STATISTICAL INFERENCE FOR HIGH-DIMENSIONAL LINEAR MODELS

Zijian Guo

A DISSERTATION

in

Statistics

For the Graduate Group in Managerial Science and Applied Economics

Presented to the Faculties of the University of Pennsylvania

in

Partial Fulfillment of the Requirements for the

Degree of Doctor of Philosophy

2017

Supervisor of Dissertation

____

T. Tony Cai
Dorothy Silberberg Professor of Statistics

Graduate Group Chairperson

____

Catherine Schrand
Celia Z. Moh Professor, Professor of Accounting

Dissertation Committee:

Dylan S. Small, Class of 1965 Wharton Professor of Statistics
Hongzhe Li, Professor of Biostatistics and Statistics
Zongming Ma, Associate Professor of Statistics

STATISTICAL INFERENCE FOR HIGH-DIMENSIONAL LINEAR MODELS

Dedicated to loving memory of my grandmother.

Acknowledgments

First and foremost, I would like to express my sincere gratitude to my advisor, Professor Tony Cai for his strong support during my PhD study. I want to thank Tony for introducing the thesis topic to me, guiding me through the literature, having extensive discussions with me and giving me full freedom to think through the problems. I am particularly appreciative to Tony for his great patience and encouragement at the whole stage of my PhD study. Tony has not just provided academic support, but also shared his wisdom in both research and life. I could not have imagined having a better advisor and mentor for my PhD study.

I would also like to thank Professor Dylan Small for motivating me to pursue PhD in Statistics and supporting me academically and emotionally during both my undergraduate and PhD study. I met Dylan while I was an undergraduate student in his applied econometrics course. I am grateful to Dylan for being willing to supervise me when I was a junior student. His encouragement and appreciation of my undergraduate research motivate me to pursue a PhD in statistics. I want to thank Dylan for introducing me to causal inference, encouraging me to explore my research interests and providing support all the time.

I would also like to thank Professor Hongzhe Li and Professor Zongming Ma for being part of my thesis committee. To Hongzhe, thank you for introducing me to the world of statistical genetics and sharing with me the insights about the connection between statistics methods and genetics applications. To Zongming, thank you for teaching me many interesting techniques in your multivariate analysis course.

I would also like to thank Professor Jing Cheng for giving me the opportunity to be involved with mediation analysis in dental study. I am thankful to Jing and Dylan for their financial support of five-semester research assistantship.

My special thanks will also go to Hyunseung Kang, Xiaodong Li and Anru Zhang for helpful discussions and suggestions during my PhD study. To Hyunseung, it is said that only a few collaborators can write three or more papers together. You are one of them. To Xiaodong, I really appreciate your sharing your own wisdom in research. It encourages me to pursue my own research along the right direction. To Anru, thank you for always sharing your two-year ahead experience with me. It is fresh and useful.

I would also like to thank all my friends at Penn and other graduates in the department, whose companion has made my graduate life entertaining and enjoyable. Special thanks to Anao, Chao, Zhuang, Xin Lu, Sam and Junyang.

Finally, I would like to thank all of my family. Words cannot express my gratitude for all of your support and love. To my parents, you are my wonderful parents and wonderful friends. Thank you so much for your unconditional love and constant support! To Xiaocan, my wife, I am so fortunate to meet you at the beginning of my PhD study. Thank you for your companion, encouragement, support and love. To my parents-in-law, thank you for your understanding and support. All of you are my source of strength. I know you are always there for me.

ABSTRACT

STATISTICAL INFERENCE FOR HIGH-DIMENSIONAL LINEAR MODELS

Zijian Guo

T. Tony Cai

High-dimensional linear models play an important role in the analysis of modern data sets. Although the estimation problem has been well understood, there is still a paucity of methods and theories on the inference problem for high-dimensional linear models. This thesis focuses on statistical inference for high-dimensional linear models and consists of the following three parts.

- The first part of the thesis considers confidence intervals for linear functionals in high-dimensional linear regression. We first establish the convergence rates of the minimax expected length for confidence intervals. Furthermore, we investigate the problem of adaptation to sparsity for the construction of confidence intervals and identify the regimes in which it is possible to construct adaptive confidence intervals.

- In the second part of the thesis, we consider point and interval estimation of the ℓ_q loss of a given estimator in high-dimensional linear regression. For the class of rate-optimal estimators, we establish the minimax rates for estimating

their ℓ_q losses, the minimax expected length of confidence intervals for their ℓ_q losses and the possibility of adaptivity of confidence intervals for their ℓ_q losses.

- In the third part of the thesis, we consider the problem in the framework of high-dimensional instrumental variable regression and construct confidence intervals for the treatment effect in the presence of possibly invalid instrumental variables. We develop a novel selection procedure, Two-Stage Hard Thresholding (TSHT) to select valid instrumental variables and construct honest confidence intervals for the treatment effect using the selected instrumental variables.

Contents

List of Tables

List of Figures

1 Introduction

1.1 Literature review

Driven by a wide range of applications, the high-dimensional linear model, where the dimension p can be much larger than the sample size n, has received significant recent attention. The linear model is

$$y = X\beta + \epsilon, \quad \epsilon \sim N(0, \sigma^2 \mathrm{I}), \tag{1.1.1}$$

where $y \in \mathbb{R}^n$, $X \in \mathbb{R}^{n \times p}$ and $\beta \in \mathbb{R}^p$. This high-dimensional linear model has been well studied in the literature, where the main focus has been on estimation of β. Several penalized/constrained ℓ_1 minimization methods, including Lasso (Tibshirani, 1996), Dantzig selector (Candès & Tao, 2007), scaled Lasso (Sun & Zhang, 2012) and square-root Lasso (Belloni et al., 2011), have been proposed. These methods have been shown to work well in applications and produce interpretable estimates of β when β is assumed to be sparse. Theoretically, with a properly chosen tuning parameter, these estimators achieve the optimal rate of convergence over collections of sparse parameter spaces. See, for example, Candès & Tao (2007); Sun & Zhang (2012); Belloni et al. (2011); Raskutti et al. (2011); Bickel et al. (2009); Bühlmann &

van de Geer (2011); Verzelen (2012).

Confidence sets play a fundamental role in statistical inference. Recently, confidence sets for high-dimensional linear models have been actively studied, where the focus is on the construction of confidence intervals for individual coordinates (Javanmard & Montanari, 2014a; van de Geer et al., 2014) and the construction of confidence balls for the whole high-dimension vector β (Nickl & van de Geer, 2013). In addition, Gautier & Tsybakov (2011); Belloni et al. (2012); Fan & Liao (2014); Chernozhukov et al. (2015a) provide honest confidence intervals for a treatment effect in the framework of high-dimensional instrumental variable regression. However, compared to the estimation problem, there is still a paucity of methods and fundamental theoretical results on the inference problem for high-dimensional linear models. In this thesis, we will focus on the statistical inference problem in high-dimensional linear models. An outline of the thesis is presented in the next subsection.

1.2 Outline of the thesis

We consider the following three statistical inference problems in high-dimensional linear models.

Confidence intervals for linear functionals

In Chapter 2, we consider confidence intervals for linear functionals in high-dimensional linear regression with random design. We first establish the convergence rates of the minimax expected length for confidence intervals in the oracle setting where the sparsity parameter is given. The focus is then on the problem of adaptation to sparsity for the construction of confidence intervals. Ideally, an adaptive confidence interval should have its length automatically adjusted to the sparsity of the unknown regression vector, while maintaining a pre-specified coverage probability. It is shown that such a goal is in general not attainable, except when the sparsity parameter is restricted to a small region over which the confidence intervals have the optimal length of the usual parametric rate. It is further demonstrated that the lack of adaptivity is not due to the conservativeness of the minimax framework, but is fundamentally caused by the difficulty of learning the bias accurately.

This chapter is joint work with T. Tony Cai.

Accuracy assessment

In Chapter 3, we consider point and interval estimation of the ℓ_q loss of a given estimator in high-dimensional linear regression with random design. We establish the minimax rate for estimating the ℓ_q loss and the minimax expected length of confidence intervals for the ℓ_q loss of rate-optimal estimators of the regression vector, including commonly used estimators such as Lasso, scaled Lasso, square-root Lasso and Dantzig Selector. Adaptivity of confidence intervals for the ℓ_q loss is also studied. Both the setting of known identity design covariance matrix and known noise level and the setting of unknown design covariance matrix and unknown noise level are studied. The results reveal interesting and significant differences between estimating the ℓ_2 loss and ℓ_q loss with $1 \leq q < 2$ as well as between the two settings. New technical tools are developed to establish rate sharp lower bounds for the minimax estimation error and the expected length of minimax and adaptive confidence intervals for the ℓ_q loss. A significant difference between loss estimation and the traditional parameter estimation is that for loss estimation the constraint is on the performance of the estimator of the regression vector, but the lower bounds are on the difficulty of estimating its ℓ_q loss. The technical tools developed in this paper can also be of independent interest.

This chapter is joint work with T. Tony Cai.

Confidence intervals for treatment effects with invalid instruments

In Chapter 4, we consider the statistical inference problem in the high-dimensional instrumental variable framework with possibly invalid instruments. The instrumental variable (IV) method is commonly used to estimate the causal effect of a treatment on an outcome by using IVs that satisfy the assumptions of association with treatment, no direct effect on the outcome and ignorability. A major challenge in IV analysis is to find said IVs, but typically one is unsure of whether all of the putative IVs are in fact valid (i.e. satisfy the assumptions). We propose a general inference procedure that provides honest inference in the presence of invalid IVs, even after controlling for a large number of covariates. The key step of our method is a novel selection procedure, which we call Two-Stage Hard Thresholding (TSHT), where we use hard thresholding to select the set of non-redundant instruments in the first stage and subsequently use hard thresholding to select the set of valid instruments in the second stage among the set of instruments selected from the first stage. TSHT allows us to not only select valid IVs, but also provide honest confidence intervals of the treatment effect at $\sqrt{n}$ rate. We establish asymptotic properties of our procedure and demonstrate that our procedure performs well in simulation studies compared to traditional IV methods, especially when the instruments are invalid.

This chapter is joint work with Hyunseung Kang, T. Tony Cai and Dylan S. Small.

2

Confidence Intervals for High-Dimensional Linear Regression: Minimax Rates and Adaptivity

2.1 Introduction

Driven by a wide range of applications, high-dimensional linear regression, where the dimension p can be much larger than the sample size n, has received significant recent attention. The linear model is

$$y = X\beta + \epsilon, \quad \epsilon \sim N(0, \sigma^2 \mathrm{I}), \tag{2.1.1}$$

where $y \in \mathbb{R}^n$, $X \in \mathbb{R}^{n \times p}$ and $\beta \in \mathbb{R}^p$. Several penalized/constrained ℓ_1 minimization methods, including Lasso (Tibshirani, 1996), Dantzig Selector (Candès & Tao, 2007), square-root Lasso (Belloni et al., 2011), and scaled Lasso (Sun & Zhang, 2012) have been proposed and studied. Under regularity conditions on the design matrix X, these methods with a suitable choice of the tuning parameter have been shown to achieve the optimal rate of convergence $k\frac{\log p}{n}$ under the squared error loss over the set of k-sparse regression coefficient vectors with $k \leq c\frac{n}{\log p}$ where $c > 0$ is a constant.

That is, there exists some constant $C > 0$ such that

$$\sup_{\|\beta\|_0 \leq k} \mathbb{P}\left(\|\widehat{\beta} - \beta\|_2^2 > Ck\frac{\log p}{n} \right) = o(1), \tag{2.1.2}$$

where $\|\beta\|_0$ denotes the number of the nonzero coordinates of a vector $\beta \in \mathbb{R}^p$. See, for example, Verzelen (2012); Bickel et al. (2009); Candès & Tao (2007); Sun & Zhang (2012). A key feature of the estimation problem is that the optimal rate can be achieved adaptively with respect to the sparsity parameter k.

Confidence sets play a fundamental role in statistical inference and confidence intervals for high-dimensional linear regression have been actively studied recently with a focus on inference for individual coordinates. But, compared to point estimation, there is still a paucity of methods and fundamental theoretical results on confidence intervals for high-dimensional regression. Zhang & Zhang (2014) was the first to introduce the idea of de-biasing for constructing a valid confidence interval for a single coordinate β_i. The confidence interval is centered at a low-dimensional projection estimator obtained through bias correction via score vector using the scaled Lasso as the initial estimator. Javanmard & Montanari (2014a); van de Geer et al. (2014) also used de-biasing for the construction of confidence intervals and van de Geer et al. (2014) established asymptotic efficiency for the proposed estimator. All the aforementioned papers, Zhang & Zhang (2014); Javanmard & Montanari (2014a); van de Geer et al. (2014), have focused on the ultra-sparse case where the sparsity $k \ll \frac{\sqrt{n}}{\log p}$ is assumed. Under such a sparsity condition, the expected length of the confidence intervals constructed in Zhang & Zhang (2014); Javanmard & Montanari (2014a); van de Geer et al. (2014) is at the parametric rate $\frac{1}{\sqrt{n}}$ and the procedures do not depend on the specific value of k.

Compared to point estimation where the sparsity condition $k \ll \frac{n}{\log p}$ is sufficient for estimation consistency (see equation (2.1.2)), the condition $k \ll \frac{\sqrt{n}}{\log p}$ for valid confidence intervals is much stronger. There are several natural questions: What happens in the region where $\frac{\sqrt{n}}{\log p} \lesssim k \lesssim \frac{n}{\log p}$? Is it still possible to construct a valid confidence interval for β_i in this case? Can one construct an adaptive honest confidence interval not depending on k?

The goal of the present paper is to address these and other related questions on confidence intervals for high-dimensional linear regression with random design. More specifically, we consider construction of confidence intervals for a linear functional $\mathrm{T}(\beta) = \xi^{\intercal}\beta$, where the loading vector $\xi \in \mathbb{R}^p$ is given and $\frac{\max_{i \in \mathrm{supp}(\xi)} |\xi_i|}{\min_{i \in \mathrm{supp}(\xi)} |\xi_i|} \leq \bar{c}$ with $\bar{c} \geq 1$ being a constant. Based on the sparsity of ξ, we focus on two specific regimes: the sparse loading regime where $\|\xi\|_0 \leq Ck$, with $C > 0$ being a constant; the dense loading regime where $\|\xi\|_0$ satisfying (2.2.7) in Section 2.2. It will be seen later that for confidence intervals, $\mathrm{T}(\beta) = \beta_i$ is a prototypical case for the general functional $\mathrm{T}(\beta) = \xi^{\intercal}\beta$ with a sparse loading ξ, and $\mathrm{T}(\beta) = \sum_{i=1}^{p} \beta_i$ is a representative case for $\mathrm{T}(\beta) = \xi^{\intercal}\beta$ with a dense loading ξ.

To illustrate the main idea, let us first focus on the two specific functionals $\mathrm{T}(\beta) = \beta_i$ and $\mathrm{T}(\beta) = \sum_{i=1}^{p} \beta_i$. We establish the convergence rate of the minimax expected length for confidence intervals in the oracle setting where the sparsity parameter k is given. It is shown that in this case the minimax expected length is of order $\frac{1}{\sqrt{n}} + k\frac{\log p}{n}$ for confidence intervals of β_i. An honest confidence interval, which depends on the sparsity k, is constructed and is shown to be minimax rate optimal. To the best of our knowledge, this is the first construction of confidence intervals in the moderate-sparse region $\frac{\sqrt{n}}{\log p} \ll k \lesssim \frac{n}{\log p}$. If the sparsity k falls into the ultra-sparse region $k \lesssim \frac{\sqrt{n}}{\log p}$, the constructed confidence interval is similar to the confidence intervals constructed in Zhang & Zhang (2014); Javanmard & Montanari (2014a); van de Geer et al. (2014). On the other hand, the convergence rate of the minimax expected length of honest confidence intervals for $\sum_{i=1}^{p} \beta_i$ in the oracle setting is shown to be $k\sqrt{\frac{\log p}{n}}$. A rate-optimal confidence interval that also depends on k is constructed. It should be noted that this confidence interval is not based on the de-biased estimator.

One drawback of the constructed confidence intervals mentioned above is that they require a prior knowledge of the sparsity k. Such knowledge of sparsity is usually unavailable in applications. A natural question is: Without knowing the sparsity k, is it possible to construct a confidence interval as good as when the sparsity k is known? This is a question about adaptive inference, which has been a major goal in nonparametric and high-dimensional statistics. Ideally, an adaptive confidence interval should have its length automatically adjusted to the true sparsity of the unknown regression vector, while maintaining a prespecified coverage probability. We show that, in marked contrast to point estimation, such a goal is in general not attainable for confidence intervals. In the case of confidence intervals for β_i, it is impossible to adapt between different sparsity levels, except when the sparsity k is restricted to the ultra-sparse region $k \lesssim \frac{\sqrt{n}}{\log p}$, over which the confidence intervals have the optimal length of the parametric rate $\frac{1}{\sqrt{n}}$, which does not depend on k. In the case of confidence intervals for $\sum_{i=1}^{p} \beta_i$, it is shown that adaptation to the sparsity is not possible at all, even in the ultra-sparse region $k \lesssim \frac{\sqrt{n}}{\log p}$.

Minimax theory is often criticized as being too conservative as it focuses on the worst case performance over a large parameter space. For confidence intervals for high dimensional linear regression, we establish strong non-adaptivity results which demonstrate that the lack of adaptivity is not due to the conservativeness of the minimax framework. It shows that for any confidence interval with guaranteed coverage probability over the set of k sparse vectors, its expected length at any given point in a large subset of the parameter space must be at least of the same order as the minimax expected length. So the confidence interval must be long at a large subset of points in the parameter space, not just at a small number of "unlucky" points. This leads directly to the impossibility of adaptation over different sparsity levels. Fundamentally, the lack of adaptivity is caused by the difficulty in accurately learning the bias of any estimator for high-dimensional linear regression.

We now turn to confidence intervals for general linear functionals. For a linear functional $\xi^\intercal\beta$ in the sparse loading regime, the rate of the minimax expected length is $\|\xi\|_2\left(\frac{1}{\sqrt{n}}+k\frac{\log p}{n}\right)$, where $\|\xi\|_2$ is the vector ℓ_2 norm of ξ. For a linear functional $\xi^\intercal\beta$ in the dense loading regime, the rate of the minimax expected length is $\|\xi\|_\infty k\sqrt{\frac{\log p}{n}}$, where $\|\xi\|_\infty$ is the vector ℓ_∞ norm of ξ. Regarding adaptivity, the phenomena observed in confidence intervals for the two special linear functionals $\mathrm{T}(\beta)=\beta_i$ and $\mathrm{T}(\beta)=\sum_{i=1}^p \beta_i$ extend to the general linear functionals. The case of confidence intervals for $\mathrm{T}(\beta)=\sum_{i=1}^p \xi_i\beta_i$ with a sparse loading ξ is similar to that of confidence intervals for β_i in the sense that rate-optimal adaptation is impossible except when the sparsity k is restricted to the ultra-sparse region $k\lesssim\frac{\sqrt{n}}{\log p}$. On the other hand, the case for a dense loading ξ is similar to that of confidence intervals for $\sum_{i=1}^p \beta_i$: adaptation to the sparsity k is not possible at all, even in the ultra-sparse region $k\lesssim\frac{\sqrt{n}}{\log p}$.

In addition to the more typical setting in practice where the covariance matrix Σ of random design and the noise level σ of the linear model are unknown, we also consider the case with the prior knowledge of $\Sigma=\mathrm{I}$ and $\sigma=\sigma_0$. It turns out that this case is strikingly different. The minimax rate for the expected length in the sparse loading regime is reduced from $\|\xi\|_2\left(\frac{1}{\sqrt{n}}+k\frac{\log p}{n}\right)$ to $\frac{\|\xi\|_2}{\sqrt{n}}$, and in particular it does not depend on the sparsity k. Furthermore, in marked contrast to the case of unknown Σ and σ, adaptation to sparsity is possible over the full range $k\lesssim\frac{n}{\log p}$. On the other hand, for linear functionals $\xi^\intercal\beta$ with a dense loading ξ, the minimax rates and impossibility for adaptive confidence intervals do not change even with the prior knowledge of $\Sigma=\mathrm{I}$ and $\sigma=\sigma_0$. However, the cost of adaptation is reduced with the prior knowledge.

The rest of the paper is organized as follows: After basic notation is introduced, Section 2.2 presents a precise formulation for the adaptive confidence interval problem. Section 2.3 establishes the minimaxity and adaptivity results for a general linear functional $\xi^\intercal\beta$ with a sparse loading ξ. Section 2.4 focuses on confidence intervals for a general linear functional $\xi^\intercal\beta$ with a dense loading ξ. Section 2.5 considers the case when there is prior knowledge of covariance matrix of the random design and the noise level of the linear model. Section 2.6 discusses connections to other work and further research directions. The proofs of the main results are given in Section 2.7. More discussion and proofs are presented in Chapter A.

2.2 Formulation for adaptive confidence interval problem

We present in this section the framework for studying the adaptivity of confidence intervals. We begin with the notation that will be used throughout the paper.

2.2.1 Notation

For a matrix $X \in \mathbb{R}^{n\times p}$, $X_{i\cdot}$, $X_{\cdot j}$, and $X_{i,j}$ denote respectively the i-th row, j-th column, and (i,j) entry of the matrix X, $X_{i,-j}$ denotes the i-th row of X excluding the j-th coordinate, and X_{-j} denotes the submatrix of X excluding the j-th column. Let $[p] = \{1, 2, \cdots, p\}$. For a subset $J \subset [p]$, X_J denotes the submatrix of X consisting of columns $X_{\cdot j}$ with $j \in J$ and for a vector $x \in \mathbb{R}^p$, x_J is the subvector of x with indices in J and x_{-J} is the subvector with indices in J^c. For a set S, $|S|$ denotes the cardinality of S. For a vector $x \in \mathbb{R}^p$, $\mathrm{supp}(x)$ denotes the support of x and the ℓ_q norm of x is defined as $\|x\|_q = (\sum_{i=1}^{q} |x_i|^q)^{\frac{1}{q}}$ for $q \geq 0$ with $\|x\|_0 = |\mathrm{supp}(x)|$ and

$\|x\|_\infty = \max_{1\le j\le p} |x_j|$. We use e_i to denote the i-th standard basis vector in $\mathbb{R}^p$. For $a \in \mathbb{R}$, $a_+ = \max\{a, 0\}$. We use $\sum \beta_i$ as a shorthand for $\sum_{i=1}^p \beta_i$, $\max \|X_{\cdot j}\|_2$ as a shorthand for $\max_{1\le j\le p} \|X_{\cdot j}\|_2$ and $\min \|X_{\cdot j}\|_2$ as a shorthand for $\min_{1\le j\le p} \|X_{\cdot j}\|_2$. For a matrix A and $1 \le q \le \infty$, $\|A\|_q = \sup_{\|x\|_q=1} \|Ax\|_q$ is the matrix ℓ_q operator norm. In particular, $\|A\|_2$ is the spectral norm. For a symmetric matrix A, $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$ denote respectively the smallest and largest eigenvalue of A. We use c and C to denote generic positive constants that may vary from place to place. For two positive sequences a_n and b_n, $a_n \lesssim b_n$ means $a_n \le C b_n$ for all n and $a_n \gtrsim b_n$ if $b_n \lesssim a_n$ and $a_n \asymp b_n$ if $a_n \lesssim b_n$ and $b_n \lesssim a_n$, and $a_n \ll b_n$ if $\limsup_{n\to\infty} \frac{a_n}{b_n} = 0$ and $a_n \gg b_n$ if $b_n \ll a_n$.

2.2.2 Framework for adaptivity of confidence intervals

We shall focus in this paper on the high-dimensional linear model with the Gaussian design,

$$y_{n\times 1} = X_{n\times p}\beta_{p\times 1} + \epsilon_{n\times 1}, \quad \epsilon \sim N_n(0, \sigma^2 \mathrm{I}), \tag{2.2.1}$$

where the rows of X satisfy $X_{i\cdot} \overset{\text{i.i.d.}}{\sim} N_p(0, \Sigma)$, $i = 1, ..., n$, and are independent of ϵ. Both Σ and the noise level σ are unknown. Let $\Omega = \Sigma^{-1}$ denote the precision matrix. The parameter $\theta = (\beta, \Omega, \sigma)$ consists of the signal β, the precision matrix Ω for the random design, and the noise level σ. The target of interest is the linear functional of β, $\mathrm{T}(\beta) = \xi^\intercal \beta$, where $\xi \in \mathbb{R}^p$ is a pre-specified loading vector. The data that we observe is $Z = (Z_1, \cdots, Z_n)^\intercal$, where $Z_i = (y_i, X_i) \in \mathbb{R}^{p+1}$ for $i = 1, \cdots, n$.

For $0 < \alpha < 1$ and a given parameter space Θ and the linear functional $\mathrm{T}(\beta)$, denote by $\mathcal{I}_\alpha(\Theta, \mathrm{T})$ the set of all $(1-\alpha)$ level confidence intervals for $\mathrm{T}(\beta)$ over the parameter space Θ,

$$\mathcal{I}_\alpha(\Theta, \mathrm{T}) = \left\{ \mathrm{CI}_\alpha(\mathrm{T}, Z) = [l(Z), u(Z)] : \inf_{\theta \in \Theta} \mathbb{P}_\theta(l(Z) \leq \mathrm{T}(\beta) \leq u(Z)) \geq 1 - \alpha \right\}. \tag{2.2.2}$$

For any confidence interval $\mathrm{CI}_\alpha(\mathrm{T}, Z) \in \mathcal{I}_\alpha(\Theta, \mathrm{T})$, the maximum expected length over a parameter space Θ is defined as

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

where for confidence interval $\mathrm{CI}_\alpha(\mathrm{T}, Z) = [l(Z), u(Z)]$, $L(\mathrm{CI}_\alpha(\mathrm{T}, Z)) = u(Z) - l(Z)$ denotes its length. For two parameter spaces $\Theta_1 \subseteq \Theta$, we define the benchmark $L^*_\alpha(\Theta_1, \Theta, \mathrm{T})$ as the infimum of the maximum expected length over Θ_1 among all $(1-\alpha)$-level confidence intervals over Θ,

$$L^*_\alpha(\Theta_1, \Theta, \mathrm{T}) = \inf_{\mathrm{CI}_\alpha(\mathrm{T}, Z) \in \mathcal{I}_\alpha(\Theta, \mathrm{T})} L(\mathrm{CI}_\alpha(\mathrm{T}, Z), \Theta_1, \mathrm{T}). \tag{2.2.3}$$

We will write $L^*_\alpha(\Theta, \mathrm{T})$ for $L^*_\alpha(\Theta, \Theta, \mathrm{T})$, which is the minimax expected length of confidence intervals over Θ.

We should emphasize that $L^*_\alpha(\Theta_1, \Theta, \mathrm{T})$ is an important quantity that measures the degree of adaptivity over the nested spaces $\Theta_1 \subset \Theta$. A confidence interval $\mathrm{CI}_\alpha(\mathrm{T}, Z)$ that is (rate-optimally) adaptive over Θ_1 and Θ should have the optimal expected length performance simultaneously over both Θ_1 and Θ while maintaining a given coverage probability over Θ, i.e., $\mathrm{CI}_\alpha(\mathrm{T}, Z) \in \mathcal{I}_\alpha(\Theta, \mathrm{T})$ such that

$$L(\mathrm{CI}_\alpha(\mathrm{T}, Z), \Theta_1, \mathrm{T}) \asymp L^*_\alpha(\Theta_1, \mathrm{T}) \quad \text{and} \quad L(\mathrm{CI}_\alpha(\mathrm{T}, Z), \Theta, \mathrm{T}) \asymp L^*_\alpha(\Theta, \mathrm{T}).$$

Note that in this case $L(\mathrm{CI}_\alpha(\mathrm{T}, Z), \Theta_1, \mathrm{T}) \geq L^*_\alpha(\Theta_1, \Theta, \mathrm{T})$. So for two parameter spaces $\Theta_1 \subset \Theta$, if $L^*_\alpha(\Theta_1, \Theta, \mathrm{T}) \gg L^*_\alpha(\Theta_1, \mathrm{T})$, then rate-optimal adaptation between

Θ_1 and Θ is impossible to achieve.

We consider the following collection of parameter spaces,

$$\Theta(k) = \left\{ \theta = (\beta, \Omega, \sigma) : \|\beta\|_0 \leq k, \frac{1}{M_1} \leq \lambda_{\min}(\Omega) \leq \lambda_{\max}(\Omega) \leq M_1, 0 < \sigma \leq M_2 \right\}, \tag{2.2.4}$$

where $M_1 > 1$ and $M_2 > 0$ are positive constants. Basically, $\Theta(k)$ is the set of all k-sparse regression vectors. $\frac{1}{M_1} \leq \lambda_{\min}(\Omega) \leq \lambda_{\max}(\Omega) \leq M_1$ and $0 < \sigma \leq M_2$ are two mild regularity conditions on the design and the noise level.

The main goal of this paper is to address the following two questions:

1. *What is the minimax length $L_\alpha^*(\Theta(k), \mathrm{T})$ in the oracle setting where the sparsity level k is known?*

2. *Is it possible to achieve rate-optimal adaptation over different sparsity levels?* More specifically, for $k_1 \ll k$, is it possible to construct a confidence interval $\mathrm{CI}_\alpha(\mathrm{T}, Z)$ that is adaptive over $\Theta(k_1)$ and $\Theta(k)$ in the sense that $\mathrm{CI}_\alpha(\mathrm{T}, Z) \in \mathcal{I}_\alpha(\Theta(k), \mathrm{T})$ and

$$\begin{aligned} L(\mathrm{CI}_\alpha(\mathrm{T}, Z), \Theta(k_1), \mathrm{T}) &\asymp L_\alpha^*(\Theta(k_1), \mathrm{T}), \\ L(\mathrm{CI}_\alpha(\mathrm{T}, Z), \Theta(k), \mathrm{T}) &\asymp L_\alpha^*(\Theta(k), \mathrm{T})? \end{aligned} \tag{2.2.5}$$

We will answer these questions by analyzing the two benchmark quantities $L_\alpha^*(\Theta(k), \mathrm{T})$ and $L_\alpha^*(\Theta(k_1), \Theta(k), \mathrm{T})$. Both lower and upper bounds will be established. If (2.2.5) can be achieved, it means that the confidence interval $\mathrm{CI}_\alpha(\mathrm{T}, Z)$ can automatically adjust its length to the sparsity level of the true regression vector β. On the other hand, if $L_\alpha^*(\Theta(k_1), \Theta(k), \mathrm{T}) \gg L_\alpha^*(\Theta(k_1), \mathrm{T})$, then such a goal is not attainable.

For ease of presentation, we calibrate the sparsity level

$$k \asymp p^\gamma \quad \text{for some } 0 \leq \gamma < \tfrac{1}{2},$$

and restrict the loading ξ to the set

$$\xi \in \Xi\left(q, \bar{c}\right) = \left\{\xi \in \mathbb{R}^p : \|\xi\|_0 = q,\ \xi \neq \mathbf{0} \ \text{ and } \ \frac{\max_{j \in \mathrm{supp}(\xi)} |\xi_j|}{\min_{j \in \mathrm{supp}(\xi)} |\xi_j|} \leq \bar{c}\right\},$$

where $\bar{c} \geq 1$ is a constant. The minimax rate and adaptivity of confidence intervals for the general linear functional $\xi^\intercal \beta$ also depends on the sparsity of ξ. We are particularly interested in the following two regimes:

1. The sparse loading regime: $\xi \in \Xi\left(q, \bar{c}\right)$ with

$$q \leq Ck. \tag{2.2.6}$$

2. The dense loading regime: $\xi \in \Xi\left(q, \bar{c}\right)$ with

$$q = cp^{\gamma_q} \quad \text{with} \quad 2\gamma < \gamma_q \leq 1. \tag{2.2.7}$$

The behavior of the problem is significantly different in these two regimes. We will consider separately the sparse loading regime in Section 2.3 and the dense loading regime in Section 2.4.

2.3 Minimax rate and adaptivity of confidence intervals for sparse loading linear functionals

In this section, we establish the rates of convergence for the minimax expected length of confidence intervals for $\xi^\intercal \beta$ with a sparse loading ξ in the oracle setting where the sparsity parameter k of the regression vector β is given. Both minimax upper and lower bounds are given. Confidence intervals for $\xi^\intercal \beta$ are constructed and shown to be minimax rate-optimal in the sparse loading regime. Finally, we establish the possibility of adaptivity for the linear functional $\xi^\intercal\beta$ with a sparse loading ξ.

2.3.1 Minimax length of confidence intervals for $\xi^\intercal\beta$ in the sparse loading regime

In this section, we focus on the sparse loading regime defined in (2.2.6). The following theorem establishes the minimax rates for the expected length of confidence intervals for $\xi^\intercal\beta$ in the sparse loading regime.

Theorem 1. *Suppose that $0 < \alpha < \frac{1}{2}$ and $k \leq c\min\{p^\gamma, \frac{n}{\log p}\}$ for some constants $c > 0$ and $0 \leq \gamma < \frac{1}{2}$. If ξ belongs to the sparse loading regime (2.2.6), the minimax expected length for $(1-\alpha)$ level confidence intervals of $\xi^\intercal\beta$ over $\Theta(k)$ satisfies*

$$L_\alpha^*\left(\Theta\left(k\right), \xi^\intercal\beta\right) \asymp \|\xi\|_2\left(\frac{1}{\sqrt{n}} + k\frac{\log p}{n}\right). \tag{2.3.1}$$

Theorem 1 is established in two separate steps.

1. Minimax upper bound: we construct a confidence interval $\mathrm{CI}_\alpha^S\left(\xi^\intercal\beta, Z\right)$ such that $\mathrm{CI}_\alpha^S\left(\xi^\intercal\beta, Z\right) \in \mathcal{I}_\alpha\left(\Theta\left(k\right), \xi^\intercal\beta\right)$ and for some constant $C > 0$

$$L\left(\mathrm{CI}_\alpha^S\left(\xi^\intercal\beta, Z\right), \Theta\left(k\right), \xi^\intercal\beta\right) \leq C\|\xi\|_2\left(\frac{1}{\sqrt{n}} + k\frac{\log p}{n}\right). \tag{2.3.2}$$

2. Minimax lower bound: we show that for some constant $c > 0$

$$L_\alpha^*\left(\Theta\left(k\right), \xi^\intercal\beta\right) \geq c\|\xi\|_2\left(\frac{1}{\sqrt{n}} + k\frac{\log p}{n}\right). \tag{2.3.3}$$

The minimax lower bound is implied by the adaptivity result given in Theorem 2. We now detail the construction of a confidence interval $\mathrm{CI}_\alpha^S\left(\xi^\intercal\beta, Z\right)$ achieving the minimax rate (2.3.1) in the sparse loading regime. The interval $\mathrm{CI}_{\alpha}^{S}\left(\xi^{\intercal}\beta, Z\right)$ is centered at a de-biased scaled Lasso estimator, which generalizes the ideas used in Zhang & Zhang (2014); Javanmard & Montanari (2014a); van de Geer et al. (2014). The construction of the (random) length is different from the aforementioned papers as the asymptotic normality result is not valid once $k \gtrsim \frac{\sqrt{n}}{\log p}$.

Let $\{\widehat{\beta}, \hat{\sigma}\}$ be the scaled Lasso estimator with $\lambda_0 = \sqrt{\frac{2.05 \log p}{n}}$,

$$\{\widehat{\beta}, \hat{\sigma}\} = \underset{\beta \in \mathbb{R}^p, \sigma \in \mathbb{R}^+}{\arg\min} \frac{\|y - X\beta\|_2^2}{2n\sigma} + \frac{\sigma}{2} + \lambda_0 \sum_{j=1}^{p} \frac{\|X_{\cdot j}\|_2}{\sqrt{n}} |\beta_j|. \tag{2.3.4}$$

Define

$$\widehat{u} = \underset{u \in \mathbb{R}^p}{\arg\min} \left\{ u^{\intercal} \widehat{\Sigma} u : \|\widehat{\Sigma} u - \xi\|_{\infty} \leq \lambda_n \right\}, \tag{2.3.5}$$

where $\widehat{\Sigma} = \frac{1}{n} X^{\intercal} X$ and $\lambda_n = 12 \|\xi\|_2 M_1^2 \sqrt{\frac{\log p}{n}}$. The confidence interval $\mathrm{CI}_{\alpha}^{S}\left(\xi^{\intercal}\beta, Z\right)$ is centered at the following de-biased estimator

$$\widetilde{\mu} = \xi^{\intercal} \widehat{\beta} + \widehat{u}^{\intercal} \frac{1}{n} X^{\intercal} \left(y - X\widehat{\beta} \right), \tag{2.3.6}$$

where $\widehat{\beta}$ is the scaled Lasso estimator given in (2.3.4) and $\widehat{u}$ is defined in (2.3.5). Before specifying the length of the confidence interval, we review the following definition of restricted eigenvalue introduced in Bickel et al. (2009),

$$\kappa(X, k, \alpha_0) = \min_{\substack{J_0 \subset \{1, \cdots, p\}, \\ |J_0| \leq k}} \min_{\substack{\delta \neq 0, \\ \|\delta_{J_0^c}\|_1 \leq \alpha_0 \|\delta_{J_0}\|_1}} \frac{\|X\delta\|_2}{\sqrt{n} \|\delta_{J_0}\|_2}. \tag{2.3.7}$$

Define

$$\rho_1(k) = \|\xi\|_2 \hat{\sigma} \min \left\{ 1.01 \sqrt{\frac{\widehat{u}^{\intercal} \widehat{\Sigma} \widehat{u}}{n \|\xi\|_2^2}} z_{\alpha/2} + C_1(X, k)\, k \frac{\log p}{n},\ \log p \left(\frac{1}{\sqrt{n}} + \frac{k \log p}{n} \right) \right\}, \tag{2.3.8}$$

where $z_{\alpha/2}$ is the $\alpha/2$ upper quantile of the standard normal distribution and

$$C_1(X,k) = 7000 M_1^2 \frac{\sqrt{n}}{\min \|X_{\cdot j}\|_2} \max \left\{ 1.25, \frac{912 \max \|X_{\cdot j}\|_2^2}{n\kappa^2 \left(X, k, 405 \left(\frac{\max \|X_{\cdot j}\|_2}{\min \|X_{\cdot j}\|_2} \right) \right)} \right\}. \quad (2.3.9)$$

Define the event

$$A = \{\hat{\sigma} \leq \log p\}. \quad (2.3.10)$$

The confidence interval $\mathrm{CI}_\alpha^S(\xi^\intercal \beta, Z)$ for $\xi^\intercal \beta$ is defined as

$$\mathrm{CI}_\alpha^S(\xi^\intercal \beta, Z) = \begin{cases} [\widetilde{\mu} - \rho_1(k), \quad \widetilde{\mu} + \rho_1(k)] & \text{on } A \\ \{0\} & \text{on } A^c \end{cases} \quad (2.3.11)$$

It will be shown in Section 2.7 that the confidence interval $\mathrm{CI}_\alpha^S(\xi^\intercal \beta, Z)$ has the desired coverage property and achieves the minimax length in (2.3.1).

Remark 1. In the special case of $\xi = e_1$, the confidence interval defined in (2.3.11) is similar to the ones based on the de-biased estimators introduced in Zhang & Zhang (2014); Javanmard & Montanari (2014a); van de Geer et al. (2014). The second term $\widehat{u}^\intercal \frac{1}{n} X^\intercal \left(y - X\widehat{\beta} \right)$ in (2.3.6) is incorporated to reduce the bias of the scaled Lasso estimator $\widehat{\beta}$. The constrained estimator $\widehat{u}$ defined in (2.3.5) is a score vector u such that the variance term $u^\intercal \widehat{\Sigma} u$ is minimized and one component of the bias term $\|\widehat{\Sigma} u - \xi\|_\infty$ is constrained by the tuning parameter λ_n. The tuning parameter λ_n is chosen as $12\|\xi\|_2 M_1^2 \sqrt{\frac{\log p}{n}}$ such that $u = \Omega\xi$ lies in the constraint set $\|\widehat{\Sigma} u - \xi\|_\infty \leq \lambda_n$ in (2.3.5) with overwhelming probability. For $C_1(X,k)$ defined in (2.3.9), it will be shown that it is upper bounded by a constant with overwhelming probability.

2.3.2 Adaptivity of confidence intervals for $\xi^\intercal\beta$ in the sparse loading regime

We have constructed a minimax rate-optimal confidence interval for $\xi^\intercal\beta$ in the oracle setting where the sparsity k is assumed to be known. A major drawback of the construction is that it requires prior knowledge of k, which is typically unavailable in practice. An interesting question is whether it is possible to construct adaptive confidence intervals that have the guaranteed coverage and automatically adjust its length to k.

We now consider the adaptivity of the confidence intervals for $\xi^\intercal\beta$. In light of the minimax expected length given in Theorem 1, the following theorem provides an answer to the adaptivity question (2.2.5) for the confidence intervals for $\xi^\intercal\beta$ in the sparse loading regime.

Theorem 2. *Suppose that $0 < \alpha < \frac{1}{2}$ and $k_1 \leq k \leq c\min\left\{p^\gamma, \frac{n}{\log p}\right\}$ for some constants $c > 0$ and $0 \leq \gamma < \frac{1}{2}$. If ξ belongs to the sparse loading regime* (2.2.6)*, then there is some constant $c_1 > 0$ such that*

$$L_\alpha^*(\Theta(k_1), \Theta(k), \xi^\intercal\beta) \geq c_1\|\xi\|_2\left(\frac{1}{\sqrt{n}} + k\frac{\log p}{n}\right). \tag{2.3.12}$$

Note that Theorem 2 implies the minimax lower bound in Theorem 1 by taking $k_1 = k$. Theorem 2 rules out the possibility of rate-optimal adaptive confidence intervals beyond the ultra-sparse region. Consider the setting where $k_1 \ll k$ and $\frac{\sqrt{n}}{\log p} \ll k \lesssim \frac{n}{\log p}$. In this case,

$$L_\alpha^*(\Theta(k_1), \Theta(k), \xi^\intercal\beta) \asymp L_\alpha^*(\Theta(k), \xi^\intercal\beta) \asymp \|\xi\|_2 k\frac{\log p}{n} \gg L_\alpha^*(\Theta(k_1), \xi^\intercal\beta).$$

So it is impossible to construct a confidence interval that is adaptive simultaneously over $\Theta(k_1)$ and $\Theta(k)$ when $\frac{\sqrt{n}}{\log p} \ll k \lesssim \frac{n}{\log p}$ and $k_1 \ll k$. For sparse loading with $q \leq Ck_1$, the only possible region for adaptation is the ultra-sparse region $k \lesssim \frac{\sqrt{n}}{\log p}$, over which the optimal expected length of confidence intervals is of order $\frac{1}{\sqrt{n}}$ and in particular does not depend on the specific sparsity level. These facts are illustrated in Figure 2.1.

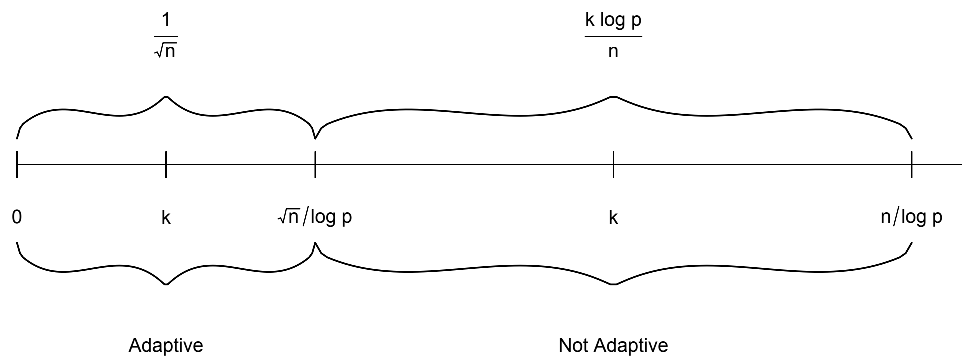

Figure 2.1: Illustration of adaptivity of confidence intervals for $\xi^{\intercal}\beta$ with a sparse loading ξ satisfying $\|\xi\|_0 \leq Ck_1$. For adaptation between $\Theta(k_1)$ and $\Theta(k)$ with $k_1 \ll k$, rate-optimal adaptation is possible if $k \lesssim \frac{\sqrt{n}}{\log p}$ and impossible otherwise.

So far the analysis is carried out within the minimax framework where the focus is on the performance in the worst case over a large parameter space. The minimax theory is often criticized as being too conservative. In the following, we establish a stronger version of the non-adaptivity result which demonstrates that the lack of adaptivity for confidence intervals is not due to the conservativeness of the minimax framework. The result shows that for any confidence interval $\mathrm{CI}_\alpha\left(\xi^{\intercal}\beta, Z\right)$, under the coverage constraint that $\mathrm{CI}_\alpha\left(\xi^{\intercal}\beta, Z\right) \in \mathcal{I}_\alpha\left(\Theta\left(k\right), \xi^{\intercal}\beta\right)$, its expected length at any given $\theta^* = (\beta^*, \mathrm{I}, \sigma) \in \Theta\left(k_1\right)$ must be of order $\|\xi\|_2 \left(\frac{1}{\sqrt{n}} + k\frac{\log p}{n}\right)$. So the confidence interval must be long at a large subset of points in the parameter space, not just at a small number of "unlucky" points.

Theorem 3. *Suppose that* $0 < \alpha < \frac{1}{2}$ *and* $k \leq c\min\{p^{\gamma}, \frac{n}{\log p}\}$ *for some constants* $c > 0$ *and* $0 \leq \gamma < \frac{1}{2}$. *Let* $k_1 \leq (1-\zeta_0)\, k - 1$ *and* $q \leq \frac{\zeta_0}{4} k$ *for some constant*

$0 < \zeta_0 < 1$. *Then for any* $\theta^* = (\beta^*, \mathrm{I}, \sigma) \in \Theta(k_1)$ *and* $\xi \in \Xi(q, \bar{c})$*, there is some constant* $c_1 > 0$ *such that*

$$\inf_{\mathrm{CI}_\alpha(\xi^\intercal\beta, Z) \in \mathcal{I}_\alpha(\Theta(k), \xi^\intercal\beta)} \mathbb{E}_{\theta^*} L\left(\mathrm{CI}_\alpha\left(\xi^\intercal\beta, Z\right)\right) \geq c_1 \|\xi\|_2 \left(k\frac{\log p}{n} + \frac{1}{\sqrt{n}}\right)\sigma. \tag{2.3.13}$$

Note that no supremum is taken over the parameter θ^* in (2.3.13). Theorem 3 illustrates that if a confidence interval $\mathrm{CI}_\alpha\left(\xi^\intercal\beta, Z\right)$ is "superefficient" at any point $\theta^* = (\beta^*, \mathrm{I}, \sigma) \in \Theta(k_1)$ in the sense that

$$\mathbb{E}_{\theta^*} L\left(\mathrm{CI}_\alpha\left(\xi^\intercal\beta, Z\right)\right) \ll \|\xi\|_2 \left(\frac{1}{\sqrt{n}} + k\frac{\log p}{n}\right)\sigma,$$

then the confidence interval $\mathrm{CI}_\alpha\left(\xi^\intercal\beta, Z\right)$ can not have the guaranteed coverage over the parameter space $\Theta(k)$.

2.3.3 Minimax rate and adaptivity of confidence intervals for β_1

We now turn to the special case $\mathrm{T}(\beta) = \beta_i$, which has been the focus of several previous papers, Zhang & Zhang (2014); Javanmard & Montanari (2014b,a); van de Geer et al. (2014). Without loss of generality, we consider β_1, the first coordinate of β, in the following discussion and the results for any other coordinate β_i are the same. The linear functional β_1 is the special case of linear functional of sparse loading regime with $\xi = e_1$.

Theorem 1 implies that the minimax expected length for $(1-\alpha)$ level confidence intervals of β_1 over $\Theta(k)$ satisfies

$$L^*_\alpha\left(\Theta(k), \beta_1\right) \asymp \frac{1}{\sqrt{n}} + k\frac{\log p}{n}. \tag{2.3.14}$$

In the ultra-sparse region with $k \lesssim \frac{\sqrt{n}}{\log p}$, the minimax expected length is of order $\frac{1}{\sqrt{n}}$. However, when k falls in the moderate-sparse region $\frac{\sqrt{n}}{\log p} \ll k \lesssim \frac{n}{\log p}$, the minimax expected length is of order $k\frac{\log p}{n}$ and in this case $k\frac{\log p}{n} \gg \frac{1}{\sqrt{n}}$. Hence the confidence intervals constructed in Zhang & Zhang (2014); Javanmard & Montanari (2014b,a); van de Geer et al. (2014), which are of parametric length $\frac{1}{\sqrt{n}}$, asymptotically have coverage probability going to 0. The condition $k \lesssim \frac{\sqrt{n}}{\log p}$ is thus necessary for the parametric rate $\frac{1}{\sqrt{n}}$. van de Geer et al. (2014) established asymptotic normality and asymptotic efficiency for a de-biased estimator under the sparsity assumption $k \ll \frac{\sqrt{n}}{\log p}$. Similar results have also been given in Ren et al. (2013) for a related problem of estimating a single entry of a p-dimensional precision matrix based on n i.i.d. samples under the same sparsity condition $k \ll \frac{\sqrt{n}}{\log p}$. It was also shown that $k \ll \frac{\sqrt{n}}{\log p}$ is necessary for the asymptotic normality and asymptotic efficiency results.

The following corollary, as a special case of Theorem 3, illustrates the strong non-adaptivity for confidence intervals of β_1 when $k \gg \frac{\sqrt{n}}{\log p}$.

Corollary 1. *Suppose that* $0 < \alpha < \frac{1}{2}$ *and* $k \leq c \min\{p^{\gamma}, \frac{n}{\log p}\}$ *for some constants* $c > 0$ *and* $0 \leq \gamma < \frac{1}{2}$. *Let* $k_1 \leq (1-\zeta_0)\, k - 1$ *for some constant* $0 < \zeta_0 < 1$. *Then for any* $\theta^* = (\beta^*, \mathrm{I}, \sigma) \in \Theta(k_1)$, *there is some constant* $c_1 > 0$ *such that*

$$\inf_{\mathrm{CI}_{\alpha}(\beta_1, Z) \in \mathcal{I}_{\alpha}(\Theta(k), \beta_1)} \mathbb{E}_{\theta^*} L\left(\mathrm{CI}_{\alpha}\left(\beta_1, Z\right)\right) \geq c_1 \left(\frac{1}{\sqrt{n}} + k\frac{\log p}{n}\right)\sigma. \tag{2.3.15}$$

2.4 Minimax rate and adaptivity of confidence intervals for dense loading linear functionals

We now turn to the setting where the loading ξ is dense in the sense of (2.2.7). We will also briefly discuss the special case $\sum_{i=1}^{p} \beta_i$ and the computationally feasible confidence intervals.

2.4.1 Minimax length of confidence intervals for $\xi^{\intercal}\beta$ in the dense loading regime

The following theorem establishes the minimax length of confidence intervals of $\xi^{\intercal}\beta$ in the dense loading regime (2.2.7).

Theorem 4. *Suppose that* $0 < \alpha < \frac{1}{2}$ *and* $k \leq c\min\{p^{\gamma}, \frac{n}{\log p}\}$ *for some constants* $c > 0$ *and* $0 \leq \gamma < \frac{1}{2}$. *If* ξ *belongs to the dense loading regime* (2.2.7), *the minimax expected length for* $(1-\alpha)$ *level confidence intervals of* $\xi^{\intercal}\beta$ *over* $\Theta(k)$ *satisfies*

$$L_{\alpha}^{*}\left(\Theta\left(k\right), \xi^{\intercal}\beta\right) \asymp \|\xi\|_{\infty} k \sqrt{\frac{\log p}{n}}. \tag{2.4.1}$$

Note that the minimax rate in (2.4.1) is significantly different from the minimax rate $\|\xi\|_2(\frac{1}{\sqrt{n}} + k\frac{\log p}{n})$ for the sparse loading case given in Theorem 1. In the following, we construct a confidence interval $\mathrm{CI}_{\alpha}^{D}\left(\xi^{\intercal}\beta, Z\right)$ achieving the minimax rate (2.4.1) in the dense loading regime. Define

$$C_2(X,k) = 822\frac{\sqrt{n}}{\min\|X_{\cdot j}\|_2} \max\left\{1.25, \frac{912\max\|X_{\cdot j}\|_2^2}{n\kappa^2\left(X, k, 405\left(\frac{\max\|X_{\cdot j}\|_2}{\min\|X_{\cdot j}\|_2}\right)\right)}\right\}. \tag{2.4.2}$$

It will be shown that $C_2(X,k)$ is upper bounded by a constant with overwhelming probability. The confidence interval $\mathrm{CI}_{\alpha}^{D}\left(\xi^{\intercal}\beta, Z\right)$ is defined to be,

$$\mathrm{CI}_{\alpha}^{D}\left(\xi^{\intercal}\beta, Z\right) = \begin{cases} \left[\xi^{\intercal}\widehat{\beta} - \|\xi\|_{\infty}\rho_2\left(k\right), \xi^{\intercal}\widehat{\beta} + \|\xi\|_{\infty}\rho_2\left(k\right)\right] & \text{on } A \\ \{0\} & \text{on } A^{c} \end{cases} \tag{2.4.3}$$

where A is defined in (2.3.10) and $\widehat{\beta}$ is the scaled Lasso estimator defined in (2.3.4) and

$$\rho_2\left(k\right) = \min\left\{C_2\left(X,k\right) k\sqrt{\frac{\log p}{n}}\hat{\sigma}, \log p\left(k\sqrt{\frac{\log p}{n}}\hat{\sigma}\right)\right\}. \tag{2.4.4}$$

The confidence interval constructed in (2.4.3) will be shown to have the desired coverage property and achieve the minimax length in (2.4.1). A major difference between the construction of $\mathrm{CI}_{\alpha}^{D}\left(\xi^{\intercal}\beta, Z\right)$ and that of $\mathrm{CI}_{\alpha}^{S}\left(\xi^{\intercal}\beta, Z\right)$ is that $\mathrm{CI}_{\alpha}^{D}\left(\xi^{\intercal}\beta, Z\right)$ is not centered at a de-biased estimator. If a de-biased estimator is used for the construction of confidence intervals for $\xi^{\intercal}\beta$ with a dense loading, its variance would be too large, which leads to a confidence interval with length much larger than the optimal length $\|\xi\|_{\infty} k\sqrt{\frac{\log p}{n}}$.

2.4.2 Adaptivity of confidence intervals for $\xi^{\intercal}\beta$ in the dense loading regime

In this section, we investigate the possibility of adaptive confidence intervals for $\xi^{\intercal}\beta$ in the dense loading regime. The following theorem leads directly to an answer to the adaptivity question (2.2.5) for confidence intervals for $\xi^{\intercal}\beta$ in the dense loading regime.

Theorem 5. *Suppose that* $0 < \alpha < \frac{1}{2}$ *and* $k_1 \leq k \leq c \min\left\{p^{\gamma}, \frac{n}{\log p}\right\}$ *for some constants* $c > 0$ *and* $0 \leq \gamma < \frac{1}{2}$*. If* ξ *belongs to the dense loading regime* (2.2.7)*, then there is some constant* $c_1 > 0$ *such that*

$$L_{\alpha}^{*}\left(\Theta\left(k_1\right), \Theta\left(k\right), \xi^{\intercal}\beta\right) \geq c_1 \|\xi\|_{\infty} k\sqrt{\frac{\log p}{n}}. \tag{2.4.5}$$

Theorem 5 implies the minimax lower bound in Theorem 4 by taking $k_1 = k$. If $k_1 \ll k$, (2.4.5) implies

$$L_{\alpha}^{*}\left(\Theta\left(k_1\right), \Theta\left(k\right), \xi^{\intercal}\beta\right) \geq c\|\xi\|_{\infty} k\sqrt{\frac{\log p}{n}} \gg L_{\alpha}^{*}\left(\Theta\left(k_1\right), \xi^{\intercal}\beta\right), \tag{2.4.6}$$

which shows that rate-optimal adaptation over two different sparsity levels k_1 and k

is not possible at all for any $k_1 \ll k$. In contrast, in the case of the sparse loading regime, Theorem 2 shows that it is possible to construct an adaptive confidence interval in the ultra-sparse region $k \lesssim \frac{\sqrt{n}}{\log p}$, although adaptation is not possible in the moderate-sparse region $\frac{\sqrt{n}}{\log p} \ll k \lesssim \frac{n}{\log p}$.

Similarly to Theorem 3, the following theorem establishes the strong non-adaptivity results for $\xi^\intercal \beta$ in the dense loading regime.

Theorem 6. *Suppose that $0 < \alpha < \frac{1}{2}$ and $k \leq c \min\{p^\gamma, \frac{n}{\log p}\}$ for some constants $c > 0$ and $0 \leq \gamma < \frac{1}{2}$. Let q satisfy (2.2.7) and $k_1 \leq (1 - \zeta_0) k - 1$ for some positive constant $0 < \zeta_0 < 1$. Then for any $\theta^* = (\beta^*, \mathrm{I}, \sigma) \in \Theta(k_1)$ and $\xi \in \Xi(q, \bar{c})$, there is some constant $c_1 > 0$ such that*

$$\inf_{\mathrm{CI}_\alpha(\xi^\intercal\beta, Z) \in \mathcal{I}_\alpha(\Theta(k), \xi^\intercal\beta)} \mathbb{E}_{\theta^*} L\left(\mathrm{CI}_\alpha\left(\xi^\intercal\beta, Z\right)\right) \geq c_1 \|\xi\|_\infty k \sqrt{\frac{\log p}{n}} \sigma. \tag{2.4.7}$$

2.4.3 Minimax length and adaptivity of confidence intervals for $\sum_{i=1}^p \beta_i$

We now turn to to the special case of $\mathrm{T}(\beta) = \sum_{i=1}^p \beta_i$, the sum of all regression coefficients. Theorem 4 implies that the minimax expected length for $(1 - \alpha)$ level confidence intervals of $\sum_{i=1}^p \beta_i$ over $\Theta(k)$ satisfies

$$L_\alpha^*\left(\Theta(k), \sum \beta_i\right) \asymp k \sqrt{\frac{\log p}{n}}. \tag{2.4.8}$$

The following impossibility of adaptivity result for confidence intervals for $\sum_{i=1}^p \beta_i$ is a special case of Theorem 6.

Corollary 2. *Suppose that $0 < \alpha < \frac{1}{2}$ and $k \leq c \min\{p^\gamma, \frac{n}{\log p}\}$ for some constants $c > 0$ and $0 \leq \gamma < \frac{1}{2}$. Let $k_1 \leq (1 - \zeta_0) k - 1$ for some constant $0 < \zeta_0 < 1$. Then for* *any* $\theta^* = (\beta^*, \mathrm{I}, \sigma) \in \Theta(k_1)$,

$$\inf_{\mathrm{CI}_\alpha(\sum \beta_i, Z) \in \mathcal{I}_\alpha(\Theta(k), \sum \beta_i)} \mathbb{E}_{\theta^*} L\left(\mathrm{CI}_\alpha\left(\sum \beta_i, Z\right)\right) \geq c_1 k \sqrt{\frac{\log p}{n}} \sigma, \tag{2.4.9}$$

for some constant $c_1 > 0$.

Remark 2. In the Gaussian sequence model, minimax estimation of the sum of sparse means has been considered in Cai & Low (2004) and construction of confidence intervals for the sum was studied in Cai & Low (2005). In particular, minimax estimation rate and minimax expected length of confidence intervals are given in Cai & Low (2004) and Cai & Low (2005), respectively. A more refined non-asymptotic analysis for the minimax estimation of the sum of sparse means was given in a recent paper Collier et al. (2015).

2.4.4 Computationally feasible confidence intervals

A major drawback of the minimax rate-optimal confidence intervals $\mathrm{CI}_\alpha^S(\xi^\intercal \beta, Z)$ given in (2.3.11) and $\mathrm{CI}_\alpha^D(\xi^\intercal \beta, Z)$ given in (2.4.3) is that they are not computationally feasible as both depend on restricted eigenvalue $\kappa(X, k, \alpha_0)$, which is difficult to evaluate. In this section, we assume the prior knowledge of the sparsity k and discuss how to construct a computationally feasible confidence interval.

The main idea is to replace the term involved with restricted eigenvalue by a computationally feasible lower bound function $\omega(\Omega, X, k)$ defined by

$$\omega(\Omega, X, k) = \left(\frac{1}{4\sqrt{\lambda_{\max}(\Omega)}} - \frac{9\left(1 + 405\frac{\max \|X_{\cdot j}\|_2}{\min \|X_{\cdot j}\|_2}\right)}{\sqrt{\lambda_{\min}(\Omega)}} \sqrt{k \frac{\log p}{n}}\right)_+^2. \tag{2.4.10}$$

The lower bound relation is established by Lemma 13 in Chapter A, which is based on the concentration inequality for Gaussian design in Raskutti et al. (2010). Except for

$\lambda_{\min}(\Omega)$ and $\lambda_{\max}(\Omega)$, all terms in (2.4.10) are based on the data (X, y) and the prior knowledge of k. To construct a data-dependent computationally feasible confidence interval, we make the following assumption,

$$\sup_{\Omega \in \mathcal{G}_\Omega} \mathbb{P}_X \left(\max \left\{ \left| \widetilde{\lambda_{\min}(\Omega)} - \lambda_{\min}(\Omega) \right|, \left| \widetilde{\lambda_{\max}(\Omega)} - \lambda_{\max}(\Omega) \right| \right\} \geq C a_{n,p} \right) = o(1), \tag{2.4.11}$$

where $\limsup a_{n,p} = 0$ and $\mathcal{G}_\Omega$ is a pre-specified parameter space for Ω and $\mathbb{P}_X$ denotes the probability distribution with respect to X.

Remark 3. We assume $\mathcal{G}_\Omega$ is a subspace of the precision matrix defined in (2.2.4), $\left\{\Omega : \frac{1}{M_1} \leq \lambda_{\min}(\Omega) \leq \lambda_{\max}(\Omega) \leq M_1\right\}$. By assuming $\mathcal{G}_\Omega$ is the set of precision matrix of special structure, we can find estimators satisfying (2.4.11). For example, if $\mathcal{G}_\Omega$ is assumed to be the set of sparse precision matrices, the precision matrix Ω can be estimated by the CLIME estimator $\widetilde{\Omega}$ proposed in Cai et al. (2011). Under a proper sparsity assumption on Ω, the plugin estimator $\left(\widetilde{\lambda_{\min}(\Omega)}, \widetilde{\lambda_{\max}(\Omega)}\right) = \left(\lambda_{\min}\left(\widetilde{\Omega}\right), \lambda_{\max}\left(\widetilde{\Omega}\right)\right)$ satisfies (2.4.11). Other special structures can also be assumed, for example, the covariance matrix Σ is sparse. We can use the plugin estimator of the thresholding estimators proposed in Cai & Liu (2011); Cai & Zhou (2012).

With $\widetilde{\lambda_{\min}(\Omega)}$ and $\widetilde{\lambda_{\max}(\Omega)}$, we define $\widetilde{\omega}(\Omega, X, k)$ as

$$\widetilde{\omega}(\Omega, X, k) = \left(\frac{1}{4\sqrt{\widetilde{\lambda_{\max}(\Omega)}}} - \frac{9\left(1 + 405\frac{\max \|X_{\cdot j}\|_2}{\min \|X_{\cdot j}\|_2}\right)}{\sqrt{\widetilde{\lambda_{\min}(\Omega)}}} \sqrt{k\frac{\log p}{n}} \right)_+^2 .$$

and construct computationally feasible confidence intervals by replacing

$$\kappa^2\left(X, k, 405\left(\frac{\max \|X_{\cdot j}\|_2}{\min \|X_{\cdot j}\|_2}\right)\right)$$

in (2.3.11) and (2.4.3) with $\widetilde{\omega}(\Omega, X, k)$.

2.5 Confidence intervals for linear functionals with prior knowledge $\Omega = \mathrm{I}$ and $\sigma = \sigma_0$

We have so far focused on the setting where both the precision matrix Ω and the noise level σ are unknown, which is the case in most statistical applications. It is still of theoretical interest to study the problem when Ω and σ are known. It is interesting to contrast the results with the ones when Ω and σ are unknown. In this case, we consider the setting where it is known a priori that $\Omega = \mathrm{I}$ and $\sigma = \sigma_0$ and specify the parameter space as

$$\Theta(k, \mathrm{I}, \sigma_0) = \{\theta = (\beta, \mathrm{I}, \sigma_0) : \|\beta\|_0 \leq k\}. \tag{2.5.1}$$

We will discuss separately the minimax rates and adaptivity of confidence intervals for the linear functionals in the sparse loading regime and dense loading regime over the parameter space $\Theta(k, \mathrm{I}, \sigma_0)$.

2.5.1 Confidence intervals for linear functionals in the sparse loading regime

The following theorem establishes the minimax rate of confidence intervals for linear functionals in the sparse loading regime when there is prior knowledge that $\Omega = \mathrm{I}$ and $\sigma = \sigma_0$.

Theorem 7. *Suppose that $0 < \alpha < \frac{1}{2}$ and $k \leq c \min\{p^{\gamma}, \frac{n}{\log p}\}$ for some constants $c > 0$ and $0 \leq \gamma < \frac{1}{2}$. If ξ belongs to the sparse loading regime* (2.2.6)*, the minimax* *expected length for* $(1-\alpha)$ *level confidence intervals of* $\xi^\intercal\beta$ *over* $\Theta(k, \mathrm{I}, \sigma_0)$ *satisfies*

$$L_\alpha^*\left(\Theta(k, \mathrm{I}, \sigma_0), \xi^\intercal\beta\right) \asymp \frac{\|\xi\|_2}{\sqrt{n}}. \tag{2.5.2}$$

Compared with the minimax rate $\frac{\|\xi\|_2}{\sqrt{n}} + \|\xi\|_2 k \frac{\log p}{n}$ for the unknown Ω and σ case given in Theorem 1, the minimax rate in (2.5.2) is significantly different. With the prior knowledge of $\Omega = \mathrm{I}$ and $\sigma = \sigma_0$, the above theorem shows that the minimax expected length of confidence intervals for $\xi^\intercal\beta$ is always of the parametric rate and in particular does not depend on the sparsity parameter k. In this case, adaptive confidence intervals for $\xi^\intercal\beta$ is possible over the full range $k \leq c\frac{n}{\log p}$. A similar result for confidence intervals covering all β_i was given in a recent paper Javanmard & Montanari (2015). The focus of Javanmard & Montanari (2015) is on individual coordinates, not general linear functionals.

The proof of Theorem 7 involves establishment of both minimax lower and upper bounds. The lower bound follows from the same proof for the parametric lower bound in Theorem 1. As both Ω and σ are known, the upper bound analysis is easier than the unknown Ω and σ case and is similar to the one given in Javanmard & Montanari (2015). For completeness, we detail the construction of a confidence interval achieving the minimax length in (2.5.2) using the de-biasing method. We first randomly split the samples (X, y) into two subsamples $\left(X^{(1)}, y^{(1)}\right)$ and $\left(X^{(2)}, y^{(2)}\right)$ with sample sizes n_1 and n_2, respectively. Without loss of generality, we assume that n is even and $n_1 = n_2 = \frac{n}{2}$. Let $\widehat{\beta}$ denote the Lasso estimator defined based on the sample $\left(X^{(1)}, y^{(1)}\right)$ with the proper tuning parameter $\lambda = \sqrt{\frac{2.05\log p}{n_1}}\sigma_0$,

$$\widehat{\beta} = \operatorname*{arg\,min}_{\beta\in\mathbb{R}^p} \frac{\|y^{(1)} - X^{(1)}\beta\|_2^2}{2n_1} + \lambda \sum_{j=1}^{p} \frac{\|X_{\cdot j}^{(1)}\|_2}{\sqrt{n_1}} |\beta_j|. \tag{2.5.3}$$

We define the following estimator of $\xi^\intercal \beta$,

$$\bar{\mu} = \xi^\intercal \widehat{\beta} + \frac{1}{n_2} \xi^\intercal \left(X^{(2)}\right)^\intercal \left(y^{(2)} - X^{(2)} \widehat{\beta}\right). \tag{2.5.4}$$

Based on the estimator, we construct the following confidence interval

$$\mathrm{CI}_\alpha^{\mathrm{I}}\left(\xi^\intercal \beta, Z\right) = \left[\bar{\mu} - 1.01 \frac{\|\xi\|_2}{\sqrt{n_2}} z_{\alpha_0/2} \sigma_0,\ \bar{\mu} + 1.01 \frac{\|\xi\|_2}{\sqrt{n_2}} z_{\alpha_0/2} \sigma_0\right], \tag{2.5.5}$$

where $\alpha_0 = \gamma_0 \alpha$ with $0 < \gamma_0 < 1$. It will be shown in Chapter A that the confidence interval proposed in (2.5.5) has the nominal coverage probability asymptotically and achieves the minimax length in (2.5.2).

2.5.2 Confidence intervals for linear functionals in the dense loading regime

The following theorem establishes the adaptivity lower bound in the dense loading regime.

Theorem 8. *Suppose that $0 < \alpha < \frac{1}{2}$ and $k_1 \le k \le c \min\left\{p^\gamma, \frac{n}{\log p}\right\}$ for some constants $c > 0$ and $0 \le \gamma < \frac{1}{2}$. If ξ belongs to the dense loading regime* (2.2.7)*, then there is some constant $c_1 > 0$ such that*

$$\begin{aligned} &L_\alpha^*\left(\Theta\left(k_1, \mathrm{I}, \sigma_0\right), \Theta\left(k, \mathrm{I}, \sigma_0\right), \xi^\intercal \beta\right) \\ &\qquad \ge c_1 \|\xi\|_\infty \sigma_0 \max\left\{\sqrt{k k_1} \sqrt{\frac{\log p}{n}}, \min\left\{k \sqrt{\frac{\log p}{n}}, \frac{\sqrt{k}}{n^{\frac{1}{4}}}\right\}\right\}. \end{aligned} \tag{2.5.6}$$

Remark 4. There are two parts in the lower bound given in (2.5.6), which are established separately. The lower bound $\min\left\{k\sqrt{\frac{\log p}{n}}, \frac{\sqrt{k}}{n^{\frac{1}{4}}}\right\}$ is obtained using well known techniques by testing a simple null against a composite alternative. The construction of the least favorable set is quite standard. For example, such a construction of least favorable set has been used under the Gaussian sequence model in Baraud (2002) for signal detection and in Cai & Low (2004, 2005) for estimation and confidence intervals for linear functionals. The technique has also been used more recently in Ingster et al. (2010); Nickl & van de Geer (2013) for detection and confidence ball in sparse linear regression. On the other hand, the other lower bound, $\sqrt{kk_1}\sqrt{\frac{\log p}{n}}$, cannot be established using a similar argument and a novel comparison of two composite least favorable spaces is introduced to establish this lower bound.

The lower bound given in (2.5.6) immediately yields the minimax lower bound for the expected length of confidence intervals over $\Theta\left(k, \mathrm{I}, \sigma_0\right)$,

$$L_\alpha^*\left(\Theta\left(k, \mathrm{I}, \sigma_0\right), \xi^\intercal \beta\right) \geq c_1 \|\xi\|_\infty k \sqrt{\frac{\log p}{n}} \sigma_0,$$

by simply setting $k_1 = k$ in (2.5.6). Since this lower bound can be achieved by the confidence interval constructed in (2.4.3), we have established the minimax convergence rate $L_\alpha^*\left(\Theta\left(k_1, \mathrm{I}, \sigma_0\right), \xi^\intercal \beta\right) \asymp \|\xi\|_\infty k \sqrt{\frac{\log p}{n}} \sigma_0$, which is the same as the minimax rate established in Theorem 4 for the case of unknown Ω and σ. Thus, in marked contrast to the sparse loading regime, the prior knowledge of $\Omega = \mathrm{I}$ and $\sigma = \sigma_0$ does not improve the minimax rate in the dense loading regime. Under the framework (2.2.5), adaptive confidence intervals are still impossible, since for $k_1 \ll k$,

$$L_\alpha^*\left(\Theta\left(k_1, \mathrm{I}, \sigma_0\right), \Theta\left(k, \mathrm{I}, \sigma_0\right), \xi^\intercal \beta\right) \gg L_\alpha^*\left(\Theta\left(k_1, \mathrm{I}, \sigma_0\right), \xi^\intercal \beta\right).$$

However, compared with Theorem 5, we observe that the cost of adaptation is reduced with the prior knowledge of $\Omega = \mathrm{I}$ and $\sigma = \sigma_0$.

2.6 Discussion

In the present paper we studied the minimaxity and adaptivity of confidence intervals for general linear functionals $\xi^{\intercal}\beta$ with a sparse or dense loading ξ for the setting where Ω and σ are unknown as well as the setting with the prior knowledge of $\Omega = \mathrm{I}$ and $\sigma = \sigma_0$. In the more typical case in practice where Ω and σ are unknown, the adaptivity results are quite negative: With the exception of the ultra-sparse region for confidence intervals for $\xi^{\intercal}\beta$ with a sparse loading ξ, it is necessary to know the true sparsity k in order to have guaranteed coverage probability and rate-optimal expected length. In contrast to estimation, knowledge of the sparsity k is crucial to constructing honest confidence intervals. In this sense, the problem of constructing confidence intervals is much harder than the estimation problem.

The case of known $\Omega = \mathrm{I}$ and $\sigma = \sigma_0$ is strikingly different. The minimax expected length in the sparse loading regime is of order $\frac{\|\xi\|_2}{\sqrt{n}}$ and in particular does not depend on k and adaptivity can be achieved over the full range of sparsity $k \lesssim \frac{n}{\log p}$. So in this case, the knowledge of Ω and σ is very useful. On the other hand, in the dense loading regime the information on Ω and σ is of limited use. In this case, the minimax rate and lack of adaptivity remain unchanged, compared with the unknown Ω and σ case, although the cost of adaptation is reduced.

Regarding the construction of confidence intervals, there is a significant difference between the sparse and dense loading regimes. The de-biasing method is useful in the sparse loading regime since such a procedure reduces the bias but does not dramatically increase the variance. However, the de-biasing construction is not applicable to the dense loading regime since the cost of obtaining a near-unbiased estimator is to significantly increase the variance which would lead to an unnecessarily long confidence interval. An interesting open problem is the construction of a confidence interval for $\xi^{\intercal}\beta$ achieving the minimax length where the sparsity q of the loading ξ is in the middle regime with $cp^{\gamma} \leq q \leq cp^{2\gamma+\varsigma}$ for some $0 < \varsigma < 1 - 2\gamma$.

In addition to constructing confidence intervals for linear functionals, another interesting problem is constructing confidence balls for the whole vector β. Such has been considered in Nickl & van de Geer (2013), where the impossibility of adaptive confidence balls for sparse linear regression was established. These problems are connected, but each has its own special features and the behaviors of the problems are different from each other. The connections and differences in adaptivity among various forms of confidence sets have also been observed in nonparametric function estimation problems. See, for example, Cai & Low (2005) for adaptive confidence intervals for linear functionals, Hoffmann & Nickl (2011); Cai et al. (2014) for adaptive confidence bands, and Cai & Low (2006); Robins & van der Vaart (2006) for adaptive confidence balls.

In the context of nonparametric function estimation, a general adaptation theory for confidence intervals for an arbitrary linear functional was developed in Cai & Low (2005) over a collection of convex parameter spaces. It was shown that the key quantity that determines adaptivity is a geometric quantity called the between-class modulus of continuity. The convexity assumption on the parameter space in Cai & Low (2005) is crucial for the adaptation theory. In high-dimensional linear regression, the parameter space is highly non-convex. The adaptation theory developed in Cai & Low (2005) does not apply to the present setting of high-dimensional linear regression. It would be of significant interest to develop a general adaptation theory for confidence intervals in such a non-convex setting.

2.7 Proofs

In this section, we prove three main results, Theorem 1, Theorem 2 and Theorem 3. For reasons of space, the proofs of Theorems 4-8 are given in Chapter A.

A key technical tool for the proof of the lower bound results is the following lemma which establishes the adaptivity over two nested parameter spaces. Such a formulation has been considered in Cai & Low (2005) in the context of adaptive confidence intervals over convex parameter spaces under the Gaussian sequence model. However, the parameter space $\Theta(k)$ considered in the high dimension setting is highly non-convex. The following lemma can be viewed as a generalization of Cai & Low (2005) to the non-convex parameter space, where the lower bound argument requires testing composite hypotheses.

Suppose that we observe a random variable Z which has a distribution $\mathbf{P}_\theta$ where the parameter θ belongs to the parameter space $\mathcal{H}$. Let $\mathrm{CI}_\alpha\left(\mathrm{T}, Z\right)$ be the confidence interval for the linear functional $\mathrm{T}\left(\theta\right)$ with the guaranteed coverage $1-\alpha$ over the parameter space $\mathcal{H}$. Let $\mathcal{H}_0$ and $\mathcal{H}_1$ be subsets of the parameter space $\mathcal{H}$ where $\mathcal{H} = \mathcal{H}_0 \cup \mathcal{H}_1$. Let $\pi_{\mathcal{H}_i}$ denote the prior distribution supported on the parameter space $\mathcal{H}_i$ for $i = 0, 1$. Let $f_{\pi_{\mathcal{H}_i}}\left(z\right)$ denote the density function of the marginal distribution of Z with the prior $\pi_{\mathcal{H}_i}$ on $\mathcal{H}_i$ for $i = 0, 1$. More specifically, $f_{\pi_{\mathcal{H}_i}}\left(z\right) = \int f_\theta\left(z\right) \pi_{\mathcal{H}_i}\left(\theta\right) d\theta$, for $i = 0, 1$.

Denote by $\mathbb{P}_{\pi_{\mathcal{H}_i}}$ the marginal distribution of Z with the prior $\pi_{\mathcal{H}_i}$ on $\mathcal{H}_i$ for $i = 0, 1$. For any function g, we write $\mathbb{E}_{\pi_{\mathcal{H}_0}}\left(g\left(Z\right)\right)$ for the expectation of $g\left(Z\right)$ with respect to the marginal distribution of Z with the prior $\pi_{\mathcal{H}_0}$ on $\mathcal{H}_0$. We define the χ^2 distance between two density functions f_1 and f_0 by

$$\chi^2(f_1, f_0) = \int \frac{\left(f_1(z) - f_0(z)\right)^2}{f_0(z)} dz = \int \frac{f_1^2(z)}{f_0(z)} dz - 1 \tag{2.7.1}$$

and the total variation distance by $\mathrm{TV}(f_1, f_0) = \int \left|f_1(z) - f_0(z)\right| dz$. It is well known that

$$\mathrm{TV}(f_1, f_0) \leq \sqrt{\chi^2(f_1, f_0)}. \tag{2.7.2}$$

Lemma 1. *Assume* $\mathrm{T}(\theta) = \mu_0$ *for* $\theta \in \mathcal{H}_0$ *and* $\mathrm{T}(\theta) = \mu_1$ *for* $\theta \in \mathcal{H}_1$ *and* $\mathcal{H} = \mathcal{H}_0 \cup \mathcal{H}_1$. *For any* $\mathrm{CI}_\alpha(\mathrm{T}, Z) \in \mathcal{I}_\alpha(\mathrm{T}, \mathcal{H})$,

$$L\left(\mathrm{CI}_\alpha\left(\mathrm{T}, Z\right), \mathcal{H}\right) \geq L\left(\mathrm{CI}_\alpha\left(\mathrm{T}, Z\right), \mathcal{H}_0\right) \geq \left|\mu_1 - \mu_0\right| \left(1 - 2\alpha - \mathrm{TV}\left(f_{\pi_{\mathcal{H}_1}}, f_{\pi_{\mathcal{H}_0}}\right)\right)_+ . \tag{2.7.3}$$

2.7.1 Proof of Lemma 1

The supremum risk over $\mathcal{H}_0$ is lower bounded by the Bayesian risk with the prior $\pi_{\mathcal{H}_0}$ on $\mathcal{H}_0$,

$$\sup_{\theta \in \mathcal{H}_0} \mathbb{E}_\theta L\left(\mathrm{CI}_\alpha\left(\mathrm{T}, Z\right)\right) \geq \int_\theta \mathbb{E}_\theta L\left(\mathrm{CI}_\alpha\left(\mathrm{T}, Z\right)\right) \pi_{\mathcal{H}_0}\left(\theta\right) d\theta = \mathbb{E}_{\pi_{\mathcal{H}_0}} L\left(\mathrm{CI}_\alpha\left(\mathrm{T}, Z\right)\right). \tag{2.7.4}$$

By the definition of $\mathrm{CI}_\alpha(\mathrm{T}, Z) \in \mathcal{I}_\alpha(\mathrm{T}, \mathcal{H})$, we have

$$\mathbb{P}_{\pi_{\mathcal{H}_i}}\left(\mu_i \in \mathrm{CI}_\alpha\left(\mathrm{T}, Z\right)\right) = \int_\theta \mathbb{P}_\theta\left(\mu_i \in \mathrm{CI}_\alpha\left(\mathrm{T}, Z\right)\right) \pi_{\mathcal{H}_i}\left(\theta\right) d\theta \geq 1 - \alpha, \tag{2.7.5}$$

for $i = 0, 1$. By the following inequality

$$\left|\mathbb{P}_{\pi_{\mathcal{H}_1}}\left(\mu_1 \in \mathrm{CI}_\alpha\left(\mathrm{T}, Z\right)\right) - \mathbb{P}_{\pi_{\mathcal{H}_0}}\left(\mu_1 \in \mathrm{CI}_\alpha\left(\mathrm{T}, Z\right)\right)\right| \leq \mathrm{TV}\left(f_{\pi_{\mathcal{H}_1}}, f_{\pi_{\mathcal{H}_0}}\right),$$

then we have $\mathbb{P}_{\pi_{\mathcal{H}_0}}\left(\mu_1 \in \mathrm{CI}_\alpha\left(\mathrm{T}, Z\right)\right) \geq 1 - \alpha - \mathrm{TV}(f_{\pi_{\mathcal{H}_1}}, f_{\pi_{\mathcal{H}_0}})$. This together with (2.7.5) yields $\mathbb{P}_{\pi_{\mathcal{H}_0}}\left(\mu_0, \mu_1 \in \mathrm{CI}_\alpha\left(\mathrm{T}, Z\right)\right) \geq 1 - 2\alpha - \mathrm{TV}(f_{\pi_{\mathcal{H}_1}}, f_{\pi_{\mathcal{H}_0}})$, which leads to $\mathbb{P}_{\pi_{\mathcal{H}_0}}\left(L\left(\mathrm{CI}_\alpha\left(\mathrm{T}, Z\right)\right) \geq |\mu_1 - \mu_0|\right) \geq 1 - 2\alpha - \mathrm{TV}(f_{\pi_{\mathcal{H}_1}}, f_{\pi_{\mathcal{H}_0}})$. Hence, $\mathbb{E}_{\pi_{\mathcal{H}_0}} L(\mathrm{CI}_\alpha\left(\mathrm{T}, Z\right)) \geq |\mu_1 - \mu_0|(1 - 2\alpha - \mathrm{TV}(f_{\pi_{\mathcal{H}_1}}, f_{\pi_{\mathcal{H}_0}}))_+$. The lower bound (2.7.3) then follows from inequality (2.7.4). $\square$

2.7.2 Proof of Theorem 3

The lower bound in (2.3.13) can divided into the following two lower bounds,

$$\inf_{\mathrm{CI}_\alpha(\xi^\intercal\beta,Z)\in\mathcal{I}_\alpha(\Theta(k),\xi^\intercal\beta)} \mathbb{E}_{\theta^*} L\left(\mathrm{CI}_\alpha\left(\xi^\intercal\beta, Z\right)\right) \geq c\|\xi\|_2 k\frac{\log p}{n}\sigma, \tag{2.7.6}$$

and

$$\inf_{\mathrm{CI}_\alpha(\xi^\intercal\beta,Z)\in\mathcal{I}_\alpha(\Theta(k),\xi^\intercal\beta)} \mathbb{E}_{\theta^*} L\left(\mathrm{CI}_\alpha\left(\xi^\intercal\beta, Z\right)\right) \geq c\frac{\|\xi\|_2}{\sqrt{n}}\sigma, \tag{2.7.7}$$

for some constant $c > 0$. We will establish the lower bounds (2.7.6) and (2.7.7) separately.

