb{P}\left\{\left|\|P_A^\perp y\|^2 - Q(A)\right| \geq \frac{1}{2}Q(A)\right\} \\
&\leq \frac{\mathbb{E}\left(\|P_A^\perp y\|^2 - Q(A)\right)^4}{\left(\frac{1}{2}Q(A)\right)^4} \qquad \text{by Chebyshev inequality} \\
&\leq 16D_1\frac{\sigma^4(n-k)^2 + \|P_A^\perp \mu\|^4}{Q(A)^4} \qquad \text{by (2.48)} \\
&\leq \frac{16D_1}{(n-k)^2}.
\end{aligned}$$

Similarly, using (2.49) and (2.50), we can also show that both the second and the third terms are bounded by $D_2/(\sigma^2\delta^4)$ for some $D_2 > 0$ depending only on d. Lastly, the proof is finished by letting $D = (16D_1) \vee (D_2/\sigma^2)$. □

Proof of Lemma 5. Let

$$\begin{aligned}
g(\sigma^2) &= \frac{\sqrt{n}}{\sigma^2}\left(\sigma^2\hat{L} - n^{-1}\|\hat{\mu} - \mu\|^2\right) \\
&= \frac{\sqrt{n}}{\sigma^2}\left(\sigma^2 - \frac{1}{n}\|\varepsilon\|^2 + \frac{2\sigma^2}{\|y\|^2}\left(\langle \varepsilon, \mu\rangle + \|\varepsilon\|^2 - \sigma^2 n\right)\right) \\
&= \sqrt{n}\left(1 - \frac{\|\varepsilon\|^2}{n\sigma^2} + \frac{2}{\|y\|^2}\left(\langle \varepsilon, \mu\rangle + \|\varepsilon\|^2 - \sigma^2 n\right)\right).
\end{aligned} \tag{2.51}$$

First, we find a constant M_2 to bound $|g(\hat{\sigma}^2) - g(\sigma^2)|$. To be exact, we show that for any $\xi > 0$, there exists N and M_2 such that for any $n > N$ and $\mu \in \mathbb{R}^p$,

$$\mathbb{P}\left\{|g(\hat{\sigma}^2) - g(\sigma^2)| \leq M_2\right\} \geq 1 - \xi. \tag{2.52}$$

Note that

$$g(\hat{\sigma}^2) - g(\sigma^2) = \frac{\|\varepsilon\|^2}{\sqrt{n}} \left(\frac{1}{\hat{\sigma}^2} - \frac{1}{\sigma^2} \right) + \frac{2n^{3/2}}{\|y\|^2} (\hat{\sigma}^2 - \sigma^2). \tag{2.53}$$

It is easier to separately bound the two terms on the right side of (2.53). The first term

$$\frac{\|\varepsilon\|^2}{\sqrt{n}} \left| \frac{1}{\hat{\sigma}^2} - \frac{1}{\sigma^2} \right| = \frac{\|\varepsilon\|^2}{n} \frac{|\sigma^2 - \hat{\sigma}^2|/\sqrt{n}}{\sigma^2 \hat{\sigma}^2} = O_p(1).$$

Therefore, there exists an upper bound M_3 so that

$$\mathbb{P} \left\{ \frac{\|\varepsilon\|^2}{\sqrt{n}} \left| \frac{1}{\hat{\sigma}^2} - \frac{1}{\sigma^2} \right| \le M_3 \right\} \ge 1 - \xi/2 \tag{2.54}$$

as $n > N_1$ for some large integer N_1. For the second term, we have

$$\begin{aligned} \mathbb{P} \left\{ \frac{2n^{3/2}}{\|y\|^2} |\hat{\sigma}^2 - \sigma^2| \le M_4 \right\} &= 1 - \mathbb{P} \left\{ \frac{2\sqrt{n}|\hat{\sigma}^2 - \sigma^2|}{M_4} \ge \frac{\|y\|^2}{n} \right\} \\ &\ge 1 - \mathbb{P} \left\{ \frac{2\sqrt{n}|\hat{\sigma}^2 - \sigma^2|}{M_4} \ge \frac{\|\varepsilon\|^2}{n} \right\}, \end{aligned} \tag{2.55}$$

for any constant $M_4 > 0$. The inequality (2.55) holds for any μ for the following reason. For any R, $\mathbb{P}\{R^2 \ge \|y\|^2\}$ is the integral of the standard normal density $\phi(z)$ over the ball $B(-\mu, R)$ centering at $-\mu$ with radius R, while the ball is $B(0, R)$ for $\mathbb{P}\{R^2 \ge \|\varepsilon\|^2\}$. For any $\epsilon_1 \in B(-\mu, R) \setminus B(0, R)$ and any $\epsilon_2 \in B(0, R) \setminus B(-\mu, R)$, we always have $\phi(\epsilon_1) \le \phi(\epsilon_2)$. Consequently, the integral over $B(-\mu, R)$ is no greater than that over $B(0, R)$, which implies the last inequality. Since $\|\varepsilon\|^2/(n\sigma^2) = 1 + O_p(\frac{1}{\sqrt{n}})$, we can find a large M_4 so that $\frac{2M_1}{M_4} < 1$. In this way, for any $\xi/2 > 0$, there exists some N_2 so that (2.55) is at least $1 - \xi/2$ for any $\mu \in \mathbb{R}^n$ as $n > N_2$.

Letting $N = \max\{N_1, N_2\}$ and $M_2 = M_3 + M_4$, we have shown that $|g(\hat{\sigma}^2) - g(\sigma^2)| \le M_2$ with probability at least $1 - \xi$ uniformly for all $\mu \in \mathbb{R}^n$ when $n > N$. Now choose $c_s'(\alpha) = c(\alpha - \xi) + M_2$. It follows that

$$\begin{aligned} &\sup_{\mu \in \mathbb{R}^n} \mathbb{P} \left\{ g(\hat{\sigma}^2) > c_s'(\alpha) \right\} \\ &\le \sup_{\mu \in \mathbb{R}^n} \mathbb{P} \left[\left\{ g(\hat{\sigma}^2) > c_s'(\alpha) \right\} \cap \left\{ |g(\hat{\sigma}^2) - g(\sigma^2)| \le M_2 \right\} \right] + \sup_{\mu \in \mathbb{R}^n} \mathbb{P} \left\{ |g(\hat{\sigma}^2) - g(\sigma^2)| > M_2 \right\} \\ &\le \sup_{\mu \in \mathbb{R}^n} \mathbb{P} \left\{ g(\sigma^2) > c_s'(\alpha) - M_2 \right\} + \xi \\ &= \sup_{\mu \in \mathbb{R}^n} \mathbb{P} \left\{ g(\sigma^2) > c_s(\alpha - \xi) \right\} + \xi. \end{aligned} \tag{2.56}$$

Then the conclusion follows from Lemma 1 by taking upper limit on both sides as $n \to \infty$. □

Remark 7. Based on (2.53) in the above proof of Lemma 5, we propose an empirical method to estimate $c'_s(\alpha)$. First consider $\hat{\sigma}^2$. Although Lemma 5 assumes a fixed sequence of $\hat{\sigma}^2$, $\hat{\sigma}^2$ is estimated from a data set (X', y') independent of (X, y), when constructing a confidence set in practice. Without loss of generality, assume the size of (X', y') is n. Sun and Zhang (2012) prove that $\hat{\sigma}^2$ estimated by scaled lasso satisfies the central limit theorem under certain conditions

$$n^{1/2}\left(\frac{\hat{\sigma}}{\sigma} - 1\right) \to \mathcal{N}_1(0, \frac{1}{2}). \tag{2.57}$$

Second, let $g(\sigma^2, y; \mu) = g(\sigma^2)$ based on (2.51). Regarding $g(\hat{\sigma}^2, y; \mu)$ as a random variable, $c'_s(\alpha)$ is approximately the supremum of the $(1-\alpha)$-quantile of $g(\hat{\sigma}^2, y; \mu)$, denoted as $c'_s(\alpha; \mu)$, over $\mu \in \mathbb{R}^n$. Given a fixed μ, $c'_s(\alpha; \mu)$ can be easily estimated by sampling from $g(\hat{\sigma}^2, y; \mu)$. Note that $g(\hat{\sigma}^2, y; \mu) = g(\sigma^2, y; \mu) + [g(\hat{\sigma}^2) - g(\sigma^2)]$, the first term of which is a function with respect to y/σ through $\sigma^2\hat{L}$ and $\hat{\mu}$, and the second term of which is approximately a function with respect to y/σ and $(\hat{\sigma} - \sigma)/\sigma$. Write $g(y/\sigma, (\hat{\sigma} - \sigma)/\sigma) = g(\hat{\sigma}^2, y; \mu)$, where the two arguments are independent. In conclusion, we independently draw $y/\sigma \sim \mathcal{N}_n(\mu/\sigma, \mathbf{I}_n)$ and $(\hat{\sigma} - \sigma)/\sigma \sim \mathcal{N}_1(0, \frac{1}{2})$ and get desired samples through $g(y/\sigma, (\hat{\sigma} - \sigma)/\sigma)$. Lastly, we argue that it is unnecessary to take supremum of $c'_s(\alpha; \mu)$ over $\mu \in \mathbb{R}^n$ to estimate $c'_s(\alpha)$ but simply let use $c'_s(\alpha; 0)$. Because we only need $c'_s(\alpha)$ for weak signals and $\|\mu\|$ is usually close to 0 with a well estimated candidate set.

Proof of Theorem 6. Without ambiguity, all σ^2 in expressions are replaced by $\hat{\sigma}$ for r_A (2.7), $r_\perp$ (2.13), (2.8) and (2.14) in the rest of this proof. To complete the proof, it suffices to prove (2.8) and (2.14), the latter of which is a direct result from Lemma 5.

For (2.8), one can derive

$$\begin{aligned}\mathbb{P}\left\{\frac{\|P_A\mu - \hat{\mu}_A\|^2}{nr_A^2} \le 1/c_1\right\} &= \mathbb{P}\left\{\frac{\chi_k^2\sigma^2}{\chi_{k,\alpha/2}^2\hat{\sigma}^2} \le 1\right\} \\ &= F_{\chi_k^2}\left(\frac{\hat{\sigma}^2}{\sigma^2}\chi_{k,\alpha/2}^2\right) = F_{\chi_k^2}\left(\left[1 + \frac{\hat{\sigma}^2 - \sigma^2}{\sigma^2}\right]\chi_{k,\alpha/2}^2\right),\end{aligned} \tag{2.58}$$

where χ_k^2 is χ^2 distribution with k degrees of freedom, $F_{\chi_k^2}(x)$ is the cumulative distribution function of χ_k^2 and $\chi_{k,\alpha/2}^2$ is the $(1-\alpha/2)$-quantile. Let $f_{\chi_k^2}(x)$ be probability density function

of χ^2_k. We can further bound (2.58) by Taylor expansion as follows. Note that the maximum of $f_{\chi^2_k}(x)$ is at $x = k-2$. Then we have

$$\begin{aligned} F_{\chi^2_k}\left(\left[1+\frac{\hat{\sigma}^2-\sigma^2}{\sigma^2}\right]\chi^2_{k,\alpha/2}\right) &\leq F_{\chi^2_k}(\chi^2_{k,\alpha/2}) + f_{\chi^2_k}(k-2)\frac{|\hat{\sigma}^2-\sigma^2|}{\sigma^2}\chi^2_{k,\alpha/2} \\ &= \frac{\alpha}{2} + \frac{1}{2^{k/2}\Gamma(k/2)}(k-2)^{k/2-1}e^{-(k-2)/2}\frac{\hat{\sigma}^2-\sigma^2}{\sigma^2}\chi^2_{k,\alpha/2} \\ &= \frac{\alpha}{2} + \frac{1}{2\Gamma(k/2)}(\frac{k-2}{2e})^{k/2-1}\frac{|\hat{\sigma}^2-\sigma^2|}{\sigma^2}\chi^2_{k,\alpha/2}. \end{aligned} \tag{2.59}$$

Approximating $\Gamma(k/2)$ by Stirling's formula, one can derive

$$\begin{aligned} \Gamma(k/2) &= \left(\frac{k}{2}-1\right)\left(\frac{k}{2}-2\right)\ldots 1\cdot\Gamma(1) = \left(\frac{k}{2}-1\right)! \\ &= \sqrt{\pi(k-2)}\left(\frac{k-2}{2e}\right)^{k/2-1}e^{\theta_k}, \end{aligned} \tag{2.60}$$

if k is even, and

$$\begin{aligned} \Gamma(k/2) &= \left(\frac{k}{2}-1\right)\left(\frac{k}{2}-2\right)\ldots\frac{1}{2}\cdot\Gamma\left(\frac{1}{2}\right) = \frac{(k-2)!!}{2^{(k-1)/2}}\Gamma\left(\frac{1}{2}\right) \\ &= \frac{(k-1)!}{(k-1)!!2^{(k-1)/2}}\Gamma\left(\frac{1}{2}\right) = \frac{(k-1)!}{\left(\frac{k-1}{2}\right)!2^{(k-1)}}\Gamma\left(\frac{1}{2}\right) \\ &= \sqrt{2}\left(\frac{k-1}{2e}\right)^{(k-1)/2}\Gamma\left(\frac{1}{2}\right)e^{\theta_k}, \end{aligned} \tag{2.61}$$

if k is odd. Here, θ_k is a real number that depends on k and satisfies $\lim_{k\to\infty}\theta_k = 0$.

Based on $\chi^2_{k,\alpha/2} \asymp k$ (see the proof of Lemma 2), it follows from (2.59), (2.60) and (2.61) that for some constant M_5,

$$F_{\chi^2_k}\left(\left[1+\frac{\hat{\sigma}^2-\sigma^2}{\sigma^2}\right]\chi^2_{k,\alpha/2}\right) \leq \frac{\alpha}{2} + M_5\sqrt{k}(\hat{\sigma}^2-\sigma^2). \tag{2.62}$$

Since $\sqrt{k}(\hat{\sigma}^2-\sigma^2) = O(\sqrt{k/n}) = o(1)$ by assumption, the honesty for strong signals is proved. □

Proof of Lemma 6. Note that in this proof, r_A and $r_\perp$ are calculated with $\hat{\sigma}^2$ in place of σ^2.

Based on the rate of r_A with true σ, we have

$$\begin{aligned} r_A^2 &= c_1\hat{\sigma}^2\frac{\chi^2_{k,\alpha/2}}{n} = \frac{\hat{\sigma}^2}{n}(k+O(\sqrt{k})) \\ &= \frac{\sigma^2}{n}\frac{\hat{\sigma}^2}{\sigma^2}(k+O(\sqrt{k})) \\ &= \frac{\sigma^2}{n}\left(1+O(\frac{1}{n})\right)(k+O(\sqrt{k})) \\ &= \frac{\sigma^2}{n}\left(k+O(\sqrt{k})\right). \end{aligned}$$

For the radius of weak signals

$$r_\perp^2 = c_2\frac{n-k}{n}\hat{\sigma}^2\left\{1-\frac{(n-k)\hat{\sigma}^2}{\|y_\perp\|^2}+c_s(\alpha/2)(n-k)^{-1/2}\right\}. \tag{2.63}$$

The reciprocal of the stein shrinkage factor is given by

$$\begin{aligned} \frac{\|y_\perp\|^2}{(n-k)\hat{\sigma}^2} &= \frac{\|\varepsilon_\perp\|^2+2\langle\mu_\perp,\varepsilon_\perp\rangle+\|\mu_\perp\|^2}{(n-k)\hat{\sigma}^2} \\ &= \frac{\sigma^2}{\hat{\sigma}^2}\left[1+O_p\left(\frac{1}{\sqrt{n-k}}\right)+O_p\left(\frac{\|\mu_\perp\|}{n-k}\right)+\frac{\|\mu_\perp\|^2}{(n-k)\sigma^2}\right] \\ &= \frac{\sigma^2}{\hat{\sigma}^2}\left[1+O_p\left(\frac{1}{\sqrt{n-k}}\right)+O_p\left(\frac{\|\mu_\perp\|}{n-k}\right)\right] \\ &= 1+O_p\left(\frac{1}{\sqrt{n-k}}\right)+O_p\left(\frac{\|\mu_\perp\|}{n-k}\right). \end{aligned} \tag{2.64}$$

Plugging (2.64) back into (2.63), we further get

$$\begin{aligned} r_\perp^2 &= c_2\frac{n-k}{n}\sigma^2\frac{\hat{\sigma}^2}{\sigma^2}\left[O_p\left(\frac{1}{\sqrt{n-k}}\right)+O_p\left(\frac{\|\mu_\perp\|}{n-k}\right)+c_s(\alpha/2)(n-k)^{-1/2}\right] \\ &= \frac{n-k}{n}\left[O_p\left(\frac{1}{\sqrt{n-k}}\right)+O_p\left(\frac{\|\mu_\perp\|}{n-k}\right)\right]. \end{aligned}$$

□

CHAPTER 3

Empirical Study: Two-Step Stein Method v.s. Other Competitors

To demonstrate the advantages of our method, we develop in Section 3.1 a few competing methods making use of the lasso prediction or the oracle of the true sparsity. Then we provide extensive numerical comparisons in Section 3.2 to show the superior performance of our two-step Stein method, relative to the competitors, in a variety of sparsity settings, including when β is quite dense. Lastly, we end this chapter with a real-data example in Section 3.3.