Proof of (2.7.6) Without loss of generality, we assume $\mathrm{supp}(\xi) = \{1, \cdots, q\}$, where $q = \|\xi\|_0$. We generate the orthogonal matrix $M \in \mathbb{R}^{q\times q}$ such that its first row is $\frac{1}{\|\xi\|_2}\xi_{\mathrm{supp}(\xi)}$ and define the orthogonal matrix Q as $Q = \begin{pmatrix} M & \mathbf{0} \\ \mathbf{0} & \mathrm{I}\end{pmatrix}$. We transform both the design matrix X and the regression vector β and view the linear model (3.2.1) as $y = V\psi + \epsilon$, where $V = XQ^\intercal$ and $\psi = Q\beta$. The transformed coefficient vector $\psi^* = Q\beta^* = \begin{pmatrix} M\beta^*_{\mathrm{supp}(\xi)} \\ \beta^*_{-\mathrm{supp}(\xi)}\end{pmatrix}$ is of sparsity at most $q + k_1$. The first coefficient ψ_1 of ψ is $\frac{1}{\|\xi\|_2}\xi^\intercal\beta$. The covariance matrix Ψ of $V_{1\cdot}$ is $Q\Sigma Q^\intercal$ and its corresponding precision matrix is $\Gamma = Q\Omega Q^\intercal$. To represent the transformed observed data and parameter, we abuse the notation slightly and also use $Z_i = (y_i, V_{i\cdot})$ and $\theta^* = (\psi^*, \mathrm{I}, \sigma)$. We define the parameter space $\mathcal{G}\left(k\right)$ of (ψ, Γ, σ) as

$$\mathcal{G}\left(k\right) = \left\{(\psi, \Gamma, \sigma) : \psi = Q\beta,\ \Gamma = Q\Omega Q^\intercal \text{ for } (\beta, \Omega, \sigma) \in \Theta\left(k\right)\right\}. \tag{2.7.8}$$

For a given Q, there exists a bijective mapping between $\Theta\left(k\right)$ and $\mathcal{G}\left(k\right)$. To show that $(\psi, \Gamma, \sigma) \in \mathcal{G}\left(k\right)$, it is equivalent to show $(Q^\intercal\psi, Q^\intercal\Gamma Q, \sigma) \in \Theta\left(k\right)$. Let $\mathcal{I}_\alpha\left(\mathcal{G}\left(k\right), \psi_1\right)$ denote the set of confidence intervals for $\psi_1 = \frac{1}{\|\xi\|_2}\xi^\intercal\beta$ with guaranteed coverage over $\mathcal{G}(k)$. If $\mathrm{CI}_\alpha(\psi_1, Z) \in \mathcal{I}_\alpha(\mathcal{G}(k), \psi_1)$, then $\|\xi\|_2 \mathrm{CI}_\alpha(\psi_1, Z) \in \mathcal{I}_\alpha(\Theta(k), \xi^\intercal\beta)$; If $\mathrm{CI}_\alpha(\xi^\intercal\beta, Z) \in \mathcal{I}_\alpha(\Theta(k), \xi^\intercal\beta)$, then $\frac{1}{\|\xi\|_2}\mathrm{CI}_\alpha(\xi^\intercal\beta, Z) \in \mathcal{I}_\alpha(\mathcal{G}(k), \psi_1)$. Because of such one to one correspondence, we have

$$\inf_{\mathrm{CI}_\alpha(\xi^\intercal\beta, Z)\in\mathcal{I}_\alpha(\Theta(k),\xi^\intercal\beta)} \mathbb{E}_{\theta^*} L\left(\mathrm{CI}_\alpha\left(\xi^\intercal\beta, Z\right)\right) = \|\xi\|_2 \inf_{\mathrm{CI}_\alpha(\psi_1, Z)\in\mathcal{I}_\alpha(\mathcal{G}(k),\psi_1)} \mathbb{E}_{\theta^*} L\left(\mathrm{CI}_\alpha\left(\psi_1, Z\right)\right). \tag{2.7.9}$$

By (2.7.6) and (2.7.9), we reduce the problem to

$$\inf_{\mathrm{CI}_\alpha(\psi_1, Z)\in\mathcal{I}_\alpha(\mathcal{G}(k),\psi_1)} \mathbb{E}_{\theta^*} L\left(\mathrm{CI}_\alpha\left(\psi_1, Z\right)\right) \geq ck\frac{\log p}{n}\sigma. \tag{2.7.10}$$

Under the Gaussian random design model, $Z_i = (y_i, V_{i\cdot}) \in \mathbb{R}^{p+1}$ follows a joint Gaussian distribution with mean 0. Let Σ^z denotes the covariance matrix of Z_i. Decompose Σ^z into blocks $\begin{pmatrix} \Sigma^z_{yy} & \left(\Sigma^z_{vy}\right)^\intercal \\ \Sigma^z_{vy} & \Sigma^z_{vv} \end{pmatrix}$, where Σ^z_{yy}, Σ^z_{vv} and Σ^z_{vy} denote the variance of y, the variance of V and the covariance of y and V, respectively. We define the function $h : \Sigma^z \to (\psi, \Gamma, \sigma)$ as $h(\Sigma^z) = \left(\left(\Sigma^z_{vv}\right)^{-1}\Sigma^z_{vy}, \left(\Sigma^z_{vv}\right)^{-1}, \Sigma^z_{yy} - \left(\Sigma^z_{vy}\right)^\intercal\left(\Sigma^z_{vv}\right)^{-1}\Sigma^z_{vy}\right)$. The function h is bijective and its inverse mapping $h^{-1} : (\psi, \Gamma, \sigma) \to \Sigma^z$ is

$$h^{-1}\left((\psi, \Gamma, \sigma)\right) = \begin{pmatrix} \psi^\intercal\Gamma^{-1}\psi + \sigma^2 & \psi^\intercal\Gamma^{-1} \\ \Gamma^{-1}\psi & \Gamma^{-1} \end{pmatrix}.$$

The null space is taken as $\mathcal{H}_0 = \{(\psi^*, \mathrm{I}, \sigma)\}$ and $\pi_{\mathcal{H}_0}$ denotes the point mass prior at this point. The proof is divided into three steps:

1. Construct $\mathcal{H}_1$ and show that $\mathcal{H}_1 \subset \mathcal{G}(k)$;

2. Control the distribution distance $\mathrm{TV}\left(f_{\pi_{\mathcal{H}_1}}, f_{\pi_{\mathcal{H}_0}}\right)$;

3. Calculate the distance $\mu_1 - \mu_0$ where $\mu_0 = \psi_1^*$ and $\mu_1 = \psi_1$ with $(\psi, \Gamma, \sigma) \in \mathcal{H}_1$. We show that $\mu_1 = \psi_1$ is a fixed constant for all $(\psi, \Gamma, \sigma) \in \mathcal{H}_1$ and then apply

Lemma 1.

Step 1. We construct the alternative hypothesis parameter space $\mathcal{H}_1$. Let Σ_0^z denote the covariance matrix of Z_i corresponding to $(\psi^*, \mathrm{I}, \sigma) \in \mathcal{H}_0$. Let $S_1 = \mathrm{supp}\,(\psi^*) \cup \{1\}$ and $S = S_1 \backslash \{1\}$. Let k_* denote the size of S and p_1 denote the size of S_1^c and we have $k_* \leq k_1 + q$ and $p_1 \geq p - k_* - 1 \geq cp$. Without loss of generality, let $S = \{2, \cdots, k_* + 1\}$. We have the following expression for the covariance matrix of Z_i under the null,

$$\Sigma_0^z = \left(\begin{array}{c|c|c|c} \|\psi^*\|_2^2 + \sigma^2 & \psi_1^* & (\psi_S^*)^\intercal & \mathbf{0}_{1\times p_1} \\ \hline \psi_1^* & 1 & \mathbf{0}_{1\times k_*} & \mathbf{0}_{1\times p_1} \\ \hline \psi_S^* & \mathbf{0}_{k_*\times 1} & \mathrm{I}_{\mathrm{k_*\times k_*}} & \mathbf{0}_{k_*\times p_1} \\ \hline \mathbf{0}_{p_1\times 1} & \mathbf{0}_{p_1\times 1} & \mathbf{0}_{p_1\times k_*} & \mathrm{I}_{p_1\times p_1} \end{array}\right). \tag{2.7.11}$$

To construct $\mathcal{H}_1$, we define the following set,

$$\ell\left(p_1, \frac{\zeta_0}{2}k, \rho\right) = \left\{\boldsymbol{\delta} : \boldsymbol{\delta} \in \mathbb{R}^{p_1},\ \|\boldsymbol{\delta}\|_0 = \frac{\zeta_0}{2}k,\ \boldsymbol{\delta}_i \in \{0, \rho\} \text{ for } 1 \leq i \leq p_1\right\}. \tag{2.7.12}$$

Define the parameter space $\mathcal{F}$ for Σ^z by $\mathcal{F} = \left\{\Sigma_{\boldsymbol{\delta}}^z : \boldsymbol{\delta} \in \ell\left(p_1, \frac{\zeta_0}{2}k, \rho\right)\right\}$, where

$$\Sigma_{\boldsymbol{\delta}}^z = \left(\begin{array}{c|c|c|c} \|\psi^*\|_2^2 + \sigma^2 & \psi_1^* & (\psi_S^*)^\intercal & \rho_0\boldsymbol{\delta}^\intercal \\ \hline \psi_1^* & 1 & \mathbf{0}_{1\times k_*} & \boldsymbol{\delta}^\intercal \\ \hline \psi_S^* & \mathbf{0}_{k_*\times 1} & \mathrm{I}_{\mathrm{k_*\times k_*}} & \mathbf{0}_{k_*\times p_1} \\ \hline \rho_0\boldsymbol{\delta} & \boldsymbol{\delta} & \mathbf{0}_{p_1\times k_*} & \mathrm{I}_{\mathrm{p_1\times p_1}} \end{array}\right). \tag{2.7.13}$$

Then we construct the alternative hypothesis space $\mathcal{H}_1$ for (ψ, Γ, σ), which is induced by the mapping h and the parameter space $\mathcal{F}$,

$$\mathcal{H}_1 = \{(\psi, \Gamma, \sigma)\ :\ (\psi, \Gamma, \sigma) = h\left(\Sigma^z\right) \quad \text{for } \Sigma^z \in \mathcal{F}\}. \tag{2.7.14}$$

In the following, we show that $\mathcal{H}_1 \subset \mathcal{G}(k)$. It is necessary to identify $(\psi, \Gamma, \sigma) = h(\Sigma^z)$ for $\Sigma^z \in \mathcal{F}$ and show $(Q^\intercal \psi, Q^\intercal \Gamma Q, \sigma) \in \Theta(k)$. Firstly, we identify the expression $\mathbb{E}(y_i \mid V_{i,\cdot})$ under the alternative joint distribution (2.7.13). Assuming $y_i = V_{i1}\psi_1 + V_{i,S}\psi_S + V_{i,S_1^c}\psi_{S_1^c} + \epsilon_i'$, we have

$$\psi_1 = \frac{-\|\boldsymbol{\delta}\|_2^2 \rho_0 + \psi_1^*}{1 - \|\boldsymbol{\delta}\|_2^2}, \ \psi_S = \psi_S^*, \ \psi_{S_1^c} = (\rho_0 - \psi_1)\, \boldsymbol{\delta}, \tag{2.7.15}$$

and

$$\operatorname{Var}(\epsilon_i') = \sigma^2 - \frac{\|\boldsymbol{\delta}\|_2^2 (\rho_0 - \psi_1^*)^2}{1 - \|\boldsymbol{\delta}\|_2^2} \leq \sigma^2 \leq M_2. \tag{2.7.16}$$

Based on (2.7.15), the sparsity of ψ in the alternative hypothesis space is upper bounded by $1 + |\operatorname{supp}(\psi_S^*)| + |\operatorname{supp}(\boldsymbol{\delta})| \leq \left(1 - \frac{\zeta_0}{4}\right) k$, and hence the sparsity of the corresponding $\beta = Q^\intercal \psi$ is controlled by

$$\|\beta\|_0 \leq \left(1 - \frac{\zeta_0}{4}\right) k + q \leq k. \tag{2.7.17}$$

Secondly, we show that $\Omega = Q^\intercal \Gamma Q$ satisfies the condition $\frac{1}{M_1} \leq \lambda_{\min}(\Omega) \leq \lambda_{\max}(\Omega) \leq M_1$. The covariance matrix Ψ of $V_{i,\cdot}$ in the alternative hypothesis parameter space is expressed as

$$\Psi = \left(\begin{array}{c|c|c} 1 & \mathbf{0}_{1\times k_*} & \mathbf{0}_{k_*\times p_1} \\ \hline \mathbf{0}_{k_*\times 1} & \mathrm{I}_{\mathrm{k_*\times k_*}} & \mathbf{0}_{k_*\times p_1} \\ \hline \mathbf{0}_{p_1\times 1} & \mathbf{0}_{p_1\times k_*} & \mathrm{I}_{\mathrm{p_1\times p_1}} \end{array}\right) + \left(\begin{array}{c|c|c} 0 & \mathbf{0}_{1\times k_*} & \boldsymbol{\delta}^\intercal \\ \hline \mathbf{0}_{k_*\times 1} & \mathbf{0}_{k_*\times k_*} & \mathbf{0}_{k_*\times p_1} \\ \hline \boldsymbol{\delta} & \mathbf{0}_{p_1\times k_*} & \mathbf{0}_{p_1\times p_1} \end{array}\right). \tag{2.7.18}$$

Since the second matrix on the above equation is of spectral norm $\|\boldsymbol{\delta}\|_2$, Weyl's inequality leads to $\max\{|\lambda_{\min}(\Psi) - 1|, |\lambda_{\max}(\Psi) - 1|\} \leq \|\boldsymbol{\delta}\|_2$. When $\|\boldsymbol{\delta}\|_2$ is chosen such that $\|\boldsymbol{\delta}\|_2 \leq \min\left\{1 - \frac{1}{M_1}, M_1 - 1\right\}$, then we have $\frac{1}{M_1} \leq \lambda_{\min}(\Psi) \leq \lambda_{\max}(\Psi) \leq M_1$. Since Ω and $\Gamma = Q\Omega Q^\intercal$ have the same eigenvalues, we have $\frac{1}{M_1} \leq \lambda_{\min}(\Omega) \leq$

$\lambda_{\max}(\Omega) \leq M_1$. Combined with (2.7.16) and (2.7.17), we show that $\mathcal{H}_1 \subset \mathcal{G}(k)$.

Step 2. To control $\mathrm{TV}\left(f_{\pi_{\mathcal{H}_1}}, f_{\pi_{\mathcal{H}_0}}\right)$, it is sufficient to control $\chi^2\left(f_{\pi_{\mathcal{H}_1}}, f_{\pi_{\mathcal{H}_0}}\right)$ and apply (2.7.2). Let π denote the uniform prior on $\boldsymbol{\delta}$ over $\ell\left(p_1, \frac{\zeta_0}{2}k, \rho\right)$. Note that this uniform prior π induces a prior distribution $\pi_{\mathcal{H}_1}$ over the parameter space $\mathcal{H}_1$. Let $\mathbb{E}_{\boldsymbol{\delta},\widetilde{\boldsymbol{\delta}}}$ denote the expectation with respect to the independent random variables $\boldsymbol{\delta}, \widetilde{\boldsymbol{\delta}}$ with uniform prior π over the parameter space $\ell\left(p_1, \frac{\zeta_0}{2}k, \rho\right)$. The following lemma controls the χ^2 distance between the null and the mixture over the alternative distribution.

Lemma 2. *Let* $f_1 = \left(\sigma^2 + (\psi_1^*)^2 - \rho_0\psi_1^*\right)$. *Then*

$$\chi^2\left(f_{\pi_{\mathcal{H}_1}}, f_{\pi_{\mathcal{H}_0}}\right) + 1 = \mathbb{E}_{\boldsymbol{\delta},\widetilde{\boldsymbol{\delta}}}\left(1 - \frac{1}{\sigma^2}\left(\rho_0\left(\rho_0 - \psi_1^*\right) + f_1\right)\boldsymbol{\delta}^\intercal\widetilde{\boldsymbol{\delta}}\right)^{-n}. \tag{2.7.19}$$

The following lemma is useful in controlling the right hand side of (2.7.19).

Lemma 3. *Let* J *be a* $\text{Hypergeometric}\,(p, k, k)$ *variable with* $\mathbb{P}(J = j) = \frac{\binom{k}{j}\binom{p-k}{k-j}}{\binom{p}{k}}$, *then*

$$\mathbb{E}\exp(tJ) \leq e^{\frac{k^2}{p-k}}\left(1 - \frac{k}{p} + \frac{k}{p}\exp(t)\right)^k. \tag{2.7.20}$$

Taking $\rho_0 = \psi_1^* + \sigma$, we have $\frac{1}{\sigma^2}\left(\rho_0\left(\rho_0 - \psi_1^*\right) + f_1\right) = 2$ and by Lemma 2,

$$\chi^2\left(f_{\pi_{\mathcal{H}_1}}, f_{\pi_{\mathcal{H}_0}}\right) + 1 = \mathbb{E}_{\boldsymbol{\delta},\widetilde{\boldsymbol{\delta}}}\left(1 - 2\boldsymbol{\delta}^\intercal\widetilde{\boldsymbol{\delta}}\right)^{-n}.$$

By the inequality $\frac{1}{1-x} \leq \exp(2x)$ for $x \in \left[0, \frac{\log 2}{2}\right]$, if $\boldsymbol{\delta}^\intercal\widetilde{\boldsymbol{\delta}} \leq \frac{\zeta_0}{2}k\rho^2 < \frac{\log 2}{4}$, then $\left(1 - 2\boldsymbol{\delta}^\intercal\widetilde{\boldsymbol{\delta}}\right)^{-n} \leq \exp\left(4n\boldsymbol{\delta}^\intercal\widetilde{\boldsymbol{\delta}}\right)$. By Lemma 3, we further have

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{\delta},\widetilde{\boldsymbol{\delta}}}\exp\left(4n\boldsymbol{\delta}^\intercal\widetilde{\boldsymbol{\delta}}\right) &= \mathbb{E}\exp\left(4Jn\rho^2\right) \leq e^{\frac{\zeta_0^2k^2}{4p_1 - 2\zeta_0k}}\left(1 - \frac{\zeta_0k}{2p_1} + \frac{\zeta_0k}{2p_1}\exp\left(4n\rho^2\right)\right)^{\frac{\zeta_0}{2}k} \\
&\leq e^{\frac{\zeta_0^2k^2}{4p_1 - 2\zeta_0k}}\left(1 - \frac{\zeta_0k}{2p_1} + \frac{\zeta_0k}{2p_1}\sqrt{\frac{4p_1}{\zeta_0^2k^2}}\right)^{\frac{\zeta_0}{2}k} \leq e^{\frac{c^2\zeta_0^2p^{2\gamma}}{4p_1 - 2c\zeta_0p^\gamma}}\left(1 + \frac{1}{\sqrt{p_1}}\right)^{\frac{c\zeta_0}{2}p^\gamma},
\end{aligned}$$

where the second inequality follows by plugging in $\rho = \sqrt{\frac{\log \frac{4p_1}{\zeta_0^2 k^2}}{8n}}$ and the last inequality follows by $k \le cp^{\gamma}$. If $k \le c\left\{\frac{n}{\log p}, p^{\gamma}\right\}$, where $0 \le \gamma < \frac{1}{2}$ and c is a sufficient small positive constant, then $k\rho^2 < \min\left\{\frac{\log 2}{2\zeta_0}, \left(1 - \frac{1}{M_1}\right)^2, 1\right\}$ and hence

$$\chi^2\left(f_{\pi_{\mathcal{H}_1}}, f_{\pi_{\mathcal{H}_0}}\right) \le \left(\frac{1}{2} - \alpha\right)^2 \quad \text{and} \quad \mathrm{TV}\left(f_{\pi_{\mathcal{H}_1}}, f_{\pi_{\mathcal{H}_0}}\right) \le \frac{1}{2} - \alpha. \tag{2.7.21}$$

Step 3. We calculate the distance between μ_1 and μ_0. Under $\mathcal{H}_0$, $\mu_0 = \psi_1^*$. Under $\mathcal{H}_1$, $\mu_1 = \psi_1 = \frac{-\|\boldsymbol{\delta}\|_2^2 \rho_0 + \psi_1^*}{1 - \|\boldsymbol{\delta}\|_2^2}$. For $\boldsymbol{\delta} \in \ell\left(p_1, \frac{\zeta_0}{2}k, \rho\right)$, $\|\boldsymbol{\delta}\|_2^2 = \frac{\zeta_0}{2}k\rho^2$ and $\mu_1 = \psi_1 = \frac{-\frac{\zeta_0}{2}k\rho^2\left(\psi_1^* + \sigma\right) + \psi_1^*}{1 - \frac{\zeta_0}{2}k\rho^2}$. Since ρ is selected as fixed, $\mu_1 = \psi_1$ is a fixed constant for $(\psi, \Gamma, \sigma) \in \mathcal{H}_1$. Note that $\mu_1 - \mu_0 = \frac{\|\boldsymbol{\delta}\|_2^2\left(\psi_1^* - \rho_0\right)}{1 - \|\boldsymbol{\delta}\|_2^2} = \frac{-\sigma\|\boldsymbol{\delta}\|_2^2}{1 - \|\boldsymbol{\delta}\|_2^2}$, and it follows that $|\mu_1 - \mu_0| = \sigma\frac{\|\boldsymbol{\delta}\|_2^2}{1 - \|\boldsymbol{\delta}\|_2^2} \ge ck\frac{\log \frac{4p_1}{\zeta_0^2 k^2}}{n}\sigma$. Combined with (2.7.2) and (2.7.21), Lemma 1 leads to (2.7.10). By (2.7.9), we establish (2.3.13).

<u>**Proof of** (2.7.7)</u> Similar to the proof of (2.7.6), the proof is divided into three steps.

<u>The first step.</u> We construct alternative hypothesis parameter space $\mathcal{H}_1$. For a given ξ, β^* and a small positive constant $\bar{\epsilon}$, we select β such that

$$\beta_{-\mathrm{supp}(\xi)} = \beta^*_{-\mathrm{supp}(\xi)}, \quad \|\beta_{\mathrm{supp}(\xi)} - \beta^*_{\mathrm{supp}(\xi)}\|_2 = \sigma\frac{\bar{\epsilon}}{\sqrt{n}}. \tag{2.7.22}$$

and

$$\xi^{\intercal}\left(\beta - \beta^*\right) = \sum_{i \in \mathrm{supp}(\xi)} \xi_i\left(\beta_i - \beta_i^*\right) = \|\xi\|_2\|\beta - \beta^*\|_2. \tag{2.7.23}$$

The sparsity of β is controlled by $\|\beta\|_0 \le \|\beta^*\|_0 + \|\xi\|_0 \le k$, and hence $(\beta, \mathrm{I}, \sigma) \in \Theta(k)$. We consider the parameter spaces $\mathcal{H}_0 = \{\theta^* = (\beta^*, \mathrm{I}, \sigma)\}$ and $\mathcal{H}_1 = \{(\beta, \mathrm{I}, \sigma)\}$.

<u>The second step.</u> Let $\pi_{\mathcal{H}_0}$ denote the point mass prior on the point $(\beta^*, \mathrm{I}, \sigma)$ and $\pi_{\mathcal{H}_1}$ denote the point mass prior on the point $(\beta, \mathrm{I}, \sigma)$. Let $f_{\pi_{\mathcal{H}_0}}(y \mid X)$ denote the conditional density function of the marginal distribution of y given X with the pa-

rameter $\pi_{\mathcal{H}_i}$ on $\mathcal{H}_i$ for $i=0,1$. The χ^2 distance between the conditional distributions $f_{\pi_{\mathcal{H}_1}}(y \mid X)$ and $f_{\pi_{\mathcal{H}_0}}(y \mid X)$ is

$$\chi^2\left(f_{\pi_{\mathcal{H}_1}}(y \mid X), f_{\pi_{\mathcal{H}_0}}(y \mid X)\right)+1=\exp\left(\frac{1}{\sigma^2}\|X(\beta-\beta^*)\|_2^2\right). \tag{2.7.24}$$

Let $\mathbb{E}_X$ denote the expectation with respect to X, where $X_{i\cdot} \overset{\text{i.i.d.}}{\sim} N_p(0,\mathrm{I})$, $i=1,...,n$, then we have

$$\begin{aligned}\chi^2\left(f_{\pi_{\mathcal{H}_1}}(y,X), f_{\pi_{\mathcal{H}_0}}(y,X)\right)&=\mathbb{E}_X\left(\chi^2\left(f_{\pi_{\mathcal{H}_1}}(y \mid X), f_{\pi_{\mathcal{H}_0}}(y \mid X)\right)\right)\\&=\mathbb{E}_X\exp\left(\frac{1}{\sigma^2}\|X(\beta-\beta^*)\|_2^2\right)-1.\end{aligned}$$

If $\frac{2\|\beta^*-\beta\|_2^2}{\sigma^2}<\frac{\log 2}{2}$, we have

$$\chi^2\left(f_{\pi_{\mathcal{H}_1}}(y,X), f_{\pi_{\mathcal{H}_0}}(y,X)\right)=\left(1-\frac{2\|\beta^*-\beta\|_2^2}{\sigma^2}\right)^{-\frac{n}{2}}-1\le\exp\left(\frac{2n\|\beta^*-\beta\|_2^2}{\sigma^2}\right)-1, \tag{2.7.25}$$

where the first equality follows from the moment generating function of χ^2 distribution and the second inequality follows from the inequality $\frac{1}{1-x}\le\exp(2x)$ for $x\in\left[0,\frac{\log 2}{2}\right]$.

The third step. We calculate the distance between $\mu_1=\mathrm{T}\beta$ and $\mu_0=\mathrm{T}\beta^*$. Note that μ_0 and μ_1 are fixed constants under the simple null and alternative hypothesis. By Lemma 1, the construction (2.7.22) and (2.7.23) and the control of χ^2 distance (2.7.25) lead to

$$\mathbb{E}_{\theta^*}\left(L(\mathrm{CI}_\alpha(\xi^\intercal\beta,Z))\right)\ge\sigma\frac{\bar{\epsilon}}{\sqrt{n}}\left(1-2\alpha-\sqrt{\exp(2\bar{\epsilon}^2)-1}\right). \quad \square$$

2.7.3 Proof of Theorem 2

Theorem 2 follows from Theorem 3. Given $0<\zeta_0<1$, we define $k_1^*=\min\{k_1,(1-\zeta_0)k-1\}$ and $q^*=\min\{\frac{\zeta_0}{4}k,\|\xi\|_0\}$. Let J denote the subset of $\{1,\cdots,p\}$ corresponding to the q^* largest in absolute value coordinates of ξ. Define the parameter space $\Theta_\xi(k) = \{\theta \in \Theta(k) : \beta_{\mathrm{supp}(\xi)\setminus J} = 0\}$, which is a subspace of $\Theta(k)$ setting β to be zero on the set $\mathrm{supp}(\xi)\setminus J$. Define the vector $\bar{\xi}$ such that $\bar{\xi}_j = \xi_j$ for $j \in J$ and $\bar{\xi}_j = 0$ for $j \notin J$. By the fact that $\xi^\intercal\beta = \bar{\xi}^\intercal\beta$ for $\beta \in \Theta_\xi(k)$, we have

$$\inf_{\mathrm{CI}_\alpha(\xi^\intercal\beta,Z)\in\mathcal{I}_\alpha\left(\Theta_\xi(k),\xi^\intercal\beta\right)} \mathbb{E}_{\theta^*} L\left(\mathrm{CI}_\alpha(\xi^\intercal\beta, Z)\right) = \inf_{\mathrm{CI}_\alpha(\bar{\xi}^\intercal\beta,Z)\in\mathcal{I}_\alpha\left(\Theta_\xi(k),\bar{\xi}^\intercal\beta\right)} \mathbb{E}_{\theta^*} L\left(\mathrm{CI}_\alpha(\bar{\xi}^\intercal\beta, Z)\right).$$

It then follows from the same argument as the proof of Theorem 3 that

$$\inf_{\mathrm{CI}_\alpha(\bar{\xi}^\intercal\beta,Z)\in\mathcal{I}_\alpha\left(\Theta_\xi(k),\bar{\xi}^\intercal\beta\right)} \mathbb{E}_{\theta^*} L\left(\mathrm{CI}_\alpha(\bar{\xi}^\intercal\beta, Z)\right) \geq c\|\bar{\xi}\|_2\left(\frac{1}{\sqrt{n}} + \frac{k\log p}{n}\right).$$

By taking $\theta^* \in \Theta_\xi(k_1^*)$, we have

$$L_\alpha^*(\Theta_\xi(k_1^*), \Theta_\xi(k), \xi^\intercal\beta) \geq \inf_{\mathrm{CI}_\alpha(\xi^\intercal\beta,Z)\in\mathcal{I}_\alpha\left(\Theta_\xi(k),\xi^\intercal\beta\right)} \mathbb{E}_{\theta^*} L\left(\mathrm{CI}_\alpha\left(\xi^\intercal\beta, Z\right)\right).$$

Since $\Theta_\xi(k_1^*) \subset \Theta(k_1)$, $\Theta_\xi(k) \subset \Theta(k)$ and $\|\bar{\xi}\|_2 \geq c\|\xi\|_2$, we have established Theorem 2. □

2.7.4 Proof of Theorem 1

The lower bound of Theorem 1 follows from Theorem 2 by taking $k_1 = k$. The minimax upper bound follows from the following proposition, which establishes the coverage property and the expected length of the confidence interval constructed in (2.3.11). Such a confidence interval achieves the minimax length in (2.3.1).

Proposition 1. *Suppose that $k \leq c_* \frac{n}{\log p}$, where c_* is a small positive constant, then*

$$\liminf_{n,p\to\infty} \inf_{\theta\in\Theta(k)} \mathbb{P}_\theta\left(\xi^\intercal\beta \in \mathrm{CI}_\alpha^S\left(\xi^\intercal\beta, Z\right)\right) \geq 1-\alpha, \tag{2.7.26}$$

and

$$L\left(\mathrm{CI}_{\alpha}^{S}\left(\xi^{\intercal}\beta, Z\right), \Theta\left(k\right)\right) \leq C\|\xi\|_2\left(k\frac{\log p}{n} + \frac{1}{\sqrt{n}}\right), \tag{2.7.27}$$

for some constant $C > 0$.

In the following, we are going to prove Proposition 1. By normalizing the columns of X and the true sparse vector β, the linear regression model can be expressed as

$$y = Wd + \epsilon, \quad \text{with } \; W = XD, \; d = D^{-1}\beta \text{ and } \epsilon \sim N(0, \sigma^2\mathrm{I}), \tag{2.7.28}$$

where

$$D = \operatorname{diag}\left(\frac{\sqrt{n}}{\|X_{\cdot j}\|_2}\right)_{j\in[p]} \tag{2.7.29}$$

denotes the $p \times p$ diagonal matrix with (j, j) entry to be $\frac{\sqrt{n}}{\|X_{\cdot j}\|_2}$. Take $\delta_0 = 1.0048$ and $\eta_0 = 0.01$, and we have $\lambda_0 = (1+\eta_0)\sqrt{\frac{2\delta_0 \log p}{n}}$. Take $\epsilon_0 = \frac{2.01}{\eta_0} + 1 = 202$, $\nu_0 = 0.01$, $C_1 = 2.25$, $c_0 = \frac{1}{6}$ and $C_0 = 3$. Rather than use the constants directly in the following discussion, we use $\delta_0, \eta_0, \epsilon_0, \nu_0, C_1, C_0$ and c_0 to represent the above fixed constants in the following discussion. We also assume that $\frac{\log p}{n} \leq \frac{1}{25}$ and $\delta_0 \log p > 2$. Define the l_1 cone invertibility factor (CIF_1) as follows,

$$CIF_1\left(\alpha_0, K, W\right) = \inf\left\{\frac{|K|\|\frac{W^{\intercal}W}{n}u\|_{\infty}}{\|u_K\|_1} : \|u_{K^c}\|_1 \leq \alpha_0\|u_K\|_1, u \neq 0\right\}, \tag{2.7.30}$$

where K is an index set. Define $\sigma^{ora} = \frac{1}{\sqrt{n}}\|y - X\beta\|_2 = \frac{1}{\sqrt{n}}\|y - Wd\|_2$,

$$T = \{k : |d_k| \geq \lambda_0\sigma^{ora}\}, \; \tau = \left(1+\epsilon_0\right)\lambda_0 \max\left\{\frac{4}{\sigma^{ora}}\|d_{T^c}\|_1, \frac{8|T|}{CIF_1\left(2\epsilon_0 + 1, T, W\right)}\right\}. \tag{2.7.31}$$

To facilitate the proof, we define the following events for the random design X and the error ϵ,

$$
\begin{aligned}
G_1 &= \left\{ \frac{2}{5}\frac{1}{\sqrt{M_1}} < \frac{\|X_{\cdot j}\|_2}{\sqrt{n}} < \frac{7}{5}\sqrt{M_1} \text{ for } 1 \le j \le p \right\}, \\
G_2 &= \left\{ \left| \frac{(\sigma^{ora})^2}{\sigma^2} - 1 \right| \le 2\sqrt{\frac{\log p}{n}} + 2\frac{\log p}{n} \right\}, \\
G_3 &= \left\{ \max\left\{ \left| \frac{\xi^\intercal \widehat{\Sigma} \xi}{\xi^\intercal \Sigma \xi} - 1 \right|, \left| \frac{u^\intercal \widehat{\Sigma} u}{\xi^\intercal \Omega \xi} - 1 \right| \right\} \le 2\sqrt{\frac{\log p}{n}} + 2\frac{\log p}{n} \right\}, \text{ where } u = \Omega\xi, \\
G_4 &= \left\{ \kappa(X, k, \alpha) \ge \frac{1}{4\sqrt{\lambda_{\max}(\Omega)}} - \frac{9}{\sqrt{\lambda_{\min}(\Omega)}}(1+\alpha)\sqrt{k\frac{\log p}{n}} \right\}, \\
G_5 &= \left\{ \frac{\|W^\intercal \epsilon\|_\infty}{n} \le \sigma\sqrt{\frac{2\delta_0 \log p}{n}} \right\}, \\
S_1 &= \left\{ \frac{\|W^\intercal \epsilon\|_\infty}{n} \le \sigma^{ora}\lambda_0 \frac{\epsilon_0 - 1}{\epsilon_0 + 1}(1-\tau) \right\}, \\
S_2 &= \left\{ (1-\nu_0)\,\hat{\sigma} \le \sigma \le (1+\nu_0)\hat{\sigma} \right\}, \\
B_1 &= \left\{ \|\xi^\intercal \Omega \hat{\Sigma} - \xi^\intercal\|_\infty \le \lambda_n \right\}, \quad \text{where} \quad \lambda_n = 4C_0 M_1^2 \|\xi\|_2 \sqrt{\frac{\log p}{n}}.
\end{aligned}
$$

Define $G = \cap_{i=1}^5 G_i$ and $S = \cap_{i=1}^2 S_i$. The following lemmas control the probability of events G, S and B_1. The detailed proofs of Lemma 4, 5 and 6 are in the supplement.

Lemma 4.

$$
\mathbb{P}_\theta(G) \ge 1 - \frac{6}{p} - 2p^{1-C_1} - \frac{1}{2\sqrt{\pi\delta_0 \log p}} p^{1-\delta_0} - c' \exp(-cn), \tag{2.7.32}
$$

and

$$
\mathbb{P}_\theta(B_1) \ge 1 - 2p^{1-c_0 C_0^2}, \tag{2.7.33}
$$

where c and c' are universal positive constants. If $k \le c_* \frac{n}{\log p}$, then

$$
\mathbb{P}_\theta(G \cap S) \ge \mathbb{P}_\theta(G) - 2\exp\left(-\left(\frac{g_0 + 1 - \sqrt{2g_0 + 1}}{2} \right) n \right) - c'' \frac{1}{\sqrt{\log p}} p^{1-\delta_0}, \tag{2.7.34}
$$

where c_* *and* c'' *are universal positive constants and* $g_0 = \frac{\nu_0}{2+3\nu_0}$.

The following lemma establishes a data-dependent upper bound for the term $\|\widehat{\beta} - \beta\|_1$.

Lemma 5. *On the event* $G \cap S$,

$$\|\widehat{\beta} - \beta\|_1 \leq (2 + 2\epsilon_0) \frac{\sqrt{n}}{\min \|X_{\cdot j}\|_2} l(Z, k), \tag{2.7.35}$$

where

$$l(Z, k) = \max \left\{ k\lambda_0 \sigma^{ora}, \frac{(2 + 2\epsilon_0) \max \|X_{\cdot j}\|_2^2 \left(\sigma \sqrt{\frac{2\delta_0 \log p}{n}} + \lambda_0 \hat{\sigma} \right) k}{n\kappa^2 \left(X, k, (1 + 2\epsilon_0) \left(\frac{\max \|X_{\cdot j}\|_2}{\min \|X_{\cdot j}\|_2} \right) \right)} \right\}. \tag{2.7.36}$$

The following lemma controls the radius of the confidence interval.

Lemma 6. *On the event* $G \cap S \cap B_1$, *there exists* p_0 *such that if* $p \geq p_0$,

$$\rho_1(k) \leq C\|\xi\|_2 \left(\frac{1}{\sqrt{n}} + \frac{k \log p}{n} \right) \sigma \leq \|\xi\|_2 \log p \left(\frac{1}{\sqrt{n}} + \frac{k \log p}{n} \right) \hat{\sigma}, \tag{2.7.37}$$

$$\rho_2(k) \leq Ck\sqrt{\frac{\log p}{n}} \sigma \leq \log p \left(k\sqrt{\frac{\log p}{n}} \hat{\sigma} \right). \tag{2.7.38}$$

In the following, we establish the coverage property of the proposed confidence interval. By the definition of $\widetilde{\mu}$ in (2.3.6), we have

$$\widetilde{\mu} - \xi^\intercal \beta = \frac{1}{n} \widehat{u}^\intercal X^\intercal \epsilon + \left(\xi^\intercal - \widehat{u}^\intercal \widehat{\Sigma} \right) \left(\widehat{\beta} - \beta \right). \tag{2.7.39}$$

We now construct a confidence interval for the variance term $\frac{1}{n}\widehat{u}^\intercal X^\intercal \epsilon$ by normal distribution and a high probability upper bound for the bias term $\left(\xi^\intercal - \widehat{u}^\intercal \widehat{\Sigma} \right) \left(\widehat{\beta} - \beta \right)$.

Since ϵ is independent of X and $\widehat{u}$ and $\widehat{\Sigma}$ is a function of X, we have $\frac{1}{n}\widehat{u}^\intercal X^\intercal \epsilon \mid X \sim N\left(0, \sigma^2 \frac{\widehat{u}^\intercal \widehat{\Sigma} \widehat{u}}{n}\right)$, and

$$\mathbb{P}_{\epsilon|X}\left(\frac{1}{n}\widehat{u}^\intercal X^\intercal \epsilon \in \left(-\sqrt{\frac{\widehat{u}^\intercal \widehat{\Sigma} \widehat{u}}{n}}\sigma z_{\alpha/2}, \sqrt{\frac{\widehat{u}^\intercal \widehat{\Sigma} \widehat{u}}{n}}\sigma z_{\alpha/2}\right)\middle| X\right) = 1-\alpha.$$

By (2.7.39), we have $\mathbb{P}_{\epsilon|X}\left(\xi^\intercal \beta \in \mathrm{CI}_0\left(Z,k\right)|X\right) = 1-\alpha$, where

$$\begin{aligned}\mathrm{CI}_0(Z,k) = \Bigg[&\widetilde{\mu} - \left(\xi^\intercal - \widehat{u}^\intercal \widehat{\Sigma}\right)\left(\widehat{\beta}-\beta\right) - \sqrt{\frac{\widehat{u}^\intercal \widehat{\Sigma} \widehat{u}}{n}}\sigma z_{\alpha/2}, \\ &\widetilde{\mu} - \left(\xi^\intercal - \widehat{u}^\intercal \widehat{\Sigma}\right)\left(\widehat{\beta}-\beta\right) + \sqrt{\frac{\widehat{u}^\intercal \widehat{\Sigma} \widehat{u}}{n}}\sigma z_{\alpha/2}\Bigg].\end{aligned}$$

Integrating with respect to X, we have

$$\mathbb{P}_\theta\left(\xi^\intercal \beta \in \mathrm{CI}_0\left(Z,k\right)\right) = \int \mathbb{P}_{\epsilon|x}\left(\xi^\intercal \beta \in \mathrm{CI}_0\left(Z,k\right)|x\right) f(x)dx = 1-\alpha. \tag{2.7.40}$$

Since $\left|\left(\xi^\intercal - \widehat{u}^\intercal \widehat{\Sigma}\right)\left(\widehat{\beta}-\beta\right)\right| \leq \|\xi^\intercal - \widehat{u}^\intercal \widehat{\Sigma}\|_\infty \|\widehat{\beta}-\beta\|_1$, on the event $S \cap G$, Lemma 5 and the constraint in (2.3.5) lead to

$$\|\xi^\intercal - \widehat{u}^\intercal \widehat{\Sigma}\|_\infty \|\widehat{\beta}-\beta\|_1 \leq \lambda_n \left(2+2\epsilon_0\right) \frac{\sqrt{n}}{\min \|X_{\cdot j}\|_2} l\left(Z,k\right), \tag{2.7.41}$$

where $l\left(Z,k\right)$ is defined in (2.7.36). On the event $G \cap S$, we also have $\sigma \leq \left(1+\nu_0\right)\hat{\sigma}$ and $\sigma^{ora} \leq \left(1+\nu_0\right)\sqrt{1+2\sqrt{\frac{\log p}{n}}+2\frac{\log p}{n}}\hat{\sigma}$. We define the following confidence interval to facilitate the discussion, $\mathrm{CI}_1\left(Z,k\right) = \left(\widetilde{\mu} - l_k, \widetilde{\mu} + l_k\right)$, where

$$l_k = \left(1+\nu_0\right)\sqrt{\frac{\widehat{u}^\intercal \widehat{\Sigma} \widehat{u}}{n}} z_{\alpha/2}\hat{\sigma} + C_1\left(X,k\right)\|\xi\|_2 k \frac{\log p}{n}\hat{\sigma}.$$

On the event $G \cap S$, we have

$$\mathrm{CI}_0\left(Z,k\right) \subset \mathrm{CI}_1\left(Z,k\right). \tag{2.7.42}$$

On the event S_2, if $p \geq \exp(2M_2)$, then $\hat{\sigma} \leq \frac{1}{1-\nu_0}\sigma \leq \frac{1}{1-\nu_0}M_2 < \log p$. Hence, the event A holds and $\mathrm{CI}_{\alpha}^{S}(\xi^{\intercal}\beta, Z) = [\widetilde{\mu} - \rho_1(k), \quad \widetilde{\mu} + \rho_1(k)]$. By Lemma 6, on the event $G \cap S \cap B_1$, if $p \geq \max\{p_0, \exp(2M_2)\}$, we have $\rho_1(k) = l_k$, and hence

$$\mathrm{CI}_1(Z, k) = \mathrm{CI}_{\alpha}^{S}(\xi^{\intercal}\beta, Z). \tag{2.7.43}$$

We have the following bound on the coverage probability,

$$\begin{aligned}
&\mathbb{P}_{\theta}\left(\left\{\xi^{\intercal}\beta \in \mathrm{CI}_{\alpha}^{S}(\xi^{\intercal}\beta, Z)\right\}\right) \geq \mathbb{P}_{\theta}\left(\left\{\xi^{\intercal}\beta \in \mathrm{CI}_0(Z, k)\right\} \cap S \cap G \cap B_1\right) \\
\geq &\mathbb{P}_{\theta}\left(\left\{\xi^{\intercal}\beta \in \mathrm{CI}_0(Z, k)\right\}\right) - \mathbb{P}_{\theta}\left((S \cap G \cap B_1)^c\right) = 1 - \alpha - \mathbb{P}_{\theta}\left((S \cap G \cap B_1)^c\right) \\
= &\mathbb{P}_{\theta}\left(S \cap G \cap B_1\right) - \alpha,
\end{aligned}$$

where the first inequality follows from (2.7.42) and (2.7.43) and the first equality follows from (2.7.40). Combined with Lemma 4, we establish (2.7.26). We control the expected length as follows,

$$\begin{aligned}
&\mathbb{E}_{\theta}L\left(\mathrm{CI}_{\alpha}^{S}(\xi^{\intercal}\beta, Z)\right) = \mathbb{E}_{\theta}L\left(\mathrm{CI}_{\alpha}^{S}(\xi^{\intercal}\beta, Z)\right)\mathbf{1}_{A} \\
= &\mathbb{E}_{\theta}L\left(\mathrm{CI}_{\alpha}^{S}(\xi^{\intercal}\beta, Z)\right)\mathbf{1}_{A\cap(S\cap G\cap B_1)} + \mathbb{E}_{\theta}L\left(\mathrm{CI}_{\alpha}^{S}(\xi^{\intercal}\beta, Z)\right)\mathbf{1}_{A\cap(S\cap G\cap B_1)^c} \\
\leq &C\|\xi\|_2\left(k\frac{\log p}{n} + \frac{1}{\sqrt{n}}\right)\sigma + \|\xi\|_2(\log p)^2\left(\frac{1}{\sqrt{n}} + \frac{k\log p}{n}\right)\mathbb{P}_{\theta}\left((S \cap G \cap B_1)^c\right) \\
\leq &C\|\xi\|_2\left(k\frac{\log p}{n} + \frac{1}{\sqrt{n}}\right)\left(\sigma + C\left(p^{1-\min\left\{\delta_0, C_1, c_0 C_0^2\right\}} + c'\exp(-cn)\right)(\log p)^2\right),
\end{aligned} \tag{2.7.44}$$

where the first inequality follows from (2.7.37) and second inequality follows from Lemma 4. If $\frac{\log p}{n} \leq c$, then $\left(p^{1-\min\left\{\delta_0, C_1, c_0 C_0^2\right\}} + c'\exp(-cn)\right)(\log p)^2 \to 0$, and hence $\mathbb{E}_{\theta}L\left(\mathrm{CI}_{\alpha}^{S}(\xi^{\intercal}\beta, Z)\right) \leq C\|\xi\|_2\left(k\frac{\log p}{n} + \frac{1}{\sqrt{n}}\right)M_2$. □

3

Accuracy Assessment for High-dimensional Linear Regression

3.1 Introduction

In many applications, the goal of statistical inference is not only to construct a good estimator, but also to provide a measure of accuracy for this estimator. In classical statistics, when the parameter of interest is one-dimensional, this is achieved in the form of a standard error or a confidence interval. A prototypical example is the inference for a binomial proportion, where often not only an estimate of the proportion but also its margin of error are given. Accuracy measures of an estimation procedure have also been used as a tool for the empirical selection of tuning parameters. A well known example is Stein's Unbiased Risk Estimate (SURE), which has been an effective tool for the construction of data-driven adaptive estimators in normal means estimation, nonparametric signal recovery, covariance matrix estimation, and other problems. See, for instance, Stein (1981); Li (1985); Donoho & Johnstone (1995); Cai & Zhou (2009); Yi & Zou (2013). The commonly used cross-validation methods can also be viewed as a useful tool based on the idea of empirical assessment of accuracy.

In this paper, we consider the problem of estimating the loss of a given estimator in the setting of high-dimensional linear regression, where one observes (X, y) with

$X \in \mathbb{R}^{n\times p}$ and $y \in \mathbb{R}^n$, and for $1 \le i \le n$,

$$y_i = X_{i\cdot}\beta + \epsilon_i.$$

Here $\beta \in \mathbb{R}^p$ is the regression vector, $X_{i\cdot} \overset{iid}{\sim} N_p(0, \Sigma)$ are the rows of X, and the errors $\epsilon_i \overset{iid}{\sim} N(0, \sigma^2)$ are independent of X. This high-dimensional linear model has been well studied in the literature, where the main focus has been on estimation of β. Several penalized/constrained ℓ_1 minimization methods, including Lasso (Tibshirani, 1996), Dantzig selector (Candès & Tao, 2007), scaled Lasso (Sun & Zhang, 2012) and square-root Lasso (Belloni et al., 2011), have been proposed. These methods have been shown to work well in applications and produce interpretable estimates of β when β is assumed to be sparse. Theoretically, with a properly chosen tuning parameter, these estimators achieve the optimal rate of convergence over collections of sparse parameter spaces. See, for example, Candès & Tao (2007); Sun & Zhang (2012); Belloni et al. (2011); Raskutti et al. (2011); Bickel et al. (2009); Bühlmann & van de Geer (2011); Verzelen (2012).

For a given estimator $\widehat{\beta}$, the ℓ_q loss $\|\widehat{\beta} - \beta\|_q^2$ with $1 \le q \le 2$ is commonly used as a metric of accuracy for $\widehat{\beta}$. We consider in the present paper both point and interval estimation of the ℓ_q loss $\|\widehat{\beta} - \beta\|_q^2$ for a given $\widehat{\beta}$. Note that the loss $\|\widehat{\beta} - \beta\|_q^2$ is a random quantity, depending on both the estimator $\widehat{\beta}$ and the parameter β. For such a random quantity, prediction and prediction interval are ususally used for point and interval estimation, respectively. However, we slightly abuse the terminologies in the present paper by using estimation and confidence interval to represent the point and interval estimators of the loss $\|\widehat{\beta} - \beta\|_q^2$. Since the ℓ_q loss depends on the estimator $\widehat{\beta}$, it is necessary to specify the estimator in the discussion of loss estimation. Throughout this paper, we restrict our attention to a broad collection of estimators $\widehat{\beta}$ that perform well at least at one interior point or a small subset of the parameter space. This collection of estimators includes most state-of-art estimators such as Lasso, Dantzig selector, scaled Lasso and square-root Lasso.

High-dimensional linear regression has been well studied in two settings. One is the setting with known design covariance matrix $\Sigma = \mathrm{I}$, known noise level $\sigma = \sigma_0$ and sparse β. See for example, Donoho et al. (2011); Bayati & Montanari (2012); Nickl & van de Geer (2013); Verzelen (2012); Thrampoulidis et al. (2015); Cai & Guo (2016b); Arias-Castro et al. (2011); Ingster et al. (2010). Another commonly considered setting is sparse β with unknown Σ and σ. We study point and interval estimation of the ℓ_q loss $\|\widehat{\beta} - \beta\|_q^2$ in both settings. Specifically, we consider the parameter space $\Theta_0(k)$ introduced in (3.2.3), which consists of k-sparse signals β with known design covariance matrix $\Sigma = \mathrm{I}$ and known noise level $\sigma = \sigma_0$, and $\Theta(k)$ defined in (3.2.4), which consists of k-sparse signals with unknown Σ and σ.

3.1.1 Our contributions

The present paper studies the minimax and adaptive estimation of the loss $\|\widehat{\beta} - \beta\|_q^2$ for a given estimator $\widehat{\beta}$ and the minimax expected length and adaptivity of confidence intervals for the loss. A major step in our analysis is to establish rate sharp lower bounds for the minimax estimation error and the minimax expected length of confidence intervals for the ℓ_q loss over $\Theta_0(k)$ and $\Theta(k)$ for a broad class of estimators of β, which contains the subclass of rate-optimal estimators. We then focus on the estimation of the loss of rate-optimal estimators and take the Lasso and scaled Lasso estimators as generic examples. For these rate-optimal estimators, we propose procedures for point estimation as well as confidence intervals for their ℓ_q losses. It is shown that the proposed procedures achieve the corresponding lower bounds up to a constant factor. These results together establish the minimax rates for estimating the ℓ_q loss of rate-optimal estimators over $\Theta_0(k)$ and $\Theta(k)$. The analysis shows interesting and significant differences between estimating the ℓ_2 loss and ℓ_q loss with $1 \le q < 2$ as well as between the two parameter spaces $\Theta(k)$ and $\Theta_0(k)$.

- The minimax rate for estimating $\|\widehat{\beta} - \beta\|_2^2$ over $\Theta_0(k)$ is $\min\left\{\frac{1}{\sqrt{n}}, k\frac{\log p}{n}\right\}$ and over $\Theta(k)$ is $k\frac{\log p}{n}$. So loss estimation is much easier with the prior information $\Sigma = \mathrm{I}$ and $\sigma = \sigma_0$ when $\frac{\sqrt{n}}{\log p} \ll k \lesssim \frac{n}{\log p}$.

- The minimax rate for estimating $\|\widehat{\beta} - \beta\|_q^2$ with $1 \le q < 2$ over both $\Theta_0(k)$ and $\Theta(k)$ is $k^{\frac{2}{q}}\frac{\log p}{n}$.

In the regime $\frac{\sqrt{n}}{\log p} \ll k \lesssim \frac{n}{\log p}$, a practical loss estimator is proposed for estimating the ℓ_2 loss and shown to achieve the optimal convergence rate $\frac{1}{\sqrt{n}}$ adaptively over $\Theta_0(k)$. We say *estimation of loss is impossible* if the minimax rate can be achieved by the trivial estimator 0, which means that the estimation accuracy of the loss is at least of the same order as the loss itself. In all other considered cases, estimation of loss is shown to be impossible. These results indicate that loss estimation is difficult.

We then turn to the construction of confidence intervals for the ℓ_q loss. A confidence interval for the loss is useful even when it is "impossible" to estimate the loss, as a confidence interval can provide non-trivial upper and lower bounds for the loss. In terms of convergence rate over $\Theta_0(k)$ or $\Theta(k)$, the minimax rate of the expected length of confidence intervals for the ℓ_q loss, $\|\widehat{\beta} - \beta\|_q^2$, of any rate-optimal estimator $\widehat{\beta}$ coincides with the minimax estimation rate. We also consider the adaptivity of confidence intervals for the ℓ_q loss of any rate-optimal estimator $\widehat{\beta}$. (The framework for adaptive confidence intervals is discussed in detail in Section 3.3.1.) Regarding confidence intervals for the ℓ_2 loss in the case of known $\Sigma = \mathrm{I}$ and $\sigma = \sigma_0$, a procedure is proposed and is shown to achieve the optimal length $\frac{1}{\sqrt{n}}$ adaptively over $\Theta_0(k)$ for $\frac{\sqrt{n}}{\log p} \lesssim k \lesssim \frac{n}{\log p}$. Furthermore, it is shown that this is the only regime where adaptive confidence intervals exist, even over two given parameter spaces. For example, when $k_1 \ll \frac{\sqrt{n}}{\log p}$ and $k_1 \ll k_2$, it is impossible to construct a confidence interval for the ℓ_2

loss with guaranteed coverage probability over $\Theta_0(k_2)$ (consequently also over $\Theta_0(k_1)$) and with the expected length automatically adjusted to the sparsity. Similarly, for the ℓ_q loss with $1 \leq q < 2$, construction of adaptive confidence intervals is impossible over $\Theta_0(k_1)$ and $\Theta_0(k_2)$ for $k_1 \ll k_2 \lesssim \frac{n}{\log p}$. Regarding confidence intervals for the ℓ_q loss with $1 \leq q \leq 2$ in the case of unknown Σ and σ, the impossibility of adaptivity also holds over $\Theta(k_1)$ and $\Theta(k_2)$ for $k_1 \ll k_2 \lesssim \frac{n}{\log p}$.

Establishing rate-optimal lower bounds requires the development of new technical tools. One main difference between loss estimation and the traditional parameter estimation is that for loss estimation the constraint is on the performance of the estimator $\widehat{\beta}$ of the regression vector β, but the lower bound is on the difficulty of estimating its loss $\|\widehat{\beta} - \beta\|_q^2$. We introduce useful new lower bound techniques for the minimax estimation error and the expected length of adaptive confidence intervals for the loss $\|\widehat{\beta} - \beta\|_q^2$. In several important cases, it is necessary to test a composite null against a composite alternative in order to establish rate sharp lower bounds. The technical tools developed in this paper can also be of independent interest.

In addition to $\Theta_0(k)$ and $\Theta(k)$, we also study an intermediate parameter space where the noise level σ is known and the design covariance matrix Σ is unknown but of certain structure. Lower bounds for the expected length of minimax and adaptive confidence intervals for $\|\widehat{\beta} - \beta\|_q^2$ over this parameter space are established for a broad collection of estimators $\widehat{\beta}$ and are shown to be rate sharp for the class of rate-optimal estimators. Furthermore, the lower bounds developed in this paper have wider implications. In particular, it is shown that they lead immediately to minimax lower bounds for estimating $\|\beta\|_q^2$ and the expected length of confidence intervals for $\|\beta\|_q^2$ with $1 \leq q \leq 2$.

3.1.2 Comparison with other works

Statistical inference on the loss of specific estimators of β has been considered in the recent literature. The papers Donoho et al. (2011); Bayati & Montanari (2012) established, in the setting $\Sigma = \mathrm{I}$ and $n/p \to \delta \in (0, \infty)$, the limit of the normalized loss $\frac{1}{p}\|\widehat{\beta}(\lambda) - \beta\|_2^2$ where $\widehat{\beta}(\lambda)$ is the Lasso estimator with a pre-specified tuning parameter λ. Although Donoho et al. (2011); Bayati & Montanari (2012) provided an exact asymptotic expression of the normalized loss, the limit itself depends on the unknown β. In a similar setting, the paper Thrampoulidis et al. (2015) established the limit of a normalized ℓ_2 loss of the square-root Lasso estimator. These limits of the normalized losses help understand the properties of the corresponding estimators of β, but they do not lead to an estimate of the loss. Our results imply that although these normalized losses have a limit under certain regularity conditions, such losses cannot be estimated well in most settings.

A recent paper, Janson et al. (2015), constructed a confidence interval for $\|\widehat{\beta} - \beta\|_2^2$ in the case of known $\Sigma = \mathrm{I}$, unknown noise level σ, and moderate dimension where $n/p \to \xi \in (0, 1)$ and no sparsity is assumed on β. While no sparsity assumption on β is imposed, their method requires the assumption of $\Sigma = \mathrm{I}$ and $n/p \to \xi \in (0, 1)$. In contrast, in this paper, we consider both unknown Σ and known $\Sigma = \mathrm{I}$ settings, while allowing $p \gg n$ and assuming sparse β.

Honest adaptive inference has been studied in the nonparametric function estimation literature, including Cai & Low (2005) for adaptive confidence intervals for linear functionals, Hoffmann & Nickl (2011); Cai et al. (2014) for adaptive confidence bands, and Cai & Low (2006); Robins & van der Vaart (2006) for adaptive confidence balls, and in the high-dimensional linear regression literature, including Nickl & van de Geer (2013) for adaptive confidence set and Cai & Guo (2016b) for adaptive confidence interval for linear functionals. In this paper, we develop new lower bound tools, Theorems 16 and 17, to establish the possibility of adaptive confidence intervals for $\|\widehat{\beta}-\beta\|_q^2$. The connection between ℓ_2 loss considered in the current paper and the work Nickl & van de Geer (2013) is discussed in more detail in Section 3.3.2.

3.1.3 Organization

Section 3.2 establishes the minimax lower bounds of estimating the loss $\|\widehat{\beta}-\beta\|_q^2$ with $1 \leq q \leq 2$ over both $\Theta_0(k)$ and $\Theta(k)$ and shows that these bounds are rate sharp for the Lasso and scaled Lasso estimators, respectively. We then turn to interval estimation of $\|\widehat{\beta}-\beta\|_q^2$. Sections 3.3 and 3.4 present the minimax and adaptive minimax lower bounds for the expected length of confidence intervals for $\|\widehat{\beta}-\beta\|_q^2$ over $\Theta_0(k)$ and $\Theta(k)$. For Lasso and scaled Lasso estimators, we show that the lower bounds can be achieved and investigate the possibility of adaptivity. Section 3.5 considers the rate-optimal estimators and establishes the minimax convergence rate of estimating their ℓ_q losses. Section 3.6 presents new minimax lower bound techniques for estimating the loss $\|\widehat{\beta}-\beta\|_q^2$. Section 3.7 discusses the minimaxity and adaptivity in another setting, where the noise level σ is known and the design covariance matrix Σ is unknown but of certain structure. Section 3.8 applies the newly developed lower bounds to establish lower bounds for a related problem, that of estimating $\|\beta\|_q^2$. Section 3.9 proves the main results and additional proofs are given in Chapter B.

3.1.4 Notation

For a matrix $X \in \mathbb{R}^{n\times p}$, $X_{i\cdot}$, $X_{\cdot j}$, and $X_{i,j}$ denote respectively the i-th row, j-th column, and (i,j) entry of the matrix X. For a subset $J \subset \{1,2,\cdots,p\}$, $|J|$ denotes the cardinality of J, J^c denotes the complement $\{1,2,\cdots,p\}\backslash J$, X_J denotes the submatrix of X consisting of columns $X_{\cdot j}$ with $j \in J$ and for a vector $x \in \mathbb{R}^p$, x_J

is the subvector of x with indices in J. For a vector $x \in \mathbb{R}^p$, $\text{supp}(x)$ denotes the support of x and the ℓ_q norm of x is defined as $\|x\|_q = (\sum_{i=1}^{p} |x_i|^q)^{\frac{1}{q}}$ for $q \geq 0$ with $\|x\|_0 = |\text{supp}(x)|$ and $\|x\|_\infty = \max_{1\leq j\leq p} |x_j|$. For $a \in \mathbb{R}$, $a_+ = \max\{a, 0\}$. We use $\max \|X_{\cdot j}\|_2$ as a shorthand for $\max_{1\leq j\leq p} \|X_{\cdot j}\|_2$ and $\min \|X_{\cdot j}\|_2$ as a shorthand for $\min_{1\leq j\leq p} \|X_{\cdot j}\|_2$. For a matrix A, we define the spectral norm $\|A\|_2 = \sup_{\|x\|_2=1} \|Ax\|_2$ and the matrix ℓ_1 norm $\|A\|_{L_1} = \sup_{1\leq j\leq p} \sum_{i=1}^{p} |A_{ij}|$; For a symmetric matrix A, $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$ denote respectively the smallest and largest eigenvalue of A. We use c and C to denote generic positive constants that may vary from place to place. For two positive sequences a_n and b_n, $a_n \lesssim b_n$ means $a_n \leq C b_n$ for all n and $a_n \gtrsim b_n$ if $b_n \lesssim a_n$ and $a_n \asymp b_n$ if $a_n \lesssim b_n$ and $b_n \lesssim a_n$, and $a_n \ll b_n$ if $\limsup_{n\to\infty} \frac{a_n}{b_n} = 0$ and $a_n \gg b_n$ if $b_n \ll a_n$.

3.2 Minimax estimation of the ℓ_q loss

We begin by presenting the minimax framework for estimating the ℓ_q loss, $\|\widehat{\beta}-\beta\|_q^2$, of a given estimator $\widehat{\beta}$, and then establish the minimax lower bounds for the estimation error for a broad collection of estimators $\widehat{\beta}$. We also show that such minimax lower bounds can be achieved for the Lasso and scaled Lasso estimators.

3.2.1 Problem formulation

Recall the high-dimensional linear model,

$$y_{n\times 1} = X_{n\times p}\beta_{p\times 1} + \epsilon_{n\times 1}, \quad \epsilon \sim N_n(0, \sigma^2 \mathrm{I}). \tag{3.2.1}$$

We focus on the random design with $X_{i\cdot} \overset{iid}{\sim} N(0, \Sigma)$ and $X_{i\cdot}$ and ϵ_i are independent. Let $Z = (X, y)$ denote the observed data and $\widehat{\beta}$ be a given estimator of β. Denoting by $\widehat{L}_q(Z)$ any estimator of the loss $\|\widehat{\beta} - \beta\|_q^2$, the minimax rate of convergence for estimating $\|\widehat{\beta} - \beta\|_q^2$ over a parameter space Θ is defined as the largest quantity $\gamma_{\widehat{\beta},\ell_q}(\Theta)$ such that

$$\inf_{\widehat{L}_q} \sup_{\theta \in \Theta} \mathbb{P}_\theta \left(|\widehat{L}_q(Z) - \|\widehat{\beta} - \beta\|_q^2| \geq \gamma_{\widehat{\beta},\ell_q}(\Theta) \right) \geq \delta, \tag{3.2.2}$$

for some constant $\delta > 0$ not depending on n or p. We shall write $\widehat{L}_q$ for $\widehat{L}_q(Z)$ when there is no confusion.

We denote the parameter by $\theta = (\beta, \Sigma, \sigma)$, which consists of the signal β, the design covariance matrix Σ and the noise level σ. For a given $\theta = (\beta, \Sigma, \sigma)$, we use $\beta(\theta)$ to denote the corresponding β. Two settings are considered: The first is known design covariance matrix $\Sigma = \mathrm{I}$ and known noise level $\sigma = \sigma_0$ and the other is unknown Σ and σ. In the first setting, we consider the following parameter space that consists of k-sparse signals,

$$\Theta_0(k) = \left\{ (\beta, \mathrm{I}, \sigma_0) : \|\beta\|_0 \leq k \right\}, \tag{3.2.3}$$

and in the second setting, we consider

$$\Theta(k) = \left\{ (\beta, \Sigma, \sigma) : \|\beta\|_0 \leq k,\ \frac{1}{M_1} \leq \lambda_{\min}(\Sigma) \leq \lambda_{\max}(\Sigma) \leq M_1,\ 0 < \sigma \leq M_2 \right\}, \tag{3.2.4}$$

where $M_1 \geq 1$ and $M_2 > 0$ are constants. The parameter space $\Theta_0(k)$ is a subset of $\Theta(k)$, which consists of k-sparse signals with unknown Σ and σ.

The minimax rate $\gamma_{\widehat{\beta},\ell_q}(\Theta)$ for estimating $\|\widehat{\beta} - \beta\|_q^2$ also depends on the estimator $\widehat{\beta}$. Different estimators $\widehat{\beta}$ could lead to different losses $\|\widehat{\beta} - \beta\|_q^2$ and in general the difficulty of estimating the loss $\|\widehat{\beta} - \beta\|_q^2$ varies with $\widehat{\beta}$. We first recall the properties of some state-of-art estimators and then specify the collection of estimators on which we focus in this paper. As shown in Candès & Tao (2007); Bickel et al. (2009);

Belloni et al. (2011); Sun & Zhang (2012), Lasso, Dantzig Selector, scaled Lasso and square-root Lasso satisfy the following property if the tuning parameter is properly chosen,

$$\sup_{\theta \in \Theta(k)} \mathbb{P}_\theta \left(\|\widehat{\beta} - \beta\|_q^2 \geq C k^{\frac{2}{q}} \frac{\log p}{n} \right) \to 0, \tag{3.2.5}$$

where $C > 0$ is a constant. The minimax lower bounds established in Verzelen (2012); Raskutti et al. (2011); Ye & Zhang (2010) imply that $k^{\frac{2}{q}} \frac{\log p}{n}$ is the optimal rate for estimating β over the parameter space $\Theta(k)$. It should be stressed that all of these algorithms do not require knowledge of the sparsity k and are thus adaptive to the sparsity provided $k \lesssim \frac{n}{\log p}$. We consider a broad collection of estimators $\widehat{\beta}$ satisfying one of the following two assumptions.

(A1) The estimator $\widehat{\beta}$ satisfies, for some $\theta_0 = (\beta^*, \mathrm{I}, \sigma_0)$,

$$\mathbb{P}_{\theta_0} \left(\|\widehat{\beta} - \beta^*\|_q^2 \geq C^* \|\beta^*\|_0^{\frac{2}{q}} \frac{\log p}{n} \sigma_0^2 \right) \leq \alpha_0, \tag{3.2.6}$$

where $0 \leq \alpha_0 < \frac{1}{4}$ and $C^* > 0$ are constants.

(A2) The estimator $\widehat{\beta}$ satisfies

$$\sup_{\{\theta = (\beta^*, \mathrm{I}, \sigma): \sigma \leq 2\sigma_0\}} \mathbb{P}_\theta \left(\|\widehat{\beta} - \beta^*\|_q^2 \geq C^* \|\beta^*\|_0^{\frac{2}{q}} \frac{\log p}{n} \sigma^2 \right) \leq \alpha_0, \tag{3.2.7}$$

where $0 \leq \alpha_0 < \frac{1}{4}$ and $C^* > 0$ are constants and $\sigma_0 > 0$ is given.

In view of the minimax rate given in (3.2.5), Assumption (A1) requires $\widehat{\beta}$ to be a good estimator of β at at least one point $\theta_0 \in \Theta_0(k)$. Assumption (A2) is slightly stronger than (A1) and requires $\widehat{\beta}$ to estimate β well for a single β^* but over a range of noise levels $\sigma \leq 2\sigma_0$ while $\Sigma = \mathrm{I}$. Of course, any estimator $\widehat{\beta}$ satisfying (3.2.5) satisfies both (A1) and (A2). In addition to Assumptions (A1) and (A2), we also introduce the following sparsity assumptions that will be used in various theorems.

(B1) Let c_0 be the constant defined in (3.9.14). The sparsity levels k and k_0 satisfy $k \leq c_0 \min\{p^{\gamma}, \frac{n}{\log p}\}$ for some constant $0 \leq \gamma < \frac{1}{2}$ and $k_0 \leq c_0 \min\{k, \frac{\sqrt{n}}{\log p}\}$.

(B2) The sparsity levels k_1, k_2 and k_0 satisfy $k_1 \leq k_2 \leq c_0 \min\{p^{\gamma}, \frac{n}{\log p}\}$ for some constant $0 \leq \gamma < \frac{1}{2}$ and $c_0 > 0$ and $k_0 \leq c_0 \min\{k_1, \frac{\sqrt{n}}{\log p}\}$.

3.2.2 Minimax estimation of the ℓ_q loss over $\Theta_0(k)$

The following theorem establishes the minimax lower bounds for estimating the loss $\|\widehat{\beta} - \beta\|_q^2$ over the parameter space $\Theta_0(k)$.

Theorem 9. *Suppose that the sparsity levels k and k_0 satisfy Assumption* (B1). *For any estimator $\widehat{\beta}$ satisfying Assumption* (A1) *with $\|\beta^*\|_0 \leq k_0$,*

$$\inf_{\widehat{L}_2} \sup_{\theta \in \Theta_0(k)} \mathbb{P}_{\theta} \left(|\widehat{L}_2 - \|\widehat{\beta} - \beta\|_2^2| \geq c \min\left\{ k\frac{\log p}{n}, \frac{1}{\sqrt{n}} \right\} \sigma_0^2 \right) \geq \delta. \tag{3.2.8}$$

For any estimator $\widehat{\beta}$ satisfying Assumption (A2) *with $\|\beta^*\|_0 \leq k_0$,*

$$\inf_{\widehat{L}_q} \sup_{\theta \in \Theta_0(k)} \mathbb{P}_{\theta} \left(|\widehat{L}_q - \|\widehat{\beta} - \beta\|_q^2| \geq c k^{\frac{2}{q}} \frac{\log p}{n} \sigma_0^2 \right) \geq \delta, \quad \text{for } 1 \leq q < 2, \tag{3.2.9}$$

where $\delta > 0$ and $c > 0$ are constants.

Remark 5. Assumption (A1) restricts our focus to estimators that can perform well at at least one point $(\beta^*, \mathrm{I}, \sigma_0) \in \Theta_0(k)$. This weak condition makes the established lower bounds widely applicable as the benchmark for evaluating estimators of the ℓ_q loss of any $\widehat{\beta}$ that performs well at a proper subset, or even a single point of the whole parameter space.

In this paper, we focus on estimating the loss $\|\widehat{\beta} - \beta\|_q^2$ with $1 \leq q \leq 2$. Similar results can be established for the loss in the form of $\|\widehat{\beta} - \beta\|_q^q$ with $1 \leq q \leq 2$; Under the same assumptions as those in Theorem 9, the lower bounds for estimating the loss

$\|\widehat{\beta}-\beta\|_q^q$ hold with replacing the convergence rates with their $\frac{q}{2}$ power; that is, (3.2.8) remains the same while the convergence rate $k^{\frac{2}{q}}(\sqrt{\log p/n}\sigma_0)^2$ in (3.2.9) is replaced by $k(\sqrt{\log p/n}\sigma_0)^q$. Similarly, all the results established in the rest of the paper for $\|\widehat{\beta}-\beta\|_q^2$ hold for $\|\widehat{\beta}-\beta\|_q^q$ with corresponding convergence rates replaced by their $\frac{q}{2}$ power.

Theorem 9 establishes the minimax lower bounds for estimating the ℓ_2 loss $\|\widehat{\beta}-\beta\|_2^2$ of any estimator $\widehat{\beta}$ satisfying Assumption (A1) and the ℓ_q loss $\|\widehat{\beta}-\beta\|_q^2$ with $1 \leq q < 2$ of any estimator $\widehat{\beta}$ satisfying Assumption (A2). We will take the Lasso estimator as an example and demonstrate the implications of the above theorem. We randomly split $Z = (y, X)$ into subsamples $Z^{(1)} = \left(y^{(1)}, X^{(1)}\right)$ and $Z^{(2)} = \left(y^{(2)}, X^{(2)}\right)$ with sample sizes n_1 and n_2, respectively. The Lasso estimator $\widehat{\beta}^L$ based on the first subsample $Z^{(1)} = \left(y^{(1)}, X^{(1)}\right)$ is defined as

$$\widehat{\beta}^L = \underset{\beta\in\mathbb{R}^p}{\arg\min} \frac{\|y^{(1)} - X^{(1)}\beta\|_2^2}{n_1} + \lambda \sum_{j=1}^{p} \frac{\|X^{(1)}_{\cdot j}\|_2}{\sqrt{n_1}} |\beta_j|, \tag{3.2.10}$$

where $\lambda = A\sqrt{\log p/n_1}\sigma_0$ with $A > \sqrt{2}$ being a pre-specified constant. Without loss of generality, we assume $n_1 \asymp n_2$. For the case $1 \leq q < 2$, (3.2.5) and (3.2.9) together imply that the estimation of the ℓ_q loss $\|\widehat{\beta}^L - \beta\|_q^2$ is impossible since the lower bound can be achieved by the trivial estimator of the loss, 0. That is, $\sup_{\theta\in\Theta_0(k)} \mathbb{P}_\theta \left(|0 - \|\widehat{\beta}^L - \beta\|_q^2| \geq Ck^{\frac{2}{q}} \frac{\log p}{n}\right) \to 0$.

For the case $q = 2$, in the regime $k \ll \frac{\sqrt{n}}{\log p}$, the lower bound $\frac{k \log p}{n}$ in (3.2.8) can be achieved by the zero estimator and hence estimation of the loss $\|\widehat{\beta}^L-\beta\|_2^2$ is impossible. However, the interesting case is when $\frac{\sqrt{n}}{\log p} \lesssim k \lesssim \frac{n}{\log p}$, the loss estimator $\widetilde{L}_2$ proposed in (3.2.11) achieves the minimax lower bound $\frac{1}{\sqrt{n}}$ in (3.2.8), which cannot be achieved by the zero estimator. We now detail the construction of the loss estimator $\widetilde{L}_2$. Based on the second half sample $Z^{(2)} = \left(y^{(2)}, X^{(2)}\right)$, we propose the following estimator,

$$\widetilde{L}_2 = \left(\frac{1}{n_2}\left\|y^{(2)} - X^{(2)}\widehat{\beta}^L\right\|_2^2 - \sigma_0^2\right)_+. \tag{3.2.11}$$

Note that the first subsample $Z^{(1)} = \left(y^{(1)}, X^{(1)}\right)$ is used to produce the Lasso estimator $\widehat{\beta}^L$ in (3.2.10) and the second subsample $Z^{(2)} = \left(y^{(2)}, X^{(2)}\right)$ is retained to evaluate the loss $\|\widehat{\beta}^L - \beta\|_2^2$. Such sample splitting technique is similar to cross-validation and has been used in Nickl & van de Geer (2013) for constructing confidence sets for β and in Janson et al. (2015) for confidence intervals for the ℓ_2 loss.

The following proposition establishes that the estimator $\widetilde{L}_2$ achieves the minimax lower bound of (3.2.8) over the regime $\frac{\sqrt{n}}{\log p} \lesssim k \lesssim \frac{n}{\log p}$.

Proposition 2. *Suppose that $k \lesssim \frac{n}{\log p}$ and $\widehat{\beta}^L$ is the Lasso estimator defined in* (3.2.10) *with $A > \sqrt{2}$, then the estimator of loss proposed in* (3.2.11) *satisfies, for any sequence $\delta_{n,p} \to \infty$,*

$$\limsup_{n,p\to\infty} \sup_{\theta\in\Theta_0(k)} \mathbb{P}_\theta\left(\left|\widetilde{L}_2 - \|\widehat{\beta}^L - \beta\|_2^2\right| \geq \delta_{n,p}\frac{1}{\sqrt{n}}\right) = 0. \tag{3.2.12}$$

3.2.3 Minimax estimation of the ℓ_q loss over $\Theta(k)$

We now turn to the case of unknown Σ and σ and establish the minimax lower bound for estimating the ℓ_q loss over the parameter space $\Theta(k)$.

Theorem 10. *Suppose that the sparsity levels k and k_0 satisfy Assumption* (B1)*. For any estimator $\widehat{\beta}$ satisfying Assumption* (A1) *with $\|\beta^*\|_0 \leq k_0$,*

$$\inf_{\widehat{L}_q} \sup_{\theta\in\Theta(k)} \mathbb{P}_\theta\left(|\widehat{L}_q - \|\widehat{\beta} - \beta\|_q^2| \geq ck^{\frac{2}{q}}\frac{\log p}{n}\right) \geq \delta, \quad 1 \leq q \leq 2, \tag{3.2.13}$$

where $\delta > 0$ and $c > 0$ are constants.

Theorem 10 provides a minimax lower bound for estimating the ℓ_q loss of any estimator $\widehat{\beta}$ satisfying Assumption (A1), including the scaled Lasso estimator defined as

$$\{\widehat{\beta}^{SL}, \hat{\sigma}\} = \underset{\beta \in \mathbb{R}^p, \sigma \in \mathbb{R}^+}{\arg\min} \frac{\|y - X\beta\|_2^2}{2n\sigma} + \frac{\sigma}{2} + \lambda_0 \sum_{j=1}^{p} \frac{\|X_{\cdot j}\|_2}{\sqrt{n}} |\beta_j|, \tag{3.2.14}$$

where $\lambda_0 = A\sqrt{\log p/n}$ with $A > \sqrt{2}$. Note that for the scaled Lasso estimator, the lower bound in (3.2.13) can be achieved by the trivial loss estimator 0, in the sense, $\sup_{\theta \in \Theta(k)} \mathbb{P}_\theta \left(|0 - \|\widehat{\beta}^{SL} - \beta\|_q^2| \geq C k^{\frac{2}{q}} \frac{\log p}{n} \right) \to 0$, and hence estimation of loss is impossible in this case.

3.3 Minimaxity and adaptivity of confidence intervals over $\Theta_0(k)$

We focused in the last section on point estimation of the ℓ_q loss and showed the impossibility of loss estimation except for one regime. The results naturally lead to another question: Is it possible to construct "useful" confidence intervals for $\|\widehat{\beta} - \beta\|_q^2$ that can provide non-trivial upper and lower bounds for the loss? In this section, after introducing the framework for minimaxity and adaptivity of confidence intervals, we consider the case of known $\Sigma = \mathrm{I}$ and $\sigma = \sigma_0$ and establish the minimaxity and adaptivity lower bounds for the expected length of confidence intervals for the ℓ_q loss of a broad collection of estimators over the parameter space $\Theta_0(k)$. We also show that such minimax lower bounds can be achieved for the Lasso estimator and then discuss the possibility of adaptivity using the Lasso estimator as an example. The case of unknown Σ and σ will be the focus of the next section.

3.3.1 Framework for minimaxity and adaptivity of confidence intervals

In this section, we introduce the following decision theoretical framework for confidence intervals of the loss $\|\widehat{\beta}-\beta\|_q^2$. Given $0<\alpha<1$ and the parameter space Θ and the loss $\|\widehat{\beta}-\beta\|_q^2$, denote by $\mathcal{I}_\alpha\left(\Theta,\widehat{\beta},\ell_q\right)$ the set of all $(1-\alpha)$ level confidence intervals for $\|\widehat{\beta}-\beta\|_q^2$ over Θ,

$$\mathcal{I}_\alpha\left(\Theta,\widehat{\beta},\ell_q\right)=\left\{\mathrm{CI}_\alpha\left(\widehat{\beta},\ell_q,Z\right)=\left[l\left(Z\right),u\left(Z\right)\right]:\inf_{\theta\in\Theta}\mathbb{P}_\theta\left(\|\widehat{\beta}-\beta(\theta)\|_q^2\in \mathrm{CI}_\alpha\left(\widehat{\beta},\ell_q,Z\right)\right)\geq 1-\alpha\right\}. \tag{3.3.1}$$

We will write CI_α for $\mathrm{CI}_\alpha\left(\widehat{\beta},\ell_q,Z\right)$ when there is no confusion. For any confidence interval $\mathrm{CI}_\alpha\left(\widehat{\beta},\ell_q,Z\right)=\left[l\left(Z\right),u\left(Z\right)\right]$, its length is denoted by $\mathbf{R}\left(\mathrm{CI}_\alpha\left(\widehat{\beta},\ell_q,Z\right)\right)=u\left(Z\right)-l\left(Z\right)$ and the maximum expected length over a parameter space Θ_1 is defined as

$$\mathbf{R}\left(\mathrm{CI}_\alpha\left(\widehat{\beta},\ell_q,Z\right),\Theta_1\right)=\sup_{\theta\in\Theta_1}\mathbb{E}_\theta\mathbf{R}\left(\mathrm{CI}_\alpha\left(\widehat{\beta},\ell_q,Z\right)\right). \tag{3.3.2}$$

For two nested parameter spaces $\Theta_1\subseteq\Theta_2$, we define the benchmark $\mathbf{R}_\alpha^*\left(\Theta_1,\Theta_2,\widehat{\beta},\ell_q\right)$, measuring the degree of adaptivity over the nested spaces $\Theta_1\subset\Theta_2$,

$$\mathbf{R}_\alpha^*\left(\Theta_1,\Theta_2,\widehat{\beta},\ell_q\right)=\inf_{\mathrm{CI}_\alpha\left(\widehat{\beta},\ell_q,Z\right)\in\mathcal{I}_\alpha\left(\Theta_2,\widehat{\beta},\ell_q\right)}\sup_{\theta\in\Theta_1}\mathbb{E}_\theta\mathbf{R}\left(\mathrm{CI}_\alpha\left(\widehat{\beta},\ell_q,Z\right)\right). \tag{3.3.3}$$

We will write $\mathbf{R}_\alpha^*\left(\Theta_1,\widehat{\beta},\ell_q\right)$ for $\mathbf{R}_\alpha^*\left(\Theta_1,\Theta_1,\widehat{\beta},\ell_q\right)$, which is the minimax expected length of confidence intervals for $\|\widehat{\beta}-\beta\|_q^2$ over Θ_1. The benchmark $\mathbf{R}_\alpha^*\left(\Theta_1,\Theta_2,\widehat{\beta},\ell_q\right)$ is the infimum of the maximum expected length over Θ_1 among all $(1-\alpha)$-level confidence intervals over Θ_2. In contrast, $\mathbf{R}_\alpha^*\left(\Theta_1,\widehat{\beta},\ell_q\right)$ is considering all $(1-\alpha)$-level confidence intervals over Θ_1. In words, if there is prior information that the parameter lies in the smaller parameter space Θ_1, $\mathbf{R}_\alpha^*\left(\Theta_1,\widehat{\beta},\ell_q\right)$ measures the benchmark length of confidence intervals over the parameter space Θ_1, which is illustrated in the left of Figure 3.1; however, if there is only prior information that the parameter lies in the larger parameter space Θ_2, $\mathbf{R}^*_\alpha\left(\Theta_1, \Theta_2, \widehat{\beta}, \ell_q\right)$ measures the benchmark length of confidence intervals over the parameter space Θ_1, which is illustrated in the right of Figure 3.1.

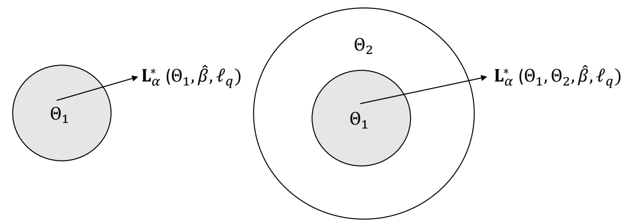

Figure 3.1: The plot demonstrates definitions of $\mathbf{R}^*_\alpha\left(\Theta_1, \widehat{\beta}, \ell_q\right)$ and $\mathbf{R}^*_\alpha\left(\Theta_1, \Theta_2, \widehat{\beta}, \ell_q\right)$.

Rigorously, we define a confidence interval CI^* to be simultaneously adaptive over Θ_1 and Θ_2 if $\mathrm{CI}^* \in \mathcal{I}_\alpha\left(\Theta_2, \widehat{\beta}, \ell_q\right)$,

$$\mathbf{R}\left(\mathrm{CI}^*, \Theta_1\right) \asymp \mathbf{R}^*_\alpha\left(\Theta_1, \widehat{\beta}, \ell_q\right), \quad \text{and} \quad \mathbf{R}\left(\mathrm{CI}^*, \Theta_2\right) \asymp \mathbf{R}^*_\alpha\left(\Theta_2, \widehat{\beta}, \ell_q\right). \tag{3.3.4}$$

The condition (3.3.4) means that the confidence interval CI^*, which has coverage over the larger parameter space Θ_2, achieves the minimax rate over both Θ_1 and Θ_2. Note that $\mathbf{R}\left(\mathrm{CI}^*, \Theta_1\right) \geq \mathbf{R}^*_\alpha\left(\Theta_1, \Theta_2, \widehat{\beta}, \ell_q\right)$. If $\mathbf{R}^*_\alpha\left(\Theta_1, \Theta_2, \widehat{\beta}, \ell_q\right) \gg \mathbf{R}^*_\alpha\left(\Theta_1, \widehat{\beta}, \ell_q\right)$, then the rate-optimal adaptation (3.3.4) is impossible to achieve for $\Theta_1 \subset \Theta_2$. Otherwise, it is possible to construct confidence intervals simultaneously adaptive over parameter spaces Θ_1 and Θ_2. The possibility of adaptation over parameter spaces Θ_1 and Θ_2 can thus be answered by investigating the benchmark quantities $\mathbf{R}^*_\alpha\left(\Theta_1, \widehat{\beta}, \ell_q\right)$ and $\mathbf{R}^*_\alpha\left(\Theta_1, \Theta_2, \widehat{\beta}, \ell_q\right)$. Such framework has already been introduced in Cai & Guo (2016b), which studies the minimaxity and adaptivity of confidence intervals for linear functionals in high-dimensional linear regression.

We will adopt the minimax and adaptation framework discussed above and establish the minimax expected length $\mathbf{R}_\alpha^*\left(\Theta_0(k), \widehat{\beta}, \ell_q\right)$ and the adaptation benchmark $\mathbf{R}_\alpha^*\left(\Theta_0(k_1), \Theta_0(k_2), \widehat{\beta}, \ell_q\right)$. In terms of the minimax expected length and the adaptivity behavior, there exist fundamental differences between the case $q = 2$ and $1 \le q < 2$. We will discuss them separately in the following two subsections.

3.3.2 Confidence intervals for the ℓ_2 loss over $\Theta_0(k)$

The following theorem establishes the minimax lower bound for the expected length of confidence intervals of $\|\widehat{\beta} - \beta\|_2^2$ over the parameter space $\Theta_0(k)$.

Theorem 11. *Suppose that $0 < \alpha < \frac{1}{4}$ and the sparsity levels k and k_0 satisfy Assumption* (B1). *For any estimator $\widehat{\beta}$ satisfying Assumption* (A1) *with $\|\beta^*\|_0 \le k_0$, then there is some constant $c > 0$ such that*

$$\mathbf{R}_\alpha^*\left(\Theta_0(k), \widehat{\beta}, \ell_2\right) \ge c \min\left\{\frac{k \log p}{n}, \frac{1}{\sqrt{n}}\right\} \sigma_0^2. \tag{3.3.5}$$

In particular, if $\widehat{\beta}^L$ is the Lasso estimator defined in (3.2.10) *with $A > \sqrt{2}$, then the minimax expected length for $(1-\alpha)$ level confidence intervals of $\|\widehat{\beta}^L - \beta\|_2^2$ over $\Theta_0(k)$ is*

$$\mathbf{R}_\alpha^*\left(\Theta_0(k), \widehat{\beta}^L, \ell_2\right) \asymp \min\left\{\frac{k \log p}{n}, \frac{1}{\sqrt{n}}\right\} \sigma_0^2. \tag{3.3.6}$$

We now consider adaptivity of confidence intervals for the ℓ_2 loss. The following theorem gives the lower bound for the benchmark $\mathbf{R}_\alpha^*\left(\Theta_0(k_1), \Theta_0(k_2), \widehat{\beta}, \ell_2\right)$. We will then discuss Theorems 11 and 12 together.

Theorem 12. *Suppose that $0 < \alpha < \frac{1}{4}$ and the sparsity levels k_1, k_2 and k_0 satisfy Assumption* (B2). *For any estimator $\widehat{\beta}$ satisfying Assumption* (A1) *with $\|\beta^*\|_0 \le k_0$, then there is some constant $c > 0$ such that*

$$\mathbf{R}_\alpha^*\left(\Theta_0(k_1), \Theta_0(k_2), \widehat{\beta}, \ell_2\right) \ge c \min\left\{\frac{k_2 \log p}{n}, \frac{1}{\sqrt{n}}\right\} \sigma_0^2. \tag{3.3.7}$$

In particular, if $\widehat{\beta}^L$ *is the Lasso estimator defined in* (3.2.10) *with* $A > \sqrt{2}$*, the above lower bound can be achieved.*

The lower bound established in Theorem 12 implies that of Theorem 11 and both lower bounds hold for a general class of estimators satisfying Assumption (A1). There is a phase transition for the lower bound of the benchmark $\mathbf{R}_{\alpha}^{*}\left(\Theta_0(k_1), \Theta_0(k_2), \widehat{\beta}, \ell_2\right)$. In the regime $k_2 \ll \frac{\sqrt{n}}{\log p}$, the lower bound in (3.3.7) is $\frac{k_2 \log p}{n}\sigma_0^2$; when $\frac{\sqrt{n}}{\log p} \lesssim k_2 \lesssim \frac{n}{\log p}$, the lower bound in (3.3.7) is $\frac{1}{\sqrt{n}}\sigma_0^2$. For the Lasso estimator $\widehat{\beta}^L$ defined in (3.2.10), the lower bound $\frac{k \log p}{n}\sigma_0^2$ in (3.3.5) and $\frac{k_2 \log p}{n}\sigma_0^2$ in (3.3.7) can be achieved by the confidence intervals $\mathrm{CI}_{\alpha}^{0}(Z, k, 2)$ and $\mathrm{CI}_{\alpha}^{0}(Z, k_2, 2)$ defined in (3.3.15), respectively. Applying a similar idea to (3.2.11), we show that the minimax lower bound $\frac{1}{\sqrt{n}}\sigma_0^2$ in (3.3.6) and (3.3.7) can be achieved by the following confidence interval,

$$\mathrm{CI}_{\alpha}^{1}(Z) = \left(\left(\frac{\psi(Z)}{\frac{1}{n_2}\chi^2_{1-\frac{\alpha}{2}}(n_2)} - \sigma_0^2\right)_+, \left(\frac{\psi(Z)}{\frac{1}{n_2}\chi^2_{\frac{\alpha}{2}}(n_2)} - \sigma_0^2\right)_+\right), \tag{3.3.8}$$

where $\chi^2_{1-\frac{\alpha}{2}}(n_2)$ and $\chi^2_{\frac{\alpha}{2}}(n_2)$ are the $1-\frac{\alpha}{2}$ and $\frac{\alpha}{2}$ quantiles of χ^2 random variable with n_2 degrees of freedom, respectively, and

$$\psi(Z) = \min\left\{\frac{1}{n_2}\left\|y^{(2)} - X^{(2)}\widehat{\beta}^L\right\|_2^2, \sigma_0^2 \log p\right\}. \tag{3.3.9}$$

Note that the two-sided confidence interval (3.3.8) is simply based on the observed data Z, not depending on any prior knowledge of the sparsity k. Furthermore, it is a two-sided confidence interval, which tells not only just an upper bound, but also a lower bound for the loss. The coverage property and the expected length of $\mathrm{CI}_{\alpha}^{1}(Z)$ are established in the following proposition.

Proposition 3. *Suppose* $k \lesssim \frac{n}{\log p}$ *and* $\widehat{\beta}^L$ *is the estimator defined in* (3.2.10) *with*

$A > \sqrt{2}$. *Then* $\mathrm{CI}_\alpha^1(Z)$ *defined in* (3.3.8) *satisfies,*

$$\liminf_{n,p\to\infty} \inf_{\theta\in\Theta_0(k)} \mathbb{P}_\theta\left(\|\widehat{\beta}^L-\beta\|_2^2 \in \mathrm{CI}_\alpha^1(Z)\right) \geq 1-\alpha, \tag{3.3.10}$$

and

$$\mathbf{R}\left(\mathrm{CI}_\alpha^1(Z), \Theta_0(k)\right) \lesssim \frac{1}{\sqrt{n}}\sigma_0^2. \tag{3.3.11}$$

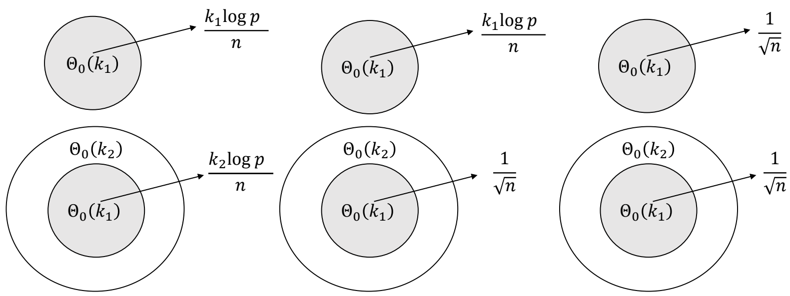

Figure 3.2: Illustration of $\mathbf{R}_\alpha^*\left(\Theta_0(k_1), \widehat{\beta}^L, \ell_2\right)$ (top) and $\mathbf{R}_\alpha^*\left(\Theta_0(k_1), \Theta_0(k_2), \widehat{\beta}^L, \ell_2\right)$(bottom) over regimes $k_1 \leq k_2 \lesssim \frac{\sqrt{n}}{\log p}$ (leftmost), $k_1 \lesssim \frac{\sqrt{n}}{\log p} \lesssim k_2 \lesssim \frac{n}{\log p}$(middle) and $\frac{\sqrt{n}}{\log p} \lesssim k_1 \leq k_2 \lesssim \frac{n}{\log p}$ (rightmost).

Regarding the Lasso estimator $\widehat{\beta}^L$ defined in (3.2.10), we will discuss the possibility of adaptivity of confidence intervals for $\|\widehat{\beta}^L-\beta\|_2^2$. The adaptivity behavior of confidence intervals for $\|\widehat{\beta}^L-\beta\|_2^2$ is demonstrated in Figure 3.2. As illustrated in the rightmost plot of Figure 3.2, in the regime $\frac{\sqrt{n}}{\log p} \lesssim k_1 \leq k_2 \lesssim \frac{n}{\log p}$, we obtain $\mathbf{R}_\alpha^*\left(\Theta_0(k_1), \Theta_0(k_2), \widehat{\beta}^L, \ell_2\right) \asymp \mathbf{R}_\alpha^*\left(\Theta_0(k_1), \widehat{\beta}^L, \ell_2\right) \asymp \frac{1}{\sqrt{n}}$, which implies that adaptation is possible over this regime. As shown in Proposition 3, the confidence interval $\mathrm{CI}_\alpha^1(Z)$ defined in (3.3.8) is fully adaptive over the regime $\frac{\sqrt{n}}{\log p} \lesssim k \lesssim \frac{n}{\log p}$ in the sense of (3.3.4).

Illustrated in the leftmost and middle plots of Figure 3.2, it is impossible to construct an adaptive confidence interval for $\|\widehat{\beta}^L-\beta\|_2^2$ over regimes $k_1 \leq k_2 \lesssim \frac{\sqrt{n}}{\log p}$ and $k_1 \ll \frac{\sqrt{n}}{\log p} \lesssim k_2 \lesssim \frac{n}{\log p}$ since $\mathbf{R}_\alpha^*\left(\Theta_0(k_1), \Theta_0(k_2), \widehat{\beta}^L, \ell_2\right) \gg \mathbf{R}_\alpha^*\left(\Theta_0(k_1), \widehat{\beta}^L, \ell_2\right)$ if $k_1 \ll \frac{\sqrt{n}}{\log p}$ and $k_1 \ll k_2$. To sum up, adaptive confidence intervals for $\|\widehat{\beta}^L-\beta\|_2^2$ is only possible over the regime $\frac{\sqrt{n}}{\log p} \lesssim k \lesssim \frac{n}{\log p}$.

Comparison with confidence balls

We should note that the problem of constructing confidence intervals for $\|\widehat{\beta}-\beta\|_2^2$ is related to but different from that of constructing confidence sets for β itself. Confidence balls constructed in Nickl & van de Geer (2013) are of form $\left\{\beta : \ \|\beta - \widehat{\beta}\|_2^2 \leq u_n\left(Z\right)\right\}$, where $\widehat{\beta}$ can be the Lasso estimator and $u_n\left(Z\right)$ is a data dependent squared radius. See Nickl & van de Geer (2013) for further details. A naive application of this confidence ball leads to a one-sided confidence interval for the loss $\|\widehat{\beta} - \beta\|_2^2$,

$$\mathrm{CI}_{\alpha}^{\text{induced}}\left(Z\right) = \left\{\|\widehat{\beta} - \beta\|_2^2 : \ \|\widehat{\beta} - \beta\|_2^2 \leq u_n\left(Z\right)\right\}. \tag{3.3.12}$$

Due to the reason that confidence sets for β were sought for in Theorem 1 in Nickl & van de Geer (2013), confidence sets in the form $\left\{\beta : \ \|\beta - \widehat{\beta}\|_2^2 \leq u_n\left(Z\right)\right\}$ will suffice to achieve the optimal length. However, since our goal is to characterize $\|\widehat{\beta} - \beta\|_2^2$, we apply the unbiased risk estimation discussed in Theorem 1 of Nickl & van de Geer (2013) and construct the two-sided confidence interval in (3.3.8). Such a two-sided confidence interval is more informative than the one-sided confidence interval (3.3.12) since the one-sided confidence interval does not contain the information whether the loss is close to zero or not. Furthermore, as shown in Nickl & van de Geer (2013), the length of confidence interval $\mathrm{CI}_{\alpha}^{\text{induced}}\left(Z\right)$ over the parameter space $\Theta_0(k)$ is of order $\frac{1}{\sqrt{n}} + \frac{k \log p}{n}$. The two-sided confidence interval $\mathrm{CI}_{\alpha}^{1}\left(Z\right)$ constructed in (3.3.8) is of expected length $\frac{1}{\sqrt{n}}$, which is much shorter than $\frac{1}{\sqrt{n}} + \frac{k \log p}{n}$ in the regime $k \gg \frac{\sqrt{n}}{\log p}$. That is, the two-sided confidence interval (3.3.8) provides a more accurate interval estimator of the ℓ_2 loss. This is illustrated in Figure 3.3.

The lower bound technique developed in the literature of adaptive confidence sets Nickl & van de Geer (2013) can also be used to establish some of the lower bound

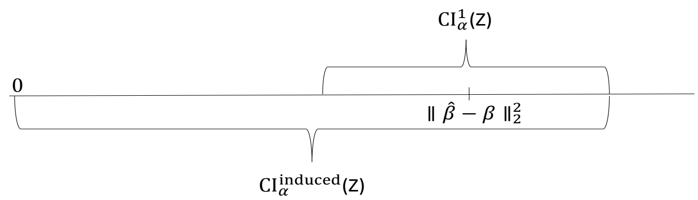

Figure 3.3: Comparison of the two-sided confidence interval $\mathrm{CI}_\alpha^1(Z)$ with the one-sided confidence interval $\mathrm{CI}_\alpha^{\text{induced}}(Z)$.

results for the case $q = 2$ given in the present paper. However, new techniques are needed in order to establish the rate sharp lower bounds for the minimax estimation error (3.2.9) in the region $\frac{\sqrt{n}}{\log p} \leq k \lesssim \frac{n}{\log p}$ and for the expected length of the confidence intervals (3.3.18) and (3.7.3) in the region $\frac{\sqrt{n}}{\log p} \lesssim k_1 \leq k_2 \lesssim \frac{n}{\log p}$, where it is necessary to test a composite null against a composite alternative in order to establish rate sharp lower bounds.

3.3.3 Confidence intervals for the ℓ_q loss with $1 \leq q < 2$ over $\Theta_0(k)$

We now consider the case $1 \leq q < 2$ and investigate the minimax expected length and adaptivity of confidence intervals for $\|\widehat{\beta} - \beta\|_q^2$ over the parameter space $\Theta_0(k)$. The following theorem characterizes the minimax convergence rate for the expected length of confidence intervals.

Theorem 13. *Suppose that $0 < \alpha < \frac{1}{4}$, $1 \leq q < 2$ and the sparsity levels k and k_0 satisfy Assumption* (B1)*. For any estimator $\widehat{\beta}$ satisfying Assumption* (A2) *with $\|\beta^*\|_0 \leq k_0$, then there is some constant $c > 0$ such that*

$$\mathbf{R}_\alpha^*\left(\Theta_0(k), \widehat{\beta}, \ell_q\right) \geq c k^{\frac{2}{q}} \frac{\log p}{n} \sigma_0^2. \tag{3.3.13}$$

In particular, if $\widehat{\beta}^L$ is the Lasso estimator defined in (3.2.10) *with $A > 4\sqrt{2}$, then the minimax expected length for $(1 - \alpha)$ level confidence intervals of $\|\widehat{\beta}^L - \beta\|_q^2$ over*

$\Theta_0(k)$ *is*

$$\mathbf{R}_{\alpha}^{*}\left(\Theta_0(k), \widehat{\beta}^{L}, \ell_q\right) \asymp k^{\frac{2}{q}} \frac{\log p}{n} \sigma_0^2. \tag{3.3.14}$$

We now construct the confidence interval achieving the minimax convergence rate in (3.3.14),

$$\mathrm{CI}_{\alpha}^{0}\left(Z, k, q\right) = \left(0, C^{*}(A, k) k^{\frac{2}{q}} \frac{\log p}{n}\right), \tag{3.3.15}$$

where $C^*(A,k) = \max\left\{\frac{(22A\sigma_0)^2}{\left(\frac{1}{4}-42\sqrt{\frac{2k\log p}{n_1}}\right)^4}, \frac{\left(\frac{3\eta_0}{\eta_0+1}A\sigma_0\right)^2}{\left(\frac{1}{4}-(9+11\eta_0)\sqrt{\frac{2k\log p}{n_1}}\right)^4}\right\}$ with $\eta_0 = 1.01\frac{\sqrt{A}+\sqrt{2}}{\sqrt{A}-\sqrt{2}}$. The following proposition establishes the coverage property and the expected length of $\mathrm{CI}_{\alpha}^{0}\left(Z, k, q\right)$.

Proposition 4. *Suppose* $k \lesssim \frac{n}{\log p}$ *and* $\widehat{\beta}^{L}$ *is the estimator defined in* (3.2.10) *with* $A > 4\sqrt{2}$. *For* $1 \leq q \leq 2$, *the confidence interval* $\mathrm{CI}_{\alpha}^{0}\left(Z, k, q\right)$ *defined in* (3.3.15) *satisfies*

$$\liminf_{n,p\to\infty} \inf_{\theta\in\Theta_0(k)} \mathbb{P}_{\theta}\left(\|\widehat{\beta}-\beta\|_q^2 \in \mathrm{CI}_{\alpha}^{0}\left(Z, k, q\right)\right) = 1, \tag{3.3.16}$$

and

$$\mathbf{R}\left(\mathrm{CI}_{\alpha}^{0}\left(Z, k, q\right), \Theta_0\left(k\right)\right) \lesssim k^{\frac{2}{q}} \frac{\log p}{n} \sigma_0^2. \tag{3.3.17}$$

In particular, for the case $q = 2$, (3.3.16) *and* (3.3.17) *also hold for the estimator* $\widehat{\beta}^{L}$ *defined in* (3.2.10) *with* $A > \sqrt{2}$.

This result shows that the confidence interval $\mathrm{CI}_{\alpha}^{0}\left(Z, k, q\right)$ achieves the minimax rate given in (3.3.14). In contrast to the ℓ_2 loss where the two-sided confidence interval (3.3.8) is significantly shorter than the one-sided interval and achieves the optimal rate over the regime $\frac{\sqrt{n}}{\log p} \lesssim k \lesssim \frac{n}{\log p}$, for the ℓ_q loss with $1 \leq q < 2$, the one-sided confidence interval achieves the optimal rate given in (3.3.14).

We now consider adaptivity of confidence intervals. The following theorem establishes the lower bounds for $\mathbf{R}_{\alpha}^{*}\left(\Theta_0(k_1), \Theta_0(k_2), \widehat{\beta}, \ell_q\right)$ with $1 \leq q < 2$.

Theorem 14. *Suppose $0 < \alpha < \frac{1}{4}$, $1 \leq q < 2$ and the sparsity levels k_1, k_2 and k_0 satisfy Assumption* (B2). *For any estimator $\widehat{\beta}$ satisfying Assumption* (A2) *with $\|\beta^*\|_0 \leq k_0$, then there is some constant $c > 0$ such that*

$$\mathbf{R}_{\alpha}^{*}\left(\Theta_0(k_1), \Theta_0(k_2), \widehat{\beta}, \ell_q\right) \geq \begin{cases} ck_2^{\frac{2}{q}} \frac{\log p}{n} \sigma_0^2 & \textit{if} \quad k_1 \leq k_2 \lesssim \frac{\sqrt{n}}{\log p}; \\ ck_2^{\frac{2}{q}-1} \frac{1}{\sqrt{n}} \sigma_0^2 & \textit{if} \quad k_1 \lesssim \frac{\sqrt{n}}{\log p} \lesssim k_2 \lesssim \frac{n}{\log p}; \\ ck_2^{\frac{2}{q}-1} k_1 \frac{\log p}{n} \sigma_0^2 & \textit{if} \quad \frac{\sqrt{n}}{\log p} \lesssim k_1 \leq k_2 \lesssim \frac{n}{\log p}. \end{cases} \tag{3.3.18}$$

In particular, if $p \geq n$ and $\widehat{\beta}^L$ is the Lasso estimator defined in (3.2.10) *with $A > 4\sqrt{2}$, the above lower bounds can be achieved.*

The lower bounds of Theorem 14 imply that of Theorem 13 and both lower bounds hold for a general class of estimators satisfying Assumption (A2). However, the lower bound (3.3.18) in Theorem 14 has a significantly different meaning from (3.3.13) in Theorem 13 where (3.3.18) quantifies the cost of adaptation without knowing the sparsity level. For the Lasso estimator $\widehat{\beta}^L$ defined in (3.2.10), by comparing Theorem 13 and Theorem 14, we obtain $\mathbf{R}_{\alpha}^{*}\left(\Theta_0(k_1), \Theta_0(k_2), \widehat{\beta}^L, \ell_q\right) \gg \mathbf{R}_{\alpha}^{*}\left(\Theta_0(k_1), \widehat{\beta}^L, \ell_q\right)$ if $k_1 \ll k_2$, which implies the impossibility of constructing adaptive confidence intervals for the case $1 \leq q < 2$. There exists marked difference between the case $1 \leq q < 2$ and the case $q = 2$, where it is possible to construct adaptive confidence intervals over the regime $\frac{\sqrt{n}}{\log p} \lesssim k \lesssim \frac{n}{\log p}$.

For the Lasso estimator $\widehat{\beta}^L$ defined in (3.2.10), it is shown in Proposition 4 that the confidence interval $\mathrm{CI}_{\alpha}^{0}\left(Z, k_2, q\right)$ defined in (3.3.15) achieves the lower bound $k_2^{\frac{2}{q}} \frac{\log p}{n} \sigma_0^2$ of (3.3.18). The lower bounds $k_2^{\frac{2}{q}-1} k_1 \frac{\log p}{n} \sigma_0^2$ and $k_2^{\frac{2}{q}-1} \frac{1}{\sqrt{n}} \sigma_0^2$ of (3.3.18) can be achieved by the following proposed confidence interval,

$$\mathrm{CI}_{\alpha}^{2}\left(Z,k_2,q\right)=\left(\left(\frac{\psi\left(Z\right)}{\frac{1}{n_2}\chi^2_{1-\frac{\alpha}{2}}\left(n_2\right)}-\sigma_0^2\right)_+,\left(16k_2\right)^{\frac{2}{q}-1}\left(\frac{\psi\left(Z\right)}{\frac{1}{n_2}\chi^2_{\frac{\alpha}{2}}\left(n_2\right)}-\sigma_0^2\right)_+\right), \tag{3.3.19}$$

where $\psi\left(Z\right)$ is given in (3.3.9). The above claim is verified in Proposition 5. Note that the confidence interval $\mathrm{CI}_{\alpha}^{1}\left(Z\right)$ defined in (3.3.8) is a special case of $\mathrm{CI}_{\alpha}^{2}\left(Z,k_2,q\right)$ with $q=2$.

Proposition 5. *Suppose $p\geq n$, $k_1\leq k_2\lesssim\frac{n}{\log p}$ and $\widehat{\beta}^L$ is defined in* (3.2.10) *with $A>4\sqrt{2}$. Then $\mathrm{CI}_{\alpha}^{2}\left(Z,k_2,q\right)$ defined in* (3.3.19) *satisfies,*

$$\liminf_{n,p\to\infty}\inf_{\theta\in\Theta_0(k_2)}\mathbb{P}_{\theta}\left(\|\widehat{\beta}-\beta\|_q^2\in\mathrm{CI}_{\alpha}^{2}\left(Z,k_2,q\right)\right)\geq 1-\alpha, \tag{3.3.20}$$

and

$$\mathbf{R}\left(\mathrm{CI}_{\alpha}^{2}\left(Z,k_2,q\right),\Theta_0\left(k_1\right)\right)\lesssim k_2^{\frac{2}{q}-1}\left(k_1\frac{\log p}{n}+\frac{1}{\sqrt{n}}\right)\sigma_0^2. \tag{3.3.21}$$

3.4 Minimaxity and adaptivity of confidence intervals over $\Theta(k)$

In this section, we focus on the case of unknown Σ and σ and establish the minimax expected length of confidence intervals for $\|\widehat{\beta}-\beta\|_q^2$ with $1\leq q\leq 2$ over $\Theta(k)$ defined in (3.2.4). We also study the possibility of adaptivity of confidence intervals for $\|\widehat{\beta}-\beta\|_q^2$. The following theorem establishes the lower bounds for the benchmark quantities $\mathbf{R}_{\alpha}^{*}\left(\Theta\left(k_i\right),\widehat{\beta},\ell_q\right)$ with $i=1,2$ and $\mathbf{R}_{\alpha}^{*}\left(\Theta\left(k_1\right),\Theta\left(k_2\right),\widehat{\beta},\ell_q\right)$.

Theorem 15. *Suppose that $0<\alpha<\frac{1}{4}$, $1\leq q\leq 2$ and the sparsity levels k_1,k_2 and k_0 satisfy Assumption* (B2). *For any estimator $\widehat{\beta}$ satisfying Assumption* (A1) *at*

$\theta_0 = (\beta^*, \mathrm{I}, \sigma_0)$ *with* $\|\beta^*\|_0 \leq k_0$*, there is a constant* $c > 0$ *such that*

$$\mathbf{R}^*_\alpha\left(\Theta\left(k_i\right), \widehat{\beta}, \ell_q\right) \geq c k_i^{\frac{2}{q}} \frac{\log p}{n}, \quad \textit{for} \quad i = 1, 2; \tag{3.4.1}$$

$$\mathbf{R}^*_\alpha\left(\{\theta_0\}, \Theta\left(k_2\right), \widehat{\beta}, \ell_q\right) \geq c k_2^{\frac{2}{q}} \frac{\log p}{n}. \tag{3.4.2}$$

In particular, if $\widehat{\beta}^{SL}$ *is the scaled Lasso estimator defined in* (3.2.14) *with* $A > 2\sqrt{2}$*, then the above lower bounds can be achieved.*

The lower bounds (3.4.1) and (3.4.2) hold for any $\widehat{\beta}$ satisfying Assumption (A1) at an interior point $\theta_0 = (\beta^*, \mathrm{I}, \sigma_0)$, including the scaled Lasso estimator as a special case. We demonstrate the impossibility of adaptivity of confidence intervals for the ℓ_q loss of the scaled Lasso estimator $\widehat{\beta}^{SL}$. Since $\mathbf{R}^*_\alpha\left(\Theta\left(k_1\right), \Theta\left(k_2\right), \widehat{\beta}^{SL}, \ell_q\right) \geq \mathbf{R}^*_\alpha\left(\{\theta_0\}, \Theta\left(k_2\right), \widehat{\beta}^{SL}, \ell_q\right)$, by (3.4.2), we have

$$\mathbf{R}^*_\alpha\left(\Theta\left(k_1\right), \Theta\left(k_2\right), \widehat{\beta}^{SL}, \ell_q\right) \gg \mathbf{R}^*_\alpha\left(\Theta\left(k_1\right), \widehat{\beta}^{SL}, \ell_q\right) \text{ if } k_1 \ll k_2.$$

The comparison of $\mathbf{R}^*_\alpha\left(\Theta\left(k_1\right), \widehat{\beta}^{SL}, \ell_q\right)$ and $\mathbf{R}^*_\alpha\left(\Theta\left(k_1\right), \Theta\left(k_2\right), \widehat{\beta}^{SL}, \ell_q\right)$ is illustrated in Figure 3.4. Referring to the adaptivity defined in (3.3.4), it is impossible to construct adaptive confidence intervals for $\|\widehat{\beta}^{SL} - \beta\|_q^2$.

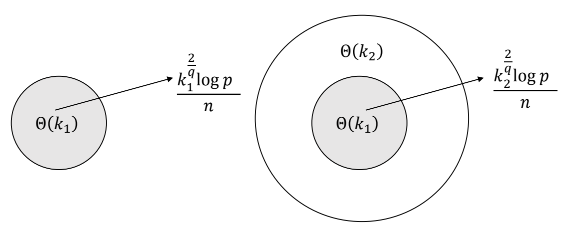

Figure 3.4: Illustration of $\mathbf{R}^*_\alpha\left(\Theta\left(k_1\right), \widehat{\beta}^{SL}, \ell_q\right)$ (left) and $\mathbf{R}^*_\alpha\left(\Theta\left(k_1\right), \Theta\left(k_2\right), \widehat{\beta}^{SL}, \ell_q\right)$ (right).

Theorem 15 shows that for any confidence interval $\mathrm{CI}_\alpha\left(\widehat{\beta}, \ell_q, Z\right)$ for the loss of any given estimator $\widehat{\beta}$ satisfying Assumption (A1), under the coverage constraint that

$\mathrm{CI}_\alpha\left(\widehat{\beta},\ell_q,Z\right)\in\mathcal{I}_\alpha\left(\Theta\left(k_2\right),\widehat{\beta},\ell_q\right)$, its expected length at any given $\theta_0=(\beta^*,\mathrm{I},\sigma)\in\Theta\left(k_0\right)$ must be of order $k_2^{\frac{2}{q}}\frac{\log p}{n}$. In contrast to Theorem 12 and 14, Theorem 15 demonstrates that confidence intervals must be long at a large subset of points in the parameter space, not just at a small number of "unlucky" points. Therefore, the lack of adaptivity for confidence intervals is not due to the conservativeness of the minimax framework.

In the following, we detail the construction of confidence intervals for $\|\widehat{\beta}^{SL}-\beta\|_q^2$. The construction of confidence intervals is based on the following definition of restricted eigenvalue, which is introduced in Bickel et al. (2009),

$$\kappa(X,k,s,\alpha_0)=\min_{\substack{J_0\subset\{1,\cdots,p\},\\|J_0|\le k}}\ \min_{\substack{\delta\neq 0,\\ \|\delta_{J_0^c}\|_1\le\alpha_0\|\delta_{J_0}\|_1}}\frac{\|X\delta\|_2}{\sqrt{n}\|\delta_{J_{01}}\|_2},\tag{3.4.3}$$

where J_1 denotes the subset corresponding to the s largest in absolute value coordinates of δ outside of J_0 and $J_{01}=J_0\cup J_1$. Define the event $\mathcal{B}=\{\hat{\sigma}\le\log p\}$. The confidence interval for $\|\widehat{\beta}^{SL}-\beta\|_q^2$ is defined as

$$\mathrm{CI}_\alpha\left(Z,k,q\right)=\begin{cases}[0,\ \varphi\left(Z,k,q\right)] & \text{on } \mathcal{B}\\ \{0\} & \text{on } \mathcal{B}^c,\end{cases}\tag{3.4.4}$$

where

$$\varphi\left(Z,k,q\right)=\min\left\{\left(\frac{16A\max\|X_{\cdot j}\|_2^2\widehat{\sigma}}{n\kappa^2\left(X,k,k,3\left(\frac{\max\|X_{\cdot j}\|_2}{\min\|X_{\cdot j}\|_2}\right)\right)}\right)^2 k^{\frac{2}{q}}\frac{\log p}{n},\ \left(k^{\frac{2}{q}}\frac{\log p}{n}\log p\right)\widehat{\sigma}^2\right\}.$$

Remark 6. The restricted eigenvalue $\kappa^2\left(X,k,k,3\left(\frac{\max\|X_{\cdot j}\|_2}{\min\|X_{\cdot j}\|_2}\right)\right)$ is computationally infeasible. For design covariance matrix Σ of special structures, the restricted eigenvalue can be replaced by its lower bound and a computationally feasible confidence interval can be constructed. See Section 4.4 in Cai & Guo (2016b) for more details.

Properties of $\mathrm{CI}_\alpha\left(Z,k,q\right)$ are established as follows.

Proposition 6. *Suppose* $k \lesssim \frac{n}{\log p}$ *and* $\widehat{\beta}^{SL}$ *is the estimator defined in* (3.2.14) *with* $A > 2\sqrt{2}$. *For* $1 \leq q \leq 2$, *then* $\mathrm{CI}_\alpha(Z, k, q)$ *defined in* (3.4.4) *satisfies the following properties,*

$$\liminf_{n,p\to\infty} \inf_{\theta\in\Theta(k)} \mathbb{P}_\theta\left(\|\widehat{\beta}-\beta\|_q^2 \in \mathrm{CI}_\alpha(Z,k,q)\right) = 1, \tag{3.4.5}$$

$$\mathbf{R}\left(\mathrm{CI}_\alpha(Z,k,q), \Theta(k)\right) \lesssim k^{\frac{2}{q}} \frac{\log p}{n}. \tag{3.4.6}$$

Proposition 6 shows that the confidence interval $\mathrm{CI}_\alpha(Z, k_i, q)$ defined in (3.4.4) achieves the lower bound in (3.4.1), for $i = 1, 2$, and the confidence interval $\mathrm{CI}_\alpha(Z, k_2, q)$ defined in (3.4.4) achieves the lower bound in (3.4.2).

3.5 Estimation of the ℓ_q loss of rate-optimal estimators

We have established minimax lower bounds for the estimation accuracy of the loss of a broad class of estimators $\widehat{\beta}$ satisfying (A1) or (A2) and also demonstrated that such minimax lower bounds are sharp for the Lasso and scaled Lasso estimators. We now show that the minimax lower bounds are sharp for the class of rate-optimal estimators satisfying the following Assumption (A).

(A) The estimator $\widehat{\beta}$ satisfies,

$$\sup_{\theta\in\Theta(k)} \mathbb{P}_\theta\left(\|\widehat{\beta}-\beta\|_q^2 \geq C^*\|\beta\|_0^{\frac{2}{q}} \frac{\log p}{n}\right) \leq Cp^{-\delta}, \tag{3.5.1}$$

for all $k \ll \frac{n}{\log p}$, where $\delta > 0$, $C^* > 0$ and $C > 0$ are constants not depending on k, n, or p.

We say an estimator $\widehat{\beta}$ is rate-optimal if it satisfies Assumption (A). As shown in Candès & Tao (2007); Bickel et al. (2009); Belloni et al. (2011); Sun & Zhang (2012), Lasso, Dantzig Selector, scaled Lasso and square-root Lasso are rate-optimal when the tuning parameter is chosen properly. We shall stress that Assumption (A) implies Assumptions (A1) and (A2). Assumption (A) requires the estimator $\widehat{\beta}$ to perform well over the whole parameter space $\Theta(k)$ while Assumptions (A1) and (A2) only require $\widehat{\beta}$ to perform well at a single point or over a proper subset. The following proposition shows that the minimax lower bounds established in Theorem 9 to Theorem 15 can be achieved for the class of rate-optimal estimators.

Proposition 7. *Let $\widehat{\beta}$ be an estimator satisfying Assumption* (A)*.*

1. *There exist (point or interval) estimators of the loss $\|\widehat{\beta} - \beta\|_q^2$ with $1 \leq q < 2$ achieving, up to a constant factor, the minimax lower bounds* (3.2.9) *in Theorem 9 and* (3.3.13) *in Theorem 13 and estimators of loss $\|\widehat{\beta} - \beta\|_q^2$ with $1 \leq q \leq 2$ achieving, up to a constant factor, the minimax lower bounds* (3.2.13) *in Theorem 10 and* (3.4.1) *and* (3.4.2) *in Theorem 15.*

2. *Suppose that the estimator $\widehat{\beta}$ is constructed based on the subsample $Z^{(1)} = \left(y^{(1)}, X^{(1)}\right)$, then there exist estimators of the loss $\|\widehat{\beta} - \beta\|_2^2$ achieving, up to a constant factor, the minimax lower bounds* (3.2.8) *in Theorem 9,* (3.3.5) *in Theorem 11 and* (3.3.7) *in Theorem 12.*

3. *Suppose the estimator $\widehat{\beta}$ is constructed based on the subsample $Z^{(1)} = \left(y^{(1)}, X^{(1)}\right)$ and it satisfies Assumption (A) with $\delta > 2$ and*

$$\sup_{\theta \in \Theta(k)} \mathbb{P}_\theta \left(\|(\widehat{\beta} - \beta)_{S^c}\|_1 \geq c^* \|(\widehat{\beta} - \beta)_S\|_1 \text{ where } S = \operatorname{supp}(\beta) \right) \leq C p^{-\delta}, \quad (3.5.2)$$

for all $k \ll \frac{n}{\log p}$. Then for $p \geq n$ there exist estimators of the loss $\|\widehat{\beta} - \beta\|_q^2$ with $1 \leq q < 2$ achieving the lower bounds given in (3.3.18) *in Theorem 14.*

For reasons of space, we do not discuss the detailed construction for the point and interval estimators achieving these minimax lower bounds here and postpone the construction to the proof of Proposition 7.

Remark 7. Sample splitting has been widely used in the literature. For example, the condition that $\widehat{\beta}$ is constructed based on the subsample $Z^{(1)} = \left(y^{(1)}, X^{(1)}\right)$ has been introduced in Nickl & van de Geer (2013) for constructing confidence sets for β and in Janson et al. (2015) for constructing confidence intervals for the ℓ_2 loss. Such a condition is imposed purely for technical reasons to create independence between the estimator $\widehat{\beta}$ and the subsample $Z^{(2)} = \left(y^{(2)}, X^{(2)}\right)$, which is useful to evaluate the ℓ_q loss of the estimator $\widehat{\beta}$. As shown in Bickel et al. (2009), the assumption (3.5.2) is satisfied for Lasso and Dantzig Selector. This technical assumption is imposed such that $\|\widehat{\beta} - \beta\|_1^2$ can be tightly controlled by $\|\widehat{\beta} - \beta\|_2^2$.

3.6 General tools for minimax lower bounds

A major step in our analysis is to establish rate sharp lower bounds for the estimation error and the expected length of confidence intervals for the ℓ_q loss. We introduce in this section new technical tools that are needed to establish these lower bounds.

A significant distinction of the lower bound results given in the previous sections from those for the traditional parameter estimation problems is that the constraint is on the performance of the estimator $\widehat{\beta}$ of the regression vector β, but the lower bounds are on the difficulty of estimating its loss $\|\widehat{\beta} - \beta\|_q^2$. It is necessary to develop new lower bound techniques to establish rate-optimal lower bounds for the estimation error and the expected length of confidence intervals for the loss $\|\widehat{\beta} - \beta\|_q^2$. These technical tools may also be of independent interest.

We begin with notation. Let Z denote a random variable whose distribution is indexed by some parameter $\theta \in \Theta$ and let π denote a prior on the parameter space Θ. We will use $f_\theta(z)$ to denote the density of Z given θ and $f_\pi(z)$ to denote the marginal density of Z under the prior π. Let $\mathbb{P}_\pi$ denote the distribution of Z corresponding to $f_\pi(z)$, i.e., $\mathbb{P}_\pi(\mathcal{A}) = \int 1_{z\in\mathcal{A}} f_\pi(z)\,dz$, where $1_{z\in\mathcal{A}}$ is the indicator function. For a function g, we write $\mathbb{E}_\pi(g(Z))$ for the expectation under f_π. More specifically, $f_\pi(z) = \int f_\theta(z)\pi(\theta)\,d\theta$ and $\mathbb{E}_\pi(g(Z)) = \int g(z) f_\pi(z)\,dz$. The L_1 distance between two probability distributions with densities f_0 and f_1 is given by $\mathrm{TV}(f_1, f_0) = \int |f_1(z) - f_0(z)|\,dz$. The following theorem establishes the minimax lower bounds for the estimation error and the expected length of confidence intervals for the ℓ_q loss, under the constraint that $\widehat{\beta}$ is a good estimator at at least one interior point.

Theorem 16. *Suppose $0 < \alpha, \alpha_0 < \frac{1}{4}$, $1 \leq q \leq 2$, Σ_0 is positive definite, $\theta_0 = (\beta^*, \Sigma_0, \sigma_0) \in \Theta$, and $\mathcal{F} \subset \Theta$. Define $d = \min_{\theta\in\mathcal{F}} \|\beta(\theta) - \beta^*\|_q$. Let π denote a prior over the parameter space $\mathcal{F}$. If an estimator $\widehat{\beta}$ satisfies*

$$\mathbb{P}_{\theta_0}\left(\|\widehat{\beta} - \beta^*\|_q^2 \leq \frac{1}{16}d^2\right) \geq 1 - \alpha_0, \tag{3.6.1}$$

then

$$\inf_{\widehat{L}_q} \sup_{\theta\in\{\theta_0\}\cup\mathcal{F}} \mathbb{P}_\theta\left(|\widehat{L}_q - \|\widehat{\beta} - \beta\|_q^2| \geq \frac{1}{4}d^2\right) \geq \bar{c}_1, \tag{3.6.2}$$

and

$$\mathbf{R}^*_\alpha\left(\{\theta_0\}, \Theta, \widehat{\beta}, \ell_q\right) = \inf_{\mathrm{CI}_\alpha\left(\widehat{\beta},\ell_q,Z\right)\in\mathcal{I}_\alpha\left(\Theta,\widehat{\beta},\ell_q\right)} \mathbb{E}_{\theta_0}\mathbf{R}\left(\mathrm{CI}_\alpha\left(\widehat{\beta}, \ell_q, Z\right)\right) \geq c_2^* d^2, \tag{3.6.3}$$

where $\bar{c}_1 = \min\left\{\frac{1}{10}, \left(\frac{9}{10} - \alpha_0 - \mathrm{TV}(f_\pi, f_{\theta_0})\right)_+\right\}$ and $c_2^* = \frac{1}{2}\left(1 - 2\alpha - \alpha_0 - 2\mathrm{TV}(f_\pi, f_{\theta_0})\right)_+$.

Remark 8. The minimax lower bound (3.6.2) for the estimation error and (3.6.3) for the expected length of confidence intervals hold as long as the estimator $\widehat{\beta}$ estimates β well at an interior point θ_0. Besides Condition (3.6.1), another key ingredient for the lower bounds (3.6.2) and (3.6.3) is to construct the least favorable space $\mathcal{F}$ with the prior π such that the marginal distributions f_π and f_{θ_0} are non-distinguishable. For the estimation lower bound (3.6.2), constraining that $\|\widehat{\beta}-\beta^*\|_q^2$ can be well estimated at θ_0, due to the non-distinguishability between f_π and f_{θ_0}, we can establish that the loss $\|\widehat{\beta}-\beta\|_q^2$ cannot be estimated well over $\mathcal{F}$. For the lower bound (3.6.3), by Condition (3.6.1) and the non-distinguishability between f_π and f_{θ_0}, we will show that $\|\widehat{\beta}-\beta\|_q^2$ over $\mathcal{F}$ is much larger than $\|\widehat{\beta}-\beta^*\|_q^2$ and hence the honest confidence intervals must be sufficiently long.

Theorem 16 is used to establish the minimax lower bounds for both the estimation error and the expected length of confidence intervals of the ℓ_q loss over $\Theta(k)$. By taking $\theta_0 \in \Theta(k_0)$ and $\Theta = \Theta(k)$, Theorem 10 follows from (3.6.2) with a properly constructed subset $\mathcal{F} \subset \Theta(k)$. By taking $\theta_0 \in \Theta(k_0)$ and $\Theta = \Theta(k_2)$, the lower bound (3.4.2) in Theorem 15 follows from (3.6.3) with a properly constructed $\mathcal{F} \subset \Theta(k_2)$. In both cases, Assumption (A1) implies Condition (3.6.1).

Several minimax lower bounds over $\Theta_0(k)$ can also be implied by Theorem 16. For the estimation error, the minimax lower bounds (3.2.8) and (3.2.9) over the regime $k \lesssim \frac{\sqrt{n}}{\log p}$ in Theorem 9 follow from (3.6.2). For the expected length of confidence intervals, the minimax lower bounds (3.3.7) in Theorem 12 and (3.3.18) in the regions $k_1 \leq k_2 \lesssim \frac{\sqrt{n}}{\log p}$ and $k_1 \lesssim \frac{\sqrt{n}}{\log p} \lesssim k_2 \lesssim \frac{n}{\log p}$ in Theorem 14 follow from (3.6.3). In these cases, Assumption (A1) or (A2) can guarantee that Condition (3.6.1) is satisfied. However, the minimax lower bounds for estimation error (3.2.9) in the region $\frac{\sqrt{n}}{\log p} \leq k \lesssim \frac{n}{\log p}$ and for the expected length of confidence intervals (3.3.18) in the region $\frac{\sqrt{n}}{\log p} \lesssim k_1 \leq k_2 \lesssim \frac{n}{\log p}$ cannot be established using the above theorem. The following theorem, which requires testing a composite null against a composite alternative, establishes the refined minimax lower bounds over $\Theta_0(k)$.

Theorem 17. *Let $0 < \alpha, \alpha_0 < \frac{1}{4}$, $1 \leq q \leq 2$, and $\theta_0 = (\beta^*, \Sigma_0, \sigma_0)$ where Σ_0 is a* *positive definite matrix. Let k_1 and k_2 be two sparsity levels. Assume that for $i = 1, 2$ there exist parameter spaces $\mathcal{F}_i \subset \{(\beta, \Sigma_0, \sigma_0) : \|\beta\|_0 \leq k_i\}$ such that for given* dist_i *and d_i*

$$\sqrt{(\beta(\theta) - \beta^*)^{\intercal}\Sigma_0(\beta(\theta) - \beta^*)} = \mathrm{dist}_i \quad \text{and} \quad \|\beta(\theta) - \beta^*\|_q = d_i, \quad \text{for all } \theta \in \mathcal{F}_i.$$

Let π_i denote a prior over the parameter space $\mathcal{F}_i$ for $i = 1, 2$. Suppose that for $\theta_1 = \left(\beta^*, \Sigma_0, \sigma_0^2 + \mathrm{dist}_1^2\right)$ and $\theta_2 = \left(\beta^*, \Sigma_0, \sigma_0^2 + \mathrm{dist}_2^2\right)$, there exist constants $c_1, c_2 > 0$ such that

$$\mathbb{P}_{\theta_i}\left(\|\widehat{\beta} - \beta^*\|_q^2 \leq c_i^2 d_i^2\right) \geq 1 - \alpha_0, \quad \textit{for} \quad i = 1, 2. \tag{3.6.4}$$

Then we have

$$\inf_{\widehat{L}_q} \sup_{\theta \in \mathcal{F}_1 \cup \mathcal{F}_2} \mathbb{P}_{\theta}\left(|\widehat{L}_q - \|\widehat{\beta} - \beta\|_q^2| \geq c_3^* d_2^2\right) \geq \bar{c}_3, \tag{3.6.5}$$

and

$$\mathbf{R}_{\alpha}^*\left(\Theta_0(k_1), \Theta_0(k_2), \widehat{\beta}, \ell_q\right) \geq c_4^*\left((1 - c_2)^2 d_2^2 - (1 + c_1)^2 d_1^2\right)_+, \tag{3.6.6}$$

where

$$c_3^* = \min\left\{\frac{1}{4}, \left((1 - c_2)^2 - \frac{1}{4} - (1 + c_1)^2 \frac{d_1^2}{d_2^2}\right)_+\right\},$$

$$c_4^* = \left(1 - 2\alpha_0 - 2\alpha - \sum_{i=1}^{2} \mathrm{TV}\left(f_{\pi_i}, f_{\theta_i}\right) - 2\mathrm{TV}\left(f_{\pi_2}, f_{\pi_1}\right)\right)_+,$$

$$\bar{c}_3 = \min\left\{\frac{1}{10}, \left(\frac{9}{10} - 2\alpha_0 - \sum_{i=1}^{2} \mathrm{TV}\left(f_{\pi_i}, f_{\theta_i}\right) - 2\mathrm{TV}\left(f_{\pi_2}, f_{\pi_1}\right)\right)_+\right\}.$$

Remark 9. As long as the estimator $\widehat{\beta}$ performs well at two points, θ_1 and θ_2, the minimax lower bounds (3.6.5) for the estimation error and (3.6.6) for the expected length of confidence intervals hold. Note that θ_i in the above theorem does not belong to the parameter space $\{(\beta, \Sigma_0, \sigma_0) : \|\beta\|_0 \leq k_i\}$, for $i = 1, 2$. In contrast to

Theorem 16, Theorem 17 compares composite hypotheses $\mathcal{F}_1$ and $\mathcal{F}_2$, which will lead to a sharper lower bound than comparing the simple null $\{\theta_0\}$ with the composite alternative $\mathcal{F}$. For simplicity, we construct least favorable parameter spaces $\mathcal{F}_i$ such that the points in $\mathcal{F}_i$ is of fixed (generalized) ℓ_2 distance and fixed ℓ_q distance to β^*, for $i = 1, 2$, respectively. More importantly, we construct $\mathcal{F}_1$ with the prior π_1 and $\mathcal{F}_2$ with the prior π_2 such that f_{π_1} and f_{π_2} are not distinguishable, where θ_1 and θ_2 are introduced to facilitate the comparison. By Condition (3.6.4) and the construction of $\mathcal{F}_1$ and $\mathcal{F}_2$, we establish that the ℓ_q loss cannot be simultaneously estimated well over $\mathcal{F}_1$ and $\mathcal{F}_2$. For the lower bound (3.6.6), under the same conditions, it is shown that the ℓ_q loss over $\mathcal{F}_1$ and $\mathcal{F}_2$ are far apart and any confidence interval with guaranteed coverage probability over $\mathcal{F}_1 \cup \mathcal{F}_2$ must be sufficiently long. Due to the prior information $\Sigma = \mathrm{I}$ and $\sigma = \sigma_0$, the lower bound construction over $\Theta_0(k)$ is more involved than that over $\Theta(k)$. We shall stress that the construction of $\mathcal{F}_1$ and $\mathcal{F}_2$ and the comparison between composite hypotheses are of independent interest.

The minimax lower bound (3.2.9) in the region $\frac{\sqrt{n}}{\log p} \lesssim k \lesssim \frac{n}{\log p}$ follows from (3.6.5) and the minimax lower bound (3.3.18) in the region $\frac{\sqrt{n}}{\log p} \lesssim k_1 \leq k_2 \lesssim \frac{n}{\log p}$ for the expected length of confidence intervals follows from (3.6.6). In these cases, Σ_0 is taken as I and Assumption (A2) implies Condition (3.6.4).

3.7 An intermediate setting with known $\sigma = \sigma_0$ and unknown Σ

The results given in Sections 3.3 and 3.4 show the significant difference between $\Theta_0(k)$ and $\Theta(k)$ in terms of minimaxity and adaptivity of confidence intervals for $\|\widehat{\beta} - \beta\|_q^2$. $\Theta_0(k)$ is for the simple setting with known design covariance matrix $\Sigma = \mathrm{I}$ and known noise level $\sigma = \sigma_0$, and $\Theta(k)$ is for unknown Σ and σ. In this section, we further consider minimaxity and adaptivity of confidence intervals for $\|\widehat{\beta}-\beta\|_q^2$ in an intermediate setting where the noise level $\sigma=\sigma_0$ is known and Σ is unknown but of certain structure. Specifically, we consider the following parameter space,

$$\Theta_{\sigma_0}(k,s)=\left\{(\beta,\Sigma,\sigma_0):\begin{array}{l}\|\beta\|_0\leq k,\quad \frac{1}{M_1}\leq\lambda_{\min}(\Sigma)\leq\lambda_{\max}(\Sigma)\leq M_1\\ \|\Sigma^{-1}\|_{L_1}\leq M,\quad \max_{1\leq i\leq p}\|\left(\Sigma^{-1}\right)_{i\cdot}\|_0\leq s\end{array}\right\}, \tag{3.7.1}$$

for some constants $M_1\geq 1$ and $M>0$. $\Theta_{\sigma_0}(k,s)$ basically assumes known noise level σ and imposes sparsity conditions on the precision matrix of the random design. This parameter space is similar to those used in the literature of sparse linear regression with random design van de Geer et al. (2014); Chernozhukov et al. (2015a,b). $\Theta_{\sigma_0}(k,s)$ has two sparsity parameters where k represents the sparsity of β and s represents the maximum row sparsity of the precision matrix Σ^{-1}. Note that $\Theta_0(k)\subset\Theta_{\sigma_0}(k,s)\subset\Theta(k)$ and $\Theta_0(k)$ is a special case of $\Theta_{\sigma_0}(k,s)$ with $M_1=1$.

Under the assumption $s\ll\sqrt{n/\log p}$, the minimaxity and adaptivity lower bounds for the expected length of confidence intervals for $\|\widehat{\beta}-\beta\|_q^2$ with $1\leq q<2$ over $\Theta_{\sigma_0}(k,s)$ are the same as those over $\Theta_0(k)$. That is, Theorems 13 and 14 hold with $\Theta_0(k_1)$, $\Theta_0(k_2)$, and $\Theta_0(k)$ replaced by $\Theta_{\sigma_0}(k_1,s)$, $\Theta_{\sigma_0}(k_2,s)$, and $\Theta_{\sigma_0}(k,s)$, respectively. For the case $q=2$, the following theorem establishes the minimaxity and adaptivity lower bounds for the expected length of confidence intervals for $\|\widehat{\beta}-\beta\|_2^2$ over $\Theta_{\sigma_0}(k,s)$.

Theorem 18. *Suppose* $0<\alpha,\alpha_0<1/4$, $M_1>1$, $s\ll\sqrt{n/\log p}$ *and the sparsity levels* k_1,k_2 *and* k_0 *satisfy Assumption* (B2) *with the constant* c_0 *replaced by* c_0^* *defined in* (3.9.14). *For any estimator* $\widehat{\beta}$ *satisfying*

$$\sup_{\theta\in\Theta(k_0)}\mathbb{P}_\theta\left(\|\widehat{\beta}-\beta^*\|_q^2\geq C^*\|\beta^*\|_0^{\frac{2}{q}}\frac{\log p}{n}\sigma^2\right)\leq\alpha_0, \tag{3.7.2}$$

with a constant $C^* > 0$, then there is some constant $c > 0$ such that

$$\mathbf{R}_\alpha^*\left(\Theta_{\sigma_0}(k_1, s), \Theta_{\sigma_0}(k_2, s), \widehat{\beta}, \ell_2\right) \geq c \min\left\{k_2 \frac{\log p}{n}, \max\left\{k_1 \frac{\log p}{n}, \frac{1}{\sqrt{n}}\right\}\right\} \sigma_0^2 \quad (3.7.3)$$

and

$$\mathbf{R}_\alpha^*\left(\Theta_{\sigma_0}(k_i, s), \widehat{\beta}, \ell_2\right) \geq c \frac{k_i \log p}{n} \sigma_0^2 \quad \textit{and} \quad i = 1, 2. \quad (3.7.4)$$

In particular, if $p \geq n$ and $\widehat{\beta}$ is constructed based on the subsample $Z^{(1)} = \left(y^{(1)}, X^{(1)}\right)$ and satisfies Assumption (A) with $\delta > 2$, the above lower bounds can be attained.

In contrast to Theorems 11 and 12, the lower bounds for the case $q = 2$ change in the absence of the prior knowledge $\Sigma = \mathrm{I}$ but the possibility of adaptivity of confidence intervals over $\Theta_{\sigma_0}(k, s)$ is similar to that over $\Theta_0(k)$. Since the Lasso estimator $\widehat{\beta}^L$ defined in (3.2.10) with $A > 4\sqrt{2}$ satisfies Assumption (A) with $\delta > 2$, by Theorem 18, the minimax lower bounds (3.7.3) and (3.7.4) can be attained for $\widehat{\beta}^L$. For $\widehat{\beta}^L$, only when $\frac{\sqrt{n}}{\log p} \lesssim k_1 \leq k_2 \lesssim \frac{n}{\log p}$, $\mathbf{R}_\alpha^*\left(\Theta_{\sigma_0}(k_1, s), \widehat{\beta}^L, \ell_2\right) \asymp \mathbf{R}_\alpha^*\left(\Theta_{\sigma_0}(k_1, s), \Theta_{\sigma_0}(k_2, s), \widehat{\beta}, \ell_2\right) \asymp \frac{k_1 \log p}{n}$ and adaptation between $\Theta_{\sigma_0}(k_1, s)$ and $\Theta_{\sigma_0}(k_2, s)$ is possible. In other regimes, if $k_1 \ll k_2$, then $\mathbf{R}_\alpha^*\left(\Theta_{\sigma_0}(k_1, s), \widehat{\beta}^L, \ell_2\right) \ll \mathbf{R}_\alpha^*\left(\Theta_{\sigma_0}(k_1, s), \Theta_{\sigma_0}(k_2, s), \widehat{\beta}, \ell_2\right)$ and adaptation between $\Theta_{\sigma_0}(k_1, s)$ and $\Theta_{\sigma_0}(k_2, s)$ is impossible. For reasons of space, more discussion on $\Theta_{\sigma_0}(k, s)$, including the construction of adaptive confidence intervals over the regime $\frac{\sqrt{n}}{\log p} \lesssim k_1 \leq k_2 \lesssim \frac{n}{\log p}$, is postponed to Chapter B.

3.8 Minimax lower bounds for estimating $\|\beta\|_q^2$ with $1 \leq q \leq 2$

The lower bounds developed in this paper have broader implications. In particular, the established results imply the minimax lower bounds for estimating $\|\beta\|_q^2$ and the expected length of confidence intervals for $\|\beta\|_q^2$ with $1 \leq q \leq 2$. To build the connection, it is sufficient to note that the trivial estimator $\widehat{\beta} = 0$ satisfies Assumptions (A1) and (A2) with $\beta^* = 0$. Then we can apply the lower bounds (3.2.8), (3.2.9) and (3.2.13) to the estimator $\widehat{\beta} = 0$ and establish the minimax lower bounds of estimating $\|\beta\|_q^2$,

$$\inf_{\widehat{L}_2} \sup_{\theta \in \Theta_0(k)} \mathbb{P}_\theta \left(|\widehat{L}_2 - \|\beta\|_2^2| \geq c \min\left\{ k\frac{\log p}{n}, \frac{1}{\sqrt{n}} \right\} \sigma_0^2 \right) \geq \delta; \tag{3.8.1}$$

$$\inf_{\widehat{L}_q} \sup_{\theta \in \Theta_0(k)} \mathbb{P}_\theta \left(|\widehat{L}_q - \|\beta\|_q^2| \geq c k^{\frac{2}{q}} \frac{\log p}{n} \sigma_0^2 \right) \geq \delta, \quad \text{for } 1 \leq q < 2, \tag{3.8.2}$$

$$\inf_{\widehat{L}_q} \sup_{\theta \in \Theta(k)} \mathbb{P}_\theta \left(|\widehat{L}_q - \|\beta\|_q^2| \geq c k^{\frac{2}{q}} \frac{\log p}{n} \right) \geq \delta, \quad \text{for } 1 \leq q \leq 2, \tag{3.8.3}$$

for some constants $\delta > 0$ and $c > 0$. Similarly, all the lower bounds for the expected length of confidence intervals for $\|\widehat{\beta} - \beta\|_q^2$ established in Theorem 11 to Theorem 15 imply corresponding lower bounds for $\|\beta\|_q^2$. The lower bound $\min\{k\frac{\log p}{n}, \frac{1}{\sqrt{n}}\}\sigma_0^2$ in (3.8.1) is the same as the detection boundary in the sparse linear regression for the case $\Sigma = \mathrm{I}$ and $\sigma = 1$; See Ingster et al. (2010) and Arias-Castro et al. (2011) for more details. Estimation of $\|\beta\|_2^2$ in high-dimensional linear regression has been considered in Guo et al. (2016) under the general setting where Σ and σ are unknown and the lower bound (3.8.3) with $q = 2$ leads to one key component of the lower bound $ck\frac{\log p}{n}$ for estimating $\|\beta\|_2^2$.

3.9 Proofs

This section presents the proofs of the lower bound results. We first establish the general lower bound result, Theorem 16, in Section 3.9.1. By applying Theorems 16 and 17, we prove Theorems 12 and 14 in Section 3.9.2. For reasons of space, the proofs of other main results, Theorems 9, 10, 11, 13, 15, 17, 18 as well as Propositions 2, 3, 4, 5, 6, 7 and the proofs of technical lemmas are postponed to Chapter B.

We define the χ^2 distance between two density functions f_1 and f_0 by $\chi^2(f_1, f_0) = \int \frac{(f_1(z)-f_0(z))^2}{f_0(z)} dz = \int \frac{f_1^2(z)}{f_0(z)} dz - 1$, and it is well known that

$$\mathrm{TV}(f_1, f_0) \leq \sqrt{\chi^2(f_1, f_0)}. \tag{3.9.1}$$

We follow the same notation used in Section 3.6. Let $\mathbb{P}_{Z,\theta\sim\pi}$ be the joint probability of Z and θ with the joint density function $f(\theta, z) = f_\theta(z)\pi(\theta)$. The following lemma, which is proved in Chapter B, is needed in the proofs of Theorem 16 and Theorem 17.

Lemma 7. *For any event $\mathcal{A}$, we have*

$$\mathbb{P}_\pi(Z \in \mathcal{A}) = \mathbb{P}_{Z,\theta\sim\pi}(Z \in \mathcal{A}), \tag{3.9.2}$$

$$|\mathbb{P}_{\pi_1}(Z \in \mathcal{A}) - \mathbb{P}_{\pi_2}(Z \in \mathcal{A})| \leq \mathrm{TV}(f_{\pi_2}, f_{\pi_1}). \tag{3.9.3}$$

We will write $\mathbb{P}_\pi(\mathcal{A})$ and $\mathbb{P}_{Z,\theta\sim\pi}(\mathcal{A})$ for $\mathbb{P}_\pi(Z \in \mathcal{A})$ and $\mathbb{P}_{Z,\theta\sim\pi}(Z \in \mathcal{A})$ respectively. Recall that $\widehat{L}_q(Z)$ denotes a data-dependent loss estimator and $\beta(\theta)$ denotes the corresponding β of the parameter θ.

3.9.1 Proof of Theorem 16

We set $c_0 = \frac{1}{4}$ and $\alpha_1 = \frac{1}{10}$.

Proof of (3.6.2)

We assume

$$\mathbb{P}_{\theta_0}\left(\left|\widehat{L}_q(Z) - \|\widehat{\beta}(Z) - \beta^*\|_q^2\right| \leq \frac{1}{4}d^2\right) \geq 1 - \alpha_1. \tag{3.9.4}$$

Otherwise, we have

$$\mathbb{P}_{\theta_0}\left(\left|\widehat{L}_q(Z) - \|\widehat{\beta}(Z) - \beta^*\|_q^2\right| \geq \frac{1}{4}d^2\right) \geq \alpha_1, \tag{3.9.5}$$

and hence (3.6.2) follows. Define the event

$$\mathcal{A}_0 = \left\{ z : \|\widehat{\beta}(z) - \beta^*\|_q^2 \leq c_0^2 d^2 \ , \ \left| \widehat{L}_q(z) - \|\widehat{\beta}(z) - \beta^*\|_q^2 \right| \leq \frac{1}{4} d^2 \right\}. \tag{3.9.6}$$

By (3.6.1) and (3.9.4), we have $\mathbb{P}_{\theta_0}(\mathcal{A}_0) \geq 1 - \alpha_0 - \alpha_1$. By (3.9.3), we obtain

$$\mathbb{P}_\pi(\mathcal{A}_0) \geq 1 - \alpha_0 - \alpha_1 - \int |f_{\theta_0}(z) - f_\pi(z)| \, dz. \tag{3.9.7}$$

For $z \in \mathcal{A}_0$ and $\theta \in \mathcal{F}$, by triangle inequality,

$$\|\widehat{\beta}(z) - \beta(\theta)\|_q \geq \left| \|\beta(\theta) - \beta^*\|_q - \|\widehat{\beta}(z) - \beta^*\|_q \right| \geq (1 - c_0)\, d. \tag{3.9.8}$$

For $z \in \mathcal{A}_0$ and $\theta \in \mathcal{F}$, then

$$\begin{aligned}
&\left| \widehat{L}_q(z) - \|\widehat{\beta}(z) - \beta(\theta)\|_q^2 \right| \\
&\geq \left| \|\widehat{\beta}(z) - \beta(\theta)\|_q^2 - \|\widehat{\beta}(z) - \beta^*\|_q^2 \right| - \left| \widehat{L}_q(z) - \|\widehat{\beta}(z) - \beta^*\|_q^2 \right| \\
&\geq (1 - 2c_0 - \frac{1}{4}) d^2,
\end{aligned}$$

where the first inequality follows from triangle inequality and the last inequality follows from (3.9.6) and (3.9.8). Hence, for $z \in \mathcal{A}_0$, we obtain

$$\inf_{\theta \in \mathcal{F}} \left| \widehat{L}_q(z) - \|\widehat{\beta}(z) - \beta(\theta)\|_q^2 \right| \geq (1 - 2c_0 - \frac{1}{4}) d^2. \tag{3.9.9}$$

Note that $\sup_{\theta \in \mathcal{F}} \mathbb{P}_\theta \left(\left| \widehat{L}_q(Z) - \|\widehat{\beta}(Z) - \beta(\theta)\|_q^2 \right| \geq (1 - 2c_0 - \frac{1}{4}) d^2 \right) \geq$ $\sup_{\theta \in \mathcal{F}} \mathbb{P}_\theta \left(\inf_{\theta \in \mathcal{F}} \left| \widehat{L}_q(Z) - \|\widehat{\beta}(Z) - \beta(\theta)\|_q^2 \right| \geq (1 - 2c_0 - \frac{1}{4}) d^2 \right)$. Since the max risk is lower bounded by the Bayesian risk, we can further lower bound the last term by $\mathbb{P}_\pi \left(\inf_{\theta \in \mathcal{F}} \left| \widehat{L}_q(Z) - \|\widehat{\beta}(Z) - \beta(\theta)\|_q^2 \right| \geq (1 - 2c_0 - \frac{1}{4}) d^2 \right)$. Combined with (3.9.9), we establish

$$\sup_{\theta\in\mathcal{F}}\mathbb{P}_\theta\left(\left|\widehat{L}_q\left(Z\right)-\|\widehat{\beta}(Z)-\beta\left(\theta\right)\|_q^2\right|\geq(1-2c_0-\frac{1}{4})d^2\right)\geq\mathbb{P}_\pi(\mathcal{A}_0). \tag{3.9.10}$$

Combining (3.9.5), (3.9.7) and (3.9.10), we establish (3.6.2).

Proof of (3.6.3)

For $\mathrm{CI}_\alpha\left(\widehat{\beta},\ell_q,Z\right)\in\mathcal{I}_\alpha\left(\Theta,\widehat{\beta},\ell_q\right)$, we have

$$\inf_{\theta\in\Theta}\mathbb{P}_\theta\left(\|\widehat{\beta}(Z)-\beta\left(\theta\right)\|_q^2\in\mathrm{CI}_\alpha\left(\widehat{\beta},\ell_q,Z\right)\right)\geq1-\alpha. \tag{3.9.11}$$

Define the event $\mathcal{A}=\left\{z:\|\widehat{\beta}(z)-\beta^*\|_q<c_0d,\ \ \|\widehat{\beta}(z)-\beta^*\|_q^2\in\mathrm{CI}_\alpha\left(\widehat{\beta},L,z\right)\right\}$. By (3.6.1) and (3.9.11), we have $\mathbb{P}_{\theta_0}\left(\mathcal{A}\right)\geq1-\alpha-\alpha_0$. (3.9.2) and (3.9.3) imply

$$\mathbb{P}_{Z,\theta\sim\pi}\left(\mathcal{A}\right)=\mathbb{P}_\pi\left(\mathcal{A}\right)\geq1-\alpha-\alpha_0-\mathrm{TV}\left(f_\pi,f_{\theta_0}\right). \tag{3.9.12}$$

Define the event $\mathcal{B}_\theta=\left\{z:\|\widehat{\beta}(z)-\beta\left(\theta\right)\|_q^2\in\mathrm{CI}_\alpha\left(\widehat{\beta},\ell_q,z\right)\right\}$ and $\mathcal{M}=\cup_{\theta\in\mathcal{F}}\mathcal{B}_\theta$. By (3.9.11), we have

$$\mathbb{P}_{Z,\theta\sim\pi}\left(\mathcal{M}\right)=\int\left(\int 1_{z\in\mathcal{M}}f_\theta(z)dz\right)\pi\left(\theta\right)d\theta\geq\int\left(\int 1_{z\in\mathcal{B}_\theta}f_\theta(z)dz\right)\pi\left(\theta\right)d\theta\geq1-\alpha.$$

Combined with (3.9.12), we have $\mathbb{P}_{Z,\theta\sim\pi}\left(\mathcal{A}\cap\mathcal{M}\right)\geq1-2\alpha-\alpha_0-\mathrm{TV}\left(f_\pi,f_{\theta_0}\right)$. For $z\in\mathcal{M}$, there exists $\bar{\theta}\in\mathcal{F}$ such that $\|\widehat{\beta}(z)-\beta(\bar{\theta})\|_q^2\in\mathrm{CI}_\alpha\left(\widehat{\beta},\ell_q,z\right)$; For $z\in\mathcal{A}$, we have $\|\widehat{\beta}(z)-\beta^*\|_q^2\in\mathrm{CI}_\alpha\left(\widehat{\beta},\ell_q,z\right)$ and $\|\widehat{\beta}(z)-\beta^*\|_q<c_0d$. Hence, for $z\in\mathcal{A}\cap\mathcal{M}$, we have $\|\widehat{\beta}(z)-\beta(\bar{\theta})\|_q^2,\|\widehat{\beta}(z)-\beta^*\|_q^2\in\mathrm{CI}_\alpha\left(\widehat{\beta},\ell_q,z\right)$ and $\|\widehat{\beta}(z)-\beta(\bar{\theta})\|_q\geq\|\beta(\bar{\theta})-\beta^*\|_q-\|\widehat{\beta}(z)-\beta^*\|_q\geq\left(1-c_0\right)d$ and hence

$$\mathbf{R}\left(\mathrm{CI}_\alpha\left(\widehat{\beta},\ell_q,z\right)\right)\geq\left(1-2c_0\right)d^2. \tag{3.9.13}$$

Define the event $\mathcal{C}=\left\{z:\mathbf{R}\left(\mathrm{CI}_{\alpha}\left(\widehat{\beta},\ell_q,z\right)\right)\geq\left(1-2c_0\right)d^2\right\}$. By (3.9.13), we have $\mathbb{P}_{\pi}\left(\mathcal{C}\right)=\mathbb{P}_{Z,\theta\sim\pi}\left(\mathcal{C}\right)\geq\mathbb{P}_{Z,\theta\sim\pi}\left(\mathcal{A}\cap\mathcal{M}\right)\geq 1-2\alpha-\alpha_0-\mathrm{TV}\left(f_{\pi},f_{\theta_0}\right)$. By (3.9.3), we establish $\mathbb{P}_{\theta_0}\left(\mathcal{C}\right)\geq 1-2\alpha-\alpha_0-2\mathrm{TV}\left(f_{\pi},f_{\theta_0}\right)$ and hence (3.6.3).

3.9.2 Proof of Theorems 12 and 14

We first specify some constants used in the proof. Let C^* be given in (3.2.6). Define $\epsilon_1=\frac{1-2\alpha-2\alpha_0}{12}$ and

$$c_0=\min\left\{\frac{1}{2},32\log\left(1+\epsilon_1^2\right),\frac{2}{3}\sqrt{\log(1+\epsilon_1^2)},\frac{1-2\gamma}{16C^*},\left(\frac{1-2\gamma}{16C^*}\right)^2\right\},\ c_0^*=\min\left\{c_0,\frac{\sqrt{M_1}-1}{C^*M_1+\sqrt{M_1}-1}\right\}. \tag{3.9.14}$$

Theorems 12 and 14 follow from Theorem 19 below.

Theorem 19. *Suppose $0<\alpha<\frac{1}{4}$, $1\leq q\leq 2$ and the sparsity levels k_1,k_2 and k_0 satisfy Assumption* (B2). *Suppose that $\widehat{\beta}$ satisfies Assumption* (A2) *with $\|\beta^*\|_0\leq k_0$.*

1. *If $k_2\lesssim\frac{\sqrt{n}}{\log p}$, then there is some constant $c>0$ such that*

$$\mathbf{R}_{\alpha}^*\left(\Theta_0\left(k_1\right),\Theta_0\left(k_2\right),\widehat{\beta},\ell_q\right)\geq ck_2^{\frac{2}{q}}\frac{\log p}{n}\sigma_0^2. \tag{3.9.15}$$

2. *If $\frac{\sqrt{n}}{\log p}\lesssim k_2\lesssim\frac{n}{\log p}$, then there is some constant $c>0$ such that*

$$\begin{aligned}&\mathbf{R}_{\alpha}^*\left(\Theta_0\left(k_1\right),\Theta_0\left(k_2\right),\widehat{\beta},\ell_q\right)\\&\geq c\max\left\{\left((1-c_2)^2k_2^{\frac{2}{q}-1}k_1\frac{\log p}{n}-(1+c_1)^2k_1^{\frac{2}{q}}\frac{\log p}{n}\right)_+,\frac{k_2^{\frac{2}{q}-1}}{\sqrt{n}}\right\}\sigma_0^2,\end{aligned} \tag{3.9.16}$$

where $c_1=\frac{C^*k_0^{\frac{1}{q}}}{(k_1-k_0)^{\frac{1}{q}}}$ and $c_2=\frac{C^*k_0^{\frac{1}{q}}}{(k_2-k_0)^{\frac{1}{q}-\frac{1}{2}}(k_1-k_0)^{\frac{1}{2}}}$.

In particular, the minimax lower bound (3.9.15) *and the term $\frac{k_2^{\frac{2}{q}-1}}{\sqrt{n}}\sigma_0^2$ in* (3.9.16) *can be established under the weaker assumption* (A1) *with $\|\beta^*\|_0\leq k_0$.*

By Theorem 19, we establish (3.3.7) in Theorem 12 and (3.3.18) in Theorem 14. In the regime $k_2 \lesssim \frac{\sqrt{n}}{\log p}$, the lower bound (3.3.7) for $q = 2$ and (3.3.18) for $1 \leq q < 2$ follow from (3.9.15). For the case $q = 2$, in the regime $\frac{\sqrt{n}}{\log p} \lesssim k_2 \lesssim \frac{n}{\log p}$, the first term of the right hand side of (3.9.16) is 0 while the second term is $\frac{1}{\sqrt{n}}\sigma_0^2$, which leads to (3.3.7). For $1 \leq q < 2$, let $k_1^* = \min\{k_1, \zeta_0 k_2\}$ for some constant $0 < \zeta_0 < 1$, an application of (3.9.16) leads to $\mathbf{R}_\alpha^*\left(\Theta_0\left(k_1^*\right), \Theta_0\left(k_2\right), \widehat{\beta}, \ell_q\right) \geq c \max\left\{k_2^{\frac{2}{q}-1} k_1^* \frac{\log p}{n}, \frac{k_2^{\frac{2}{q}-1}}{\sqrt{n}}\right\}\sigma_0^2$. By this result, if $k_1 \leq \zeta_0 k_2$, then $k_1^* = k_1$ and the lower bounds (3.3.18) in the regions $k_1 \lesssim \frac{\sqrt{n}}{\log p} \lesssim k_2 \lesssim \frac{n}{\log p}$ and $\frac{\sqrt{n}}{\log p} \lesssim k_1 \leq k_2 \lesssim \frac{n}{\log p}$ follow; if $\zeta_0 k_2 < k_1 \leq k_2$, then $k_1^* = \zeta_0 k_2 \geq \zeta_0 k_1$. By the fact that $\mathbf{R}_\alpha^*\left(\Theta_0\left(k_1\right), \Theta_0\left(k_2\right), \widehat{\beta}, \ell_q\right) \geq \mathbf{R}_\alpha^*\left(\Theta_0\left(k_1^*\right), \Theta_0\left(k_2\right), \widehat{\beta}, \ell_q\right)$, the lower bounds (3.3.18) over the regions $k_1 \lesssim \frac{\sqrt{n}}{\log p} \lesssim k_2 \lesssim \frac{n}{\log p}$ and $\frac{\sqrt{n}}{\log p} \lesssim k_1 \leq k_2 \lesssim \frac{n}{\log p}$ follow. The following lemma shows that (3.3.7) holds for $\widehat{\beta}^L$ defined in (3.2.10) with $A > \sqrt{2}$ by verifying Assumption (A1) and (3.3.18) holds for $\widehat{\beta}^L$ defined in (3.2.10) with $A > 4\sqrt{2}$ by verifying Assumption (A2). Its proof can be found in Chapter B.

Lemma 8. *If $A > 4\sqrt{2}$, then we have*

$$\sup_{\{\theta=(\beta^*, \mathrm{I}, \sigma):\sigma \leq 2\sigma_0\}} \mathbb{P}_\theta\left(\|\widehat{\beta}^L - \beta^*\|_q^2 \geq C\|\beta^*\|_0^{\frac{2}{q}} \frac{\log p}{n}\sigma^2\right) \leq c\exp\left(-c'n\right) + p^{-c}.$$

In particular, the above result holds for $q = 2$ under the assumption $A > \sqrt{2}$.

4

Confidence Interval for Causal Effects with Invalid Instruments using Two-Stage Hard Thresholding

4.1 Introduction

4.1.1 Motivation: invalid instruments even after controlling for (potentially many) confounders

Instrumental variables (IV) analysis is a popular method to deduce causal effects in the presence of unmeasured confounding. An IV analysis requires variables called instruments that (A1) are related to the exposure (A2) have no direct pathway to the outcome and (A3) are not related to unmeasured variables that affect the exposure and the outcome (see Section 4.2.2 for details). Variables that satisfy these assumptions are referred to as valid instruments. A major challenge in IV analysis is to find valid instruments.

In practice, it is often the case that potential candidate instruments become more plausible as valid instruments after controlling for some, possibly high dimensional, covariates (see Hernán & Robins (2006) and Baiocchi et al. (2014) for discussion on control for covariates for an instrument to be valid). For example, a long-standing interest in economics is the causal effect of education on earnings and often, IV analysis is used to deduce the effect (Angrist & Krueger, 1991; Card, 1993, 1999). A popular instrument in this analysis is a person's proximity to a college when growing up (Card, 1999, 1993). However, proximity to a college may be related to a person's socioeconomic status, characteristics of a person's high school and other covariates that may affect a person's earnings. Thus, these covariates need to be controlled for in order for proximity to college to be a valid IV and with the growing trend toward collecting large data sets with many variables, this approach of finding instrumental variables that are valid after conditioning on covariates has increasing promise (Hernán & Robins, 2006; Swanson & Hernán, 2013; Baiocchi et al., 2014; Varian, 2014; Imbens, 2014).

Yet, despite the promise that large data sets may bring in terms of finding valid instruments by conditioning on potentially many covariates, some IVs may still turn out to be invalid and subsequent analysis assuming that all the IVs are valid after conditioning can be misleading (Murray, 2006). For example, suppose for studying the causal effect of education on earnings, we used proximity as an IV and to make sure the IV satisfies (A3), we control for confounders like high school test scores of the student, high school size, individual's genetic makeup, family education, and family's socioeconomic status. But, if living close to college had other benefits beyond getting more education, say by being exposed to many programs available to high school students for job preparation and employers who come to the area to discuss employment opportunities for college students, then the IV, proximity to college, can directly affect individual's earning potential and violate (A2) (Card, 1999). This problem is also widely prevalent in other applications of instrumental variables, most notably in Mendelian randomization (Davey Smith & Ebrahim, 2003, 2004) where the instruments are genetic in nature and some instruments are likely to be invalid due to having pleiotropic effects (Lawlor et al., 2008; Burgess et al., 2015).

This paper tackles the problem of constructing confidence intervals for causal effects that are robust to invalid instruments even after controlling for possibly high dimensional covariates.

4.1.2 Prior work

In non-IV settings with many high dimensional covariates, Zhang & Zhang (2014); Javanmard & Montanari (2014a); van de Geer et al. (2014); Belloni et al. (2014) and Cai & Guo (2016b) provide honest confidence intervals for a treatment effect. In IV settings with high dimensional covariates (or IVs), Gautier & Tsybakov (2011); Belloni et al. (2012); Fan & Liao (2014) and Chernozhukov et al. (2015a) provide honest confidence intervals for a treatment effect, under the assumption that all the IVs are valid after controlling for said covariates. In invalid IV settings, Kolesár et al. (2015) and Bowden et al. (2015) provide inferential methods for treatment effects. However, the method requires that the effects of the IVs on the treatment be orthogonal to their direct effects on the outcome, a stringent assumption. Bowden et al. (2016); Burgess et al. (2016); Kang et al. (2016b) and Windmeijer et al. (2016) also work on the invalid IV setting, but without making the stringent orthogonality assumption. Unfortunately, all these papers focuses on the low dimensional setting and some only work in the case where the IVs are completely uncorrelated/orthogonal to each other unless modifications are made (Bowden et al., 2015; Burgess et al., 2016). Furthermore, all the previous work only provides a consistent estimator of the treatment effect without any theoretical guarantees on inference; in fact, one of the simplest consistent estimators in this setting, the median estimator (Bowden et al., 2016; Burgess et al., 2016; Windmeijer et al., 2016) has been shown to be consistent, but not $\sqrt{n}$ consistent (Windmeijer et al., 2016).

There are two major challenges in obtaining valid confidence intervals in our problem: (i) potentially high-dimensional covariates and (ii) the invalid IVs. The problem related with high-dimensional covariates can be dealt with by applying recent debiasing methods developed in Zhang & Zhang (2014); Javanmard & Montanari (2014a); van de Geer et al. (2014); Cai & Guo (2016b). However, the general idea behind debiasing does not inherently resolve the invalid IV problem as even a single IV that is improperly assumed as valid while it is truly invalid can make these debiased estimates useless. To put it another way, debiasing as a method is only meant to asymptotically remove the bias of regression coefficients from ℓ_1 shrinkage estimators and to conduct proper inference on these de-biased coefficients. This methodological goal is different than in the invalid IV problem where the goal is to properly estimate a set of valid IVs, as even a single error of declaring an IV that is invalid as valid can lead to dishonest inference. In fact, the methodological challenge is not only to correctly select IVs, but also once selected, to do robust inference using the selected IVs.

4.1.3 Our contributions

Although there are existing methods for estimating the treatment effect in the presence of possibly invalid IVs, there is a paucity of procedures for selecting the set of valid instruments and forming confidence intervals for the treatment effects with theoretical coverage guarantees. In this paper, we propose a novel two-stage hard thresholding (TSHT) procedure to estimate the set of valid instruments and form confidence intervals with theoretical coverage guarantee. As the name suggests, a key component of TSHT is the two sequential steps of hard-thresholding procedures common in high dimensional inference (Donoho & Johnstone, 1994; Donoho, 1995) to simultaneously allow for invalid IVs and endogeneity of the treatment. Specifically, in the first thresholding stage, we select non-redundant IVs (see Definition 2 for details) and in the second thresholding stage, we use the thresholded estimates from the first thresholding step as pilot estimates to guide the selection of the set of valid instruments; see Section 4.3.3 for details. Using our two-stage variant of thresholding properly accounts for the selection of IVs and leads to $1/\sqrt{n}$ rate confidence intervals with desired coverage in both low and high dimensional settings where invalid IVs are present and without knowing a priori which of these IVs are invalid. Also, for the low dimensional covariate setting, our procedure is the first to have theoretical guarantees that it performs as well asymptotically as the oracle procedure that knows which instruments are valid. For the high dimensional covariate setting, our procedure is the first available procedure for forming confidence intervals with desired coverage when there may be invalid IVs.

The outline of the paper is as follows. After describing the model setup in Section 4.2, we formulate our TSHT procedure in Section 4.3. In Section 4.4, we develop the theoretical properties of our procedure. In Section 4.5, we investigate the performance of our procedure in a large simulation study and find that our confidence interval performs very similarly with respect to the oracle even if some of the underlying theoretical assumptions made in Section 4.4 are violated (see Sections 4.4 and 4.5 for details). In Section 4.6, we present an empirical study where we revisit the question of the causal effect of years of schooling on income using data from the Wisconsin Longitudinal Study. We provide conclusions and discussion in Section 4.7.

4.2 Model

4.2.1 Notation

To define causal effects, the potential outcome approach (Neyman, 1923; Rubin, 1974) for instruments laid out in Holland (1988) is used. For each individual $i \in \{1, \ldots, n\}$, let $Y_i^{(d,\mathbf{z})} \in \mathbb{R}$ be the potential outcome if the individual were to have exposure $d \in \mathbb{R}$ and instruments $\mathbf{z} \in \mathbb{R}^{p_z}$. Let $D_i^{(\mathbf{z})} \in \mathbb{R}$ be the potential exposure if the individual had instruments $\mathbf{z} \in \mathbb{R}^{p_z}$. For each individual, only one possible realization of $Y_i^{(d,\mathbf{z})}$ and $D_i^{(\mathbf{z})}$ is observed, denoted as Y_i and D_i, respectively, based on his observed instrument values $\mathbf{Z}_{i.} \in \mathbb{R}^{p_z}$ and exposure D_i. We also denote pre-instrument covariates for each individual i as $\mathbf{X}_{i.} \in \mathbb{R}^{p_x}$. In total, n sets of outcome, exposure, and instruments, denoted as $(Y_i, D_i, \mathbf{Z}_{i.}, \mathbf{X}_{i.})$, are observed in an i.i.d. fashion.

We denote $\mathbf{Y} = (Y_1, \ldots, Y_n)$ to be an n-dimensional vector of observed outcomes, $\mathbf{D} = (D_1, \ldots, D_n)$ to be an n-dimensional vector of observed exposures/treatment, $\mathbf{Z}$ to be a n by p_z matrix of instruments where row i consists of $\mathbf{Z}_{i.}$, and $\mathbf{X}$ to be an n by p_x matrix of covariates where row i consists of $\mathbf{X}_{i.}$. Let $\mathbf{W}$ be an n by $p = p_z + p_x$ matrix where $\mathbf{W}$ is a result of concatenating the matrices $\mathbf{Z}$ and $\mathbf{X}$ and $\mathbf{\Sigma}^* = \mathbb{E}\left(\mathbf{W}_{i\cdot}\mathbf{W}_{i\cdot}^\intercal\right)$ is positive definite. For any vector $\mathbf{v} \in \mathbb{R}^p$, let $\mathbf{v}_j$ denote the jth element of $\mathbf{v}$. Let $\|\mathbf{v}\|_1$, $\|\mathbf{v}\|_2$, and $\|\mathbf{v}\|_\infty$ denote the usual $1, 2$ and ∞-norms, respectively. Let $\|\mathbf{v}\|_0$ denote the number of non-zero elements in $\mathbf{v}$ and $\text{supp}(\mathbf{v}) \subseteq \{1, \ldots, p\}$, is defined as $\{j : \mathbf{v}_j \neq 0\}$.

For any n by p matrix $\mathbf{M} \in \mathbb{R}^{n \times p}$, we denote the (i, j) element of matrix $\mathbf{M}$ as M_{ij}, the ith row as $\mathbf{M}_{i.}$, and the jth column as $\mathbf{M}_{.j}$. Let $\mathbf{M}^\intercal$ be the transpose of $\mathbf{M}$ and $\|\mathbf{M}\|_\infty$ represent the element-wise matrix sup norm of matrix $\mathbf{M}$. For a sequence of random variables X_n, we use $X_n \xrightarrow{p} X$ and $X_n \xrightarrow{d} X$ to represent that X_n converges to X in probability and in distribution, respectively. Finally, for any two sequences a_n and b_n, we will write $a_n \gg b_n$ if $\limsup \frac{b_n}{a_n} = 0$ and write $a_n \ll b_n$ if $b_n \gg a_n$. Also, for a set J, $|J|$ denotes its cardinality.

4.2.2 Model and instrumental variables assumptions

We consider the Additive LInear, Constant Effects (ALICE) model of Holland (1988) and extend it to allow for multiple valid and possibly invalid instruments as in Small (2007) and Kang et al. (2016b). For two possible values of the exposure d', d and instruments $\mathbf{z}', \mathbf{z}$, we assume the following potential outcomes model

$$Y_i^{(d',\mathbf{z}')} - Y_i^{(d,\mathbf{z})} = (\mathbf{z}' - \mathbf{z})^\intercal \boldsymbol{\kappa}^* + (d' - d)\beta^*, \quad \mathbb{E}(Y_i^{(0,\mathbf{0})} \mid \mathbf{Z}_{i.}, \mathbf{X}_{i.}) = \mathbf{Z}_{i.}^\intercal \boldsymbol{\eta}^* + \mathbf{X}_{i.}^\intercal \boldsymbol{\phi}^* \quad (4.2.1)$$

where $\boldsymbol{\kappa}^*, \beta^*, \boldsymbol{\eta}^*$, and $\boldsymbol{\phi}^*$ are unknown parameters. The parameter β^* represents the causal parameter of interest, the causal effect (divided by $d' - d$) of changing the exposure from d' to d on the outcome. The parameter $\boldsymbol{\phi}^*$ represents the impact of covariates on the baseline potential outcome $Y_i^{(0,\mathbf{0})}$. The parameter $\boldsymbol{\kappa}^*$ represents violation of (A2), the direct effect of the instruments on the outcome. If (A2) holds, then $\boldsymbol{\kappa}^* = 0$. The parameter $\boldsymbol{\eta}^*$ represents violation of (A3), the presence of unmeasured confounding between the instrument and the outcome. If (A3) holds, then $\boldsymbol{\eta}^* = 0$.

Let $\boldsymbol{\pi}^* = \boldsymbol{\kappa}^* + \boldsymbol{\eta}^*$ and $\epsilon_{i1} = Y_i^{(0,\mathbf{0})} - \mathbb{E}(Y_i^{(0,\mathbf{0})} \mid \mathbf{Z}_{i.}, \mathbf{X}_{i.})$. When we combine equation (4.2.1) along with the definition of ϵ_{i1}, the observed data model becomes

$$Y_i = \mathbf{Z}_{i.}^\intercal \boldsymbol{\pi}^* + D_i \beta^* + \mathbf{X}_{i.}^\intercal \boldsymbol{\phi}^* + \epsilon_{i1}, \quad \mathbb{E}(\epsilon_{i1} \mid \mathbf{Z}_{i.}, \mathbf{X}_{i.}) = 0 \qquad (4.2.2)$$

and we denote $\sigma^2 = \mathrm{Var}(\epsilon_{i1} | \mathbf{Z}_{i.}, \mathbf{X}_{i.})$. The observed model is also known as the under-identified single-equation linear model in econometrics (page 83 of Wooldridge (2010)). This model is not a usual regression model because D_i might be correlated with ϵ_{i1}. In particular, the parameter β^* measures the causal effect of changing D on Y rather than an association. Also, the parameter $\boldsymbol{\pi}^*$ in model (4.2.2) combines both the violation of (A2), represented by $\boldsymbol{\kappa}^*$, and the violation of (A3), represented by $\boldsymbol{\eta}^*$. If both (A2) and (A3) are satisfied for IVs, typically referred to as valid IVs

(Murray, 2006), then $\boldsymbol{\kappa}^* = \boldsymbol{\eta}^* = 0$ and $\boldsymbol{\pi}^* = 0$. Hence, $\boldsymbol{\pi}^*$ captures invalid IVs, i..e the violations of (A2) and (A3). We formalize this notion with the following definition.

Definition 1. *Suppose we have* p_{z} *candidate instruments along with the models* (4.2.1)–(4.2.2). *We say that instrument* $j = 1, \ldots, p_z$ *is valid, i.e. satisfies* (A2) *and* (A3)*, if* $\boldsymbol{\pi}_j^* = 0$.