3.1 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 lasso, we review an error bound for lasso prediction according to Bickel et al. (2009).

3.1.1 Lasso prediction error

Given X, y and $\lambda > 0$, consider the lasso estimator $\hat{\beta} = \hat{\beta}(y, X; \lambda)$ defined as in (2.18). Let $\omega(X) = \max_j(\|X_j\|^2/n)$. Error bounds of lasso prediction have been established under the restricted eigenvalue assumption (Bickel et al., 2009). For $S \subseteq [p]$ and $c_0 > 0$, define the cone

$$\mathscr{C}(S, c_0) := \left\{ \delta \in \mathbb{R}^p : \sum_{j \in S^c} |\delta_j| \leq c_0 \sum_{j \in S} |\delta_j| \right\}. \tag{3.1}$$

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

$$\kappa(s, c_0; X) := \min_{|S| \le s} \min_{\delta \neq 0} \left\{ \frac{\|X\delta\|}{\sqrt{n}\|\delta_S\|} : \delta \in \mathscr{C}(S, c_0) \right\} > 0. \tag{3.2}$$

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

$$\mathbb{P}\left\{ \|X(\hat{\beta} - \beta)\|^2 \le \frac{16K^2\sigma^2\omega(X)}{\kappa^2(s, 3; X)} s \log p \right\} \ge 1 - p^{1-K^2/8}. \tag{3.3}$$

Remark 8. The original theorem in Bickel et al. (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 $\omega(X)$.

3.1.2 Another adaptive method

Here we develop another adaptive method following the procedure in Section 3 of Robins and van der Vaart (2006), which constructs a confidence set for μ from $y \sim \mathcal{N}_n(\mu, \sigma^2\mathbf{I}_n)$ via sample splitting. The basic idea is introduced in Section 1.1. Applied to the linear model (2.1), the method can be described as follows. Split the original data set into (X', y') and (X, y), of which the former is used to obtain an initial lasso estimate $\hat{\beta} = \hat{\beta}(y', X'; \lambda)$ (2.18), and the latter is used to compute two quantities

$$R_n = \frac{1}{n}\|y - X\hat{\beta}\|^2 - \sigma^2, \qquad \hat{\tau}_n^2 = \frac{2\sigma^4}{n} + \frac{4\sigma^2}{n^2}\|X\beta - X\hat{\beta}\|^2, \tag{3.4}$$

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

$$\widehat{C}_a = \left\{ \mu \in \mathbb{R}^n : \frac{R_n - n^{-1}\|\mu - X\hat{\beta}\|^2}{\hat{\tau}_n} \ge -z_\alpha \right\}, \tag{3.5}$$

where z_α is the $(1-\alpha)$-quantile of the standard normal distribution. Note that $\hat{\tau}_n$ in (3.5) 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 $\mathrm{RE}(s, c_0)$, on the restricted maximum eigenvalue of $X^\mathsf{T}X/n$ over the cone $\mathscr{C}(S, c_0)$ (3.1). 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 8. *The $(1-\alpha)$ confidence set $\widehat{C}_a$ (3.5) is honest for all $\beta \in \mathbb{R}^p$. Suppose $s \log p = o(n)$, the sequence $X = X(n)$ satisfies*

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

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

$$|\widehat{C}_a| = O_p\left(n^{-1/4} + \sqrt{s \log p/n}\right). \tag{3.6}$$

These properties have been informally discussed in the introduction (Section 2.1). 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 (2.16) 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 3.2.4 demonstrate that $|\widehat{C}_a|$ can be 10 times larger than the diameter of our two-step Stein method when β is not sparse.

Proof of Theorem 8. The honesty of $\widehat{C}_a$ in (3.5) is guaranteed by Theorem 3.1 and Proposition 2.1 in Robins and van der Vaart (2006) with the only assumption

$$y/\sqrt{n} \sim \mathcal{N}_n(\mu/\sqrt{n}, \sigma^2 \mathbf{I}_n/n).$$

It is not difficult to verify that (X', y') satisfies all the conditions in Corollary B.2 and Theorem 7.2 of Bickel et al. (2009). Thus, with probability approaching one, we have $\|\hat{\beta} - \beta\|^2 = O(s \log p/n)$ and $(\hat{\beta} - \beta) \in \mathscr{C}(A_0, 3)$, as defined in (3.1), with $A_0 = \text{supp}(\beta)$. By the definition of $\zeta(s, 3; X)$, this implies that

$$\frac{1}{n}\|X(\beta - \hat{\beta})\|^2 \leq \zeta\|\hat{\beta} - \beta\|^2 = O_p(s \log p/n) = o_p(1). \tag{3.7}$$

Again, by Theorem 3.1 in Robins and van der Vaart (2006), we have

$$|\widehat{C}_a|^2 = O_p\left(n^{-1/2} + \frac{1}{n}\|X(\beta - \hat{\beta})\|^2\right) = O_p\left(n^{-1/2} + s \log p/n\right),$$

which completes the proof. □

3.1.3 An oracle lasso method

We calculate the lasso $\hat{\beta} = \hat{\beta}(y, X; \lambda)$ from the whole data set 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

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

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 3.1.4 for a two-step lasso method. Although there are sharper upper bounds, e.g. $O(s_\beta \log(p/s_\beta)/n)$, for lasso prediction error (e.g. Chapter 11 in Hastie et al. (2015)), 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.

3.1.4 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 $k = \mathrm{rank}(X_A)$ and further assume A contains strong signals only, that is, $A \subseteq \mathrm{supp}(\beta)$. We use the same method to find $\hat{\mu}_A$ and r_A (2.7) 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

$$\lambda_2 = K\sigma\sqrt{\log(p-k)/n}, \qquad K > 2\sqrt{2}, \tag{3.8}$$

to find the estimate

$$\tilde{\beta} = \tilde{\beta}(\lambda_2) = \underset{\beta\in\mathbb{R}^p}{\mathrm{argmin}} \left[\frac{1}{2n}\|y_\perp - P_A^\perp X\beta\|^2 + \lambda_2\|\beta\|_1\right]. \tag{3.9}$$

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 9. *Let $n \geq 1$ and $p \geq 2$. Suppose that $\|\beta\|_0 \leq s$ and Assumption* $\mathrm{RE}(s,3)$ *holds for X. Choose λ_2 as in* (3.8)*. Then for any fixed $A \subseteq \mathrm{supp}(\beta)$ with $k = \mathrm{rank}(X_A) < s$, we have*

$$\mathbb{P}\left\{\|P_A^\perp X(\tilde{\beta}-\beta)\|^2 \leq \frac{16K^2\sigma^2\omega(X)}{\kappa^2(s,3;X)}(s-k)\log(p-k)\right\} \geq 1-(p-k)^{1-K^2/8}. \tag{3.10}$$

Accordingly, the radius for weak signals is chosen as

$$r_\perp^2 = c_2\tilde{r}_\perp^2 = c_2c_l(\alpha/2)\sigma^2\frac{(s_\beta-k)\log(p-k)}{n}, \tag{3.11}$$

where $c_l(\alpha/2) = c_l(\alpha/2; P_A^\perp X)$ is a constant. Lastly, we combine $(\hat{\mu}_\perp, r_\perp)$ with $(\hat{\mu}_A, r_A)$ as in (2.5) 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 (2.18) 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 9. Upper bounds on $\|\hat{\beta} - \beta\|_\infty$ are available under certain conditions; see, for example, Theorem 11.3 in Hastie et al. (2015).

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 (2.29) and then choose the confidence set with the minimum volume or diameter. The main difference is how to approximate $c_l(\alpha)$ in (3.11), 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 (2.18) 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 $\nu \leq 1$ is a pre-determined 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 (2.29) 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 9 requires $|A_m| < s$.

Proof of Corollary 9. Rewrite orthogonal matrix $P_A^\perp = VV^\mathsf{T}$, where $V \in \mathbb{R}^{n\times(n-k)}$ consists of orthogonal unit column vectors. Write the lasso estimate in (3.9) as $\tilde{\beta} = F(y_\perp, P_A^\perp X; n\lambda_2)$, where F is understood as a mapping with a parameter $n\lambda_2 > 0$. Since $P_A^\perp X_A = 0$, the loss in (3.9) becomes

$$\begin{aligned}\frac{1}{2n}\|y_\perp - P_A^\perp X\beta\|^2 + \lambda_2\|\beta\|_1 &= \frac{1}{2n}\|y_\perp - P_A^\perp X_{A^c}\beta_{A^c}\|^2 + \lambda_2\|\beta\|_1 \\ &= \frac{1}{2n}\|V^\mathsf{T}y - V^\mathsf{T}X_{A^c}\beta_{A^c}\|^2 + \lambda_2\|\beta\|_1,\end{aligned}$$

which demonstrates that $\tilde{\beta}_A = 0$ and $\tilde{\beta}_{A^c} = F(V^\mathsf{T}y, V^\mathsf{T}X_{A^c}; n\lambda_2)$. Moreover, we have

$$\|V^\mathsf{T}X_{A^c}(\tilde{\beta}_{A^c} - \beta_{A^c})\| = \|P_A^\perp X(\tilde{\beta} - \beta)\|. \tag{3.12}$$

We will verify that the lasso problem, $\tilde{\beta}_{A^c} = F(V^\mathsf{T}y, V^\mathsf{T}X_{A^c}; n\lambda_2)$, satisfies all the assumptions in Lemma 7 so that we can apply (3.3) to bound the prediction error on the left side of (3.12). Since $A \subseteq \mathrm{supp}(\beta)$, we have $\|\beta_{A^c}\|_0 \le s - k$. Next, we show $V^\mathsf{T}X_{A^c} \in \mathbb{R}^{(n-k)\times(p-k)}$ satisfies $\mathrm{RE}(s-k, 3)$. Let D be any subset of $[p-k]$ such that $|D| \le (s-k)$. For any nonzero $\gamma \in \mathbb{R}^{p-k}$ in the cone $\mathscr{C}(D, 3)$, a vector $\delta = (\eta, \gamma) \in \mathbb{R}^p$ can always be constructed satisfying

$$X_A\eta + P_A X_{A^c}\gamma = 0,$$

since $P_A X_{A^c}\gamma \in \mathrm{span}(X_A)$. Define a mapping $g : i \mapsto i + |A|$ for $i \in [p]$ and let $B = [|A|] \cup g(D) \subset [p]$. Then $|B| = |A| + |D| \le s$, and $\delta \in \mathscr{C}(B, 3)$ because

$$\sum_{i\in B^c}|\delta_i| = \sum_{i\in D^c}|\gamma_i| \le 3\sum_{i\in D}|\gamma_i| \le 3\sum_{i\in B}|\delta_i|,$$

where the second step is due to $\gamma \in \mathscr{C}(D, 3)$. Based on that X satisfies $\mathrm{RE}(s, 3)$, we arrive

at the following inequality:

$$\frac{\|V^\mathsf{T} X_{A^c}\gamma\|}{\sqrt{n-k}\|\gamma_D\|} = \frac{\|X_A\eta + P_A X_{A^c}\gamma + P_A^\perp X_{A^c}\gamma\|}{\sqrt{n-k}\|\gamma_D\|}$$
$$= \frac{\sqrt{n}}{\sqrt{n-k}}\frac{\|X\delta\|}{\sqrt{n}\|\gamma_D\|} \geq \frac{\sqrt{n}}{\sqrt{n-k}}\frac{\|X\delta\|}{\sqrt{n}\|\delta_B\|} \geq \frac{\sqrt{n}}{\sqrt{n-k}}\kappa(s,3;X),$$

which shows that $\mathrm{RE}(s-k,3)$ holds for $V^\mathsf{T} X_{A^c}$ and

$$\kappa(s-k,3;V^\mathsf{T} X_{A^c}) \geq \sqrt{n/(n-k)}\kappa(s,3;X).$$

Lastly, $n\lambda_2 = K\sigma\sqrt{n\log(p-k)} \geq K\sigma\sqrt{(n-k)\log(p-k)}$, as required in Lemma 7.

So far, we have shown that $(V^\mathsf{T} X_{A^c}, V^\mathsf{T} y)$ and λ_2 satisfy all the conditions in Lemma 7, which with (3.12) implies that

$$\mathbb{P}\left\{\|P_A^\perp X(\tilde{\beta}-\beta)\|^2 \leq \frac{16nK^2\sigma^2\omega(V^\mathsf{T} X_{A^c})}{(n-k)\kappa^2(s-k,3;V^\mathsf{T} X_{A^c})}(s-k)\log(p-k)\right\}$$
$$\geq 1-(p-k)^{1-K^2/8},$$

for any $A \subseteq \mathrm{supp}(\beta)$. Then inequality (3.10) immediately follows by noting that $\omega(V^\mathsf{T} X_{A^c}) \leq \omega(X)$ and substituting $\kappa(s-k,3;V^\mathsf{T} X_{A^c})$ with $\sqrt{n/(n-k)}\kappa(s,3;X)$. □

3.2 Numerical results

We will first compare our method with the above competing methods when β is sparse relative to the sample size, i.e., s/n is small, and then consider the more challenging settings in which the sparsity s is comparable to n.

3.2.1 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 $\mathcal{U}(-b, b)$ and the other half of the components follow $\mathcal{U}(-0.2, 0.2)$, so there are two signal strengths under this setting.

Lastly, 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 data sets were generated independently, so that the total number of data sets used in this simulation study was 6,000.

The confidence level $1 - \alpha$ was set to 0.95. The threshold values $\{a_m\}$ in (2.29) 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 et al. (2009), $\lambda_{val} = 2\sqrt{2}\sigma\sqrt{\log p/n}$, 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 $c_o(\alpha; X, \lambda_{cv}) = \eta_1 c_o(\alpha; X, \lambda_{val})$ and $c_o(\alpha; X, \lambda_{1se}) = \eta_2 c_o(\alpha; X, \lambda_{val})$, where the factors η_k were chosen such that the overall coverage rate across data sets simulated with $b > 0.3$ was around the desired level.

Unlike the adaptive method in Section 3.1.2 and our two-step methods, the oracle lasso method does not require sample splitting. Consequently, a confidence set is constructed

based on the whole data set 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.

3.2.2 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 (2.30) with upper bound $E = 10$.

Figure 3.1 compares the geometric average radius $\bar{r}$ among the three methods against the signal strength b under the first way of drawing β. Every point in a panel was computed by averaging $\bar{r}$ from 100 data sets 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 3.2 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

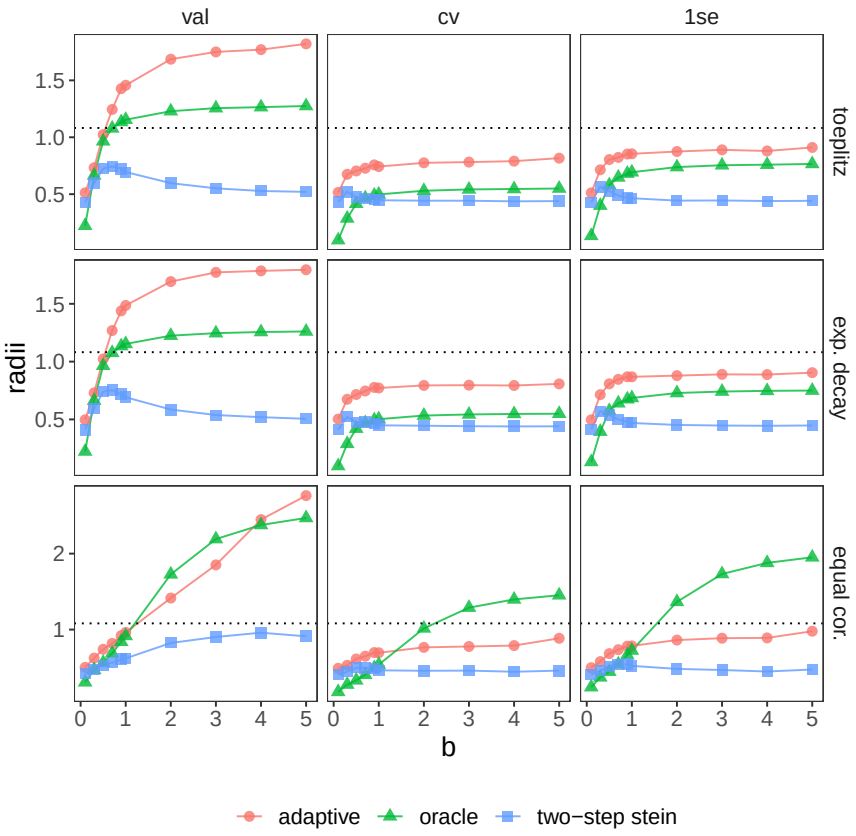

Figure 3.1: Geometric average radius against b under the first way of generating β. Each panel reports the results for one type of design (row) and one way of choosing λ (column), where the dashed line indicates the naive χ^2 radius.

properly distinguish strong signals and weak signals.