We also assume a linear association/observational model between the endogenous variable D_i, the instruments $\mathbf{Z}_{i.}$, and the covariates $\mathbf{X}_{i.}$,

$$D_i = \mathbf{Z}_{i.}^{\intercal}\boldsymbol{\gamma}^* + \mathbf{X}_{i.}^{\intercal}\boldsymbol{\psi}^* + \epsilon_{i2}, \quad \mathbb{E}(\epsilon_{i2}|\mathbf{Z}_{i.}, \mathbf{X}_{i.}) = 0. \tag{4.2.3}$$

Each element $\boldsymbol{\gamma}_j^*$ is the partial correlation between the jth instrument and D. The parameter $\boldsymbol{\psi}^*$ represents the association between the covariates and D_i. Also, unlike the models (4.2.1)-(4.2.2), we do not need a causal model between D_i, $\mathbf{Z}_{i.}$, and $\mathbf{X}_{i.}$; only the association model (4.2.3) is sufficient for our method. Finally, for notation, we let $s_{\mathrm{z2}} = \|\boldsymbol{\pi}^*\|_0$, $s_{\mathrm{x2}} = \|\boldsymbol{\phi}^*\|_0$, $s_{\mathrm{z1}} = \|\boldsymbol{\gamma}^*\|_0$, $s_{\mathrm{x1}} = \|\boldsymbol{\psi}^*\|_0$ and $s = \max\{s_{\mathrm{z2}}, s_{\mathrm{x2}}, s_{\mathrm{z1}}, s_{\mathrm{x1}}\}$. Finally, in both models, the instruments and the covariates are exogenous to the error terms; see Wooldridge (2010) for textbook discussion on exogeneity.

Based on model (4.2.3), we can define a set of instruments that satisfy (A1), or sometimes referred to as non-redundant instruments in the econometrics literature (Cheng & Liao, 2015).

Definition 2. *Suppose we have* p_{z} *candidate instruments along with the model* (4.2.3)*. We say that instrument* $j = 1, \ldots, p_z$ *satisfies* (A1)*, or is a non-redundant IV, if* $\boldsymbol{\gamma}_j^* \neq 0$ *and denote* $\mathcal{S}^*$ *to be the set of these instruments.*

Typically, satisfying (A1) has been defined in a global sense where (A1) is satisfied if $\boldsymbol{\gamma}^* \neq 0$ (Wooldridge, 2010). However, this global definition can be misleading in the presence of multiple candidate instruments. For example, it is possible that $\gamma_1^* \gg 0$ while $\gamma_j^* = 0$ for all $j \neq 1$ so that only the first instrument has an effect on the exposure while the rest do not. Using the global definition would imply that all the p_z instruments satisfy (A1) while Definition 2 makes it explicit and, perhaps less ambiguous, that it is only the first instrument $j = 1$ that satisfies (A1). Nevertheless, both the traditional global definition and Definition 2 are equivalent if $\gamma_j^* \neq 0$ for all j, that is where we only include relevant instruments, which is typically the case in practice and is the scenario studied by Kang et al. (2016b). Finally, when there is only one candidate instrument so that $p_{\mathrm{z}} = 1$, both definitions are equivalent to the definition presented in Holland (1988) and both become a special case of the definition presented in Angrist et al. (1996) under an additive, linear, constant effects model. In short, Definition 2 agrees with most definitions of satisfying (A1) in the literature.

Combining Definitions 1 and 2, we can formally define the usual three core conditions, i.e. (A1)-(A3), that define instruments.

Definition 3. *Suppose we have p_{z} candidate instruments along with the models* (4.2.1)–(4.2.3). *We say that the $Z_{.j}$, $j = 1, \ldots, p_z$, is an instrument if* (A1) $-$ (A3) *are satisfied, i.e. if $\pi_j^* = 0$ and $\gamma_j^* \neq 0$. Let $\mathcal{V}^*$ be the set of instruments.*

When there is only one instrument, $p_z = 1$, Definition 3 of an instrument is identical to the definition of an instrument in Holland (1988). Specifically, Definition 2 satisfies assumption (A1) that the instrument is related to the exposure. Also, assumption (A2), the exclusion restriction, which means $Y_i^{(d,\mathbf{z})} = Y_i^{(d,\mathbf{z}')}$ for all $d, \mathbf{z}, \mathbf{z}'$, is equivalent to $\boldsymbol{\kappa}^* = 0$ and assumption (A3), no unmeasured confounding, which means $Y_i^{(d,\mathbf{z})}$ and $D_i^{(\mathbf{z})}$ are independent of $Z_{i.}$ for all d and $\mathbf{z}$, is equivalent to $\boldsymbol{\eta}^* = 0$, implying $\boldsymbol{\pi}^* = \boldsymbol{\kappa}^* + \boldsymbol{\eta}^* = \mathbf{0}$. Definition 3 is also a special case of the definition of an instrument in Angrist et al. (1996) where here we assume the model is additive, linear, and has a constant treatment effect β^*. Hence, when multiple instruments,

$p_z > 1$, are present, our models (4.2.1)–(4.2.3) and Definition 3 can be viewed as a generalization of the definition of instruments in Holland (1988).

Note that the models presented above are commonly used in applications of IVs in econometrics (Wooldridge, 2010) and applications of IVs in genetic epidemiology and Mendelian randomization (Didelez & Sheehan, 2007). However, we generalize these widely used models in two important ways: (i) the model in (4.2.2) allows for possibly invalid instruments and (ii) we allow the number of covariates p_x (and even the number of instruments p_z) to be larger than the sample size n.

4.3 Confidence interval estimation via Two-Stage Hard Thresholding

4.3.1 General approach

The construction of our confidence interval can be broken down into two parts. The first part, detailed in Section 4.3.2, is estimating ITT effects based on the models (4.2.2) and (4.2.3). As we will see, the first part primarily deals with the problem posed by potentially high dimensional covariates, specifically the bias that comes from penalized estimators for high dimensional regression. The second part, which is elaborated in Section 4.3.3, tackles the heart of the problem in this paper, the presence of invalid IVs even after conditioning on high dimensional controls. Here, we take a novel two-stage hard thresholding approach to correctly select the valid IVs. Specifically, in the first step, we estimate the set of IVs that satisfy (A1) and in the second step, we use these IVs as initial guides to find IVs that satisfy (A2) and (A3) using the estimated set in the first step. Combining the two parts gives our confidence interval estimation procedure and is summarized in Procedure 1.

Procedure 1 provides a general recipe to construct confidence intervals in the pres-

Procedure 1 Two-Stage Hard Thresholding (TSHT) for Confidence Interval for β^* under Invalid IVs with High-Dimensional Covariates

Input: Outcome $\mathbf{Y}$, treatment $\mathbf{D}$, instrument $\mathbf{Z}$, covariates $\mathbf{X}$, significance level α

STEP 1: Estimate ITT effects (i.e. $\widetilde{\mathbf{\Gamma}}$, $\widetilde{\boldsymbol{\gamma}}$) via debiased scaled Lasso in (4.3.3)-(4.3.6)

STEP 2: Select valid IVs (i.e. $\widetilde{V}$) via two-stage hard thresholding

STEP 2a: Estimate IVs satisfying (A1) (i.e. $\widetilde{\mathcal{S}}$) via hard thresholding $\widetilde{\boldsymbol{\gamma}}$ in (4.3.7)

STEP 2b: For each IV satisfying (A1) (i.e. $j \in \widetilde{\mathcal{S}}$), estimate IVs satisfying (A2) and (A3) via hard-thresholding $\widetilde{\boldsymbol{\pi}}^{[j]}$ in (4.3.8)-(4.3.9)

STEP 3: Combine STEP 1 and STEP 2 via (4.3.10)-(4.3.12) to obtain confidence interval

Output: $1 - \alpha$ Confidence interval for β^*

ence of invalid IVs and high dimensional covariates. The key step in the procedure is STEP 2, where we utilize two-stage hard thresholding, to deal with the problem posed by invalid IVs; as such, we call our procedure TSHT procedure, akin to the acronym for two-stage least squares (TSLS) procedure in IV, arguably the most popular IV estimator in the literature. We also note that the Procedure 1 as stated can handle (i) low dimensional covariates, (ii) high dimensional covariates, and (iii) settings with all IVs having no direct effect and no unmeasured confounding, which is unrealistic in practice, and can still obtain valid confidence intervals. However, depending on particular data sets one may have, the procedure can be modified for simplicity and, in some cases, efficiency; Sections 4.3.5 and 4.3.6 discusses these cases in detail.

4.3.2 Estimating lTT effects

The first part of the confidence interval procedure involves estimation of ITT effects. Specifically, given the observed models (4.2.2) and (4.2.3), we can write reduced-forms models where both models are only functions of $\mathbf{Z}_{i.}$ and $\mathbf{X}_{i.}$,

$$Y_i = \mathbf{Z}_{i.}^{\intercal}\mathbf{\Gamma}^* + \mathbf{X}_{i.}^{\intercal}\Psi^* + e_{i1}, \tag{4.3.1}$$

$$D_i = \mathbf{Z}_{i.}^{\intercal}\boldsymbol{\gamma}^* + \mathbf{X}_{i.}^{\intercal}\boldsymbol{\psi}^* + \epsilon_{i2}. \tag{4.3.2}$$

Here, $\mathbf{\Gamma}^* = \beta^* \boldsymbol{\gamma}^* + \boldsymbol{\pi}^*$ and $\Psi^* = \boldsymbol{\phi}^* + \beta^* \boldsymbol{\psi}^*$ are the parameters of the reduced-form model with $\mathbf{\Gamma}^*$ representing the ITT effect of the instruments on the outcome and $\boldsymbol{\gamma}^*$ representing the ITT effect of the instruments on the exposure. The term $e_{i1} = \beta^* \epsilon_{i2} + \epsilon_{i1}$ is the reduced-form error term in (4.3.1). The errors have the property that $\mathbb{E}(e_{i1} | \mathbf{Z}_{i.}, \mathbf{X}_{i.}) = 0$ and $\mathbb{E}(\epsilon_{i2} | \mathbf{Z}_{i.}, \mathbf{X}_{i.}) = 0$ with the variances $\Theta_{11}^* = \text{Var}(e_{i1} | \mathbf{Z}_{i.}, \mathbf{X}_{i.})$, $\Theta_{22}^* = \text{Var}(\epsilon_{i2} | \mathbf{Z}_{i.}, \mathbf{X}_{i.})$, and $\Theta_{12}^* = \text{Cov}(e_{i1}, \epsilon_{i2} | \mathbf{Z}_{i.}, \mathbf{X}_{i.})$. Thus, each equation in the reduced-form model is a usual (high dimensional) regression model with (high dimensional) covariates $\mathbf{Z}_{i.}$ and $\mathbf{X}_{i.}$ and outcomes Y_i and D_i, respectively.

There are many methods in the literature to estimate the parameters of high dimensional regression models like the reduced-form models in (4.3.1) and (4.3.2). One approach is the scaled Lasso estimator proposed by Sun & Zhang (2012),

$$\begin{aligned} &\{\widehat{\mathbf{\Gamma}}, \widehat{\mathbf{\Psi}}, \widehat{\Theta}_{11}\} \\ &= \underset{\mathbf{\Gamma} \in \mathbb{R}^{p_z}, \mathbf{\Psi} \in \mathbb{R}^{p_x}, \Theta_{11} \in \mathbb{R}^+}{\operatorname{argmin}} \frac{\|\mathbf{Y} - \mathbf{Z}\mathbf{\Gamma} - \mathbf{X}\mathbf{\Psi}\|_2^2}{2n\sqrt{\Theta_{11}}} + \frac{\sqrt{\Theta_{11}}}{2} + \frac{\lambda_0}{\sqrt{n}} \left(\sum_{j=1}^{p_z} \|\mathbf{Z}_{.j}\|_2 |\Gamma_j| + \sum_{j=1}^{p_x} \|\mathbf{X}_{.j}\|_2 |\Psi_j| \right) \end{aligned} \tag{4.3.3}$$

for the reduced model in (4.3.1) and

$$\begin{aligned} &\{\widehat{\gamma}, \widehat{\psi}, \widehat{\Theta}_{22}\} \\ &= \underset{\mathbf{\Gamma} \in \mathbb{R}^{p_z}, \mathbf{\Psi} \in \mathbb{R}^{p_x}, \Theta_{22} \in \mathbb{R}^+}{\operatorname{argmin}} \frac{\|\mathbf{D} - \mathbf{Z}\boldsymbol{\gamma} - \mathbf{X}\boldsymbol{\psi}\|_2^2}{2n\sqrt{\Theta_{22}}} + \frac{\sqrt{\Theta_{22}}}{2} + \frac{\lambda_0}{\sqrt{n}} \left(\sum_{j=1}^{p_z} \|\mathbf{Z}_{.j}\|_2 |\gamma_j| + \sum_{j=1}^{p_x} \|\mathbf{X}_{.j}\|_2 |\psi_j| \right) \end{aligned} \tag{4.3.4}$$

for the reduced model in (4.3.2). The term λ_0 in both estimation problems (4.3.3) and (4.3.4) represents the penalty term in the scaled Lasso estimator and we choose $\lambda_0 = \sqrt{a_0 \log p / n}$ for some constant $a_0 > 2$; in practice, we find that setting $a_0 = 2$ or 2.05 works well. Also, we can estimate Θ_{12}^* from the estimation problems (4.3.3) and (4.3.4) by $\widehat{\Theta}_{12} = 1/n \left(\mathbf{Y} - \mathbf{Z}\widehat{\mathbf{\Gamma}} - \mathbf{X}\widehat{\mathbf{\Psi}} \right)^\intercal \left(\mathbf{D} - \mathbf{Z}\widehat{\boldsymbol{\gamma}} - \mathbf{X}\widehat{\boldsymbol{\psi}} \right)$.

Unfortunately, most penalized estimators for high dimensional regression problems are biased and the scaled Lasso estimators are no exception. In our case, using the estimates, say $\widehat{\mathbf{\Gamma}}$ and $\widehat{\boldsymbol{\gamma}}$, are biased for the parameters that they estimate $\mathbf{\Gamma}^*$ and

$\boldsymbol{\gamma}^*$. Thankfully, recent works by Zhang & Zhang (2014); Javanmard & Montanari (2014a); van de Geer et al. (2014) and Cai & Guo (2016b) allow us to debias these biased estimates. Specifically, let $\mathbf{W}$ be the concatenated matrix of the instruments $\mathbf{Z}$ and the covariates $\mathbf{X}$. Suppose we solve p_z optimization problems where the solution to each p_z optimization problem, denoted as $\widehat{\mathbf{u}}^{[j]} \in \mathbb{R}^p$, $j = 1, \ldots, p_z$, is

$$\widehat{\mathbf{u}}^{[j]} = \operatorname*{argmin}_{\mathbf{u} \in \mathbb{R}^p} \frac{1}{n} \|\mathbf{W}\mathbf{u}\|_2^2 \quad \text{s.t.} \quad \|\frac{1}{n}\mathbf{W}^\intercal \mathbf{W}\mathbf{u} - \mathbf{I}_{.j}\|_\infty \leq \lambda_n, \tag{4.3.5}$$

with $\mathbf{I}_{.j}$ denoting the j-th column of the identity matrix $\mathbf{I}$. The tuning parameter λ_n is chosen to be $12M_1^2\sqrt{\log p/n}$ with M_1 defined as the largest eigenvalue of $\boldsymbol{\Sigma}^*$. Let $\widehat{\mathbf{U}}$ denote the concatenation of the p_z solutions to the optimization problem, i.e. $\widehat{\mathbf{U}} = (\widehat{\mathbf{u}}^{[1]}, \ldots, \widehat{\mathbf{u}}^{[p_z]})^\intercal$. Then, the debiased estimates of $\widehat{\boldsymbol{\Gamma}}$ and $\widehat{\boldsymbol{\gamma}}$, denoted as $\widetilde{\boldsymbol{\Gamma}}$ and $\widetilde{\boldsymbol{\gamma}}$, are

$$\widetilde{\boldsymbol{\Gamma}} = \widehat{\boldsymbol{\Gamma}} + \frac{1}{n}\widehat{\mathbf{U}}\mathbf{W}^\intercal \left(\mathbf{Y} - \mathbf{Z}\widehat{\boldsymbol{\Gamma}} - \mathbf{X}\widehat{\boldsymbol{\Psi}}\right), \quad \widetilde{\boldsymbol{\gamma}} = \widehat{\boldsymbol{\gamma}} + \frac{1}{n}\widehat{\mathbf{U}}\mathbf{W}^\intercal \left(\mathbf{D} - \mathbf{Z}\widehat{\boldsymbol{\gamma}} - \mathbf{X}\widehat{\boldsymbol{\psi}}\right). \tag{4.3.6}$$

In short, we used scaled Lasso along with de-biasing methods on the reduced-form models to obtain de-biased estimates $\widetilde{\boldsymbol{\Gamma}}$ and $\widetilde{\boldsymbol{\gamma}}$ of the intent-to-treat effects of the instruments on the outcome and the exposure, respectively.

4.3.3 Two-Stage Hard Thresholding

The second part of the confidence interval procedure deals with the problem posed by invalid IVs. Specifically, we need to select valid IVs among p_z candidate IVs that satisfy all (A1)-(A3) assumptions, that is the set $\mathcal{V}^*$ in Definition 3. As discussed before, we do this by first, finding IVs that satisfy (A1), that is the set $\mathcal{S}^*$ in Definition

2 consisting of js where $\gamma_j^* \neq 0$, by thresholding the de-biased estimate $\widetilde{\boldsymbol{\gamma}}$

$$\widetilde{\mathcal{S}} = \left\{ j : |\widetilde{\gamma}_j| \geq \frac{\sqrt{\widehat{\Theta}_{22}} \|\mathbf{W}\widehat{\mathbf{u}}^{[j]}\|_2}{\sqrt{n}} \sqrt{\frac{a_0 \log p_{\mathrm{z}}}{n}} \right\}, \tag{4.3.7}$$

where $\widetilde{\mathcal{S}}$ denotes an estimate of $\mathcal{S}^*$. The threshold is based on the noise level of $\widetilde{\gamma}_j$ in (4.3.6) (represented by $\sqrt{\widehat{\Theta}_{22}} \|\mathbf{W}\widehat{\mathbf{u}}^{[j]}\|_2/n$), adjusted by dimensionality of the instrument size (represented by $\sqrt{a_0 \log p_{\mathrm{z}}}$).

The second thresholding step involves selecting IVs that satisfy (A2) and (A3). Specifically, by Definition 1, the set of instruments that satisfy (A2) and (A3) are those js where $\pi_j^* = 0$. Consequently, to estimate $\boldsymbol{\pi}^*$, we take each instrument j in $\widetilde{\mathcal{S}}$ that satisfy (A1) and we define $\widehat{\beta}^{[j]}$ to be a "pilot" estimate of β^* by using this IV and dividing the reduced-form estimates, i.e. $\widehat{\beta}^{[j]} = \widetilde{\Gamma}_j / \widetilde{\gamma}_j$, and $\widehat{\boldsymbol{\pi}}^{[j]}$ to be the estimate of $\boldsymbol{\pi}^*$ using this jth instrument's estimate of β^*, i.e. $\widehat{\boldsymbol{\pi}}^{[j]} = \widetilde{\boldsymbol{\Gamma}} - \widehat{\beta}^{[j]} \widetilde{\boldsymbol{\gamma}}$; we also construct corresponding pilot estimates of σ^2, i.e. $\widehat{\sigma^2}^{[j]} = \widehat{\Theta}_{11} + (\widehat{\beta}^{[j]})^2 \widehat{\Theta}_{22} - 2\widehat{\beta}^{[j]} \widehat{\Theta}_{12}$. Then, for each $\widehat{\boldsymbol{\pi}}^{[j]}$ in $j \in \widetilde{\mathcal{S}}$, we threshold each element of $\widehat{\boldsymbol{\pi}}^{[j]}$ to create the thresholded estimate $\widetilde{\boldsymbol{\pi}}^{[j]}$,

$$\widetilde{\pi}_k^{[j]} = \widehat{\pi}_k^{[j]} \mathbf{1} \left(k \in \widetilde{\mathcal{S}} \, \cap \, |\widehat{\pi}_k^{[j]}| \geq a_0 \sqrt{\widehat{\sigma^2}^{[j]}} \frac{\|\mathbf{W}(\widehat{\mathbf{u}}^{[k]} - \frac{\widetilde{\gamma}_k}{\widetilde{\gamma}_j} \widehat{\mathbf{u}}^{[j]})\|_2}{\sqrt{n}} \sqrt{\frac{\log p_{\mathrm{z}}}{n}} \right) \tag{4.3.8}$$

for all $1 \leq k \leq p_{\mathrm{z}}$. Each thresholded estimate $\widetilde{\boldsymbol{\pi}}^{[j]}$ is obtained by looking at the elements of the un-thresholded estimate, $\widehat{\boldsymbol{\pi}}^{[j]}$, and examining whether each element of it exceeds the noise threshold, denoted by the term $\sqrt{\widehat{\sigma^2}^{[j]}} \|\mathbf{W}(\widehat{\mathbf{u}}^{[k]} - \frac{\widetilde{\gamma}_k}{\widetilde{\gamma}_j} \widehat{\mathbf{u}}^{[j]})\|_2/n$, adjusting for the multiplicity of the selection procedure by the term $a_0 \sqrt{\log p_{\mathrm{z}}}$. Among the $|\widetilde{\mathcal{S}}|$ candidate estimates of $\boldsymbol{\pi}^*$ based on each instrument in $\widetilde{\mathcal{S}}$, i.e. $\widetilde{\boldsymbol{\pi}}^{[j]}$, and we choose $\widetilde{\boldsymbol{\pi}}^{[j]}$ with most valid instruments, or equivalently choose $j^* \in \widetilde{\mathcal{S}}$ where $j^* = \text{argmin} \, \|\widetilde{\boldsymbol{\pi}}^{[j]}\|_0$; if there is a non-unique solution, we choose $\widetilde{\boldsymbol{\pi}}^{[j]}$ with the smallest

ℓ_1 norm, the closest convex norm of ℓ_0.

Intuitively, the second-stage thresholding selects the invalid IVs and valid IVs as follows. Among the $|\widetilde{\mathcal{S}}|$ pilot estimates $\widetilde{\boldsymbol{\pi}}^{[j]}$, the best estimate of π^* is the one that uses a valid IV from the set $\widetilde{\mathcal{S}}$. In particular, if the jth pilot estimate is actually based on a valid IV, then all the invalid IVs will be included in the support of the thresholded estimate $\widetilde{\boldsymbol{\pi}}^{[j]}$ because their π^* will be away from zero and all the valid instruments will be excluded from the support because their π^* are zero. On the other hand, if the jth pilot estimate is based on an invalid IV, the pilot estimate $\widetilde{\boldsymbol{\pi}}^{[j]}$ will be biased in the sense that the valid IVs will no longer have $\widetilde{\boldsymbol{\pi}}^{[j]}$ that will be thresholded to zero and most of the elements of $\widetilde{\boldsymbol{\pi}}^{[j]}$ will be away from zero. Consequently, many IVs will be declared invalid based on $\widetilde{\boldsymbol{\pi}}^{[j]}$ and when we minimize with respect to the number of non-zero elements of the vector, i.e. $\min \|\widetilde{\boldsymbol{\pi}}^{[j]}\|_0$ among all pilot estimates, we should be able to select the best estimate of π^*. We remark that the latter ℓ_0 minimization is reminiscent of Theorem 1 in Kang et al. (2016b) where a necessary and sufficient condition for identification of β^* under invalid instruments is by looking at the largest subset of valid instruments that converge on a unique (i.e. identified) value; the search for the largest subset of valid IVs is essentially a minimization of ℓ_0 norm, which counts the number of invalid IVs, and hence, there is some sense that our procedure is both sufficient and necessary way to estimate β^*.

Finally, we note that it is crucial to construct pilot estimates of π^* from the IVs in the first thresholding step, that is $\widetilde{\mathcal{S}}$, as each of these IVs represent strong IVs and have non-zero effects on the exposure; using IVs that are not in $\widetilde{\mathcal{S}}$ may lead to poor estimates of the direct effect on the outcome since a redundant instrument j, whose true γ_j^* is zero, can lead to a large, unstable value of $|\widetilde{\gamma}_k/\widetilde{\gamma}_j|$ and the threshold value in (4.3.8), which will make it difficult to distinguish truly invalid IVs from noise.

4.3.4 Confidence Interval Estimation

After the two thresholding steps, we estimate the set of valid instruments $\widetilde{\mathcal{V}} \subseteq \{1, \ldots, p_z\}$ as those elements of $\widetilde{\boldsymbol{\pi}}^{[j^*]}$ that are zero,

$$\widetilde{\mathcal{V}} = \widetilde{\mathcal{S}} \setminus \operatorname{supp}\left(\widetilde{\boldsymbol{\pi}}^{[j^*]}\right) \tag{4.3.9}$$

Then, using the estimated $\widetilde{\mathcal{V}}$, we obtain our estimate of β^*

$$\widehat{\beta} = \frac{\sum_{j \in \widetilde{\mathcal{V}}} \widetilde{\gamma}_j \widetilde{\Gamma}_j}{\sum_{j \in \widetilde{\mathcal{V}}} \widetilde{\gamma}_j^2}, \tag{4.3.10}$$

along with an estimate of its standard error

$$\widehat{\mathrm{V}} = \frac{\left\| \sum_{j \in \widetilde{\mathcal{V}}} \widetilde{\gamma}_j \frac{1}{\sqrt{n}} \mathbf{W} \widehat{\mathbf{u}}^{[j]} \right\|_2^2}{\left(\sum_{j \in \widetilde{\mathcal{V}}} \widetilde{\gamma}_j^2 \right)^2} \widehat{\sigma^2} \quad \text{and} \quad \widehat{\sigma^2} = \widehat{\Theta}_{11} + \widehat{\beta}^2 \widehat{\Theta}_{22} - 2\widehat{\beta}\widehat{\Theta}_{12}, \tag{4.3.11}$$

and the usual form for the confidence interval for β^*,

$$\left(\widehat{\beta} - z_{1-\alpha/2} \sqrt{\widehat{\mathrm{V}}/n}, \quad \widehat{\beta} + z_{1-\alpha/2} \sqrt{\widehat{\mathrm{V}}/n} \right), \tag{4.3.12}$$

where $z_{1-\alpha/2}$ is the $1 - \alpha/2$ quantile of the standard normal distribution.

In the equations for $\widehat{\beta}$ and the standard error $\widehat{\mathrm{V}}$, we see some familiar expressions from the traditional IV literature. First, $\widehat{\beta}$ has the "correct" form in that if, by chance, we correctly estimated the set of valid instruments $\mathcal{V}^*$ and our debiased estimates of the reduced-form parameters, $\widetilde{\boldsymbol{\Gamma}}$ and $\widetilde{\boldsymbol{\gamma}}$, are perfect estimates of the reduced-form parameters, our estimate of β^* in (4.3.10) would become $\widehat{\beta} = \sum_{j \in \widetilde{\mathcal{V}}} \widetilde{\gamma}_j \widetilde{\Gamma}_j / \sum_{j \in \widetilde{\mathcal{V}}} \widetilde{\gamma}_j^2 = \sum_{j \in \mathcal{V}^*} \gamma_j^{*2} \beta^* / \sum_{j \in \mathcal{V}^*} \gamma_j^{*2} = \beta^*$. Hence, our estimator in (4.3.10) would identify β^*. Clearly, we would never have a perfect estimate of the set $\mathcal{V}^*$ or of the reduced-form parameters in finite sample and Section 4.4 describes the properties of our estimate

$\widehat{\beta}$ under these uncertainties. Second, in the standard error formula in (4.3.11), the $\widehat{\Theta}_{11} + \widehat{\beta}^2\widehat{\Theta}_{22} - 2\widehat{\beta}\widehat{\Theta}_{12}$ is of the similar form to the usual IV estimator of σ^2, variance of the error term in our original model (4.2.2). But, our standard error estimator is scaled by terms that depend on the estimated set of valid instruments $\widetilde{\mathcal{V}}$.

4.3.5 Special case of procedure 1: valid IVs after controlling for high dimensional covariates

To better understand the components of our inference procedure 1, it is instructive to go through some specific cases of estimating β^* that is common in the literature as these special cases can greatly simplify the procedure and remove unnecessary components. The first case is when the instruments are assumed to be valid (i.e. no direct effect and no unmeasured confounding) after conditioning on high dimensional covariates. This setup was considered in Gautier & Tsybakov (2011); Belloni et al. (2012); Fan & Liao (2014); Chernozhukov et al. (2015a). Under this case, our procedure doesn't have to go through STEP 2b, the estimation of π^*, illustrated in Section 4.3.3. Instead, we can simply replace STEP 2b with $\widetilde{\mathcal{V}} = \widetilde{\mathcal{S}}$ and the resulting estimator for β^* is

$$\widehat{\beta}_H = \frac{\sum_{j \in \widetilde{\mathcal{S}}} \widetilde{\boldsymbol{\gamma}}_j \widetilde{\boldsymbol{\Gamma}}_j}{\sum_{j \in \widetilde{\mathcal{S}}} \widetilde{\boldsymbol{\gamma}}_j^2}. \tag{4.3.13}$$

The corresponding confidence interval for β^* would be

$$\left(\widehat{\beta}_H - z_{1-\alpha/2}\sqrt{\widehat{\mathrm{V}}_H/n}, \quad \widehat{\beta}_H + z_{1-\alpha/2}\sqrt{\widehat{\mathrm{V}}_H/n}\right). \tag{4.3.14}$$

Here, $\widehat{\mathrm{V}}_H$ is $\widehat{\mathrm{V}}$ in (4.3.11) except we replace $\widetilde{\mathcal{V}} = \widetilde{\mathcal{S}}$ and $\widehat{\beta}$ with $\widehat{\beta}_H$.

4.3.6 Special case of procedure 1: invalid instruments after controlling for low dimensional covariates

The second special case worth examining is the problem of invalid instruments after controlling for low dimensional covariates. While the plausibility of candidate IVs adherence to assumptions (A1)-(A3), especially (A3), is higher with many covariates, the low dimensional setting has recently received much attention and is discussed in Bowden et al. (2015, 2016); Burgess et al. (2016); Kang et al. (2016a), and Windmeijer et al. (2016). As we will see below, our procedure simplifies greatly and with a minor modification, our estimator, unlike the estimators proposed in said prior literature, achieves optimal performance.

Specifically, under the low-dimensional scenario, there is no need to use the debiased scaled lasso in STEP 1 of Procedure 1. Instead, we can replace STEP 1 with the simple ordinary least square (OLS) estimates of the reduced-forms, $(\widetilde{\mathbf{\Gamma}}, \widetilde{\mathbf{\Psi}})^\intercal = (\mathbf{W}^\intercal\mathbf{W})^{-1}\mathbf{W}^\intercal\mathbf{Y}$ and $(\widetilde{\gamma}, \widetilde{\boldsymbol{\psi}})^\intercal = (\mathbf{W}^\intercal\mathbf{W})^{-1}\mathbf{W}^\intercal\mathbf{D}$, and of the covariance terms

$$\widehat{\Theta}_{11} = \left\|\mathbf{Y} - \mathbf{Z}\widehat{\mathbf{\Gamma}} - \mathbf{X}\widehat{\mathbf{\Psi}}\right\|_2^2/n, \quad \widehat{\Theta}_{22} = \left\|\mathbf{D} - \mathbf{Z}\widehat{\boldsymbol{\gamma}} - \mathbf{X}\widehat{\boldsymbol{\psi}}\right\|_2^2/n,$$

and

$$\widehat{\Theta}_{12} = \left(\mathbf{Y} - \mathbf{Z}\widehat{\mathbf{\Gamma}} - \mathbf{X}\widehat{\mathbf{\Psi}}\right)^\intercal \left(\mathbf{D} - \mathbf{Z}\widehat{\boldsymbol{\gamma}} - \mathbf{X}\widehat{\boldsymbol{\psi}}\right)/n.$$

As a result of using OLS in STEP 1, we need to replace $\widehat{\mathbf{u}}^{[j]}$ from (4.3.5) with $\widehat{\mathbf{u}}^{[j]} = (\widehat{\mathbf{\Sigma}})^{-1}_{\cdot j}$, $\widehat{\mathbf{\Sigma}} = \mathbf{W}^\intercal\mathbf{W}/n$, and replace the log terms in our thresholds in (4.3.7) and (4.3.8) from $\sqrt{\log p_{\mathrm{z}}}$ to $\sqrt{\log n}$.

We can then proceed to use the estimator defined in (4.3.10). Alternatively, we can use a modified version of $\widehat{\beta}$, denoted as $\widehat{\beta}_E$, using the weighting matrix $\mathbf{A} =$

$\widehat{\boldsymbol{\Sigma}}_{\widetilde{\mathcal{V}},\widetilde{\mathcal{V}}} - \widehat{\boldsymbol{\Sigma}}_{\widetilde{\mathcal{V}},\widetilde{\mathcal{V}}^c} \left(\widehat{\boldsymbol{\Sigma}}_{\widetilde{\mathcal{V}}^c,\widetilde{\mathcal{V}}^c}\right)^{-1} \widehat{\boldsymbol{\Sigma}}_{\widetilde{\mathcal{V}}^c,\widetilde{\mathcal{V}}}$

$$\widehat{\beta}_E = \frac{\widetilde{\boldsymbol{\gamma}}_{\widetilde{\mathcal{V}}}^{\intercal} A \widetilde{\boldsymbol{\Gamma}}_{\widetilde{\mathcal{V}}}}{\widetilde{\boldsymbol{\gamma}}_{\widetilde{\mathcal{V}}}^{\intercal} A \widetilde{\boldsymbol{\gamma}}_{\widetilde{\mathcal{V}}}} \tag{4.3.15}$$

along with the estimated standard error $\widehat{\mathrm{V}}_E = \widehat{\sigma}^2 / \widetilde{\boldsymbol{\gamma}}_{\widetilde{\mathcal{V}}}^{\intercal} A \widetilde{\boldsymbol{\gamma}}_{\widetilde{\mathcal{V}}}$ where $\widehat{\sigma}^2 = \widehat{\Theta}_{11} + \widehat{\beta}_E^2 \widehat{\Theta}_{22} - 2\widehat{\beta}_E \widehat{\Theta}_{12}$ and confidence interval

$$\left(\widehat{\beta}_E - z_{1-\alpha/2}\sqrt{\widehat{\mathrm{V}}_E/n}, \quad \widehat{\beta}_E + z_{1-\alpha/2}\sqrt{\widehat{\mathrm{V}}_E/n}\right). \tag{4.3.16}$$

Note that $\widehat{\beta}_E$ in (4.3.15) is reduced to $\widehat{\beta}$ in (4.3.10) by setting $A = \mathrm{I}$. As we will see in Section 4.4.1, our estimator $\widehat{\beta}_E$, compared to other estimators in prior work, achieves optimal performance in the sense that our performance is asymptotically identical to the TSLS estimator for β^* that knows which IVs are valid a priori, i.e. the set $\mathcal{V}^*$.

4.4 Theoretical results

In this section, we investigate the properties of the confidence interval proposed in Procedure 1. We first consider in Section 4.4.1 the coverage property in the case of invalid IVs with low dimensional covariates where p_x and p_z are fixed. In Section 4.4.2, we establish the coverage property for the general case, invalid IVs even after controlling for many covariates.

4.4.1 Invalid IVs after controlling for low dimensional covariates

We state the following mild assumption commonly used in the invalid IV literature

(IN1) (50% Rule) The number of valid IVs is more than half of the number of non-redundant IVs, that is $|\mathcal{V}^*| > \frac{1}{2}|\mathcal{S}^*|$.

We denote the assumption as "IN" since the assumption is specific to the case of invalid IVs. Assumption (IN1) is the generalization of the 50% rule in Kang et al. (2016b) and Han (2008) in the presence of possibly redundant IVs. In a nutshell, (IN1) states that if the number of invalid instruments is not too large, then we can detect the invalid IVs from valid IVs, without knowing a priori which IVs are valid or invalid; see Kang et al. (2016b) for a detailed discussion of this assumption and how this type of proportion-based assumption is a necessary component for identification of model parameters under invalid instruments.

Under the 50% assumption alone, Theorem 20 states that we can show that our procedure produces confidence intervals with desired coverage and optimal length in low dimensional settings where p_x and p_z are fixed.

Theorem 20. *Suppose that the assumption* (IN1) *holds. Then the following property holds for the estimator* $\widehat{\beta}_E$ *defined in* (4.3.15),

$$\sqrt{n}\left(\widehat{\beta}_E - \beta^*\right) \xrightarrow{d} N\left(0, \frac{\sigma^2}{\boldsymbol{\gamma}_{\mathcal{V}^*}^{*\intercal}\left(\boldsymbol{\Sigma}^*_{\mathcal{V}^*\mathcal{V}^*} - \boldsymbol{\Sigma}^*_{\mathcal{V}^*(\mathcal{V}^*)^C}\boldsymbol{\Sigma}^{*-1}_{(\mathcal{V}^*)^C(\mathcal{V}^*)^C}\boldsymbol{\Sigma}^*_{(\mathcal{V}^*)^C\mathcal{V}^*}\right)\boldsymbol{\gamma}^*_{\mathcal{V}^*}}\right). \quad (4.4.1)$$

Consequently, the confidence interval given in (4.3.16) *has asymptotically coverage probability* $1-\alpha$*, i.e.,*

$$\mathbf{P}\left\{\beta \in \left(\widehat{\beta}_E - z_{1-\alpha/2}\sqrt{\widehat{\mathrm{V}}_E/n}, \quad \widehat{\beta}_E + z_{1-\alpha/2}\sqrt{\widehat{\mathrm{V}}_E/n}\right)\right\} \to 1-\alpha. \quad (4.4.2)$$

We note that the proposed estimator $\widehat{\beta}_E$ has the same asymptotic variance as the oracle TSLS estimator with the prior knowledge of $\mathcal{V}^*$, which is shown to be efficient under the homoskedastic variance assumption (Theorem 5.2 in Wooldridge (2010)); consequently, our confidence interval asymptotically performs like the oracle TSLS confidence interval and is of optimal length. But, unlike TSLS, we achieve this oracle performance without prior knowledge of $\mathcal{V}^*$. We also note the estimators proposed in prior work, Bowden et al. (2015, 2016); Burgess et al. (2016); Kang et al. (2016a); Windmeijer et al. (2016), do not achieve oracle performance and TSLS-like efficiency.

4.4.2 Invalid IVs after controlling for high dimensional covariates

We now consider the coverage property for the general case, invalid IVs even after controlling for many confounders. We first introduce the usual regularity assumptions used in high-dimensional statistical inference (Bickel et al., 2009; Bühlmann & van de Geer, 2011; Cai & Guo, 2016b).

(R1) (Coherence): The matrix $\mathbf{\Sigma}^*$ satisfies $1/M_1 \leq \lambda_{\min}(\mathbf{\Sigma}^*) \leq \lambda_{\max}(\mathbf{\Sigma}^*) \leq M_1$ for some constant $M_1 > 1$ and has bounded sub-Gaussian norm.

(R2) (Normality): The error terms in (4.3.1) and (4.3.2) follow a bivariate normal distribution.

(R3) (Global IV Strength): The IVs are globally strong with $\|\boldsymbol{\gamma}^*_{\mathcal{V}^*}\|_2 = \sqrt{\sum_{j \in \mathcal{V}^*} \gamma_j^2} \geq \delta \gg s_{z1} \log p/\sqrt{n}$, where $\mathcal{V}^*$ is the set of valid IVs defined in Definition 3.

Assumption (R1) places a condition on the spectrum of the design matrix $\mathbf{W}$ and the tail distribution of $\mathbf{W}_{i,\cdot}$, which is related to the restricted eigenvalue condition in Bickel et al. (2009). For simplicity, we also assume that the sub-Gaussian norm of $\mathbf{W}_{i\cdot}$ is upper bounded by M_1, that is, $\sup_{\boldsymbol{v} \in S^{p-1}} \sup_{q \geq 1} (\mathbb{E}|\boldsymbol{v}^\intercal \mathbf{W}_{i\cdot}|^q/q)^{\frac{1}{q}} \leq M_1$ where S^{p-1} is the unit sphere in $\mathbb{R}^p$; see Vershynin (2012) for details on sub-Gaussian random variables and bounds. Assumption (R2) states that the errors (e_{i1}, ϵ_{i2}) are bivariate normal. Here, we make the normality assumption out of simplicity, similar to the work on inference in weak IV literature where error terms are typically assumed to be normal (e.g. Section 2 of Moreira (2003) and Section 2.2.1 of Andrews et al. (2007)).

Finally, Assumption (R3) states that the global strength of instruments, measured by the ℓ_2 norm of $\boldsymbol{\gamma}^*$ among valid IVs $\mathcal{V}^*$, is bounded away from zero. This type of global strength assumption is commonly made in the IV literature under the guise as a concentration parameter, which is a measure of strength of the instrument (see Section 4.5 for details) and is the weighted ℓ_2 norm of $\boldsymbol{\gamma}^*_{\mathcal{V}^*}$, and is often referred to as "traditional/strong" asymptotics (Stock et al., 2002; Wooldridge, 2010). Recent works by Belloni et al. (2012) and Chernozhukov et al. (2015a), which considered the setting where all IVs were valid after conditioning on high dimensional covariates, also make this type of assumption, specifically condition SM in Belloni et al. (2012) and condition RF in the supplementary materials of Chernozhukov et al. (2015a). Essentially, both these works require $\|\boldsymbol{\gamma}^*\|_2$ to be bounded away from zero by a constant and are actually stronger than our (R3). In practice, (R3) is satisfied so long as there is at least one IV that has a constant non-zero effect on the treatment, or a non-zero effect that doesn't diminish with sample size. However, if the IVs are arbitrary weak in the sense of Staiger & Stock (1997), then (R3), let alone the said assumptions in high dimensional valid IV literature (Belloni et al., 2012), do not hold, and we leave this as a future topic of research to deal with arbitrary weak IVs in invalid IV settings.

Section C.1 in Chapter C shows that if the IVs are valid after conditioning on many covariates, then Asumptions (R1)-(R3) are sufficient for the confidence interval proposed in (4.3.14) to have correct coverage. However, when IVs are invalid after conditioning on said controls, we need to make two additional assumptions that are not in the usual high dimensional inference or instrumental variables literature and may be of theoretical interest in future work.

(IN2) (Individual IV Strength) For IVs in $\mathcal{S}^*$, $\min_{j \in \mathcal{S}^*} \left|\gamma_j^*\right| \geq \delta_{\min} \gg \sqrt{\log p / n}$.

(IN3) (Strong violation) Among IVs in the set $\mathcal{S}^* \setminus \mathcal{V}^*$, we have

$$\min_{j \in \mathcal{S}^* \setminus \mathcal{V}^*} \left| \frac{\pi_j^*}{\gamma_j^*} \right| \geq \frac{12(1 + |\beta^*|)}{\delta_{\min}} \sqrt{\frac{M_1 \log p_{\mathrm{z}}}{\lambda_{\min}(\mathbf{\Theta}^*) n}}. \tag{4.4.3}$$

Assumption (IN2) requires individual IV strength to be bounded away from zero so that all IVs in selected $\widetilde{\mathcal{S}}$ are strong. This assumption is needed primarily for cleaner technical exposition and our simulation studies in Section 4.5 demonstrates that (IN2) is largely unnecessary for our confidence interval to guarantee coverage. In the literature, (IN2) is similar to the "beta-min" condition assumption in high dimensional linear regression without IVs, with the exception that this condition is not imposed on our inferential quantity of interest, β^*. Also, (IN2) is different from Assumption (R3), where (R3) only requires the global IV strength to be bounded away from zero. Next, Assumption (IN3) requires the ratios π_j^*/γ_j^* for invalid IVs to be large and this assumption is needed to correctly identify IVs that violate (A2) and (A3). Specifically, for any IV with $|\pi_j^*/\gamma_j^*|$ being non-zero but small, it's difficult to distinguish such a weakly invalid IV from valid IVs where $\pi_j^*/\gamma_j^* = 0$. If a weakly invalid IV is mistakenly declared as valid, the bias from this mistake is of the order $\sqrt{\log p_{\mathrm{z}}/n}$, which has consequences, not for consistency of the point estimate, but for a $\sqrt{n}$ confidence interval; see Theorem 21 and Section 4.7 for more discussions.

With (R1)-(R3) and (IN1)-(IN3), our general Procedure 1 produces a consistent and asymptotic normal estimate of β^* even if IVs are invalid after conditioning on high dimensional controls.

Theorem 21. *Suppose the assumptions* (R1) − (R3) *and* (IN1) − (IN2) *hold. As* $\sqrt{s_{\mathrm{z}1}}s \log p/\sqrt{n} \to 0$, *with probability larger than* $1 - c\left(p^{-c} + \exp\left(-cn\right)\right)$,

$$\left|\widehat{\beta} - \beta^*\right| \leq C \frac{1}{\delta_{\min}} \sqrt{\frac{\log p_{\mathrm{z}}}{n}}, \tag{4.4.4}$$

where $c, C > 0$ are constants independent of n and p. In addition, if (IN3) *holds, we have*

$$\sqrt{n}\left(\widehat{\beta}-\beta^*\right)=T^{\beta^*}+\Delta^{\beta^*} \tag{4.4.5}$$

where $T^{\beta^*} \mid \mathbf{W} \sim N(0, \mathrm{V})$, $\mathrm{V} = \sigma^2/\left(\sum_{\mathrm{j}\in\mathcal{V}^*}(\gamma_{\mathrm{j}}^*)^2\right)^2\left\|\sum_{\mathrm{j}\in\mathcal{V}^*}\gamma_{\mathrm{j}}^*\mathbf{W}\widehat{\mathbf{u}}^{[\mathrm{j}]}/\sqrt{\mathrm{n}}\right\|_2^2$, *and* $\Delta^{\beta^*}/\sqrt{\mathrm{V}} \xrightarrow{p} 0$ *as* $\sqrt{s_{\mathrm{z}1}}s\log p/\sqrt{n} \to 0$. *Consequently, the confidence interval given in* (4.3.12) *has asymptotically coverage probability* $1-\alpha$, *i.e.,*

$$\mathbf{P}\left\{\beta^* \in \left(\widehat{\beta}-z_{1-\alpha/2}\sqrt{\widehat{\mathrm{V}}/n}, \quad \widehat{\beta}+z_{1-\alpha/2}\sqrt{\widehat{\mathrm{V}}/n}\right)\right\} \to 1-\alpha. \tag{4.4.6}$$

In Theorem 21, the consistency of our estimator in (4.4.4) is established without (IN3) because the bias term $\sqrt{\log p_{\mathrm{z}}/n}$ discussed above is still going to zero. However, for $\sqrt{n}$ asymptotic normality, Theorem 21 requires Assumption (IN3) to eliminate said bias so that our confidence interval (4.3.12) has correct coverage even with invalid IVs and high dimensional controls.

4.5 Simulation

4.5.1 Setup

In addition to the theoretical analysis of our method in Section 4.4, we also conduct a simulation study to investigate (i) the performance of our method and other comparators and (ii) sensitivity of our method to violations of the regularity assumptions mentioned above, most notably (IN2) and (IN3). The data generating process for the simulation follows the models (4.2.2) and (4.2.3) in Section 4.2.2 with $p_{\mathrm{z}} = 100$ instruments and $p_x = 150$ covariates where $\mathbf{W}_{i.}$ is a multivariate normal with mean zero and covariance $\Sigma_{ij}^* = 0.5^{|i-j|}$ for $1 \le i, j \le 250$. The parameters for the models are: $\beta^* = 1$, $\phi^* = (0.6, 0.7, 0.8, \cdots, 1.5, 0, 0, \cdots, 0) \in \mathbb{R}^{150}$ so that $s_{\mathrm{x}1} = 10$, $\psi^* =$

$(1.1, 1.2, 1.3, \cdots, 2.0, 0, 0, \cdots, 0) \in \mathbb{R}^{150}$ so that $s_{x2} = 10$, and variance-covariance of the error terms are $\mathrm{Var}(\epsilon_{i1}) = \mathrm{Var}(\epsilon_{i2}) = 1.5$, and $\mathrm{Cov}(\epsilon_{i1}, \epsilon_{i2}) = 0.75$. Instruments that satisfy Assumption (A1) are $\mathcal{S}^* = \{1, \ldots, 7\}$ and instruments that satisfy all three IV assumptions (A1)-(A3) are $\mathcal{V}^* = \{1, 2, 3, 4, 5\}$; thus instruments 6 and 7 only satisfy (A1), but do not satisfy (A2) and (A3). We fix these values throughout the entire simulation study.

The parameters we vary in the simulation study are: the sample size n, the strength of IVs via $\boldsymbol{\gamma}^*$, and violations of (A2) and (A3) via $\boldsymbol{\pi}^*$. For sample size, we let $n = (100, 200, 300, 1000, 3000)$. For IV strength, we set $\boldsymbol{\gamma}^*_{\mathcal{V}^*} = K\,(1, 1, 1, 1, \rho_1)$ and $\boldsymbol{\gamma}^*_{\mathcal{S}^* \setminus \mathcal{V}^*} = K\,(1, 1)$ and $\boldsymbol{\gamma}_{(\mathcal{S}^*)^C} = \mathbf{0}$, where we vary K (to be discussed later) and $\rho_1 = (0, 0.1, 0.2)$ across simulations. The value K controls the global strength of instruments, with higher $|K|$ indicating strong instruments in a global sense. The value ρ_1 controls the relative individual strength of instruments, specifically between the first four instruments in $\mathcal{V}^*$ and the fifth instrument. For example, $\rho_1 = 0.2$ implies that the fifth IV's individual strength is only 20% of the other four valid instruments, i.e IVs 1 to 4. Also, varying ρ_1 would simulate the adherence of regularity assumption (IN2).

To specify K across simulations, we introduce a quantity we call the oracle concentration parameter (OCP) denoted as $C\,(\boldsymbol{\gamma}^*, \mathcal{V}^*, n)$

$$C\,(\boldsymbol{\gamma}^*, \mathcal{V}^*, n) = n \frac{\boldsymbol{\gamma}^{*\intercal}_{\mathcal{V}^*} \left(\boldsymbol{\Sigma}^*_{\mathcal{V}^*\mathcal{V}^*} - \boldsymbol{\Sigma}^*_{\mathcal{V}^*(\mathcal{V}^*)^C} \boldsymbol{\Sigma}^{*-1}_{(\mathcal{V}^*)^C(\mathcal{V}^*)^C} \boldsymbol{\Sigma}^*_{(\mathcal{V}^*)^C\mathcal{V}^*} \right) \boldsymbol{\gamma}^*_{\mathcal{V}^*}}{|\mathcal{V}^*|\Theta^*_{22}}, \qquad (4.5.1)$$

where $\boldsymbol{\Sigma}^*_{IJ}$ denotes the submatrix containing Σ^*_{ij} for $i \in I$ and $j \in J$ and $\boldsymbol{\gamma}^*_{\mathcal{V}^*}$ denotes the subvector containing γ^*_j for $j \in \mathcal{V}^*$. We define the OCP because the usual concentration parameter can be misleading when there are unknown redundant and invalid instruments and the OCP serves as a proxy for the usual concentration parameter.

Having defined the OCP, we can specify K as a function of n and $C\left(\boldsymbol{\gamma}^{*}, \mathcal{V}^{*}, n\right)$. Specifically, if n is set at a baseline of 100 and the simulation parameters $\mathcal{V}^{*}, \rho_{1}$, $\boldsymbol{\Sigma}^{*}$ and Θ_{22}^{*} are specified as above, we can find K for a particular value of the expected oracle concentration parameter $C\left(\boldsymbol{\gamma}^{*}, \mathcal{V}^{*}, 100\right)$. Thus, by varying $C\left(\boldsymbol{\gamma}^{*}, \mathcal{V}^{*}, 100\right) = (50, 100, 150, 200, 250, 500, 1000)$, we vary K.

Finally, we vary $\boldsymbol{\pi}^{*}$, which controls the validity of the IVs by defining $\pi_{j}^{*} = \rho_{2} \gamma_{j}^{*}$ for $j = 6, 7$ and $\pi_{j}^{*} = 0$ for all other j so that ρ_{2} controls the magnitude of the violation of IV assumptions (A2) and (A3) from the 6th and 7th instruments. In the ideal case, we would have $\rho_{2} = 0$ so that $\mathcal{S}^{*} = \mathcal{V}^{*} = \{1, 2, 3, 4, 5, 6, 7\}$. But, $\rho_{2} \neq 0$ implies that the last two instruments do not satisfy (A2) and (A3). As such, we vary $\boldsymbol{\pi}^{*}$ by varying $\rho_{2} = (0, 1, 2)$. Also, varying ρ_{2} would simulate the adherence of regularity assumption (IN3).

In summary, we vary n, the strength of IVs via $\boldsymbol{\gamma}^{*}$, and violations of (A2) and (A3) via $\boldsymbol{\pi}^{*}$ in our simulation study, with ρ_{1} and ρ_{2} simulating the adherence to the new regularity assumptions in the paper, (IN2), and (IN3), respectively. For the setting $n \leq p$, we compare our procedure to $\widehat{\beta}_{H}$, which assumes IVs are valid. For the setting $n \geq p$, we add two additional comparators, the two-stage least squares (TSLS) and OLS. TSLS is the most popular IV method where one regresses $\mathbf{D}$ on $\mathbf{Z}$ and $\mathbf{X}$, and uses the predicted value of $\mathbf{D}$ in the regression of $\mathbf{Y}$ on $\mathbf{X}$ and $\mathbf{D}$. Note that the way we implement TSLS mimics most practitioners' use of TSLS by simply assuming all the instruments $\mathbf{Z}$ are valid. OLS is defined as where one regresses $\mathbf{Y}$ on $\mathbf{D}$ and $\mathbf{X}$. OLS will be biased because of confounding on $\mathbf{D}$. Finally, for both low and high dimensional settings, we have the oracle TSLS where an oracle provides us with the true set of valid IVs, which will not occur in practice. Our simulations are repeated 500 times.

4.5.2 Results

We present the most representative results from our simulation study. First, Figure 4.1 considers the high dimensional setting with $n = 200$ and three comparators, our procedure $\widehat{\beta}$ that is robust to invalid IVs, our procedure $\widehat{\beta}_H$ that assumes all valid IVs, and the oracle TSLS. Columns "Weak" and "Strong" in the figure represent cases where $\rho_1 = 0.2$ and $\rho_1 = 0$, respectively. Columns "Valid" and "Invalid" represent cases where $\rho_2 = 0$ and $\rho_2 = 2$, respectively. The row "MAE" in the figure represents the median absolute error of the estimators, which measures the performance of the point estimators. The row "Coverage" represents the coverage performance of the confidence intervals. Finally, the row "Length" represents the average length of confidence intervals across simulations.

Both estimators $\widehat{\beta}$ and $\widehat{\beta}_H$ perform well in terms of estimation accuracy, coverage and length of confidence intervals and have similar performance to the benchmark, $\widehat{\beta}_{\text{oracle}}$, when all the instruments are valid (i.e. first and second columns of Figure 4.1). For example, in the MAE and length plots, the solid lines, which represent our estimator, the dashed lines, which represent our estimator assuming all valid IVs after conditioning on covariates, and the dotted lines, which represent the oracle, overlap with each other. However, if the instruments are invalid (i.e. the third and fourth columns of Figure 4.1), $\widehat{\beta}_H$ is not consistent and loses coverage, which makes sense since $\widehat{\beta}_H$ assumes all the IVs are valid after conditioning. However, our proposed estimator $\widehat{\beta}$ allows for possibly invalid instruments and performs as well as the oracle in terms of estimation accuracy and coverage. The average length of our robust confidence interval is only slightly larger than that of the oracle.

Figure 4.2 represents the same setting as Figure 4.1 except we now consider a larger sample size $n = 1000$. Even though n is larger than p, we still consider this to be in the many controls/high dimensional setting because the ratio of p to n is away from zero at 1/4. As expected, the estimators $\widehat{\beta}$ and $\widehat{\beta}_H$ along with the traditional TSLS estimator perform similarly to the oracle benchmark in terms of estimation accuracy, coverage and the length of confidence intervals when all the instruments are actually valid. For example, in the MAE plot of Figure 4.2, the solid, dashed, green and dotted lines, representing $\widehat{\beta}$, $\widehat{\beta}_H$, TSLS and the oracle, respectively, overlap with each other. Note that OLS cannot deal with confounding and hence, produces a biased estimate. However, when the instruments are invalid, the traditional TSLS estimator and $\widehat{\beta}_H$ are biased and fail to have the correct coverage. In contrast, the proposed estimator $\widehat{\beta}$ performs as well as the oracle estimator in terms of estimation accuracy and coverage, with the length of the proposed estimator being slightly longer than that for the oracle.

Finally, Figure 4.3 represents the setting where invalid instruments are present after conditioning on low dimensional covariates where $p_{\mathrm{z}} = 9$ and $p_x = 10$ so that no coefficients for ϕ^* and ψ^* are zero and the sample size is $n = 1000$. If we use the estimator $\widehat{\beta}_E$ defined in (4.3.15) and the confidence interval (4.3.16), the proposed procedure performs almost the same as the oracle in terms of accuracy, coverage property and length, which supports the theory established in Theorem 20. Note that the performance of our procedure under the low dimensional setting with invalid IVs does not rely on assumptions (R1)-(R3) and, more importantly, (IN2)-(IN3).

4.6 Application: causal effect of years of education on annual earnings

To demonstrate our procedure 1 in real settings, we analyze the causal effect of years of education on yearly earnings, which has been studied extensively in economics using IV methods (Angrist & Krueger, 1991; Card, 1993, 1999). The data comes

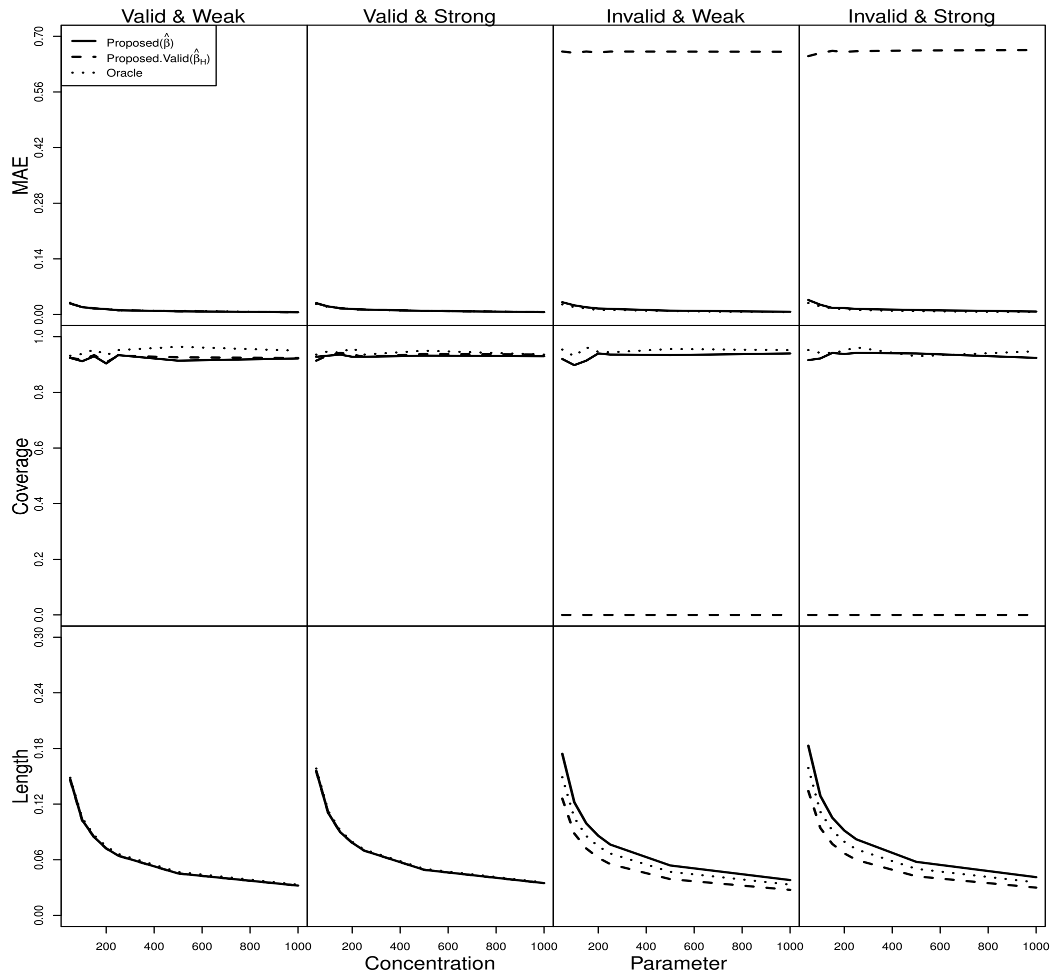

Figure 4.1: Comparison of different methods when $p_z = 100$, $p_x = 150$ and $n = 200$. The x-axis represents the concentration parameter. On the y-axis, MAE represents Median Absolute Error of the estimators, Coverage represents coverage of the confidence intervals and Length represents the average length of confidence intervals. Proposed is our method allowing for invalid IVs and is represented by the solid line. Proposed.valid is our method that assumes all the IVs are valid and is represented by the dashed line. Oracle is the method that knows exactly which instruments are valid and is represented by the dotted line. The column labeled with Valid & Weak represents the case $\rho_1 = 0.2$ and $\rho_2 = 0$. The column labeled with Valid &Strong represents the case $\rho_1 = 0$ and $\rho_2 = 0$. The column labeled with Invalid &Weak represents the case $\rho_1 = 0.2$ and $\rho_2 = 2$. Finally, the column labeled with Invalid & Strong represents the case $\rho_1 = 0$ and $\rho_2 = 2$.

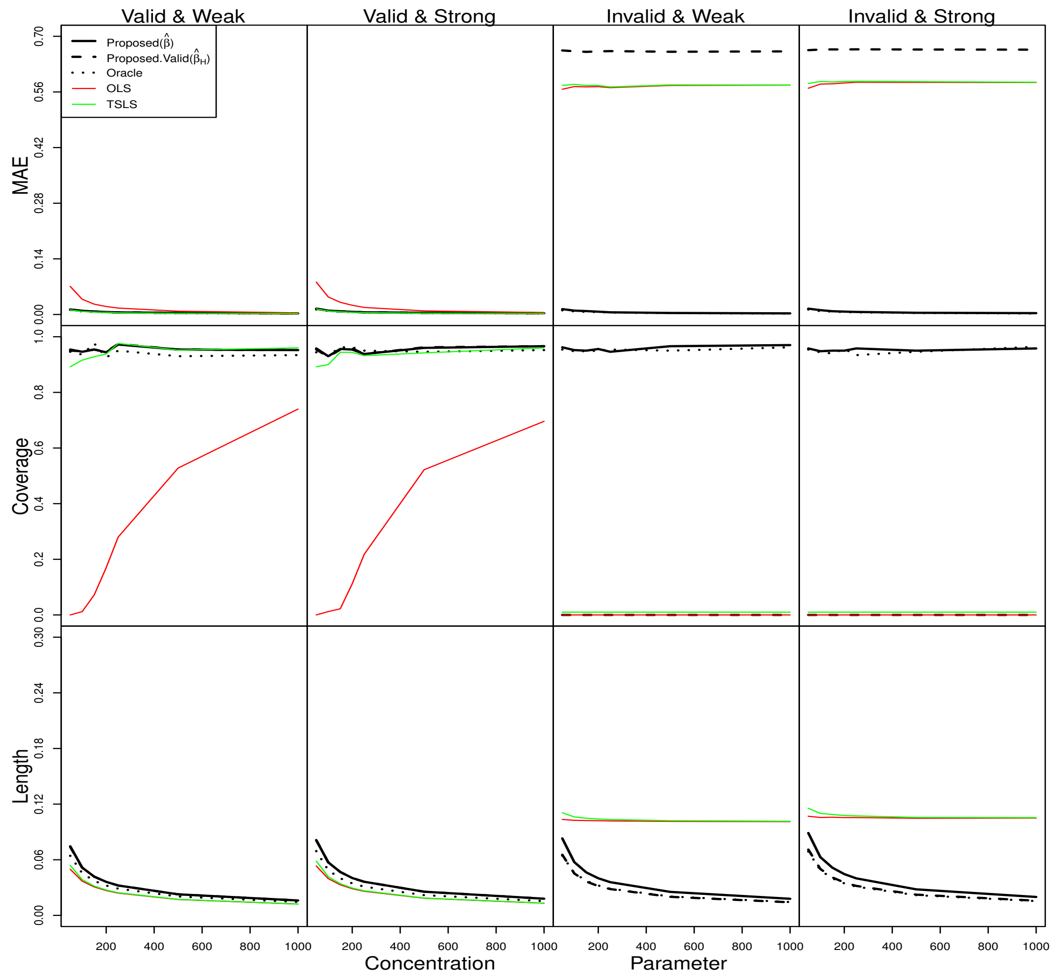

Figure 4.2: Comparison of different methods when $p_z = 100$, $p_x = 150$ and $n = 1000$. The x-axis represents the concentration parameter. On the y-axis, MAE represents Median Absolute Error of the estimators, Coverage represents coverage of confidence intervals and Length represents the average length of confidence intervals. Proposed is our method allowing for invalid IVs and is represented by the solid line. Proposed.valid is our method that assumes all the IVs are valid and is represented by the dashed line. Oracle is the method that knows exactly which instruments are valid and is represented by the dotted line. The column labeled with Valid & Weak represents the case $\rho_1 = 0.2$ and $\rho_2 = 0$. The column labeled with Valid &Strong represents the case $\rho_1 = 0$ and $\rho_2 = 0$. The column labeled with Invalid &Weak represents the case $\rho_1 = 0.2$ and $\rho_2 = 2$. Finally, the column labeled with Invalid & Strong represents the case $\rho_1 = 0$ and $\rho_2 = 2$.

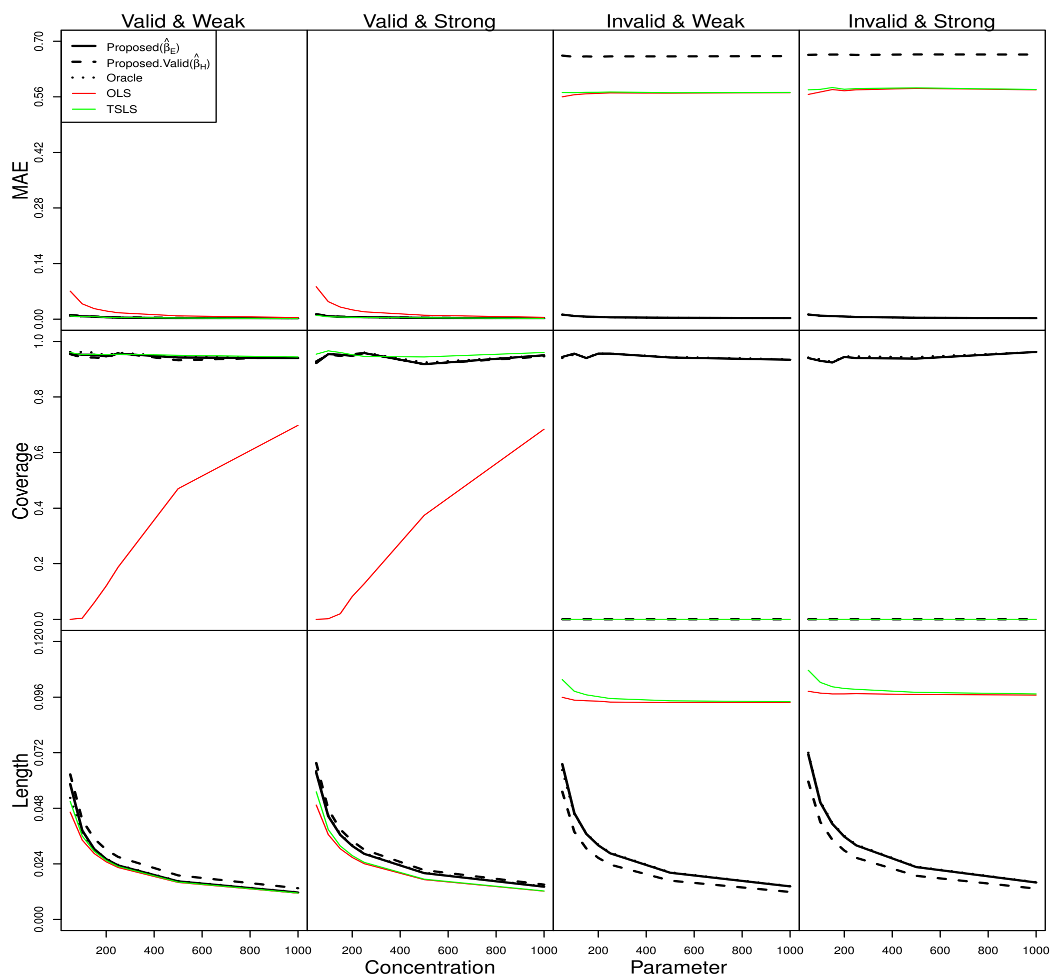

Figure 4.3: Comparison of different methods when $p_z = 9$, $p_x = 10$ and $n = 1000$. The x-axis represents the concentration parameter. On the y-axis, MAE represents Median Absolute Error of the estimators, Coverage represents coverage of confidence intervals and Length represents the average length of confidence intervals. Proposed is our method allowing for invalid IVs and is represented by the solid line. Proposed.valid is our method that assumes all the IVs are valid and is represented by the dashed line. Oracle is the method that knows exactly which instruments are valid and is represented by the dotted line. The column labeled with Valid & Weak represents the case $\rho_1 = 0.2$ and $\rho_2 = 0$. The column labeled with Valid &Strong represents the case $\rho_1 = 0$ and $\rho_2 = 0$. The column labeled with Invalid &Weak represents the case $\rho_1 = 0.2$ and $\rho_2 = 2$. Finally, the column labeled with Invalid & Strong represents the case $\rho_1 = 0$ and $\rho_2 = 2$.

from the Wisconsin Longitudinal Study (WLS), a longitudinal study that has kept track of American high school graduates from Wisconsin since 1957, and we examine the relationship between graduates' earnings and education from the 1974 survey (Hauser, 2005), roughly 20 years after they graduated from high school. Our analysis includes $N = 3772$ individuals, 1784 males and 1988 females. For our outcome, we use imputed log total yearly earnings prepared by WLS (see WLS documentation and Hauser (2005) for details) and for the treatment, we use the total years of education, all from the 1974 survey. The median total earnings is $9,200 with a 25% quartile of $1,000 and a 75% quartile of $15,320 in 1974 dollars. The mean years of total education is 13.7 years with a standard deviation of 2.3 years.

We incorporate many covariates, including sex, graduate's hometown population, educational attainment of graduates' parents, graduates' family income, relative income in graduates' hometown, graduates' high school denomination, high school class size, all measured in 1957 when the participants were high school seniors. We also include 81 genetic covariates, specifically single nucleotide polymorphisms (SNPs), that were part of WLS to further control for potential variations between graduates. In summary, our data analysis includes 7 non-genetic covariates and 81 genetic covariates.

We used five instruments in our analysis, all derived from past studies of education on earnings (Card, 1993; Blundell et al., 2005; Gary-Bobo et al., 2006). They are (i) total number of sisters, (ii) total number of brothers, (iii) individual's birth order in the family, all from Gary-Bobo et al. (2006), (iv) proximity to college from Card (1993), and (v) teacher's interest in individual's college education from Blundell et al. (2005), all measured in 1957. Although all these IVs have been suggested to be valid with varying explanations as to why they satisfy (A2) and (A3) after controlling for the aforementioned covariates, in practice, we are always uncertain due to the lack of complete socioeconomic knowledge about the effect of these IVs. Our method should provide some protection against this uncertainty compared to traditional methods where they simply assume that all five IVs are valid. Also, the first-stage F-test produces an F-statistic of 90.3 with a p-value less than 10^{-16}, which indicates very strong set of instruments.

Table 4.1 summarizes the results of our data analysis. OLS refers to running a regression of the treatment and the covariates on the outcome and looking at the slope coefficient of the treatment variable. TSLS refers to running two-stage least squares as described in Section 4.5 under the operating assumption that all the five instruments are valid; this is the usual and most popular analysis in the IV literature. Finally, we run the Pocedure 1.

Method	Point Estimate	95% Confidence Interval
OLS	0.097	(0.051, 0.143)
TSLS	0.169	(0.029, 0.301)
TSHT	0.062	(0.046, 0.077)

Table 4.1: Estimates of the Effect of Years of Education on Log Earnings. OLS is ordinary least squares, TSLS is two-stage least squares, and TSHT is Procedure 1.

The OLS estimate suggests a positive association between education and earnings, with statistically significant result at $\alpha = 0.05$ level. This agrees with previous literature which suggests a statistically significant positive association between years of education and log earnings (Card, 1999). However, OLS does not completely control for confounding even after controlling for covariates. TSLS provides an alternative method of controlling for confounding by using instruments so long as all the instruments satisfy the three core assumptions and the inclusion of covariates helps make these assumptions more plausible. Unfortunately, we notice that the TSLS estimate in Table 4.1 is inconsistent with previous studies' estimates among individuals from the U.S. between 1950s to 1970s, which range from 0.06 to 0.13 (see Table 4 in Card (1999)). Our method, which addresses the concern for invalid instruments with TSLS, provides an estimate of 0.062, which is more consistent with previous studies' estimates of the effect of years of education on earnings.

The data analysis suggests that our method can be a useful tool in IV analysis when there is concern for invalid instruments, even after attempting to mitigate this problem via covariates. Our method provides much more accurate estimates of the returns on education than TSLS, which naively assumes all the instruments are valid.

4.7 Conclusion and discussion

We present a method to estimate the effect of the treatment on the outcome using instrumental variables where we do not make the assumption that all the instruments are valid. Our approach is based on the novel TSHT procedure, which is shown to succeed in selecting valid IVs in the presence of possibly invalid IVs. Our approach provides robust confidence intervals in the presence of invalid IVs even after controlling for many covariates. In simulation and in real data settings, our approach provides a more robust analysis than the traditional IV approaches, most notably TSLS, by providing some protection against possibly invalid instruments.

As discussed in Section 4.4.2, our theoretical analysis for the case of invalid IVs even after controlling for high-dimensional covariates require Assumptions (IN2) and (IN3). While (IN2) is not crucial in practice as our simulation study demonstrates and is made for a cleaner technical exposition, we believe (IN3) is most likely necessary for invalid IV problems and this is echoed in the model selection literature by Leeb & Pötscher (2005) who pointed out that "in general no model selector can be uniformly consistent for the most parsimonious true model" and hence the post-model-selection inference is generally non-uniform. Consequently, the set of competing models has to be "well separated" such that we can consistently select a correct model and Assumption (IN3) serves as this "well separated" condition in our invalid IV problem. While some recent work in high dimensional inference (Zhang & Zhang, 2014; Javanmard & Montanari, 2014a; van de Geer et al., 2014; Chernozhukov et al., 2015a; Cai & Guo, 2016b) do not make this "well separated" assumption, as we stressed before, our invalid IV problem is of different nature than the prior work because a single invalid IV declared as valid can ruin inference while said prior works assume covariates are exogenous and moments are known perfectly.

Finally, in practice, we believe that violation of (IN3) in high dimensions will not drastically harm inference and our CI will still have coverage around $1 - \alpha$, which is much better than TSLS and prior work assuming valid IVs after conditioning on many covariates, which have no coverage. In particular, our empirical investigations generally show that the under-coverage is no more than 5% and we think this is partly due to the fact that (i) our procedure will still pick up the strongly invalid IVs and (ii) if the instruments are weakly invalid, the bias from them via π^* will be relatively small. It is certainly possible that advanced methods can weaken (IN3) and we leave this as a direction for further research.

A Supplement for Chapter 2

A.1 Proofs of Theorems

In this section, we provide detailed proofs of Theorem 4, Theorem 5, Theorem 6, Theorem 7 and Theorem 8.

A.1.1 Proof of Theorem 6

The proof is divided into three steps.

The first step. We construct the alternative hypothesis parameter space $\mathcal{H}_1$. Let $\mathcal{H}_0 = \{\theta^* = (\beta^*, \mathrm{I}, \sigma)\} \subset \Theta(k_1)$, $S_2 = \mathrm{supp}(\beta^*)$ and $S_3 = \mathrm{supp}(\xi) \backslash S_2$. Let k^* denote the size of S_2 and p_2 denote the size of S_2^c and p_3 denote the size of S_3. We have $S_3 \subset S_2^c$, $k^* \leq k_1$ and $p_3 \geq q - k_1$. We have the following expression for the covariance matrix Σ_0^z of $(y_1, X_{1\cdot})$ corresponding to the null parameter space $\mathcal{H}_0 = \{\theta^* = (\beta^*, \mathrm{I}, \sigma)\}$,

$$\Sigma_0^z = \left(\begin{array}{c|c|c} \|\beta^*\|_2^2 + \sigma^2 & \left(\beta^*_{S_2}\right)^\intercal & \mathbf{0}_{1\times p_2} \\ \hline \beta^*_{S_2} & \mathrm{I}_{k^*\times k^*} & \mathbf{0}_{k^*\times p_2} \\ \hline \mathbf{0}_{p_2\times 1} & \mathbf{0}_{p_2\times k^*} & \mathrm{I}_{p_2\times p_2} \end{array}\right). \tag{A.1.1}$$

Let $\xi^* = \min_{i\in S_3} |\xi_i| > 0$ and we define the following set,

$$\ell_1(p_2, \zeta_0 k, \rho) = \left\{ \boldsymbol{\delta} : \boldsymbol{\delta} \in \mathbb{R}^{p_2}, \operatorname{supp}(\boldsymbol{\delta}) \subset S_3, \|\boldsymbol{\delta}\|_0 = \zeta_0 k, \boldsymbol{\delta}_i \frac{\xi_i}{\xi^*} \in \{0, \rho\} \text{ for } i \in S_3 \right\}, \tag{A.1.2}$$

and then construct a parameter space $\mathcal{F}_1$ for Σ^z, $\mathcal{F}_1 = \{\Sigma^z_{\boldsymbol{\delta}} : \boldsymbol{\delta} \in \ell_1(p_2, \zeta_0 k, \rho)\}$ with

$$\Sigma^z_{\boldsymbol{\delta}} = \left(\begin{array}{c|c|c} \|\beta^*\|_2^2 + \sigma^2 & \left(\beta^*_{S_2}\right)^\intercal & \boldsymbol{\delta}^\intercal \\ \hline \beta^*_{S_2} & \mathrm{I}_{k^* \times k^*} & \mathbf{0}_{k^* \times p_2} \\ \hline \boldsymbol{\delta} & \mathbf{0}_{p_2 \times k^*} & \mathrm{I}_{p_2 \times p_2} \end{array} \right). \tag{A.1.3}$$

We construct the alternative hypothesis space $\mathcal{H}_1$ for (β, Ω, σ), which is induced by the mapping h and the parameter space $\mathcal{F}_1$,

$$\mathcal{H}_1 = \left\{ (\beta, \Omega, \sigma) \ : \ (\beta, \Omega, \sigma) = h(\Sigma^z) \quad \text{for } \Sigma^z \in \mathcal{F}_1 \right\}. \tag{A.1.4}$$

Under the alternative joint distribution (A.1.3), for the linear model expression $y_i = X_{i,S_2}\beta_{S_2} + X_{i,S_2^c}\beta_{S_2^c} + \epsilon'_i$, we have

$$\beta_{S_2} = \beta^*_{S_2}, \quad \beta_{S_2^c} = \boldsymbol{\delta}, \quad \text{and} \quad \operatorname{Var}(\epsilon'_i) = \sigma^2 - \|\boldsymbol{\delta}\|_2^2 \leq M_2. \tag{A.1.5}$$

Based on (A.1.5), the sparsity of β under the alternative is upper bounded by $|\operatorname{supp}(\beta^*)| + |\operatorname{supp}(\boldsymbol{\delta})| \leq (1 - \zeta_0) k + \zeta_0 k = k$. Since we do not perturb the covariance matrix for $X_{1,\cdot}$, the precision matrix for $X_{1,\cdot}$ in the alternative hypothesis space satisfies the conditions in $\Theta(k)$. Hence, we show that $\mathcal{H}_1 \subset \Theta(k)$.

The second step Let π denote the uniform prior on $\boldsymbol{\delta}$ over $\ell_1(p_2, \zeta_0 k, \rho)$. Note that this uniform prior π induces a prior distribution $\pi_{\mathcal{H}_1}$ over the parameter space $\mathcal{H}_1$. The following lemma controls the χ^2 distance between the null and the alternative.

Lemma 9.

$$\chi^2\left(f_{\pi_{\mathcal{H}_1}}, f_{\pi_{\mathcal{H}_0}}\right) + 1 = \mathbb{E}_{\boldsymbol{\delta},\widetilde{\boldsymbol{\delta}}}\left(1 - \frac{1}{\sigma^2}\boldsymbol{\delta}^\intercal\widetilde{\boldsymbol{\delta}}\right)^{-n}.$$

By the inequality $\frac{1}{1-x} \le \exp(2x)$ for $x \in \left[0, \frac{\log 2}{2}\right]$, for $\frac{1}{\sigma^2}\boldsymbol{\delta}^\intercal\widetilde{\boldsymbol{\delta}} \le \frac{1}{\sigma^2}\zeta_0 k\rho^2 < \frac{\log 2}{2}$, then we have $\left(1 - \frac{1}{\sigma^2}\boldsymbol{\delta}^\intercal\widetilde{\boldsymbol{\delta}}\right)^{-n} \le \exp\left(\frac{2}{\sigma^2}n\boldsymbol{\delta}^\intercal\widetilde{\boldsymbol{\delta}}\right)$. By Lemma 3 in Chapter 2 , we further have

$$\begin{aligned}
&\mathbb{E}\exp\left(\frac{2}{\sigma^2}n\boldsymbol{\delta}^\intercal\widetilde{\boldsymbol{\delta}}\right) = \mathbb{E}\exp\left(\frac{1}{\sigma^2}2Jn\rho^2\right) \le e^{\frac{\zeta_0^2k^2}{p_3-\zeta_0 k}}\left(1 - \frac{\zeta_0 k}{p_3} + \frac{\zeta_0 k}{p_3}\exp\left(\frac{1}{\sigma^2}2n\rho^2\right)\right)^{\zeta_0 k} \\
&\le e^{\frac{\zeta_0^2k^2}{p_3-\zeta_0 k}}\left(1 - \frac{\zeta_0 k}{p_3} + \frac{\zeta_0 k}{p_3}\sqrt{\frac{p_3}{\zeta_0^2k^2}}\right)^{\zeta_0 k} \le e^{\frac{c^2\zeta_0^2p^{2\gamma}}{p_3-c\zeta_0p^\gamma}}\left(1 + \frac{1}{\sqrt{p_3}}\right)^{c\zeta_0 p^\gamma},
\end{aligned} \tag{A.1.6}$$

where the second inequality follows by plugging $\rho = \sqrt{\frac{\log\frac{p_3}{\zeta_0^2k^2}}{4n}}\sigma$. If $k \le c\min\left\{\frac{n}{\log p}, p^\gamma\right\}$ for a sufficiently small positive constant c, we have $k\rho^2 < \frac{\log 2}{2\zeta_0}\sigma^2$. Since $p_3 \ge q - k_1 \ge cp^{2\gamma+\varsigma} - k_1$ with $0 < \varsigma < 1 - 2\gamma$, we have $\chi^2\left(f_{\pi_{\mathcal{H}_1}}, f_{\pi_{\mathcal{H}_0}}\right) \le \left(\frac{1}{2} - \alpha\right)^2$ and hence $\mathrm{TV}\left(f_{\pi_{\mathcal{H}_1}}, f_{\pi_{\mathcal{H}_0}}\right) \le \frac{1}{2} - \alpha$.

The third step. We calculate the distance between μ_1 and μ_0. Under $\mathcal{H}_0$, $\mu_0 = \xi_{S_2}^\intercal\beta_{S_2}^*$. Under $\mathcal{H}_1$, we have $\mu_1 = \xi_{S_2}^\intercal\beta_{S_2}^* + \xi_{S_2^c}^\intercal\boldsymbol{\delta}$. Note that for $\boldsymbol{\delta} \in \ell_1(p_2, \zeta_0 k, \rho)$, we have $\mu_1 = \xi_{S_2}^\intercal\beta_{S_2}^* + \xi^*\zeta_0 k\rho$. Hence, μ_1 is a fixed constant for a given ρ. For $\xi \in \Xi(q, \bar{c})$,

$$|\mu_1 - \mu_0| = \xi^*\zeta_0 k\rho \ge c\|\xi\|_\infty\zeta_0 k\sqrt{\frac{\log\frac{p_3}{\zeta_0^2k^2}}{n}}\sigma \ge c\|\xi\|_\infty k\sqrt{\frac{\log p}{n}}\sigma,$$

where the last inequality follows from $p_3 \ge q - k_1 \ge cp^{2\gamma+\varsigma} - k_1$ with $0 < \varsigma < 1 - 2\gamma$. By (2.7.3) of Lemma 1 in Chapter 2, we establish (2.4.7) in the main paper.

A.1.2 Proof of Theorem 5

Theorem 5 follows from Theorem 6. Given $0 < \zeta_0 < 1$, we define $k_1^* = \min\{k_1, (1-\zeta_0)k - 1\}$. By taking $\theta^* \in \Theta(k_1^*)$, we have

$$L_\alpha^*(\Theta(k_1^*), \Theta(k), \xi^\intercal\beta) \geq \inf_{\mathrm{CI}_\alpha(\xi^\intercal\beta, Z) \in \mathcal{I}_\alpha(\Theta(k), \xi^\intercal\beta)} \mathbb{E}_{\theta^*} L\left(\mathrm{CI}_\alpha\left(\xi^\intercal\beta, Z\right)\right)$$

and hence Theorem 5 follows from the fact that $L_\alpha^*(\Theta(k_1), \Theta(k), \xi^\intercal\beta) \geq L_\alpha^*(\Theta(k_1^*), \Theta(k), \xi^\intercal\beta)$.

A.1.3 Proof of Theorem 4

The minimax lower bound of Theorem 4 follows from Theorem 5 with taking $k_1 = k$. The following proposition establishes the minimax upper bound of Theorem 4.

Proposition 8. *Suppose that $k \leq c_* \frac{n}{\log p}$, where c_* is a small positive constant, then*

$$\liminf_{n,p\to\infty} \inf_{\theta \in \Theta(k)} \mathbb{P}_\theta\left(\xi^\intercal\beta \in \mathrm{CI}^D_{\xi^\intercal\beta}\left(Z, k\right)\right) = 1, \tag{A.1.7}$$

and

$$L\left(\mathrm{CI}^D_{\xi^\intercal\beta}\left(Z, k\right), \Theta\left(k\right)\right) \leq C\|\xi\|_\infty k\sqrt{\frac{\log p}{n}}, \tag{A.1.8}$$

for some constant $C > 0$.

In the following, we will establish Proposition 8. On the event $S \cap G$, we have

$$\left|\xi^\intercal\widehat{\beta} - \xi^\intercal\beta\right| \leq \|\xi\|_\infty \left\|\widehat{\beta} - \beta\right\|_1 \leq \|\xi\|_\infty \left(2 + 2\epsilon_0\right) \frac{\sqrt{n}}{\min\|X_{\cdot j}\|_2} l\left(Z, k\right), \tag{A.1.9}$$

where $l\left(Z, k\right)$ is defined in (2.7.36) in Chapter 2. On event S_2, if $p \geq \exp\left(2M_2\right)$, then $\hat{\sigma} < \log p$. Hence, the event A holds and

$$\mathrm{CI}^D_\alpha\left(\xi^\intercal\beta, Z\right) = \left[\xi^\intercal\widehat{\beta} - \|\xi\|_\infty\rho_2(k),\ \xi^\intercal\widehat{\beta} + \|\xi\|_\infty\rho_2(k)\right].$$

By (2.7.38) in Chapter 2, on the event $G \cap S$, if $p \geq p_0$, $\|\xi\|_\infty \rho_2(k)$ is equal to the right hand side of (A.1.9), and hence $\xi^\intercal\beta \in \mathrm{CI}_\alpha^D(\xi^\intercal\beta, Z)$. We have the following inequality about the coverage probability, $\mathbb{P}_\theta\left(\xi^\intercal\beta \in \mathrm{CI}_\alpha^D(\xi^\intercal\beta, Z)\right) \geq \mathbb{P}_\theta(S \cap G)$. By Lemma 4, we establish (A.1.7). We control the expected length as follows,

$$\begin{aligned}
&\mathbb{E}_\theta L\left(\mathrm{CI}_\alpha^D(\xi^\intercal\beta, Z)\right) = \mathbb{E}_\theta L\left(\mathrm{CI}_\alpha^D(\xi^\intercal\beta, Z)\right)\mathbf{1}_A \\
=&\mathbb{E}_\theta L\left(\mathrm{CI}_\alpha^D(\xi^\intercal\beta, Z)\right)\mathbf{1}_{A\cap S\cap G} + \mathbb{E}_\theta L\left(\mathrm{CI}_\alpha^D(\xi^\intercal\beta, Z)\right)\mathbf{1}_{A\cap(S\cap G)^c} \\
\leq& C\|\xi\|_\infty\left(k\sqrt{\frac{\log p}{n}}\sigma + (\log p)^2 k\sqrt{\frac{\log p}{n}}\mathbb{P}_\theta\left((S\cap G)^c\right)\right) \\
\leq& C\|\xi\|_\infty k\sqrt{\frac{\log p}{n}}\left(\sigma + C\left(p^{1-\min\left\{\delta_0, C_1, c_0C_0^2\right\}} + c'\exp(-cn)\right)(\log p)^2\right),
\end{aligned} \tag{A.1.10}$$

where the first inequality follows from (2.7.38) in Chapter 2 and second inequality follows from Lemma 4. If $\frac{\log p}{n} \leq c$, then $\mathbb{E}_\theta L\left(\mathrm{CI}_\alpha^D(\xi^\intercal\beta, Z)\right) \leq C\|\xi\|_\infty k\sqrt{\frac{\log p}{n}}M_2$.

A.1.4 Proof of Theorem 7

The following proposition shows that the confidence interval proposed in (2.5.5) in Chapter 2 has the desired coverage property and achieves the minimax lower bound of Theorem 7.

Proposition 9. *Suppose $k \leq c_* \frac{n}{\log p}$, where c_* is a small positive constant, then*

$$\liminf_{n,p\to\infty} \inf_{\theta\in\Theta(k,\mathrm{I},\sigma_0)} \mathbb{P}_\theta\left(\xi^\intercal\beta \in \mathrm{CI}_\alpha^\mathrm{I}(\xi^\intercal\beta, Z)\right) \geq 1-\alpha, \tag{A.1.11}$$

and

$$L\left(\mathrm{CI}_\alpha^\mathrm{I}(\xi^\intercal\beta, Z), \Theta(k, \mathrm{I}, \sigma_0)\right) \leq C\frac{\|\xi\|_2}{\sqrt{n}}. \tag{A.1.12}$$

for some constant $C > 0$.

The control of the expected length (A.1.12) follows from the construction (2.5.5).

In the following, we control the coverage property (A.1.11). Let $\widehat{\Sigma}^{(2)} = \frac{1}{n_2}\left(X^{(2)}\right)^{\intercal} X^{(2)}$. We have the following decomposition,

$$\bar{\mu} - \xi^{\intercal}\beta = \xi^{\intercal}\left(\mathrm{I} - \widehat{\Sigma}^{(2)}\right)\left(\widehat{\beta} - \beta\right) + \frac{1}{n_2}\xi^{\intercal}\left(X^{(2)}\right)^{\intercal}\epsilon^{(2)}. \tag{A.1.13}$$

with $\frac{1}{n_2}\xi^{\intercal}\left(X^{(2)}\right)^{\intercal}\epsilon \mid X \sim N\left(0, \sigma_0^2\frac{\xi^{\intercal}\widehat{\Sigma}^{(2)}\xi}{n_2}\right)$. Before controlling the terms in the right hand side of (A.1.13), we introduce the following definitions. We state the definition of $\kappa(X, k, s, \alpha_0)$, which was introduced in Bickel et al. (2009),

$$\kappa(X, k, s, \alpha_0) = \min_{\substack{J_0 \subset \{1,\cdots,p\},\\ |J_0| \leq k}} \min_{\substack{\delta \neq 0,\\ \|\delta_{J_0^c}\|_1 \leq \alpha_0\|\delta_{J_0}\|_1}} \frac{\|X\delta\|_2}{\sqrt{n}\|\delta_{J_{01}}\|_2}, \tag{A.1.14}$$

where J_1 denotes the subset corresponding to the s largest in absolute value coordinates of δ outside of J_0 and $J_{01} = J_0 \cup J_1$. Let $W_{\cdot j}^{(1)} = X_{\cdot j}^{(1)}\frac{\sqrt{n_1}}{\|X_{\cdot j}^{(1)}\|_2}$ for $j = 1, \cdots, p$. Define the following events

$$\begin{aligned}
\bar{G}_1 &= \left\{\frac{2}{5}\frac{1}{\sqrt{M_1}} < \frac{\|X_{j\cdot}^{(1)}\|_2}{\sqrt{n_1}} < \frac{7}{5}\sqrt{M_1} \text{ for } 1 \leq j \leq p\right\},\\
\bar{G}_2 &= \left\{\kappa(X^{(1)}, k, k, \alpha) \geq \frac{1}{4\sqrt{\lambda_{\max}(\Omega)}} - \frac{9}{\sqrt{\lambda_{\min}(\Omega)}}(1+\alpha)\sqrt{2k\frac{\log p}{n_1}}\right\},\\
\bar{G}_3 &= \left\{\frac{\|\left(W^{(1)}\right)^{\intercal}\epsilon^{(1)}\|_\infty}{n_1} \leq \sigma_0\sqrt{\frac{2\delta_0\log p}{n_1}}\right\},\\
\bar{G}_4 &= \left\{\left|\frac{\xi^{\intercal}\widehat{\Sigma}^{(2)}\xi}{\xi^{\intercal}\xi} - 1\right| \leq 2\sqrt{\frac{\log p}{n_2}} + 2\frac{\log p}{n_2}\right\},\\
\bar{G}_5 &= \left\{\left|\xi^{\intercal}\left(\mathrm{I} - \widehat{\Sigma}^{(2)}\right)\left(\widehat{\beta} - \beta\right)\right| \leq 8\sqrt{6\log\frac{2}{(1-\gamma_0)\,\alpha}}\|\xi\|_2\|\widehat{\beta} - \beta\|_2\frac{1}{\sqrt{n_2}}M_1\right\}.
\end{aligned}$$

The following lemma controls the probability of the events $\bar{G}_i$ with $1 \leq i \leq 5$,

Lemma 10. *If $k \leq c\frac{n}{\log p}$, then*

$$\min\left\{\mathbb{P}_{\theta}\left(\bar{G}_1 \cap \bar{G}_2 \cap \bar{G}_3\right), \mathbb{P}_{\theta}\left(\bar{G}_4\right)\right\} \geq 1 - c\exp\left(-c'n\right) - cp^{-c''}, \tag{A.1.15}$$

where c, c' and c'' are positive constants. We also have

$$\mathbb{P}_{\theta}\left(\bar{G}_5\right) \geq 1 - \left(1 - \gamma_0\right)\alpha \tag{A.1.16}$$

By the proof of Theorem 7.2 in Bickel et al. (2009) and Theorem 3 in Ye & Zhang (2010), on the event $\bar{G}_1 \cap \bar{G}_2 \cap \bar{G}_3$, $\|\widehat{\beta} - \beta\|_2 \leq C\sqrt{\frac{k\log p}{n}}\sigma_0$. On the event $\bar{G}_1 \cap \bar{G}_2 \cap \bar{G}_3 \cap \bar{G}_5$, we have

$$\left|\xi^{\intercal}\left(\mathrm{I} - \widehat{\Sigma}^{(2)}\right)\left(\widehat{\beta} - \beta\right)\right| \leq C\|\xi\|_2\frac{\sqrt{k\log p}}{n_2}\sigma_0 \leq 0.005\frac{\|\xi\|_2}{\sqrt{n_2}}z_{\alpha_0/2}\sigma_0,$$

where the last inequality follows from the assumption that $k \leq c\frac{n}{\log p}$. On the event $\bar{G}_4$, we have $\sigma_0^2\frac{\xi^{\intercal}\widehat{\Sigma}^{(2)}\xi}{n_2} \leq 1.001\frac{\|\xi\|_2^2}{n_2}\sigma_0^2$. To sum up, if $k \leq c\frac{n}{\log p}$, we have $\mathbb{P}_{\theta}\left(\xi^{\intercal}\beta \in \mathrm{CI}_{\alpha}^{\mathrm{I}}\left(\xi^{\intercal}\beta, Z\right)\right) \geq \mathbb{P}_{\theta}\left(\cap_{i=1}^{5}\bar{G}_i\right) \geq 1 - \alpha - c\exp\left(-c'n\right) - cp^{-c''}$.

A.1.5 Proof of Theorem 8

It is sufficient to establish the following lower bounds,

$$L_{\alpha}^{*}\left(\Theta\left(k_1, \mathrm{I}, \sigma_0\right), \Theta\left(k, \mathrm{I}, \sigma_0\right), \xi^{\intercal}\beta\right) \geq c_1\|\xi\|_{\infty}\sigma_0\sqrt{kk_1}\sqrt{\frac{\log p}{n}}; \tag{A.1.17}$$

and

$$L_{\alpha}^{*}\left(\Theta\left(k_1, \mathrm{I}, \sigma_0\right), \Theta\left(k, \mathrm{I}, \sigma_0\right), \xi^{\intercal}\beta\right) \geq c_1\|\xi\|_{\infty}\sigma_0\min\left\{k\sqrt{\frac{\log p}{n}}, \frac{\sqrt{k}}{n^{\frac{1}{4}}}\right\}. \tag{A.1.18}$$

Proof of (A.1.17)

It is sufficient to establish (A.1.17) for $k_1 \leq (1-\zeta_0)k - 1$ with $0 < \zeta_0 < 1$ being a constant. Set $k_1^* = \min\{k_1, (1-\zeta_0)k - 1\}$. If $k_1^* \leq k_1 \leq k$, we can establish (A.1.17) by the following observation,

$$
\begin{aligned}
&L_{\alpha}^{*}\left(\Theta\left(k_1, \mathrm{I}, \sigma_0\right), \Theta\left(k, \mathrm{I}, \sigma_0\right), \xi^{\intercal}\beta\right) \geq L_{\alpha}^{*}\left(\Theta\left(k_1^*, \mathrm{I}, \sigma_0\right), \Theta\left(k, \mathrm{I}, \sigma_0\right), \xi^{\intercal}\beta\right) \\
&\geq c_1 \|\xi\|_{\infty} \sigma_0 \sqrt{k k_1^*} \sqrt{\frac{\log p}{n}} \geq c_1 \|\xi\|_{\infty} \sigma_0 \sqrt{k k_1} \sqrt{\frac{\log p}{n}},
\end{aligned}
$$

where the first inequality follows from $k_1^* \leq k_1$, the second inequality follows from the assumption that (A.1.17) holds for $k_1^* \leq (1-\zeta_0)k - 1$ and the last inequality follows from $k_1^* \geq (1-\zeta_0)k_1 - 1$.

In the following, we will establish (A.1.17) for $k_1 \leq (1-\zeta_0)k - 1$ with $0 < \zeta_0 < 1$ being a constant. The proof of (A.1.17) is more complicated than the previous lower bound argument since we need to compare two composite hypothesis.

<u>The first step.</u> Let $\xi^* = \min_{i \in \mathrm{supp}(\xi)} |\xi_i| > 0$ and we define the following set,

$$
\ell_2\left(p, k, \rho\right) = \left\{\boldsymbol{\delta} : \boldsymbol{\delta} \in \mathbb{R}^p, \|\boldsymbol{\delta}\|_0 = k, \boldsymbol{\delta}_i \frac{\xi_i}{\xi^*} \in \{0, \rho\} \text{ for } i \in \mathrm{supp}(\xi)\right\}. \tag{A.1.19}
$$

We construct the following two parameter spaces for Σ^z,

$$
\mathcal{F}_0 = \left\{\Sigma_{\boldsymbol{\nu}}^z : \boldsymbol{\nu} \in \ell_2\left(p, k_1, \rho\right)\right\}, \quad \text{where} \quad \Sigma_{\boldsymbol{\nu}}^z = \left(\begin{array}{c|c} \|\boldsymbol{\nu}\|_2^2 + \sigma_0^2 & \boldsymbol{\nu}^{\intercal} \\ \hline \boldsymbol{\nu} & \mathrm{I}_{p\times p} \end{array}\right). \tag{A.1.20}
$$

and

$$
\mathcal{F}_1 = \left\{\Sigma_{\boldsymbol{\delta}}^z : \boldsymbol{\delta} \in \ell_2\left(p, k, \sqrt{\frac{k_1}{k}}\rho\right)\right\}, \quad \text{where} \quad \Sigma_{\boldsymbol{\delta}}^z = \left(\begin{array}{c|c} \|\boldsymbol{\delta}\|_2^2 + \sigma_0^2 & \boldsymbol{\delta}^{\intercal} \\ \hline \boldsymbol{\delta} & \mathrm{I}_{p\times p} \end{array}\right). \tag{A.1.21}
$$

Then we construct the parameter spaces $\mathcal{H}_i$, which is induced by the mapping h and the parameter space $\mathcal{F}_i$ for $i = 0, 1$,

$$\mathcal{H}_i = \{(\beta, \Omega, \sigma) \ : \ (\beta, \Omega, \sigma) = h\left(\Sigma^z\right) \quad \text{for } \Sigma^z \in \mathcal{H}_i\} \quad \text{for} \quad i = 1, 2. \tag{A.1.22}$$

By (A.1.20) and (A.1.21), we have

$$\begin{aligned}\mathcal{H}_0 &= \{(\boldsymbol{\nu}, \mathrm{I}, \sigma_0) \ : \ \boldsymbol{\nu} \in \ell_2\left(p, k_1, \rho\right)\} \subset \Theta\left(k_1, \sigma_0, \mathrm{I}\right); \\ \mathcal{H}_1 &= \left\{(\boldsymbol{\delta}, \mathrm{I}, \sigma_0) \ : \ \boldsymbol{\delta} \in \ell_2\left(p, k, \sqrt{\frac{k_1}{k}}\rho\right)\right\} \subset \Theta\left(k, \sigma_0, \mathrm{I}\right).\end{aligned} \tag{A.1.23}$$

<u>The second step</u> Let π_0 denote the uniform prior over $\ell_2\left(p, k_1, \rho\right)$ and π_1 denote the uniform prior over $\ell_2\left(p, k, \sqrt{\frac{k_1}{k}}\rho\right)$. Let $\pi_{\mathcal{H}_0}$ denote the uniform prior on $\mathcal{H}_0$ induced by the prior π_0 and $\pi_{\mathcal{H}_1}$ denote the uniform prior on $\mathcal{H}_1$ induced by the prior π_1. To calculate the distance $\mathrm{TV}\left(f_{\pi_{\mathcal{H}_1}}, f_{\pi_{\mathcal{H}_0}}\right)$, we introduce $\mathcal{H} = \{(0, \mathrm{I}, \sigma_0^2 + k_1\rho^2)\}$ with $\pi_{\mathcal{H}}$ denoting the mass prior at this point. Since

$$\mathrm{TV}\left(f_{\pi_{\mathcal{H}_1}}, f_{\pi_{\mathcal{H}_0}}\right) \leq \mathrm{TV}\left(f_{\pi_{\mathcal{H}_0}}, f_{\pi_{\mathcal{H}}}\right) + \mathrm{TV}\left(f_{\pi_{\mathcal{H}_1}}, f_{\pi_{\mathcal{H}}}\right) \leq \sum_{i=0}^{1} \sqrt{\chi^2\left(f_{\pi_{\mathcal{H}_0}}, f_{\pi_{\mathcal{H}}}\right)},$$

it is sufficient to control $\chi^2\left(f_{\pi_{\mathcal{H}_i}}, f_{\pi_{\mathcal{H}_0}}\right)$ for $i = 0, 1$. Applying Lemma 9 with $S_2 = \emptyset$, we have

$$\begin{aligned}\chi^2\left(f_{\pi_{\mathcal{H}_0}}, f_{\pi_{\mathcal{H}}}\right) + 1 &= \mathbb{E}_{\boldsymbol{\nu}, \widetilde{\boldsymbol{\nu}}}\left(1 - \frac{1}{\sigma_0^2 + k_1\rho^2}\boldsymbol{\nu}^{\intercal}\widetilde{\boldsymbol{\nu}}\right)^{-n}. \\ \chi^2\left(f_{\pi_{\mathcal{H}_1}}, f_{\pi_{\mathcal{H}}}\right) + 1 &= \mathbb{E}_{\boldsymbol{\delta}, \widetilde{\boldsymbol{\delta}}}\left(1 - \frac{1}{\sigma_0^2 + k_1\rho^2}\boldsymbol{\delta}^{\intercal}\widetilde{\boldsymbol{\delta}}\right)^{-n}.\end{aligned} \tag{A.1.24}$$

In the following, we will control $\chi^2\left(f_{\pi_{\mathcal{H}_1}}, f_{\pi_{\mathcal{H}}}\right)+1$ and the argument for $\chi^2\left(f_{\pi_{\mathcal{H}_0}}, f_{\pi_{\mathcal{H}}}\right)+1$ is similar. By the inequality $\frac{1}{1-x} \leq \exp(2x)$ for $x \in \left[0, \frac{\log 2}{2}\right]$, if $\frac{1}{\sigma_0^2+k_1\rho^2}\boldsymbol{\delta}^{\intercal}\widetilde{\boldsymbol{\delta}} \leq \bar{c}^2\frac{k_1\rho^2}{\sigma_0^2} < \frac{\log 2}{2}$, then $\left(1 - \frac{1}{\sigma_0^2+k_1\rho^2}\boldsymbol{\delta}^{\intercal}\widetilde{\boldsymbol{\delta}}\right)^{-n} \leq \exp\left(\frac{2n}{\sigma_0^2}\boldsymbol{\delta}^{\intercal}\widetilde{\boldsymbol{\delta}}\right)$. By Lemma 3 in Chapter 2, we further have

$$\mathbb{E}\exp\left(\frac{2n}{\sigma_0^2}\boldsymbol{\delta}^\intercal\widetilde{\boldsymbol{\delta}}\right)\leq\mathbb{E}\exp\left(\frac{2nk_1}{\sigma_0^2 k}\bar{c}^2 J\rho^2\right)\leq e^{\frac{k^2}{q-k}}\left(1-\frac{k}{q}+\frac{k}{q}\exp\left(\frac{2\bar{c}^2 nk_1}{\sigma_0^2 k}\rho^2\right)\right)^{\zeta_0 k}$$
$$\leq e^{\frac{k^2}{q-k}}\left(1-\frac{k}{q}+\frac{k}{q}\sqrt{\frac{q}{k^2}}\right)^k\leq e^{\frac{k^2}{q-k}}\left(1+\frac{1}{\sqrt{q}}\right)^k,$$

where the second inequality follows by plugging $\rho=\sqrt{\frac{\log\frac{q}{k^2}}{4\bar{c}^2 n}}\sigma_0$. If $k\leq c\min\left\{\frac{n}{\log p},p^\gamma\right\}$ for a sufficiently small positive constant c, we have $k\rho^2<\frac{\log 2}{2}\sigma_0^2$. Since $q-k\geq cp^{2\gamma+\varsigma}$ with $0<\varsigma<1-2\gamma$, we have $\chi^2\left(f_{\pi_{\mathcal{H}_1}},f_{\pi_{\mathcal{H}}}\right)\leq\left(\frac{1}{4}-\frac{\alpha}{2}\right)^2$ and $\mathrm{TV}\left(f_{\pi_{\mathcal{H}_1}},f_{\pi_{\mathcal{H}}}\right)\leq\frac{1}{4}-\frac{\alpha}{2}$. Similarly, we can establish $\chi^2\left(f_{\pi_{\mathcal{H}_0}},f_{\pi_{\mathcal{H}}}\right)\leq\left(\frac{1}{4}-\frac{\alpha}{2}\right)^2$ and hence $\mathrm{TV}\left(f_{\pi_{\mathcal{H}_1}},f_{\pi_{\mathcal{H}_0}}\right)\leq\frac{1}{2}-\alpha$.

The third step. We calculate the distance between μ_1 and μ_0. Under $\mathcal{H}_0$, $\mu_0=\xi^* k_1\rho$. Under $\mathcal{H}_1$, we have $\mu_1=\xi^* k\sqrt{\frac{k_1}{k}}\rho$. Hence, μ_0 and μ_1 are fixed constants for a given ρ. For $\xi\in\Xi\left(q,\bar{c}\right)$ and $k_1\leq(1-\zeta_0)k-1$, we have $|\mu_1-\mu_0|\geq c\|\xi\|_\infty\sqrt{kk_1}\sqrt{\frac{\log p}{n}}\sigma_0$. By (2.7.3) of Lemma 1 in Chapter 2, we establish (A.1.17).

Proof of (A.1.18)

The first step. We construct the null parameter space $\mathcal{H}_0=\{(0,\mathrm{I},\sigma_0^2)\}$. We construct the following parameter space for Σ^z,

$$\mathcal{F}_1=\left\{\Sigma^z_{\boldsymbol{\delta}}:\boldsymbol{\delta}\in\ell_2\left(p,k,\rho\right)\right\},\quad\text{where}\quad\Sigma^z_{\boldsymbol{\delta}}=\left(\begin{array}{c|c}\|\boldsymbol{\delta}\|_2^2+\sigma_0^2 & \boldsymbol{\delta}^\intercal\\ \hline \boldsymbol{\delta} & \mathrm{I}_{p\times p}\end{array}\right).\tag{A.1.25}$$

Then we construct the parameter space $\mathcal{H}_1$, which is induced by the mapping h and the parameter space $\mathcal{F}_1$, $\mathcal{H}_1=\left\{(\beta,\Omega,\sigma):(\beta,\Omega,\sigma)=h\left(\Sigma^z\right)\text{ for }\Sigma^z\in\mathcal{F}_1\right\}$. By (A.1.20) and (A.1.21), we have $\mathcal{H}_1=\left\{(\boldsymbol{\delta},\mathrm{I},\sigma_0):\boldsymbol{\delta}\in\ell_2\left(p,k,\rho\right)\right\}\subset\Theta\left(k,\mathrm{I},\sigma_0\right)$.

The second step Let π_0 denote the mass prior at the point $(0,\mathrm{I},\sigma_0^2)$ and π_1 denote the uniform prior over $\ell_2\left(p,k,\rho\right)$. Let $\pi_{\mathcal{H}_0}$ denote the mass prior on $\mathcal{H}_0$ induced by the prior π_0 and $\pi_{\mathcal{H}_1}$ denote the uniform prior on $\mathcal{H}_1$ induced by the prior π_1. To calculate the distance $\mathrm{TV}\left(f_{\pi_{\mathcal{H}_1}}, f_{\pi_{\mathcal{H}_0}}\right)$, we introduce $\mathcal{H}=\{(0, \mathrm{I}, \sigma_0^2+k\rho^2)\}$ with $\pi_{\mathcal{H}}$ denoting the mass prior at this point. Since $\mathrm{TV}\left(f_{\pi_{\mathcal{H}_1}}, f_{\pi_{\mathcal{H}_0}}\right) \leq \sqrt{\chi^2\left(f_{\pi_{\mathcal{H}_0}}, f_{\pi_{\mathcal{H}}}\right)} + \sqrt{\chi^2\left(f_{\pi_{\mathcal{H}_1}}, f_{\pi_{\mathcal{H}}}\right)}$, it is sufficient to control $\chi^2\left(f_{\pi_{\mathcal{H}_1}}, f_{\pi_{\mathcal{H}}}\right)$ and $\chi^2\left(f_{\pi_{\mathcal{H}_0}}, f_{\pi_{\mathcal{H}}}\right)$. Applying Lemma 9 with $S_2=\emptyset$, we have

$$\chi^2\left(f_{\pi_{\mathcal{H}_1}}, f_{\pi_{\mathcal{H}}}\right)+1=\mathbb{E}_{\boldsymbol{\delta},\widetilde{\boldsymbol{\delta}}}\left(1-\frac{1}{\sigma_0^2+k\rho^2}\boldsymbol{\delta}^{\intercal}\widetilde{\boldsymbol{\delta}}\right)^{-n}. \tag{A.1.26}$$

We also have

$$\chi^2\left(f_{\pi_{\mathcal{H}_0}}, f_{\pi_{\mathcal{H}}}\right)+1=\mathbb{E}_{\boldsymbol{\delta},\widetilde{\boldsymbol{\delta}}}\left(1-\frac{\left(k\rho^2\right)^2}{\sigma_0^4}\right)^{-\frac{n}{2}}. \tag{A.1.27}$$

By the similar argument with (A.1.24), if $\rho=c\sqrt{\frac{\log p}{n}}\sigma_0$, then $\chi^2\left(f_{\pi_{\mathcal{H}_1}}, f_{\pi_{\mathcal{H}}}\right) \leq \left(\frac{1}{4}-\frac{\alpha}{2}\right)^2$. If $\rho=c\frac{1}{\sqrt{k}n^{\frac{1}{4}}}\sigma_0$, then $\chi^2\left(f_{\pi_{\mathcal{H}_0}}, f_{\pi_{\mathcal{H}}}\right) \leq \left(\frac{1}{4}-\frac{\alpha}{2}\right)^2$. If we take

$$\rho=c\min\left\{\sqrt{\frac{\log p}{n}}\sigma_0, \frac{1}{\sqrt{k}n^{\frac{1}{4}}}\sigma_0\right\},$$

then we have $\mathrm{TV}\left(f_{\pi_{\mathcal{H}_1}}, f_{\pi_{\mathcal{H}_0}}\right) \leq \frac{1}{2}-\alpha$.

The third step. We calculate the distance between μ_1 and μ_0. Under $\mathcal{H}_0$, $\mu_0=0$. Under $\mathcal{H}_1$, we have $\mu_1=\xi^* k\rho$. Hence, μ_0 and μ_1 are fixed constants for a given ρ. For $\xi \in \Xi\left(q, \bar{c}\right)$, $|\mu_1-\mu_0| \geq c\|\xi\|_\infty k\rho = c\|\xi\|_\infty \min\left\{k\sqrt{\frac{\log p}{n}}\sigma_0, \frac{\sqrt{k}}{n^{\frac{1}{4}}}\sigma_0\right\}$. By (2.7.3) of Lemma 1 in Chapter 2 , we establish (A.1.18).

A.2 Proof of lemmas

In this section, we prove Lemma 2, 3, 4, 5, 6 and 9. We prove Lemma 2, 3 and 9 in Section A.2.2 , prove Lemma 4 in Section A.2.3, prove Lemma 5 in Section A.2.4, prove Lemma 6 in Section A.2.5 and prove Lemma 10 in Section A.2.6.

A.2.1 Technical lemmas

We introduce the following technical lemmas. The first lemma (Theorem 2.3 in Boucheron et al. (2013)) is a concentration result of χ^2 random variable.

Lemma 11. *Let χ_n^2 denote the χ^2 random variable with n degrees of freedom, then we have the following concentration inequality,*

$$\mathbb{P}\left(\left|\chi_n^2 - E\chi_n^2\right| > 2\sqrt{nt} + 2t\right) \leq 2\exp(-t).$$

The following lemma (Theorem 1 in Raskutti et al. (2010)) establishes the concentration result for restricted eigenvalue in the Gaussian design.

Lemma 12. *For any Gaussian random design $X \in \mathbb{R}^{n\times p}$ with i.i.d $N(0,\Sigma)$ rows and define $\rho(\Sigma) = \sqrt{\max_{j=1,\cdots,p} \Sigma_{jj}}$, there are universal positive constants c, c' such that*

$$\frac{\|Xv\|_2}{\sqrt{n}} \geq \frac{1}{4}\|\Sigma^{\frac{1}{2}}v\|_2 - 9\rho(\Sigma)\sqrt{\frac{\log p}{n}}\|v\|_1 \quad \text{for all} \quad v \in \mathbb{R}^p, \tag{A.2.1}$$

with probability at least $1 - c'\exp(-cn)$.

The following lemmas are useful in controlling the χ^2 distance between the null and the alternative hypothesis. The first lemma is established in Cai & Zhou (2012); Ren et al. (2013).

Lemma 13. *Let g_i be the density function of $N(0,\Sigma_i)$ for $i = 0, 1, 2$, respectively. Then*

$$\int \frac{g_1 g_2}{g_0} = \left(\det\left(I - \Sigma_0^{-1}(\Sigma_1 - \Sigma_0)\Sigma_0^{-1}(\Sigma_2 - \Sigma_0)\right)\right)^{-\frac{1}{2}}.$$

Lemma 14.

$$\begin{aligned}&\chi^2\left(f_{\pi_{\mathcal{H}_1}}, f_{\pi_{\mathcal{H}_0}}\right)\\&=\int\left(\det\left(I-\left(\Sigma_0^z\right)^{-1}\left(\Sigma_{\widetilde{\boldsymbol{\delta}}}^z-\Sigma_0^z\right)\left(\Sigma_0^z\right)^{-1}\left(\Sigma_{\boldsymbol{\delta}}^z-\Sigma_0^z\right)\right)\right)^{-\frac{n}{2}}\pi\left(\boldsymbol{\delta}\right)\pi\left(\widetilde{\boldsymbol{\delta}}\right)d\boldsymbol{\delta}d\widetilde{\boldsymbol{\delta}},\end{aligned} \tag{A.2.2}$$

Lemma 14 follows from the definition of $f_{\pi_{\mathcal{H}_1}}, f_{\pi_{\mathcal{H}_0}}$, Fubini's theorem and Lemma 13. For reasons of space, the proof is omitted here.

The following lemma establishes that restricted eigenvalue is a lower bound for CIF_1 and the function ω defined in (2.4.10) in Chapter 2 is a further lower bound for the restricted eigenvalue.

Lemma 15. *The ℓ_1 cone invertibility factor CIF_1 is lower bounded by restricted eigenvalue,*

$$CIF_1\left(1+2\epsilon_0, T, W\right) \geq \frac{n}{\left(2+2\epsilon_0\right)\max\|X_{\cdot j}\|_2^2}\kappa^2\left(X, k, \left(1+2\epsilon_0\right)\left(\frac{\max\|X_{\cdot j}\|_2}{\min\|X_{\cdot j}\|_2}\right)\right). \tag{A.2.3}$$

On the event G_4,

$$\kappa^2\left(X, k, \left(1+2\epsilon_0\right)\left(\frac{\max\|X_{\cdot j}\|_2}{\min\|X_{\cdot j}\|_2}\right)\right) \geq \omega\left(\Omega, X, k\right), \tag{A.2.4}$$

where $\omega\left(\Omega, X, k\right)$ is defined in (2.4.10) *in Chapter 2. On the event $G_1 \cap G_4$, if $k \leq c\frac{n}{\log p}$, then*

$$CIF_1\left(1+2\epsilon_0, T, W\right) \geq \frac{n}{\left(2+2\epsilon_0\right)\max\|X_{\cdot j}\|_2^2}\omega\left(\Omega, X, k\right) \geq C\left(M_1\right). \tag{A.2.5}$$

where $C\left(M_1\right) = \frac{25}{3136\left(2+2\epsilon_0\right)M_1^2}$.

Proof of Lemma 15. We first prove the following useful inequalities,

$$\kappa^2\left(W,k,\alpha_0\right) \geq \frac{n}{\max\|X_{\cdot j}\|_2^2}\kappa^2\left(X,k,\alpha_0\left(\frac{\max\|X_{\cdot j}\|_2}{\min\|X_{\cdot j}\|_2}\right)\right), \tag{A.2.6}$$

and

$$CIF_1\left(2\epsilon_0+1,T,W\right) \geq \frac{1}{2+2\epsilon_0}\kappa^2\left(W,k,1+2\epsilon_0\right). \tag{A.2.7}$$

Proof of (A.2.6). By the normalization, $W = XD$ where D is defined in (2.7.29) in Chapter 2. For fixed δ and J_0 such that $\|\delta_{J_0^c}\|_1 \leq \alpha_0\|\delta_{J_0}\|_1$, we have

$$\frac{\|W\delta\|_2^2}{n\|\delta_{J_0}\|_2^2} = \frac{\|XD\delta\|_2^2}{n\|\delta_{J_0}\|_2^2} \geq \frac{\|XD\delta\|_2^2\left(\min\frac{\sqrt{n}}{\|X_{\cdot j}\|_2}\right)^2}{n\|D_{J_0\times J_0}\delta_{J_0}\|_2^2},$$

where $D_{J_0\times J_0}$ is the submatrix of D with row indices J_0 and column indices J_0. Let $u = D\delta$, since $\|\delta_{J_0^c}\|_1 \leq \alpha_0\|\delta_{J_0}\|_1, \|u_{J_0^c}\|_1 \leq \max\frac{\sqrt{n}}{\|X_{\cdot j}\|_2}\|\delta_{J_0^c}\|_1$ and $\|u_{J_0}\|_1 \geq \min\frac{\sqrt{n}}{\|X_{\cdot j}\|_2}\|\delta_{J_0}\|_1$, we have $\|u_{J_0^c}\|_1 \leq \alpha_0\left(\frac{\max\|X_{\cdot j}\|_2}{\min\|X_{\cdot j}\|_2}\right)\|u_{J_0}\|_1$. Hence,

$$\begin{aligned}
\kappa^2(W,k,\alpha_0) &= \min_{\substack{J_0\subset\{1,\cdots,p\},\\ |J_0|\leq k}} \min_{\substack{\delta\neq 0,\\ \|\delta_{J_0^c}\|_1\leq\alpha_0\|\delta_{J_0}\|_1}} \frac{\|W\delta\|_2^2}{n\|\delta_{J_0}\|_2^2} \\
&\geq \min_{\substack{J_0\subset\{1,\cdots,p\},\\ |J_0|\leq k}} \min_{\substack{u\neq 0,\\ \|u_{J_0^c}\|_1\leq\alpha_0\left(\frac{\max\|X_{\cdot j}\|_2}{\min\|X_{\cdot j}\|_2}\right)\|u_{J_0}\|_1}} \left(\min_{1\leq i\leq p}\frac{\sqrt{n}}{\|X_{\cdot j}\|_2}\right)^2 \frac{\|Xu\|_2^2}{n\|u_{J_0}\|_2^2} \\
&\geq \frac{n}{\max\|X_{\cdot j}\|_2^2}\kappa^2\left(X,k,\alpha_0\left(\frac{\max\|X_{\cdot j}\|_2}{\min\|X_{\cdot j}\|_2}\right)\right).
\end{aligned}$$

Proof of (A.2.7).

$$\begin{aligned}
&\min_{\substack{J_0\subset\{1,\cdots,p\},\\ |J_0|\leq |T|}} \min_{\substack{\delta\neq 0,\\ \|\delta_{J_0^c}\|_1\leq\alpha_0\|\delta_{J_0}\|_1}} \frac{\|W\delta\|_2^2}{n\|\delta_{J_0}\|_2^2} \leq \min_{\substack{\delta\neq 0,\\ \|\delta_{T^c}\|_1\leq\alpha_0\|\delta_T\|_1}} \frac{\|W\delta\|_2^2}{n\|\delta_T\|_2^2} \leq \min_{\substack{\delta\neq 0,\\ \|\delta_{T^c}\|_1\leq\alpha_0\|\delta_T\|_1}} \frac{\|\delta\|_1\|\frac{W^\intercal W}{n}\delta\|_\infty}{\|\delta_T\|_2^2} \\
&\leq \min_{\substack{\delta\neq 0,\\ \|\delta_{T^c}\|_1\leq\alpha_0\|\delta_T\|_1}} \frac{\left(1+\alpha_0\right)\|\delta_T\|_1\|\frac{W^\intercal W}{n}\delta\|_\infty}{\|\delta_T\|_2^2} \leq \min_{\substack{\delta\neq 0,\\ \|\delta_{T^c}\|_1\leq\alpha_0\|\delta_T\|_1}} \frac{\left(1+\alpha_0\right)|T|\|\frac{W^\intercal W}{n}\delta\|_\infty}{\|\delta_T\|_1}.
\end{aligned}$$

Since $|T| \leq k$, the definitions of $\kappa^2(W, |T|, \alpha_0)$ and $CIF_1(\alpha_0, T, W)$ lead to

$$\kappa^2(W, |T|, \alpha_0) \geq \kappa^2(W, k, \alpha_0),$$

and hence $CIF_1(1 + 2\epsilon_0, T, W) \geq \frac{1}{2+2\epsilon_0}\kappa^2(W, k, 1 + 2\epsilon_0)$. Combining (A.2.6) and (A.2.7), we establish (A.2.3). The lower bound (A.2.4) follows from the definition of the event G_4. On the event G_1, we have $\frac{\max \|X_{\cdot j}\|_2}{\min \|X_{\cdot j}\|_2} \leq \frac{7}{2}M_1$. If $k \leq c\frac{n}{\log p}$ with a sufficient small constant c, then we have

$$\begin{aligned} &\frac{1}{4\sqrt{\lambda_{\max}(\Omega)}} - 9\frac{1}{\sqrt{\lambda_{\min}(\Omega)}}\left(1 + (1 + 2\epsilon_0)\left(\frac{\max \|X_{\cdot j}\|_2}{\min \|X_{\cdot j}\|_2}\right)\right)\sqrt{k\frac{\log p}{n}} \\ \geq &\frac{1}{4\sqrt{M_1}} - 9\sqrt{M_1}\left(1 + (1 + 2\epsilon_0)\frac{7}{2}M_1\right)\sqrt{k\frac{\log p}{n}} > \frac{1}{8\sqrt{M_1}}. \end{aligned} \tag{A.2.8}$$

Combined with (A.2.3) and (A.2.4), we establish (A.2.5).

A.2.2 Proof of lemmas for the lower bound

We prove the Lemma 2, 9 and for lower bound in this section.

Proof of Lemma 2. Let g_0 denote the density function of $N(0, \Sigma_0^z)$, g_1 denote the density function of $N(0, \Sigma_{\boldsymbol{\delta}}^z)$ and g_2 denote the density function of $N(0, \Sigma_{\tilde{\boldsymbol{\delta}}}^z)$. By plugging into Lemma 13, we have

$$\int \frac{g_1 g_2}{g_0} = \left(\det\left(I - (\Sigma_0^z)^{-1}(\Sigma_{\boldsymbol{\delta}}^z - \Sigma_0^z)(\Sigma_0^z)^{-1}\left(\Sigma_{\tilde{\boldsymbol{\delta}}}^z - \Sigma_0^z\right)\right)\right)^{-\frac{1}{2}}. \tag{A.2.9}$$

Note that

$$(\Sigma_0^z)^{-1} = \left(\begin{array}{c|c|c|c} \frac{1}{\sigma^2} & -\frac{1}{\sigma^2}\psi_1^* & \left(-\frac{1}{\sigma^2}\psi_S^*\right)^\intercal & \mathbf{0}_{1\times p_1} \\ \hline -\frac{1}{\sigma^2}\psi_1^* & 1+\frac{1}{\sigma^2}\left(\psi_1^*\right)^2 & \frac{1}{\sigma^2}\psi_1^*\left(\psi_S^*\right)^\intercal & \mathbf{0}_{1\times p_1} \\ \hline -\frac{1}{\sigma^2}\psi_S^* & \frac{1}{\sigma^2}\psi_1^*\psi_S^* & \mathrm{I}+\frac{1}{\sigma^2}\left(\psi_S^*\right)\left(\psi_S^*\right)^\intercal & \mathbf{0}_{k_*\times p_1} \\ \hline \mathbf{0}_{p_1\times 1} & \mathbf{0}_{p_1\times 1} & \mathbf{0}_{p_1\times k_*} & \mathrm{I}_{p_1\times p_1} \end{array}\right),$$

and

$$(\Sigma_0^z)^{-1}\left(\Sigma_{\boldsymbol{\delta}}^z - \Sigma_0^z\right) = \left(\begin{array}{c|c|c|c} 0 & 0 & \mathbf{0}_{1\times k_*} & \frac{1}{\sigma^2}\left(\rho_0-\psi_1^*\right)\boldsymbol{\delta}^\intercal \\ \hline 0 & 0 & \mathbf{0}_{1\times k_*} & \frac{1}{\sigma^2}\left(\sigma^2+\left(\psi_1^*\right)^2-\rho_0\psi_1^*\right)\boldsymbol{\delta}^\intercal \\ \hline \mathbf{0}_{k_*\times 1} & \mathbf{0}_{k_*\times 1} & \mathbf{0}_{k_*\times k_*} & \frac{1}{\sigma^2}\left(\left(-\rho_0+\psi_1^*\right)\psi_S^*\right)\boldsymbol{\delta}^\intercal \\ \hline \rho_0\boldsymbol{\delta} & \boldsymbol{\delta} & \mathbf{0}_{p_1\times k_*} & \mathbf{0}_{p_1\times p_1} \end{array}\right),$$

and

$$\begin{aligned}&(\Sigma_0^z)^{-1}\left(\Sigma_{\boldsymbol{\delta}}^z - \Sigma_0^z\right)(\Sigma_0^z)^{-1}\left(\Sigma_{\widetilde{\boldsymbol{\delta}}}^z - \Sigma_0^z\right)\\ &= \left(\begin{array}{c|c|c|c} \frac{1}{\sigma^2}\rho_0\left(\rho_0-\psi_1^*\right)\boldsymbol{\delta}^\intercal\widetilde{\boldsymbol{\delta}} & \frac{1}{\sigma^2}\left(\rho_0-\psi_1^*\right)\boldsymbol{\delta}^\intercal\widetilde{\boldsymbol{\delta}} & \mathbf{0}_{1\times k_*} & \mathbf{0}_{1\times p_1} \\ \hline \frac{1}{\sigma^2}\rho_0 f_1\boldsymbol{\delta}^\intercal\widetilde{\boldsymbol{\delta}} & \frac{1}{\sigma^2}f_1\boldsymbol{\delta}^\intercal\widetilde{\boldsymbol{\delta}} & \mathbf{0}_{1\times k_*} & \mathbf{0}_{1\times p_1} \\ \hline \frac{1}{\sigma^2}\rho_0 f_2\boldsymbol{\delta}^\intercal\widetilde{\boldsymbol{\delta}} & \frac{1}{\sigma^2}f_2\boldsymbol{\delta}^\intercal\widetilde{\boldsymbol{\delta}} & \mathbf{0}_{k_*\times k_*} & \mathbf{0}_{k_*\times p_1} \\ \hline \mathbf{0}_{p_1\times 1} & \mathbf{0}_{p_1\times 1} & \mathbf{0}_{p_1\times k_*} & \frac{1}{\sigma^2}\left(\rho_0\left(\rho_0-\psi_1^*\right)+f_1\right)\boldsymbol{\delta}\widetilde{\boldsymbol{\delta}}^\intercal \end{array}\right),\end{aligned}$$

where $f_1 = \left(\sigma^2+\left(\psi_1^*\right)^2-\rho_0\psi_1^*\right)$ and $f_2 = \left(-\rho_0+\psi_1^*\right)\psi_S^*$. Hence, the matrix

$$(\Sigma_0^z)^{-1}\left(\Sigma_{\boldsymbol{\delta}}^z - \Sigma_0^z\right)(\Sigma_0^z)^{-1}\left(\Sigma_{\widetilde{\boldsymbol{\delta}}}^z - \Sigma_0^z\right)$$

is of two equal eigenvalues $\frac{1}{\sigma^2}\left(\rho_0\left(\rho_0-\psi_1^*\right)+f_1\right)\boldsymbol{\delta}^\intercal\widetilde{\boldsymbol{\delta}}$. By Lemma 14, we establish $\chi^2\left(f_{\pi_{\mathcal{H}_1}}, f_{\pi_{\mathcal{H}_0}}\right)+1=\mathbb{E}_{\boldsymbol{\delta},\widetilde{\boldsymbol{\delta}}}\left(1-\frac{1}{\sigma^2}\left(\rho_0\left(\rho_0-\psi_1^*\right)+f_1\right)\boldsymbol{\delta}^\intercal\widetilde{\boldsymbol{\delta}}\right)^{-n}$.

Since the proof of Lemma 9 is similar to that of Lemma 2, we prove Lemma 9 here.

Proof of Lemma 9. Let g_0 denote the density function of $N(0,\Sigma_0^z)$, g_1 denote the density function of $N(0,\Sigma_{\boldsymbol{\delta}}^z)$ and g_2 denote the density function of $N(0,\Sigma_{\widetilde{\boldsymbol{\delta}}}^z)$. By plugging into Lemma 13, we have

$$\int\frac{g_1g_2}{g_0}=\left(\det\left(I-\left(\Sigma_0^z\right)^{-1}\left(\Sigma_{\boldsymbol{\delta}}^z-\Sigma_0^z\right)\left(\Sigma_0^z\right)^{-1}\left(\Sigma_{\widetilde{\boldsymbol{\delta}}}^z-\Sigma_0^z\right)\right)\right)^{-\frac{1}{2}}. \tag{A.2.10}$$

Note that

$$\left(\Sigma_0^z\right)^{-1}=\left(\begin{array}{c|c|c}\frac{1}{\sigma^2} & \left(-\frac{1}{\sigma^2}\beta_{\mathrm{S}_2}^*\right)^\intercal & \mathbf{0}_{1\times p_2}\\ \hline -\frac{1}{\sigma^2}\beta_{\mathrm{S}_2}^* & \mathrm{I}_{k^*\times k^*}+\frac{1}{\sigma^2}\beta_{S_2}^*\left(\beta_{S_2}^*\right)^\intercal & \mathbf{0}_{k^*\times p_2}\\ \hline \mathbf{0}_{p_2\times 1} & \mathbf{0}_{p_2\times k^*} & \mathrm{I}_{p_2\times p_2}\end{array}\right),$$

and

$$\left(\Sigma_0^z\right)^{-1}\left(\Sigma_{\boldsymbol{\delta}}^z-\Sigma_0^z\right)=\left(\begin{array}{c|c|c}0 & \mathbf{0}_{1\times k^*} & \frac{1}{\sigma^2}\boldsymbol{\delta}^\intercal\\ \hline \mathbf{0}_{k^*\times 1} & \mathbf{0}_{k^*\times k^*} & -\frac{1}{\sigma^2}\beta_{S_2}^*\boldsymbol{\delta}^\intercal\\ \hline \boldsymbol{\delta} & \mathbf{0}_{p_2\times k^*} & \mathbf{0}_{p_2\times p_2}\end{array}\right),$$

Hence, we have

$$\left(\Sigma_0^z\right)^{-1}\left(\Sigma_{\boldsymbol{\delta}}^z-\Sigma_0^z\right)\left(\Sigma_0^z\right)^{-1}\left(\Sigma_{\widetilde{\boldsymbol{\delta}}}^z-\Sigma_0^z\right)=\left(\begin{array}{c|c|c}\frac{1}{\sigma^2}\boldsymbol{\delta}^\intercal\widetilde{\boldsymbol{\delta}} & \mathbf{0}_{1\times k^*} & \mathbf{0}_{1\times p_2}\\ \hline -\frac{1}{\sigma^2}\beta_{S_2}^*\boldsymbol{\delta}^\intercal\widetilde{\boldsymbol{\delta}} & \mathbf{0}_{k^*\times k^*} & \mathbf{0}_{k^*\times p_2}\\ \hline \mathbf{0}_{p_2\times 1} & \mathbf{0}_{p_2\times k^*} & \frac{1}{\sigma^2}\boldsymbol{\delta}\widetilde{\boldsymbol{\delta}}^\intercal\end{array}\right).$$

The matrix $\left(\Sigma_0^z\right)^{-1}\left(\Sigma_{\boldsymbol{\delta}}^z-\Sigma_0^z\right)\left(\Sigma_0^z\right)^{-1}\left(\Sigma_{\widetilde{\boldsymbol{\delta}}}^z-\Sigma_0^z\right)$ is of two equal eigenvalues $\frac{1}{\sigma^2}\boldsymbol{\delta}^\intercal\widetilde{\boldsymbol{\delta}}$. By Lemma 14, we establish $\chi^2\left(f_{\pi_{\mathcal{H}_1}}, f_{\pi_{\mathcal{H}_0}}\right)+1=\mathbb{E}_{\boldsymbol{\delta},\widetilde{\boldsymbol{\delta}}}\left(1-\frac{1}{\sigma^2}\boldsymbol{\delta}^\intercal\widetilde{\boldsymbol{\delta}}\right)^{-n}$. Lemma 3 is shown in the proofs of Cai & Low (2005). We prove it here to be self-contained.

Proof of Lemma 3. By the fact that $\mathbb{P}(J=j) \le \binom{k}{j}\left(\frac{k}{p}\right)^j\left(1-\frac{k}{p}\right)^{k-j}\left(1-\frac{k}{p}\right)^{-k}$ and $\left(1-\frac{k}{p}\right)^{-k} \le e^{\frac{k^2}{p-k}}$, we have $\mathbb{P}(J=j) \le e^{\frac{k^2}{p-k}}\left(\binom{k}{j}\left(\frac{k}{p}\right)^j\left(1-\frac{k}{p}\right)^{k-j}\right)$, where $\binom{k}{j}\left(\frac{k}{p}\right)^j\left(1-\frac{k}{p}\right)^{k-j}$ is the pdf of binomial distribution with $\left(k,\frac{k}{p}\right)$. Hence $\mathbb{E}\exp(tJ) \le e^{\frac{k^2}{p-k}}\left(1-\frac{k}{p}+\frac{k}{p}\exp(t)\right)^k$.

A.2.3 Proof of Lemma 4

In the following, we prove the three inequalities (2.7.32), (2.7.33) and (2.7.34) in Chapter 2 .

Proof of (2.7.32)**.** By Lemma 26, we have

$$\mathbb{P}_\theta\left(\left|\frac{\|X_{\cdot j}\|_2^2}{n\Sigma_{jj}}-1\right| \ge 2\sqrt{\frac{t}{n}}+2\frac{t}{n}\right) \le 2\exp(-t),$$

By taking $t=C_1\log p$ with $C_1=2.25$ and the inequality

$$\frac{2}{5}<\sqrt{\left(1-2\sqrt{\frac{C_1\log p}{n}}-2\frac{C_1\log p}{n}\right)}<\sqrt{\left(1+2\sqrt{\frac{C_1\log p}{n}}+2\frac{C_1\log p}{n}\right)}<\frac{7}{5},$$

the union bound leads to $\mathbb{P}_\theta(G_1^c) \le 2p^{1-C_1}$. By Lemma 26, we have

$$\mathbb{P}_\theta\left(\left|\frac{(\sigma^{ora})^2}{\sigma^2}-1\right| \ge 2\sqrt{\frac{t}{n}}+2\frac{t}{n}\right) \le 2\exp(-t),$$

$$\mathbb{P}_\theta\left(\left|\frac{\xi^\intercal\widehat{\Sigma}\xi}{\xi^\intercal\Sigma\xi}-1\right| \ge 2\sqrt{\frac{t}{n}}+2\frac{t}{n}\right) \le 2\exp(-t),$$

$$\mathbb{P}_\theta\left(\left|\frac{u^\intercal\widehat{\Sigma}u}{\xi^\intercal\Omega\xi}-1\right| \ge 2\sqrt{\frac{t}{n}}+2\frac{t}{n}\right) \le 2\exp(-t).$$

Taking $t = \log p$, we have $\mathbb{P}_\theta\left(G_2^c\right) \leq \frac{2}{p}$ and $\mathbb{P}_\theta(G_3^c) \leq \frac{4}{p}$. By Lemma 12, for $\|v_{S^c}\|_1 \leq \alpha_0 \|v_S\|_1$, with probability larger than $1 - c' \exp\left(-cn\right)$, we have

$$\frac{\|Xv\|_2}{\sqrt{n}\|v_S\|_2} \geq \frac{1}{4\sqrt{\lambda_{\max}\left(\Omega\right)}} - 9\frac{1}{\sqrt{\lambda_{\min}\left(\Omega\right)}}\left(1+\alpha_0\right)\sqrt{k\frac{\log p}{n}}.$$

By the definition of $\kappa(X, k, \alpha_0)$,

$$\kappa(X, k, \alpha_0) \geq \frac{1}{4\sqrt{\lambda_{\max}\left(\Omega\right)}} - 9\frac{1}{\sqrt{\lambda_{\min}\left(\Omega\right)}}\left(1+\alpha_0\right)\sqrt{k\frac{\log p}{n}},$$

with probability larger than $1 - c' \exp\left(-cn\right)$. Since $\|W_i\|_2 = \sqrt{n}$, by the union bound of p standard gaussian random variable, we have

$$\mathbb{P}_{\epsilon|X}\left(\frac{\|W^\intercal \epsilon\|_\infty}{n} > \sigma\sqrt{\frac{2\delta_0 \log p}{n}}\middle| X\right) \leq \frac{1}{2\sqrt{\pi\delta_0 \log p}} p^{1-\delta_0},$$

and

$$\mathbb{P}_\theta\left(G_5^c\right) = \mathbb{P}_\theta\left(\frac{\|W^\intercal \epsilon\|_\infty}{n} > \sigma\sqrt{\frac{2\delta_0 \log p}{n}}\right) \leq \frac{1}{2\sqrt{\pi\delta_0 \log p}} p^{1-\delta_0}.$$

The union bound will lead to (2.7.32) in Chapter 2.

The high probability statement (2.7.33) in Chapter 2 is a generalization of Lemma 6.2 in Javanmard & Montanari (2014a). Before the proof, we introduce the following definitions. The sub-gaussian norm of a random variable U is defined as $\|U\|_{\psi_2} = \sup_{q\geq 1} \frac{1}{\sqrt{q}} \left(\mathbb{E}|U|^q\right)^{\frac{1}{q}}$, and the sub-gaussian norm of a random vector $U \in \mathbb{R}^p$ is defined as $\|U\|_{\psi_2} = \sup_{v\in S^{p-1}} \|\langle v, U\rangle\|_{\psi_2}$, where S^{p-1} is the unit sphere in $\mathbb{R}^p$. The sub-exponential norm of a random variable U is defined as $\|U\|_{\psi_1} = \sup_{q\geq 1} \frac{1}{q} \left(\mathbb{E}|U|^q\right)^{\frac{1}{q}}$, and the sub-exponential norm of a random vector $U \in \mathbb{R}^p$ is defined as $\|U\|_{\psi_1} = \sup_{v\in S^{p-1}} \|\langle v, U\rangle\|_{\psi_1}$.

Proof of (2.7.33)**.** It is sufficient to show that with probability larger than $1 -$

$2p^{1-c_0C_0^2}$,

$$\|\xi^\intercal\Omega\widehat{\Sigma}-\xi^\intercal\|_\infty \le 2C_0 e\|\xi^\intercal\Omega\|_2\|X_{i\cdot}\|_{\psi_2}^2\sqrt{\frac{\log p}{n}}, \tag{A.2.11}$$

where $\|X_{i\cdot}\|_{\psi_2}^2 \le \frac{2}{e}M_1$. We define $d^\intercal = \xi^\intercal\Omega\hat{\Sigma}-\xi^\intercal$ and $d_j = \frac{1}{n}\sum_{i=1}^n \left(\xi^\intercal\Omega X_{i\cdot}^\intercal\right)(X_{ij})-\xi_j$. We define that $q_{ij} = \left(\xi^\intercal\Omega X_{i\cdot}^\intercal\right)(X_{ij})$ and we have the following properties of q_{ij}:

1. $\mathbb{E}q_{ij} = \xi_j$ and q_{ij} is sub-exponential random variable with

$$\|q_{ij}\|_{\psi_1} \le 2\|\xi^\intercal\Omega\|_2\|X_{i\cdot}\|_{\psi_2}^2. \tag{A.2.12}$$

2. $q_{ij}-v_j$ is sub-exponential random variable with

$$\|q_{ij}-\xi_j\|_{\psi_1} \le 2\|q_{ij}\|_{\psi_1} \le 4\|\xi^\intercal\Omega\|_2\|X_{i\cdot}\|_{\psi_2}^2,$$

where the first inequality follows from Remark 5.18 in Vershynin (2012) and Javanmard & Montanari (2014a) and the second inequality follows from (A.2.12).

To show (A.2.12), we have

$$\begin{aligned}
&\|q_{ij}\|_{\psi_1} = \sup_{p\ge1}\frac{1}{p}\left(\mathbb{E}|q_{ij}|^p\right)^{\frac{1}{p}} \le \sup_{p\ge1}\frac{1}{p}\left(\mathbb{E}|\xi^\intercal\Omega X_{i\cdot}^\intercal|^{2p}E|X_{ij}|^{2p}\right)^{\frac{1}{2p}}\\
&\le 2\sup_{p\ge1}\frac{1}{\sqrt{2p}}\left(\mathbb{E}|\xi^\intercal\Omega X_{i\cdot}|^{2p}\right)^{\frac{1}{2p}}\sup_{p\ge1}\frac{1}{\sqrt{2p}}\left(\mathbb{E}|X_{ij}|^{2p}\right)^{\frac{1}{2p}}\\
&\le \|\xi^\intercal\Omega\|_2 2\sup_{p\ge1}\frac{1}{\sqrt{2p}}\left(\mathbb{E}|\frac{\xi^\intercal\Omega}{\|\xi^\intercal\Omega\|_2}X_{i\cdot}^\intercal|^{2p}\right)^{\frac{1}{2p}}\sup_{p\ge1}\frac{1}{\sqrt{2p}}\left(\mathbb{E}|e_j^\intercal X_{i\cdot}^\intercal|^{2p}\right)^{\frac{1}{2p}} \le 2\|\xi^\intercal\Omega\|_2\|X_{i\cdot}\|_{\psi_2}^2.
\end{aligned}$$

By the property of sum of independent centered subexponential random variables (Vershynin, 2012; Javanmard & Montanari, 2014a), we have

$$\mathbb{P}_\theta\left(\frac{1}{n}|\sum_{i=1}^n q_{ij}-\xi_j|\ge\epsilon\right) \le 2\exp\left(-c_0 n\min\left(\frac{\epsilon}{K},\frac{\epsilon^2}{K^2}\right)\right), \tag{A.2.13}$$

where $K = 2e\|\xi^\intercal\Omega\|_2\|X_{i\cdot}\|_{\psi_2}^2$ and $c_0 = \frac{1}{6}$. By taking $\epsilon = 2eC_0\|\xi^\intercal\Omega\|_2\|X_{i\cdot}\|_{\psi_2}^2\sqrt{\frac{\log p}{n}} \leq 2e\|\xi^\intercal\Omega\|_2\|X_{i\cdot}\|_{\psi_2}^2$, then $\mathbb{P}_\theta\left(\frac{1}{n}|\sum_{i=1}^n q_{ij} - \xi_j| \geq \epsilon\right) \leq 2\exp\left(-c_0C_0^2\log p\right)$. Hence

$$\mathbb{P}_\theta\left(\max_j |d_j| \geq 2eC_0\|\xi^\intercal\Sigma^{-1}\|_2\|X_{i\cdot}\|_{\psi_2}^2\sqrt{\frac{\log p}{n}}\right) \leq 2p^{1-c_0C_0^2},$$

which leads to (A.2.11).

The proof of (2.7.34) in Chapter 2 relies on the results in Ren et al. (2013).

Proof of (2.7.34). We will control the probability $P(S^c \cap G)$. Define

$$\tau_0 = (1+\epsilon_0)\,\lambda_0 \max\left\{4k\lambda_0, \frac{8\lambda_0 k}{C\,(M_1)}\right\}$$

and $D_1 = \left\{\frac{\|W^\intercal\epsilon\|_\infty}{n} > \sigma^{ora}\lambda_0\frac{\epsilon_0-1}{\epsilon_0+1}(1-\tau_0)\right\}$. By (A.2.5), on the event $G_1 \cap G_4$, if $k \leq c\frac{n}{\log p}$, then $\tau \leq \tau_0$, where τ is defined in (2.7.31). Hence, on the event $G_1 \cap G_4$, we have $S_1^c = \left\{\frac{\|W^\intercal\epsilon\|_\infty}{n} > \sigma^{ora}\lambda_0\frac{\epsilon_0-1}{\epsilon_0+1}(1-\tau)\right\} \subset D_1$. If $k \leq c\frac{n}{\log p}$, by the definition of τ_0 and $\epsilon_0 = \frac{2.01}{\eta_0} + 1$, we have $\frac{\lambda_0\frac{\epsilon_0-1}{\epsilon_0+1}(1-\tau_0)}{\sqrt{\frac{2\delta_0\log p}{n}}} = \frac{2.01+2.01\eta_0}{2.01+2\eta_0}(1-\tau_0) \geq 1$. If $\delta_0\log p > 2$ and $k \leq c\frac{n}{\log p}$, as discussed in inequality (100) in Ren et al. (2013), we have

$$\begin{aligned}
&\mathbb{P}_\theta\left(G_1 \cap G_4 \cap S_1^c\right) \leq P(D_1) = \mathbb{P}_\theta\left(\frac{\|W^\intercal\epsilon\|_\infty}{n\sigma^{ora}} > \lambda_0\frac{\epsilon_0-1}{\epsilon_0+1}(1-\tau_0)\right)\\
\leq&\mathbb{P}_\theta\left(\frac{\|W^\intercal\epsilon\|_\infty}{n\sigma^{ora}} > \sqrt{\frac{2\delta_0\log p}{n}}\right) \leq c\frac{1}{\sqrt{\pi\delta_0\log p}}p^{1-\delta_0}.
\end{aligned}$$

In the following, we control $G \cap S_1 \cap S_2^c$. On the event $G_1 \cap G_4$, if $\sigma > (1+\nu_0)\hat{\sigma}$, we have $\frac{\hat{\sigma}}{\sigma} - 1 < \frac{-\nu_0}{1+\nu_0}$; If $\sigma < (1-\nu_0)\hat{\sigma}$, we have $\frac{\hat{\sigma}}{\sigma} - 1 > \frac{\nu_0}{1-\nu_0}$, and hence $\frac{\nu_0}{1+\nu_0} < \left|\frac{\hat{\sigma}}{\sigma} - 1\right|$. On the event S_1, $\left|\frac{\hat{\sigma}}{\sigma^{ora}} - 1\right| \leq \tau \leq \tau_0$, and hence

$$\left|\frac{\hat{\sigma}}{\sigma} - 1\right| \leq \left|\left(\frac{\hat{\sigma}}{\sigma^{ora}} - 1\right)\frac{\sigma^{ora}}{\sigma}\right| + \left|\frac{\sigma^{ora}}{\sigma} - 1\right| \leq \tau_0\left(\left|\frac{\sigma^{ora}}{\sigma} - 1\right| + 1\right) + \left|\frac{\sigma^{ora}}{\sigma} - 1\right|.$$

By solving the above inequality for $\left|\frac{\sigma^{ora}}{\sigma} - 1\right|$, we have $\left|\frac{\sigma^{ora}}{\sigma} - 1\right| > \frac{\frac{\nu_0}{1+\nu_0} - \tau_0}{1+\tau_0}$, and

$\left|\frac{(\sigma^{ora})^2}{\sigma^2}-1\right| > \frac{\frac{\nu_0}{1+\nu_0}-\tau_0}{1+\tau_0}$. On the event $G_1 \cap G_4 \cap S_1$, if $k \le c\frac{n}{\log p}$, we have $\frac{\frac{\nu_0}{1+\nu_0}-\tau_0}{1+\tau_0} \ge \frac{\frac{\nu_0}{1+\nu_0}}{2\left(1+\frac{\nu_0}{2(1+\nu_0)}\right)} = \frac{\nu_0}{2+3\nu_0}$, and hence

$$S_2^c = \{\sigma > (1+\nu_0)\hat{\sigma}, \sigma < (1-\nu_0)\hat{\sigma}\} \subset D_2 = \left\{\left|\frac{(\sigma^{ora})^2}{\sigma^2}-1\right| > \frac{\nu_0}{2+3\nu_0}\right\},$$

and

$$\mathbb{P}_\theta\left(S_2^c \cap S_1 \cap G\right) \le \mathbb{P}_\theta\left(D_2\right) \le 2\exp\left(-\left(\frac{g_0+1-\sqrt{2g_0+1}}{2}\right)n\right), \tag{A.2.14}$$

where $g_0 = \frac{\nu_0}{2+3\nu_0}$, and the last inequality follows from the concentration of χ^2 distribution. Combining the above inequalities (A.2.3) and (A.2.14), we have

$$\begin{aligned}
&\mathbb{P}_\theta\left(S \cap G\right) = \mathbb{P}_\theta\left(S_1 \cap G\right) - \mathbb{P}_\theta\left(S_2^c \cap S_1 \cap G\right) \\
=&\mathbb{P}_\theta\left(G\right) - \mathbb{P}_\theta\left(S_1^c \cap G\right) - \mathbb{P}_\theta\left(S_2^c \cap S_1 \cap G\right) \\
\ge&\mathbb{P}_\theta\left(G\right) - c\frac{1}{\sqrt{\pi\delta_0\log p}}p^{1-\delta_0} - 2\exp\left(-\left(\frac{g_0+1-\sqrt{2g_0+1}}{2}\right)n\right).
\end{aligned} \tag{A.2.15}$$

A.2.4 Proof of Lemma 5.

By the normalization (2.7.28) in Chapter 2, the scaled Lasso algorithm (2.3.4) in Chapter 2 can be expressed as

$$\{\widehat{d}, \hat{\sigma}\} = \underset{d\in\mathbb{R}^p, \sigma\in\mathbb{R}}{\arg\min} \frac{\|y-Wd\|_2^2}{2n\sigma} + \frac{\sigma}{2} + \lambda_0\sum_{j=1}^{p}|d_j|. \tag{A.2.16}$$

For fixed μ, we also define $\widehat{d}(\mu) = \arg\min_{d\in\mathbb{R}^p} \frac{\|y-Wd\|_2^2}{2n} + \mu\sum_{j=1}^{p}|d_j|$. Note that

$$\widehat{d} = \widehat{d}(\lambda_0\hat{\sigma}) \quad \text{and} \quad \widehat{d}_j = \widehat{\beta}_j\frac{\|X_{\cdot j}\|_2}{\sqrt{n}} \quad \text{for } j \in [p]. \tag{A.2.17}$$

The proof of Lemma 5 depends on the following lemmas. The first two lemmas are Proposition 1 and Proposition 2 in an arXiv version of Ren et al. (2013).

Lemma 16. *For any $\epsilon_0 > 1$, on the event $\left\{\frac{\|W^\intercal \epsilon\|_\infty}{n} \leq \mu \frac{\epsilon_0 - 1}{\epsilon_0 + 1}\right\}$, we have*

$$\|\widehat{d}(\mu) - d\|_1 \leq (2 + 2\epsilon_0) \max\left\{\|d_{T^c}\|_1, \frac{\left(\frac{\|W^\intercal \epsilon\|_\infty}{n} + \mu\right)|T|}{\mathrm{CIF}_1(2\epsilon_0 + 1, \mathrm{T}, \mathrm{W})}\right\}, \tag{A.2.18}$$

where T is defined in (2.7.31).

Lemma 17. *Let $\{\widehat{d}, \hat{\sigma}\}$ be the solution of the scaled Lasso* (A.2.16). *For any $\epsilon_0 > 1$, on the event $S_1 = \left\{\frac{\|W^\intercal \epsilon\|_\infty}{n} \leq \sigma^{ora} \lambda_0 \frac{\epsilon_0 - 1}{\epsilon_0 + 1}(1 - \tau)\right\}$, we have*

$$\left|\frac{\hat{\sigma}}{\sigma^{ora}} - 1\right| \leq \tau. \tag{A.2.19}$$

where τ is defined in (2.7.31).

By Lemma 17, on the event $S_1 = \left\{\frac{\|W^\intercal \epsilon\|_\infty}{n} \leq \sigma^{ora} \lambda_0 \frac{\epsilon_0 - 1}{\epsilon_0 + 1}(1 - \tau)\right\}$, we have

$$\frac{\|W^\intercal \epsilon\|_\infty}{n} \leq \lambda_0 \hat{\sigma} \frac{\epsilon_0 - 1}{\epsilon_0 + 1}.$$

By Lemma 16 with $\mu = \lambda_0 \hat{\sigma}$ and the relation (A.2.17), we have

$$\|\widehat{\beta} - \beta\|_1 \leq (2 + 2\epsilon_0) \frac{\sqrt{n}}{\min \|X_{\cdot j}\|_2} \max\left\{\|d_{T^c}\|_1, \frac{\left(\frac{\|W^\intercal \epsilon\|_\infty}{n} + \lambda_0 \hat{\sigma}\right)|T|}{CIF_1(2\epsilon_0 + 1, T, W)}\right\}.$$

By the definition of the index set T, the event G_4 and $\|\beta\|_0 \leq k$,

$$\|\widehat{\beta} - \beta\|_1 \leq (2 + 2\epsilon_0) \frac{\sqrt{n}}{\min \|X_{\cdot j}\|_2} \max\left\{k \lambda_0 \sigma^{ora}, \frac{\left(\sigma \sqrt{\frac{2\delta_0 \log p}{n}} + \lambda_0 \hat{\sigma}\right) k}{CIF_1(2\epsilon_0 + 1, T, W)}\right\}.$$

By replacing $CIF_1\left(1+2\epsilon_0, T, W\right)$ with its lower bound in (A.2.3), we establish (2.7.35) in Chapter 2.

A.2.5 Proof of Lemma 6

By (A.2.11), on the event B_1, we choose $\lambda_n = 4C_0M^2\|\xi\|_2\sqrt{\frac{\log p}{n}}$ and $\xi^\intercal\Omega$ belongs to the feasible set of (2.3.5) in Chapter 2. Hence, $\widehat{u}^\intercal\widehat{\Sigma}\widehat{u} \leq u^\intercal\widehat{\Sigma}u = \frac{1}{n}\|Xu\|_2^2$, where $Xu \in \mathbb{R}^n$ follows i.i.d Gaussian with variance $u^\intercal\Sigma u = \xi^\intercal\Omega\xi$. On the event G_3, we have

$$u^\intercal\widehat{\Sigma}u \leq \xi^\intercal\Omega\xi\left(1+2\sqrt{\frac{\log p}{n}}+2\frac{\log p}{n}\right) \leq \frac{49}{25}M_1\|\xi\|_2^2. \tag{A.2.20}$$

On the event $G\cap S$,

$$C_2\left(X,k\right)k\sqrt{\frac{\log p}{n}}\hat{\sigma} \leq C_1(M_1)k\sqrt{\frac{\log p}{n}}\sigma,$$

where $C_1(M_1) = 2500\sqrt{M_1}\max\left\{1.25, \frac{2}{C(M_1)}\right\}$. Hence, on the event $G\cap S\cap B_1$, we have $C_1\left(X,k\right)\frac{k\log p}{n}\hat{\sigma} \leq 4C_0M_1^2C_1\left(M_1\right)\frac{k\log p}{n}\sigma$ and $1.01\sqrt{\frac{\widehat{u}^\intercal\widehat{\Sigma}\widehat{u}}{n}}\hat{\sigma}z_{\alpha/2} \leq \frac{8\sqrt{M_1}}{5\sqrt{n}}\|\xi\|_2z_{\alpha/2}\sigma$. There exists a large positive integer p_0 such that if $p\geq p_0$,

$$\log p \geq 1.01\max\left\{4C_0M_1^2C_1\left(M_1\right), \frac{8\sqrt{M_1}}{5}z_{\alpha/2}, C_1(M_1)\right\}.$$

For $p\geq p_0$, we establish the inequalities (2.7.38) and (2.7.37) in Chapter 2.

A.2.6 Proof of Lemma 10

The control of probability of $\bar{G}_1, \bar{G}_2, \bar{G}_3$ and $\bar{G}_4$ follows from the similar argument of Lemma A.2.3. In the following, we will control $\mathbb{P}_\theta\left(\bar{G}_5\right)$. We have the decomposition $\xi^\intercal\left(\mathrm{I}-\widehat{\Sigma}^{(2)}\right)\left(\widehat{\beta}-\beta\right) = \frac{1}{n_2}\sum_{j=1}^{n_2}\delta_j$ with $\delta_j = \xi^\intercal\left(\mathrm{I}-\left(X_j^{(2)}\right)^\intercal X_j^{(2)}\right)\left(\widehat{\beta}-\beta\right)$. Similarly to the proof of (2.7.33), we can show that $\mathbb{E}\delta_j = 0$ and δ_j is sub-exponential random variable with $\|\delta_j\|_{\psi_1} \le 4\|\xi\|_2\|\widehat{\beta} - \beta\|_2\|X_{j\cdot}^{(2)}\|_{\psi_2}^2 \le \frac{8}{e}\|\xi\|_2\|\widehat{\beta} - \beta\|_2 M_1$. By the property of sum of independent centered subexponential random variables (Vershynin, 2012; Javanmard & Montanari, 2014a), we have

$$\mathbb{P}_\theta\left(\frac{1}{n}|\sum_{i=1}^{n}\delta_j| \ge \epsilon\right) \le 2\exp\left(-c_0 n \min\left(\frac{\epsilon}{K}, \frac{\epsilon^2}{K^2}\right)\right),$$

where $K = 8\|\xi\|_2\|\widehat{\beta} - \beta\|_2 M_1$ and $c_0 = \frac{1}{6}$. By taking $\epsilon = 8\sqrt{6\log\frac{2}{(1-\gamma_0)\alpha}}\|\xi\|_2\|\widehat{\beta} - \beta\|_2\frac{1}{\sqrt{n}}M_1 \le K$, we have (A.1.16).

$\mathcal{B}$

Supplement for Chapter 3

B.1 Difference between $\Theta(k)$ and $\Theta_0(k)$

In Chapter 3, we have investigated the minimax estimation rate, minimax expected length and adaptivity of confidence intervals for the loss $\|\widehat{\beta}-\beta\|_q^2$ with $1 \leq q \leq 2$ over the parameter spaces $\Theta_0(k)$ and $\Theta(k)$. It is interesting to compare the minimaxity and adaptivity behaviors between loss estimation over the parameter spaces $\Theta(k)$ and $\Theta_0(k)$. The comparison shows significant differences between estimating the ℓ_2 loss and the ℓ_q loss with $1 \leq q < 2$ as well as the differences between the two parameter spaces $\Theta(k)$ and $\Theta_0(k)$.

In terms of the minimax estimation rate and minimax expected length of confidence intervals, the prior information $\Sigma = \mathrm{I}$ and $\sigma = \sigma_0$ reduces the convergence rate for the ℓ_2 loss $\|\widehat{\beta}-\beta\|_2^2$ from $\frac{k\log p}{n}$ to $\min\left\{\frac{k\log p}{n}, \frac{1}{\sqrt{n}}\right\}$. With this prior information, the adaptive estimation of ℓ_2 loss is made possible over the regime $\frac{\sqrt{n}}{\log p} \ll k \lesssim \frac{n}{\log p}$. In contrast, even with such prior information, the minimax convergence rate remains unchanged for the case $1 \leq q < 2$.

Regarding adaptivity of confidence intervals, the prior knowledge $\Sigma = \mathrm{I}$ and $\sigma = \sigma_0$ is extremely useful for the construction of adaptive confidence intervals for the ℓ_2 loss in the regime $\frac{\sqrt{n}}{\log p} \lesssim k \lesssim \frac{n}{\log p}$. Though adaptivity is still impossible outside this regime, we have seen that, with this prior knowledge, the expected length of optimal confidence intervals over the regime $k_1 \lesssim \frac{\sqrt{n}}{\log p} \lesssim k_2 \lesssim \frac{n}{\log p}$ is reduced from $k_2 \frac{\log p}{n}$ to $\frac{1}{\sqrt{n}}$.

In contrast, for ℓ_q loss with $1 \leq q < 2$, even with this prior information, it is still impossible to construct adaptive confidence intervals for $\|\widehat{\beta} - \beta\|_q^2$ with $1 \leq q < 2$. However, a comparison of Theorem 14 in Chapter 3 with Theorem 15 in Chapter 3 reveals that the expected length of confidence intervals is reduced with such prior knowledge, from $k_2^{\frac{2}{q}} \frac{\log p}{n}$ to $k_2^{\frac{2}{q}-1} \frac{1}{\sqrt{n}}$ in the regime $k_1 \lesssim \frac{\sqrt{n}}{\log p} \lesssim k_2 \lesssim \frac{n}{\log p}$ and from $k_2^{\frac{2}{q}} \frac{\log p}{n}$ to $k_2^{\frac{2}{q}-1} k_1 \frac{\log p}{n}$ in the regime $\frac{\sqrt{n}}{\log p} \lesssim k_1 \leq k_2 \lesssim \frac{n}{\log p}$.

B.2 Minimaxity and adaptivity of confidence intervals for $\|\widehat{\beta} - \beta\|_q^2$ over $\Theta_{\sigma_0}(k, s)$

In Chapter 3, we have shown that there is significant difference between $\Theta_0(k)$ and $\Theta(k)$ in terms of the minimax convergence rates and the adaptivity behaviors. As discussed in Section 3.7 in Cai & Guo (2016a), the parameter space $\Theta_0(k)$ is relatively simple and in this section, we consider a more general parameter space for (β, Σ, σ),

$$\Theta_{\sigma_0}(k, s) = \left\{ (\beta, \Sigma, \sigma_0) : \begin{array}{c} \|\beta\|_0 \leq k, \quad \frac{1}{M_1} \leq \lambda_{\min}(\Sigma) \leq \lambda_{\max}(\Sigma) \leq M_1 \\ \|\Sigma^{-1}\|_{L_1} \leq M, \quad \max_{1 \leq i \leq p} \| \left(\Sigma^{-1}\right)_{i\cdot} \|_0 \leq s \end{array} \right\}, \quad \text{(B.2.1)}$$

for some positive constant $M_1 \geq 1$ and $M > 0$. We will present the lower bound results over $\Theta_{\sigma_0}(k, s)$ in Section B.2.1 and the upper bound results in Section B.2.2.

B.2.1 Minimax lower bounds

In this section, we first establish the minimax lower bounds over the parameter $\Theta_{\sigma_0}(k,s)$.

Theorem 22. *Suppose $0<\alpha,\alpha_0<1/4$, $s \ll \sqrt{n/\log p}$ and the sparsity levels k_1, k_2 and k_0 satisfy Assumption* (B2) *in Chapter 3 with the constant c_0 defined in* (3.9.14) *in Chapter 3. For any estimator $\widehat{\beta}$ satisfying*

$$\sup_{\theta\in\Theta(k_0)} \mathbb{P}_{\theta}\left(\|\widehat{\beta}-\beta^*\|_q^2 \geq C^*\|\beta^*\|_0^{\frac{2}{q}}\frac{\log p}{n}\sigma^2\right) \leq \alpha_0, \tag{B.2.2}$$

with a constant $C^*>0$.

1. *If $k_2 \lesssim \frac{\sqrt{n}}{\log p}$, then there is some constant $c>0$ such that*

$$\mathbf{R}_{\alpha}^{*}\left(\Theta_{\sigma_0}\left(k_1,s\right),\Theta_{\sigma_0}\left(k_2,s\right),\widehat{\beta},\ell_q\right) \geq c k_2^{\frac{2}{q}}\frac{\log p}{n}\sigma_0^2. \tag{B.2.3}$$

2. *If $\frac{\sqrt{n}}{\log p} \lesssim k_2 \lesssim \frac{n}{\log p}$, then there is some constant $c>0$ such that*

$$\begin{aligned}&\mathbf{R}_{\alpha}^{*}\left(\Theta_{\sigma_0}\left(k_1,s\right),\Theta_{\sigma_0}\left(k_2,s\right),\widehat{\beta},\ell_q\right)\\ &\geq c\max\left\{\left((1-c_2)^2 M_1 k_2^{\frac{2}{q}-1}k_1\frac{\log p}{n}-(1+c_1^2)\frac{1}{M_1}k_1^{\frac{2}{q}}\frac{\log p}{n}\right)_+, \frac{k_2^{\frac{2}{q}-1}}{\sqrt{n}}\right\}\sigma_0^2,\end{aligned} \tag{B.2.4}$$

where $c_1=\frac{C^*\sqrt{M_1}k_0^{\frac{1}{q}}}{(k_1-k_0)^{\frac{1}{q}}}$ and $c_2=\frac{C^* k_0^{\frac{1}{q}}}{\sqrt{M_1}(k_2-k_0)^{\frac{1}{q}-\frac{1}{2}}(k_1-k_0)^{\frac{1}{2}}}$.

Consequently, we have

$$\mathbf{R}_{\alpha}^{*}\left(\Theta_{\sigma_0}(k_i,s),\widehat{\beta},\ell_2\right) \geq c\frac{k_i^{\frac{2}{q}}\log p}{n}\sigma_0^2 \quad \textit{for} \quad 1\leq q\leq 2,\ i=1,2. \tag{B.2.5}$$

Remark 10. The minimax lower bound (B.2.5) with $i=1$ follows from (B.2.3) and (B.2.4) by taking $k_2=k_1$. The minimax lower bound (B.2.5) with $i=2$ follows from

(B.2.3) and (B.2.4) by taking $k_1 = \frac{1}{2}k_2$ and the fact that $\mathbf{R}_\alpha^*\left(\Theta_{\sigma_0}(k_2, s), \widehat{\beta}, \ell_2\right) \geq \mathbf{R}_\alpha^*\left(\Theta_{\sigma_0}\left(k_1, s\right), \Theta_{\sigma_0}\left(k_2, s\right), \widehat{\beta}, \ell_q\right)$. Theorem 18 is the special case of the above with $q = 2$.

B.2.2 Minimax upper bounds

In this section, we will focus on the estimator $\widehat{\beta}$ constructed based on the subsample $Z^{(1)} = \left(y^{(1)}, X^{(1)}\right)$ and satisfying Assumption (A) in Chapter 3 with $\delta > 2$ and demonstrate the lower bounds (B.2.3) and (B.2.4) in Theorem 22 can be achieved. The Lasso estimator $\widehat{\beta}^L$ defined in (3.2.10) in Chapter 3 with $A > 4\sqrt{2}$ is an example of such estimators. To simply the notation, we use Ω to denote Σ^{-1}. Let $\widehat{\Omega}$ denote the CLIME estimator (Cai et al., 2011) of Ω and $\lambda_{\max}(\widehat{\Omega})$ and $\lambda_{\min}(\widehat{\Omega})$ denote the maximum and minimum eigenvalue of the estimator $\widehat{\Omega}$.

The constructions of confidence intervals are very similar to the confidence intervals, $\mathrm{CI}_\alpha^1\left(Z\right)$ in (3.3.8), and $\mathrm{CI}_\alpha^2\left(Z, k_2, q\right)$ in (3.3.19) in Chapter 3. The only difference here is that there is no prior knowledge $\Sigma = \mathrm{I}$ and we need to estimate $\Omega = \Sigma^{-1}$ based on the data Z. In the following, we will detail the modification of $\mathrm{CI}_\alpha^1\left(Z\right)$ in (3.3.8) in Chapter 3 and $\mathrm{CI}_\alpha^2\left(Z, k_2, q\right)$ in (3.3.19) in Chapter 3.

We modify the construction of $\mathrm{CI}_\alpha^1\left(Z\right)$ proposed in (3.3.8) in Chapter 3 as follows,

$$\mathrm{CI}_\alpha^3\left(Z\right) = \left(0.99\lambda_{\min} \times \left(\frac{\psi\left(Z\right)}{\frac{1}{n_2}\chi^2_{1-\frac{\alpha}{2}}\left(n_2\right)} - \sigma_0^2\right)_+, 1.01\lambda_{\max} \times \left(\frac{\psi\left(Z\right)}{\frac{1}{n_2}\chi^2_{\frac{\alpha}{2}}\left(n_2\right)} - \sigma_0^2\right)_+\right), \tag{B.2.6}$$

where $\lambda_{\max} = \max\left\{\lambda_{\max}(\widehat{\Omega}), \log p\right\}$ and $\lambda_{\min} = \max\left\{\lambda_{\min}(\widehat{\Omega}), \log p\right\}$ and

$$\psi\left(Z\right) = \min\left\{\frac{1}{n_2}\left\|y^{(2)} - X^{(2)}\widehat{\beta}\right\|_2^2, \sigma_0^2 \log p\right\}. \tag{B.2.7}$$

Before modifying the construction of $\mathrm{CI}_\alpha^2\left(Z, k_2, q\right)$ in (3.3.19) in Chapter 3, we restrict our attention to the set of estimator $\widehat{\beta}$ satisfying the following assumption,

$$\sup_{\theta \in \Theta(k)} \mathbb{P}_\theta \left(\|(\widehat{\beta} - \beta)_{S^c}\|_1 \geq c^* \|(\widehat{\beta} - \beta)_S\|_1 \text{ with } S = \text{supp}(\beta) \right) \leq C p^{-\delta}, \qquad \text{(B.2.8)}$$

for all $k \ll \frac{n}{\log p}$. For $\widehat{\beta}$ satisfying (B.2.8), we modify the construction of $\text{CI}_\alpha^2(Z, k_2, q)$ in (3.3.19) in Chapter 3 as follows,

$$\text{CI}_\alpha^4(Z, k_2, q) = \left(0.99 \lambda_{\min} \left(\frac{\psi(Z)}{\frac{1}{n_2} \chi^2_{1-\frac{\alpha}{2}}(n_2)} - \sigma_0^2 \right)_+, 1,01 \lambda_{\max} \left((1 + c^*)^2 k_2 \right)^{\frac{2}{q}-1} \left(\frac{\psi(Z)}{\frac{1}{n_2} \chi^2_{\frac{\alpha}{2}}(n_2)} - \sigma_0^2 \right)_+ \right). \qquad \text{(B.2.9)}$$

The following proposition shows that the minimax lower bound (B.2.4) can be achieved by the confidence interval $\text{CI}_\alpha^3(Z)$ in (B.2.6) for the case $q = 2$ and $\text{CI}_\alpha^4(Z, k_2, q)$ in (B.2.9) for the case $1 \leq q < 2$.