The coverage rates, each computed from 100 data sets, for each of the three ways of choosing λ are summarized in Figure 3.3. 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}

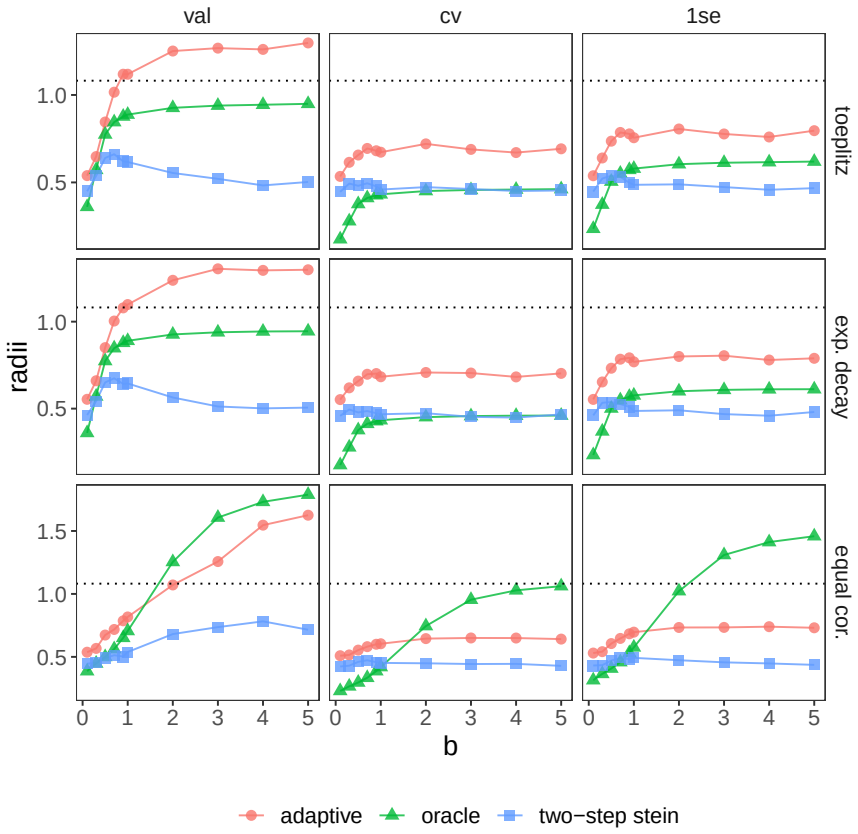

Figure 3.2: Average radius $\bar{r}$ against b in the second scenario of generating β.

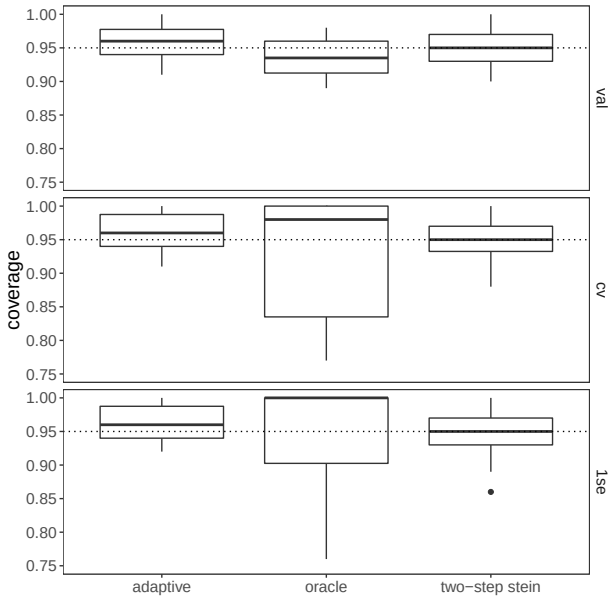

Figure 3.3: Box plots of coverage rates for each choice of λ, pooling data from three designs. The dashed lines indicate the desired confidence level of 95%.

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 (3.3) even with a known sparsity. Together with the results in Figures 3.1 and 3.2, 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 3.1.2) for all simulation settings and all choices of λ.

3.2.3 Comparison with the two-step lasso method

We discussed in Section 2.5 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-

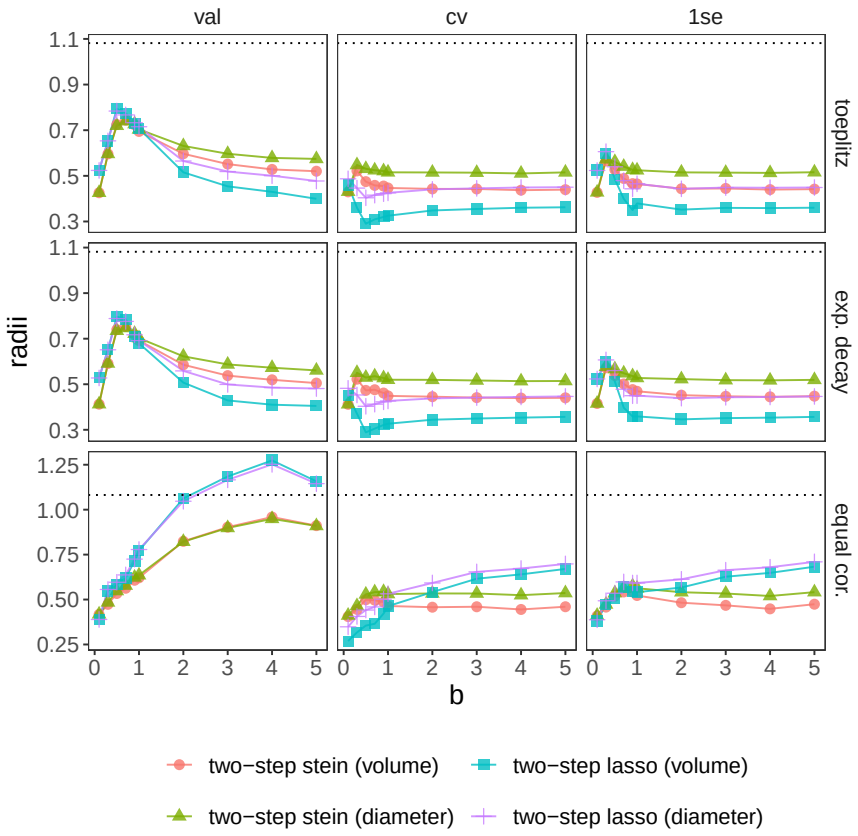

Figure 3.4: Average radius $\bar{r}$ against b in the first scenario of generating β.

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 3.4 shows the plots of radius against b by the four methods under different settings, while Figure 3.5 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 (3.9) 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

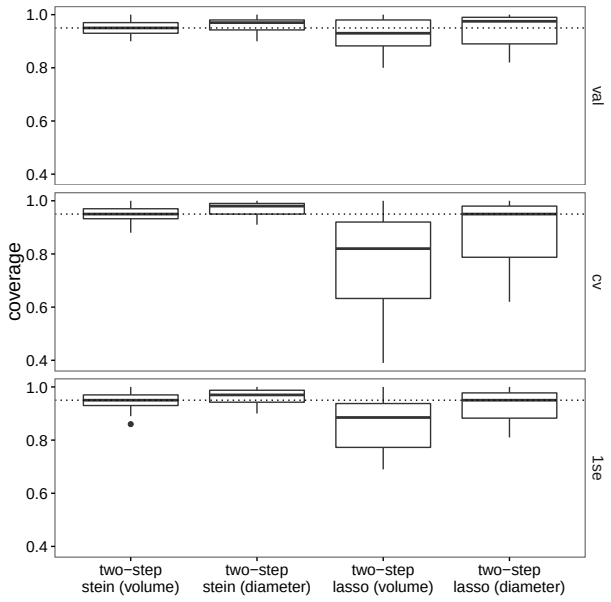

Figure 3.5: Box plots of coverage rates for each choice of λ. The dashed lines indicate the desired confidence level of 95%.

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 3.5 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 (3.3) with a known sparsity. The coverage rates of the two methods were quite comparable as reported in Figures 3.3 and 3.5. The geometric average radius of the oracle lasso method (Figure 3.1) was 2 to 5 times that of TLD (Figure 3.4). 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.

3.2.4 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 data sets under each setting. The tuning parameter λ was selected as λ_{1se} for all the results here.

Figure 3.6 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

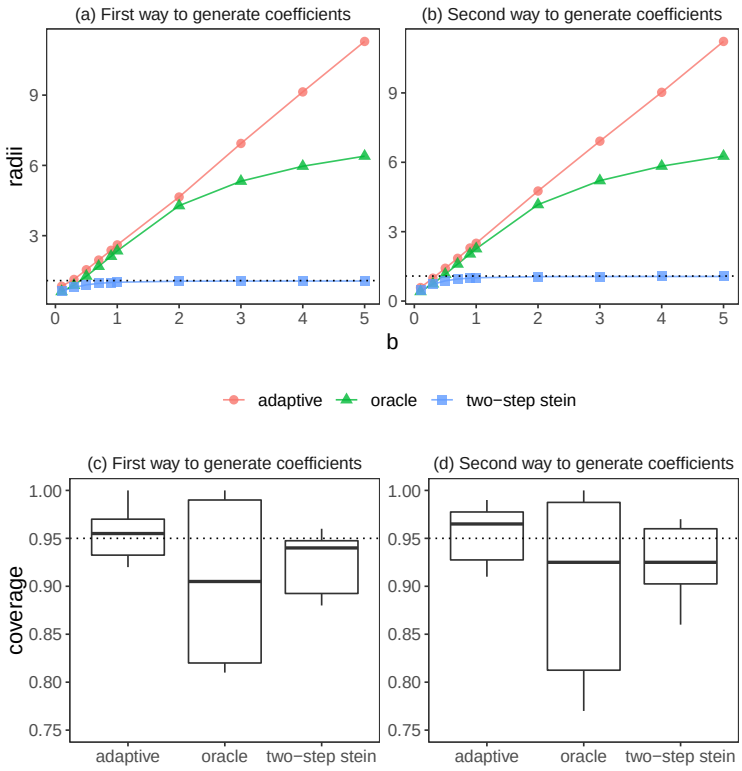

Figure 3.6: Comparison results under dense signal settings. (a) and (b) Geometric average radius against b. (c) and (d) Box plots of the coverage rates.

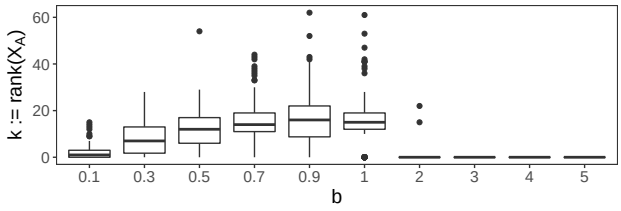

Figure 3.7: The box plot of k across data sets for each value of b

approached the naive 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 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 data sets here, the upper bounds for the diameters of these two methods, $|\widehat{C}_o| = O_p(\sqrt{s \log p/n})$ 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 3.6(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 strong signals in the set A, i.e., $k = |A|$. Figure 3.7 displays the box plot of k across 100 data sets 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 $b \geq 2$, i.e., $A = \varnothing$. Note that the radius of our method will be close to the naive χ^2 radius when $k = n$ or $k = 0$; see (2.15) 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 $\hat{\mu}(y; 0)$ (2.9) with radius estimated via the SURE.

3.2.5 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-squares 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 squares 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 3.2.1. In particular, $\hat{\sigma}^2$ was used in place of σ^2 to calculate the theoretical value λ_{val}. We input the same $\hat{\sigma}^2$ for 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 3.2.1. The average radii and coverage rates are reported in Figures 3.8 and 3.9, respectively. It is seen from Figure 3.8 that the trend of $\bar{r}$ against the signal strength b is quite similar to Figure 3.1 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 designs 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 3.9, 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 λ_{1se}, two practical ways of choosing the lasso tuning parameters. There are some outliers in the box-plots, 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 λ_{1se} our method had fewer outliers.

As expected, the coverage rates here in Figure 3.9 are somewhat lower than those reported

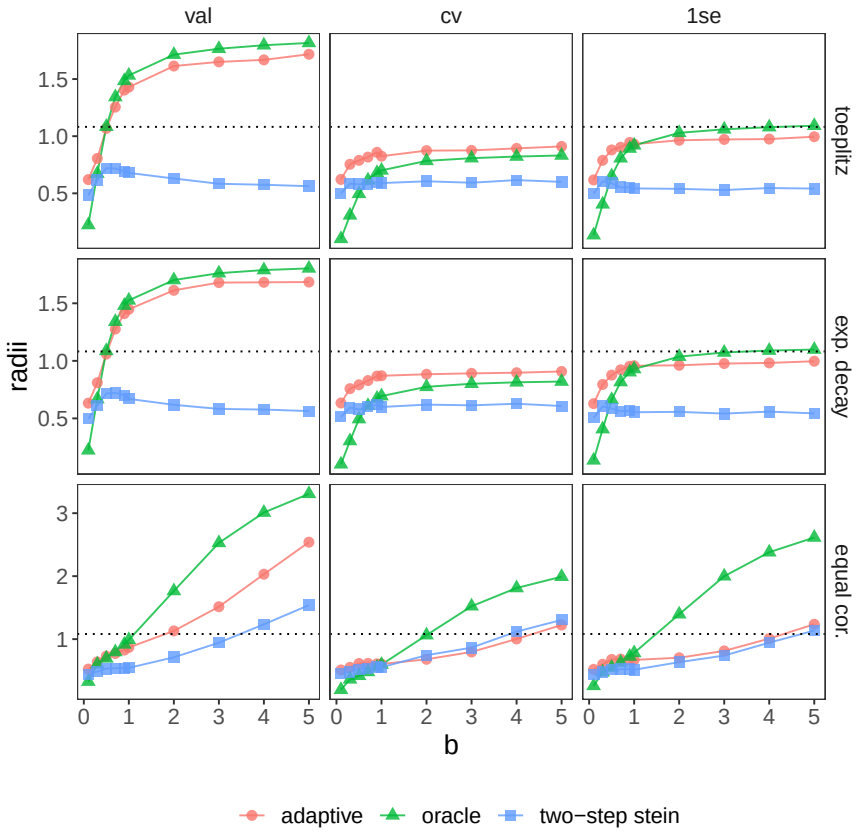

Figure 3.8: Average radius $\bar{r}$ against b with estimated error variance $\hat{\sigma}^2$.

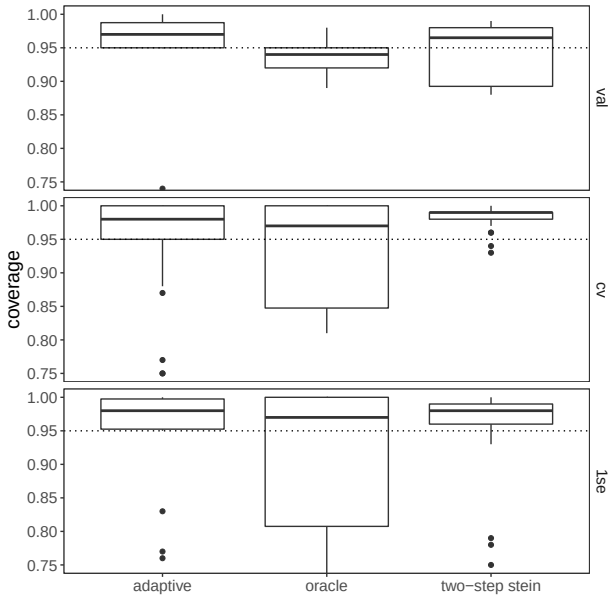

Figure 3.9: Box plots of coverage rates for each choice of λ with estimated $\hat{\sigma}^2$. The dashed lines indicate the desired confidence level of 95%. Outliers below 0.75 are truncated.

in Figure 3.3 assuming σ^2 is known. Among those data sets 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 were either < 0.8 or > 1.2 (recall $\sigma^2 = 1$), suggesting that the lower coverage was mostly caused by 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 3.1 for known σ^2 and Figure 3.8 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.

3.2.6 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 subsection, 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 forth 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 3.2.2. 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 3.2.5. For simplicity, our method still used a single candidate set $A = \text{supp}(\hat{\beta})$ in the comparison.

The average radii and coverage rates of the constructed confidence sets are summarized in Figure 3.10. 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 i.i.d. normal errors (lower panel of Figure 3.9). 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 3.10) in all the four settings, demonstrating the higher relative efficiency of our confidence sets when model assumptions are not satisfied, or even severely violated.

As shown in Figure 3.10, as b increased, the average radius of the adaptive method approached or exceeded an estimated naive χ^2 radius, $\hat{\sigma}(\chi^2_{n,\alpha}/n)^{1/2}$, under i.i.d. normal

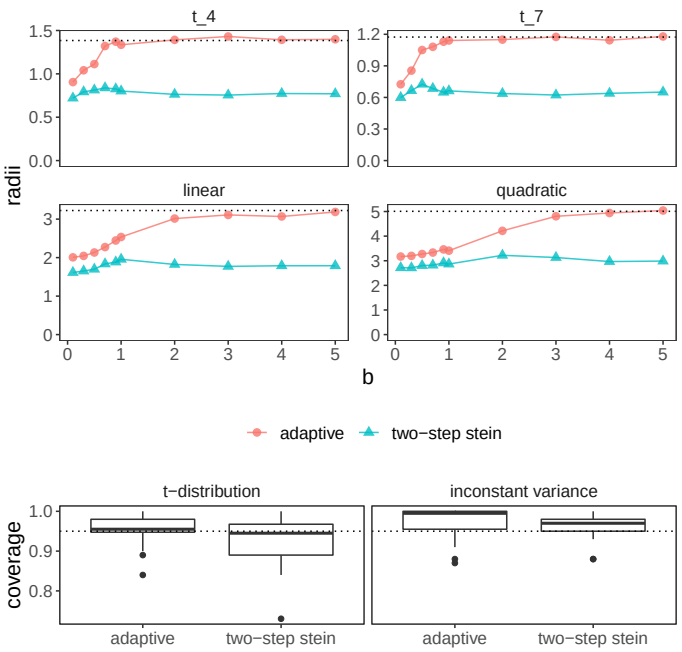

Figure 3.10: (Upper) Average radius $\bar{r}$ against b and (lower) box-plots of coverage rates of all settings under t-distributions or heterogeneous error variance. The dashed lines in the four top panels indicate the average naive χ^2 radius. The dashed lines in the box-plots indicate the nominal coverage level of 95%.

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 (2.11) plays a vital role against heterogeneity. Note that the left-hand side of the inequality in (2.12) 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 (2.12) still holds in spite of error heterogeneity, guaranteeing good coverage for our method.

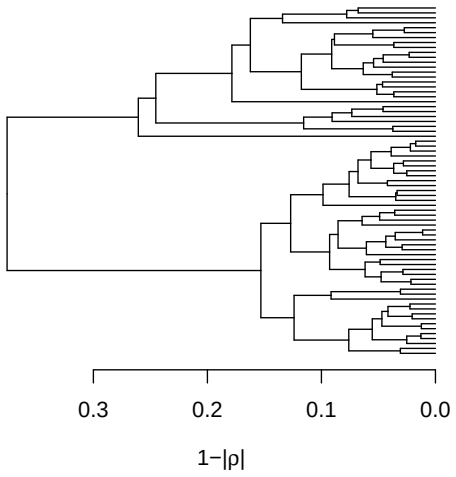

Figure 3.11: Hierarchical clustering of gene expression vectors among 72 individuals.

3.3 Real data analysis

We apply the two-step Stein method on the riboflavin data set compiled by Bühlmann et al. (2014) to demonstrate its practical significance. This data set 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 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.

Since the true riboflavin production rate is unknown, we conducted a simulation based on the real data set to verify the performance of our method. First, we estimated the error standard deviation $\tilde{\sigma} = 0.320$ from (X, y) by least-squares after scaled lasso selection. Next, we perturbed y to simulate $y^* \sim \mathcal{N}_n(y, \tilde{\sigma}^2 \mathbf{I}_n)$. In what follows, we will apply an inference method on the perturbed data (X, y^*) to construct a confidence set and check if it covers

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 than those simulation studies in the previous section. Again, we split (X, y^*) into two subsamples. One of them is used to calculate an initial lasso estimate $\hat{\beta}$ (2.18) 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 $\hat{\beta}$ is chosen by the one standard error rule. The other subsample will be used to construct a confidence set. In our analysis, two ways of sample splitting were considered. The first way is to randomly split the whole data set into two even halves, while the second is to split according to the gene clustering pattern of the n individuals. 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 3.11, 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. When splitting by clustering, we also swap the two subsamples to build two confidence sets, one for each subsample.

The whole process, starting from the simulation of y^*, was repeated 400 times. The average results are summarized in Table 3.1. Compared with the adaptive method, our method achieved much smaller average radius $\bar{r}$ with higher coverage that is above or close to the nominal level of 95% for both ways of sample splitting, randomly or by clusters. Besides, the $\bar{r}$ of the adaptive method was greater than the average radius of the naive χ^2 set, making it not practical useful, while the $\bar{r}$'s of our method were all below the naive radius. This is a very satisfactory result given that the relationship between y and X is noisy and could be nonlinear, as we mentioned above, and that the sample sizes here $n \leq 44$ are much smaller than $p > 4000$. For such a small sample size, the candidate set A calculated from the initial lasso estimate may not be stable. Therefore, we also tested our method with $A = \varnothing$, i.e. using only the weak signals to construct a confidence set. The results are reported in Table 3.1 as well. In this case, the $\bar{r} = r_\perp$ of our method was slightly smaller than that with $A = \text{supp}(\hat{\beta})$, showing that our method was quite robust with respect to the

Table 3.1: Comparison between the two-step Stein method and the adaptive method over 400 data sets based on the riboflavin data set.

		split evenly	split by clustering	
	group size	36	27	44
	χ^2 radius	0.530	0.541	0.424
adaptive	$\bar{r}$	0.781	0.745	0.985
	coverage	0.819	0.898	0.942
two-step Stein	$\bar{r}$	0.490	0.530	0.411
	r_s	0.615	0.631	0.482
$A = \text{supp}(\hat{\beta})$	$r_\perp$	0.450	0.478	0.404
	coverage	0.968	0.96	0.933
$A = \varnothing$	$r_\perp$	0.489	0.522	0.406
	coverage	0.968	1.000	0.915

candidate set A.

Next, we applied both our two-step Stein method and the adaptive method to the riboflavin data set to construct confidence sets. The results are summarized in Table 3.2. For random splitting, we repeated the process 100 times independently and report the average results over the 100 random subsamples. For our method, we chose the candidate set $A = \text{supp}(\hat{\beta})$ or $A = \varnothing$, like what we did in Table 3.1. One sees that, for both ways of sample splitting, the $\bar{r}$ of the adaptive method was substantially greater than the $\bar{r}$ of our method regardless of how A was chosen, consistent with the results on the perturbed data. Considering $\|y\|/\sqrt{n} = 7.038$, the confidence sets constructed by our method achieved substantial reduction in the uncertainty in μ, especially given the small sample sizes (≤ 44) after sample splitting and the large number $p > 4000$ of covariates. We observe that $\bar{r}$ of $A = \text{supp}(\hat{\beta})$ was smaller than $\bar{r}$ of $A = \varnothing$ for random splitting, while it was greater for cluster-based splitting. One reason for this observation is that the strong signals can be better detected with a random subsample, since the variance among the gene expression vectors (covariates) reduces when the individuals are partitioned into clusters.

Table 3.2: Confidence sets constructed on the riboflavin data set.

		split evenly	split by clustering	
	group size	36	27	44
adaptive	$\bar{r}$	0.657	0.641	0.943
two-step Stein $A = \text{supp}(\hat{\beta})$	$\bar{r}$	0.389	0.372	0.337
	r_s	0.449	0.404	0.175
	$r_\perp$	0.358	0.349	0.347
$A = \varnothing$	$r_\perp$	0.399	0.351	0.289

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 individuals (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 one aspect of the practical significance of our inference method.

CHAPTER 4

Post-Selection Inference with Estimator Augmentation

We have introduced post-selection inference in Section 1.4 and emphasized its difficulty regarding the restriction of sampling distribution to an irregular subset of $\mathbb{R}^n$. Min and Zhou (2019) showed that a randomization step and estimator augmentation can effectively construct an honest confidence set conditioning on any active set $\mathcal{A} = A$. However, the lack of theoretical justifications casts doubt on whether the nominal significance level can be achieved, whether the honesty can be maintained over the full parameter space, and whether the diameter of the confidence set converges to 0 under the asymptotic framework. This chapter intends to explore the potential answers to the aforementioned questions. We mainly consider the linear model

$$y = X\beta + \varepsilon, \tag{4.1}$$

where $y = y(n) \in \mathbb{R}^n$, $X = X(n) \in \mathbb{R}^{n\times p}$, $\beta = \beta(n) \in \mathbb{R}^p$ and $\varepsilon \sim \mathcal{N}_n(0, \sigma^2\mathbf{I}_n)$ under the asymptotic framework. Throughout this chapter, the active set $\mathcal{A}$ is defined by lasso estimator, $\mathcal{A} = \text{supp}(\hat{\beta})$ with $\hat{\beta}$ defined as

$$\hat{\beta} := \underset{\beta\in\mathbb{R}^p}{\text{argmin}} \frac{1}{2n}\|y - X\beta\|_2^2 + \lambda\sum_{j=1}^{p}|\beta_j|,$$

where $\lambda > 0$ is a tuning parameter, unless otherwise noted. We propose a new framework to construct confidence sets, which is organized as follows:

1. Derive the distribution of $\hat{\beta}$ conditioning on $\mathcal{A} = A$, i.e., $\pi(\hat{\beta}|\mathcal{A} = A; \beta)$.

2. Design a prior of β (or μ) $\pi(\beta)$ and then derive the conditional posterior $\pi(\beta|\hat{\beta}, \mathcal{A})$ under Bayesian framework.

3. Construct a credible set based on the conditional posterior $\pi(\beta|\hat{\beta}, \mathcal{A})$.

4. Study the properties of the credible set in the frequentist view. One of the most important question is whether $\widehat{C}$ is a confidence set satisfying

$$\liminf_{n\to\infty} \inf_{\beta\in\mathbb{R}^p} \mathbb{P}\{\beta \in \widehat{C}|\mathcal{A} = A\} \geq 1 - \alpha.$$

The rest of this chapter is organized as follows: Section 4.1 introduces estimator augmentation (Zhou, 2014), an effective method to derive a closed-form probability density function of $\hat{\beta}$ conditioning on $\mathcal{A}$. Section 4.2 presents our current progress regarding the posterior and the construction of credible sets in the Bayesian view. We propose a decision-theoretic framework to generalize this problem to generalized linear models (GLMs) in Section 4.3. Section 4.4 discusses the potential of applying this framework with $\mathcal{A}$ defined by the block lasso estimator.

4.1 Estimator augmentation

The first step in our framework has been solved by estimator augmentation (Zhou, 2014). We present his idea, as it is also associated to the generalization of GLMs later. In this section, we consider a more general distribution of ε with mean zero and variance σ^2, and redefine $\hat{\beta}$ as a general ℓ_1-penalized estimator given by the minimizer of the loss function

$$\ell(\beta) = \frac{1}{2n}\|y - X\beta\|_2^2 + \lambda \sum_{j=1}^{p} w_j|\beta_j|,$$

where $w_j > 0$, $j = 1, \ldots, p$ and $\lambda > 0$. By Karush–Kuhn–Tucker (KKT) conditions, $\hat{\beta}$ is characterized by

$$\frac{1}{n}X^{\mathsf{T}}y = \frac{1}{n}X^{\mathsf{T}}X\hat{\beta} + \lambda W S, \tag{4.2}$$

where $W = diag(w_1, w_2, \ldots, w_p)$ and S is the subgradient of $\|\hat{\beta}\|_1$, namely, for $j = 1, 2, \ldots, p$,

$$\begin{cases} S_j = sgn(\hat{\beta}_j) & \text{if } \hat{\beta}_j \neq 0, \\ S_j \in [-1, 1] & \text{if } \hat{\beta}_j = 0, \end{cases}$$

where $sgn(.)$ is the sign function. The idea is to find out the distribution of the augmented estimator $(\hat{\beta}, S)$ instead of $\hat{\beta}$. Further, let $\mathcal{A} = \text{supp}(\hat{\beta})$ be the set of nonzero coefficients and $\mathcal{I} = \{1, 2, \ldots, p\} \setminus \mathcal{A}$ be its complement. Then $(\hat{\beta}, S)$ can be equivalently represented by the triplet $(\hat{\beta}_{\mathcal{A}}, S_{\mathcal{I}}, \mathcal{A})$ and vice versa, by noting $\hat{\beta}_{\mathcal{I}}$ is a zero-vector and $S_{\mathcal{A}} = sgn(\hat{\beta}_{\mathcal{A}})$. It can be directly seen that the triplet $(\hat{\beta}_{\mathcal{A}}, S_{\mathcal{I}}, \mathcal{A})$ lies in the set

$$\Omega = \{(b_A, s_I, A) : A \subseteq \{1, 2, \ldots, p\}, b_A \in (\mathbb{R} \setminus \{0\})^{|A|}, s_I \in [-1, 1]^{p-|A|}\}, \tag{4.3}$$

where $I = \{1, 2, \ldots, p\} \setminus A$. Clearly, Ω is a subset of the product space of $\mathbb{R}^p$ and a finite discrete space, i.e., $\mathbb{R}^P \times 2^{\{1,2,\ldots,p\}}$, where $2^{\{1,2,\ldots,p\}}$ is the collection of all subsets of $\{1, 2, \ldots, p\}$. Let $C = \frac{1}{n} X^\mathsf{T} X$ and $U = \frac{1}{n} X^\mathsf{T} \varepsilon = \frac{1}{n} X^\mathsf{T} y - C\beta$. Rewrite (4.2) as

$$\begin{aligned} U &= \left(C_{\mathcal{A}} \,\middle|\, C_{\mathcal{I}}\right) \begin{pmatrix} \hat{\beta}_{\mathcal{A}} \\ 0 \end{pmatrix} + \lambda \left(W_{\mathcal{A}} \,\middle|\, W_{\mathcal{I}}\right) \begin{pmatrix} S_{\mathcal{A}} \\ S_{\mathcal{I}} \end{pmatrix} - C\beta \\ &= D(\mathcal{A}) \begin{pmatrix} \hat{\beta}_{\mathcal{A}} \\ S_{\mathcal{I}} \end{pmatrix} + \lambda W_{\mathcal{A}} sgn(\hat{\beta}_{\mathcal{A}}) - C\beta := H(\hat{\beta}_{\mathcal{A}}, S_{\mathcal{I}}, \mathcal{A}; \beta), \end{aligned} \tag{4.4}$$

where $D(\mathcal{A}) = \left(C_{\mathcal{A}} \,\middle|\, \lambda W_{\mathcal{I}}\right)$. By permuting the rows of $D(\mathcal{A})$, its determinant is

$$det(D(\mathcal{A})) = \begin{vmatrix} C_{\mathcal{A}\mathcal{A}} & 0 \\ C_{\mathcal{I}\mathcal{A}} & \lambda W_{\mathcal{I}\mathcal{I}} \end{vmatrix} = \lambda^{|\mathcal{I}|} det(C_{\mathcal{A}\mathcal{A}}) \prod_{j \in \mathcal{I}} w_j. \tag{4.5}$$

If $C_{\mathcal{A}\mathcal{A}}$ has full rank, $|det(D(\mathcal{A}))| > 0$.

In the *low-dimensional* setting ($p \leq n$), Lemma 2 in Zhou (2014) proves that if $rank(X) = p \leq n$, the mapping $H : \Omega \to \mathbb{R}^p$ defined in (4.4) between Ω (4.3) and $\mathbb{R}^p$ is bijective. Based on this lemma, one is able to find the sampling distribution of $(\hat{\beta}_{\mathcal{A}}, S_{\mathcal{I}}, \mathcal{A})$. Let ξ_k denote k-dimensional Lebesgue measure. Denote by $\phi_k(z; \mu, \Sigma)$ the probability density function of k-variate normal distribution with mean μ and covariance matrix Σ.

Theorem 10 (Theorem 1 and Corollary 1 in Zhou (2014))**.** *Assume $rank(X) = p$ and let f_U be the probability density function of U with respect to ξ_p. For $(b_A, s_I, A) \in \Omega$, the joint distribution of $(\hat{\beta}_{\mathcal{A}}, S_{\mathcal{I}}, \mathcal{A})$ is given by*

$$\mathbb{P}(\hat{\beta}_{\mathcal{A}} \in db_A, S_{\mathcal{I}} \in s_I, \mathcal{A} = A) = f_U(H(b_A, s_I, A; \beta)) |det(D(A))| \xi_p(db_A ds_I), \tag{4.6}$$

and the distribution of $(\hat{\beta}_{\mathcal{A}}, \mathcal{A})$ *is a marginal distribution given by*

$$\mathbb{P}(\hat{\beta}_{\mathcal{A}} \in db_A, \mathcal{A} = A) = \left[\int_{[-1,1]^{p-|A|}} f_U(H(b_A, s_I, A; \beta)) |det(D(A))| \xi_{p-|A|}(ds_I) \right] \xi_{|A|}(db_A). \tag{4.7}$$

Furthermore, if $\varepsilon \sim \mathcal{N}_n(0, \sigma^2 \mathbf{I}_n)$*, the joint density of* $(\hat{\beta}_{\mathcal{A}}, S_{\mathcal{I}}, \mathcal{A})$ *is*

$$\pi(b_A, s_I, A) = \phi_k(z; \mu(A, s_A; \beta), \Sigma(A; \sigma^2)) \mathbf{1}((z, A) \in \Omega), \tag{4.8}$$

where $z = (b_A, s_I)$*,* $\mathbf{1}(.)$ *is an indicator function and*

$$\mu(A, s_A; y) = D(A)^{-1}(C\beta - \lambda W_A s_A),$$
$$\Sigma(A; \sigma^2) = \frac{\sigma^2}{n} D(A)^{-1} C D(A)^{-\mathsf{T}}.$$