Proposition 10. *Suppose $p \geq n$, $k_1 \leq k_2 \lesssim \frac{n}{\log p}$ $s \ll \sqrt{\frac{n}{\log p}}$ and $\widehat{\beta}$ is constructed based on the subsample $Z^{(1)} = \left(y^{(1)}, X^{(1)}\right)$ and satisfies Assumption (A) in Chapter 3. Then $\text{CI}_\alpha^3(Z)$ defined in* (B.2.6) *satisfies,*

$$\liminf_{n,p \to \infty} \inf_{\theta \in \Theta_{\sigma_0}(k_2, s)} \mathbb{P}_\theta \left(\|\widehat{\beta} - \beta\|_2^2 \in \text{CI}_\alpha^3(Z) \right) \geq 1 - \alpha, \qquad \text{(B.2.10)}$$

and

$$\mathbf{R}\left(\text{CI}_\alpha^3(Z), \Theta_{\sigma_0}(k_1, s)\right) \lesssim \left(k_1 \frac{\log p}{n} + \frac{1}{\sqrt{n}} \right) \sigma_0^2. \qquad \text{(B.2.11)}$$

In addition, if $\widehat{\beta}$ satisfies Assumption (A) in Chapter 3 and the assumption (B.2.8) *with $\delta > 2$, then $\text{CI}_\alpha^4(Z, k_2, q)$ defined in* (B.2.9) *satisfies,*

$$\liminf_{n,p \to \infty} \inf_{\theta \in \Theta_{\sigma_0}(k_2, s)} \mathbb{P}_\theta \left(\|\widehat{\beta} - \beta\|_q^2 \in \text{CI}_\alpha^4(Z, k_2, q) \right) \geq 1 - \alpha, \qquad \text{(B.2.12)}$$

and

$$\mathbf{R}\left(\text{CI}_\alpha^4(Z, k_2, q), \Theta_{\sigma_0}(k_1, s)\right) \lesssim k_2^{\frac{2}{q}-1} \left(k_1 \frac{\log p}{n} + \frac{1}{\sqrt{n}} \right) \sigma_0^2. \qquad \text{(B.2.13)}$$

Similar to the construction in (B.2.9) and (B.2.6), we can modify the construction of confidence interval $\mathrm{CI}_{\alpha}^{0}\left(Z,k,q\right)$ in (3.3.15) defined in Chapter 3 by replacing the known minimum and maximum eigenvalues with the corresponding estimates $\lambda_{\min}(\widehat{\Omega})$ and $\lambda_{\max}(\widehat{\Omega})$. The minimax lower bounds (B.2.5) and (B.2.3) can be achieved by such construction of confidence intervals.

B.3 Additional lower bound analysis

In this Section, we prove the lower bound results, Theorem 9, Theorem 10, Theorem 11, Theorem 13, Theorem 15, Theorem 17, Theorem 18, Theorem 19 and Theorem 22.

B.3.1 Proof of Theorem 17

Proof of (3.6.6)

By (3.6.4) and (3.9.3) in Chapter 3, we have $\mathbb{P}_{\pi_i}\left(\|\widehat{\beta}\left(Z\right)-\beta^{*}\|_q\leq c_i d_i\right)\geq 1-\alpha_0-\mathrm{TV}\left(f_{\pi_i},f_{\theta_i}\right),$ for $i=1,2$. Then by (3.9.2) in Chapter 3, we have

$$\mathbb{P}_{Z,\theta\sim\pi_i}\left(\|\widehat{\beta}\left(Z\right)-\beta^{*}\|_q\leq c_i d_i\right)\geq 1-\alpha_0-\mathrm{TV}\left(f_{\pi_i},f_{\theta_i}\right),\ \text{for } i=1,2. \tag{B.3.1}$$

Define the following events

$$\mathcal{A}_i=\left\{z:\left(1-c_i\right)d_i\leq\inf_{\theta\in\mathcal{F}_i}\|\widehat{\beta}\left(z\right)-\beta\left(\theta\right)\|_q\leq\sup_{\theta\in\mathcal{F}_i}\|\widehat{\beta}\left(z\right)-\beta\left(\theta\right)\|_q\leq\left(1+c_i\right)d_i\right\},\ \text{for } i=1,2. \tag{B.3.2}$$

If $\|\widehat{\beta}(z)-\beta^{*}\|_q\leq c_i d_i$ and $\theta\in\mathcal{F}_i$ where $i=1,2$, then

$$\|\widehat{\beta}\left(z\right)-\beta\left(\theta\right)\|_q\geq\|\beta\left(\theta\right)-\beta^{*}\|_q-\|\widehat{\beta}(z)-\beta^{*}\|_q\geq\left(1-c_i\right)d_i, \tag{B.3.3}$$

and

$$\|\widehat{\beta}\left(z\right)-\beta\left(\theta\right)\|_q\leq\|\beta\left(\theta\right)-\beta^{*}\|_q+\|\widehat{\beta}(z)-\beta^{*}\|_q\leq\left(1+c_i\right)d_i. \tag{B.3.4}$$

By (B.3.1), (B.3.3) and (B.3.4), we have $\mathbb{P}_{Z,\theta\sim\pi_i}\left(\mathcal{A}_i\right)\geq 1-\alpha_0-\mathrm{TV}\left(f_{\pi_i},f_{\theta_i}\right),$ for $i=1,2$. Applying (3.9.2) in Chapter 3, we obtain

$$\mathbb{P}_{\pi_1}\left(\mathcal{A}_1\right)\geq 1-\alpha_0-\mathrm{TV}\left(f_{\pi_1},f_{\theta_1}\right),\quad\text{and}\quad\mathbb{P}_{\pi_2}\left(\mathcal{A}_2\right)\geq 1-\alpha_0-\mathrm{TV}\left(f_{\pi_2},f_{\theta_2}\right). \tag{B.3.5}$$

By the second inequality of (B.3.5) and (3.9.3) in Chapter 3, we have $\mathbb{P}_{\pi_1}\left(\mathcal{A}_2\right)\geq 1-\alpha_0-\mathrm{TV}\left(f_{\pi_2},f_{\theta_2}\right)-\mathrm{TV}\left(f_{\pi_2},f_{\pi_1}\right).$ Combined with the first inequality of (B.3.5), we further have

$$\mathbb{P}_{\pi_1}\left(\mathcal{A}_1\cap\mathcal{A}_2\right)\geq 1-2\alpha_0-\sum_{i=1}^{2}\mathrm{TV}\left(f_{\pi_i},f_{\theta_i}\right)-\mathrm{TV}\left(f_{\pi_2},f_{\pi_1}\right). \tag{B.3.6}$$

By the coverage property, we have

$$\inf_{\theta\in\mathcal{F}_1\cup\mathcal{F}_2}\mathbb{P}_{\theta}\left(\|(\widehat{\beta}\left(Z\right)-\beta\left(\theta\right)\|_q^2\in\mathrm{CI}_{\alpha}\left(\widehat{\beta},\ell_q,Z\right)\right)\geq 1-\alpha. \tag{B.3.7}$$

Define the following event indexed with θ, $\mathcal{B}_{\theta}=\left\{z:\|(\widehat{\beta}\left(z\right)-\beta\left(\theta\right)\|_q^2\in\mathrm{CI}_{\alpha}\left(\widehat{\beta},\ell_q,z\right)\right\}.$ Define $\mathcal{M}_1=\cup_{\theta\in\mathcal{F}_1}\mathcal{B}_{\theta}$ and $\mathcal{M}_2=\cup_{\theta\in\mathcal{F}_2}\mathcal{B}_{\theta}$. We then obtain

$$\mathbb{P}_{Z,\theta\sim\pi_1}\left(\mathcal{M}_1\right)=\int\left(\int\mathbf{1}_{\mathcal{M}_1}f_{\theta}\left(z\right)dz\right)\pi_1\left(\theta\right)d\theta\geq\int\left(\int\mathbf{1}_{\mathcal{B}_{\theta}}f_{\theta}\left(z\right)dz\right)\pi_1\left(\theta\right)d\theta\geq 1-\alpha,$$

where the last inequality follows from (B.3.7). Similarly, we can establish $\mathbb{P}_{Z,\theta\sim\pi_2}\left(\mathcal{M}_2\right)\geq 1-\alpha$. By (3.9.2) in Chapter 3, we further have

$$\mathbb{P}_{\pi_1}\left(\mathcal{M}_1\right)=\mathbb{P}_{Z,\theta\sim\pi_1}\left(\mathcal{M}_1\right)\geq 1-\alpha\quad\text{and}\quad\mathbb{P}_{\pi_2}\left(\mathcal{M}_2\right)=\mathbb{P}_{Z,\theta\sim\pi_2}\left(\mathcal{M}_2\right)\geq 1-\alpha. \tag{B.3.8}$$

By the second inequality of (B.3.8) with (3.9.3) in Chapter 3, we have $\mathbb{P}_{\pi_1}(\mathcal{M}_2) \geq 1 - \alpha - \mathrm{TV}(f_{\pi_2}, f_{\pi_1})$. Combined with the first inequality of (B.3.8), we have

$$\mathbb{P}_{\pi_1}(\mathcal{M}_1 \cap \mathcal{M}_2) \geq 1 - 2\alpha - \mathrm{TV}(f_{\pi_2}, f_{\pi_1}). \tag{B.3.9}$$

Combining (B.3.6) and (B.3.9), we obtain

$$\mathbb{P}_{\pi_1}(\mathcal{A}_1 \cap \mathcal{A}_2 \cap \mathcal{M}_1 \cap \mathcal{M}_2) \geq 1 - 2\alpha_0 - 2\alpha - \sum_{i=1}^{2} \mathrm{TV}(f_{\pi_i}, f_{\theta_i}) - 2\mathrm{TV}(f_{\pi_2}, f_{\pi_1}). \tag{B.3.10}$$

For $z \in \mathcal{M}_1 \cap \mathcal{M}_2$, there exists $\bar{\theta}_1 \in \mathcal{F}_1$ and $\bar{\theta}_2 \in \mathcal{F}_2$ such that

$$\|\widehat{\beta}(z) - \beta(\bar{\theta}_1)\|_q^2 \in \mathrm{CI}_\alpha\left(\widehat{\beta}, \ell_q, z\right) \text{ and } \|\widehat{\beta}(z) - \beta(\bar{\theta}_2)\|_q^2 \in \mathrm{CI}_\alpha\left(\widehat{\beta}, \ell_q, z\right). \tag{B.3.11}$$

Since $z \in \mathcal{A}_1 \cap \mathcal{A}_2$, we have

$$(1-c_1)d_1 \leq \|\widehat{\beta}(z) - \beta(\bar{\theta}_1)\|_q \leq (1+c_1)d_1, \quad \text{and} \quad (1-c_2)d_2 \leq \|\widehat{\beta}(z) - \beta(\bar{\theta}_2)\|_q \leq (1+c_2)d_2. \tag{B.3.12}$$

For $z \in \mathcal{A}_1 \cap \mathcal{A}_2 \cap \mathcal{M}_1 \cap \mathcal{M}_2$, (B.3.11) and (B.3.12) lead to

$$\mathbf{R}\left(\mathrm{CI}_\alpha\left(\widehat{\beta}, \ell_q, z\right)\right) \geq (1-c_2)^2 d_2^2 - (1+c_1)^2 d_1^2.$$

Combined with (B.3.10), we establish

$$\mathbb{E}_{\pi_1}\mathbf{R}\left(\mathrm{CI}_\alpha\left(\widehat{\beta}, \ell_q, Z\right)\right) \geq c_4^*\left((1-c_2)^2 d_2^2 - (1+c_1)^2 d_1^2\right).$$

Since the maximum risk is lower bounded by the Bayesian risk, we establish (3.6.6) in Chapter 3.

Proof of (3.6.5)

The proof of (3.6.5) in Chapter 3 combines the proof ideas of (3.6.2) and (3.6.6) in

Chapter 3. Assume that

$$\sup_{\theta \in \mathcal{F}_1} \mathbb{P}_\theta \left(\left| \widehat{L}_q(Z) - \|\widehat{\beta}(Z) - \beta(\theta)\|_q^2 \right| \geq \frac{1}{4} d_2^2 \right) \leq \alpha_1, \text{ with } \alpha_1 = \frac{1}{10}. \tag{B.3.13}$$

Otherwise, we can establish (3.6.5) in Chapter 3 by having

$$\sup_{\theta \in \mathcal{F}_1} \mathbb{P}_\theta \left(\left| \widehat{L}_q(Z) - \|\widehat{\beta}(Z) - \beta(\theta)\|_q^2 \right| \geq \frac{1}{4} d_2^2 \right) \geq \alpha_1. \tag{B.3.14}$$

By (B.3.13), we have $\sup_{\theta \in \mathcal{F}_1} \mathbb{P}_\theta \left(\min_{\theta \in \mathcal{F}_1} \left| \widehat{L}_q(Z) - \|\widehat{\beta}(Z) - \beta(\theta)\|_q^2 \right| \geq \frac{1}{4} d_2^2 \right) \leq \alpha_1$ and hence

$$\mathbb{P}_{\pi_1} \left(\min_{\theta \in \mathcal{F}_1} \left| \widehat{L}_q(Z) - \|\widehat{\beta}(Z) - \beta(\theta)\|_q^2 \right| \leq \frac{1}{4} d_2^2 \right) \geq 1 - \alpha_1.$$

Define $\mathcal{M}_0 = \left\{ z : \min_{\theta \in \mathcal{F}_1} \left| \widehat{L}_q(z) - \|\widehat{\beta}(z) - \beta(\theta)\|_q^2 \right| \leq \frac{1}{4} d_2^2 \right\}$. By (3.9.3) in Chapter 3, we obtain

$$\mathbb{P}_{\pi_2} \left(\mathcal{M}_0 \right) \geq 1 - \alpha_1 - \mathrm{TV} \left(f_{\pi_2}, f_{\pi_1} \right). \tag{B.3.15}$$

Define $\mathcal{A}_i$ with $i = 1, 2$ as in (B.3.2). Similar to (B.3.6), we can establish the following control of probability

$$\mathbb{P}_{\pi_2} \left(\mathcal{A}_1 \cap \mathcal{A}_2 \right) \geq 1 - 2\alpha_0 - \sum_{i=1}^{2} \mathrm{TV} \left(f_{\pi_i}, f_{\theta_i} \right) - \mathrm{TV} \left(f_{\pi_2}, f_{\pi_1} \right). \tag{B.3.16}$$

By combining (B.3.16) and (B.3.15), we establish that

$$\mathbb{P}_{\pi_2} \left(\mathcal{A}_1 \cap \mathcal{A}_2 \cap \mathcal{M}_0 \right) \geq 1 - \alpha_1 - 2\alpha_0 - \sum_{i=1}^{2} \mathrm{TV} \left(f_{\pi_i}, f_{\theta_i} \right) - 2\mathrm{TV} \left(f_{\pi_2}, f_{\pi_1} \right). \tag{B.3.17}$$

For $z \in \mathcal{M}_0$, there exist $\bar{\theta} \in \mathcal{F}_1$ such that

$$\left| \widehat{L}_q \left(z \right) - \|\widehat{\beta}(z) - \beta(\bar{\theta})\|_q^2 \right| \leq \frac{1}{4} d_2^2. \tag{B.3.18}$$

For $z \in \mathcal{A}_1 \cap \mathcal{A}_2$, we have the following results for $\bar{\theta}$ and any $\theta \in \mathcal{F}_2$,

$$(1-c_1)d_1 \leq \|\widehat{\beta}(z)-\beta(\bar{\theta})\|_q \leq (1+c_1)d_1, \quad \text{and} \quad (1-c_2)d_2 \leq \|\widehat{\beta}\,(z)-\beta\,(\theta)\,\|_q \leq (1+c_2)d_2. \tag{B.3.19}$$

Hence, for $z \in \mathcal{A}_1 \cap \mathcal{A}_2 \cap \mathcal{M}_0$ and $\theta \in \mathcal{F}_2$, we have

$$\begin{aligned}
&\left|\widehat{L}_q\,(z) - \|\widehat{\beta}(z) - \beta(\theta)\|_q^2\right| \\
\geq &\left|\|\widehat{\beta}(z) - \beta(\theta)\|_q^2 - \|\widehat{\beta}(z) - \beta(\bar{\theta})\|_q^2\right| - \left|\widehat{L}_q\,(z) - \|\widehat{\beta}(z) - \beta(\bar{\theta})\|_q^2\right| \\
\geq &(1-c_2)^2 d_2^2 - (1+c_1)^2 d_1^2 - \frac{1}{4}d_2^2,
\end{aligned} \tag{B.3.20}$$

where the last inequality follows from (B.3.18) and (B.3.19). That is, for $z \in \mathcal{A}_1 \cap \mathcal{A}_2 \cap \mathcal{M}_0$, we obtain

$$\min_{\theta \in \mathcal{F}_2} \left|\widehat{L}_q\,(z) - \|\widehat{\beta}(z) - \beta(\theta)\|_q^2\right| \geq ((1-c_2)^2 - \frac{1}{4})d_2^2 - (1+c_1)^2 d_1^2. \tag{B.3.21}$$

Note that

$$\begin{aligned}
&\sup_{\theta \in \mathcal{F}_2} \mathbb{P}_\theta \left(\left|\widehat{L}_q\,(Z) - \|\widehat{\beta}(Z) - \beta(\theta)\|_q^2\right| \geq ((1-c_2)^2 - \frac{1}{4})d_2^2 - (1+c_1)^2 d_1^2\right) \\
&\geq \sup_{\theta \in \mathcal{F}_2} \mathbb{P}_\theta \left(\min_{\theta \in \mathcal{F}_2} \left|\widehat{L}_q\,(Z) - \|\widehat{\beta}(Z) - \beta(\theta)\|_q^2\right| \geq ((1-c_2)^2 - \frac{1}{4})d_2^2 - (1+c_1)^2 d_1^2\right).
\end{aligned}$$

Since the max risk is lower bounded by the Bayesian risk, the final term of the last inequality can be further bounded by

$$\begin{aligned}
&\mathbb{P}_{\pi_2} \left(\min_{\theta \in \mathcal{F}_2} \left|\widehat{L}_q\,(Z) - \|\widehat{\beta}(Z) - \beta(\theta)\|_q^2\right| \geq ((1-c_2)^2 - \frac{1}{4})d_2^2 - (1+c_1)^2 d_1^2\right) \\
&\geq \mathbb{P}_{\pi_2}\left(\mathcal{A}_1 \cap \mathcal{A}_2 \cap \mathcal{M}_0\right) \geq 1 - \alpha_1 - 2\alpha_0 - \sum_{i=1}^{2} \mathrm{TV}\left(f_{\pi_i}, f_{\theta_i}\right) - 2\mathrm{TV}\left(f_{\pi_2}, f_{\pi_1}\right),
\end{aligned} \tag{B.3.22}$$

where the inequality follows from (B.3.21) and (B.3.17). Combing (B.3.14) and (B.3.22), we establish (3.6.5) in Chapter 3.

B.3.2 Proof of Theorem 19 and Theorem 22

To establish the theorems, we need to establish the following three lower bounds,

$$\mathbf{R}_{\alpha}^{*}\left(\Theta_0\left(k_1\right), \Theta_0\left(k_2\right), \widehat{\beta}, \ell_q\right) \geq c k_2^{\frac{2}{q}} \frac{\log p}{n} \sigma_0^2, \quad \text{for} \quad k_2 \lesssim \frac{\sqrt{n}}{\log p} \tag{B.3.23}$$

and for $\frac{\sqrt{n}}{\log p} \lesssim k_2 \lesssim \frac{n}{\log p}$,

$$\mathbf{R}_{\alpha}^{*}\left(\Theta_0\left(k_1\right), \Theta_0\left(k_2\right), \widehat{\beta}, \ell_q\right) \geq c k_2^{\frac{2}{q}-1} \frac{1}{\sqrt{n}} \sigma_0^2, \tag{B.3.24}$$

$$\begin{aligned} &\mathbf{R}_{\alpha}^{*}\left(\Theta_{\sigma_0}\left(k_1, s\right), \Theta_{\sigma_0}\left(k_2, s\right), \widehat{\beta}, \ell_q\right) \\ &\geq c\left(\left(1-c_2\right)^2 M_1 (k_2-k_0)^{\frac{2}{q}-1} (k_1-k_0)\rho^2 - \frac{1}{M_1}\left(1+c_1\right)^2\left(k_1-k_0\right)^{\frac{2}{q}}\rho^2\right)_{+}. \end{aligned} \tag{B.3.25}$$

By the fact that $\mathbf{R}_{\alpha}^{*}\left(\Theta_{\sigma_0}\left(k_1, s\right), \Theta_{\sigma_0}\left(k_2, s\right), \widehat{\beta}, \ell_q\right) \geq \mathbf{R}_{\alpha}^{*}\left(\Theta_0\left(k_1\right), \Theta_0\left(k_2\right), \widehat{\beta}, \ell_q\right)$, Theorem 22 follows from (B.3.23), (B.3.24) and (B.3.25). Theorem 19 follows from (B.3.23), (B.3.24) and (B.3.25) with $M_1 = 1$. In the following, we will prove (B.3.23), (B.3.24) and (B.3.25), separately.

Proof of (B.3.23)

The proof of (B.3.23) is an application of (3.6.3) of Theorem 16 in Chapter 3. For $\theta_0 = (\beta^*, \mathrm{I}, \sigma_0) \in \Theta_0\left(k_0\right)$, we construct

$$\mathcal{F} = \left\{(\beta^* + \boldsymbol{\delta}, \mathrm{I}, \sigma_0) \;:\; \boldsymbol{\delta} \in \ell\left(\beta^*, k_2-k_0, \rho\right)\right\} \subset \Theta_0\left(k_2\right), \tag{B.3.26}$$

where

$$\ell\left(\beta^{*}, k_{2}-k_{0}, \rho\right)=\left\{\boldsymbol{\delta}: \operatorname{supp}(\boldsymbol{\delta}) \subset \operatorname{supp}\left(\beta^{*}\right)^{c},\|\boldsymbol{\delta}\|_{0}=k_{2}-k_{0},\ \boldsymbol{\delta}_{i} \in\{0, \rho\}\right\}. \tag{B.3.27}$$

Let $S=\operatorname{supp}(\beta^*)$. Without loss of generality, we assume $S=\{1,2,\cdots,k_0\}$. Let p_1 denote the size of S^c and hence $p_1=p-k_0$. Let π denote the uniform prior on the parameter space $\mathcal{F}$, which is induced by the uniform prior of $\boldsymbol{\delta}$ on $\ell(\beta^*, k_2-k_0, \rho)$. Under the Gaussian random design model, $Z_i=(y_i, X_{i\cdot})\in\mathbb{R}^{p+1}$ follows a joint Gaussian distribution with mean 0. Let Σ^z denote the covariance matrix of Z_i. For the indices of Σ^z, we use 0 as the index of y_i and $\{1,\cdots,p\}$ as the indices for $(X_{i1},\cdots,X_{ip})\in\mathbb{R}^p$. Decompose Σ^z into blocks $\begin{pmatrix}\Sigma^z_{yy} & \left(\Sigma^z_{xy}\right)^\intercal \\ \Sigma^z_{xy} & \Sigma^z_{xx}\end{pmatrix}$, where Σ^z_{yy}, Σ^z_{xx} and Σ^z_{xy} denote the variance of y, the variance of X and the covariance of y and X, respectively. There exists a bijective function $h:\Sigma^z\to(\beta,\Sigma,\sigma)$ and the inverse mapping $h^{-1}:(\beta,\Sigma,\sigma)\to\Sigma^z$, where $h^{-1}\left((\beta,\Sigma,\sigma)\right)=\begin{pmatrix}\beta^\intercal\Sigma\beta+\sigma^2 & \beta^\intercal\Sigma \\ \Sigma\beta & \Sigma\end{pmatrix}$ and

$$h(\Sigma^z)=\left(\left(\Sigma^z_{xx}\right)^{-1}\Sigma^z_{xy}, \Sigma^z_{xx}, \Sigma^z_{yy}-\left(\Sigma^z_{xy}\right)^\intercal\left(\Sigma^z_{xx}\right)^{-1}\Sigma^z_{xy}\right). \tag{B.3.28}$$

Based on the bijection, the control of $\chi^2(f_\pi, f_{\theta_0})$ is reduced to the control of the χ^2 distance between two multivariate Gaussian distributions.

The parameter spaces for Σ^z corresponding to $\{\theta_0\}$ and $\mathcal{F}$ are

$$\mathcal{H}_1=\left\{\Sigma^z_0\right\}, \quad \text{where} \quad \Sigma^z_0=\left(\begin{array}{c|c|c}\|\beta^*\|_2^2+\sigma_0^2 & (\beta^*_S)^\intercal & \mathbf{0}_{1\times p_1} \\ \hline \beta^*_S & \mathrm{I}_{k_0\times k_0} & \mathbf{0}_{k_0\times p_1} \\ \hline \mathbf{0}_{p_1\times 1} & \mathbf{0}_{p_1\times k_0} & \mathrm{I}_{p_1\times p_1}\end{array}\right),$$

and

$$\mathcal{H}_2 = \left\{\Sigma^z_{\boldsymbol{\delta}} : \boldsymbol{\delta} \in \ell\left(\beta^*, k_2 - k_0, \rho\right)\right\}, \text{ where } \Sigma^z_{\boldsymbol{\delta}} = \left(\begin{array}{c|c|c} \|\beta^*\|_2^2 + \|\boldsymbol{\delta}\|_2^2 + \sigma_0^2 & (\beta^*_S)^{\intercal} & \boldsymbol{\delta}^{\intercal} \\ \hline \beta^*_S & \mathrm{I}_{k_0 \times k_0} & \mathbf{0}_{k_0 \times p_1} \\ \hline \boldsymbol{\delta} & \mathbf{0}_{p_1 \times k_0} & \mathrm{I}_{p_1 \times p_1} \end{array}\right).$$

Define $\theta_1 = (\beta^*, \mathrm{I}, \sigma_0^2 + \|\boldsymbol{\delta}\|_2^2)$. For $\boldsymbol{\delta} \in \ell\left(\beta^*, k_2 - k_0, \rho\right)$, we have $\|\boldsymbol{\delta}\|_2^2 = (k_2 - k_0)\rho^2$ and hence $\theta_1 = (\beta^*, \mathrm{I}, \sigma_0^2 + (k_2 - k_0)\rho^2)$. By $\mathrm{TV}\left(f_\pi, f_{\theta_0}\right) \leq \mathrm{TV}\left(f_\pi, f_{\theta_1}\right) + \mathrm{TV}\left(f_{\theta_1}, f_{\theta_0}\right)$, it is sufficient to control $\mathrm{TV}\left(f_\pi, f_{\theta_1}\right)$ and $\mathrm{TV}\left(f_{\theta_1}, f_{\theta_0}\right)$. By (3.9.1) in Chapter 3, it is sufficient to establish $\chi^2\left(f_\pi, f_{\theta_1}\right) \leq \epsilon_1^2$ and $\chi^2\left(f_{\theta_1}, f_{\theta_0}\right) \leq \epsilon_1^2$.

Let $\mathbb{E}_{\boldsymbol{\delta}, \widetilde{\boldsymbol{\delta}}}$ denote the expectation with respect to the independent random variables $\boldsymbol{\delta}, \widetilde{\boldsymbol{\delta}}$ with a uniform prior over the parameter space $\ell\left(\beta^*, k_2 - k_0, \rho\right)$. The following two lemmas are useful to control $\chi^2\left(f_\pi, f_{\theta_1}\right)$ and $\chi^2\left(f_{\theta_1}, f_{\theta_0}\right)$. The proof of Lemma 18 is given in Section B.5.2.

Lemma 18.

$$\chi^2\left(f_\pi, f_{\theta_1}\right) + 1 = \mathbb{E}_{\boldsymbol{\delta}, \widetilde{\boldsymbol{\delta}}} \left(1 - \frac{\boldsymbol{\delta}^{\intercal}\widetilde{\boldsymbol{\delta}}}{\|\boldsymbol{\delta}\|_2^2 + \sigma_0^2}\right)^{-\frac{n}{2}} \left(1 - \frac{\boldsymbol{\delta}^{\intercal}\widetilde{\boldsymbol{\delta}}}{\|\widetilde{\boldsymbol{\delta}}\|_2^2 + \sigma_0^2}\right)^{-\frac{n}{2}}. \tag{B.3.29}$$

$$\chi^2\left(f_{\theta_1}, f_{\theta_0}\right) + 1 = \mathbb{E}_{\boldsymbol{\delta}, \widetilde{\boldsymbol{\delta}}} \left(1 - \frac{\|\boldsymbol{\delta}\|_2^2 \|\widetilde{\boldsymbol{\delta}}\|_2^2}{\sigma_0^4}\right)^{-\frac{n}{2}}. \tag{B.3.30}$$

The following lemma (Lemma 3 in Cai & Guo (2016b)) controls the right hand side of (B.3.29).

Lemma 19. *Suppose the random variable J follows* $\mathrm{Hypergeometric}\left(p, k, k\right)$ *with* $\mathbb{P}\left(J = j\right) = \frac{\binom{k}{j}\binom{p-k}{k-j}}{\binom{p}{k}}$, *then we have*

$$\mathbb{E} \exp\left(tJ\right) \leq e^{\frac{k^2}{p-k}} \left(1 - \frac{k}{p} + \frac{k}{p}\exp\left(t\right)\right)^k. \tag{B.3.31}$$

By the construction (B.3.26), we have $\|\widetilde{\boldsymbol{\delta}}\|_2^2 = \|\boldsymbol{\delta}\|_2^2 = (k_2 - k_0)\rho^2$ and $\boldsymbol{\delta}^\intercal \widetilde{\boldsymbol{\delta}} \leq (k_2 - k_0)\rho^2$. By the inequality $\frac{1}{1-x} \leq \exp(2x)$ for $x \in \left[0, \frac{\log 2}{2}\right]$, if $\frac{(k_2-k_0)\rho^2}{\sigma_0^2} < \frac{\log 2}{2}$, we have

$$\left(1 - \frac{\boldsymbol{\delta}^\intercal \widetilde{\boldsymbol{\delta}}}{\|\delta\|_2^2 + \sigma_0^2}\right)^{-\frac{n}{2}} \left(1 - \frac{\boldsymbol{\delta}^\intercal \widetilde{\boldsymbol{\delta}}}{\|\widetilde{\delta}\|_2^2 + \sigma_0^2}\right)^{-\frac{n}{2}} \leq \exp\left(2n \frac{\boldsymbol{\delta}^\intercal \widetilde{\boldsymbol{\delta}}}{(k_2 - k_0)\rho^2 + \sigma_0^2}\right) \leq \exp\left(2n \frac{\boldsymbol{\delta}^\intercal \widetilde{\boldsymbol{\delta}}}{\sigma_0^2}\right). \tag{B.3.32}$$

Let J denote the hypergeometric distribution with parameters $(p_1, k_2 - k_0, k_2 - k_0)$. We further have

$$\begin{aligned} \mathbb{E} \exp\left(2n \frac{\boldsymbol{\delta}^\intercal \widetilde{\boldsymbol{\delta}}}{\sigma_0^2}\right) &= \mathbb{E}_J \exp\left(2n \frac{J\rho^2}{\sigma_0^2}\right) \leq e^{\frac{(k_2-k_0)^2}{p_1-(k_2-k_0)}} \left(1 - \frac{k_2-k_0}{p_1} + \frac{k_2-k_0}{p_1} \exp\left(\frac{1}{\sigma_0^2} 2n\rho^2\right)\right)^{(k_2-k_0)} \\ &\leq e^{\frac{(k_2-k_0)^2}{p_1-(k_2-k_0)}} \left(1 - \frac{k_2-k_0}{p_1} + \frac{k_2-k_0}{p_1} \sqrt{\frac{p_1}{(k_2-k_0)^2}}\right)^{(k_2-k_0)} \leq e^{\frac{(k_2-k_0)^2}{p_1-(k_2-k_0)}} \left(1 + \frac{1}{\sqrt{p_1}}\right)^{(k_2-k_0)}, \end{aligned} \tag{B.3.33}$$

where the first inequality applies Lemma 19 and the second inequality follows by plugging $\rho = \frac{1}{2}\sqrt{\frac{\log \frac{p_1}{(k_2-k_0)^2}}{n}}\sigma_0$. If $(k_2 - k_0) \leq c_0 \min\left\{\frac{n}{\log p}, p^\gamma\right\}$, we have $(k_2 - k_0)\rho^2 < \frac{\log 2}{2}\sigma_0^2$. Since $(k_2 - k_0) \leq c_0 p^\gamma$ with $0 \leq \gamma < \frac{1}{2}$, we have $\chi^2(f_\pi, f_{\theta_1}) \leq \epsilon_1^2$. Since $\frac{\|\boldsymbol{\delta}\|_2^2}{\sigma_0^2} = \frac{\|\widetilde{\boldsymbol{\delta}}\|_2^2}{\sigma_0^2} = \frac{(k_2-k_0)\rho^2}{\sigma_0^2} < \frac{\log 2}{2}$, we have the following control of (B.3.30),

$$\mathbb{E}_{\boldsymbol{\delta}, \widetilde{\boldsymbol{\delta}}} \left(1 - \frac{\|\boldsymbol{\delta}\|_2^2 \|\widetilde{\boldsymbol{\delta}}\|_2^2}{\sigma_0^4}\right)^{-\frac{n}{2}} \leq \exp\left(n \left(\frac{(k_2-k_0)\rho^2}{\sigma_0^2}\right)^2\right) = \exp\left(\frac{\left((k_2-k_0)\log \frac{p_1}{(k_2-k_0)^2}\right)^2}{16n}\right), \tag{B.3.34}$$

where the inequality follows from $\frac{1}{1-x} \leq \exp(2x)$ for $x \in \left[0, \frac{\log 2}{2}\right]$ and the equality follows by plugging in $\rho = \frac{1}{2}\sqrt{\frac{\log \frac{p_1}{(k_2-k_0)^2}}{n}}\sigma_0$. Under the assumption $(k_2 - k_0) \leq c\frac{\sqrt{n}}{\log p}$, we have $\chi^2(f_{\theta_1}, f_{\theta_0}) \leq \epsilon_1^2$ and hence $\mathrm{TV}(f_\pi, f_{\theta_1}) \leq \epsilon_1$, $\mathrm{TV}(f_{\theta_1}, f_{\theta_0}) \leq \epsilon_1$ and $\mathrm{TV}(f_\pi, f_{\theta_0}) \leq 2\epsilon_1$. Note that $d = (k_2 - k_0)^{\frac{1}{q}}\rho$. By (3.2.6) in Chapter 3 and $\|\beta^*\|_0 \leq k_0$, we establish

$$\mathbb{P}_{\theta_0}\left(\|\widehat{\beta} - \beta^*\|_q^2 \leq C\left(\frac{k_0}{k_2 - k_0}\right)^{\frac{2}{q}} d^2\right) \geq 1 - \alpha_0.$$

By (3.6.3) in Chapter 3 and the fact $C\left(\frac{k_0}{k_2-k_0}\right)^{\frac{1}{q}} \leq \frac{1}{16}$, we establish (B.3.23).

Proof of (B.3.24)

The proof of (B.3.24) is based on the exactly same argument with (B.3.23) by taking $\rho = (\log(1+\epsilon_1^2))^{\frac{1}{4}} \frac{1}{n^{\frac{1}{4}}\sqrt{k_2-k_0}}\sigma_0$. Since $k_2 \geq C\frac{\sqrt{n}}{\log p}$ and $k_0 \leq c_0 k_2$, then

$$(\log\left(1+\epsilon_1^2\right))^{\frac{1}{4}} \frac{1}{n^{\frac{1}{4}}\sqrt{k_2-k_0}}\sigma_0 \leq \frac{1}{2}\sqrt{\frac{\log\frac{p_1}{(k_2-k_0)^2}}{n}}\sigma_0$$

and hence (B.3.33) holds. It is sufficient to control the following term

$$\mathbb{E}_{\boldsymbol{\delta},\widetilde{\boldsymbol{\delta}}}\left(1-\frac{\|\boldsymbol{\delta}\|_2^2\|\widetilde{\boldsymbol{\delta}}\|_2^2}{\sigma_0^4}\right)^{-\frac{n}{2}} \leq \exp\left(n\left(\frac{(k_2-k_0)\rho^2}{\sigma_0^2}\right)^2\right) \leq 1+\epsilon_1^2. \tag{B.3.35}$$

In this case, $d^2 = \sqrt{\log(1+\epsilon_1^2)}(k_2-k_0)^{\frac{2}{q}-1}\frac{1}{\sqrt{n}}\sigma_0^2$ and

$$\mathbb{P}_{\theta_0}\left(\|\widehat{\beta}-\beta^*\|_q^2 \leq C\frac{k_0^{\frac{2}{q}}\frac{\log p}{n}}{(k_2-k_0)^{\frac{2}{q}-1}\frac{1}{\sqrt{n}}}d^2\right) \geq 1-\alpha_0.$$

Since $k_0 \leq c_0\min\{k_1, \frac{\sqrt{n}}{\log p}\}$ and $k_1 \leq k_2$, we have $C\frac{k_0^{\frac{2}{q}}\frac{\log p}{n}}{(k_2-k_0)^{\frac{2}{q}-1}\frac{1}{\sqrt{n}}} \leq \frac{1}{16}$ and the lower bound (B.3.24) follows from (3.6.3) of Theorem 16 in Chapter 3.

Proof of (B.3.25) The proof of (B.3.25) is an application of (3.6.6) of Theorem 17 in Chapter 3. The key is to construct parameter spaces $\mathcal{F}_1$, $\mathcal{F}_2$ and the points θ_1 and θ_2 and then control the distribution distances between the density functions. Let $S = \mathrm{supp}(\beta^*)$. Without loss of generality, we assume $S = \{1, 2, \cdots, k_0\}$. Define p_0 denote the largest integer smaller than $\frac{p-k_0}{2}$, $\mathcal{I}_1 = \{k_0+1, k_0+2, \cdots, k_0+p_0\}$ and $\mathcal{I}_2 = \{k_0+p_0+1, k_0+p_0+2, \cdots, p\}$. Define $\Sigma_0 = \left(\begin{array}{c|c|c} \mathrm{I}_{S\times S} & \mathbf{0} & \mathbf{0} \\ \hline \mathbf{0} & M_1\mathrm{I}_{\mathcal{I}_1\times\mathcal{I}_1} & \mathbf{0} \\ \hline \mathbf{0} & \mathbf{0} & \frac{1}{M_1}\mathrm{I}_{\mathcal{I}_2\times\mathcal{I}_2} \end{array}\right)$.

For $\theta_0 = (\beta^*, \mathrm{I}, \sigma_0)$, we construct

$$
\begin{aligned}
\mathcal{F}_1 &= \left\{ (\beta^* + \boldsymbol{\nu}, \Sigma_0, \sigma_0) \ : \ \boldsymbol{\nu} \in \ell\left(\mathcal{I}_1, k_1 - k_0, \frac{1}{\sqrt{M_1}}\rho \right) \right\} \subset \Theta_0\left(k_1\right); \\
\mathcal{F}_2 &= \left\{ (\beta^* + \boldsymbol{\delta}, \mathrm{I}, \sigma_0) \ : \ \boldsymbol{\delta} \in \ell\left(\mathcal{I}_2, k_2 - k_0, \sqrt{M_1}\sqrt{\frac{k_1 - k_0}{k_2 - k_0}}\rho \right) \right\} \subset \Theta_0\left(k_2\right),
\end{aligned}
\tag{B.3.36}
$$

where

$$
\ell\left(\mathcal{I}_1, k_2 - k_0, \frac{1}{\sqrt{M_1}}\rho\right) = \left\{ \boldsymbol{\nu} : \mathrm{supp}\left(\boldsymbol{\nu}\right) \subset \mathcal{I}_1, \|\boldsymbol{\nu}\|_0 = k_1 - k_0, \ \boldsymbol{\nu}_i \in \frac{1}{\sqrt{M_1}}\{0, \rho\} \right\}, \tag{B.3.37}
$$

and

$$
\ell\left(\mathcal{I}_2, k_2 - k_0, \sqrt{M_1}\sqrt{\frac{k_1 - k_0}{k_2 - k_0}}\rho\right) = \left\{ \boldsymbol{\delta} : \mathrm{supp}\left(\boldsymbol{\delta}\right) \subset \mathcal{I}_2, \|\boldsymbol{\delta}\|_0 = k_2 - k_0, \ \boldsymbol{\delta}_i \in \sqrt{M_1}\sqrt{\frac{k_1 - k_0}{k_2 - k_0}}\{0, \rho\} \right\}. \tag{B.3.38}
$$

Let π_i denote the uniform prior on the parameter space $\mathcal{F}_i$ for $i = 1, 2$. The corresponding parameter spaces for Σ^z corresponding to $\mathcal{F}_1$ and $\mathcal{F}_2$ are

$$
\mathcal{H}_1 = \left\{ \Sigma^z_{\boldsymbol{\nu}} : \boldsymbol{\nu} \in \ell\left(\mathcal{I}_1, k_1 - k_0, \frac{1}{\sqrt{M_1}}\rho\right) \right\} \text{ where } \Sigma^z_{\boldsymbol{\nu}} = \left(\begin{array}{c|c} \|\beta^*\|_2^2 + \left(k_1 - k_0\right)\rho^2 + \sigma_0^2 & (\beta^* + \boldsymbol{\nu})^\intercal \\ \hline \beta^* + \boldsymbol{\nu} & \Sigma_0 \end{array} \right),
$$

and

$$
\mathcal{H}_2 = \left\{ \Sigma^z_{\boldsymbol{\delta}} : \boldsymbol{\delta} \in \ell\left(\mathcal{I}_2, k_2 - k_0, \sqrt{M_1}\sqrt{\frac{k_1 - k_0}{k_2 - k_0}}\rho\right) \right\} \text{ where } \Sigma^z_{\boldsymbol{\delta}} = \left(\begin{array}{c|c} \|\beta^*\|_2^2 + \left(k_1 - k_0\right)\rho^2 + \sigma_0^2 & (\beta^* + \boldsymbol{\delta})^\intercal \\ \hline \beta^* + \boldsymbol{\delta} & \Sigma_0 \end{array} \right).
$$

Since $\mathrm{dist}_1^2 = M_1\|\boldsymbol{\nu}\|_2^2 = (k_1 - k_0)\rho^2$ and $\mathrm{dist}_2^2 = \frac{1}{M_1}\|\boldsymbol{\delta}\|_2^2 = (k_1 - k_0)\rho^2$, we have $\theta_1 = \theta_2 = (\beta^*, \Sigma_0, \sigma_0^2 + (k_1 - k_0)\rho^2)$. In this case, we have $\mathrm{TV}\left(f_{\theta_2}, f_{\theta_1}\right) = 0$. By the same argument of (B.3.29) in Lemma 18, we have

$$
\chi^2\left(f_{\pi_1}, f_{\theta_1}\right) + 1 = \mathbb{E}_{\boldsymbol{\nu}, \widetilde{\boldsymbol{\nu}}} \left(1 - \frac{M_1 \boldsymbol{\nu}^\intercal \widetilde{\boldsymbol{\nu}}}{\left(k_1 - k_0\right)\rho^2 + \sigma_0^2}\right)^{-\frac{n}{2}} \left(1 - \frac{M_1 \boldsymbol{\nu}^\intercal \widetilde{\boldsymbol{\nu}}}{\left(k_1 - k_0\right)\rho^2 + \sigma_0^2}\right)^{-\frac{n}{2}},
$$

and

$$\chi^2\left(f_{\pi_2}, f_{\theta_2}\right) + 1 = \mathbb{E}_{\boldsymbol{\delta},\widetilde{\boldsymbol{\delta}}}\left(1 - \frac{\frac{1}{M_1}\boldsymbol{\delta}^{\intercal}\widetilde{\boldsymbol{\delta}}}{(k_1-k_0)\,\rho^2+\sigma_0^2}\right)^{-\frac{n}{2}}\left(1 - \frac{\frac{1}{M_1}\boldsymbol{\delta}^{\intercal}\widetilde{\boldsymbol{\delta}}}{(k_1-k_0)\,\rho^2+\sigma_0^2}\right)^{-\frac{n}{2}}.$$

Taking $\rho = \frac{1}{2}\sqrt{\frac{\log\frac{p_1}{(k_2-k_0)^2}}{n}}\sigma_0$, a similar argument to (B.3.32) and (B.3.33) leads to $\mathrm{TV}\left(f_{\pi_i}, f_{\theta_i}\right) \le \epsilon_1$ for $i = 1, 2$. Note that $d_1^2 = \frac{1}{M_1}(k_1-k_0)^{\frac{2}{q}}\rho^2$, $d_2^2 = M_1(k_2-k_0)^{\frac{2}{q}-1}(k_1-k_0)\rho^2$. The assumption (3.2.7) leads to $\mathbb{P}_{\theta_i}\left(\|\widehat{\beta}-\beta^*\|_q^2 \le c_i^2 d_i^2\right) \ge 1-\alpha_0$, for $i = 1, 2$, where $c_1 = \frac{C^*\sqrt{M_1}k_0^{\frac{1}{q}}}{(k_1-k_0)^{\frac{1}{q}}}$ and $c_2 = \frac{C^* k_0^{\frac{1}{q}}}{\sqrt{M_1}(k_2-k_0)^{\frac{1}{q}-\frac{1}{2}}(k_1-k_0)^{\frac{1}{2}}}$. By (3.6.6) in Chapter 3, we obtain

$$\begin{aligned}&\mathbf{R}_{\alpha}^{*}\left(\Theta_0\left(k_1\right), \Theta_0\left(k_2\right), \widehat{\beta}, \ell_q\right) \\ &\ge c\left(\left(1-c_2\right)^2 M_1(k_2-k_0)^{\frac{2}{q}-1}(k_1-k_0)\rho^2 - \frac{1}{M_1}\left(1+c_1\right)^2\left(k_1-k_0\right)^{\frac{2}{q}}\rho^2\right)_+.\end{aligned} \tag{B.3.39}$$

B.3.3 Proof of Theorem 18

Theorem 18 is implied by Theorem 22. Since $k_0 \le c_0^* k_1$, we have $(1-c_2)^2 \ge \frac{1}{\sqrt{M_1}}$ and $(1+c_1)^2 \le \sqrt{M_1}$. By $k_0 \le c_0 \min\{k_1, \frac{\sqrt{n}}{\log p}\}$, (B.2.4) implies the lower bound $\min\{\frac{k_1 \log p}{n}, \frac{1}{\sqrt{n}}\}\sigma_0^2$. Combined with (B.2.3), we can establish (3.7.3) in Chapter 3.

B.3.4 Proof of Theorems 11 and 13

The minimax lower bound of Theorem 11 follows from Theorem 12. We take $k_1 = k_2 = k$ and (3.3.5) in Chapter 3 follows from (3.3.7) in Chapter 3. The minimax lower bound (3.3.6) in Chapter 3 follows from (3.3.5) in Chapter 3 and Lemma 8. The minimax lower bound of Theorem 13 follows from Theorem 14. We take $k_1 = k_2 = k$ and (3.3.13) follows from (3.3.18). The minimax lower bound (3.3.14) in Chapter 3 follows from (3.3.13) in Chapter 3 and Lemma 8.

B.3.5 Proof of Theorems 10 and 15

The proofs of Theorem 10 and 15 are applications of the minimax lower bounds (3.6.2) and (3.6.3) of Theorem 16 in Chapter 3, respectively. To apply Theorem 16, it is sufficient to construct the least favorable set $\mathcal{F}$ corresponding to the point $\theta_0 = (\beta^*, \mathrm{I}, \sigma_0)$ such that the distribution distance $\mathrm{TV}\left(f_\pi, f_{\theta_0}\right)$ or $\chi^2\left(f_\pi, f_{\theta_0}\right)$ is controlled and the functional distance $d = \min_{\theta \in \mathcal{F}} \|\beta(\theta) - \beta^*\|_q$ is maximized. In the following, we first establish Theorem 15 by constructing $\mathcal{F}$ with the prior π and control the distance $\chi^2\left(f_\pi, f_{\theta_0}\right)$. Taking $\theta_0 = (\beta^*, \mathrm{I}, \sigma_0)$ with $\|\beta^*\|_0 \leq k_0$ and $\Theta = \Theta(k_2)$, we define $\mathcal{F} = \mathcal{F}(\theta_0, k_2 - k_0, \rho)$ as

$$\mathcal{F}(\theta_0, k_2 - k_0, \rho) = \left\{\theta = (\beta^* + \boldsymbol{\delta}, \mathrm{I}, \sigma) : \boldsymbol{\delta} \in \ell(\beta^*, k_2 - k_0, \rho)\ , \sigma = \sqrt{\sigma_0^2 - (k_2 - k_0)\rho^2}\right\}, \tag{B.3.40}$$

where $\ell(\beta^*, k_2 - k_0, \rho)$ is defined in (B.3.27). Note that $\mathcal{F}(\theta_0, k_2 - k_0, \rho) \subset \Theta(k_2)$. The prior π on $\mathcal{F}(\theta_0, k_2 - k_0, \rho)$ is induced by the uniform prior of $\boldsymbol{\delta}$ on $\ell(\beta^*, k_2 - k_0, \rho)$. In the following, we will control $\chi^2\left(f_\pi, f_{\theta_0}\right)$.

Let Σ_0^z denote the covariance matrix of (y_i, X_i) corresponding to $\theta_0 = (\beta^*, \mathrm{I}, \sigma) \in \Theta(k_0)$. Let $S = \mathrm{supp}(\beta^*)$. Without loss of generality, we assume $S = \{1, 2, \cdots, k_0\}$. Let p_1 denote the size of S^c and hence $p_1 = p - k_0$. We have the expression of the covariance matrix Σ_0^z corresponding to θ_0,

$$\Sigma_0^z = \left(\begin{array}{c|c|c} \|\beta^*\|_2^2 + \sigma_0^2 & (\beta_S^*)^\intercal & \mathbf{0}_{1\times p_1} \\ \hline \beta_S^* & \mathrm{I}_{k_0\times k_0} & \mathbf{0}_{k_0\times p_1} \\ \hline \mathbf{0}_{p_1\times 1} & \mathbf{0}_{p_1\times k_0} & \mathrm{I}_{p_1\times p_1} \end{array}\right). \tag{B.3.41}$$

The corresponding set of Σ^z to $\mathcal{F}=\mathcal{F}\left(\theta_0,k_2-k_0,\rho\right)$ is

$$\mathcal{H}=\left\{\Sigma^z_{\boldsymbol{\delta}}:\boldsymbol{\delta}\in\ell\left(\beta^*,k_2-k_0,\rho\right)\right\},\ \text{with}\ \ \Sigma^z_{\boldsymbol{\delta}}=\left(\begin{array}{c|c|c}\|\beta^*\|_2^2+\sigma_0^2 & (\beta^*_S)^{\intercal} & \boldsymbol{\delta}^{\intercal}\\ \hline \beta^*_S & \mathrm{I}_{k_0\times k_0} & \mathbf{0}_{k_0\times p_1}\\ \hline \boldsymbol{\delta} & \mathbf{0}_{p_1\times k_0} & \mathrm{I}_{p_1\times p_1}\end{array}\right). \tag{B.3.42}$$

The following lemma (Lemma 7 in Cai & Guo (2016c)) controls the χ^2 distance between f_π and f_{θ_0}.

Lemma 20.

$$\chi^2\left(f_\pi,f_{\theta_0}\right)+1=\mathbb{E}_{\boldsymbol{\delta},\widetilde{\boldsymbol{\delta}}}\left(1-\frac{1}{\sigma_0^2}\boldsymbol{\delta}^{\intercal}\widetilde{\boldsymbol{\delta}}\right)^{-n}. \tag{B.3.43}$$

By the construction (B.3.40), we have $\|\widetilde{\boldsymbol{\delta}}\|_2^2=\|\boldsymbol{\delta}\|_2^2=\left(k_2-k_0\right)\rho^2$ and $\boldsymbol{\delta}^{\intercal}\widetilde{\boldsymbol{\delta}}\leq\left(k_2-k_0\right)\rho^2$. By the inequality $\frac{1}{1-x}\leq\exp(2x)$ for $x\in\left[0,\frac{\log 2}{2}\right]$, if $\frac{(k_2-k_0)\rho^2}{\sigma_0^2}<\frac{\log 2}{2}$, we have $\left(1-\frac{1}{\sigma_0^2}\boldsymbol{\delta}^{\intercal}\widetilde{\boldsymbol{\delta}}\right)^{-n}\leq\exp\left(\frac{2}{\sigma_0^2}n\boldsymbol{\delta}^{\intercal}\widetilde{\boldsymbol{\delta}}\right)$. Let J denote the hypergeometric distribution with parameters (p_1,k_2-k_0,k_2-k_0). We further have

$$\begin{aligned}\mathbb{E}\exp\left(\frac{2}{\sigma_0^2}n\boldsymbol{\delta}^{\intercal}\widetilde{\boldsymbol{\delta}}\right)&=\mathbb{E}\exp\left(\frac{1}{\sigma_0^2}2Jn\rho^2\right)\leq e^{\frac{(k_2-k_0)^2}{p_1-(k_2-k_0)}}\left(1-\frac{k_2-k_0}{p_1}+\frac{k_2-k_0}{p_1}\exp\left(\frac{1}{\sigma_0^2}2n\rho^2\right)\right)^{k_2-k_0}\\ &\leq e^{\frac{(k_2-k_0)^2}{p_1-(k_2-k_0)}}\left(1-\frac{k_2-k_0}{p_1}+\frac{k_2-k_0}{p_1}\sqrt{\frac{p_1}{\left(k_2-k_0\right)^2}}\right)^{k_2-k_0}\leq e^{\frac{(k_2-k_0)^2}{p_1-(k_2-k_0)}}\left(1+\frac{1}{\sqrt{p_1}}\right)^{k_2-k_0},\end{aligned} \tag{B.3.44}$$

where the first inequality applies Lemma 19 and the second inequality follows by plugging $\rho=\frac{1}{2}\sqrt{\frac{\log\frac{p_1}{(k_2-k_0)^2}}{n}}\sigma_0$. If $k_2\leq c_0\min\left\{\frac{n}{\log p},p^{\gamma}\right\}$, we have $\frac{(k_2-k_0)\rho^2}{\sigma_0^2}<\frac{\log 2}{2}$ and establish $\chi^2\left(f_\pi,f_{\theta_0}\right)\leq\epsilon_1^2$ by (B.3.44) and $\mathrm{TV}(f_\pi,f_{\theta_0})\leq\epsilon_1$ by (3.9.1) in Chapter 3, where $\epsilon_1=\frac{1-2\alpha-2\alpha_0}{12}$.

To establish Theorem 15, we apply Theorem 16 and compute

$$d=\frac{1}{2}(k_2-k_0)^{\frac{1}{q}}\sqrt{\frac{\log\frac{p_1}{(k_2-k_0)^2}}{n}}\sigma_0.$$

By (3.2.6) in Chapter 3 and $\|\beta^*\|_0 \leq k_0$, we establish

$$\mathbb{P}_{\theta_0}\left(\|\widehat{\beta}-\beta^*\|_q^2 \leq C^*\left(\frac{k_0}{k_2-k_0}\right)^{\frac{2}{q}} d^2\right) \geq 1-\alpha_0. \tag{B.3.45}$$

By the fact $C^*\left(\frac{k_0}{k_2-k_0}\right)^{\frac{2}{q}} \leq \frac{1}{16}$, we establish (3.6.1) in Chapter 3. By applying (3.6.3) of Theorem 16, we establish (3.4.2) in Chapter 3. Since $\theta_0 \in \Theta(k_2)$ and $\mathbf{R}_\alpha^*\left(\Theta\left(k_2\right), \widehat{\beta}, \ell_q\right) \geq \mathbf{R}_\alpha^*\left(\{\theta_0\}, \Theta\left(k_2\right), \widehat{\beta}, \ell_q\right)$, the lower bound (3.4.1) in Chapter 3 with $i=2$ follows from (3.4.2) in Chapter 3. For (3.4.1) in Chapter 3 with $i=1$, the lower bound is established using the above argument with k_2 replaced by k_1. The following lemma shows that $\widehat{\beta}^{SL}$ with $A > 2\sqrt{2}$ satisfying the assumption (A2) and hence the lower bounds (3.4.1) and (3.4.2) in Chapter 3 hold for $\widehat{\beta}^{SL}$ with $A > 2\sqrt{2}$.

Lemma 21. *If $A > 2\sqrt{2}$, then we have*

$$\mathbb{P}_{\theta_0}\left(\|\widehat{\beta}^{SL}-\beta^*\|_q^2 \geq C\|\beta^*\|_0^{\frac{2}{q}}\frac{\log p}{n}\sigma_0^2\right) \leq c\exp\left(-c'n\right)+p^{-c}.$$

To establish Theorem 10, we apply the general lower bound (3.6.2) in Chapter 3 and the same argument between (B.3.40) and (B.3.45) by replacing k_2 with k.

B.3.6 Proof of Theorem 9

The proof of Theorem 9 is similar to the proof of Theorem 19, which is presented in Section B.3.2. For the case $k \lesssim \frac{\sqrt{n}}{\log p}$, the proof is similar to (B.3.23). Taking θ_0, $\mathcal{F}$ and ρ as defined in the proof of (B.3.23), with $k_2 = k$, we apply (3.6.2) in Theorem 16, we establish (3.2.8) and (3.2.9) in the regime $k \lesssim \frac{\sqrt{n}}{\log p}$ in Chapter 3.

The proof of (3.2.8) in the case $\frac{\sqrt{n}}{\log p} \ll k \lesssim \frac{n}{\log p}$ is similar to that of (B.3.24). Taking θ_0, $\mathcal{F}$ and ρ as defined in the proof of (B.3.24), with $k_2 = k$, we establish (3.2.8) in Chapter 3 for $\frac{\sqrt{n}}{\log p} \ll k \lesssim \frac{n}{\log p}$. The proof of (3.2.9) in Chapter 3 in the case $\frac{\sqrt{n}}{\log p} \ll k \lesssim \frac{n}{\log p}$ is similar to that of (B.3.25). Taking $\theta_0, \theta_1, \theta_2$, $\mathcal{F}_1, \mathcal{F}_2$ and ρ as defined in the proof of (B.3.25), with $k_1 = \frac{1}{3}k$ and $k_2 = k$, we apply (3.6.5) of Theorem 17 to establish (3.2.9) in Chapter 3 for $\frac{\sqrt{n}}{\log p} \ll k \lesssim \frac{n}{\log p}$.

B.4 Upper bound analysis

In Section B.4.1, we establish minimax upper bounds of Theorems 11, 12, 13, 14 and 15 based on Propositions 3, 4, 5 and 6. In later sections, we establish Propositions 2, 3, 4, 5, 6, 7 and 10.

B.4.1 Proof of upper bounds of Theorems

In the following, we will establish the minimax upper bounds in the main paper based on Propositions 3, 4, 5 and 6.

Proof of the upper bound of Theorem 11 By Proposition 4, the minimax convergence rate (3.3.6) in Chapter 3 over $k \lesssim \frac{\sqrt{n}}{\log p}$ is achieved by the confidence interval $\mathrm{CI}_{\alpha}^{0}(Z, k, 2)$ defined in (3.3.15) in Chapter 3. By Proposition 3, the minimax convergence rate (3.3.6) in Chapter 3 over $\frac{\sqrt{n}}{\log p} \ll k \lesssim \frac{n}{\log p}$ is achieved by the confidence interval $\mathrm{CI}_{\alpha}^{1}(Z)$ defined in (3.3.8) in Chapter 3.

Proof of the upper bound of Theorem 12 By Proposition 4, the minimax lower bound (3.3.7) in Chapter 3 in the region $k_2 \lesssim \frac{\sqrt{n}}{\log p}$ is achieved by the confidence interval $\mathrm{CI}_{\alpha}^{0}(Z, k_2, 2)$. By proposition 3, the minimax lower bound (3.3.7) in Chapter 3 over $\frac{\sqrt{n}}{\log p} \ll k_2 \lesssim \frac{n}{\log p}$ is achieved by the confidence interval $\mathrm{CI}_{\alpha}^{1}(Z)$.

Proof of the upper bound of Theorem 13 By Proposition 4, the minimax convergence rate (3.3.14) in Chapter 3 is achieved by the confidence interval $\mathrm{CI}_{\alpha}^{0}(Z, k, q)$.

Proof of the upper bound of Theorem 14 By Proposition 4, the minimax lower bound (3.3.18) in Chapter 3 in the regime $k_1 \leq k_2 \lesssim \frac{\sqrt{n}}{\log p}$ is achieved by the confidence interval $\mathrm{CI}_\alpha^0(Z, k_2, q)$. By Proposition 5, the minimax lower bounds (3.3.18) in Chapter 3 in the regime $k_1 \lesssim \frac{\sqrt{n}}{\log p} \lesssim k_2 \lesssim \frac{n}{\log p}$ and $\frac{\sqrt{n}}{\log p} \lesssim k_1 \leq k_2 \lesssim \frac{n}{\log p}$ are achieved by the confidence interval $\mathrm{CI}_\alpha^2(Z, k_2, q)$ defined in (3.3.19) in Chapter 3.

Proof of the upper bound of Theorem 15 By Proposition 6, the minimax lower bounds (3.4.1) in Chapter 3 are achieved by the confidence interval $\mathrm{CI}_\alpha(Z, k_i, q)$ defined in (3.4.4) in Chapter 3 for $i = 1, 2$ and the lower bound in (3.4.2) in Chapter 3 is achieved by the confidence interval $\mathrm{CI}_\alpha(Z, k_2, q)$.

B.4.2 Proof of Proposition 6

The following argument is similar to the upper bound argument in Cai & Guo (2016b,c), which also relies on the results from Bickel et al. (2009); Ren et al. (2013); Sun & Zhang (2012). We first normalize the columns of X and the true sparse vector β and the linear regression model can be expressed as

$$y = Wd + \epsilon, \quad \text{with } \; W = XD, \; d = D^{-1}\beta \text{ and } \epsilon \sim N(0, \sigma^2 \mathrm{I}), \tag{B.4.1}$$

where $D = \mathrm{diag}\left(\frac{\sqrt{n}}{\|X_{\cdot j}\|_2}\right)_{j\in[p]}$ denotes the $p \times p$ diagonal matrix with (j,j) entry to be $\frac{\sqrt{n}}{\|X_{\cdot j}\|_2}$. Setting $\delta_0 = \frac{A}{\sqrt{2}}$ and $\eta_0 = (\frac{A}{\sqrt{2}})^{\frac{1}{2}} - 1$, we have $\lambda_0 = (1+\eta_0)\sqrt{\frac{2\delta_0 \log p}{n}}$. Take $\epsilon_0 = \frac{2.01}{\eta_0} + 1$, $\nu_0 = 0.01$ and $C_1 = 2.25$. Rather than use the constants directly in the following discussion, we use $\delta_0, \eta_0, \epsilon_0, \nu_0$ and C_1 to represent the above fixed constants in the following discussion. We also assume that $\frac{\log p}{n} \leq \frac{1}{25}$ and $\delta_0 \log p > 2$.

Define the l_1 cone invertibility factor (CIF_1) as follows,

$$CIF_1(\alpha_0, K, W) = \inf\left\{\frac{|K|\|\frac{W^\intercal W}{n}u\|_\infty}{\|u_K\|_1} : \|u_{K^c}\|_1 \leq \alpha_0\|u_K\|_1, u \neq 0\right\}, \tag{B.4.2}$$

where K is an index set. Define $\sigma^{ora} = \frac{1}{\sqrt{n}}\|y - X\beta\|_2 = \frac{1}{\sqrt{n}}\|y - Wd\|_2$,

$$T = \{k : |d_k| \geq \lambda_0 \sigma^{ora}\}, \quad \tau = (1+\epsilon_0)\,\lambda_0 \max\left\{\frac{4}{\sigma^{ora}}\|d_{T^c}\|_1, \frac{8\lambda_0|T|}{CIF_1\,(2\epsilon_0+1, T, W)}\right\}. \tag{B.4.3}$$

To facilitate the proof, we define the following events for the random design X and the error ϵ,

$$\begin{aligned}
\mathcal{G}_1 &= \left\{\frac{2}{5}\frac{1}{\sqrt{M_1}} < \frac{\|X_{\cdot j}\|_2}{\sqrt{n}} < \frac{7}{5}\sqrt{M_1} \text{ for } 1 \leq j \leq p\right\},\\
\mathcal{G}_2 &= \left\{\left|\frac{(\sigma^{ora})^2}{\sigma^2} - 1\right| \leq 2\sqrt{\frac{\log p}{n}} + 2\frac{\log p}{n}\right\},\\
\mathcal{G}_3 &= \left\{\kappa(X, k, k, \alpha) \geq \frac{1}{4\sqrt{\lambda_{\max}\,(\Omega)}} - \frac{9}{\sqrt{\lambda_{\min}\,(\Omega)}}\,(1+\alpha)\sqrt{k\frac{\log p}{n}}\right\},\\
\mathcal{G}_4 &= \left\{\frac{\|W^\intercal \epsilon\|_\infty}{n} \leq \sigma\sqrt{\frac{2\delta_0 \log p}{n}}\right\},\\
\mathcal{S}_1 &= \left\{\frac{\|W^\intercal \epsilon\|_\infty}{n} \leq \sigma^{ora}\lambda_0 \frac{\epsilon_0 - 1}{\epsilon_0 + 1}\,(1-\tau)\right\},\\
\mathcal{S}_2 &= \{(1-\nu_0)\,\hat{\sigma} \leq \sigma \leq (1+\nu_0)\hat{\sigma}\}.
\end{aligned}$$

Define $\mathcal{G} = \cap_{i=1}^4 \mathcal{G}_i$ and $\mathcal{S} = \cap_{i=1}^2 \mathcal{S}_i$. We introduce the following lemma to control the probability of events $\mathcal{G}$ and $\mathcal{S}$. Lemma 22 was established as Lemma 4 in Cai & Guo (2016b), which relies on the results in Ren et al. (2013).

Lemma 22.

$$\mathbb{P}_\theta\,(\mathcal{G}) \geq 1 - \frac{6}{p} - 2p^{1-C_1} - \frac{1}{2\sqrt{\pi\delta_0 \log p}}p^{1-\delta_0} - c'\exp\,(-cn)\,, \tag{B.4.4}$$

where c and c' are universal positive constants. If $k \leq c\frac{n}{\log p}$, then

$$\mathbb{P}_\theta\left(\mathcal{G} \cap \mathcal{S}\right) \geq \mathbb{P}_\theta\left(\mathcal{G}\right) - 2\exp\left(-\left(\frac{g_0 + 1 - \sqrt{2g_0+1}}{2}\right)n\right) - c''\frac{1}{\sqrt{\log p}}p^{1-\delta_0}, \quad \text{(B.4.5)}$$

where c, c' and c'' are universal positive constants and $g_0 = \frac{\nu_0}{2+3\nu_0}$.

The following lemma establishes a data-dependent upper bound for the term $\|\widehat{\beta} - \beta\|_q^2$ with $1 \leq q \leq 2$. The proof of this lemma is in Section B.5.3.

Lemma 23. *On the event $\mathcal{G} \cap \mathcal{S}$,*

$$\|\widehat{\beta}^{SL} - \beta\|_q^2 \leq \left(\frac{16A\max\|X_{\cdot j}\|_2^2\widehat{\sigma}}{n\kappa^2\left(X, k, k, 3\left(\frac{\max\|X_{\cdot j}\|_2}{\min\|X_{\cdot j}\|_2}\right)\right)}\right)^2 k^{\frac{2}{q}}\frac{\log p}{n}. \quad \text{(B.4.6)}$$

On the event $\mathcal{G} \cap \mathcal{S}$, we have $\hat{\sigma} \leq (1+\nu_0)\sigma < \log p$ and there exists p_0 such that if $p \geq p_0$, then we have

$$\left(\frac{16A\max\|X_{\cdot j}\|_2^2\widehat{\sigma}}{n\kappa^2\left(X, k, k, 3\left(\frac{\max\|X_{\cdot j}\|_2}{\min\|X_{\cdot j}\|_2}\right)\right)}\right)^2 k^{\frac{2}{q}}\frac{\log p}{n} \leq C\left(k^{\frac{2}{q}}\frac{\log p}{n}\right)\widehat{\sigma}^2 \ll \left(k^{\frac{2}{q}}\frac{\log p}{n}\log p\right)\widehat{\sigma}^2. \quad \text{(B.4.7)}$$

By Lemma 23, we have $\mathbb{P}_\theta\left(\|\widehat{\beta} - \beta\|_q^2 \in \mathrm{CI}_\alpha\left(Z, k, q\right)\right) \geq \mathbb{P}_\theta(\mathcal{G} \cap \mathcal{S})$. Then the coverage property (3.4.5) in Chapter 3 follows from Lemma 22. Recall that $\mathcal{B} = \{\widehat{\sigma} \leq \log p\}$. The expected length is controlled as follows,

$$\begin{aligned}
&\mathbb{E}_\theta L\left(\mathrm{CI}_\alpha\left(Z,k,q\right)\right) = \mathbb{E}_\theta L\left(\mathrm{CI}_\alpha\left(Z,k,q\right)\right)\mathbf{1}_{\mathcal{B}} \\
=&\mathbb{E}_\theta L\left(\mathrm{CI}_\alpha\left(Z,k,q\right)\right)\mathbf{1}_{\mathcal{B}\cap(\mathcal{S}\cap\mathcal{G})} + \mathbb{E}_\theta L\left(\mathrm{CI}_\alpha\left(Z,k,q\right)\right)\mathbf{1}_{\mathcal{B}\cap(\mathcal{S}\cap\mathcal{G})^c} \\
\leq&Ck^{\frac{2}{q}}\frac{\log p}{n}\sigma^2 + k^{\frac{2}{q}}\frac{\log p}{n}\left(\log p\right)^3\mathbb{P}_\theta\left((\mathcal{S}\cap\mathcal{G})^c\right) \\
\leq&Ck^{\frac{2}{q}}\frac{\log p}{n}\left(\sigma^2 + C\left(p^{1-\min\{\delta_0, C_1\}} + c'\exp\left(-cn\right)\right)\left(\log p\right)^3\right),
\end{aligned} \quad \text{(B.4.8)}$$

where the first inequality follows from (B.4.7) and second inequality follows from

Lemma 22. If $\frac{\log p}{n} \leq c$, then $\left(p^{1-\min\{\delta_0, C_1\}} + c' \exp\left(-cn\right)\right) \left(\log p\right)^3 \to 0$ and hence (3.4.6) in Chapter 3 follows.

B.4.3 Proof of Proposition 4

For the split samples $y^{(1)} = X^{(1)}\beta + \epsilon^{(1)}$ and $y^{(2)} = X^{(2)}\beta + \epsilon^{(2)}$, we define the following events

$$
\begin{aligned}
\bar{\mathcal{G}}_1 &= \left\{ 0.9 < \frac{\|X^{(1)}_{j\cdot}\|_2}{\sqrt{n_1}} < 1.1 \, \text{for} \, 1 \leq j \leq p \right\}, \\
\bar{\mathcal{G}}_2 &= \left\{ \kappa(X^{(1)}, k, k, \alpha) \geq \frac{1}{4\sqrt{\lambda_{\max}\left(\Omega\right)}} - \frac{9}{\sqrt{\lambda_{\min}\left(\Omega\right)}} \left(1 + \alpha\right) \sqrt{2k \frac{\log p}{n_1}} \right\}, \\
\bar{\mathcal{G}}_3 &= \left\{ 2 \frac{\|\left(W^{(1)}\right)^\intercal \epsilon^{(1)}\|_\infty}{n_1} \leq A \sqrt{\frac{\log p}{n_1}} \sigma_0 \right\}, \\
\bar{\mathcal{G}}_4 &= \left\{ \frac{\|\left(W^{(1)}\right)^\intercal \epsilon^{(1)}\|_\infty}{n_1} \leq \frac{\eta_0 - 1}{\eta_0 + 1} A \sqrt{\frac{\log p}{n_1}} \sigma_0 \right\}, \\
\bar{\mathcal{G}}_5 &= \left\{ \|\widehat{\beta}^L - \beta\|_q^2 \leq C_1^*\left(A, k\right) k^{\frac{2}{q}} \frac{\log p}{n} \right\}, \\
\bar{\mathcal{G}}_6 &= \left\{ \|\widehat{\beta}^L - \beta\|_2^2 \leq C_2^*\left(A, k\right) k \frac{\log p}{n} \right\}, \\
\bar{\mathcal{G}}_7 &= \left\{ \frac{1}{n_2} \left\| y^{(2)} - X^{(2)} \widehat{\beta}^L \right\|_2^2 \leq \left(\sigma_0^2 + \|\widehat{\beta}^L - \beta\|_2^2\right) \left(1 + 2\sqrt{\frac{2\log p}{n_2}} + 2\frac{\log p}{n_2}\right) \right\}, \\
\bar{\mathcal{G}}_8 &= \left\{ \frac{1}{n_2} \left\| y^{(2)} - X^{(2)} \widehat{\beta}^L \right\|_2^2 \leq \sigma_0^2 \left(1 + \frac{k \log p}{n_1}\right) \left(1 + 2\sqrt{\frac{2\log p}{n_2}} + 2\frac{\log p}{n_2}\right) \right\},
\end{aligned}
$$

where $W^{(1)}_{\cdot j} = X^{(1)}_{\cdot j} \frac{\sqrt{n_1}}{\|X^{(1)}_{\cdot j}\|_2}$ for $j = 1, \cdots, p$ and $\eta_0 = 1.01 \frac{\sqrt{A}+\sqrt{2}}{\sqrt{A}-\sqrt{2}}$, $C_1^*(A, k) = \frac{\left(22A\sigma_0\right)^2}{\left(\frac{1}{4} - 42\sqrt{\frac{2k\log p}{n_1}}\right)^4}$ and $C_2^*(A, k) = \frac{\left(\frac{3\eta_0}{\eta_0+1} A\sigma_0\right)^2}{\left(\frac{1}{4} - (9+11\eta_0)\sqrt{\frac{2k\log p}{n_1}}\right)^4}$. We introduce the following lemma to control the probability $\mathbb{P}_\theta(\bar{\mathcal{G}}_i)$ for $1 \leq i \leq 8$. The lemma is proved in Section B.5.4.

Lemma 24. *Suppose $k \leq c\frac{n}{\log p}$ and $\theta \in \Theta_0(k)$. If $A > 4\sqrt{2}$, we have*

$$\mathbb{P}_\theta\left(\bar{\mathcal{G}}_5\right) \geq \mathbb{P}_\theta\left(\bar{\mathcal{G}}_1 \cap \bar{\mathcal{G}}_2 \cap \bar{\mathcal{G}}_3\right) \geq 1 - c\exp\left(-c'n\right) - cp^{1-\frac{A^2}{8}}. \tag{B.4.9}$$

If $A > \sqrt{2}$, we have

$$\mathbb{P}_\theta\left(\bar{\mathcal{G}}_6\right) \geq \mathbb{P}_\theta\left(\bar{\mathcal{G}}_1 \cap \bar{\mathcal{G}}_2 \cap \bar{\mathcal{G}}_4\right) \geq 1 - c\exp\left(-c'n\right) - p^{-c}, \tag{B.4.10}$$

and

$$\mathbb{P}_\theta\left(\bar{\mathcal{G}}_7\right) \geq 1 - p^{-2} \quad \text{and} \quad \mathbb{P}_\theta\left(\bar{\mathcal{G}}_8\right) \geq \mathbb{P}_\theta\left(\bar{\mathcal{G}}_6\right)(1 - cp^{-2}), \tag{B.4.11}$$

where c and c' are positive constants.

The coverage property (3.3.16) in Chapter 3 follows from the fact that

$$\mathbb{P}_\theta\left(\|\widehat{\beta}^L - \beta\|_q^2 \in \mathrm{CI}_\alpha^0\left(Z, k, q\right)\right) \geq \mathbb{P}_\theta(\bar{\mathcal{G}}_5).$$

For the case $q = 2$, the coverage property follows from

$$\mathbb{P}_\theta\left(\|\widehat{\beta}^L - \beta\|_2^2 \in \mathrm{CI}_\alpha^0\left(Z, k, 2\right)\right) \geq \mathbb{P}_\theta(\bar{\mathcal{G}}_6).$$

The expected length (3.3.17) in Chapter 3 follows from the definition of $\mathrm{CI}_\alpha^0\left(Z, k, q\right)$.

B.4.4 Proof of Propositions 2, 3, 5 and 10

We will first introduce the following lemma, which establishes conditional distribution of $\|y^{(2)} - X^{(2)}\widehat{\beta}\|_2^2$.

Lemma 25. *Suppose the estimator $\widehat{\beta}$ is constructed based on the subsample $Z^{(1)} =$*

$\left(y^{(1)}, X^{(1)}\right)$, *then*

$$\|y^{(2)} - X^{(2)}\widehat{\beta}\|_2^2 \mid \left(y^{(1)}, X^{(1)}\right) \sim \left(\|\Sigma^{1/2}(\beta - \widehat{\beta})\|_2^2 + \sigma_0^2\right) \chi^2\left(n_2\right), \tag{B.4.12}$$

and

$$\mathbb{P}_\theta\left(\chi^2_{\frac{\alpha}{2}}\left(n_2\right) \le \frac{\|y^{(2)} - X^{(2)}\widehat{\beta}\|_2^2}{\|\Sigma^{1/2}(\beta - \widehat{\beta})\|_2^2 + \sigma_0^2} \le \chi^2_{1-\frac{\alpha}{2}}\left(n_2\right)\right) = 1 - \alpha. \tag{B.4.13}$$

The above Lemma follows from the observation that conditioning on $\left(y^{(1)}, X^{(1)}\right)$, $y_i^{(2)} - X_i^{(2)}\widehat{\beta} = X_i^{(2)}\left(\beta - \widehat{\beta}\right) + \epsilon_i^{(2)} \sim N\left(0, \|\Sigma^{1/2}(\beta - \widehat{\beta})\|_2^2 + \sigma_0^2\right)$.

Proof of Proposition 2

We first introduce the following lemma (Theorem 2.3 in Boucheron et al. (2013)) about concentration of χ^2 random variable.

Lemma 26. *Let χ_n^2 denote the χ^2 random variable with n degrees of freedom, then we have the following concentration inequality,*

$$\mathbb{P}\left(\left|\chi_n^2 - E\chi_n^2\right| > 2\sqrt{nt} + 2t\right) \le 2\exp(-t).$$

Since $\|\widehat{\beta}^L - \beta\|_2^2$ is non-negative, we have

$$\mathbb{P}_\theta\left(\left|\widetilde{L}_2 - \|\widehat{\beta}^L - \beta\|_2^2\right| \ge \delta_{n,p}\frac{1}{\sqrt{n}}\right) \le \mathbb{P}_\theta\left(\left|\frac{1}{n_2}\|y^{(2)} - X^{(2)}\widehat{\beta}^L\|_2^2 - \sigma_0^2 - \|\widehat{\beta}^L - \beta\|_2^2\right| \ge \delta_{n,p}\frac{1}{\sqrt{n}}\right).$$

By (B.4.12) (with $\Sigma = \mathrm{I}$), we establish that

$$
\begin{aligned}
&\mathbb{P}_\theta\left(\left|\frac{1}{n_2}\|y^{(2)}-X^{(2)}\widehat{\beta}^L\|_2^2-\sigma_0^2-\|\widehat{\beta}^L-\beta\|_2^2\right|\geq \delta_{n,p}\frac{1}{\sqrt{n}}\right)\\
\leq&\mathbb{P}_\theta\left(\left\{\left|\frac{1}{n_2}\|y^{(2)}-X^{(2)}\widehat{\beta}^L\|_2^2-\sigma_0^2-\|\widehat{\beta}^L-\beta\|_2^2\right|\geq \delta_{n,p}\frac{1}{\sqrt{n}}\right\}\cap\bar{\mathcal{G}}_6\right)+\mathbb{P}_\theta(\bar{\mathcal{G}}_6^c)\\
\leq&\exp\left(-\frac{C\delta_{n,p}^2}{\sigma_0^4}\right)\mathbb{P}_\theta(\bar{\mathcal{G}}_6)+\mathbb{P}_\theta(\bar{\mathcal{G}}_6^c)\leq\exp\left(-\frac{C\delta_{n,p}^2}{\sigma_0^4}\right)+cp^{-c}+c\exp(-c'n),
\end{aligned}
\tag{B.4.14}
$$

where the last inequality follows from Lemma 26. Taking supremum on both sides of (B.4.14), we establish (3.2.12) in Chapter 3.

Proof of coverage properties in Propositions 3, 5 and 10

We first introduce the following confidence intervals,

$$
\bar{\mathrm{CI}}_\alpha^1(Z)=\left(\frac{\psi(Z)}{\frac{1}{n_2}\chi^2_{1-\frac{\alpha}{2}}(n_2)}-\sigma_0^2,\frac{\psi(Z)}{\frac{1}{n_2}\chi^2_{\frac{\alpha}{2}}(n_2)}-\sigma_0^2\right),
$$

$$
\bar{\mathrm{CI}}_\alpha^2(Z,k_2,q)=\left(\frac{\psi(Z)}{\frac{1}{n_2}\chi^2_{1-\frac{\alpha}{2}}(n_2)}-\sigma_0^2,(16k_2)^{\frac{2}{q}-1}\left(\frac{\psi(Z)}{\frac{1}{n_2}\chi^2_{\frac{\alpha}{2}}(n_2)}-\sigma_0^2\right)\right),
$$

$$
\bar{\mathrm{CI}}_\alpha^3(Z)=\left(0.99\lambda_{\min}\left(\frac{\psi(Z)}{\frac{1}{n_2}\chi^2_{1-\frac{\alpha}{2}}(n_2)}-\sigma_0^2\right),1.01\lambda_{\max}\left(\frac{\psi(Z)}{\frac{1}{n_2}\chi^2_{\frac{\alpha}{2}}(n_2)}-\sigma_0^2\right)\right),
$$

and

$$
\begin{aligned}
&\bar{\mathrm{CI}}_\alpha^4(Z,k_2,q)\\
=&\left(0.99\lambda_{\min}\left(\frac{\psi(Z)}{\frac{1}{n_2}\chi^2_{1-\frac{\alpha}{2}}(n_2)}-\sigma_0^2\right),1,01\lambda_{\max}\left((1+c^*)^2k_2\right)^{\frac{2}{q}-1}\left(\frac{\psi(Z)}{\frac{1}{n_2}\chi^2_{\frac{\alpha}{2}}(n_2)}-\sigma_0^2\right)\right).
\end{aligned}
$$

Since the loss is positive, the coverage property of CI_α^i is the same with that of $\bar{\mathrm{CI}}_\alpha^i$, for $i = 1, 2, 3, 4$. We also have

$$\mathbb{E}L\left(\mathrm{CI}_\alpha^i\right) \leq \mathbb{E}L\left(\bar{\mathrm{CI}}_\alpha^i\right), \text{ for } i = 1, 2, 3, 4. \tag{B.4.15}$$

On the event $\bar{\mathcal{G}}_8$, we have $\psi(Z) = \frac{1}{n_2}\|y^{(2)} - X^{(2)}\widehat{\beta}^L\|_2^2$ and hence

$$\begin{aligned}
&\mathbb{P}_\theta\left(\|\widehat{\beta}^L - \beta\|_2^2 \in \bar{\mathrm{CI}}_\alpha^1(Z)\right) \geq \mathbb{P}_\theta\left(\left\{\chi^2_{\frac{\alpha}{2}}(n_2) \leq \frac{\|y^{(2)} - X^{(2)}\widehat{\beta}^L\|_2^2}{\|\beta - \widehat{\beta}^L\|_2^2 + \sigma_0^2} \leq \chi^2_{1-\frac{\alpha}{2}}(n_2)\right\} \cap \bar{\mathcal{G}}_8\right) \\
\geq &\mathbb{P}_\theta\left(\left\{\chi^2_{\frac{\alpha}{2}}(n_2) \leq \frac{\|y^{(2)} - X^{(2)}\widehat{\beta}^L\|_2^2}{\|\beta - \widehat{\beta}^L\|_2^2 + \sigma_0^2 \leq \chi^2_{1-\frac{\alpha}{2}}(n_2)}\right\}\right) + \mathbb{P}_\theta\left(\bar{\mathcal{G}}_8\right) - 1 \\
\geq &1 - \alpha - \exp(-c'n) - cp^{-c},
\end{aligned}$$

where the last inequality follows from (B.4.13) and Lemma 24. The coverage property (3.3.10) in Chapter 3 over $\Theta_0(k)$ follows from taking infimum over both sides of above inequality.

To establish the coverage property in Proposition 5, we introduce the following lemma, which establishes an upper bound for $\|a\|_q^q$ where $1 \leq q \leq 2$ and $a \in \mathbb{R}^p$.

Lemma 27.

$$\|a\|_q^q \leq \left(\sum_{j=1}^p |a_j|\right)^{2-q}\left(\sum_{j=1}^p a_j^2\right)^{q-1}. \tag{B.4.16}$$

The above lemma is established in the proof of Theorem 7.1 in Bickel et al. (2009). We introduce the following confidence interval, Define the events

$$\begin{aligned}
\bar{\mathcal{G}}_6' &= \left\{\|\widehat{\beta}^L - \beta\|_2^2 \leq C_2^*(A, k_1)\, k_1 \frac{\log p}{n}\right\}, \\
\bar{\mathcal{G}}_8' &= \left\{\frac{1}{n_2}\left\|y^{(2)} - X^{(2)}\widehat{\beta}^L\right\|_2^2 \leq \sigma_0^2\left(1 + \frac{k_1 \log p}{n_1}\right)\left(1 + 2\sqrt{\frac{\log p}{n_2}} + \frac{2\log p}{n_2}\right)\right\},
\end{aligned}$$

and similar to Lemma 24, we have

$$\min_{\theta\in\Theta_0(k_1)} \mathbb{P}_\theta(\bar{\mathcal{G}}_6' \cap \bar{\mathcal{G}}_8') \geq 1 - c\exp\left(-c'n\right) - cp^{1-\frac{A^2}{8}} - cp^{-2}. \tag{B.4.17}$$

Note that

$$\|\widehat{\beta}^L - \beta\|_2^2 \leq \|\widehat{\beta}^L - \beta\|_q^2, \quad \text{for } 1 \leq q < 2. \tag{B.4.18}$$

For the Lasso estimator $\widehat{\beta}^L$ with $A > 4\sqrt{2}$, let S denote the support of β, then on the event $\bar{\mathcal{G}}_3$,

$$\|\widehat{\beta}^L - \beta\|_1^2 \leq 16\|(\widehat{\beta}^L - \beta)_S\|_1^2 \leq 16k_2\|\widehat{\beta}^L - \beta\|_2^2. \tag{B.4.19}$$

By Lemma 27, on the event $\bar{\mathcal{G}}_3$, we have $\|\widehat{\beta}^L - \beta\|_q^2 \leq (16k_2)^{\frac{2}{q}-1}\|\widehat{\beta}^L - \beta\|_2^2$. Combined with (B.4.4) and (B.4.18), we establish the coverage property (3.3.20) in Chapter 3.

The proof of coverage properties in Proposition 10 is a generalization of those in Propositions 3 and 5. We also define the following extra event to facilitate the discussion,

$$\bar{\mathcal{G}}_9 = \left\{\max\left\{\left|\frac{\lambda_{\min}(\widehat{\Omega})}{\lambda_{\min}(\Omega)} - 1\right|, \left|\frac{\lambda_{\max}(\widehat{\Omega})}{\lambda_{\max}(\Omega)} - 1\right|\right\} \leq 0.01\right\}. \tag{B.4.20}$$

By Theorem 1 in Cai et al. (2011), with a proper chosen tuning parameter, we have $\mathbb{P}_\theta(\bar{\mathcal{G}}_9) \geq 1 - p^{-2}$. On the event $\bar{\mathcal{G}}_8$, we have $\psi\left(Z\right) = \frac{1}{n_2}\|y^{(2)} - X^{(2)}\widehat{\beta}\|_2^2$. On the event $\bar{\mathcal{G}}_9$, we have $\lambda_{\max} = \lambda_{\max}(\widehat{\Omega})$ and $\lambda_{\min} = \lambda_{\min}(\widehat{\Omega})$ and then

$$\begin{aligned}
&\mathbb{P}_\theta\left(\|\widehat{\beta} - \beta\|_2^2 \in \bar{\mathrm{CI}}_\alpha^3\left(Z\right)\right)\\
\geq&\mathbb{P}_\theta\left(\left\{\chi^2_{\frac{\alpha}{2}}\left(n_2\right) \leq \frac{\|y^{(2)} - X^{(2)}\widehat{\beta}\|_2^2}{\|\Sigma^{1/2}(\beta - \widehat{\beta})\|_2^2 + \sigma_0^2} \leq \chi^2_{1-\frac{\alpha}{2}}\left(n_2\right)\right\} \cap \bar{\mathcal{G}}_8 \cap \bar{\mathcal{G}}_9\right)\\
\geq&\mathbb{P}_\theta\left(\left\{\chi^2_{\frac{\alpha}{2}}\left(n_2\right) \leq \frac{\|y^{(2)} - X^{(2)}\widehat{\beta}\|_2^2}{\|\Sigma^{1/2}(\beta - \widehat{\beta})\|_2^2 + \sigma_0^2} \leq \chi^2_{1-\frac{\alpha}{2}}\left(n_2\right)\right\}\right) + \mathbb{P}_\theta\left(\bar{\mathcal{G}}_8 \cap \bar{\mathcal{G}}_9\right) - 1\\
\geq&1 - \alpha - cp^{-c}.
\end{aligned}$$

The coverage property (B.2.10) over $\Theta_{\sigma_0}(k,s)$ follows from taking infimum over both sides of above inequality.