In the *high-dimensional* setting $(n > p)$, $U = \frac{1}{n} X^{\mathsf{T}} \varepsilon$ only lies in an n-dimensional subspace of $\mathbb{R}^p$, i.e., $row(X)$. Consequently, some constraint must be imposed on Ω to obtain a bijective mapping. Let $v_j \in \mathbb{R}^p$, $j = 1 \ldots, n$, be eigenvectors of $C = \frac{1}{n} X^{\mathsf{T}} X$, which also form a basis for $row(X)$. Select orthonormal vectors $v_{n+1}, \ldots, v_p$ to form a basis for $null(X)$. Naturally, $V = (v_1 | \ldots | v_p)$ forms a basis for $\mathbb{R}^p$. Let $R = \{1, \ldots, n\}$ and $V = \{n+1, \ldots, p\}$ be two index sets corresponding to the columns of V that form a basis for $row(X)$ and $null(X)$, respectively. Since $U \in Row(X)$, it follows from $V_N^{\mathsf{T}} U = 0$ and (4.4) that

$$0 = \lambda V_N^{\mathsf{T}} W S = \lambda \left(V_{\mathcal{A}N}^{\mathsf{T}} W_{\mathcal{A}\mathcal{A}} S_{\mathcal{A}} + V_{\mathcal{I}N}^{\mathsf{T}} W_{\mathcal{I}\mathcal{I}} S_{\mathcal{I}} \right), \tag{4.9}$$

which implies that WS must lie in $row(X)$. Therefore, Ω must be restricted to

$$\Omega_r = \{(b_A, s_I, A) \in \Omega : V_{AN}^{\mathsf{T}} sgn(b_A) + V_{IN}^{\mathsf{T}} W_{II} s_I = 0\}$$

for the augmented estimator $(\hat{\beta}_{\mathcal{A}}, S_{\mathcal{I}}, \mathcal{A})$. According to Lemma 3 (Zhou, 2014), the restriction of the mapping H to Ω_r, denoted as $H|\Omega_r : \Omega_r \to \text{row}(X)$, is bijective. Represent U by coordinates with respect to V_R and let $R = V_R^{\mathsf{T}} U$ to get

$$R = V_R^{\mathsf{T}} C_{\mathcal{A}} \hat{\beta}_{\mathcal{A}} + \lambda V_{\mathcal{A}R}^{\mathsf{T}} W_{\mathcal{A}\mathcal{A}} S_{\mathcal{A}} + \lambda V_{\mathcal{I}R}^{\mathsf{T}} W_{\mathcal{I}\mathcal{I}} S_{\mathcal{I}} - V_R^{\mathsf{T}} C\beta := H_r(\hat{\beta}_{\mathcal{A}}, S_{\mathcal{I}}, A; \beta). \tag{4.10}$$

Differentiating (4.9) and (4.10) with respect to $(\hat{\beta}_A, S_\mathcal{I})$, respectively, we have

$$dR = V_R^\mathsf{T} C_\mathcal{A} d\hat{\beta}_A + \lambda V_{\mathcal{I}R}^\mathsf{T} W_{\mathcal{I}\mathcal{I}} dS_\mathcal{I}, \tag{4.11}$$

$$V_{\mathcal{I}N}^\mathsf{T} W_{\mathcal{I}\mathcal{I}} W dS_\mathcal{I} = 0, \tag{4.12}$$

implying that $dS_\mathcal{I}$ is in $null(V_{\mathcal{I}N}^\mathsf{T} W_{\mathcal{I}\mathcal{I}})$. Under a mild assumption on X, the dimension of $null(V_{\mathcal{I}N}^\mathsf{T} W_{\mathcal{I}\mathcal{I}})$ is $n - |\mathcal{A}| \geq 0$ so there exists a $B(\mathcal{I}) \in \mathbb{R}^{|\mathcal{I}| \times (n-|\mathcal{A}|)}$, which is an orthonormal basis for $null(V_{\mathcal{I}N}^\mathsf{T} W_{\mathcal{I}\mathcal{I}})$. Let $d\tilde{S}$ denote coordinates so that $dS = B(\mathcal{I}) d\tilde{S}$. Then (4.11) becomes

$$dR = V_R^\mathsf{T} C_\mathcal{A} d\hat{\beta}_\mathcal{A} + \lambda V_{\mathcal{I}R}^\mathsf{T} W_{\mathcal{I}\mathcal{I}} B(\mathcal{I}) d\tilde{S} = T(\mathcal{A}) \begin{pmatrix} d\hat{\beta}_\mathcal{A} \\ d\tilde{S} \end{pmatrix},$$

where $T(\mathcal{A}) = \left(V_R^\mathsf{T} C_\mathcal{A} \;\middle|\; \lambda V_{\mathcal{I}R}^\mathsf{T} W_{\mathcal{I}\mathcal{I}} B(\mathcal{I}) d\tilde{S}. \right)$ is the Jacobian of the map H_r.

Theorem 11 (Theorem 2 in Zhou (2014))**.** *Assume that $p > n$ and every n columns of X are linearly independent and every $(p-n)$ rows of V_N are linearly independent. Let f_R be the probability density of R with respect to ξ_n. For $(b_A, s_I.A) \in \Omega_r$, the joint distribution of $(\hat{\beta}_\mathcal{A}, S_\mathcal{I}, \mathcal{A})$ is given by*

$$\mathbb{P}(\hat{\beta}_\mathcal{A} \in db_A, S_\mathcal{I} \in ds_I, \mathcal{A} = A) = f_R(H_r(b_A, s_I, A; \beta)) |det(T(A))| \xi_n(db_A d\tilde{s}). \tag{4.13}$$

If $\varepsilon \sim \mathcal{N}_n(0, \sigma^2 \mathbf{I}_n)$, then $f_R(\cdot) = \phi_n(\cdot; 0, \frac{\sigma^2}{n} V_R^\mathsf{T} C V_R)$. Finally, it is straightforward to derive the conditional density of $[\hat{\beta}_\mathcal{A}, S_\mathcal{I} | \mathcal{A} = A; \beta]$ from (4.6) or (4.13).

4.2 Posterior distribution and inference after model selection

The second step, the third step and the forth step in our framework are closely related. We are currently researching on these three steps. For simplicity, we consider the low-dimensional setting $(n > p)$ with $\varepsilon \sim \mathcal{N}_n(0, \sigma^2 \mathbf{I}_n)$ and known σ^2.

4.2.1 Conditional posterior distribution

It follows from (4.8) that the probability of $\mathcal{A} = A$ is given by

$$Z_A(\beta) := \int_{\Omega_A} \phi_k(H(b_A, s_I, A; \beta); \mu(A, s_A; \beta), \Sigma(A; \sigma^2))\xi_p(db_A ds_I), \tag{4.14}$$

where Ω_A is the subspace of Ω (4.3) restricted to $\mathcal{A} = A$, i.e.,

$$\Omega_A = \{(b_A, s_I) \in \mathbb{R}^p : b_A \in (\mathbb{R} \setminus \{0\})^{|A|}, s_I \in [-1, 1]^{p-|A|}\}. \tag{4.15}$$

So the conditional density of $(\hat{\beta}_A, S_I)$ given $\mathcal{A} = A$ is

$$\pi(b_A, s_I|\mathcal{A} = A; \beta) = \frac{1}{Z_A(\beta)}\pi(b_A, s_I, A) = \frac{1}{Z_A(\beta)}\phi_k(H(b_A, s_I, A; \beta); \mu(A, s_A; \beta), \Sigma(A; \sigma^2)) \tag{4.16}$$

According to the Bayesian view, we should find out a proper prior of β, $\pi(z)$, in order to obtain the posterior and then to construct credible sets. However, unlike the Bayesian view, which requires the prior is independent of the data set (X, y), we can design the prior based on $(X, y), n, p, \mathcal{A} = A$ or any quantities related to (X, y), since our eventual objective is to prove that the credible sets constructed from the posterior distribution are also $(1 - \alpha)$ confidence sets according to frequentist view. As a preliminary study, let $\pi(z) \propto 1$. Then one can derive the implied posterior distribution of β given the triplet $(\hat{\beta}_{\mathcal{A}}, S_{\mathcal{I}}, \mathcal{A})$, for any $z \in \mathbb{R}^p$

$$\begin{aligned}
&\pi(z|\hat{\beta}_{\mathcal{A}} = b_A, S_{\mathcal{I}} = s_I, \mathcal{A} = A) \propto \pi(z)\pi(b_A, s_I|\mathcal{A} = A; \beta = z) \\
&\propto \frac{1}{Z_A(z)} \exp\left(-\frac{1}{2}\left[\begin{pmatrix} b_A \\ s_I \end{pmatrix} - \mu(A, s_A; z)\right]^{\mathsf{T}} \Sigma(A; \sigma^2)^{-1} \left[\begin{pmatrix} b_A \\ s_I \end{pmatrix} - \mu(A, s_A; z)\right]\right) \\
&\propto \frac{1}{Z_A(z)} \exp\left(-\frac{n}{2\sigma^2}\left[D(A)\begin{pmatrix} b_A \\ s_I \end{pmatrix} + \lambda W_A s_A - Cz\right]^{\mathsf{T}} C^{-1} \left[D(A)\begin{pmatrix} b_A \\ s_I \end{pmatrix} + \lambda W_A s_A - Cz\right]\right) \\
&\propto \frac{1}{Z_A(z)} \exp\left(-\frac{n}{2\sigma^2}[Cb + \lambda Ws - Cz]^{\mathsf{T}} C^{-1}[Cb + \lambda Ws - Cz]\right) \\
&\propto \frac{1}{Z_A(z)} \exp\left(-\frac{n}{2\sigma^2}[z - (b + \lambda C^{-1}Ws)]^{\mathsf{T}} C[z - \lambda(b + C^{-1}Ws)]\right) \\
&\propto \frac{1}{Z_A(z)} \phi_p(z; b + \lambda C^{-1}Ws, \frac{\sigma^2}{n}C^{-1}).
\end{aligned} \tag{4.17}$$

Note that $\frac{1}{Z_A(z)}$ depends on z and thus cannot be dropped. In fact, without conditioning on $\mathcal{A} = A$, the posterior distribution of β becomes

$$\pi(z|\hat{\beta}_{\mathcal{A}} = b_{\mathcal{A}}, S_{\mathcal{I}} = s_{\mathcal{I}}) \propto \phi_p(z; b + \lambda C^{-1} W s, \frac{\sigma^2}{n} C^{-1}) \propto \phi_p(z; \hat{\beta}^{(ols)}, \frac{\sigma^2}{n} C^{-1}),$$

where $\hat{\beta}^{(ols)}$ is the ordinary lest square estimator of (X, y), by the equation

$$\frac{1}{n} X^\mathsf{T} X \hat{\beta}^{(ols)} = \frac{1}{n} X^\mathsf{T} y = \frac{1}{n} X^\mathsf{T} X \hat{\beta} + \lambda W S.$$

Note that the respective dimensions of $b_{\mathcal{A}}$ and $s_{\mathcal{I}}$ in $\pi(z|\hat{\beta}_{\mathcal{A}} = b_{\mathcal{A}}, S_{\mathcal{I}} = s_{\mathcal{I}})$ are not fixed and any n-dimensional vector could be values of different (b_A, s_I)'s with different $\mathcal{A} = A$, which is a significant difference from $\pi(z|\hat{\beta}_{\mathcal{A}} = b_A, S_{\mathcal{I}} = s_I, \mathcal{A} = A)$ where A is fixed. Therefore, the construction of credible sets based on $\pi(z|\hat{\beta}_{\mathcal{A}} = b_A, S_{\mathcal{I}} = s_I, \mathcal{A} = A)$ is a non-trivial and even more complicate problem due to the irregular function $Z_A(z)$.

To further simplify this problem, we assume $C = \frac{1}{n} X^\mathsf{T} X = \mathbf{I}_n$ and let $\hat{\beta}$ be the lasso estimator (i.e., $w_j = 1$ for $j = 1, 2, \ldots, p$). Under this assumption, the expression (4.17) is rewritten as

$$\pi(z|\hat{\beta}_{\mathcal{A}} = b_A, S_{\mathcal{I}} = s_I, \mathcal{A} = A) \propto$$
$$\prod_{j:b_j>0} \frac{\phi(z_j; b_j + \lambda, \sigma^2/n)}{\Phi\left(\frac{z_j - \lambda}{\sqrt{\sigma^2/n}}\right)} \prod_{j:b_j<0} \frac{\phi(z_j; b_j - \lambda, \sigma^2/n)}{\Phi\left(\frac{-\lambda - z_j}{\sqrt{\sigma^2/n}}\right)} \prod_{j:-1\le s_j\le 1} \frac{\phi(z_j; \lambda s_j, \sigma^2/n)}{\Phi\left(\frac{\lambda - z_j}{\sqrt{\sigma^2/n}}\right) - \Phi\left(\frac{-\lambda - z_j}{\sqrt{\sigma^2/n}}\right)}, \tag{4.18}$$

where $\phi(\cdot;\cdot,\cdot)$ is $\phi_1(\cdot;\cdot,\cdot)$ and Φ is the cumulative distribution of the standard normal distribution. Note that (4.18) is the product of p terms, each of which is a function with respect to a single value of b_j, which means all $[\beta_j|\hat{\beta}_{\mathcal{A}}, S_{\mathcal{I}}, \mathcal{A}]$ are independent of each other. Usually, the inference on the active set is of more interest (e.g., ν in (1.21)), so we next look at a single coefficient $[\beta_j|\hat{\beta}_{\mathcal{A}}, S_{\mathcal{I}}, \mathcal{A}]$, where the corresponding estimated coefficient is positive, i.e., $\hat{\beta}_j > 0$, and study the closed-form marginal posterior distributions.

We frequently use the result (4.19) in the following paragraphs. The tail probability of the standard normal distribution, $\Phi(z)^c$, is bounded by

$$\frac{z}{z^2+1} \frac{e^{-z^2/2}}{\sqrt{2\pi}} \le \Phi(z)^c = 1 - \Phi(z) \le \frac{1}{z} \frac{e^{-z^2/2}}{\sqrt{2\pi}}, \tag{4.19}$$

for $z > 0$. In other words, $\Phi^c(z) \propto (\frac{1}{z} + O(\frac{1}{z^3}))\frac{e^{-z^2/2}}{\sqrt{2\pi}}$ for $z > 0$.

Checking the tail probability of $[\beta_j|\hat{\beta}_{\mathcal{A}}, S_{\mathcal{I}}, \mathcal{A}]$ for $\hat{\beta}_j > 0$, we have, if $z_j \to -\infty$,

$$\begin{aligned}
\pi(z_j|\hat{\beta} = b, S = s, \mathcal{A} = A, \hat{\beta}_j > 0) &\propto e^{-\frac{n}{2\sigma^2}(z_j - b_j - \lambda)^2} \Big/ \Phi(\frac{z_j - \lambda}{\sqrt{\sigma^2/n}}) \\
&\propto e^{-\frac{n}{2\sigma^2}((z_j - b_j - \lambda)^2} \Big/ \left[\frac{1}{|z_j - \lambda|\sqrt{n/\sigma^2}} + O(\frac{1}{|z_j - \lambda|^3 (n/\sigma^2)^{3/2}})\right] e^{-\frac{n}{2\sigma^2}(z_j - \lambda)^2} \\
&\propto \frac{1}{\sigma} e^{-\frac{nb_j^2}{2\sigma^2}} \sqrt{n}(\lambda - z_j) e^{-\frac{n}{\sigma^2} b_j (\lambda - z_j)} \left(1 + O(\frac{1}{(z_j - \lambda)^2 n/\sigma^2})\right), \qquad (4.20)
\end{aligned}$$

and, if $z_j \to +\infty$,

$$\begin{aligned}
\pi(z_j|\hat{\beta} = b, S = s, \mathcal{A} = A, \hat{\beta}_j > 0) &\propto e^{-\frac{n}{2\sigma^2}(z_j - b_j - \lambda)^2} \Big/ \left[1 - \Phi^c(\frac{z_j - \lambda}{\sqrt{\sigma^2/n}})\right] \\
&\propto e^{-\frac{n}{2\sigma^2}(z_j - b_j - \lambda)^2} \Big/ \left[1 - \left(\frac{1}{|z_j - \lambda|\sqrt{n/\sigma^2}} + O(\frac{1}{|z_j - \lambda|^3 (n/\sigma^2)^{3/2}})\right) e^{-\frac{n}{2\sigma^2}(z_j - \lambda)^2}\right] \\
&\propto e^{-\frac{n}{2\sigma^2}(z_j - b_j - \lambda)^2} \left(1 + O(\frac{1}{|z_j - \lambda|\sqrt{n/\sigma^2}} e^{-\frac{n}{2\sigma^2}(z_j - \lambda)^2})\right). \qquad (4.21)
\end{aligned}$$

One can immediately see from the left and right tails that $[\beta_j|\hat{\beta}_{\mathcal{A}}, S_{\mathcal{I}}, \mathcal{A}]$ for $\hat{\beta}_j > 0$ is approximately the mixture of a Gamma distribution $\Gamma(2, 1)$ and a normal distribution. This observation can be verified by the simulation below.

Set $n = 500$, $p = 3$, $\sigma = 1$. We numerically computed the density function of the posterior distribution $[\beta_j|\hat{\beta}_{\mathcal{A}}, S_{\mathcal{I}}, \mathcal{A}]$ for $\hat{\beta}_j > 0$ associated with the flat prior in Figure 4.1. Here, $\hat{\beta}_j$ took values of 0.005, 0.01 and 0.1, respectively. It can be seen from $\hat{\beta}_j = 0.005$ that the left tail of the density function decayed at a rate slower than the right tail, which matches our derivation in (4.20) and (4.21). Though $[\beta_j|\hat{\beta}_{\mathcal{A}}, S_{\mathcal{I}}, \mathcal{A}]$ is approximately a mixture of a Gamma distribution and a normal distribution, the weight of each component varies and depends on how close $\hat{\beta}_j$ is to zero. If $\hat{\beta}_j$ is close to zero, the Gamma distribution dominates. In contrast, if $\hat{\beta}_j$ is much greater than 0, say $\hat{\beta} = 0.1$, the normal distribution dominates. This uncertainty of the weights causes great trouble in the construction of credible intervals. We also tested a Gaussian prior and the result is similar to Figure 4.1, indicating that a non-trivial and informative prior must be provided in order to overcome such uncertainty.