Combining (B.2.10) and the fact that with probability greater than $1-p^{-\delta}$,

$$\|\widehat{\beta}-\beta\|_2^2 \leq \|\widehat{\beta}-\beta\|_q^2 \leq \left((1+c^*)^2 k_2\right)^{\frac{2}{q}-1} \|\widehat{\beta}-\beta\|_2^2, \quad \text{for } 1\leq q<2, \tag{B.4.21}$$

we establish the coverage property (B.2.12).

Proof of expected lengths in Propositions 3, 5 and 10

In the following, we control expected lengths of confidence intervals $\bar{\mathrm{CI}}_\alpha^i$ for $i=1,2,3,4$. By (B.4.15), these upper bounds are also upper bounds for expected lengths of CI_α^i for $i=1,2,3,4$.

We start with the expected length of $\bar{\mathrm{CI}}_\alpha^1$ over $\Theta_0(k)$. Note that

$$\frac{\psi(Z)}{\frac{1}{n_2}\chi^2_{\frac{\alpha}{2}}(n_2)} - \frac{\psi(Z)}{\frac{1}{n_2}\chi^2_{1-\frac{\alpha}{2}}(n_2)} = \psi(Z)\, f(n_2),$$

where $f(n_2)=\left(\frac{\frac{1}{n_2}\chi^2_{1-\frac{\alpha}{2}}(n_2)-\frac{1}{n_2}\chi^2_{\frac{\alpha}{2}}(n_2)}{\frac{1}{n_2}\chi^2_{\frac{\alpha}{2}}(n_2)\frac{1}{n_2}\chi^2_{1-\frac{\alpha}{2}}(n_2)}\right)$. On the event $\bar{\mathcal{G}}_8$, we have

$$\psi(Z)=\frac{1}{n_2}\|y^{(2)}-X^{(2)}\widehat{\beta}^L\|_2^2$$

and then obtain

$$\begin{aligned}
&\mathbb{E}\left|\frac{\psi(Z)}{\frac{1}{n_2}\chi^2_{\frac{\alpha}{2}}(n_2)} - \frac{\psi(Z)}{\frac{1}{n_2}\chi^2_{1-\frac{\alpha}{2}}(n_2)}\right| \mathbf{1}_{\bar{\mathcal{G}}_6\cap\bar{\mathcal{G}}_8} = \left(\mathbb{E}\frac{1}{n_2}\|y^{(2)}-X^{(2)}\widehat{\beta}^L\|_2^2\mathbf{1}_{\bar{\mathcal{G}}_6\cap\bar{\mathcal{G}}_8}\right) f(n_2)\\
&\leq \left(\mathbb{E}\frac{1}{n_2}\|y^{(2)}-X^{(2)}\widehat{\beta}^L\|_2^2\mathbf{1}_{\bar{\mathcal{G}}_6}\right) f(n_2).
\end{aligned}$$

Conditioning on $(X^{(1)},y^{(1)})$, by taking expectation of right hand side of above equation with respect to $(X^{(2)}, y^{(2)})$, the right hand side is equal to

$$\mathbb{E}\left(\left(\|\beta-\widehat{\beta}^{L}\|_2^2+\sigma_0^2\right)\mathbf{1}_{\bar{\mathcal{G}}_6}\right)f(n_2)\leq\left(Ck\frac{\log p}{n}+1\right)\sigma_0^2 f(n_2).$$

Based on Lemma 26, we have $1-2\sqrt{\frac{2\log\frac{4}{\alpha}}{n}}-4\frac{\log\frac{4}{\alpha}}{n}\leq\frac{1}{n_2}\chi^2_{\frac{\alpha}{2}}(n_2)\leq\frac{1}{n_2}\chi^2_{1-\frac{\alpha}{2}}(n_2)\leq 1+2\sqrt{\frac{2\log\frac{4}{\alpha}}{n}}+4\frac{\log\frac{4}{\alpha}}{n}$ and hence $f(n_2)\leq\frac{C}{\sqrt{n}}$. Hence, we have

$$\mathbb{E}\left|\frac{\psi(Z)}{\frac{1}{n_2}\chi^2_{\frac{\alpha}{2}}(n_2)}-\frac{\psi(Z)}{\frac{1}{n_2}\chi^2_{1-\frac{\alpha}{2}}(n_2)}\right|\mathbf{1}_{\bar{\mathcal{G}}_6\cap\bar{\mathcal{G}}_8}\leq\frac{C}{\sqrt{n}}\left(Ck\frac{\log p}{n}+1\right)\sigma_0^2. \tag{B.4.22}$$

Note that

$$\begin{aligned}&\mathbb{E}\left|\frac{\psi(Z)}{\frac{1}{n_2}\chi^2_{\frac{\alpha}{2}}(n_2)}-\frac{\psi(Z)}{\frac{1}{n_2}\chi^2_{1-\frac{\alpha}{2}}(n_2)}\right|\mathbf{1}_{\left(\bar{\mathcal{G}}_6\cap\bar{\mathcal{G}}_8\right)^c}\leq\sigma_0^2\log p\times f(n_2)\mathbb{P}_\theta\left(\left(\bar{\mathcal{G}}_6\cap\bar{\mathcal{G}}_8\right)^c\right)\\&\leq\frac{1}{\sqrt{n}}\sigma_0^2\log p\left(cp^{-c}+c\exp\left(-c'n\right)\right).\end{aligned}$$

Combined with (B.4.22), we establish $\mathbb{E}L\left(\bar{\mathrm{CI}}^1_\alpha\right)\lesssim\frac{1}{\sqrt{n}}\sigma_0^2$ and hence (3.3.11) in Chapter 3.

To control the expected length of $\bar{\mathrm{CI}}^2_\alpha(Z,k_2,q)$ over $\Theta_0(k_1)$, we decompose the length as

$$L\left(\bar{\mathrm{CI}}^2_\alpha(Z,k_2,q)\right)=L\left(\bar{\mathrm{CI}}^1_\alpha(Z)\right)+\left((16k_2)^{\frac{2}{q}-1}-1\right)\left(\frac{\psi(Z)}{\frac{1}{n_2}\chi^2_{\frac{\alpha}{2}}(n_2)}-\sigma_0^2\right). \tag{B.4.23}$$

Since we have established $\mathbb{E}L\left(\bar{\mathrm{CI}}^1_\alpha\right)\lesssim\frac{1}{\sqrt{n}}\sigma_0^2$, it is sufficient to control the term

$\mathbb{E}\left|\frac{\psi(Z)}{\frac{1}{n_2}\chi^2_{\frac{\alpha}{2}}(n_2)}-\sigma_0^2\right|$. On the event $\bar{\mathcal{G}}_8'$, we have $\psi\left(Z\right)=\frac{1}{n_2}\|y^{(2)}-X^{(2)}\widehat{\beta}^L\|_2^2$ and

$$\begin{aligned}
&\frac{\psi\left(Z\right)}{\frac{1}{n_2}\chi^2_{\frac{\alpha}{2}}\left(n_2\right)}-\sigma_0^2=\frac{\frac{1}{n_2}\|y^{(2)}-X^{(2)}\widehat{\beta}^L\|_2^2-\sigma_0^2-\left(\frac{1}{n_2}\chi^2_{\frac{\alpha}{2}}\left(n_2\right)-1\right)\sigma_0^2}{\frac{1}{n_2}\chi^2_{\frac{\alpha}{2}}\left(n_2\right)}\\
&=\frac{\frac{1}{n_2}\|y^{(2)}-X^{(2)}\widehat{\beta}^L\|_2^2-\sigma_0^2-\|\widehat{\beta}^L-\beta\|_2^2}{\frac{1}{n_2}\chi^2_{\frac{\alpha}{2}}\left(n_2\right)}-\frac{\left(\frac{1}{n_2}\chi^2_{\frac{\alpha}{2}}\left(n_2\right)-1\right)\sigma_0^2-\|\widehat{\beta}^L-\beta\|_2^2}{\frac{1}{n_2}\chi^2_{\frac{\alpha}{2}}\left(n_2\right)}.
\end{aligned}$$

Hence, we have

$$\begin{aligned}
&\mathbb{E}\left|\frac{\psi\left(Z\right)}{\frac{1}{n_2}\chi^2_{\frac{\alpha}{2}}\left(n_2\right)}-\sigma_0^2\right|\mathbf{1}_{\bar{\mathcal{G}}_6'\cap\bar{\mathcal{G}}_8'}\leq\mathbb{E}\left|\frac{\frac{1}{n_2}\|y^{(2)}-X^{(2)}\widehat{\beta}^L\|_2^2-\sigma_0^2-\|\widehat{\beta}^L-\beta\|_2^2}{\frac{1}{n_2}\chi^2_{\frac{\alpha}{2}}\left(n_2\right)}\right|\mathbf{1}_{\bar{\mathcal{G}}_6'}\\
&+\mathbb{E}\frac{\left|\frac{1}{n_2}\chi^2_{\frac{\alpha}{2}}\left(n_2\right)-1\right|\sigma_0^2+\|\widehat{\beta}^L-\beta\|_2^2}{\frac{1}{n_2}\chi^2_{\frac{\alpha}{2}}\left(n_2\right)}\mathbf{1}_{\bar{\mathcal{G}}_6'}\leq C\left(\frac{1}{\sqrt{n}}+k_1\frac{\log p}{n}\right)\sigma_0^2,
\end{aligned}\tag{B.4.24}$$

and

$$\begin{aligned}
&\mathbb{E}\left|\frac{\psi\left(Z\right)}{\frac{1}{n_2}\chi^2_{\frac{\alpha}{2}}\left(n_2\right)}-\sigma_0^2\right|\mathbf{1}_{\left(\bar{\mathcal{G}}_6'\cap\bar{\mathcal{G}}_8'\right)^c}\leq\sigma_0^2\log p\,\mathbb{P}_\theta\left(\left(\bar{\mathcal{G}}_6'\cap\bar{\mathcal{G}}_8'\right)^c\right)\\
&\leq C\sigma_0^2\log p\left(cp^{1-\frac{A^2}{8}}+cp^{-2}+c\exp\left(-c'n\right)\right)\leq C\left(\frac{k_1\log p}{n}+\frac{1}{\sqrt{n}}\right)\sigma_0^2,
\end{aligned}\tag{B.4.25}$$

where the last inequality follows from the inequality (B.4.17) and the fact that $A>4\sqrt{2}$ and $n\leq p$. The control of length (3.3.21) in Chapter 3 follows from (B.4.23), (B.4.24) and (B.4.25).

In the following, we control the expected lengths of $\bar{\mathrm{CI}}_\alpha^3$ and $\bar{\mathrm{CI}}_\alpha^4$ over $\Theta_{\sigma_0}(k,s)$. Applying the similar argument as (B.4.24) and (B.4.25), we have

$$\mathbb{E}\lambda_{\max}\left|\frac{\psi\left(Z\right)}{\frac{1}{n_2}\chi^2_{\frac{\alpha}{2}}\left(n_2\right)}-\sigma_0^2\right|\mathbf{1}_{\bar{\mathcal{G}}_6'\cap\bar{\mathcal{G}}_8'\cap\bar{\mathcal{G}}_9}\leq C\left(\frac{1}{\sqrt{n}}+k_1\frac{\log p}{n}\right)\sigma_0^2,$$

and

$$\mathbb{E}\lambda_{\max}\left|\frac{\psi\left(Z\right)}{\frac{1}{n_2}\chi^2_{\frac{\alpha}{2}}\left(n_2\right)}-\sigma_0^2\right|\mathbf{1}_{\left(\bar{\mathcal{G}}_6'\cap\bar{\mathcal{G}}_8'\cap\bar{\mathcal{G}}_9\right)^c}\leq\sigma_0^2\log p^2\mathbb{P}_\theta\left(\left(\bar{\mathcal{G}}_6'\cap\bar{\mathcal{G}}_8'\right)^c\right)$$
$$\leq C\sigma_0^2\log p\left(cp^{-\delta}+cp^{-2}+c\exp\left(-c'n\right)\right)\leq C\left(\frac{k_1\log p}{n}+\frac{1}{\sqrt{n}}\right)\sigma_0^2,$$

where the last inequality follows from the inequality (B.4.17) and the fact that $\delta>2$ and $n\leq p$. The above two inequalities lead to the control of expected lengths in (B.2.11) and (B.2.13).

B.4.5 Proof of Proposition 7

We will consider the estimators $\widehat{\beta}$ satisfying Assumption (A) introduced in (3.5.1) in Chapter 3. The minimax lower bounds (3.2.9) in Theorem 9, (3.2.13) in Theorem 10 and the minimax lower bound $\frac{k\log p}{n}$ of (3.2.8) in Theorem 9 in Chapter 3 can be achieved by the trivial estimator 0.

For estimators $\widehat{\beta}$ constructed using the subsample $Z^{(1)}=\left(y^{(1)},X^{(1)}\right)$, the minimax lower bound $\frac{1}{\sqrt{n}}\sigma_0^2$ of (3.2.8) in Theorem 9 in Chapter 3 can be achieved by the estimator as defined in (3.2.11) in Chapter 3, $\widetilde{L}_2=\left(\frac{1}{n_2}\left\|y^{(2)}-X^{(2)}\widehat{\beta}\right\|_2^2-\sigma_0^2\right)_+$. Applying the proof of Proposition 2, we can establish, for any sequence $\delta_{n,p}\to\infty$,

$$\limsup_{n,p\to\infty}\sup_{\theta\in\Theta_0(k)}\mathbb{P}_\theta\left(\left|\widetilde{L}_2-\|\widehat{\beta}-\beta\|_2^2\right|\geq\delta_{n,p}\frac{1}{\sqrt{n}}\right)=0.$$

The minimax lower bound $\frac{k\log p}{n}\sigma_0^2$ in (3.3.5) in Theorem 11 in Chapter 3 can be achieved by the confidence interval $(0,C^*k\frac{\log p}{n}\sigma_0^2)$. The minimax lower bound $\frac{k_2\log p}{n}\sigma_0^2$ of (3.3.7) in Theorem 12 in Chapter 3 can be achieved by the confidence interval $(0,C^*k_2\frac{\log p}{n}\sigma_0^2)$; For estimators $\widehat{\beta}$ constructed using the subsample $Z^{(1)}=\left(y^{(1)},X^{(1)}\right)$, the minimax lower bound $\frac{1}{\sqrt{n}}\sigma_0^2$ of (3.3.5) in Theorem 11 in Chapter 3 and (3.3.7) in Theorem 12 in Chapter 3 can be achieved by the confidence interval $\mathrm{CI}_\alpha^1\left(Z\right)$ as defined in (3.3.8) in Chapter 3 with $\psi(Z) = \min\left\{\frac{1}{n_2}\left\|y^{(2)} - X^{(2)}\widehat{\beta}\right\|_2^2, \sigma_0^2 \log p\right\}$. Applying the proof of Proposition 3, we can establish that the constructed confidence interval satisfies

$$\liminf_{n,p\to\infty} \inf_{\theta\in\Theta_0(k)} \mathbb{P}\left(\|\widehat{\beta} - \beta\|_2^2 \in \mathrm{CI}_\alpha^1(Z)\right) \geq 1 - \alpha,$$

and

$$\mathbf{R}\left(\mathrm{CI}_\alpha^1(Z), \Theta_0(k)\right) \lesssim \frac{1}{\sqrt{n}}\sigma_0^2.$$

The minimax lower bound (3.3.13) in Theorem 13 in Chapter 3 can be achieved by the confidence interval $(0, C^* k^{\frac{2}{q}} \frac{\log p}{n}\sigma_0^2)$. The minimax lower bound $ck_2^{\frac{2}{q}} \frac{\log p}{n}\sigma_0^2$ of (3.3.18) in Theorem 14 in Chapter 3 can be achieved by the confidence interval $(0, C^* k_2^{\frac{2}{q}} \frac{\log p}{n}\sigma_0^2)$.

The minimax lower bounds $ck_2^{\frac{2}{q}-1}\frac{1}{\sqrt{n}}\sigma_0^2$ of (3.3.18) and $ck_2^{\frac{2}{q}-1}k_1\frac{\log p}{n}\sigma_0^2$ of (3.3.18) in Theorem 14 in Chapter 3 can be achieved by the confidence interval $\mathrm{CI}_\alpha^2(Z, k_2, q)$ as defined in (3.3.19) in Chapter 3

$$\left(\left(\frac{\psi(Z)}{\frac{1}{n_2}\chi^2_{1-\frac{\alpha}{2}}(n_2)} - \sigma_0^2\right)_+, \left((1+c^*)^2 k_2\right)^{\frac{2}{q}-1}\left(\frac{\psi(Z)}{\frac{1}{n_2}\chi^2_{\frac{\alpha}{2}}(n_2)} - \sigma_0^2\right)_+\right),$$

with $\psi(Z) = \min\left\{\frac{1}{n_2}\left\|y^{(2)} - X^{(2)}\widehat{\beta}\right\|_2^2, \sigma_0^2 \log p\right\}$. Applying the proof of Proposition 5, we can establish that

$$\liminf_{n,p\to\infty} \inf_{\theta\in\Theta_0(k_2)} \mathbb{P}_\theta\left(\|\widehat{\beta} - \beta\|_q^2 \in \mathrm{CI}_\alpha^2(Z, k_2, q)\right) \geq 1 - \alpha,$$

and

$$\mathbf{R}\left(\mathrm{CI}_\alpha^2(Z, k_2, q), \Theta_0(k_1)\right) \lesssim k_2^{\frac{2}{q}-1}\left(k_1\frac{\log p}{n} + \frac{1}{\sqrt{n}}\right)\sigma_0^2.$$

Note that the proof of Proposition 5 requires the following conditions, $\widehat{\beta}$ is constructed based on the subsample $Z^{(1)} = \left(y^{(1)}, X^{(1)}\right)$ and it satisfies Assumption (A) and (3.5.2) in Chapter 3 with $\delta > 2$.

The minimax lower bound (3.4.1) in Theorem 15 in Chapter 3 can be achieved by the confidence interval $\mathrm{CI}_{\alpha}\left(Z,k,q\right)$ as defined in (3.4.4) in Chapter 3 with $\varphi\left(Z,k,q\right) = C^{*}k^{\frac{2}{q}}\frac{\log p}{n}\widehat{\sigma}^{2}$.

B.5 Proof of extra lemmas

In this section, we prove Lemma 7, 8, 18, 21, 23 and 24.

B.5.1 Proof of Lemma 7

The equality (3.9.2) in Chapter 3 follows from

$$\begin{aligned}\mathbb{P}_{Z,\theta\sim\pi}\left(Z\in\mathcal{A}\right) &= \int\int \mathbf{1}_{z\in\mathcal{A}} f_{\theta}\left(z\right)\pi\left(\theta\right)dzd\theta = \int \mathbf{1}_{z\in\mathcal{A}}\left(\int f_{\theta}\left(z\right)\pi\left(\theta\right)d\theta\right)dz \\ &= \int \mathbf{1}_{z\in\mathcal{A}} f_{\pi}\left(z\right)dz = \mathbb{P}_{\pi}\left(Z\in\mathcal{A}\right),\end{aligned}$$

where the second equality follows from the Fubini's theorem.

The inequality (3.9.3) in Chapter 3 follows from

$$\begin{aligned}\left|\mathbb{P}_{\pi_1}\left(Z\in\mathcal{A}\right)-\mathbb{P}_{\pi_2}\left(Z\in\mathcal{A}\right)\right| &= \left|\int \mathbf{1}_{z\in\mathcal{A}} f_{\pi_1}\left(z\right)dz - \int \mathbf{1}_{z\in\mathcal{A}} f_{\pi_2}\left(z\right)dz\right| \\ &\leq \int \left|f_{\pi_1}\left(z\right)-f_{\pi_2}\left(z\right)\right|dz,\end{aligned}$$

where the equality follows from the definition of $\mathbb{P}_{\pi_i}$ and the inequality follows from the triangle inequality.

B.5.2 Proof of Lemma 18

In the following, we prove Lemma 18 in Chapter 3. The following lemmas are useful in controlling the χ^2 distance between the null and the alternative hypothesis. The first lemma is established in Cai & Zhou (2012); Ren et al. (2013).

Lemma 28. *Let g_i be the density function of $N(0,\Sigma_i)$ for $i=0,1,2$, respectively. Then*

$$\int \frac{g_1 g_2}{g_0} = \left(\det\left(I-\Sigma_0^{-1}\left(\Sigma_1-\Sigma_0\right)\Sigma_0^{-1}\left(\Sigma_2-\Sigma_0\right)\right)\right)^{-\frac{1}{2}}.$$

Lemma 29. *Let $\boldsymbol{\delta}$ and $\widetilde{\boldsymbol{\delta}}$ be independent random variables with the prior distribution π, then*

$$\begin{aligned}&\chi^2\left(f_\pi, f_0\right)+1\\ =&\int\left(\det\left(\mathrm{I}-\left(\Sigma_0^z\right)^{-1}\left(\Sigma_{\widetilde{\boldsymbol{\delta}}}^z-\Sigma_0^z\right)\left(\Sigma_0^z\right)^{-1}\left(\Sigma_{\boldsymbol{\delta}}^z-\Sigma_0^z\right)\right)\right)^{-\frac{n}{2}}\pi\left(\boldsymbol{\delta}\right)\pi\left(\widetilde{\boldsymbol{\delta}}\right)d\boldsymbol{\delta}d\widetilde{\boldsymbol{\delta}}.\end{aligned}\tag{B.5.1}$$

The proof of the above lemma can be found in Cai & Guo (2016b). We first introduce the covariance matrix of (y_i, X_i) corresponding to the parameter θ_1,

$$\Sigma_1^z=\left(\begin{array}{c|c|c}\|\beta^*\|_2^2+\|\boldsymbol{\delta}\|_2^2+\sigma_0^2 & (\beta_S^*)^\intercal & \mathbf{0}_{1\times p_1}\\ \hline \beta_S^* & \mathrm{I}_{k_0\times k_0} & \mathbf{0}_{k_0\times p_1}\\ \hline \mathbf{0}_{p_1\times 1} & \mathbf{0}_{p_1\times k_0} & \mathrm{I}_{p_1\times p_1}\end{array}\right).\tag{B.5.2}$$

Since Σ_1^z only depends on $\boldsymbol{\delta}$ through the ℓ_2 norm $\|\boldsymbol{\delta}\|_2^2=(k_2-k_0)\rho^2$, Σ_1^z is fixed for a given ρ. To control $\chi^2\left(f_\pi, f_{\theta_1}\right)$, we have the following expression for the main term of (B.5.1),

$$\left(\Sigma_1^z\right)^{-1}\left(\Sigma_{\boldsymbol{\delta}}^z-\Sigma_1^z\right)\left(\Sigma_1^z\right)^{-1}\left(\Sigma_{\widetilde{\boldsymbol{\delta}}}^z-\Sigma_1^z\right)=\left(\begin{array}{c|c|c}\frac{1}{\|\boldsymbol{\delta}\|_2^2+\sigma_0^2}\boldsymbol{\delta}^\intercal\widetilde{\boldsymbol{\delta}} & \mathbf{0}_{1\times k_0} & \mathbf{0}_{1\times p_1}\\ \hline -\frac{1}{\|\boldsymbol{\delta}\|_2^2+\sigma_0^2}\beta_S^*\boldsymbol{\delta}^\intercal\widetilde{\boldsymbol{\delta}} & \mathbf{0}_{k_0\times k_0} & \mathbf{0}_{k_0\times p_1}\\ \hline \mathbf{0}_{p_1\times 1} & \mathbf{0}_{p_1\times k_0} & \frac{1}{\|\widetilde{\boldsymbol{\delta}}\|_2^2+\sigma_0^2}\boldsymbol{\delta}\widetilde{\boldsymbol{\delta}}^\intercal\end{array}\right).$$

By applying Lemma 29, we have

$$\chi^2\left(f_\pi, f_{\theta_1}\right) + 1 = \mathbb{E}_{\boldsymbol{\delta},\widetilde{\boldsymbol{\delta}}}\left(1 - \frac{\boldsymbol{\delta}^\intercal \widetilde{\boldsymbol{\delta}}}{\|\boldsymbol{\delta}\|_2^2 + \sigma_0^2}\right)^{-\frac{n}{2}} \left(1 - \frac{\boldsymbol{\delta}^\intercal \widetilde{\boldsymbol{\delta}}}{\|\widetilde{\boldsymbol{\delta}}\|_2^2 + \sigma_0^2}\right)^{-\frac{n}{2}}.$$

To control $\chi^2\left(f_{\theta_1}, f_{\theta_0}\right)$, we have the following expression for the main term of (B.5.1),

$$\left(\Sigma_0^z\right)^{-1}\left(\Sigma_1^z - \Sigma_0^z\right)\left(\Sigma_0^z\right)^{-1}\left(\Sigma_1^z - \Sigma_0^z\right) = \left(\begin{array}{c|c|c} \frac{\|\boldsymbol{\delta}\|_2^2\|\widetilde{\boldsymbol{\delta}}\|_2^2}{\sigma_0^4} & \frac{\|\boldsymbol{\delta}\|_2^2\|\widetilde{\boldsymbol{\delta}}\|_2^2}{\sigma_0^4}(\beta_S^*)^\intercal & \mathbf{0}_{1\times p_1} \\ \hline \mathbf{0}_{k_0\times 1} & \mathbf{0}_{k_0\times k_0} & \mathbf{0}_{k_0\times p_1} \\ \hline \mathbf{0}_{p_1\times 1} & \mathbf{0}_{p_1\times k_0} & \mathbf{0}_{p_1\times p_1} \end{array}\right).$$

By applying Lemma 29, we have

$$\chi^2\left(f_{\theta_1}, f_{\theta_0}\right) + 1 = \mathbb{E}_{\boldsymbol{\delta},\widetilde{\boldsymbol{\delta}}}\left(1 - \frac{\|\boldsymbol{\delta}\|_2^2\|\widetilde{\boldsymbol{\delta}}\|_2^2}{\sigma_0^4}\right)^{-\frac{n}{2}}.$$

B.5.3 Proof of Lemma 23

By the normalization (B.4.1) in Chapter 3, the scaled Lasso algorithm can be expressed as

$$\{\widehat{d}, \hat{\sigma}\} = \underset{d\in\mathbb{R}^p, \sigma\in\mathbb{R}}{\arg\min} \frac{\|y - Wd\|_2^2}{2n\sigma} + \frac{\sigma}{2} + \lambda_0 \sum_{j=1}^{p} |d_j|. \tag{B.5.3}$$

The following lemma from an arXiv version of Ren et al. (2013) is useful to control the estimation of the noise level σ.

Lemma 30. *Let $\{\widehat{d}, \hat{\sigma}\}$ be the solution of the scaled Lasso* (B.5.3)*. For any $\epsilon_0 > 1$, on the event $\mathcal{S}_1 = \left\{\frac{\|W^\intercal \epsilon\|_\infty}{n} \leq \sigma^{ora}\lambda_0 \frac{\epsilon_0 - 1}{\epsilon_0 + 1}(1-\tau)\right\}$, we have*

$$\left|\frac{\hat{\sigma}}{\sigma^{ora}} - 1\right| \leq \tau. \tag{B.5.4}$$

where τ is defined in (B.4.3).

For fixed μ, we also define $\widehat{d}\,(\mu) = \arg\min_{d\in\mathbb{R}^p} \frac{\|y-Wd\|_2^2}{2n} + \mu\sum_{j=1}^{p}|d_j|$. Note that

$$\widehat{d} = \widehat{d}\,(\lambda_0\widehat{\sigma}) \quad \text{and} \quad \widehat{d}_j = \widehat{\beta}^{SL}_j \frac{\|X_{\cdot j}\|_2}{\sqrt{n}} \quad \text{for } j \in [p]. \tag{B.5.5}$$

Setting $A > 2\sqrt{2}$, on event $\mathcal{G}\cap\mathcal{S}$, we have the following events

$$\mathcal{B}_1 = \left\{\frac{1}{n}\|W^{\intercal}W\left(\widehat{d}-d\right)\|_\infty \leq \lambda_0\widehat{\sigma}\right\} \quad \text{and} \quad \mathcal{B}_2 = \left\{2\sqrt{\frac{2\log p}{n}}\sigma_0 \leq \lambda_0\widehat{\sigma}\right\}. \tag{B.5.6}$$

Based on the proof of Theorem 7.2 in Bickel et al. (2009), on the event $\mathcal{G}\cap\mathcal{S}$, we have

$$\|\widehat{\beta}^{SL}-\beta\|_q^2 \leq \frac{\max\|X_{\cdot j}\|_2^2}{n}\left(\frac{16A\widehat{\sigma}}{\kappa^2(W,k,k,3)}\right)^2 k^{\frac{2}{q}}\frac{\log p}{n}.$$

Similar to the proof of Lemma 13 in Cai & Guo (2016b), we have

$$\kappa^2\left(W,k,k,3\right) \geq \frac{n}{\max\|X_{\cdot j}\|_2^2}\kappa^2\left(X,k,k,3\left(\frac{\max\|X_{\cdot j}\|_2}{\min\|X_{\cdot j}\|_2}\right)\right), \tag{B.5.7}$$

and then establish (B.4.6). By the definitions of $\mathcal{G}_1$ and $\mathcal{G}_3$, we establish (B.4.7).

B.5.4 Proof of Lemma 24

The control of events $\bar{\mathcal{G}}_1, \bar{\mathcal{G}}_2, \bar{\mathcal{G}}_3$ and $\bar{\mathcal{G}}_4$ is similar to that of Lemma 22 and will be omitted here. We will present the proofs of $\mathbb{P}_\theta(\bar{\mathcal{G}}_5)$ and $\mathbb{P}_\theta(\bar{\mathcal{G}}_6)$. We establish $\mathbb{P}_\theta(\bar{\mathcal{G}}_5) \geq \mathbb{P}_\theta(\bar{\mathcal{G}}_1\cap\bar{\mathcal{G}}_2\cap\bar{\mathcal{G}}_3)$ by showing that $\bar{\mathcal{G}}_5$ holds on the event $\bar{\mathcal{G}}_1\cap\bar{\mathcal{G}}_2\cap\bar{\mathcal{G}}_3$. Based on the proof of Theorem 7.2 in Bickel et al. (2009) and (B.5.7), on the event $\bar{\mathcal{G}}_1\cap\bar{\mathcal{G}}_2\cap\bar{\mathcal{G}}_3$, we have

$$\|\widehat{\beta}^{L}-\beta\|_q^2 \leq \frac{\left(22A\sigma_0\right)^2}{\left(\frac{1}{4}-42\sqrt{\frac{2k\log p}{n_1}}\right)^4}k^{\frac{2}{q}}\frac{\log p}{n}.$$

For the case $q=2$, we can establish $\|\widehat{\beta}^{L}-\beta\|_2^2$ for the case $A>\sqrt{2}$ by applying the finer results established in Ye & Zhang (2010). By Theorem 3 and (27) in Ye &

Zhang (2010), on the event $\bar{\mathcal{G}}_1 \cap \bar{\mathcal{G}}_2 \cap \bar{\mathcal{G}}_4$, we have

$$\|\widehat{\beta}^L - \beta\|_2^2 \leq \frac{\left(\frac{3\eta_0}{\eta_0+1} A\sigma_0\right)^2}{\left(\frac{1}{4} - (9+11\eta_0)\sqrt{\frac{2k\log p}{n_1}}\right)^4} k \frac{\log p}{n}.$$

Hence, $\mathbb{P}_\theta(\bar{\mathcal{G}}_6) \geq \mathbb{P}_\theta(\bar{\mathcal{G}}_1 \cap \bar{\mathcal{G}}_2 \cap \bar{\mathcal{G}}_4)$. Let $\mathbb{P}_\theta(\cdot|(X^{(1)}, y^{(1)}))$ denote the conditional probability of $(X^{(2)}, y^{(2)})$ on $(X^{(1)}, y^{(1)})$. Note that conditioning on $\left(y^{(1)}, X^{(1)}\right)$, we have $y_i^{(2)} - X_i^{(2)}\widehat{\beta}^L = X_i^{(2)}\left(\beta - \widehat{\beta}^L\right) + \epsilon_i^{(2)} \sim N\left(0, \|\beta - \widehat{\beta}^L\|_2^2 + \sigma_0^2\right)$ and $\|y^{(2)} - X^{(2)}\widehat{\beta}^L\|_2^2 \sim \left(\|\beta - \widehat{\beta}^L\|_2^2 + \sigma_0^2\right)\chi^2\left(n_2\right)$. By Lemma 26, we have $\mathbb{P}_\theta(\bar{\mathcal{G}}_7|(X^{(1)}, y^{(1)})) \geq 1 - cp^{-2}$ and hence $\mathbb{P}_\theta(\bar{\mathcal{G}}_7) \geq 1 - cp^{-2}$. The control of $\mathbb{P}_\theta(\bar{\mathcal{G}}_8)$ follows from the fact $\mathbb{P}_\theta(\bar{\mathcal{G}}_8) \geq \mathbb{P}_\theta(\bar{\mathcal{G}}_7 \cap \bar{\mathcal{G}}_6) \geq (1 - cp^{-2})\mathbb{P}_\theta(\bar{\mathcal{G}}_6)$.

B.5.5 Proof of Lemma 8 and 21

The proof of Lemma 8 follows from (B.4.9) and (B.4.10) of Lemma 24 by taking $k = \|\beta^*\|_0$. The proof of Lemma 21 follows from Lemma 22 and 23 by taking $k = \|\beta^*\|_0$.

C Supplement for Chapter 4

C.1 Theory for valid IVs after controlling for high dimensional covariates

In this section, we state the theoretical results for valid IVs after controlling for high dimensional covariates. Under the assumptions (R1)-(R3), Theorem 23 shows if the instruments are valid after conditioning on many covariates, then the estimator $\widehat{\beta}_H$ in our procedure is consistent and asymptotically normal.

Theorem 23. *Suppose we have valid IVs, that is* $\boldsymbol{\pi}^* = 0$ *in* (4.2.2), *and the assumptions* (R1) − (R3) *hold. The following property holds for the estimator* $\widehat{\beta}_H$,

$$\sqrt{n}\left(\widehat{\beta}_H - \beta^*\right) = T^{\beta^*} + \Delta^{\beta^*}, \tag{C.1.1}$$

where $T^{\beta^*} \mid \mathbf{W} \sim N\left(0, \mathrm{V_H}\right)$, $\mathrm{V_H} = \sigma^2/\|\boldsymbol{\gamma}^*\|_2^4 \left\|\sum_{\mathrm{j}\in\mathcal{S}^*} \boldsymbol{\gamma}_\mathrm{j}^* \mathbf{W}\widehat{\mathbf{u}}^{[\mathrm{j}]}/\sqrt{\mathrm{n}}\right\|_2^2$ *and* $\Delta^{\beta^*}/\sqrt{\mathrm{V_H}} \xrightarrow{p} 0$ *as* $\sqrt{s_{\mathrm{z1}}}s\log p/\sqrt{n} \to 0$.

Theorem 23 states that if the IVs satisfy the exclusion restriction and no unmeasured confounding after conditioning on many covariates, $\widehat{\beta}_H$ defined in (4.3.13) is a consistent and the dominating part of the scaled difference $\sqrt{n}(\widehat{\beta}_H - \beta)$ is normal.

Based on the asymptotic normality established in (C.1.1), the following theorem justifies the coverage property of the confidence interval proposed in (4.3.14) under the assumption that all instruments have no direct effect and are unconfounded after conditioning on many covariates.

Theorem 24. *Suppose we have valid IVs, that is $\boldsymbol{\pi}^* = 0$ in* (4.2.2) *and the assumptions* (R1) – (R3) *hold. Assuming $\sqrt{s_{z1}} s \log p/\sqrt{n} \to 0$, the confidence interval given in* (4.3.14) *has asymptotically coverage probability $1-\alpha$, i.e.,*

$$\mathbf{P}\left\{\beta^* \in \left(\widehat{\beta}_H - z_{1-\alpha/2}\sqrt{\widehat{\mathrm{V}}_H/n}, \quad \widehat{\beta}_H + z_{1-\alpha/2}\sqrt{\widehat{\mathrm{V}}_H/n}\right)\right\} \to 1-\alpha, \qquad \text{(C.1.2)}$$

Theorem 24 is similar to a result given in Chernozhukov et al. (2015a), who studied IV estimators in high dimensional regime where all the instruments are valid after conditioning. However, there are some notable differences between our results and those in Chernozhukov et al. (2015a) in terms of sparsity and instrument-covariate modeling assumptions that are required to achieve $1-\alpha$ coverage. A simulation study is carried out in Section 4.5 to compare our procedure to that of the oracle.

C.2 Proofs of Theorems

In this section, we provide detailed proofs of Theorem 20, 23, 24 and 21. Proof of extra lemmas are presented in next section. Before presenting the proof, we will introduce the notations used throughout the proof.

C.2.1 Notations

For any vector $\mathbf{v} \in \mathbb{R}^p$, let $\mathbf{v}_j$ denote the jth element of $\mathbf{v}$. Let $\|\mathbf{v}\|_1$, $\|\mathbf{v}\|_2$, and $\|\mathbf{v}\|_\infty$ be the usual $1, 2$ and ∞-norms, respectively. Let $\|\mathbf{v}\|_0$ denote the 0-norm, i.e. the number of non-zero elements in $\mathbf{v}$. The support of $\mathbf{v}$, denoted as $\mathrm{supp}(\mathbf{v}) \subseteq \{1, \ldots, p\}$, is defined as the set containing the non-zero elements of the vector $\mathbf{v}$, i.e. $j \in \mathrm{supp}(\mathbf{v})$ if and only if $\mathbf{v}_j \neq 0$. Also, for a vector $\mathbf{v} \in \mathbb{R}^p$ and set $J \subseteq \{1, \ldots, p\}$, we denote $\mathbf{v}_J \in \mathbb{R}^p$ to be the vector where all the elements except whose indices are in J are zero. For a set J, $|J|$ denotes its cardinality.

For any n by p matrix $\mathbf{M} \in \mathbb{R}^{n\times p}$, we denote the (i,j) element of matrix $\mathbf{M}$ as $\mathbf{M}_{ij}$, the ith row as $\mathbf{M}_{i.}$, and the jth column as $\mathbf{M}_{.j}$. Let $\mathbf{M}^\intercal$ be the transpose of $\mathbf{M}$. Finally, $\|\mathbf{M}\|_\infty$ represents the element-wise matrix sup norm of matrix $\mathbf{M}$.

For a sequence of random variables X_n, we use $X_n \xrightarrow{p} X$ and $X_n \xrightarrow{d} X$ to represent that X_n converges to X in probability and in distribution, respectively. For any two sequences a_n and b_n, we will write $a_n \gg b_n$ if $\limsup \frac{b_n}{a_n} = 0$ and write $a_n \ll b_n$ if $b_n \gg a_n$. We use c and C to denote generic positive constants that may vary from place to place.

Throughout the whole proof section, we will use $\beta, \boldsymbol{\gamma}, \mathbf{\Gamma}, \boldsymbol{\psi}, \mathbf{\Psi}, \boldsymbol{\pi}, \mathbf{\Theta}_{11}, \mathbf{\Theta}_{22}, \mathbf{\Theta}_{12}$, $\mathbf{\Sigma}, T^\beta, \Delta^\beta$ to stand for $\beta^*, \boldsymbol{\gamma}^*, \mathbf{\Gamma}^*, \boldsymbol{\psi}^*, \mathbf{\Psi}^*, \boldsymbol{\pi}^*, \mathbf{\Theta}^*_{11}, \mathbf{\Theta}^*_{22}, \mathbf{\Theta}^*_{12}, \mathbf{\Sigma}^*, T^{\beta^*}, \Delta^{\beta^*}$ respectively and define

$$\widehat{\mathbf{v}}^{[j]} = \mathbf{W}^\intercal \widehat{\mathbf{u}}^{[j]} \quad \text{for} \quad 1 \leq j \leq p_{\mathrm{z}}.$$

We also introduce the notation $\mathbf{\Omega} = \mathbf{\Sigma}^{-1}$, $\sigma_1 = \sqrt{\mathbf{\Theta}_{11}}$, $\sigma_2 = \sqrt{\mathbf{\Theta}_{22}}$ and $\mathbf{\Pi}_{i\cdot} = (e_{i1}, \epsilon_{i2})$. Let $M_2 = \max\{1/\lambda_{\min}(\mathbf{\Theta}), \lambda_{\max}(\mathbf{\Theta})\}$ and hence $1/M_2 \leq \lambda_{\min}(\mathbf{\Theta}) \leq \lambda_{\max}(\mathbf{\Theta}) \leq M_2$. We normalize the columns of $\mathbf{W}$ as $\mathbf{H}_{.j} = \sqrt{n}\mathbf{W}_{.j}/\|\mathbf{W}_{.j}\|_2$ for $j \in [p]$. Let $\mathrm{Diag} = \mathrm{diag}\left(\|\mathbf{W}_{.j}\|_2/\sqrt{n}\right)_{1\leq j\leq p}$ denote the $p \times p$ diagonal matrix with (j,j) entry to be $\|\mathbf{W}_{.j}\|_2/\sqrt{n}$. We set $\lambda_0 = \sqrt{2.05 \log p/n} = (1+\gamma_0)\sqrt{2\delta_0 \log p/n}$, where $\delta_0 = \sqrt{1.025} > 1$ and $\gamma_0 = (1.025)^{\frac{1}{4}} - 1 > 0$. Take $\epsilon_0 = 2.01/\gamma_0 + 1$, $\nu_0 = 0.01$, $\tau_0 = 0.01$, $C_1 = 2.25$, $c_0 = 1/6$ and $C_0 = 3$. We also assume that $\log p/n \to 0$ and $\delta_0 \log p > 2$. Rather than use the constants directly in the following discussion, we use $\delta_0, \pi_0, \epsilon_0, \nu_0, C_1, C_0$ and c_0 to represent the above fixed constants in the following discussion. We review the following definition of restricted eigenvalue introduced in

Bickel et al. (2009),

$$\kappa(X,k,\alpha_0)=\min_{\substack{J_0\subset\{1,\cdots,p\},\\|J_0|\le k}}\ \min_{\substack{\delta\neq 0,\\\|\delta_{J_0^c}\|_1\le\alpha_0\|\delta_{J_0}\|_1}}\frac{\|X\delta\|_2}{\sqrt{n}\|\delta_{J_0}\|_2}. \tag{C.2.1}$$

Define the oracle estimator of σ_1 and σ_2 as

$$\sigma_1^{ora}=\frac{1}{\sqrt{n}}\|Y-\mathbf{Z}\boldsymbol{\Gamma}-\mathbf{X}\boldsymbol{\Psi}\|_2 \text{ and } \sigma_2^{ora}=\frac{1}{\sqrt{n}}\|D-\mathbf{Z}\boldsymbol{\gamma}-\mathbf{X}\boldsymbol{\psi}\|_2,$$

and

$$\tau=\sqrt{1+\epsilon_0}\frac{2\sqrt{s}\lambda_0}{\kappa(\boldsymbol{H},4s,1+2\epsilon_0)}. \tag{C.2.2}$$

C.2.2 Proof of Theorem 20

Define

$$A(\mathcal{V})=\widehat{\boldsymbol{\Sigma}}_{\mathcal{V},\mathcal{V}}-\widehat{\boldsymbol{\Sigma}}_{\mathcal{V},\mathcal{V}^c}\widehat{\boldsymbol{\Sigma}}^{-1}_{\mathcal{V}^c,\mathcal{V}^c}\widehat{\boldsymbol{\Sigma}}_{\mathcal{V}^c,\mathcal{V}} \quad\text{and}\quad A^*(\mathcal{V})=\boldsymbol{\Sigma}_{\mathcal{V},\mathcal{V}}-\boldsymbol{\Sigma}_{\mathcal{V},\mathcal{V}^c}\boldsymbol{\Sigma}^{-1}_{\mathcal{V}^c,\mathcal{V}^c}\boldsymbol{\Sigma}_{\mathcal{V}^c,\mathcal{V}}.$$

We introduce the following lemmas to facilitate the proof.

Lemma 31. *Under the assumptions of Theorem 20, we have*

$$\lim_{n\to\infty}\mathbf{P}\left(\widetilde{\mathcal{V}}=\mathcal{V}^*\right)=1 \tag{C.2.3}$$

Lemma 32. *Under the assumptions of Theorem 20, we have*

$$\sqrt{n}\left(\frac{\widetilde{\boldsymbol{\gamma}}_{\mathcal{V}^*}^{\intercal}A(\mathcal{V}^*)\widetilde{\boldsymbol{\Gamma}}_{\mathcal{V}^*}}{\widetilde{\boldsymbol{\gamma}}_{\mathcal{V}^*}^{\intercal}A(\mathcal{V}^*)\widetilde{\boldsymbol{\gamma}}_{\mathcal{V}^*}}-\beta\right)\xrightarrow{d}N\left(0,\frac{\boldsymbol{\Theta}_{11}+\beta^2\boldsymbol{\Theta}_{22}-2\beta\boldsymbol{\Theta}_{12}}{\gamma_{\mathcal{V}^*}^{\intercal}A^*(\mathcal{V})\gamma_{\mathcal{V}^*}}\right). \tag{C.2.4}$$

The estimator defined in (4.3.15) can be expressed as $\widehat{\beta}_E=\frac{\widetilde{\boldsymbol{\gamma}}_{\widetilde{\mathcal{V}}}^{\intercal}A(\widetilde{\mathcal{V}})\widetilde{\boldsymbol{\Gamma}}_{\widetilde{\mathcal{V}}}}{\widetilde{\boldsymbol{\gamma}}_{\widetilde{\mathcal{V}}}^{\intercal}A(\widetilde{\mathcal{V}})\widetilde{\boldsymbol{\gamma}}_{\widetilde{\mathcal{V}}}}$, and hence the difference $\sqrt{n}\left(\widehat{\beta}_E - \beta\right)$ can be expressed as

$$\sqrt{n}\left(\widehat{\beta}_E - \beta\right) = \sqrt{n}\left(\frac{\widetilde{\gamma}_{\mathcal{V}^*}^{\intercal} A(\mathcal{V}^*)\widetilde{\boldsymbol{\Gamma}}_{\mathcal{V}^*}}{\widetilde{\gamma}_{\mathcal{V}^*}^{\intercal} A(\mathcal{V}^*)\widetilde{\gamma}_{\mathcal{V}^*}} - \beta\right)\mathbf{1}_{\widetilde{\mathcal{V}}=\mathcal{V}^*} + \sum_{\mathcal{V}\neq\mathcal{V}^*}\sqrt{n}\left(\frac{\widetilde{\gamma}_{\mathcal{V}}^{\intercal} A(\mathcal{V})\widetilde{\boldsymbol{\Gamma}}_{\mathcal{V}}}{\widetilde{\gamma}_{\mathcal{V}}^{\intercal} A(\mathcal{V})\widetilde{\gamma}_{\mathcal{V}}} - \beta\right)\mathbf{1}_{\widetilde{\mathcal{V}}=\mathcal{V}} \tag{C.2.5}$$

By Lemma 31, we have $\mathbf{1}_{\widetilde{\mathcal{V}}=\mathcal{V}^*} \xrightarrow{p} 1$ and $\mathbf{1}_{\widetilde{\mathcal{V}}=\mathcal{V}} \xrightarrow{p} 0$ if $\mathcal{V} \neq \mathcal{V}^*$. Combined with Lemma 32 and Slutsky's theorem, we establish

$$\sqrt{n}\left(\frac{\widetilde{\gamma}_{\mathcal{V}^*}^{\intercal} A(\mathcal{V}^*)\widetilde{\boldsymbol{\Gamma}}_{\mathcal{V}^*}}{\widetilde{\gamma}_{\mathcal{V}^*}^{\intercal} A(\mathcal{V}^*)\widetilde{\gamma}_{\mathcal{V}^*}} - \beta\right)\mathbf{1}_{\widetilde{\mathcal{V}}=\mathcal{V}^*} \xrightarrow{d} N\left(0, \frac{\boldsymbol{\Theta}_{11} + \beta^2\boldsymbol{\Theta}_{22} - 2\beta\boldsymbol{\Theta}_{12}}{\gamma_{\mathcal{V}^*}^{\intercal} A^*(\mathcal{V}^*)\gamma_{\mathcal{V}^*}}\right).$$

Note that for any $\epsilon_0 > 0$,

$$\mathbf{P}\left(\left|\sqrt{n}\left(\widehat{\beta}_E - \beta\right) - \sqrt{n}\left(\frac{\widetilde{\gamma}_{\mathcal{V}^*}^{\intercal} A(\mathcal{V}^*)\widetilde{\boldsymbol{\Gamma}}_{\mathcal{V}^*}}{\widetilde{\gamma}_{\mathcal{V}^*}^{\intercal} A(\mathcal{V}^*)\widetilde{\gamma}_{\mathcal{V}^*}} - \beta\right)\mathbf{1}_{\widetilde{\mathcal{V}}=\mathcal{V}^*}\right| \geq \epsilon_0\right) \leq \mathbf{P}\left(\widetilde{\mathcal{V}} \neq \mathcal{V}^*\right) \tag{C.2.6}$$

and it follows from Lemma 31 that

$$\sqrt{n}\left(\widehat{\beta}_E - \beta\right) - \sqrt{n}\left(\frac{\widetilde{\gamma}_{\mathcal{V}^*}^{\intercal} A(\mathcal{V}^*)\widetilde{\boldsymbol{\Gamma}}_{\mathcal{V}^*}}{\widetilde{\gamma}_{\mathcal{V}^*}^{\intercal} A(\mathcal{V}^*)\widetilde{\gamma}_{\mathcal{V}^*}} - \beta\right)\mathbf{1}_{\widetilde{\mathcal{V}}=\mathcal{V}^*} \xrightarrow{p} 0. \tag{C.2.7}$$

By Lemma 3.7 in Wooldridge (2010), we establish (4.4.1).

C.2.3 Preliminary lemmas for high dimension case

We first define the following events for the random design $\mathbf{W}$ (the normalized $\boldsymbol{H}$) and the error $\mathbf{\Pi}$,

$$
\begin{aligned}
G_1 &= \left\{\frac{2}{5}\frac{1}{\sqrt{M_1}} < \frac{\|\mathbf{W}_{.j}\|_2}{\sqrt{n}} < \frac{7}{5}\sqrt{M_1} \text{ for } 1 \le j \le p\right\},\\
G_2 &= \left\{\left|\frac{(\sigma_i^{ora})^2}{\sigma_i^2} - 1\right| \le 2\sqrt{\frac{\log p}{n}} + 2\frac{\log p}{n} \text{ for } i = 1,2\right\},\\
G_3 &= \left\{\left|\frac{\boldsymbol{\gamma}^\intercal\widehat{\mathbf{\Sigma}}\boldsymbol{\gamma}}{\boldsymbol{\gamma}^\intercal\mathbf{\Sigma}\boldsymbol{\gamma}} - 1\right| \le 12\sqrt{\frac{\log p}{n}} \text{ and } \left|\frac{\mathbf{\Omega}_{j\cdot}{}^\intercal\widehat{\mathbf{\Sigma}}\mathbf{\Omega}_{j\cdot}}{\mathbf{\Omega}_{jj}} - 1\right| \le 12\sqrt{\frac{\log p}{n}}\ , 1 \le j \le p_{\mathrm{z}}\right\},\\
G_4 &= \left\{\kappa(\boldsymbol{H}, 4s, 1+2\epsilon_0) \ge \frac{1}{2\sqrt{M_1}}\right\},\\
G_5 &= \left\{\frac{\|\boldsymbol{H}^\intercal\mathbf{\Pi}_{i\cdot}\|_\infty}{n} \le \sigma_i\sqrt{\frac{2\delta_0\log p}{n}} \text{ for } i = 1,2\right\},\\
S_1 &= \left\{\frac{\|\boldsymbol{H}^\intercal\mathbf{\Pi}_{i\cdot}\|_\infty}{n} \le \sigma_i^{ora}\lambda_0\frac{\epsilon_0 - 1}{\epsilon_0 + 1}(1-\tau) \text{ for } i = 1,2\right\},\\
S_2 &= \left\{(1-\nu_0)\,\hat{\sigma}_i \le \sigma_i \le (1+\nu_0)\hat{\sigma}_i \text{ for } i = 1,2\right\},
\end{aligned}
\tag{C.2.8}
$$

and

$$
\begin{aligned}
A_1 &= \left\{\|e_j^\intercal\mathbf{\Omega}\hat{\mathbf{\Sigma}} - e_j^\intercal\|_\infty \le \lambda_n,\ j = 1,2,\cdots,p_{\mathrm{z}}\right\}, \quad \text{where } \lambda_n = 2eC_0M_1^2\sqrt{\frac{\log p}{n}},\\
A_2 &= \left\{|\widetilde{\boldsymbol{\gamma}}_j - \boldsymbol{\gamma}_j| \le \frac{\|\widehat{\mathbf{v}}^{[j]}\|_2\sigma_2}{\sqrt{n}}\sqrt{2.05\log p_{\mathrm{z}}} \text{ for } 1 \le j \le p_{\mathrm{z}}\right\},\\
A_3 &= \left\{\max_{1\le j\le p_{\mathrm{z}}}\left\|\frac{1}{n}(\widehat{v}^{[j]})^\intercal\mathbf{\Pi}_{.i}\right\|_\infty \le \left(1 + 12\sqrt{\frac{\log p}{n}}\right)M_1\sqrt{\frac{2.05\log p_{\mathrm{z}}}{n}}\sigma_i,\ \text{for } i = 1,2\right\},\\
A_4 &= \left\{\frac{2}{\sqrt{n}}\sum_{j\in\mathcal{S}^*}\boldsymbol{\gamma}_j\boldsymbol{v}^\intercal\mathbf{\Pi}_{.2} \le \frac{2\sqrt{\log p}}{\sqrt{n}}\left\|\sum_{j\in\mathcal{S}^*}\boldsymbol{\gamma}_j\widehat{\mathbf{v}}^{[j]}\right\|_2\sqrt{\mathbf{\Theta}_{22}}\right\},\\
A_5 &= \left\{\frac{1}{\sqrt{n}}\sum_{j\in\mathcal{S}^*}\boldsymbol{\gamma}_j\boldsymbol{v}^\intercal\left(\mathbf{\Pi}_{.1} + \beta\mathbf{\Pi}_{.2}\right) \le \frac{\sqrt{\log p}}{\sqrt{n}}\left\|\sum_{j\in\mathcal{S}^*}\boldsymbol{\gamma}_j\widehat{\mathbf{v}}^{[j]}\right\|_2\sqrt{\mathbf{\Theta}_{11} + \beta^2\mathbf{\Theta}_{22} + 2\beta\mathbf{\Theta}_{12}}\right\},
\end{aligned}
\tag{C.2.9}
$$

where $\widehat{\mathbf{\Sigma}} = \frac{1}{n}\mathbf{W}^\intercal\mathbf{W}$ and $\widehat{\mathbf{v}}^{[j]} = \mathbf{W}^\intercal\widehat{\mathbf{u}}^{[j]}$. Define

$$G = \cap_{i=1}^{5} G_i \quad \text{and} \quad S = \cap_{i=1}^{2} S_i \quad \text{and} \quad A = \cap_{i=1}^{5} A_i.$$

We introduce the following lemmas to control the probability of events G, S and A. The detailed proofs of the following lemmas are presented in Section C.3.4 and C.3.5.

Lemma 33. *If $s \leq cn/\log p$, then*

$$\mathbf{P}\left(G\right) \geq 1 - \frac{6}{p} - 2p^{1-C_1} - \frac{1}{2\sqrt{\pi\delta_0 \log p}}p^{1-\delta_0} - 2\exp\left(-\frac{c'n}{M_1^3}\right), \tag{C.2.10}$$

and

$$\mathbf{P}\left(G \cap S\right) \geq \mathbf{P}\left(G\right) - 2\exp\left(-\left(\frac{g_0 + 1 - \sqrt{2g_0+1}}{2}\right)n\right) - c''\frac{1}{\sqrt{\log p}}p^{1-\delta_0}, \tag{C.2.11}$$

where $g_0 = \nu_0/(2+3\nu_0)$ and c, c', c_* and c'' are universal positive constants, not depending on n and p. We also have

$$\mathbf{P}\left(A_1\right) \geq 1 - 2p_{\mathrm{z}}p^{1-c_0C_0^2}, \quad \textit{and} \quad \mathbf{P}(A_4 \cap A_5) \geq 1 - p^{-c}, \tag{C.2.12}$$

$$\min\{\mathbf{P}(A_2), \mathbf{P}(A_3)\} \geq \mathbf{P}\left((A_1 \cap G_1 \cap G_3)\right) - \frac{1}{2\sqrt{\pi \log p_{\mathrm{z}}}}p_{\mathrm{z}}^{-0.02}. \tag{C.2.13}$$

Lemma 34. *On the event $A_1 \cap G_1 \cap G_3$, we have*

$$\frac{\left(1-\lambda_n\right)^2}{2M_1} \leq \frac{\|\widehat{\mathbf{v}}^{[j]}\|_2^2}{n} \leq \left(1 + 12\sqrt{\frac{\log p}{n}}\right)M_1, \quad \textit{for } 1 \leq j \leq p_{\mathrm{z}}. \tag{C.2.14}$$

If $s_{z1}\sqrt{\log p/n} \to 0$, *on the event* G_3, *we have*

$$\frac{1}{n}\left\|\sum_{j\in\mathcal{S}^*}\boldsymbol{\gamma}_j\widehat{\mathbf{v}}^{[j]}\right\|_2^2 \geq \frac{M_1\|\boldsymbol{\gamma}\|_2^2\left(1-s_{z1}\lambda_n\right)^2}{1-12\sqrt{\frac{\log p}{n}}} \quad \text{and} \quad \frac{1}{n}\left\|\sum_{j\in\mathcal{V}^*}\boldsymbol{\gamma}_j\widehat{\mathbf{v}}^{[j]}\right\|_2^2 \geq \frac{M_1\|\boldsymbol{\gamma}_{\mathcal{V}^*}\|_2^2\left(1-s_{z1}\lambda_n\right)^2}{1-12\sqrt{\frac{\log p}{n}}}. \tag{C.2.15}$$

Furthermore, we have

$$\frac{M_1\left(1-s_{z1}\lambda_n\right)^2}{\|\boldsymbol{\gamma}\|_2^2\left(1-12\sqrt{\frac{\log p}{n}}\right)}\frac{1}{M_2} \leq \mathrm{V_H} \leq \frac{4\mathrm{s_{z1}}\mathrm{M}_1^2\mathrm{M}_2(1+\beta^2)}{\|\boldsymbol{\gamma}\|_2^2}, \tag{C.2.16}$$

and

$$\frac{M_1\left(1-s_{z1}\lambda_n\right)^2}{\|\boldsymbol{\gamma}_{\mathcal{V}^*}\|_2^2\left(1-12\sqrt{\frac{\log p}{n}}\right)}\frac{1}{M_2} \leq \mathrm{V} \leq \frac{4\mathrm{s_{z1}}\mathrm{M}_1^2\mathrm{M}_2(1+\beta^2)}{\|\boldsymbol{\gamma}_{\mathcal{V}^*}\|_2^2}. \tag{C.2.17}$$

C.2.4 Proof of Theorem 23

The proof of Theorem 23 is based on Lemma 35 and the following expression for the estimator $\widehat{\beta}_H$, $\widehat{\beta}_H = \widehat{\boldsymbol{\gamma}^\intercal\boldsymbol{\Gamma}}/\widehat{\|\boldsymbol{\gamma}\|_2^2}$, where $\widehat{\|\boldsymbol{\gamma}\|_2^2} = \sum_{j\in\widetilde{\mathcal{S}}}\widetilde{\boldsymbol{\gamma}}_j^2$and $\widehat{\boldsymbol{\gamma}^\intercal\boldsymbol{\Gamma}} = \sum_{j\in\widetilde{\mathcal{S}}}\widetilde{\boldsymbol{\gamma}}_j\widetilde{\boldsymbol{\Gamma}}_j$.

Lemma 35. *Suppose that* $\sqrt{s_{z1}}s\log p/\sqrt{n} \to 0$, $\boldsymbol{\pi}^* = 0$ *and the assumptions* (R1) – (R3) *hold. Then we have the following decompositions,*

$$\sqrt{n}\left(\widehat{\|\boldsymbol{\gamma}\|_2^2} - \|\boldsymbol{\gamma}\|_2^2\right) = \frac{2}{\sqrt{n}}\sum_{j\in\mathcal{S}^*}\boldsymbol{\gamma}_j\boldsymbol{v}^\intercal\boldsymbol{\Pi}_{\cdot 2} + R^{\gamma}, \tag{C.2.18}$$

and

$$\sqrt{n}\left(\widehat{\boldsymbol{\gamma}^\intercal\boldsymbol{\Gamma}} - \boldsymbol{\gamma}^\intercal\boldsymbol{\Gamma}\right) = \frac{1}{\sqrt{n}}\sum_{j\in\mathcal{S}^*}\boldsymbol{\gamma}_j\boldsymbol{v}^\intercal\left(\boldsymbol{\Pi}_{\cdot 1} + \beta\boldsymbol{\Pi}_{\cdot 2}\right) + R^{\mathrm{inter}}, \tag{C.2.19}$$

where

$$\frac{2}{\sqrt{n}}\sum_{j\in\mathcal{S}^*}\boldsymbol{\gamma}_j\boldsymbol{v}^\intercal\boldsymbol{\Pi}_{\cdot 2} \sim N\left(0, \frac{4}{n}\left\|\sum_{j\in\mathcal{S}^*}\boldsymbol{\gamma}_j\widehat{\mathbf{v}}^{[j]}\right\|_2^2\boldsymbol{\Theta}_{22}\right), \tag{C.2.20}$$

$$
\frac{1}{\sqrt{n}} \sum_{j \in \mathcal{S}^*} \boldsymbol{\gamma}_j \boldsymbol{v}^\intercal \left(\boldsymbol{\Pi}_{\cdot 1} + \beta \boldsymbol{\Pi}_{\cdot 2}\right) \sim N\left(0, \frac{1}{n}\left\| \sum_{j \in \mathcal{S}^*} \boldsymbol{\gamma}_j \widehat{\mathbf{v}}^{[j]} \right\|_2^2 \left(\boldsymbol{\Theta}_{11} + \beta^2 \boldsymbol{\Theta}_{22} + 2\beta \boldsymbol{\Theta}_{12}\right)\right), \tag{C.2.21}
$$

and on the event $A \cap S \cap G$, we have

$$
\max\left\{\left|R^{\gamma}\right|, \left|R^{\text{inter}}\right|\right\} \leq C\left(|\beta| + 1\right) \|\boldsymbol{\gamma}\|_2 \sqrt{s_{\mathrm{z}1}}\, s \frac{\log p}{\sqrt{n}} + C s_{\mathrm{z}1} \frac{\log p_{\mathrm{z}}}{\sqrt{n}}. \tag{C.2.22}
$$

Then on the event $A \cap S \cap G$, we have

$$
\max\left\{\left|\widehat{\|\boldsymbol{\gamma}\|_2^2} - \|\boldsymbol{\gamma}\|_2^2\right|, \left|\widehat{\boldsymbol{\gamma}^\intercal \boldsymbol{\Gamma}} - \boldsymbol{\gamma}^\intercal \boldsymbol{\Gamma}\right|\right\} \leq C \|\boldsymbol{\gamma}\|_2 s_{\mathrm{z}1} \sqrt{\frac{\log p}{n}} + C s_{\mathrm{z}1} \frac{\log p_{\mathrm{z}}}{n} \leq C \|\boldsymbol{\gamma}\|_2 s_{\mathrm{z}1} \sqrt{\frac{\log p}{n}}. \tag{C.2.23}
$$

In the following, we will prove (C.1.1) in the main paper. Note that

$$
\widetilde{\beta} - \beta = -\frac{\beta}{\|\boldsymbol{\gamma}\|_2^2}\left(\widehat{\|\boldsymbol{\gamma}\|_2^2} - \|\boldsymbol{\gamma}\|_2^2\right) + \frac{1}{\|\boldsymbol{\gamma}\|_2^2}\left(\widehat{\boldsymbol{\gamma}^\intercal \boldsymbol{\Gamma}} - \boldsymbol{\gamma}^\intercal \boldsymbol{\Gamma}\right) + \frac{\|\boldsymbol{\gamma}\|_2^2 - \widehat{\|\boldsymbol{\gamma}\|_2^2}}{\|\boldsymbol{\gamma}\|_2^2}\left(\frac{\widehat{\boldsymbol{\gamma}^\intercal \boldsymbol{\Gamma}}}{\widehat{\|\boldsymbol{\gamma}\|_2^2}} - \frac{\boldsymbol{\gamma}^\intercal \boldsymbol{\Gamma}}{\|\boldsymbol{\gamma}\|_2^2}\right). \tag{C.2.24}
$$

By Lemma 35, we have the following decomposition,

$$
\sqrt{n}\left(\widetilde{\beta} - \beta\right) = T^{\beta} + \Delta^{\beta}, \tag{C.2.25}
$$

where

$$
\begin{aligned}
T^{\beta} &= -\frac{\beta}{\|\boldsymbol{\gamma}\|_2^2} \frac{2}{\sqrt{n}} \sum_{j \in \mathcal{S}^*} \boldsymbol{\gamma}_j \boldsymbol{v}^\intercal \boldsymbol{\Pi}_{\cdot 2} + \frac{1}{\|\boldsymbol{\gamma}\|_2^2} \frac{1}{\sqrt{n}} \sum_{j \in \mathcal{S}^*} \boldsymbol{\gamma}_j \boldsymbol{v}^\intercal \left(\boldsymbol{\Pi}_{\cdot 1} + \beta \boldsymbol{\Pi}_{\cdot 2}\right) \\
&= \frac{1}{\|\boldsymbol{\gamma}\|_2^2} \frac{1}{\sqrt{n}} \sum_{j \in \mathcal{S}^*} \boldsymbol{\gamma}_j \boldsymbol{v}^\intercal \left(\boldsymbol{\Pi}_{\cdot 1} - \beta \boldsymbol{\Pi}_{\cdot 2}\right),
\end{aligned}
$$

and $\Delta^{\beta} = \mathrm{Res}_1 + \mathrm{Res}_2$ with

$$
\mathrm{Res}_1 = \frac{1}{\|\boldsymbol{\gamma}\|_2^2}\left(-\beta R^{\gamma} + R^{\text{inter}}\right) \text{ and } \mathrm{Res}_2 = \sqrt{n} \frac{\|\boldsymbol{\gamma}\|_2^2 - \widehat{\|\boldsymbol{\gamma}\|_2^2}}{\|\boldsymbol{\gamma}\|_2^2}\left(\frac{\widehat{\boldsymbol{\gamma}^\intercal \boldsymbol{\Gamma}}}{\widehat{\|\boldsymbol{\gamma}\|_2^2}} - \frac{\boldsymbol{\gamma}^\intercal \boldsymbol{\Gamma}}{\|\boldsymbol{\gamma}\|_2^2}\right).
$$

By the distribution of $\mathbf{\Pi}$, we establish that

$$T^{\beta} \mid \mathbf{W} \sim N\left(0, \frac{1}{n\|\boldsymbol{\gamma}\|_2^4}\left\|\sum_{j\in\mathcal{S}^*}\boldsymbol{\gamma}_j\widehat{\mathbf{v}}^{[j]}\right\|_2^2\left(\mathbf{\Theta}_{11}+\beta^2\mathbf{\Theta}_{22}-2\beta\mathbf{\Theta}_{12}\right)\right). \tag{C.2.26}$$

By Lemma 35, on the event $G \cap S \cap A$, we have

$$\frac{1}{\sqrt{\mathrm{V_H}}}\left|\mathrm{Res}_1\right| \leq C\frac{1}{\|\boldsymbol{\gamma}\|_2}\left(|\beta|\left|R^{\boldsymbol{\gamma}}\right|+\left|R^{\mathrm{inter}}\right|\right) \leq C\left(|\beta|+1\right)\sqrt{s_{\mathrm{z}1}}s\frac{\log p}{\sqrt{n}}+C\frac{1}{\|\boldsymbol{\gamma}\|_2}\frac{s_{\mathrm{z}1}\log p}{\sqrt{n}}. \tag{C.2.27}$$

Note that on the event $G \cap S \cap A$,

$$\frac{1}{\sqrt{\mathrm{V_H}}}\mathrm{Res}_2 \leq C\sqrt{n}\frac{\|\boldsymbol{\gamma}\|_2^2-\widehat{\|\boldsymbol{\gamma}\|_2^2}}{\|\boldsymbol{\gamma}\|_2} \times \frac{\left(\widehat{\boldsymbol{\gamma}^{\intercal}\mathbf{\Gamma}}-\boldsymbol{\gamma}^{\intercal}\mathbf{\Gamma}\right)+\beta\left(\|\boldsymbol{\gamma}\|_2^2-\widehat{\|\boldsymbol{\gamma}\|_2^2}\right)}{\|\boldsymbol{\gamma}\|_2^2+\left(\widehat{\|\boldsymbol{\gamma}\|_2^2}-\|\boldsymbol{\gamma}\|_2^2\right)} \leq C\frac{s_{\mathrm{z}1}^3(\log p)^{\frac{3}{2}}}{n}, \tag{C.2.28}$$

where the last inequality follows from (C.2.23). Combined with (C.2.27), by

$$\sqrt{s_{\mathrm{z}1}}s\log p/\sqrt{n} \to 0,$$

we can establish that on the event $G \cap S \cap A$,

$$\left|\Delta^{\beta}/\sqrt{\mathrm{V_H}}\right| \leq C\sqrt{s_{\mathrm{z}1}}s\frac{\log p}{\sqrt{n}}+C\frac{1}{\|\boldsymbol{\gamma}\|_2}\frac{s_{\mathrm{z}1}\log p}{\sqrt{n}}. \tag{C.2.29}$$

Since $\sqrt{s_{\mathrm{z}1}}s\log p/\sqrt{n} \to 0$, we establish $\Delta^{\beta}/\sqrt{\mathrm{V_H}} \xrightarrow{p} 0$. Combined with (C.2.26), we establish (C.1.1).

C.2.5 Proof of Theorem 24

We first introduce the following lemma to establish the coverage property.

Lemma 36. *Suppose that* $\boldsymbol{\pi}^* = 0$ *and the assumptions* (R1) − (R3) *hold. As*

$\sqrt{s_{\mathrm{z}1}} s \log p / \sqrt{n} \to 0$, *then we have*

$$\frac{\widehat{\mathrm{V}}_H}{\mathrm{V}_{\mathrm{H}}} \xrightarrow{p} 1. \tag{C.2.30}$$

By (C.2.26), we have $\frac{T^{\beta}}{\sqrt{\mathrm{V}_{\mathrm{H}}}} \sim N(0,1)$. Combined with (C.2.29) and Lemma 36, we have

$$\sqrt{n}\frac{\widehat{\beta}_H - \beta}{\sqrt{\widehat{\mathrm{V}}_H}} = \frac{T^{\beta} + \Delta^{\beta}}{\sqrt{\mathrm{V}_H}} \times \frac{\sqrt{\mathrm{V}_H}}{\sqrt{\widehat{\mathrm{V}}_H}} \xrightarrow{d} N(0,1). \tag{C.2.31}$$

and hence the coverage property (C.1.2) follows.

C.2.6 Proof of Theorem 21

The proof of the theorem follows from the following lemma, which characterizes the behavior of the selection process (4.3.7) and (4.3.9) in the main paper.

Lemma 37. *Suppose that* $\sqrt{s_{\mathrm{z}1}} s \log p / \sqrt{n} \to 0$ *and the assumptions* (R1) $-$ (R3) *and* (IN1) $-$ (IN2) *are satisfied. With probability larger than* $1 - c\left(p^{-c} + \exp\left(-cn\right)\right)$, *we have*

$$\widetilde{\mathcal{V}} \subset \left\{ i \in \mathcal{S}^* : \left| \frac{\boldsymbol{\pi}_j}{\boldsymbol{\gamma}_j} \right| \leq 2C_* \frac{1}{\delta_{\min}} \sqrt{\frac{\log p_{\mathrm{z}}}{n}} \right\} \quad \textit{and} \quad \left| \widetilde{\mathcal{V}} \right| > \frac{1}{2} \left| \mathcal{S}^* \right|. \tag{C.2.32}$$

Under the extra assumption (IN3) *in the main paper, with probability larger than* $1 - c\left(p^{-c} + \exp\left(-cn\right)\right)$

$$\widetilde{\mathcal{V}} = \mathcal{V}^*. \tag{C.2.33}$$

We have the following decomposition,

$$\begin{aligned}\widehat{\beta}-\beta &= \frac{\sum_{j\in\widetilde{\mathcal{V}}}\widetilde{\gamma}_j\widetilde{\boldsymbol{\Gamma}}_j}{\sum_{j\in\widetilde{\mathcal{V}}}\widetilde{\gamma}_j^2}-\frac{\sum_{j\in\widetilde{\mathcal{V}}}\gamma_j\boldsymbol{\Gamma}_j}{\sum_{j\in\widetilde{\mathcal{V}}}\gamma_j^2}+\frac{\sum_{j\in\widetilde{\mathcal{V}}}\gamma_j\boldsymbol{\Gamma}_j}{\sum_{j\in\widetilde{\mathcal{V}}}\gamma_j^2}-\beta \\ &= \left(\frac{\sum_{j\in\widetilde{\mathcal{V}}}\widetilde{\gamma}_j\widetilde{\boldsymbol{\Gamma}}_j}{\sum_{j\in\widetilde{\mathcal{V}}}\widetilde{\gamma}_j^2}-\frac{\sum_{j\in\widetilde{\mathcal{V}}}\gamma_j\boldsymbol{\Gamma}_j}{\sum_{j\in\widetilde{\mathcal{V}}}\gamma_j^2}\right)+\frac{\sum_{j\in\widetilde{\mathcal{V}}}\gamma_j\boldsymbol{\pi}_j}{\sum_{j\in\widetilde{\mathcal{V}}}\gamma_j^2},\end{aligned} \tag{C.2.34}$$

where the first term is taken as the variance term and the second term is taken as the bias term. In the following, we are going to analyze the bias and the variance term separately. For the bias term, we have

$$\left|\frac{\sum_{j\in\widetilde{\mathcal{V}}}\gamma_j\boldsymbol{\pi}_j}{\sum_{j\in\widetilde{\mathcal{V}}}\gamma_j^2}\right| = \left|\frac{\sum_{j\in\widetilde{\mathcal{V}}}\gamma_j^2\frac{\boldsymbol{\pi}_j}{\gamma_j}}{\sum_{j\in\widetilde{\mathcal{V}}}\gamma_j^2}\right| \leq \max_{j\in\widetilde{\mathcal{V}}}\left|\frac{\boldsymbol{\pi}_j}{\gamma_j}\right| \leq 2C_*\frac{1}{\delta_{\min}}\sqrt{\frac{\log p_{\mathrm{z}}}{n}}, \tag{C.2.35}$$

where the last inequality follows from (C.2.32). The following lemma controls the variance term.

Lemma 38. *Suppose that* $\sqrt{s_{\mathrm{z}1}}s\log p/\sqrt{n}\to 0$ *and the assumptions* (R1) – (R3) *and* (IN1) – (IN3) *are satisfied. On the event* $A\cap S\cap G$, *we have*

$$\left|\frac{\sum_{j\in\widetilde{\mathcal{V}}}\widetilde{\gamma}_j\widetilde{\boldsymbol{\Gamma}}_j}{\sum_{j\in\widetilde{\mathcal{V}}}\widetilde{\gamma}_j^2}-\frac{\sum_{j\in\widetilde{\mathcal{V}}}\gamma_j\boldsymbol{\Gamma}_j}{\sum_{j\in\widetilde{\mathcal{V}}}\gamma_j^2}\right| \leq C\frac{1}{\delta_{\min}}\sqrt{\frac{\log p_{\mathrm{z}}}{n}}. \tag{C.2.36}$$

Combining (C.2.35) and (C.2.36), we can establish (4.4.4) in the main paper. Under the stronger assumption (4.4.3) in the main paper, we can establish (C.2.33) and the decomposition (C.2.34) holds as

$$\widehat{\beta}-\beta = \left(\frac{\sum_{j\in\mathcal{V}^*}\widetilde{\gamma}_j\widetilde{\boldsymbol{\Gamma}}_j}{\sum_{j\in\mathcal{V}^*}\widetilde{\gamma}_j^2}-\frac{\sum_{j\in\mathcal{V}^*}\gamma_j\boldsymbol{\Gamma}_j}{\sum_{j\in\mathcal{V}^*}\gamma_j^2}\right).$$

Based on the this expression, (4.4.5) in Chapter 4 will follow the same argument with (C.1.1), which is presented in Section C.2.4. We introduce the following lemma to establish the coverage property.

Lemma 39. *Suppose the assumptions* (R1) − (R5) *and* (IN1) − (IN3) *are satisfied. As* $\sqrt{s_{z1}} s \log p / \sqrt{n} \to 0$*, we have*

$$\frac{\widehat{\mathrm{V}}}{\mathrm{V}} \xrightarrow{p} 1. \tag{C.2.37}$$

Similarly to the proof of Theorem 24 in Section C.2.5, we establish

$$\sqrt{n} \frac{\widehat{\beta} - \beta}{\sqrt{\widehat{\mathrm{V}}}} = \frac{T^{\beta} + \Delta^{\beta}}{\sqrt{\mathrm{V}}} \times \frac{\sqrt{\mathrm{V}}}{\sqrt{\widehat{\mathrm{V}}}} \xrightarrow{d} N(0,1), \tag{C.2.38}$$

and hence establish the coverage property (4.4.6) in the main paper.

C.3 Proof of extra lemmas

In this section, we prove extra lemmas used in the proof of main theorems.

C.3.1 Proof of Lemma 31

Define $\mathcal{I} = \{1, 2, \cdots, p_{\mathrm{z}}\}$. We first note the following expression for $\widetilde{\boldsymbol{\gamma}}_j$ and $\widetilde{\boldsymbol{\Gamma}}_j$ for $i \in \mathcal{I}$,

$$\sqrt{n} \left(\widetilde{\boldsymbol{\gamma}}_j - \boldsymbol{\gamma}_j\right) = \left(\widehat{\boldsymbol{\Sigma}}^{-1}\right)_{j,\cdot} \frac{1}{\sqrt{n}} \mathbf{W}^{\intercal} \boldsymbol{\Pi}_{\cdot 2} \quad \text{and} \quad \sqrt{n} \left(\widetilde{\boldsymbol{\Gamma}}_j - \boldsymbol{\Gamma}_j\right) = \left(\widehat{\boldsymbol{\Sigma}}^{-1}\right)_{j,\cdot} \frac{1}{\sqrt{n}} \mathbf{W}^{\intercal} \boldsymbol{\Pi}_{\cdot 1} \tag{C.3.1}$$

and the following limiting theorem (Theorem 3.1 in Wooldridge (2010)),

$$\widetilde{\boldsymbol{\gamma}} \xrightarrow{p} \boldsymbol{\gamma} \quad \text{and} \quad \widetilde{\boldsymbol{\Gamma}} \xrightarrow{p} \boldsymbol{\Gamma}, \tag{C.3.2}$$

$$\sqrt{n} \left(\widetilde{\boldsymbol{\gamma}} - \boldsymbol{\gamma}\right) \xrightarrow{d} N\left(0, \Theta_{22} \left(\boldsymbol{\Sigma}^{-1}\right)_{\mathcal{I},\mathcal{I}}\right) \quad \text{and} \quad \sqrt{n} \left(\widetilde{\boldsymbol{\Gamma}} - \boldsymbol{\Gamma}\right) \xrightarrow{d} N\left(0, \Theta_{11} \left(\boldsymbol{\Sigma}^{-1}\right)_{\mathcal{I},\mathcal{I}}\right). \tag{C.3.3}$$

Note that

$$\frac{\sqrt{\widehat{\boldsymbol{\Theta}}_{22}}\|\mathbf{W}(\widehat{\boldsymbol{\Sigma}}^{-1})_{\cdot j}\|_2}{\sqrt{n}} \xrightarrow{p} \sqrt{\Theta_{22}\left(\boldsymbol{\Sigma}^{-1}\right)_{jj}}. \tag{C.3.4}$$

We define the following events

$$\begin{aligned}
\mathcal{B}_1 &= \left\{\widetilde{\mathcal{S}} = \mathcal{S}^*\right\} \\
\mathcal{B}_2 &= \left\{\max_{j \in \mathcal{V}^*} \|\widetilde{\boldsymbol{\pi}}^{[j]}\|_0 < \frac{|\mathcal{S}^*|}{2} < \min_{j \in \mathcal{S}^* \setminus \mathcal{V}^*} \|\widetilde{\boldsymbol{\pi}}^{[j]}\|_0\right\} \\
\mathcal{B}_3 &= \left\{\operatorname{supp}\left(\widetilde{\boldsymbol{\pi}}^{[j]}_{\mathcal{S}^*}\right) = \operatorname{supp}\left(\boldsymbol{\pi}_{\mathcal{S}^*}\right) \quad \text{for } j \in \mathcal{V}^*\right\}
\end{aligned} \tag{C.3.5}$$

On the event $\mathcal{B} = \mathcal{B}_1 \cap \mathcal{B}_2 \cap \mathcal{B}_3$, we have $\widetilde{\mathcal{V}} = \mathcal{V}^*$ and it is sufficient to show that

$$\lim_{n \to \infty} \mathbf{P}(\mathcal{B}) = 0 \tag{C.3.6}$$

For $j \in \mathcal{S}$, we have

$$|\widetilde{\gamma}_j| - \frac{\sqrt{\widehat{\boldsymbol{\Theta}}_{22}}\|\mathbf{W}(\widehat{\boldsymbol{\Sigma}}^{-1})_{\cdot j}\|_2}{\sqrt{n}} \sqrt{\frac{a_0 \log n}{n}} \xrightarrow{p} |\boldsymbol{\gamma}_j| > 0, \tag{C.3.7}$$

where the convergence follows from (C.3.2) and (C.3.4). For $j \in \mathcal{S}^c$, we have

$$\sqrt{\frac{n}{a_0 \log n}} |\widetilde{\gamma}_j| - \frac{\sqrt{\widehat{\boldsymbol{\Theta}}_{22}}\|\mathbf{W}(\widehat{\boldsymbol{\Sigma}}^{-1})_{\cdot j}\|_2}{\sqrt{n}} \xrightarrow{p} -\sqrt{\Theta_{22}\left(\boldsymbol{\Sigma}^{-1}\right)_{jj}} < 0, \tag{C.3.8}$$

where the convergence follows from (C.3.3) and (C.3.4). Combining (C.3.7) and (C.3.8), we establish that

$$\lim_{n \to \infty} \mathbf{P}\left(\mathcal{B}_1\right) = 1. \tag{C.3.9}$$

In the following, we control $\lim_{n \to \infty} \mathbf{P}(\mathcal{B})$. Without loss of generality, we assume $1 \in \widetilde{\mathcal{S}}$ and focus on the case $i = 1$. In the following, we are going to analyze the performance of $\widehat{\beta}^{[1]}$ and $\widetilde{\boldsymbol{\pi}}_j^{[1]}$. Note that

$$\widehat{\beta}^{[1]} - \frac{\boldsymbol{\Gamma}_1}{\boldsymbol{\gamma}_1} \xrightarrow{p} 0. \tag{C.3.10}$$

and hence we have

$$\sqrt{\widehat{\boldsymbol{\Theta}}_{11} + (\widehat{\beta}^{[1]})^2 \widehat{\boldsymbol{\Theta}}_{22} - 2\widehat{\beta}^{[1]} \widehat{\boldsymbol{\Theta}}_{12}} \frac{\|\mathbf{W}((\widehat{\boldsymbol{\Sigma}}^{-1})_{k\cdot} - \frac{\widetilde{\gamma}_k}{\widetilde{\gamma}_1}(\widehat{\boldsymbol{\Sigma}}^{-1})_{1\cdot})\|_2}{\sqrt{n}}$$
$$\xrightarrow{p} \sqrt{\boldsymbol{\Theta}_{11} + \left(\frac{\boldsymbol{\Gamma}_1}{\boldsymbol{\gamma}_1}\right)^2 \boldsymbol{\Theta}_{22} - 2\frac{\boldsymbol{\Gamma}_1}{\boldsymbol{\gamma}_1}\boldsymbol{\Theta}_{12}} \sqrt{(\boldsymbol{\Sigma}^{-1})_{kk} + \left(\frac{\boldsymbol{\gamma}_k}{\boldsymbol{\gamma}_1}\right)^2 (\boldsymbol{\Sigma}^{-1})_{11} - 2\frac{\boldsymbol{\gamma}_k}{\boldsymbol{\gamma}_1}(\boldsymbol{\Sigma}^{-1})_{k1}}. \tag{C.3.11}$$

We also have the following expression

$$\begin{aligned}
&\widehat{\boldsymbol{\pi}}_k^{[1]} - \left(\boldsymbol{\Gamma}_k - \frac{\boldsymbol{\Gamma}_1}{\boldsymbol{\gamma}_1}\boldsymbol{\gamma}_k\right) = \left(\widetilde{\boldsymbol{\Gamma}}_k - \frac{\widetilde{\boldsymbol{\Gamma}}_1}{\widetilde{\boldsymbol{\gamma}}_1}\widetilde{\boldsymbol{\gamma}}_k\right) - \left(\boldsymbol{\Gamma}_k - \frac{\boldsymbol{\Gamma}_1}{\boldsymbol{\gamma}_1}\boldsymbol{\gamma}_k\right) \\
&= \left(\widetilde{\boldsymbol{\Gamma}}_k - \boldsymbol{\Gamma}_k\right) - \frac{\boldsymbol{\Gamma}_1}{\boldsymbol{\gamma}_1}(\widetilde{\boldsymbol{\gamma}}_k - \boldsymbol{\gamma}_k) - \frac{\boldsymbol{\gamma}_k}{\boldsymbol{\gamma}_1^2}\left(\boldsymbol{\gamma}_1\left(\widetilde{\boldsymbol{\Gamma}}_1 - \boldsymbol{\Gamma}_1\right) - \boldsymbol{\Gamma}_1(\widetilde{\boldsymbol{\gamma}}_1 - \boldsymbol{\gamma}_1)\right) \\
&+ \left(\frac{\widetilde{\boldsymbol{\Gamma}}_1}{\widetilde{\boldsymbol{\gamma}}_1} - \frac{\boldsymbol{\Gamma}_1}{\boldsymbol{\gamma}_1}\right)\left(\frac{\boldsymbol{\gamma}_k}{\boldsymbol{\gamma}_1}(\widetilde{\boldsymbol{\gamma}}_1 - \boldsymbol{\gamma}_1) - (\widetilde{\boldsymbol{\gamma}}_k - \boldsymbol{\gamma}_k)\right)
\end{aligned} \tag{C.3.12}$$

Note that

$$\begin{aligned}
&\sqrt{n}\left(\left(\widetilde{\boldsymbol{\Gamma}}_k - \boldsymbol{\Gamma}_k\right) - \frac{\boldsymbol{\Gamma}_1}{\boldsymbol{\gamma}_1}(\widetilde{\boldsymbol{\gamma}}_k - \boldsymbol{\gamma}_k) - \frac{\boldsymbol{\gamma}_k}{\boldsymbol{\gamma}_1^2}\left(\boldsymbol{\gamma}_1\left(\widetilde{\boldsymbol{\Gamma}}_1 - \boldsymbol{\Gamma}_1\right) - \boldsymbol{\Gamma}_1(\widetilde{\boldsymbol{\gamma}}_1 - \boldsymbol{\gamma}_1)\right)\right) \\
&= \left(\left(\widehat{\boldsymbol{\Sigma}}^{-1}\right)_{\cdot k} - \frac{\boldsymbol{\gamma}_j}{\boldsymbol{\gamma}_1}\left(\widehat{\boldsymbol{\Sigma}}^{-1}\right)_{\cdot 1}\right)\frac{1}{\sqrt{n}}\mathbf{W}^\intercal\left(\boldsymbol{\Pi}_{\cdot 2} - \frac{\boldsymbol{\Gamma}_1}{\boldsymbol{\gamma}_1}\boldsymbol{\Pi}_{\cdot 1}\right) \\
&\xrightarrow{d} N\left(0, \boldsymbol{\Theta}_{11} + \left(\frac{\boldsymbol{\Gamma}_1}{\boldsymbol{\gamma}_1}\right)^2 \boldsymbol{\Theta}_{22} - 2\frac{\boldsymbol{\Gamma}_1}{\boldsymbol{\gamma}_1}\boldsymbol{\Theta}_{12}(\boldsymbol{\Sigma}^{-1})_{kk} + \left(\frac{\boldsymbol{\gamma}_k}{\boldsymbol{\gamma}_1}\right)^2 (\boldsymbol{\Sigma}^{-1})_{11} - 2\frac{\boldsymbol{\gamma}_k}{\boldsymbol{\gamma}_1}(\boldsymbol{\Sigma}^{-1})_{k1}\right),
\end{aligned} \tag{C.3.13}$$

where the convergence follows from Theorem 3.1 in Wooldridge (2010). By (C.3.2) and (C.3.3), we have

$$\left(\frac{\widetilde{\boldsymbol{\Gamma}}_1}{\widetilde{\boldsymbol{\gamma}}_1} - \frac{\boldsymbol{\Gamma}_1}{\boldsymbol{\gamma}_1}\right)\left(\frac{\boldsymbol{\gamma}_k}{\boldsymbol{\gamma}_1}(\widetilde{\boldsymbol{\gamma}}_1 - \boldsymbol{\gamma}_1) - (\widetilde{\boldsymbol{\gamma}}_k - \boldsymbol{\gamma}_k)\right) \xrightarrow{p} 0.$$

Combined with (C.3.10) and (C.3.13), we have

$$\frac{\sqrt{n}}{\sqrt{\widehat{\boldsymbol{\Theta}}_{11}+(\widehat{\beta}^{[1]})^2\widehat{\boldsymbol{\Theta}}_{22}-2\widehat{\beta}^{[1]}\widehat{\boldsymbol{\Theta}}_{12}}\frac{\|\mathbf{W}((\widehat{\boldsymbol{\Sigma}}^{-1})_{k\cdot}-\frac{\widetilde{\gamma}_k}{\widetilde{\gamma}_1}(\widehat{\boldsymbol{\Sigma}}^{-1})_{1\cdot})\|_2}{\sqrt{n}}}\left(\widehat{\boldsymbol{\pi}}_k^{[1]}-\left(\boldsymbol{\Gamma}_k-\frac{\boldsymbol{\Gamma}_1}{\boldsymbol{\gamma}_1}\boldsymbol{\gamma}_k\right)\right)\xrightarrow{p}N(0,1) \tag{C.3.14}$$

and hence

$$\widehat{\boldsymbol{\pi}}_k^{[1]}\xrightarrow{p}\boldsymbol{\Gamma}_k-\frac{\boldsymbol{\Gamma}_1}{\boldsymbol{\gamma}_1}\boldsymbol{\gamma}_k. \tag{C.3.15}$$

We divide the discussion into the following three cases,

- $1\in\mathcal{S}^*\backslash\mathcal{V}^*$ and $k\in\mathcal{V}^*$;
- $1\in\mathcal{V}^*$ and $k\in\mathcal{V}^*$;
- $1\in\mathcal{V}^*$ and $k\in\mathcal{S}^*\backslash\mathcal{V}^*$.