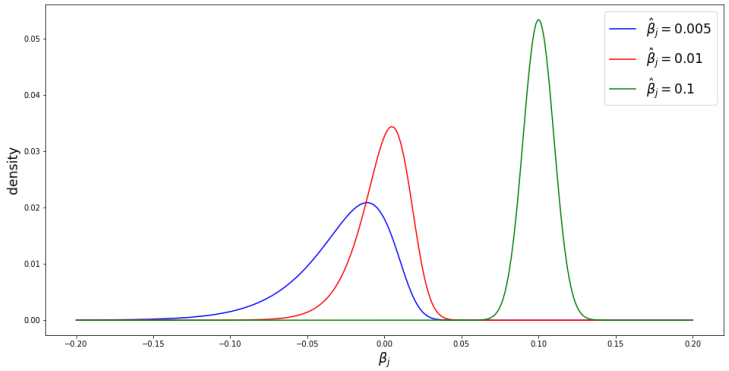

Figure 4.1: The conditional density function of $[\beta_j|\hat{\beta}_{\mathcal{A}}, S_{\mathcal{I}}, \mathcal{A}]$

4.2.2 Construction of credible sets

Based on the conditional sampling distribution derived in the last subsection, we conduct an initial investigation on the construction of credible sets and then study their significance level according to frequentist view. Throughout this subsection, we focus on the conditional marginal distribution of $[\beta_j|\hat{\beta}_{\mathcal{A}}, S_{\mathcal{I}}, \mathcal{A}]$ for $\hat{\beta}_j > 0$.

The first proposed credible interval has the form of

$$I_1(j; \hat{\beta}, S, \mathcal{A}) := \{\beta_j : \hat{\beta}_j + \lambda - \frac{\sigma}{\sqrt{n}} q_{\alpha/2} \leq \beta_j \leq \hat{\beta}_j + \lambda + \frac{\sigma}{\sqrt{n}} q_{\alpha/2}\}, \tag{4.22}$$

where $q_{\alpha/2}$ is the $(1-\alpha)$-quantile of the standard normal distribution. It is common to use $I_1(j; \hat{\beta}, S, \mathcal{A})$ as a confidence interval under the assumption of normality. Alternatively, we can construct a credible interval in the form of

$$I_2(j; \hat{\beta}, S, \mathcal{A}) := \{\beta_j : -\frac{\sigma^2}{n\hat{\beta}_j} \gamma^{(2,1)}_{\alpha/2} + \lambda \leq \beta_j \leq \hat{\beta}_j + \lambda + \frac{\sigma}{\sqrt{n}} q_{1-\frac{\alpha}{2}}\}, \tag{4.23}$$

where $\gamma^{(2,1)}_{\alpha/2}$ is the $(1-\alpha)$-quantile of the gamma distribution, $\Gamma(2,1)$. This construction is based on the approximation of the distribution $[\beta_j|\hat{\beta}_{\mathcal{A}}, S_{\mathcal{I}}, \mathcal{A}]$ in (4.20) and (4.21). Comparing the two credible intervals, one can find that the length of $I_1(j; \hat{\beta}, S, \mathcal{A})$ does not depend on $\hat{\beta}_j$ and converges to 0 as $n \to \infty$, while the length of $I_2(j; \hat{\beta}, S, \mathcal{A})$ depends on both n and $\hat{\beta}_j$ and may not converge. If $\hat{\beta}_j$ is fixed and $n \to \infty$, $I_2(j; \hat{\beta}, S, \mathcal{A})$ shrinks to $[\lambda, \hat{\beta}_j + \lambda]$.

Therefore, $I_1(j;\hat{\beta},S,\mathcal{A})$ is better in the sense of the length. However, the significance level in the frequentist framework is our main concern and should be studied for the two credible intervals as well. Let $\beta_j^* = \beta_j^*(n)$ be the jth component of the true β^*. Under the orthogonal design of X, the lasso estimator of the jth component is

$$\begin{aligned}\hat{\beta}_j &= sgn(|X_j^\mathsf{T} y|/n)\left(|X_j^\mathsf{T} y|/n - \lambda\right)_+ \\ &= sgn(\beta_j^* + \frac{\sigma}{\sqrt{n}}\epsilon_j)\left(\beta_j^* + \frac{\sigma}{\sqrt{n}}\epsilon_j - \lambda\right)_+, \end{aligned} \tag{4.24}$$

where $\epsilon_j = X_j^\mathsf{T}\varepsilon/n$ is the standard normal in distribution. Now we consider the asymptotic properties of the conditional probability

$$\mathbb{P}\{\beta_j^* \in I_k(j;\hat{\beta},S,\mathcal{A})|\mathcal{A}=A\} = \mathbb{P}\{\beta_j^* \in I_k(j;\hat{\beta},S,\mathcal{A})|\hat{\beta}_j > 0\}, \tag{4.25}$$

for $k = 1, 2$. Note that the equation in (4.25) holds because of the independence of β_j for $j = 1,\ldots,p$ and the assumption on $\hat{\beta}_j > 0$. The event $\{\hat{\beta}_j > 0\}$ is equivalent to

$$\left\{\varepsilon : \frac{X_j^\mathsf{T}\varepsilon}{n} > \lambda - \beta_j^*\right\} \equiv \left\{\epsilon_j : \epsilon_j > \frac{\sqrt{n}}{\sigma}(\lambda - \beta_j^*)\right\}. \tag{4.26}$$

Plugging (4.24) and (4.26) in (4.25) for $I_1(j;b,s,A)$, we have

$$\begin{aligned}&\mathbb{P}\left\{\hat{\beta}_j + \lambda - \frac{\sigma}{\sqrt{n}}q_{\alpha/2} \le \beta_j^* \le \hat{\beta}_j + \lambda + \frac{\sigma}{\sqrt{n}}q_{\alpha/2}\middle|\hat{\beta}_j > 0\right\} \\ &= \mathbb{P}\left\{-q_{\alpha/2} \le \epsilon_j \le q_{\alpha/2}\middle|\epsilon_j > \frac{\sqrt{n}}{\sigma}(\lambda - \beta_j^*)\right\} \\ &= \frac{\mathbb{P}\{-q_\alpha \le \epsilon_j \le q_\alpha \cap \epsilon > \frac{\sqrt{n}}{\sigma}(\lambda-\beta_j^*)\}}{\mathbb{P}\{\epsilon_j > \frac{\sqrt{n}}{\sigma}(\lambda-\beta_j^*)\}}, \end{aligned} \tag{4.27}$$

One can derive that if $\sqrt{n}(\lambda - \beta_j^*) < 0$ or $\sqrt{n}(\lambda-\beta_j^*) = o(1)$, the limit of (4.27) with respect to n is at least $(1-\alpha)$; however, if $\sqrt{n}(\lambda-\beta_j^*) \to 0$, then the probability in (4.27) decreases to 0, which means $I_1(j;\hat{\beta},S,\mathcal{A})$ can never be a $(1-\alpha)$ confidence interval for β_j^* over $\mathbb{R}$ conditioning on $\mathcal{A} = A$. Applying a similar technique to $I_2(j;\hat{\beta},S,\mathcal{A})$. It follows from (4.23) that, if $\beta_j^* - \lambda > 0$,

$$\begin{aligned}&-\frac{\sigma^2}{n\hat{\beta}_j}\gamma_{\alpha/2}^{(2,1)} + \lambda \le \beta_j^* \le \hat{\beta}_j + \lambda + \frac{\sigma}{\sqrt{n}}q_{\frac{\alpha}{2}} \\ \Leftrightarrow &\max\left(\frac{-\sigma^2\gamma_{\alpha/2}^{(2,1)}}{n(\beta_j^*-\lambda)} - (\beta_j^* - \lambda),\ -\frac{\sigma}{\sqrt{n}}q_{\alpha/2}\right) \le \frac{\sigma}{\sqrt{n}}\epsilon, \end{aligned}$$

and then (4.25) for $k = 2$ becomes

$$\begin{aligned}
&\frac{\mathbb{P}\left\{\max\left(\frac{-\sigma^2\gamma^{(2,1)}_{\alpha/2}}{n(\beta^*_j-\lambda)} - (\beta^*_j - \lambda),\ -\frac{\sigma}{\sqrt{n}}q_{\alpha/2},\ -(\beta^*_j-\lambda)\right) \le \frac{\sigma}{\sqrt{n}}\epsilon_j\right\}}{\mathbb{P}\{\epsilon_j > -\frac{\sqrt{n}}{\sigma}(\beta^*_j-\lambda)\}} \\
&\quad \ge \frac{\mathbb{P}\{\max(-q_{\alpha/2}, -\frac{\sqrt{n}}{\sigma}(\beta^*_j-\lambda)) \le \epsilon_j\}}{\mathbb{P}\{\epsilon_j > -\frac{\sqrt{n}}{\sigma}(\beta^*_j-\lambda)\}} \\
&\quad \ge 1-\alpha.
\end{aligned}$$

On the other hand, if $\beta^*_j - \lambda < 0$, the event (4.23) is

$$-\frac{\sigma}{\sqrt{n}}q_{\alpha/2} \le \frac{\sigma}{\sqrt{n}}\epsilon \le \frac{\sigma^2\gamma^{(2,1)}_{\alpha/2}}{n(\lambda-\beta^*_j)} + (\lambda-\beta^*_j)$$

and the probability (4.25) is

$$\frac{\mathbb{P}\left\{(\lambda-\beta^*_j) \le \frac{\sigma}{\sqrt{n}}\epsilon_j \le \frac{\sigma^2\gamma^{(2,1)}_{\alpha/2}}{n(\lambda-\beta^*_j)} + (\lambda-\beta^*_j)\right\}}{\mathbb{P}\{\epsilon_j > \frac{\sqrt{n}}{\sigma}(\lambda-\beta^*_j)\}} \ge const,$$

where *const* is a constant only depending on $\gamma^{(2,1)}_{\alpha/2}$ and the inequality uniformly holds for all $\beta^*_j - \lambda < 0$. If $\beta^*_j - \lambda = 0$, we can obtain a similar conclusion that (4.25) converges to 1. Consequently, we have proved that the credible interval $I_2(j;\hat{\beta},S,\mathcal{A})$ for $[\beta_j|\hat{\beta}_{\mathcal{A}}, S_{\mathcal{I}}, \mathcal{A}]$ achieves a "post-selection version" of honesty over $\mathbb{R}$ conditioning on $\mathcal{A} = A$, if $\gamma^{(2,1)}_{\alpha'/2}$ is carefully chosen, i.e.,

$$\liminf_{n\to\infty}\lim_{\beta^*_j\in\mathbb{R}}\mathbb{P}\left\{\beta^*_j \in I_2(j;b,s,A)|\mathcal{A}=A\right\} \ge 1-\alpha.$$

Through the two examples, we illustrate the idea of constructing credible intervals and how to prove that they are also confidence intervals in the frequentist view. A credible interval, like $I_1(j;\hat{\beta},S,\mathcal{A})$ which shrinks to a point, may not be honest over $\mathbb{R}$, while a credible interval, like $I_2(j;\hat{\beta},S,\mathcal{A})$ which achieves honesty, may not have its length converge to 0. As a result, there are many general questions raised from this preliminary study. Does there exist an honest confidence set for $[\beta|\hat{\beta}_{\mathcal{A}}, S_{\mathcal{I}}, \mathcal{A}]$ over the full parameter space $\mathbb{R}^p$ with the diameter converging to 0 conditioning on $\mathcal{A} = A$? If there exists, how fast the convergence of its diameter could be? If not, is the diameter unbounded and whether can

we construct an honest confidence set with the diameter converging to 0 after removing a small region from $\mathbb{R}^p$? We are still researching on these problems as well as the design of the prior $\pi(z)$ for β in Bayesian framework. Lastly, we look at another example based on our simulation.

In simulation, $p = 3$, n took values of 50 or 500 and β_j^* took values between 0 and 1. The tuning parameter λ was taken by the minimum theoretical value in Bickel et al. (2009), $\lambda_{val} = 2\sqrt{2}\sigma\sqrt{\log p/n}$. First, we only generated y, for which $\hat{\beta}_j > 0$. Next, we derived the density function of $[\beta_j|\hat{\beta}_{\mathcal{A}}, S_{\mathcal{I}}, \mathcal{A}]$ from (X, y) and applied the Metropolis–Hastings algorithm to obtain a sequence of random samples from the density. Finally, we constructed the credible intervals based on the 2.5th and 97.5th percentiles of the samples. Under each setting, 200 data sets were generated independently in order to present a reliable summary.

	β_j^*	length	coverage	$\mathbb{P}\{\hat{\beta}_j > 0\}$
$n = 50$	1	0.553	0.955	1.000
	0.1	1.886	0.875	0.365
	0.01	2.645	0.800	0.163
	0.005	3.189	0.750	0.154
	0.00001	3.234	0.735	0.146
$n = 500$	1	0.175	0.950	1.000
	0.1	0.739	0.810	0.137
	0.01	1.165	0.800	9.46×10^{-3}
	0.005	1.348	0.785	6.45×10^{-4}
	0.00001	1.427	0.765	4.34×10^{-4}

Table 4.1: Credible intervals conditioning on $\hat{\beta}_j > 0$.

The result is summarized in Table 4.1. We considered the length of credible intervals, the coverage rate and the probability of the event conditioned on. Clearly, as n was fixed and the true value $\hat{\beta}_j^*$ decreased from 1 to 0, the length of the credible intervals increased but the coverage dropped. Surprisingly, a wider credible interval even failed to maintain

the nominal significance level when β_j^* was small, which indicates that a non-trivial prior must be assumed to "regularize" the credible intervals. One can see the difference between general statistical inference and post-selection inference from the last column. If $\hat{\beta}^*$ was large, $\mathbb{P}\{\hat{\beta}_j > 0\}$ was close to 1 so the posterior distributions of $[\beta_j|\hat{\beta}_{\mathcal{A}}, S_{\mathcal{I}}, \mathcal{A}, \hat{\beta}_j > 0]$ and $[\beta_j|y, X]$ were close. This is the reason why $I_1(j; b, s, A)$ maintains the nominal significance level for $\sqrt{n}(\lambda - \beta_j^*) < 0$. However, if $\hat{\beta}^*$ was small, post-selection inference considered a rare event, which could almost be ignored by the general statistical inference. See for example $n = 500$ and $\beta_j^* = 0.00001$. It turns out that constructing an honest credible interval becomes harder if a rarer event is conditioned on. Besides, the sample size n affects the values of $\mathbb{P}\{\hat{\beta}_j > 0\}$ and in turn affects the length and the coverage through (4.26). The increase of n from 50 to 500 helped construct smaller credible intervals and improve higher coverage rate. Thus, it comes to the question of how n and the value of β_j^* together affect the size and the honesty of the credible intervals under asymptotic framework. Many questions of this new framework are unclear and we are researching on it. See Chapter 5 for our future plan.

4.3 Estimator augmentation in GLMs

While we are studying the post-selection inference in linear models, we are also exploring the possibility to generalize the idea to GLMs. The key prerequisite of constructing credible sets conditioning on $\mathcal{A} = A$ is to derive a continuous posterior distribution of the parameters. The continuity of the posterior in turn depends on the continuity of the sampling distribution of $(\hat{\beta}, S)$. In linear regression, the sampling distribution of $(\hat{\beta}, S)$ is derived through the mappings $H(\hat{\beta}_{\mathcal{A}}, S_{\mathcal{I}}, \mathcal{A}; \beta)$ in (4.4) and $H_r(\hat{\beta}_{\mathcal{A}}, S_{\mathcal{I}}, \mathcal{A}; \beta)$ in (4.10). If the noise ε and thus y have a continuous distribution, this mapping allows us to derive the continuous distribution of the augmented estimator $(\hat{\beta}, S)$. However, for a generalized linear model, the response y is often a discrete variable. Such a mapping would result in a discrete distribution of $(\hat{\beta}, S)$, which may not be desirable for making inference. Thus, we present a novel framework to solve this problem.

4.3.1 Decision-theoretic framework

Let $x_i = (x_{i1}, x_{i2}, \ldots, x_{ip})^\mathsf{T} \in \mathbb{R}^p$ be covariates for a response $y_i \in \mathbb{R}$ for $i = 1, 2, \ldots, n$. Consider a generalized linear model with a link function $\psi(\mathbb{E}[y_i|x_i]) = x_i^\mathsf{T}\beta$. Suppose the probability density or probability mass function of $[y_i|x_i]$ is $p(y_i|x_i^\mathsf{T}\beta)$. Define the prediction loss by the negative log-likelihood

$$h(y_i, x_i^\mathsf{T}\beta) := -\log p(y_i|x_i^\mathsf{T}\beta). \tag{4.28}$$

As two well-known examples, for logistic regression and Poisson regression, we have, respectively

$$\begin{aligned} h(y_i, x_i^\mathsf{T}\beta) &= -y_i(x_i^\mathsf{T}\beta) + \log\left(1 + \exp(x_i^\mathsf{T}\beta)\right), \\ h(y_i, x_i^\mathsf{T}\beta) &= -y_i(x_i^\mathsf{T}\beta) + \exp(x_i^\mathsf{T}\beta). \end{aligned}$$

Then define an ℓ_1-penalized estimator

$$\hat{\beta}^P = \underset{\beta}{\operatorname{argmin}}\, L(\beta; y, X) = \underset{\beta}{\operatorname{argmin}} \left\{ \frac{1}{n}\sum_{i=1}^{n} h(y_i, x_i^\mathsf{T}\beta) + \lambda\|\beta\|_1 \right\}. \tag{4.29}$$

Since $\hat{\beta}^P$ is a function with respect to y, it will follow a discrete distribution due to the discreteness in y.

We propose a decision-theoretic framework to define a penalized estimator for which estimator augmentation can be developed for sampling and inference. Our approach has a Bayesian interpretation and regards the parameter β as a random vector. Let $\eta \in \mathbb{R}^p$ be a decision regarding β that incurs a penalized loss,

$$\ell_B(\eta; \beta) := \mathbb{E}_\beta L(\eta; y, X) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_\beta\left[h(y_i, x_i\eta)\right] + \lambda\|\eta\|_1, \tag{4.30}$$

which is the expectation of the ℓ_1-penalized loss function in (4.29) with respect to the distribution $p(y_i|x_i^\mathsf{T}\beta)$.

Suppose that $h(y_i, x_i^\mathsf{T}\eta)$ is a linear function of y_i. It follows from (4.30) that

$$\ell_B(\eta; \beta) := \mathbb{E}_\beta L(\eta; y, X) = \frac{1}{n}\sum_{i=1}^{n} h(\psi^{-1}(x_i^\mathsf{T}\beta), x_i^\mathsf{T}\eta) + \lambda\|\eta\|_1, \tag{4.31}$$

by taking inverse of the link function $\psi(.)$ to find $\mathbb{E}(y_i|x_i)$. Here, the ℓ_1-norm is used to encourage a sparse optimal decision $\hat{\beta}$ by minimizing $\ell_B(\eta;\beta)$ over η for a given parameter β:

$$\hat{\beta} := \underset{\eta\in\mathbb{R}^p}{\operatorname{argmin}}\, \ell_B(\eta;\beta). \tag{4.32}$$

In many generalized linear models, it is easy to verify that $\ell_B(\eta)$ is convex in η. Thus, the minimizer $\hat{\beta}$ in (4.32) is characterized by KKT conditions, often in the form of

$$F(X\beta, X\hat{\beta}) + \lambda S = 0, \tag{4.33}$$

where S is the subgradient of $\|\eta\|_1$ at the minimizer $\hat{\beta}$ and $F : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}^n$ is a mapping. We call $(\hat{\beta}, S)$ the augmented estimator under this decision-theory framework. Then the KKT conditions (4.33) implicitly define a mapping $H : (\hat{\beta}, S) \to X\beta$, which plays a role similar to the mapping in (4.4) for linear regression.

Since β is a random vector in Bayesian inference, the posterior distribution of $[X\beta|y]$ determines the joint posterior distribution of the augmented estimator $(\hat{\beta}, S)$ via the above KKT conditions (4.33). Denote the posterior density by $p(\mu|y)$ for $\mu = X\beta \in \mathbb{R}^n$. In principle, a change of measure according to the mapping H will lead to the density of $[\hat{\beta}, S|y]$ at $(\hat{\beta}, S) = (b, s)$ in the form of $f(b, s|y) = p(H(b, s)|y)J(b, s)$, similar to (4.6) and (4.13), where $J(b, s)$ is the Jacobian for the mapping H. Since $\hat{\beta}$ can be sparse, due to the ℓ_1-regularizer in (4.31), we will employ the same triplet parameterization $(\hat{\beta}_{\mathcal{A}}, S_{\mathcal{I}}, \mathcal{A})$. Both the mapping H and the Jacobian J will depend on the active set $\mathcal{A}$ in the above.

When $p > n$, the posterior distribution of β under a commonly used prior is well-defined. However, what we need is instead the posterior distribution of $\mu = X\beta$, which actually exists under a few common choices of prior $\pi(\mu)$ over the mean vector. Examples of the prior include the non-informative and improper prior $\pi(\mu) \propto 1$ and a conjugate prior $\mu \sim \mathcal{N}_n(0, \tau_o^2\mathbf{I}_n)$ with τ_0 being a positive constant. To illustrate the idea of a conjugate prior, consider a Gaussian linear model for which the link function, $\psi(x) = x$, is the identity map. Since $y|\mu \sim \mathcal{N}_n(\mu, \sigma^2\mathbf{I}_n)$, it is easy to see that the posterior distribution $\mu|y \sim \mathcal{N}_n(y, \tau_n^2\mathbf{I}_n)$ under the conjugate prior, where $\tau_n = \tau_n^2(\tau_0^2, \sigma^2)$. In fact, under this prior, the estimator $\hat{\beta}$ in (4.32)

is expected to be close to that defined by (4.29) with an additional ℓ_2-regularizer like the elastic net (Zou and Hastie, 2005).

4.3.2 Exponential families

We extend Theorem 10 and Theorem 11 to generalized linear models for exponential families under the decision-theoretic framework discussed in Section 4.3.1. Consider a canonical form of a probability density (mass) function, for $i = 1, 2, \ldots, n$

$$f(y_i|\theta_i) = h(y_i, \tau) \exp\{\frac{\theta_i y_i - a(\theta_i)}{d(\tau)}\}, \tag{4.34}$$

where $y_i \in \mathbb{R}$ is a response variable, $\tau \in \mathbb{R}$ is a dispersion parameter and the canonical link function is $\theta_i = x_i^\mathsf{T}\beta$ for $x_i, \beta \in \mathbb{R}^p$. Under the canonical form, the space of θ, denoted as Θ, is a convex set and $a(\theta_i)$ is a convex function. Additionally, assume $a(\theta_i)$ is twice differentiable. Write the matrix form as $y = (y_1, \ldots, y_n)^\mathsf{T}$ and $X = (x_1, \ldots, x_n)^\mathsf{T}$.

The expectation of y_i is $\mathbb{E}[y_i|x_i^\mathsf{T}\beta] = a'(\theta_i) = a'(x_i^\mathsf{T}\beta)$, which is monotonically increasing with respect to θ_i (or $x_i^\mathsf{T}\beta$). Moreover, the variance is $\mathrm{Var}[y_i|x_i^\mathsf{T}\beta] = a''(\theta_i) = a''(x_i^\mathsf{T}\beta)$. The negative log-likelihood in (4.28) has the form of

$$h(y_i, x_i^\mathsf{T}\beta) = \frac{1}{d(\tau)}\left[-y_i(x_i^\mathsf{T}\beta) + a(x_i^\mathsf{T}\beta)\right] - \log h(y_i, \tau)$$

and the penalized loss $\ell_B(\eta; \beta)$ in (4.30) becomes

$$\ell_B(\eta; \beta) = \frac{1}{nd(\tau)} \sum_{i=1}^{n} \left[-\mathbb{E}[y_i|x_i^\mathsf{T}\beta](x_i^\mathsf{T}\eta) + a(x_i^\mathsf{T}\eta) - \mathbb{E}[\log(y_i, \tau)|x_i^\mathsf{T}\beta]\right] + \lambda \sum_{j=1}^{p} w_j|\eta_j|, \tag{4.35}$$

where $\lambda > 0$ and $w_j > 0$ for $j = 1, \ldots, p$. Naturally, $\ell_B(\eta; \beta)$ is a convex function by noting that it is the summation of convex functions with respect to η. However, one should be aware that the minimizer $\hat{\beta} = \mathrm{argmin}_\eta\, \ell_B(\eta; \beta)$ may not exist for some $X\beta \in \Theta$. For example, consider a logistic regression with $n = p = 1$. In this case, the penalized loss $\ell_B(\eta; \beta)$ is reduced to, up to a constant,

$$\ell_B(\eta_1; \beta_1) = -\frac{1}{1 + e^{-(x_{11}\beta_1)}}(x_{11}\eta_1) + \log(e^{x_{11}\eta_1} + 1) + \lambda w_1|\eta_1|,$$

where $x_{11}, \beta_1, \eta_1 \in \mathbb{R}$. When $\frac{e^{-(x_{11}\beta_1)}}{1+e^{-(x_{11}\beta_1)}} > \lambda w_1$, $\ell_B(\eta_1; \beta_1) \to -\infty$ with $\eta \to -\infty$. Therefore, unlike the linear model where any (X, y) gives at least one lasso solution, the nonexistence of $\hat{\beta}$ affects how we build the domain and the image of a bijective mapping later. Nevertheless, if such a $\hat{\beta}$ exists, $\hat{\beta}$ can be characterized by the KKT conditions

$$\frac{1}{nd(\tau)} \sum_{i=1}^{n} \left[-a'(x_i^\mathsf{T}\beta) + a'(x_i^\mathsf{T}\hat{\beta}) \right] x_i + \lambda W S = 0, \tag{4.36}$$

where $W = diag(w_1, w_2, \ldots, w_p)$ and S is the subgradient of $\|\eta\|_1$ at $\hat{\beta}$. Define $A'(\theta) = (a'(\theta_1), a'(\theta_2), \ldots, a'(\theta_n))^\mathsf{T}$. Rewrite (4.36) as a matrix expression

$$\frac{1}{n} X^\mathsf{T} A'(X\beta) = \frac{1}{n} X^\mathsf{T} A'(X\hat{\beta}) + \lambda d(\tau) W S. \tag{4.37}$$

If $y|x_i^\mathsf{T}\beta$ follows a normal distribution, the generalized linear model is exactly a linear model. In this case, $a(\theta_i) = \theta_i^2$ and (4.37) is in accordance with (4.2) with $\lambda = \lambda d(\tau)$.

Let $U = \frac{1}{n} X^\mathsf{T} A'(X\beta)$, which is a linear combination of the mean vector of y, and $\mathcal{A} = \text{supp}(\beta)$. Following the idea in Section 4.1, $(\hat{\beta}, S)$ can be equivalently represented by the triplet $(\hat{\beta}_\mathcal{A}, S_\mathcal{I}, \mathcal{A})$. Partitioning $\hat{\beta}$ as $(\hat{\beta}_\mathcal{A}, 0)$ and S as $(sgn(\hat{\beta}_\mathcal{A}), S_\mathcal{I})$, we can rewrite (4.37) as

$$U = \frac{1}{n} X^\mathsf{T} A'(X_\mathcal{A}\hat{\beta}_\mathcal{A}) + \lambda d(\tau) \left(W_\mathcal{A} sign(\hat{\beta}_\mathcal{A}) + W_\mathcal{I} S_\mathcal{I} \right) := H(\hat{\beta}_\mathcal{A}, S_\mathcal{I}, \mathcal{A}). \tag{4.38}$$

One main concern is to find Ω_1 and Ω_2, which are two respective subspaces of U and $(\hat{\beta}_\mathcal{A}, S_\mathcal{I}, \mathcal{A})$ such that the mapping $H : \Omega_2 \to \Omega_1$ is bijective. If there exists such Ω_1 and Ω_2, according to the Bayesian view, the posterior distribution of U can be learned from observational data and its prior. Subsequently, a posterior distribution of $(\hat{\beta}, S)$ can be derived through the bijection between U and $(\hat{\beta}, S)$. Finally, we can develop Monte Carlo algorithms to sample from the joint distribution $(\hat{\beta}, S)$ and obtain the sampling distribution of $\hat{\beta}$. We are presently looking for a general rule to find Ω_1 and Ω_2 due to the nonlinearity of $a'(\theta_i)$. Hereafter, we assume the bijection H can be found. Following the same idea in Section 4.1, we consider the low-dimensional setting and the high-dimensional setting separately.

4.3.3 Low-dimensional setting

In the low-dimensional setting ($n \geq p$), differentiating (4.38) with respect to $(\beta_{\mathcal{A}}, S_{\mathcal{I}}, \mathcal{A})$ gives

$$\begin{aligned} dU &= \frac{1}{n} X^{\mathsf{T}} A''(X_{\mathcal{A}}\hat{\beta}_{\mathcal{A}}) X_{\mathcal{A}} d\hat{\beta}_{\mathcal{A}} + \lambda d(\tau) W_{\mathcal{I}} dS_{\mathcal{I}}, \\ &= \left(\tfrac{1}{n} X^{\mathsf{T}} A''(X_{\mathcal{A}}\hat{\beta}_{\mathcal{A}}) X_{\mathcal{A}} \;\middle|\; \lambda d(\tau) W_{\mathcal{I}} \right) \begin{pmatrix} d\hat{\beta}_{\mathcal{A}} \\ dS_{\mathcal{I}} \end{pmatrix}, \end{aligned} \tag{4.39}$$

where $A''(X_{\mathcal{A}}\hat{\beta}_{\mathcal{A}}) = diag(a''(x_1^{\mathsf{T}}\hat{\beta}_{\mathcal{A}}), a''(x_2^{\mathsf{T}}\hat{\beta}_{\mathcal{A}})^{\mathsf{T}}, \ldots, a''(x_n^{\mathsf{T}}\hat{\beta}_{\mathcal{A}}))$ and $a''(x_i^{\mathsf{T}}\hat{\beta}_{\mathcal{A}})$ is essentially the variance of $y_i | x_i^{\mathsf{T}}\beta$. Denote the $p \times p$ matrix by $D_2(\mathcal{A})$. Permuting the rows of $D_2(\mathcal{A})$, one can see

$$\begin{aligned} |det(D_2(\mathcal{A}))| &= \begin{vmatrix} \frac{1}{n} X_{\mathcal{A}}^{\mathsf{T}} A''(X_{\mathcal{A}}\hat{\beta}_{\mathcal{A}}) X_{\mathcal{A}} & 0 \\ \frac{1}{n} X_{\mathcal{I}}^{\mathsf{T}} A''(X_{\mathcal{A}}\hat{\beta}_{\mathcal{A}}) X_{\mathcal{A}} & \lambda d(\tau) W_{\mathcal{I}\mathcal{I}} \end{vmatrix} \\ &= det(\frac{1}{n} X_{\mathcal{A}}^{\mathsf{T}} A''(X_{\mathcal{A}}\hat{\beta}_{\mathcal{A}}) X_{\mathcal{A}}) \, (\lambda d(\tau))^{|\mathcal{I}|} \prod_{j \in \mathcal{I}} w_j > 0 \end{aligned}$$

Now we can use the bijection H to derive the distribution of $(\hat{\beta}, S, \mathcal{A})$ from the distribution of U.

Proposition 12. *Assume there exists a bijective mapping $H : \Omega_2 \to \Omega_1$. Let f_U be the density of U over Ω_1. For $(b_A, s_I, A) \in \Omega_2$, the joint density distribution of $(\hat{\beta}, S, \mathcal{A})$ is given by*

$$\begin{aligned} \mathbb{P}\{\hat{\beta}_{\mathcal{A}} \in db_A, S_{\mathcal{I}} \in ds_I, \mathcal{A} = A\} &= f_U(H(\hat{\beta}_{\mathcal{A}}, S_{\mathcal{I}}, \mathcal{A}))|det(D_2(A))| \, db_A \, ds_I \\ &:= \pi(b_A, s_I, A) \, db_A \, ds_I, \end{aligned} \tag{4.40}$$

and the distribution of $(\hat{\beta}_{\mathcal{A}}, \mathcal{A})$ is a marginal distribution given by

$$\mathbb{P}\left\{\hat{\beta}_{\mathcal{A}} \in db_A, \mathcal{A} = A\right\} = \left[\int_{[-1,1]^{p-|A|}} \pi(b_A, s_I, A) \, ds_I\right] db_A.$$

Zhou (2014) enumerated two advantages of such a density $\pi(b_A, s_I, A)$. First, the density function has a closed-form expression without involving multidimensional integral as long as f_U is given. Second, the total dimension of $(\hat{\beta}, S)$ is p so that Monte Carlo algorithms avoid dealing with sampling spaces of different dimensions.

Remark 10. To be rigorous, (4.40) is derived by assuming (b_A, s_I) is an inner point of Ω_2 restricted to $\mathcal{A} = A$. This happens if and only if $|s_j| = 1$ for some $j \in I$ and the Lebesgue measure of the boundary is zero. Therefore, it causes no trouble when computing probability of any events.

Remark 11. f_U can be obtained in the Bayesian view. Suppose β has the prior $\pi(\beta)$. Then the posterior of β is

$$\pi(\beta|y) = \pi(\beta) \prod_{i=1}^{n} p(y_i|x_i^\mathsf{T}\beta; \tau).$$

Lastly, the posterior of U can be derived from $\pi(\beta|y)$ by the mapping $U = \frac{1}{n}X^\mathsf{T}A'(X\hat{\beta})$, if the inverse of $a'(.)$ exists and is differentiable. One may question why we still need (4.40) if $\pi(\beta|y)$ can be directly obtained. Note that the post-selection inference eventually conditions on $\mathcal{A} = A$, so it is easier to derive the conditional density from (4.40) by fixing $\mathcal{A} = A$ than to find the event $\{\beta : \mathcal{A}(\beta) = A\}$.

To help understanding the density π, we look at a simple example of linear regression. With the flat prior $\pi(\beta) \propto 1$, one can derive its posterior as

$$\pi(\beta|y) \propto \pi(\beta) \prod_{i=1}^{n} f(y_i|x_i^\mathsf{T}\beta) \propto \phi(\beta; (X^\mathsf{T}X)^{-1}X^\mathsf{T}y, \tau^2(X^\mathsf{T}X)^{-1}).$$

If $rank(X) = p$, then $U|y = \frac{1}{n}X^\mathsf{T}X\beta = \mathcal{N}_p(\frac{1}{n}X^\mathsf{T}y, \frac{\tau^2}{n}X^\mathsf{T}X)$ and (4.38) is simplified as

$$U = D(\mathcal{A}) \begin{pmatrix} \hat{\beta}_\mathcal{A} \\ S_\mathcal{I} \end{pmatrix} + \lambda W_\mathcal{A} sgn(\hat{\beta}_\mathcal{A}),$$

According to Proposition 12, we have

$$\pi(b_A, s_I, A|y) = \mathcal{N}_p(z; \mu(A, s_A; y), \Sigma(A; \tau^2)), \tag{4.41}$$

where

$$\mu(A, s_A; y) = D(A)^{-1}(\frac{1}{n}X^\mathsf{T}y - \lambda W_A sgn(b_A)) \tag{4.42}$$
$$\Sigma(A; \tau^2) = \frac{\tau^2}{n}D(A)^{-1}X^\mathsf{T}XD(A)^{-\mathsf{T}}.$$

This result is consistent with the result in (4.8), where a similar π to (4.41) is derived with $\frac{1}{n}X^\mathsf{T}X\beta$ in place of $\frac{1}{n}X^\mathsf{T}y$ in (4.42). If β is estimated by the least square estimator $\hat{\beta}^{(ls)} = (X^\mathsf{T}X)^{-1}X^\mathsf{T}y$, then (4.8) is exactly the same as (4.41).

4.3.4 High-dimensional setting

Under the high-dimensional setting, we assume $rank(X) = n < p$ and use the same strategy in Section 4.1. The mapping H in (4.38) is restricted to the inverse of $\Omega_2 \cap row(X)$. Under certain conditions, the new mapping is bijective. Then, left multiply (4.38) by V_R^T and V_N^T, respectively, to get

$$\begin{aligned} V_R^\mathsf{T} X^\mathsf{T} A'(X\beta) &= V_R^\mathsf{T} X^\mathsf{T} A'(X_\mathcal{A}\hat{\beta}_\mathcal{A}) + \lambda d(\tau)\left(V_{R\mathcal{A}}^\mathsf{T} W_{\mathcal{A}\mathcal{A}} sign(\hat{\beta}_\mathcal{A}) + V_{R\mathcal{I}}^\mathsf{T} W_{\mathcal{I}\mathcal{I}} S_\mathcal{I}\right) \\ &:= H_r(\hat{\beta}_\mathcal{A}, S_\mathcal{I}, \mathcal{A}) \\ V_{N\mathcal{A}}^\mathsf{T} W_{\mathcal{A}\mathcal{A}} sign(\hat{\beta}_\mathcal{A}) &+ V_{N\mathcal{I}}^\mathsf{T} W_{\mathcal{I}\mathcal{I}} S_\mathcal{I} = 0. \end{aligned}$$

Under certain conditions, we can differentiate both equations with respect to $(\beta_\mathcal{A}, S_\mathcal{I}, \mathcal{A})$ to gain

$$\begin{aligned} & V_{\mathcal{I}N}^\mathsf{T} W_{\mathcal{I}\mathcal{I}} dS_\mathcal{I} = 0, \\ & V_R^\mathsf{T} X^\mathsf{T} d(A'(X\beta)) = V_R^\mathsf{T} X^\mathsf{T} d(A'(X_\mathcal{A}\hat{\beta}_\mathcal{A})) + \lambda d(\tau) V_{\mathcal{I}R}^\mathsf{T} W_{\mathcal{I}\mathcal{I}} B(\mathcal{I}) d\tilde{S} \\ & \qquad\qquad = V_R^\mathsf{T} X^\mathsf{T} A''(X_\mathcal{A}\hat{\beta}_\mathcal{A}) X_\mathcal{A} d\hat{\beta}_\mathcal{A} + \lambda\tau V_{\mathcal{I}R}^\mathsf{T} W_{\mathcal{I}\mathcal{I}} B(\mathcal{I}) d\tilde{S} \\ & \Rightarrow d(A'(X\beta)) = A''(X_\mathcal{A}\hat{\beta}_\mathcal{A}) X_\mathcal{A} d\hat{\beta}_\mathcal{A} + \lambda d(\tau)(V_R^\mathsf{T} X^\mathsf{T})^{-1} V_{\mathcal{I}R}^\mathsf{T} W_{\mathcal{I}\mathcal{I}} B(\mathcal{I}) d\tilde{S}, \end{aligned} \tag{4.43}$$

where $B(\mathcal{I}) \in \mathbb{R}^{|\mathcal{I}|\times(n-|\mathcal{A}|)}$ is an orthonormal basis for $null(V_{\mathcal{I}N}^\mathsf{T} W_{\mathcal{I}\mathcal{I}})$ and $d\tilde{S}$ denote coordinates of $dS_\mathcal{I}$ with respect to $B(\mathcal{I})$. In the end, we derive the distribution of $(\hat{\beta}_\mathcal{A}, S_\mathcal{I}, \mathcal{A})$. Let $T_2(\mathcal{A}) = \left(A''(X_\mathcal{A}\hat{\beta}_\mathcal{A}) X_\mathcal{A} \;\middle|\; \lambda d(\tau)(V_R^\mathsf{T} X^\mathsf{T})^{-1} V_{\mathcal{I}R}^\mathsf{T} W_{\mathcal{I}\mathcal{I}} B(\mathcal{I})\right)$.

Proposition 13. *Assume there exists a bijective mapping H_r. Let f_μ be the density of the mean vector $\mu = A'(\theta)$. The joint density distribution of $(\hat{\beta}, S, \mathcal{A})$ is given by*

$$\mathbb{P}\{\hat{\beta}_\mathcal{A} \in db_A, S_\mathcal{I} \in ds_I, \mathcal{A} = A\} = f_\mu(H_r(\hat{\beta}_\mathcal{A}, S_\mathcal{I}, \mathcal{A}))|det(T_2(\mathcal{A}))|\, db_A\, ds_I,$$

for (b_A, s_I, A) in the sample space.

4.4 Post-selection inference with blocked lasso

Zhou and Min (2017) generalized estimator augmentation to blocked lasso, which provides another direction to generalize our proposed post-selection framework. Regarding this prob-

lem, we still consider the linear model (4.1). However, instead of the lasso estimator, we consider a so-called *block lasso estimator* defined via block norm regularization. Partition the predictors β into J disjoint groups $\mathcal{G}_j \subseteq [p]$ for $j = 1, \ldots, J$. Let β_j denote the jth component of β and $\beta_{(j)} = (\beta_k)_{k \in \mathcal{G}_j}$ denote the jth group. The block lasso is defined by minimizing a penalized loss function $L(\beta; \alpha)$:

$$\hat{\beta} \in \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} L(\beta; \alpha) = \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \left\{ \frac{1}{2n} \|y - X\beta\|^2 + \lambda + \sum_{j=1}^{J} w_j \|\beta_{(j)}\|_\alpha, \right\} \tag{4.44}$$

where $\|.\|_\alpha$ is the ℓ_α-norm and the weights $w_j > 0$ usually depend on the group size. If $\alpha = 1$ and $\beta_{(j)} = \beta_j$, the expression (4.44) is reduced to the lasso problem.

Similar to Section 4.1 and Section 4.3, we work with an augmented estimator $(\hat{\beta}, S)$, where S is the subgradient of $\hat{\beta}$. Let α^* be conjugate to α in the sense that $\frac{1}{\alpha} + \frac{1}{\alpha^*} = 1$. One distinct difference is that S is defined by $\eta = \eta(x) = sgn(x)|x|^{\alpha^*/\alpha}$ and

$$\begin{cases} S_{(j)} = \eta^{-1}(\hat{\beta}_{(j)} / \|\hat{\beta}_{(j)}\|_\alpha) & \text{if } \hat{\beta}_{(j)} \neq 0, \\ \|S_{(j)}\|_{\alpha^*} \leq 1 & \text{if } \hat{\beta}_{(j)} = 0. \end{cases}$$

Further, let $C = \frac{1}{n} X^\mathsf{T} X$ be the Gram matrix. By the KKT conditions, we obtain

$$U = C\hat{\beta} + \lambda W S - C\beta_0, \tag{4.45}$$

where $W = diag(w_1 \mathbf{I}_{p_1}, \ldots, w_J \mathbf{I}_{p_J})$ and $p_j = |\mathcal{G}_j|$, $j = 1, \ldots, J$. Zhou and Min (2017) proved that under certain conditions, a bijective mapping defined by (4.45) exists so that the closed-form density of the distribution of $(\hat{\beta}, S)$ can be derived from U. In this case, $(\hat{\beta}, S)$ can also be equivalently expressed by a triplet $(\hat{\beta}_\mathcal{A}, S_\mathcal{I}, \mathcal{A})$, where $\mathcal{A} \subseteq [J]$ is the active group set and $\mathcal{I} = [j] \setminus \mathcal{A}$ is the inactive group set. Lastly, the authors derived the closed-form density of $(\hat{\beta}_\mathcal{A}, S_\mathcal{I}, \mathcal{A})$ in their theorems. The closed-form density $\pi(\hat{\beta}_\mathcal{A}, S_\mathcal{I}, \mathcal{A}; \beta)$ is useful in post-selection inference. Based on our framework, we can design a prior of β and then derive the conditional posterior of $[\beta | \hat{\beta}, S_\mathcal{I}, \mathcal{A}]$, from which credible sets can be constructed.

The combination of the augmented estimator of block lasso and our post-selection framework have broader application and more robust results. When the number of predictors p is

much greater than the number of observations n, the ordinary lasso estimate could be unstable, and thus a post-selection credible set could condition on a candidate set significantly different from the true candidate set. In that case, $\mathbb{P}\{\mathcal{A} = A\}$ could be small. As we can see from Table 4.1, conditioning on a rare event can usually increase the size of the credible set as well as lower the coverage rate. On the other hand, the block lasso can alleviate this issue, since essentially the event conditioned on by the block lasso estimator is less rarer than the event conditioned on by the ordinary lasso estimator.

CHAPTER 5

Summary and Discussion

We consider constructing joint and post-selection confidence sets for high-dimensional regression throughout this dissertation.

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 $\mu = X\beta$ in (2.1) 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 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 $\|\beta\|_0 \leq s$, 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. We have also demonstrated that our method works well even when the underlying β is dense, e.g. $\|\beta\|_0 \asymp n$, which is important for practical applications. See Bradic et al. (2018) for recent theoretical results on high-dimensional inference for non-sparse β.

Another 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, based on 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). Following this idea, we propose a new post-selection framework with estimator augmentation. This framework contains Bayesian interpretation and has great flexibility to design the prior and to construct the credible sets from the conditional posterior. However, many problems regarding its theoretical properties are unclear. Meanwhile, there are a lot of generalizations we can make for this framework. Our future work is summarized as follows:

- Under the assumption of the orthogonal design of X, we first need to figure out whether there exists a better credible interval for β_j than $I_2(j; \hat{\beta}, S, \mathcal{A})$ in (4.23) by designing a proper prior, in the sense of smaller length of the interval and reaching the nominal significance level. In what follows, we will generalize this result to any design matrices X in the low-dimensional setting and later in the high-dimensional setting. Note that the low-dimensional setting and the high-dimensional setting could be essentially different. In the low-dimensional setting, we start from assuming the prior of a p-dimensional random vector such as β and U in (4.4). In contrast, we assume a prior

of n-dimensional random vector such as R in (4.10) in the high-dimensional setting, since in this case, the subgradient S of $\hat{\beta}$ only lies in a subspace of $\mathbb{R}^p$ by (4.12).

- We propose a decision-theoretic framework for generalized linear models to overcome the discreteness of observations. Though we can derive the differential equations in (4.39) and (4.43), the prerequisite is that there exists a bijective mapping between U in (4.4) and the triplet $(\hat{\beta}_{\mathcal{A}}, S_{\mathcal{I}}, \mathcal{A})$. We will work out a universal method to find such bijective mappings for a group of distributions, e.g, exponential families. Moreover, the post-selection framework with estimator augmentation needs to be justified in theory for GLMs. If a randomization step (Min and Zhou, 2019) is applied, we also need to construct the joint confidence set for the mean vector as the prior. Therefore, like the two-step Stein method, it is interesting to apply the idea of splitting β into strong and weak signals onto GLMs.

- Block lasso could be more useful than the ordinary lasso as $p \gg n$, so we can further develop the post-selection framework with estimator augmentation of the block lasso. Unlike the ordinary lasso, where the event $\{\mathcal{A} = A\}$ can be somehow represented by a truncate normal distribution (e.g., expression (4.26)), the event is more irregular for the block lasso, making it more challenging to conduct theoretical analysis.

Bibliography

Baraud, Y. (2004), "Confidence balls in Gaussian regression," *Ann. Statist.*, 32, 528–551.

Beran, R. and Dümbgen, L. (1998), "Modulation of estimators and confidence sets," *Ann. Statist.*, 26, 1826–1856.

Berk, R., Brown, L., Buja, A., Zhang, K., and Zhao, L. (2013), "Valid post-selection inference," *Ann. Statist.*, 41, 802–837.

Bickel, P. J., Ritov, Y., and Tsybakov, A. B. (2009), "Simultaneous analysis of Lasso and Dantzig selector," *Ann. Statist.*, 37, 1705–1732.

Bradic, J., Fan, J., and Zhu, Y. (2018), "Testability of high-dimensional linear models with non-sparse structures," *arXiv preprint arXiv:1802.09117*.

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.

Cai, T. T. and Guo, Z. (2017), "Confidence intervals for high-dimensional linear regression: Minimax rates and adaptivity," *Ann. Statist.*, 45, 615–646.

— (2020), "Semisupervised inference for explained variance in high dimensional linear regression and its applications," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*.

Cai, T. T. and Low, M. G. (2006), "Adaptive confidence balls," *Ann. Statist.*, 34, 202–228.

Carpentier, A. (2015), "Implementable confidence sets in high dimensional regression," in *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics*, eds. Lebanon, G. and Vishwanathan, S. V. N., San Diego, California, USA: PMLR, vol. 38 of *Proceedings of Machine Learning Research*, pp. 120–128.

Dezeure, R., Bühlmann, P., Meier, L., and Meinshausen, N. (2015), "High-Dimensional Inference: Confidence Intervals, p-Values and R-Software hdi," *Statist. Sci.*, 30, 533–558.

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

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.

Ewald, K. and Schneider, U. (2018), "Uniformly valid confidence sets based on the Lasso," *Electron. J. Statist.*, 12, 1358–1387.

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

Huang, J., Breheny, P., and Ma, S. (2012), "A Selective Review of Group Selection in High-Dimensional Models," *Statist. Sci.*, 27, 481–499.

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

Laurent, B. (1996), "Efficient estimation of integral functionals of a density," *Ann. Statist.*, 24, 659–681.

— (1997), "Estimation of integral functionals of a density and its derivatives," *Bernoulli*, 3, 181–211.

Laurent, B. and Massart, P. (2000), "Adaptive estimation of a quadratic functional by model selection," *Ann. Statist.*, 28, 1302–1338.

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

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

— (1989), "Honest Confidence Regions for Nonparametric Regression," *Ann. Statist.*, 17, 1001–1008.

Liu, K., Markovic, J., and Tibshirani, R. (2018), "More powerful post-selection inference, with application to the lasso," *arXiv preprint arXiv:1801.09037*.

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

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," *Statist. Sci.*, 27, 538–557.

Nickl, R. and van de Geer, S. (2013), "Confidence sets in sparse regression," *Ann. Statist.*, 41, 2852–2876.

Robins, J. and van der Vaart, A. (2006), "Adaptive nonparametric confidence sets," *Ann. Statist.*, 34, 229–253.

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

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

Sun, T. and Zhang, C.-H. (2012), "Scaled sparse linear regression," *Biometrika*, 99, 879–898.

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

Taylor, J. and Tibshirani, R. (2018), "Post-selection inference for -penalized likelihood models," *Canadian Journal of Statistics*, 46, 41–61.

Tian, X. and Taylor, J. (2017), "Asymptotics of Selective Inference," *Scandinavian Journal of Statistics*, 44, 480–499.

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

Tibshirani, R. J., Taylor, J., Lockhart, R., and Tibshirani, R. (2016), "Exact Post-Selection Inference for Sequential Regression Procedures," *Journal of the American Statistical Association*, 111, 600–620.

van de Geer, S., Bühlmann, P., Ritov, Y., and Dezeure, R. (2014), "On asymptotically optimal confidence regions and tests for high-dimensional models," *Ann. Statist.*, 42, 1166–1202.

van de Geer, S. and Bühlmann, P. (2009), "On the Conditions used to Prove Oracle Results for the Lasso," *Electron. J. Statist.*, 3.

Whittle, P. (1960), "Bounds for the Moments of Linear and Quadratic Forms in Independent Variables," *Theory of Probability & Its Applications*, 5, 302–305.

Zhang, C.-H. (2010), "Nearly unbiased variable selection under minimax concave penalty," *Ann. Statist.*, 38, 894–942.

Zhang, C.-H. and Huang, J. (2008), "The sparsity and bias of the Lasso selection in high-dimensional linear regression," *Ann. Statist.*, 36, 1567–1594.

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 (Statistical Methodology)*, 76, 217–242.

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

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

Zhou, Q. and Min, S. (2017), "Estimator augmentation with applications in high-dimensional group inference," *Electron. J. Statist.*, 11, 3039–3080.

Zou, H. and Hastie, T. (2005), "Regularization and variable selection via the elastic net," *Journal of the royal statistical society: series B (statistical methodology)*, 67, 301–320.