$\underline{1\in\mathcal{S}^*\backslash\mathcal{V}^* \text{ and } k\in\mathcal{V}^*}$

In this case, $\boldsymbol{\Gamma}_k-\frac{\boldsymbol{\Gamma}_1}{\boldsymbol{\gamma}_1}\boldsymbol{\gamma}_k=\frac{\boldsymbol{\pi}_1}{\boldsymbol{\gamma}_1}\boldsymbol{\gamma}_k\neq 0$. Hence, we have

$$\begin{aligned}&\left|\widehat{\boldsymbol{\pi}}_k^{[1]}\right|-2.05\sqrt{\widehat{\boldsymbol{\Theta}}_{11}+(\widehat{\beta}^{[1]})^2\widehat{\boldsymbol{\Theta}}_{22}-2\widehat{\beta}^{[1]}\widehat{\boldsymbol{\Theta}}_{12}}\frac{\|\mathbf{W}((\widehat{\boldsymbol{\Sigma}}^{-1})_{k\cdot}-\frac{\widetilde{\gamma}_k}{\widetilde{\gamma}_1}(\widehat{\boldsymbol{\Sigma}}^{-1})_{1\cdot})\|_2}{\sqrt{n}}\sqrt{\frac{\log n}{n}}\\&\xrightarrow{p}\left|\frac{\boldsymbol{\pi}_1}{\boldsymbol{\gamma}_1}\boldsymbol{\gamma}_k\right|>0\end{aligned} \tag{C.3.16}$$

and

$$\lim_{n\to\infty}\mathbf{P}\left(k\in\operatorname{supp}\left(\widetilde{\boldsymbol{\pi}}_{\mathcal{S}^*}^{[1]}\right)\right)=1. \tag{C.3.17}$$

Hence, by the assumption (IN1),

$$\lim_{n\to\infty}\mathbf{P}\left(\|\widetilde{\boldsymbol{\pi}}_{\mathcal{S}^*}^{[1]}\|_0>\frac{|\mathcal{V}^*|}{2}\right)=1. \tag{C.3.18}$$

$\underline{1\in\mathcal{V}^* \text{ and } k\in\mathcal{V}^*}$

In this case, $\boldsymbol{\Gamma}_k - \frac{\boldsymbol{\Gamma}_1}{\gamma_1}\boldsymbol{\gamma}_k = 0$. By (C.3.14), we have

$$\frac{\left|\widehat{\boldsymbol{\pi}}_k^{[1]}\right|}{2.05\sqrt{\widehat{\boldsymbol{\Theta}}_{11} + (\widehat{\beta}^{[1]})^2\widehat{\boldsymbol{\Theta}}_{22} - 2\widehat{\beta}^{[1]}\widehat{\boldsymbol{\Theta}}_{12}}\frac{\|\mathbf{W}((\widehat{\boldsymbol{\Sigma}}^{-1})_{k\cdot} - \frac{\widetilde{\gamma}_k}{\widetilde{\gamma}_1}(\widehat{\boldsymbol{\Sigma}}^{-1})_{1\cdot})\|_2}{\sqrt{n}}\sqrt{\frac{\log n}{n}}} \xrightarrow{p} 0 \tag{C.3.19}$$

and

$$\lim_{n\to\infty} \mathbf{P}\left(k \notin \operatorname{supp}\left(\widetilde{\boldsymbol{\pi}}_{\mathcal{S}^*}^{[1]}\right)\right) = 1. \tag{C.3.20}$$

Hence, by the assumption (IN1),

$$\lim_{n\to\infty} \mathbf{P}\left(\|\widetilde{\boldsymbol{\pi}}_{\mathcal{S}^*}^{[1]}\|_0 < \frac{|\mathcal{V}^*|}{2}\right) = 1. \tag{C.3.21}$$

$1 \in \mathcal{V}^*$ and $k \in \mathcal{S}^* \backslash \mathcal{V}^*$

In this case, $\boldsymbol{\Gamma}_k - \frac{\boldsymbol{\Gamma}_1}{\gamma_1}\boldsymbol{\gamma}_k = \boldsymbol{\pi}_k \neq 0$. Hence, we have

$$\left|\widehat{\boldsymbol{\pi}}_k^{[1]}\right| - 2.05\sqrt{\widehat{\boldsymbol{\Theta}}_{11} + (\widehat{\beta}^{[1]})^2\widehat{\boldsymbol{\Theta}}_{22} - 2\widehat{\beta}^{[1]}\widehat{\boldsymbol{\Theta}}_{12}}\frac{\|\mathbf{W}((\widehat{\boldsymbol{\Sigma}}^{-1})_{k\cdot} - \frac{\widetilde{\gamma}_k}{\widetilde{\gamma}_1}(\widehat{\boldsymbol{\Sigma}}^{-1})_{1\cdot})\|_2}{\sqrt{n}}\sqrt{\frac{\log n}{n}} \xrightarrow{p} |\boldsymbol{\pi}_k| > 0 \tag{C.3.22}$$

and hence

$$\lim_{n\to\infty} \mathbf{P}\left(k \in \operatorname{supp}\left(\widetilde{\boldsymbol{\pi}}_{\mathcal{S}^*}^{[1]}\right)\right) = 1. \tag{C.3.23}$$

Since we can replace the index 1 with any index $j \in \widetilde{\mathcal{V}}$, then (C.3.18) and (C.3.21) can be correspondingly replaced by

$$\lim_{n\to\infty} \mathbf{P}\left(\|\widetilde{\boldsymbol{\pi}}_{\mathcal{S}^*}^{[j]}\|_0 > \frac{|\mathcal{V}^*|}{2}\right) = 1 \quad \text{for} \quad j \in \mathcal{S}^* \backslash \mathcal{V}^*; \tag{C.3.24}$$

$$\lim_{n\to\infty} \mathbf{P}\left(\|\widetilde{\boldsymbol{\pi}}_{\mathcal{S}^*}^{[j]}\|_0 < \frac{|\mathcal{V}^*|}{2}\right) = 1 \quad \text{for} \quad j \in \mathcal{V}^*. \tag{C.3.25}$$

By (C.3.24) and (C.3.25), we can establish $\lim_{n\to\infty} \mathbf{P}\left(\mathcal{B}_1 \cap \mathcal{B}_2\right) = 1$. By (C.3.20) and (C.3.23), we have

$$\lim_{n\to\infty} \mathbf{P}\left(\operatorname{supp}\left(\widetilde{\boldsymbol{\pi}}_{\mathcal{S}^*}^{[1]}\right) = \operatorname{supp}\left(\boldsymbol{\pi}_{\mathcal{S}^*}\right)\right) = 1. \tag{C.3.26}$$

Similarly, we can obtain that

$$\lim_{n\to\infty} \mathbf{P}\left(\operatorname{supp}\left(\widetilde{\boldsymbol{\pi}}_{\mathcal{S}^*}^{[j]}\right) = \operatorname{supp}\left(\boldsymbol{\pi}_{\mathcal{S}^*}\right) \quad \text{for} \quad j \in \mathcal{V}^*\right) = 1, \tag{C.3.27}$$

and hence $\lim_{n\to\infty} \mathbf{P}\left(\mathcal{B}_1 \cap \mathcal{B}_2 \cap \mathcal{B}_3\right) = 1$.

C.3.2 Proof of Lemma 32

Note that

$$\sqrt{n}\left(\frac{\widetilde{\gamma}_{\mathcal{V}^*}^{\intercal} A(\mathcal{V}^*)\widetilde{\boldsymbol{\Gamma}}_{\mathcal{V}^*}}{\widetilde{\gamma}_{\mathcal{V}^*}^{\intercal} A(\mathcal{V}^*)\widetilde{\gamma}_{\mathcal{V}^*}} - \beta\right) = \frac{\widetilde{\gamma}_{\mathcal{V}^*}^{\intercal} A(\mathcal{V}^*)\left(\widehat{\boldsymbol{\Sigma}}^{-1}\right)_{\mathcal{V}^*\cdot}}{\widetilde{\gamma}_{\mathcal{V}^*}^{\intercal} A(\mathcal{V}^*)\widetilde{\gamma}_{\mathcal{V}^*}} \frac{1}{\sqrt{n}} \mathbf{W}^{\intercal}\left(\boldsymbol{\Pi}_{\cdot 2} - \beta \boldsymbol{\Pi}_{\cdot 1}\right). \tag{C.3.28}$$

Since

$$\frac{\widetilde{\gamma}_{\mathcal{V}^*}^{\intercal} A(\mathcal{V}^*)\left(\widehat{\boldsymbol{\Sigma}}^{-1}\right)_{\mathcal{V}^*\cdot}}{\widetilde{\gamma}_{\mathcal{V}^*}^{\intercal} A(\mathcal{V}^*)\widetilde{\gamma}_{\mathcal{V}^*}} \xrightarrow{p} \frac{\boldsymbol{\gamma}_{\mathcal{V}^*}^{\intercal} A^*(\mathcal{V}^*)\left(\boldsymbol{\Sigma}^{-1}\right)_{\mathcal{V}^*\cdot}}{\boldsymbol{\gamma}_{\mathcal{V}^*}^{\intercal} A^*(\mathcal{V}^*)\boldsymbol{\gamma}_{\mathcal{V}^*}} \tag{C.3.29}$$

and

$$\frac{1}{\sqrt{n}} \mathbf{W}^{\intercal}\left(\boldsymbol{\Pi}_{\cdot 2} - \beta \boldsymbol{\Pi}_{\cdot 1}\right) \xrightarrow{d} N\left(0, \left(\boldsymbol{\Theta}_{11} + \beta^2 \boldsymbol{\Theta}_{22} - 2\beta \boldsymbol{\Theta}_{12}\right)\boldsymbol{\Sigma}\right), \tag{C.3.30}$$

we have

$$\begin{aligned} &\frac{\widetilde{\gamma}_{\mathcal{V}^*}^{\intercal} A(\mathcal{V}^*)\left(\widehat{\boldsymbol{\Sigma}}^{-1}\right)_{\mathcal{V}^*\cdot}}{\widetilde{\gamma}_{\mathcal{V}^*}^{\intercal} A(\mathcal{V}^*)\widetilde{\gamma}_{\mathcal{V}^*}} \frac{1}{\sqrt{n}} \mathbf{W}^{\intercal}\left(\boldsymbol{\Pi}_{\cdot 2} - \beta \boldsymbol{\Pi}_{\cdot 1}\right) \\ &\xrightarrow{d} N\left(0, \frac{\boldsymbol{\gamma}_{\mathcal{V}^*}^{\intercal} A^*(\mathcal{V}^*)\left(\boldsymbol{\Sigma}^{-1}\right)_{\mathcal{V}^*\mathcal{V}^*} A^*(\mathcal{V}^*)\boldsymbol{\gamma}_{\mathcal{V}^*}}{\left(\boldsymbol{\gamma}_{\mathcal{V}^*}^{\intercal} A^*(\mathcal{V}^*)\boldsymbol{\gamma}_{\mathcal{V}^*}\right)^2}\left(\boldsymbol{\Theta}_{11} + \beta^2 \boldsymbol{\Theta}_{22} - 2\beta \boldsymbol{\Theta}_{12}\right)\right) \end{aligned} \tag{C.3.31}$$

Since $A^*(\mathcal{V}^*)\left(\boldsymbol{\Sigma}^{-1}\right)_{\mathcal{V}^*\mathcal{V}^*} A^*(\mathcal{V}^*) = A^*(\mathcal{V}^*)$, we establish (C.2.4).

C.3.3 Lemmas for scaled Lasso and de-biasing Lasso

We introduce the following lemmas for scaled Lasso and de-biasing Lasso used in the later proofs. Lemma 40 establishes the convergence rate of the scaled Lasso method, which is based on the analysis in Sun & Zhang (2012).

Lemma 40. *On the event $G \cap S$, if $s \leq cn/\log p$, then*

$$\|\widehat{\mathbf{\Gamma}}-\mathbf{\Gamma}\|_1+\|\widehat{\mathbf{\Psi}}-\mathbf{\Psi}\|_1 \leq Cs\sqrt{\frac{\log p}{n}}\sigma_1, \quad \|\widehat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}\|_1+\|\widehat{\boldsymbol{\psi}}-\boldsymbol{\psi}\|_1 \leq Cs\sqrt{\frac{\log p}{n}}\sigma_2, \tag{C.3.32}$$

$$\frac{1}{\sqrt{n}}\|\boldsymbol{Z}(\widehat{\mathbf{\Gamma}}-\mathbf{\Gamma})+\boldsymbol{X}(\widehat{\mathbf{\Psi}}-\mathbf{\Psi})\|_2 \leq C\sqrt{\frac{s\log p}{n}}\sigma_1, \tag{C.3.33}$$

and

$$\frac{1}{\sqrt{n}}\|\boldsymbol{Z}(\widehat{\boldsymbol{\gamma}}-\boldsymbol{\gamma})+\boldsymbol{X}(\widehat{\boldsymbol{\psi}}-\boldsymbol{\psi})\|_2 \leq C\sqrt{\frac{s\log p}{n}}\sigma_2. \tag{C.3.34}$$

The following lemma is the key result for the de-biasing Lasso estimator, established in Zhang & Zhang (2014); Javanmard & Montanari (2014a); van de Geer et al. (2014).

Lemma 41. *We have the following expressions for the proposed de-biased estimator,*

$$\widetilde{\mathbf{\Gamma}}-\mathbf{\Gamma} = D^{\mathbf{\Gamma}}+\Delta^{\mathbf{\Gamma}}, \tag{C.3.35}$$

where

$$D_j^{\mathbf{\Gamma}} = \frac{1}{n}\boldsymbol{v}^\intercal\mathbf{\Pi}_{\cdot 1} \quad \textit{and} \quad \Delta_j^{\mathbf{\Gamma}} = \left(\frac{1}{n}(\widehat{\mathbf{u}}^{[j]})^\intercal\widehat{\mathbf{\Sigma}}-e_j^\intercal\right)\begin{pmatrix}\widehat{\mathbf{\Gamma}}-\mathbf{\Gamma}\\ \widehat{\mathbf{\Psi}}-\mathbf{\Psi}\end{pmatrix},\ i=1,\cdots,p_{\mathrm{z}}. \tag{C.3.36}$$

We also have

$$\widetilde{\boldsymbol{\gamma}}-\boldsymbol{\gamma} = D^{\boldsymbol{\gamma}}+\Delta^{\boldsymbol{\gamma}}, \tag{C.3.37}$$

where

$$D_j^{\boldsymbol{\gamma}} = \frac{1}{n}\boldsymbol{v}^\intercal\mathbf{\Pi}_{\cdot 2} \quad \textit{and} \quad \Delta_j^{\boldsymbol{\gamma}} = \left(\frac{1}{n}(\widehat{\mathbf{u}}^{[j]})^\intercal\widehat{\mathbf{\Sigma}}-e_j^\intercal\right)\begin{pmatrix}\widehat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}\\ \widehat{\boldsymbol{\psi}}-\boldsymbol{\psi}\end{pmatrix},\ i=1,\cdots,p_{\mathrm{z}}. \tag{C.3.38}$$

On the event $S \cap G \cap A$, we have

$$\max\left\{\|\Delta^{\gamma}\|_{\infty}, \|\Delta^{\mathbf{\Gamma}}\|_{\infty}\right\} \leq Cs\frac{\log p}{n}\max\left\{\sigma_1, \sigma_2\right\}. \tag{C.3.39}$$

C.3.4 Proof of Lemma 33

The proof of Lemma 33 is a generalization of Lemma 4 in Cai & Guo (2016b). In the following, we extend the Gaussian design in Cai & Guo (2016b) to sub-gaussian design considered in this paper. Since the error of the regression is still assumed to be Gaussian, it is sufficient to establish the probability bound of G_1, G_3, G_4 and A_1 for the sub-gaussian design matrix and control the events A_2 and A_3. The probability bound of the event A_1 for the sub-gaussian design is established in Lemma 4 of Cai & Guo (2016c). By Corollary 5.17 in Vershynin (2012) and the union bound, we have

$$\mathbf{P}\left(\max_{1\leq j\leq p}\left|\frac{1}{n}\left(\|\mathbf{W}_{\cdot j}\|_2^2 - \mathbb{E}\|\mathbf{W}_{\cdot j}\|_2^2\right)\right| \geq \epsilon\right) \leq 2p\exp\left(-\frac{1}{6}\min\left\{\frac{\epsilon^2}{K^2}, \frac{\epsilon}{K}\right\}n\right),$$

where $K = 4M_1$. Taking $\epsilon = 12M_1\sqrt{\log p/n}$, we have

$$\mathbf{P}\left(\max_{1\leq j\leq p}\left|\frac{1}{n}\left(\|\mathbf{W}_{\cdot j}\|_2^2 - \mathbb{E}\|\mathbf{W}_{\cdot j}\|_2^2\right)\right| \geq 12M_1\sqrt{\frac{\log p}{n}}\right) \leq 2p^{-\frac{1}{2}} \quad \text{and} \quad \mathbf{P}\left(G_1\right) \geq 1 - 2p^{-\frac{1}{2}}. \tag{C.3.40}$$

Similarly, we have $\mathbf{P}\left(\left|\frac{1}{n}\left(\|\mathbf{W}u\|_2^2 - \mathbb{E}\|\mathbf{W}u\|_2^2\right)\right| \geq 12M_1\|u\|^2\sqrt{\log p/n}\right) \leq 2p^{-\frac{3}{2}}$, and

$$\mathbf{P}\left(\left|\frac{1}{n}\left(\frac{\|\mathbf{W}u\|_2^2}{\mathbb{E}\|\mathbf{W}u\|_2^2} - 1\right)\right| \geq 12M_1\frac{\|u\|^2}{\mathbb{E}\|\mathbf{W}u\|_2^2}\sqrt{\log p/n}\right) \leq 2p^{-\frac{3}{2}},$$

and hence

$$\mathbf{P}\left(G_3\right) \geq 1 - (p_{\mathrm{z}} + 2)p^{-\frac{3}{2}}.$$

By Theorem 1.6 in Zhou (2009), if $n \geq 1/\theta^2 \times c' M_1^3 \max\left\{12(2+\gamma_0)^2 M_1 s \log\left(5ep/4s\right), 9\log p\right\}$, then with probability at least $1 - 2\exp\left(-c\theta^2 n/M_1^3\right)$, for all δ such that there exists

$|J_0| \leq 4s$ and $\|\delta_{J_0^c}\|_1 \leq \gamma_0 \|\delta_{J_0}\|_1$, we have $\|Z\delta\|_2/(\sqrt{n}\|\mathbf{\Sigma}^{\frac{1}{2}}\delta\|_2) \geq 1-\theta$. By taking $\theta = \frac{1}{2}$, if $n \geq 4c' M_1^3 \max\left\{12\left(2+\gamma_0\right)^2 M_1 s \log\left(5ep/4s\right), 9\log p\right\}$, then $\mathbf{P}\left(G_4\right) \geq 1 - 2\exp\left(-cn/M_1^3\right)$. In the following, we control the events A_2 and A_3,

$$\begin{aligned}
\mathbf{P}(A_2^c) \leq & \mathbf{P}\left(\max_{1\leq i\leq q} \frac{|D_j^{\gamma}|}{\sqrt{\mathrm{Var}\left(D_j^{\gamma}\right)}} \geq \sqrt{2.02\log p_{\mathrm{z}}}\right) + \mathbf{P}\left(\max_{1\leq j\leq p_{\mathrm{z}}} \frac{|\Delta_j^{\gamma}|}{\sqrt{\mathrm{Var}\left(D_j^{\gamma}\right)}} \geq 0.01\sqrt{\log p_{\mathrm{z}}}\right) \\
\leq & \frac{1}{2\sqrt{\pi \log p_{\mathrm{z}}}} p_{\mathrm{z}}^{-0.02} + \mathbf{P}\left(\left(S \cap G \cap A_1\right)^c\right),
\end{aligned}$$

where the first inequality follows from (C.3.37) and the second inequality follows from (C.3.38) and (C.3.39). The control of $\mathbf{P}(A_4 \cap A_5)$ follows from (C.2.20) and (C.2.21). Note that

$$\begin{aligned}
\mathbf{P}(A_3^c) \leq & \mathbf{P}\left(\left(A_1 \cap G_1 \cap G_3\right)^c\right) + \mathbf{P}\left(A_3^c \cap A_1 \cap G_1 \cap G_3\right) \\
\leq & \mathbf{P}\left(\left(A_1 \cap G_1 \cap G_3\right)^c\right) + 2\mathbf{P}\left(\max_{1\leq j\leq p_{\mathrm{z}}} \frac{1}{\|\widehat{\mathbf{v}}^{[j]}\|_2 \sigma_1} \left|\boldsymbol{v}^{\intercal}\mathbf{\Pi}_{\cdot 1}\right| \geq \sqrt{2.05\log p_{\mathrm{z}}}\right) \\
\leq & \mathbf{P}\left(\left(A_1 \cap G_1 \cap G_3\right)^c\right) + \frac{1}{\sqrt{\pi \log p_{\mathrm{z}}}} p_{\mathrm{z}}^{-0.05},
\end{aligned}$$

where the second inequality follows from (C.2.14) and the last inequality follows from the fact that $1/(\|\widehat{\mathbf{v}}^{[j]}\|_2 \sigma_1) \times \boldsymbol{v}^{\intercal}\mathbf{\Pi}_{\cdot 1}$ conditioning on $\mathbf{W}$ is normally distributed.

C.3.5 Proof of Lemma 34

In the following, we only establish the results for $\widehat{\mathbf{v}}^{[1]}$ and the same argument extends to $\widehat{\mathbf{v}}^{[j]}$ where $1 \leq j \leq p_{\mathrm{z}}$. Since $\lambda_n = 2eC_0 M_1^2 \sqrt{\log p/n}$ is chosen such that $\mathbf{\Omega}_{1\cdot}$ belongs to the feasible set, we have

$$\frac{\|\widehat{\mathbf{v}}^{[1]}\|_2^2}{n} \leq \frac{\|\mathbf{W}\mathbf{\Omega}_{1\cdot}\|_2^2}{n}. \tag{C.3.41}$$

By Lemma 12 in Javanmard & Montanari (2014a), we have

$$\frac{\|\widehat{\mathbf{v}}^{[1]}\|_2^2}{n} \geq \frac{\left(1-\lambda_n\right)^2}{\widehat{\mathbf{\Sigma}}_{11}}. \tag{C.3.42}$$

By the definition of G_1 and G_3, we establish (C.2.14). Let $\mathcal{I} = \{1, 2, \cdots, p_z\}$ and assume that $M \in \mathbb{R}^{p \times p_z}$ belongs to the feasible set $\|\widehat{\mathbf{\Sigma}}\mathbf{\Omega} - \mathbf{I}_{\cdot\mathcal{I}}\|_\infty \le \lambda_n$, where $\mathbf{I}_{\cdot\mathcal{I}}$ denotes the sub-matrix of the identity matrix containing the column with index $i \in \mathcal{I}$, that is, $\|\widehat{\mathbf{\Sigma}}M - \mathbf{I}_{\cdot\mathcal{I}}\|_\infty \le \lambda_n$, and hence

$$\|\widehat{\mathbf{\Sigma}}M\boldsymbol{\gamma} - \boldsymbol{\gamma}\|_\infty = \|\left(\widehat{\mathbf{\Sigma}}M - \mathbf{I}_{\cdot\mathcal{I}}\right)\boldsymbol{\gamma}\|_\infty \le \|\widehat{\mathbf{\Sigma}}M - \mathbf{I}_{\cdot\mathcal{I}}\|_\infty\|\boldsymbol{\gamma}\|_1 \le \lambda_n\|\boldsymbol{\gamma}\|_1. \quad \text{(C.3.43)}$$

Note that

$$\left|\boldsymbol{\gamma}^\intercal\widehat{\mathbf{\Sigma}}M\boldsymbol{\gamma} - \|\boldsymbol{\gamma}\|_2^2\right| = \left|\boldsymbol{\gamma}^\intercal\left(\widehat{\mathbf{\Sigma}}M\boldsymbol{\gamma} - \boldsymbol{\gamma}\right)\right| \le \|\boldsymbol{\gamma}\|_1\|\widehat{\mathbf{\Sigma}}M\boldsymbol{\gamma} - \boldsymbol{\gamma}\|_\infty \le \lambda_n\|\boldsymbol{\gamma}\|_1^2, \quad \text{(C.3.44)}$$

where the last inequality follows from (C.3.43). The inequality (C.3.44) informs that $M\boldsymbol{\gamma}$ is in the feasible set

$$\left|\boldsymbol{\gamma}\widehat{\mathbf{\Sigma}}\left(M\boldsymbol{\gamma}\right) - \|\boldsymbol{\gamma}\|_2^2\right| \le \lambda_n\|\boldsymbol{\gamma}\|_1^2. \quad \text{(C.3.45)}$$

We define μ^* as

$$\begin{aligned} &\mu^* = \arg\min_{\mu}\mu^\intercal\widehat{\mathbf{\Sigma}}\mu \\ &\text{subject to} \quad \left|\boldsymbol{\gamma}^\intercal\widehat{\mathbf{\Sigma}}\mu - \|\boldsymbol{\gamma}\|_2^2\right| \le \lambda_n\|\boldsymbol{\gamma}\|_1^2 \end{aligned} \quad \text{(C.3.46)}$$

By (C.3.45), we have the following inequality,

$$\frac{1}{n}\left\|\sum_{j\in\mathcal{S}^*}\boldsymbol{\gamma}_j\widehat{\mathbf{v}}^{[j]}\right\|_2^2 = \boldsymbol{\gamma}^\intercal M^\intercal\widehat{\mathbf{\Sigma}}M\boldsymbol{\gamma} \ge (\mu^*)^\intercal\,\widehat{\mathbf{\Sigma}}\mu^*. \quad \text{(C.3.47)}$$

In the following, we will show that $(\mu^*)^\intercal\,\widehat{\mathbf{\Sigma}}\mu^* = \langle\mu^*, \widehat{\mathbf{\Sigma}}\mu^*\rangle$ is further lower bounded. Since μ^* is feasible in the constrained set of (C.3.46), we have $\|\boldsymbol{\gamma}\|_2^2 - \boldsymbol{\gamma}^\intercal\widehat{\mathbf{\Sigma}}\mu^* -$

$\lambda_n\|\boldsymbol{\gamma}\|_1^2 \leq 0$, and hence for any positive constant $c > 0$, we have

$$
\begin{aligned}
&\langle \mu^*, \widehat{\boldsymbol{\Sigma}}\mu^*\rangle \geq \langle \mu^*, \widehat{\boldsymbol{\Sigma}}\mu^*\rangle + c\left(\|\boldsymbol{\gamma}\|_2^2 - \boldsymbol{\gamma}^\intercal\widehat{\boldsymbol{\Sigma}}\mu^* - \lambda_n\|\boldsymbol{\gamma}\|_1^2\right) \\
&\geq \min_{\mu}\left(\langle \mu, \widehat{\boldsymbol{\Sigma}}\mu\rangle + c\left(\|\boldsymbol{\gamma}\|_2^2 - \boldsymbol{\gamma}^\intercal\widehat{\boldsymbol{\Sigma}}\mu - \lambda_n\|\boldsymbol{\gamma}\|_1^2\right)\right) = -\frac{c^2}{4}\langle \boldsymbol{\gamma}, \widehat{\boldsymbol{\Sigma}}\boldsymbol{\gamma}\rangle + c\left(\|\boldsymbol{\gamma}\|_2^2 - \lambda_n\|\boldsymbol{\gamma}\|_1^2\right).
\end{aligned}
\tag{C.3.48}
$$

Note that $\|\boldsymbol{\gamma}\|_1^2\lambda_n \leq s_{z1}\lambda_n\|\boldsymbol{\gamma}\|_2^2 = Cs_{z1}\sqrt{\log p/n}\|\boldsymbol{\gamma}\|_2^2 \ll \|\boldsymbol{\gamma}\|_2^2$, where the last inequality holds when $s_{z1}\sqrt{\log p/n} \to 0$. By (C.3.48), we have

$$
\begin{aligned}
\langle \mu^*, \widehat{\boldsymbol{\Sigma}}\mu^*\rangle &\geq \max_{c>0} -\frac{c^2}{4}\langle \boldsymbol{\gamma}, \widehat{\boldsymbol{\Sigma}}\boldsymbol{\gamma}\rangle + c\left(\|\boldsymbol{\gamma}\|_2^2 - \lambda_n\|\boldsymbol{\gamma}\|_1^2\right) \\
&= \frac{\left(\|\boldsymbol{\gamma}\|_2^2 - \lambda_n\|\boldsymbol{\gamma}\|_1^2\right)^2}{\langle \boldsymbol{\gamma}, \widehat{\boldsymbol{\Sigma}}\boldsymbol{\gamma}\rangle} \geq \frac{\|\boldsymbol{\gamma}\|_2^4\left(1 - s_{z1}\lambda_n\right)^2}{\langle \boldsymbol{\gamma}, \widehat{\boldsymbol{\Sigma}}\boldsymbol{\gamma}\rangle}.
\end{aligned}
\tag{C.3.49}
$$

On the event G_3, we establish (C.2.15) for $\left\|\sum_{j\in\mathcal{S}^*}\boldsymbol{\gamma}_j\widehat{\mathbf{v}}^{[j]}\right\|_2^2/n$. The same argument holds for $\left\|\sum_{j\in\mathcal{V}^*}\boldsymbol{\gamma}_j\widehat{\mathbf{v}}^{[j]}\right\|_2^2/n$. Note that

$$
\frac{1}{M_2} \leq \boldsymbol{\Theta}_{11} + \beta^2\boldsymbol{\Theta}_{22} - 2\beta\boldsymbol{\Theta}_{12} = \begin{pmatrix} 1 & -\beta \end{pmatrix}\boldsymbol{\Theta}\begin{pmatrix} 1 \\ -\beta \end{pmatrix} \leq M_2\left(1 + \beta^2\right). \tag{C.3.50}
$$

Combined with (C.2.15), we establish the first inequality of (C.2.16). Note that

$$
\frac{1}{n}\left\|\sum_{j\in\mathcal{S}^*}\boldsymbol{\gamma}_j\widehat{\mathbf{v}}^{[j]}\right\|_2^2 \leq \left(2M_2\sum_{j\in\mathcal{V}^*}|\boldsymbol{\gamma}_j|\right)^2 \leq s_{z1}\|\boldsymbol{\gamma}\|_2^2. \tag{C.3.51}
$$

Combined with (C.3.50), we establish the second inequality of (C.2.16). By the similar argument, we can establish (C.2.17).

C.3.6 Proof of Lemma 35

In the following proof, we will use the shorthand $\langle a, b\rangle_J = \sum_{j\in J} a_j b_j$. We have the following decompositions for $\widehat{\|\boldsymbol{\gamma}\|_2^2} - \|\boldsymbol{\gamma}\|_2^2$ and $\widehat{\boldsymbol{\gamma}^\intercal\boldsymbol{\Gamma}} - \boldsymbol{\gamma}^\intercal\boldsymbol{\Gamma}$,

$$\begin{aligned}\widehat{\|\boldsymbol{\gamma}\|_2^2} - \|\boldsymbol{\gamma}\|_2^2 =& 2\langle\boldsymbol{\gamma}, D^\gamma\rangle_{\widetilde{\mathcal{S}}} + 2\langle\boldsymbol{\gamma}, \Delta^\gamma\rangle_{\widetilde{\mathcal{S}}} + \langle D^\gamma, D^\gamma\rangle_{\widetilde{\mathcal{S}}} + \langle\Delta^\gamma, \Delta^\gamma\rangle_{\widetilde{\mathcal{S}}} + 2\langle D^\gamma, \Delta^\gamma\rangle_{\widetilde{\mathcal{S}}}\\ &- \left(\sum_{j\in\mathcal{S}^*\setminus\widetilde{\mathcal{S}}}\boldsymbol{\gamma}_j^2 - \sum_{j\in\widetilde{\mathcal{S}}\setminus\mathcal{S}^*}\boldsymbol{\gamma}_j^2\right),\end{aligned} \tag{C.3.52}$$

and

$$\begin{aligned}\widehat{\boldsymbol{\gamma}^\intercal\boldsymbol{\Gamma}} - \boldsymbol{\gamma}^\intercal\boldsymbol{\Gamma} =& \langle\boldsymbol{\gamma}, D^{\boldsymbol{\Gamma}}\rangle_{\widetilde{\mathcal{S}}} + \langle\boldsymbol{\Gamma}, D^\gamma\rangle_{\widetilde{\mathcal{S}}} + \langle\boldsymbol{\gamma}, \Delta^{\boldsymbol{\Gamma}}\rangle_{\widetilde{\mathcal{S}}} + \langle\boldsymbol{\Gamma}, \Delta^\gamma\rangle_{\widetilde{\mathcal{S}}} + \langle D^\gamma, D^{\boldsymbol{\Gamma}}\rangle_{\widetilde{\mathcal{S}}} + \langle\Delta^\gamma, \Delta^{\boldsymbol{\Gamma}}\rangle_{\widetilde{\mathcal{S}}}\\ &+ \langle D^\gamma, \Delta^{\boldsymbol{\Gamma}}\rangle_{\widetilde{\mathcal{S}}} + \langle D^{\boldsymbol{\Gamma}}, \Delta^\gamma\rangle_{\widetilde{\mathcal{S}}} - \left(\sum_{j\in\mathcal{S}^*\setminus\widetilde{\mathcal{S}}}\boldsymbol{\gamma}_j\boldsymbol{\Gamma}_j - \sum_{j\in\widetilde{\mathcal{S}}\setminus\mathcal{S}^*}\boldsymbol{\gamma}_j\boldsymbol{\Gamma}_j\right).\end{aligned} \tag{C.3.53}$$

Recall that $\widehat{\mathbf{v}}^{[j]} = \mathbf{W}^\intercal\widehat{\mathbf{u}}^{[j]}$, then we have the following expression

$$\langle\boldsymbol{\gamma}, D^\gamma\rangle_{\widetilde{\mathcal{S}}} = \frac{1}{n}\sum_{j\in\widetilde{\mathcal{S}}}\boldsymbol{\gamma}_j\boldsymbol{v}^\intercal\boldsymbol{\Pi}_{\cdot 2}, \tag{C.3.54}$$

and

$$\langle\boldsymbol{\gamma}, D^{\boldsymbol{\Gamma}}\rangle_{\widetilde{\mathcal{S}}} + \langle\boldsymbol{\Gamma}, D^\gamma\rangle_{\widetilde{\mathcal{S}}} = \frac{1}{n}\sum_{j\in\widetilde{\mathcal{S}}}\boldsymbol{v}^\intercal\left(\boldsymbol{\gamma}_j\boldsymbol{\Pi}_{\cdot 1} + \boldsymbol{\Gamma}_j\boldsymbol{\Pi}_{\cdot 2}\right). \tag{C.3.55}$$

Note that $\widetilde{\mathcal{S}}$ is correlated with the error $\boldsymbol{\Pi}_{\cdot 1}$ and $\boldsymbol{\Pi}_{\cdot 2}$. However, we can compare $\widetilde{\mathcal{S}}$ with the true support $\mathcal{S}^*$,

$$\begin{aligned}\langle\boldsymbol{\gamma}, D^\gamma\rangle_{\widetilde{\mathcal{S}}} - \langle\boldsymbol{\gamma}, D^\gamma\rangle_{\mathcal{S}^*} &= \frac{1}{n}\sum_{j\in\widetilde{\mathcal{S}}}\boldsymbol{\gamma}_j\boldsymbol{v}^\intercal\boldsymbol{\Pi}_{\cdot 2} - \frac{1}{n}\sum_{j\in\mathcal{S}^*}\boldsymbol{\gamma}_j\boldsymbol{v}^\intercal\boldsymbol{\Pi}_{\cdot 2}\\ &= \frac{1}{n}\sum_{j\in\widetilde{\mathcal{S}}\setminus\mathcal{S}^*}\boldsymbol{\gamma}_j\boldsymbol{v}^\intercal\boldsymbol{\Pi}_{\cdot 2} - \frac{1}{n}\sum_{j\in\mathcal{S}^*\setminus\widetilde{\mathcal{S}}}\boldsymbol{\gamma}_j\boldsymbol{v}^\intercal\boldsymbol{\Pi}_{\cdot 2},\end{aligned} \tag{C.3.56}$$

and

$$\begin{aligned}
&\left(\langle \boldsymbol{\gamma}, D^{\boldsymbol{\Gamma}}\rangle_{\widetilde{\mathcal{S}}} + \langle \boldsymbol{\Gamma}, D^{\boldsymbol{\gamma}}\rangle_{\widetilde{\mathcal{S}}}\right) - \left(\langle \boldsymbol{\gamma}, D^{\boldsymbol{\Gamma}}\rangle_{\mathcal{S}^*} + \langle \boldsymbol{\Gamma}, D^{\boldsymbol{\gamma}}\rangle_{\mathcal{S}^*}\right) \\
=&\frac{1}{n}\sum_{j\in\widetilde{\mathcal{S}}} \boldsymbol{v}^\intercal \left(\boldsymbol{\gamma}_j \boldsymbol{\Pi}_{\cdot 1} + \boldsymbol{\Gamma}_j \boldsymbol{\Pi}_{\cdot 2}\right) - \frac{1}{n}\sum_{j\in\mathcal{S}^*} \boldsymbol{v}^\intercal \left(\boldsymbol{\gamma}_j \boldsymbol{\Pi}_{\cdot 1} + \boldsymbol{\Gamma}_j \boldsymbol{\Pi}_{\cdot 2}\right) \\
=&\frac{1}{n}\sum_{j\in\widetilde{\mathcal{S}}\setminus\mathcal{S}^*} \boldsymbol{v}^\intercal \left(\boldsymbol{\gamma}_j \boldsymbol{\Pi}_{\cdot 1} + \boldsymbol{\Gamma}_j \boldsymbol{\Pi}_{\cdot 2}\right) - \frac{1}{n}\sum_{j\in\mathcal{S}^*\setminus\widetilde{\mathcal{S}}} \boldsymbol{v}^\intercal \left(\boldsymbol{\gamma}_j \boldsymbol{\Pi}_{\cdot 1} + \boldsymbol{\Gamma}_j \boldsymbol{\Pi}_{\cdot 2}\right).
\end{aligned} \tag{C.3.57}$$

Hence, the residual terms are

$$\begin{aligned}
R^{\boldsymbol{\gamma}} =& \sqrt{n}\left(2\langle \boldsymbol{\gamma}, \Delta^{\boldsymbol{\gamma}}\rangle_{\widetilde{\mathcal{S}}} + \langle D^{\boldsymbol{\gamma}}, D^{\boldsymbol{\gamma}}\rangle_{\widetilde{\mathcal{S}}} + \langle \Delta^{\boldsymbol{\gamma}}, \Delta^{\boldsymbol{\gamma}}\rangle_{\widetilde{\mathcal{S}}} + 2\langle D^{\boldsymbol{\gamma}}, \Delta^{\boldsymbol{\gamma}}\rangle_{\widetilde{\mathcal{S}}}\right) \\
&+ \sqrt{n}\left(\frac{1}{n}\sum_{j\in\widetilde{\mathcal{S}}\setminus\mathcal{S}^*} \boldsymbol{\gamma}_j \boldsymbol{v}^\intercal \boldsymbol{\Pi}_{\cdot 2} - \frac{1}{n}\sum_{j\in\mathcal{S}^*\setminus\widetilde{\mathcal{S}}} \boldsymbol{\gamma}_j \boldsymbol{v}^\intercal \boldsymbol{\Pi}_{\cdot 2}\right) - \sqrt{n}\left(\sum_{j\in\mathcal{S}^*\setminus\widetilde{\mathcal{S}}} \boldsymbol{\gamma}_j^2 - \sum_{j\in\widetilde{\mathcal{S}}\setminus\mathcal{S}^*} \boldsymbol{\gamma}_j^2\right),
\end{aligned} \tag{C.3.58}$$

and

$$\begin{aligned}
R^{\text{inter}} =& \sqrt{n}\left(\langle \boldsymbol{\gamma}, \Delta^{\boldsymbol{\Gamma}}\rangle_{\widetilde{\mathcal{S}}} + \langle \boldsymbol{\Gamma}, \Delta^{\boldsymbol{\gamma}}\rangle_{\widetilde{\mathcal{S}}} + \langle D^{\boldsymbol{\gamma}}, D^{\boldsymbol{\Gamma}}\rangle_{\widetilde{\mathcal{S}}} + \langle \Delta^{\boldsymbol{\gamma}}, \Delta^{\boldsymbol{\Gamma}}\rangle_{\widetilde{\mathcal{S}}} + \langle D^{\boldsymbol{\gamma}}, \Delta^{\boldsymbol{\Gamma}}\rangle_{\widetilde{\mathcal{S}}} + \langle D^{\boldsymbol{\Gamma}}, \Delta^{\boldsymbol{\gamma}}\rangle_{\widetilde{\mathcal{S}}}\right) \\
&+ \sqrt{n}\left(\frac{1}{n}\sum_{j\in\widetilde{\mathcal{S}}\setminus\mathcal{S}^*} \boldsymbol{v}^\intercal \left(\boldsymbol{\gamma}_j \boldsymbol{\Pi}_{\cdot 1} + \boldsymbol{\Gamma}_j \boldsymbol{\Pi}_{\cdot 2}\right) - \frac{1}{n}\sum_{j\in\mathcal{S}^*\setminus\widetilde{\mathcal{S}}} \boldsymbol{v}^\intercal \left(\boldsymbol{\gamma}_j \boldsymbol{\Pi}_{\cdot 1} + \boldsymbol{\Gamma}_j \boldsymbol{\Pi}_{\cdot 2}\right)\right) \\
&- \sqrt{n}\left(\sum_{j\in\mathcal{S}^*\setminus\widetilde{\mathcal{S}}} \boldsymbol{\gamma}_j \boldsymbol{\Gamma}_j - \sum_{j\in\widetilde{\mathcal{S}}\setminus\mathcal{S}^*} \boldsymbol{\gamma}_j \boldsymbol{\Gamma}_j\right).
\end{aligned} \tag{C.3.59}$$

Define $\mathcal{S}_0^* = \left\{j : |\boldsymbol{\gamma}_j| > \sqrt{2.05\log p_{\mathrm{z}}}\sqrt{\operatorname{Var}\left(D_j^{\boldsymbol{\gamma}}\right)}\right\}$ to be the set of strong signals, on the event A_2, we have

$$\mathcal{S}_0^* \subset \widetilde{\mathcal{S}} \subset \mathcal{S}^*, \quad \text{and} \quad \left|\widetilde{\mathcal{S}}\right| \leq s_{\mathrm{z}1}. \tag{C.3.60}$$

On the event A_3, we have

$$\max\left\{\left\|D^{\mathbf{\Gamma}}\right\|_{\infty}, \|D^{\gamma}\|_{\infty}\right\} \leq \left(1+12\sqrt{\frac{\log p}{n}}\right) M_1\sqrt{\frac{2.05\log p_{\mathrm{z}}}{n}}\max\{\sigma_1,\sigma_2\}. \quad \text{(C.3.61)}$$

On the event $S\cap G\cap A$,

$$\max\left\{\|\Delta^{\gamma}\|_{\infty}, \|\Delta^{\mathbf{\Gamma}}\|_{\infty}\right\} \leq Cs\frac{\log p}{n}\max\{\sigma_1,\sigma_2\}. \quad \text{(C.3.62)}$$

Combing (C.3.60), (C.3.61) and (C.3.62), we have on the event $S\cap G\cap A$,

$$\max\left\{\langle D^{\gamma}, D^{\gamma}\rangle_{\widetilde{S}}, \langle D^{\mathbf{\Gamma}}, D^{\mathbf{\Gamma}}\rangle_{\widetilde{S}}\right\} \leq Cs_{\mathrm{z}1}\frac{\log p_{\mathrm{z}}}{n}, \quad \text{(C.3.63)}$$

$$\max\left\{\langle \Delta^{\gamma}, \Delta^{\gamma}\rangle_{\widetilde{S}}, \langle \Delta^{\mathbf{\Gamma}}, \Delta^{\mathbf{\Gamma}}\rangle_{\widetilde{S}}\right\} \leq Cs_{\mathrm{z}1}\left(s\frac{\log p}{n}\right)^2. \quad \text{(C.3.64)}$$

Note that

$$\left|\langle D^{\gamma}, \Delta^{\gamma}\rangle_{\widetilde{S}}\right| \leq \sqrt{\langle D^{\gamma}, D^{\gamma}\rangle_{\widetilde{S}}\langle \Delta^{\gamma}, \Delta^{\gamma}\rangle_{\widetilde{S}}} \leq \frac{1}{2}\left(\langle D^{\gamma}, D^{\gamma}\rangle_{\widetilde{S}} + \langle \Delta^{\gamma}, \Delta^{\gamma}\rangle_{\widetilde{S}}\right).$$

Hence, we have

$$\left|\langle D^{\gamma}, D^{\gamma}\rangle_{\widetilde{S}} + \langle \Delta^{\gamma}, \Delta^{\gamma}\rangle_{\widetilde{S}} + 2\langle D^{\gamma}, \Delta^{\gamma}\rangle_{\widetilde{S}}\right| \leq Cs_{\mathrm{z}1}\frac{\log p_{\mathrm{z}}}{n} + Cs_{\mathrm{z}1}\left(s\frac{\log p}{n}\right)^2, \quad \text{(C.3.65)}$$

and

$$\left|\langle D^{\gamma}, D^{\mathbf{\Gamma}}\rangle_{\widetilde{S}} + \langle \Delta^{\gamma}, \Delta^{\mathbf{\Gamma}}\rangle_{\widetilde{S}} + \langle D^{\gamma}, \Delta^{\mathbf{\Gamma}}\rangle_{\widetilde{S}} + \langle D^{\mathbf{\Gamma}}, \Delta^{\gamma}\rangle_{\widetilde{S}}\right| \leq Cs_{\mathrm{z}1}\frac{\log p_{\mathrm{z}}}{n} + Cs_{\mathrm{z}1}\left(s\frac{\log p}{n}\right)^2. \quad \text{(C.3.66)}$$

We also have the following control

$$2\left|\langle \boldsymbol{\gamma}, \Delta^{\gamma}\rangle_{\widetilde{S}}\right| \leq \|\boldsymbol{\gamma}\|_2\sqrt{\langle \Delta^{\gamma}, \Delta^{\gamma}\rangle_{\widetilde{S}}} \leq C\|\boldsymbol{\gamma}\|_2\sqrt{s_{\mathrm{z}1}}s\frac{\log p}{n}, \quad \text{(C.3.67)}$$

and

$$\left|\langle\boldsymbol{\gamma},\Delta^{\boldsymbol{\Gamma}}\rangle_{\widetilde{\mathcal{S}}}+\langle\boldsymbol{\Gamma},\Delta^{\gamma}\rangle_{\widetilde{\mathcal{S}}}\right|\leq C\left(\|\boldsymbol{\gamma}\|_2+\|\boldsymbol{\Gamma}\|_2\right)\sqrt{s_{\mathrm{z}1}}s\frac{\log p}{n}. \tag{C.3.68}$$

On the event $S\cap G\cap A$, we have $\widetilde{\mathcal{S}}\backslash\mathcal{S}^*=\emptyset$ and hence $\frac{1}{n}\sum_{j\in\widetilde{\mathcal{S}}\backslash\mathcal{S}^*}\boldsymbol{\gamma}_j\boldsymbol{v}^\intercal\boldsymbol{\Pi}_{\cdot 2}=0$, $\sum_{j\in\widetilde{\mathcal{S}}\backslash\mathcal{S}^*}\boldsymbol{\gamma}_j\boldsymbol{\Gamma}_j=0$ and $\sum_{j\in\widetilde{\mathcal{S}}\backslash\mathcal{S}^*}\boldsymbol{\gamma}_j^2=0$; On the event $S\cap G\cap A$, we also have

$$\begin{aligned}&\left|\frac{1}{n}\sum_{j\in\mathcal{S}^*\backslash\widetilde{\mathcal{S}}}\boldsymbol{\gamma}_j\boldsymbol{v}^\intercal\boldsymbol{\Pi}_{\cdot 2}\right|\leq\frac{1}{n}s_{\mathrm{z}1}\max_{j\in\mathcal{S}^*\backslash\widetilde{\mathcal{S}}}|\boldsymbol{\gamma}_j|\left|\boldsymbol{v}^\intercal\boldsymbol{\Pi}_{\cdot 2}\right|\\ \leq& s_{\mathrm{z}1}\sqrt{2.05\log p_{\mathrm{z}}}\sqrt{\mathrm{Var}\left(D_j^{\gamma}\right)}\left(1+12\sqrt{\frac{\log p}{n}}\right)M_1\sqrt{\frac{2.05\log p_{\mathrm{z}}}{n}}\sigma_2\leq\frac{s_{\mathrm{z}1}\log p_{\mathrm{z}}}{n}.\end{aligned} \tag{C.3.69}$$

On the event $S\cap G\cap A$, we get

$$\left|\sum_{j\in\mathcal{S}^*\backslash\widetilde{\mathcal{S}}}\boldsymbol{\gamma}_j^2\right|\leq s_{\mathrm{z}1}\frac{\log p_{\mathrm{z}}}{n}. \tag{C.3.70}$$

By (C.3.58), (C.3.65), (C.3.67), (C.3.69) and (C.3.70), we establish that on the event $S\cap G\cap A$,

$$|R^{\gamma}|\leq Cs_{\mathrm{z}1}\frac{\log p_{\mathrm{z}}}{\sqrt{n}}+C\|\boldsymbol{\gamma}\|_2\sqrt{s_{\mathrm{z}1}}s\frac{\log p}{\sqrt{n}}. \tag{C.3.71}$$

Similarly, we can establish that and

$$\left|R^{\mathrm{inter}}\right|\leq Cs_{\mathrm{z}1}\frac{\log p_{\mathrm{z}}}{\sqrt{n}}+C\left(\|\boldsymbol{\gamma}\|_2+\|\boldsymbol{\Gamma}\|_2\right)\sqrt{s_{\mathrm{z}1}}s\frac{\log p}{\sqrt{n}}. \tag{C.3.72}$$

We can establish (C.2.19) and (C.2.22) by taking $\boldsymbol{\Gamma}_j=\beta\boldsymbol{\gamma}_j$. Note that

$$\begin{aligned}\frac{2\sqrt{\log p}}{\sqrt{n}}\left\|\sum_{j\in\mathcal{S}^*}\boldsymbol{\gamma}_j\widehat{\mathbf{v}}^{[j]}\right\|_2\sqrt{\boldsymbol{\Theta}_{22}}&=2\sqrt{\frac{\log p}{n}}\sqrt{\frac{\boldsymbol{\Theta}_{22}}{\boldsymbol{\Theta}_{11}+\beta^2\boldsymbol{\Theta}_{22}-2\beta\boldsymbol{\Theta}_{12}}}\sqrt{\mathrm{V}_{\mathrm{H}}}\|\boldsymbol{\gamma}\|_2^2,\\ \frac{\sqrt{\log p}}{\sqrt{n}}\left\|\sum_{j\in\mathcal{S}^*}\boldsymbol{\gamma}_j\widehat{\mathbf{v}}^{[j]}\right\|_2\sqrt{\boldsymbol{\Theta}_{11}+\beta^2\boldsymbol{\Theta}_{22}+2\beta\boldsymbol{\Theta}_{12}}&=\sqrt{\frac{\log p}{n}}\sqrt{\frac{\boldsymbol{\Theta}_{11}+\beta^2\boldsymbol{\Theta}_{22}+2\beta\boldsymbol{\Theta}_{12}}{\boldsymbol{\Theta}_{11}+\beta^2\boldsymbol{\Theta}_{22}-2\beta\boldsymbol{\Theta}_{12}}}\sqrt{\mathrm{V}_{\mathrm{H}}}\|\boldsymbol{\gamma}\|_2^2.\end{aligned} \tag{C.3.73}$$

By the definition of A_4 and A_5 in (C.2.9) and Lemma 34, we establish

$$\max\left\{\left|\frac{2}{n}\sum_{j\in\mathcal{S}^*}\boldsymbol{\gamma}_j\boldsymbol{v}^\intercal\boldsymbol{\Pi}_{\cdot 2}\right|, \left|\frac{1}{n}\sum_{j\in\mathcal{S}^*}\boldsymbol{\gamma}_j\boldsymbol{v}^\intercal\left(\boldsymbol{\Pi}_{\cdot 1}+\beta\boldsymbol{\Pi}_{\cdot 2}\right)\right|\right\} \leq C s_{\mathrm{z}1}\sqrt{\frac{\log p}{n}}\|\boldsymbol{\gamma}\|_2 \quad \text{(C.3.74)}$$

Combined with (C.2.22), we establish (C.2.23).

C.3.7 Proof of Lemma 38

The proof of Lemma 38 is similar to that of Lemma 35 and we will present it here. Similar to (C.3.52) and (C.3.53), we can obtain the following expressions,

$$\sum_{j\in\widetilde{\mathcal{V}}}\widetilde{\boldsymbol{\gamma}}_j^2-\sum_{j\in\widetilde{\mathcal{V}}}\boldsymbol{\gamma}_j^2=\frac{1}{n}\sum_{j\in\widetilde{\mathcal{V}}}\boldsymbol{\gamma}_j\boldsymbol{v}^\intercal\boldsymbol{\Pi}_{\cdot 2}+\langle D^\gamma,D^\gamma\rangle_{\widetilde{\mathcal{V}}}+\langle\Delta^\gamma,\Delta^\gamma\rangle_{\widetilde{\mathcal{V}}}+2\langle D^\gamma,\Delta^\gamma\rangle_{\widetilde{\mathcal{V}}}, \quad \text{(C.3.75)}$$

and

$$\begin{aligned}\sum_{j\in\widetilde{\mathcal{V}}}\widetilde{\boldsymbol{\gamma}}_j\widetilde{\boldsymbol{\Gamma}}_j-\sum_{j\in\widetilde{\mathcal{V}}}\boldsymbol{\gamma}_j\boldsymbol{\Gamma}_j=&\frac{1}{n}\sum_{j\in\widetilde{\mathcal{V}}}\boldsymbol{v}^\intercal\left(\boldsymbol{\gamma}_j\boldsymbol{\Pi}_{\cdot 1}+\boldsymbol{\Gamma}_j\boldsymbol{\Pi}_{\cdot 2}\right)+\langle\boldsymbol{\gamma},\Delta^{\boldsymbol{\Gamma}}\rangle_{\widetilde{\mathcal{V}}}+\langle\boldsymbol{\Gamma},\Delta^\gamma\rangle_{\widetilde{\mathcal{V}}},\\ &+\langle D^\gamma,D^{\boldsymbol{\Gamma}}\rangle_{\widetilde{\mathcal{V}}}+\langle\Delta^\gamma,\Delta^{\boldsymbol{\Gamma}}\rangle_{\widetilde{\mathcal{V}}}+\langle D^\gamma,\Delta^{\boldsymbol{\Gamma}}\rangle_{\widetilde{\mathcal{V}}}+\langle D^{\boldsymbol{\Gamma}},\Delta^\gamma\rangle_{\widetilde{\mathcal{V}}},\end{aligned} \quad \text{(C.3.76)}$$

On the event $A\cap S\cap G$, we have

$$\frac{1}{n}\sum_{j\in\widetilde{\mathcal{V}}}\boldsymbol{\gamma}_j\boldsymbol{v}^\intercal\boldsymbol{\Pi}_{\cdot 2}\leq\sqrt{|\widetilde{\mathcal{V}}|}\|\boldsymbol{\gamma}\|_2\max_{j\in\widetilde{\mathcal{V}}}\frac{\|\widehat{\mathbf{v}}^{[j]}\|_2}{n}\sqrt{2.05\log p_{\mathrm{z}}}\leq C\|\boldsymbol{\gamma}\|_2\sqrt{\frac{|\widetilde{\mathcal{V}}|\log p_{\mathrm{z}}}{n}}, \quad \text{(C.3.77)}$$

where the last inequality follows from (C.2.14). Combined with the fact $\boldsymbol{\Gamma}_j=\boldsymbol{\gamma}_j(\beta+\boldsymbol{\pi}_j/\boldsymbol{\gamma}_j)$ and $|\boldsymbol{\pi}_j/\boldsymbol{\gamma}_j|\leq C\sqrt{\log p_{\mathrm{z}}/n}$, we also establish that

$$\frac{1}{n}\sum_{j\in\widetilde{\mathcal{V}}}\boldsymbol{v}^\intercal\left(\boldsymbol{\gamma}_j\boldsymbol{\Pi}_{\cdot 1}+\boldsymbol{\Gamma}_j\boldsymbol{\Pi}_{\cdot 2}\right)\leq\sqrt{|\widetilde{\mathcal{V}}|}\|\boldsymbol{\gamma}\|_2\max_{j\in\widetilde{\mathcal{V}}}\frac{\|\widehat{\mathbf{v}}^{[j]}\|_2}{n}\sqrt{2.05\log p_{\mathrm{z}}}\leq C\|\boldsymbol{\gamma}\|_2\sqrt{\frac{|\widetilde{\mathcal{V}}|\log p_{\mathrm{z}}}{n}}. \quad \text{(C.3.78)}$$

By (C.3.77), (C.3.78), (C.3.65) and (C.3.66), we obtain the following inequalities

$$\begin{aligned}&\max\left\{\left|\sum_{j\in\widetilde{\mathcal{V}}}\widetilde{\boldsymbol{\gamma}}_j\widetilde{\boldsymbol{\Gamma}}_j-\sum_{j\in\widetilde{\mathcal{V}}}\boldsymbol{\gamma}_j\boldsymbol{\Gamma}_j\right|,\left|\sum_{j\in\widetilde{\mathcal{V}}}\widetilde{\boldsymbol{\gamma}}_j^2-\sum_{j\in\widetilde{\mathcal{V}}}\boldsymbol{\gamma}_j^2\right|\right\}\\&\leq Cs_{\mathrm{z}1}\frac{\log p_{\mathrm{z}}}{n}+Cs_{\mathrm{z}1}\left(s\frac{\log p}{n}\right)^2+C\|\boldsymbol{\gamma}_{\widetilde{\mathcal{V}}}\|_2\sqrt{\frac{2|\widetilde{\mathcal{V}}|\log p_{\mathrm{z}}}{n}}.\end{aligned}\tag{C.3.79}$$

Since $|\widetilde{\mathcal{V}}|>|\mathcal{V}^*|/2$ and $\min_{j\in\mathcal{S}^*}|\boldsymbol{\gamma}_j|\geq\delta_{\min}\gg\sqrt{\log p_{\mathrm{z}}/n}$, we have $\left|\sum_{j\in\widetilde{\mathcal{V}}}\widetilde{\boldsymbol{\gamma}}_j^2-\sum_{j\in\widetilde{\mathcal{V}}}\boldsymbol{\gamma}_j^2\right|\ll\sum_{j\in\widetilde{\mathcal{V}}}\boldsymbol{\gamma}_j^2$. We have the following decomposition,

$$\begin{aligned}&\left|\frac{\sum_{j\in\widetilde{\mathcal{V}}}\widetilde{\boldsymbol{\gamma}}_j\widetilde{\boldsymbol{\Gamma}}_j}{\sum_{j\in\widetilde{\mathcal{V}}}\widetilde{\boldsymbol{\gamma}}_j^2}-\frac{\sum_{j\in\widetilde{\mathcal{V}}}\boldsymbol{\gamma}_j\boldsymbol{\Gamma}_j}{\sum_{j\in\widetilde{\mathcal{V}}}\boldsymbol{\gamma}_j^2}\right|\\&\leq\frac{\left(\sum_{j\in\widetilde{\mathcal{V}}}\boldsymbol{\gamma}_j^2\right)\left|\sum_{j\in\widetilde{\mathcal{V}}}\widetilde{\boldsymbol{\gamma}}_j\widetilde{\boldsymbol{\Gamma}}_j-\sum_{j\in\widetilde{\mathcal{V}}}\boldsymbol{\gamma}_j\boldsymbol{\Gamma}_j\right|+\left|\sum_{j\in\widetilde{\mathcal{V}}}\boldsymbol{\gamma}_j\boldsymbol{\Gamma}_j\right|\left|\sum_{j\in\widetilde{\mathcal{V}}}\widetilde{\boldsymbol{\gamma}}_j^2-\sum_{j\in\widetilde{\mathcal{V}}}\boldsymbol{\gamma}_j^2\right|}{\left(\sum_{j\in\widetilde{\mathcal{V}}}\widetilde{\boldsymbol{\gamma}}_j^2\right)\left(\sum_{j\in\widetilde{\mathcal{V}}}\boldsymbol{\gamma}_j^2\right)}\\&\leq C\frac{\max\left\{\left|\sum_{j\in\widetilde{\mathcal{V}}}\widetilde{\boldsymbol{\gamma}}_j\widetilde{\boldsymbol{\Gamma}}_j-\sum_{j\in\widetilde{\mathcal{V}}}\boldsymbol{\gamma}_j\boldsymbol{\Gamma}_j\right|,\left|\sum_{j\in\widetilde{\mathcal{V}}}\widetilde{\boldsymbol{\gamma}}_j^2-\sum_{j\in\widetilde{\mathcal{V}}}\boldsymbol{\gamma}_j^2\right|\right\}}{\sum_{j\in\widetilde{\mathcal{V}}}\boldsymbol{\gamma}_j^2}\leq C\frac{1}{\delta_{\min}}\sqrt{\frac{\log p_{\mathrm{z}}}{n}},\end{aligned}\tag{C.3.80}$$

where the last inequality follows from (C.3.79) and the facts that $|\widetilde{\mathcal{V}}|>|\mathcal{V}^*|/2$, $\min_{j\in\mathcal{S}^*}|\boldsymbol{\gamma}_j|\geq\delta_{\min}\gg\sqrt{\log p_{\mathrm{z}}/n}$ and $s\log p/\sqrt{n}\to 0$.

C.3.8 Proof of Lemma 37

Since $\min_{j\in\mathcal{S}^*}|\gamma_j|\geq\delta_{\min}\gg\sqrt{\log p/n}$, on the event $A\cap S\cap G$, we have

$$\widetilde{\mathcal{S}}=\mathcal{S}^*.$$

Without loss of generality, we assume $1\in\widetilde{\mathcal{S}}$ and focus on the case $i=1$. In the following, we are going to analyze the performance of $\widehat{\beta}^{[1]}$ and $\widetilde{\boldsymbol{\pi}}_j^{[1]}$. In the following, we first analyze $\widehat{\beta}^{[1]}$,

$$\sqrt{n}\left(\widehat{\beta}^{[1]} - \left(\beta + \frac{\boldsymbol{\pi}_1}{\boldsymbol{\gamma}_1}\right)\right) = T^{\beta,1} + \Delta^{\beta,1}, \tag{C.3.81}$$

where

$$T^{\beta,1} = \frac{1}{\sqrt{n}\boldsymbol{\gamma}_1}(\widehat{\mathbf{v}}^{[1]})^{\intercal}\left(\boldsymbol{\Pi}_{\cdot 1} - \left(\beta + \frac{\boldsymbol{\pi}_1}{\boldsymbol{\gamma}_1}\right)\boldsymbol{\Pi}_{\cdot 2}\right) \quad \text{and} \quad \Delta^{\beta,1} = R_1 + R_2, \tag{C.3.82}$$

with

$$R_1 = \frac{\sqrt{n}}{\boldsymbol{\gamma}_1}\left(\Delta_1^{\boldsymbol{\Gamma}} - \left(\beta + \frac{\boldsymbol{\pi}_1}{\boldsymbol{\gamma}_1}\right)\Delta_1^{\boldsymbol{\gamma}}\right), \text{ and } R_2 = \frac{-\left(D_1^{\boldsymbol{\gamma}} + \Delta_1^{\boldsymbol{\gamma}}\right)}{\boldsymbol{\gamma}_1 + \left(D_1^{\boldsymbol{\gamma}} + \Delta_1^{\boldsymbol{\gamma}}\right)}\left(T^{\beta,1} + R_1\right). \tag{C.3.83}$$

To analyze $\widetilde{\boldsymbol{\pi}}^{[1]}$, we first analyze the following estimator,

$$\widehat{\boldsymbol{\pi}}^{[1]} = \widetilde{\boldsymbol{\Gamma}} - \widehat{\beta}^{[1]}\widetilde{\boldsymbol{\gamma}}. \tag{C.3.84}$$

Note that

$$\begin{aligned}\widehat{\boldsymbol{\pi}}_j^{[1]} - \boldsymbol{\pi}_j = &-\frac{\boldsymbol{\pi}_1}{\boldsymbol{\gamma}_1}\boldsymbol{\gamma}_j + \left(\widetilde{\boldsymbol{\Gamma}}_j - \boldsymbol{\Gamma}_j\right) - \left(\beta + \frac{\boldsymbol{\pi}_1}{\boldsymbol{\gamma}_1}\right)\left(\widetilde{\boldsymbol{\gamma}}_j - \boldsymbol{\gamma}_j\right) - \boldsymbol{\gamma}_j\left(\widehat{\beta}^{[1]} - \left(\beta + \frac{\boldsymbol{\pi}_1}{\boldsymbol{\gamma}_1}\right)\right)\\ &- \left(\widehat{\beta}^{[1]} - \left(\beta + \frac{\boldsymbol{\pi}_1}{\boldsymbol{\gamma}_1}\right)\right)\left(\widetilde{\boldsymbol{\gamma}}_j - \boldsymbol{\gamma}_j\right).\end{aligned} \tag{C.3.85}$$

By (C.3.35) and (C.3.36), we have

$$\sqrt{n}\left(\widetilde{\boldsymbol{\Gamma}}_j - \boldsymbol{\Gamma}_j\right) = \frac{1}{\sqrt{n}}\boldsymbol{v}^{\intercal}\boldsymbol{\Pi}_{\cdot 1} + \sqrt{n}\Delta_j^{\boldsymbol{\Gamma}}. \tag{C.3.86}$$

By (C.3.37) and (C.3.38), we have

$$\sqrt{n}\left(\widetilde{\boldsymbol{\gamma}}_j - \boldsymbol{\gamma}_j\right) = \frac{1}{\sqrt{n}}\boldsymbol{v}^{\intercal}\boldsymbol{\Pi}_{\cdot 2} + \sqrt{n}\Delta_j^{\boldsymbol{\gamma}}. \tag{C.3.87}$$

By plugging (C.3.82), (C.3.86) and (C.3.87) into (C.3.85), we have the following decomposition of $\widehat{\boldsymbol{\pi}}_j^{[1]} - \boldsymbol{\pi}_j$

$$\sqrt{n}\left(\widehat{\boldsymbol{\pi}}_j^{[1]} - \boldsymbol{\pi}_j\right) = -\sqrt{n}\frac{\boldsymbol{\pi}_1}{\boldsymbol{\gamma}_1}\boldsymbol{\gamma}_j + T^{\boldsymbol{\pi}_j} + \Delta^{\boldsymbol{\pi}_j}, \tag{C.3.88}$$

where

$$T^{\boldsymbol{\pi}_j} = \frac{1}{\sqrt{n}}\left(\boldsymbol{v}^{\intercal} - \frac{\boldsymbol{\gamma}_j}{\boldsymbol{\gamma}_1}(\widehat{\mathbf{v}}^{[1]})^{\intercal}\right)\left(\boldsymbol{\Pi}_{\cdot 1} - \left(\beta + \frac{\boldsymbol{\pi}_1}{\boldsymbol{\gamma}_1}\right)\boldsymbol{\Pi}_{\cdot 2}\right),$$

and

$$\Delta^{\boldsymbol{\pi}_j} = \sqrt{n}\left(\Delta_j^{\boldsymbol{\Gamma}} - \left(\beta + \frac{\boldsymbol{\pi}_1}{\boldsymbol{\gamma}_1}\right)\Delta_j^{\gamma} - \boldsymbol{\gamma}_j\Delta^{\beta,1}\right) - \left(T^{\beta,1} + \Delta^{\beta,1}\right)\left(\widetilde{\boldsymbol{\gamma}}_j - \boldsymbol{\gamma}_j\right). \tag{C.3.89}$$

Define the events for $i \in \mathcal{S}^*$,

$$F^i = \left\{\max_{j\in\mathcal{S}^*, j\neq i} |T^{\boldsymbol{\pi}_j}| \leq 2.02\sqrt{\log p_{\mathrm{z}}}\sqrt{\boldsymbol{\Theta}_{11} + \left(\beta + \frac{\boldsymbol{\pi}_i}{\boldsymbol{\gamma}_i}\right)^2\boldsymbol{\Theta}_{22} - 2\left(\beta + \frac{\boldsymbol{\pi}_i}{\boldsymbol{\gamma}_i}\right)\boldsymbol{\Theta}_{12}}\frac{\left\|\widehat{\mathbf{v}}^{[j]} - \frac{\boldsymbol{\gamma}_j}{\boldsymbol{\gamma}_i}\widehat{v}^{[i]}\right\|_2}{\sqrt{n}}.\right\}$$

Then for $F = \cap_{i\in\mathcal{S}^*} F^i$, we have

$$\mathbf{P}\left(F\right) \geq 1 - Cs_{\mathrm{z}1}^2 p_{\mathrm{z}}^{-2.04} \geq 1 - cp^{-c}. \tag{C.3.90}$$

The proof of Lemma 37 relies on the following lemmas. The following lemma provides upper bound and lower bound for the variance term and the proof of the following lemma can be found in Section C.3.11.

Lemma 42. *On the event $A \cap S \cap G$, we have*

$$\begin{aligned}&\sqrt{\boldsymbol{\Theta}_{11} + \left(\beta + \left|\frac{\boldsymbol{\pi}_1}{\boldsymbol{\gamma}_1}\right|\right)^2\boldsymbol{\Theta}_{22} - 2\left(\beta + \frac{\boldsymbol{\pi}_1}{\boldsymbol{\gamma}_1}\right)\boldsymbol{\Theta}_{12}}\frac{\left\|\widehat{\mathbf{v}}^{[j]} - \frac{\boldsymbol{\gamma}_j}{\boldsymbol{\gamma}_1}\widehat{\mathbf{v}}^{[1]}\right\|_2}{\sqrt{n}} \\ &\leq 1.1\sqrt{M_1 M_2}\left(1 + \left|\frac{\boldsymbol{\gamma}_j}{\boldsymbol{\gamma}_1}\right|\right)\sqrt{1 + \left(\beta + \frac{\boldsymbol{\pi}_1}{\boldsymbol{\gamma}_1}\right)^2},\end{aligned} \tag{C.3.91}$$

and

$$\sqrt{\boldsymbol{\Theta}_{11}+\left(\beta+\left|\frac{\boldsymbol{\pi}_1}{\boldsymbol{\gamma}_1}\right|\right)^2\boldsymbol{\Theta}_{22}-2\left(\beta+\frac{\boldsymbol{\pi}_1}{\boldsymbol{\gamma}_1}\right)\boldsymbol{\Theta}_{12}}\frac{\left\|\widehat{\mathbf{v}}^{[j]}-\frac{\gamma_j}{\gamma_1}\widehat{\mathbf{v}}^{[1]}\right\|_2}{\sqrt{n}}$$
$$\geq 0.45\sqrt{\frac{M_1}{M_2}}\left(1+\left|\frac{\gamma_j}{\gamma_1}\right|\right)\sqrt{1+\left(\beta+\frac{\boldsymbol{\pi}_1}{\boldsymbol{\gamma}_1}\right)^2}. \tag{C.3.92}$$

Lemma 43. *On the event $A\cap S\cap G\cap F^1$, for large n, we have*

$$0.995\leq\frac{\sqrt{\widehat{\boldsymbol{\Theta}}_{11}+\left(\widehat{\beta}^{[1]}\right)^2\widehat{\boldsymbol{\Theta}}_{22}-2\widehat{\beta}^{[1]}\widehat{\boldsymbol{\Theta}}_{12}}\left\|\widehat{\mathbf{v}}^{[j]}-\frac{\widetilde{\gamma}_j}{\widetilde{\gamma}_1}\widehat{\mathbf{v}}^{[1]}\right\|_2}{\sqrt{\boldsymbol{\Theta}_{11}+\left(\beta+\frac{\boldsymbol{\pi}_1}{\boldsymbol{\gamma}_1}\right)^2\boldsymbol{\Theta}_{22}-2\left(\beta+\frac{\boldsymbol{\pi}_1}{\boldsymbol{\gamma}_1}\right)\boldsymbol{\Theta}_{12}}\left\|\widehat{\mathbf{v}}^{[j]}-\frac{\gamma_j}{\gamma_1}\widehat{\mathbf{v}}^{[1]}\right\|_2}\leq 1.005. \tag{C.3.93}$$

On the event $A\cap S\cap G\cap F^1$, we have

$$\max_{j\in\mathcal{S}^*}\frac{1}{\sqrt{n}}\left|T^{\boldsymbol{\pi}_j}\right|\leq 2.02\sqrt{\frac{\log p_{\mathrm{z}}}{n}}\sqrt{\boldsymbol{\Theta}_{11}+\left(\beta+\frac{\boldsymbol{\pi}_1}{\boldsymbol{\gamma}_1}\right)^2\boldsymbol{\Theta}_{22}-2\left(\beta+\frac{\boldsymbol{\pi}_1}{\boldsymbol{\gamma}_1}\right)\boldsymbol{\Theta}_{12}}\frac{\left\|\widehat{\mathbf{v}}^{[j]}-\frac{\gamma_j}{\gamma_1}\widehat{\mathbf{v}}^{[1]}\right\|_2}{\sqrt{n}}; \tag{C.3.94}$$

and

$$\max_{j\in\mathcal{S}^*}\frac{1}{\sqrt{n}}\left|\Delta^{\boldsymbol{\pi}_j}\right|\leq \frac{1}{300}\sqrt{\frac{\log p_{\mathrm{z}}}{n}}\sqrt{\boldsymbol{\Theta}_{11}+\left(\beta+\frac{\boldsymbol{\pi}_1}{\boldsymbol{\gamma}_1}\right)^2\boldsymbol{\Theta}_{22}-2\left(\beta+\frac{\boldsymbol{\pi}_1}{\boldsymbol{\gamma}_1}\right)\boldsymbol{\Theta}_{12}}\frac{\left\|\widehat{\mathbf{v}}^{[j]}-\frac{\gamma_j}{\gamma_1}\widehat{\mathbf{v}}^{[1]}\right\|_2}{\sqrt{n}}. \tag{C.3.95}$$

The ratio $\frac{\boldsymbol{\pi}_1}{\boldsymbol{\gamma}_1}$ can be divided into the following three cases,

1. strongly invalid instrument case, $|\boldsymbol{\pi}_1/\boldsymbol{\gamma}_1|\geq C_*(1/\delta_{\min})\sqrt{\log p_{\mathrm{z}}/n}$, where $C_*=12(1+|\beta|)\sqrt{M_1/M_2}$;

2. weakly invalid instrument case, $|\boldsymbol{\pi}_1/\boldsymbol{\gamma}_1|<C_*(1/\delta_{\min})\sqrt{\log p_{\mathrm{z}}/n}$;

3. valid instrument case, $\boldsymbol{\pi}_1/\boldsymbol{\gamma}_1=0$.

We are going to show that our procedure (4.3.9) in Chapter 4 will rule out the strong invalid instrument case and a stronger assumption (4.4.3) in Chapter 4 will help use rule out the weakly invalid instrument case. In the following, we will analyze the three cases separately.

strongly invalid instrument case

In this case, we assume that $|\boldsymbol{\pi}_1/\boldsymbol{\gamma}_1| \geq C_*(1/\delta_{\min})\sqrt{\log p_{\mathrm{z}}/n}$. For $j \in \mathcal{V}^*$, (C.3.88) can be re-expressed as

$$\sqrt{n}\left(\widehat{\boldsymbol{\pi}}_j^{[1]} - 0\right) = -\sqrt{n}\frac{\boldsymbol{\pi}_1}{\boldsymbol{\gamma}_1}\boldsymbol{\gamma}_j + T^{\boldsymbol{\pi}_j} + \Delta^{\boldsymbol{\pi}_j}. \tag{C.3.96}$$

We are going to show that on the event $A \cap S \cap G \cap F^1$,

$$\|\widetilde{\boldsymbol{\pi}}^{[1]}\|_0 > \frac{|\mathcal{S}^*|}{2}. \tag{C.3.97}$$

It is sufficient to show for $j \in \mathcal{V}^*$

$$\left|-\frac{\boldsymbol{\pi}_1}{\boldsymbol{\gamma}_1}\boldsymbol{\gamma}_j + \frac{1}{\sqrt{n}}\left(T^{\boldsymbol{\pi}_j} + \Delta^{\boldsymbol{\pi}_j}\right)\right| \geq 2.05\sqrt{\widehat{\boldsymbol{\Theta}}_{11} + \left(\widehat{\beta}^{[1]}\right)^2\widehat{\boldsymbol{\Theta}}_{22} - 2\widehat{\beta}^{[1]}\widehat{\boldsymbol{\Theta}}_{12}}\sqrt{\frac{\log p_{\mathrm{z}}}{n}}\frac{\left\|\widehat{\mathbf{v}}^{[j]} - \frac{\widetilde{\boldsymbol{\gamma}}_j}{\widetilde{\boldsymbol{\gamma}}_1}\widehat{\mathbf{v}}^{[1]}\right\|_2}{\sqrt{n}}, \tag{C.3.98}$$

which can be reduced to

$$\max_{j \in \mathcal{V}^*}\frac{1}{\sqrt{n}}\left|T^{\boldsymbol{\pi}_j} + \Delta^{\boldsymbol{\pi}_j}\right| \leq 2.05\sqrt{\widehat{\boldsymbol{\Theta}}_{11} + \left(\widehat{\beta}^{[1]}\right)^2\widehat{\boldsymbol{\Theta}}_{22} - 2\widehat{\beta}^{[1]}\widehat{\boldsymbol{\Theta}}_{12}}\sqrt{\frac{\log p_{\mathrm{z}}}{n}}\frac{\left\|\widehat{\mathbf{v}}^{[j]} - \frac{\widetilde{\boldsymbol{\gamma}}_j}{\widetilde{\boldsymbol{\gamma}}_1}\widehat{\mathbf{v}}^{[1]}\right\|_2}{\sqrt{n}}; \tag{C.3.99}$$

and

$$\left|\frac{\boldsymbol{\pi}_1}{\boldsymbol{\gamma}_1}\boldsymbol{\gamma}_j\right| \geq 4.1\sqrt{\widehat{\boldsymbol{\Theta}}_{11} + \left(\widehat{\beta}^{[1]}\right)^2\widehat{\boldsymbol{\Theta}}_{22} - 2\widehat{\beta}^{[1]}\widehat{\boldsymbol{\Theta}}_{12}}\sqrt{\frac{\log p_{\mathrm{z}}}{n}}\frac{\left\|\widehat{\mathbf{v}}^{[j]} - \frac{\widetilde{\boldsymbol{\gamma}}_j}{\widetilde{\boldsymbol{\gamma}}_1}\widehat{\mathbf{v}}^{[1]}\right\|_2}{\sqrt{n}}. \tag{C.3.100}$$

By (C.3.93), (C.3.94) and (C.3.95), we establish (C.3.99). By (C.3.91) and (C.3.93), we have

$$
\begin{aligned}
&4.1\sqrt{\widehat{\mathbf{\Theta}}_{11}+\left(\widehat{\beta}^{[1]}\right)^2\widehat{\mathbf{\Theta}}_{22}-2\widehat{\beta}^{[1]}\widehat{\mathbf{\Theta}}_{12}}\sqrt{\frac{\log p_{\mathrm{z}}}{n}}\frac{\left\|\widehat{\mathbf{v}}^{[j]}-\frac{\widetilde{\gamma}_j}{\widetilde{\gamma}_1}\widehat{\mathbf{v}}^{[1]}\right\|_2}{\sqrt{n}}\\
\leq&4.1\times1.005\times1.1\sqrt{M_1M_2}\left(1+\left|\frac{\boldsymbol{\gamma}_j}{\boldsymbol{\gamma}_1}\right|\right)\sqrt{1+\left(\beta+\frac{\boldsymbol{\pi}_1}{\boldsymbol{\gamma}_1}\right)^2}\sqrt{\frac{\log p_{\mathrm{z}}}{n}}\\
\leq&|\boldsymbol{\gamma}_j|\,4.1\times1.005\times1.1\sqrt{M_1M_2}\left(\left|\frac{1}{\boldsymbol{\gamma}_j}\right|+\left|\frac{1}{\boldsymbol{\gamma}_1}\right|\right)\left(1+\left|\beta+\frac{\boldsymbol{\pi}_1}{\boldsymbol{\gamma}_1}\right|\right)\sqrt{\frac{\log p_{\mathrm{z}}}{n}}.
\end{aligned}
\tag{C.3.101}
$$

The last term can be further upper bounded by

$$
\begin{aligned}
&|\boldsymbol{\gamma}_j|\frac{1}{\delta_{\min}}8.2\times1.005\times1.1\sqrt{M_1M_2}\left(1+\left|\beta+\frac{\boldsymbol{\pi}_1}{\boldsymbol{\gamma}_1}\right|\right)\sqrt{\frac{\log p_{\mathrm{z}}}{n}}\\
\leq&|\boldsymbol{\gamma}_j|\frac{1}{\delta_{\min}}8.2\times1.005\times1.1\sqrt{M_1M_2}\left(1+|\beta|\right)\sqrt{\frac{\log p_{\mathrm{z}}}{n}}\\
&+|\boldsymbol{\gamma}_j|\frac{1}{\delta_{\min}}8.2\times1.005\times1.1\sqrt{M_1M_2}\left|\frac{\boldsymbol{\pi}_1}{\boldsymbol{\gamma}_1}\right|\sqrt{\frac{\log p_{\mathrm{z}}}{n}}\\
\leq&0.99\frac{C_*}{\delta_{\min}}|\boldsymbol{\gamma}_j|\sqrt{\frac{\log p_{\mathrm{z}}}{n}}+C\frac{\sqrt{\frac{\log p_{\mathrm{z}}}{n}}}{\delta_{\min}}\left|\frac{\boldsymbol{\pi}_1}{\boldsymbol{\gamma}_1}\boldsymbol{\gamma}_j\right|,
\end{aligned}
\tag{C.3.102}
$$

where the first inequality follows from triangle inequality and the second inequality follows from the definition of C_*. Since $\delta_{\min}\gg\sqrt{\log p/n}$ and $|\boldsymbol{\pi}_1/\boldsymbol{\gamma}_1|\geq C_*(1/\delta_{\min})\sqrt{\log p_{\mathrm{z}}/n}$, by (C.3.101) and (C.3.102), we conclude (C.3.100).

<u>weakly invalid instrument case</u>

In this case, we assume $0<|\boldsymbol{\pi}_1/\boldsymbol{\gamma}_1|<C_*(1/\delta_{\min})\sqrt{\log p_{\mathrm{z}}/n}$. We have the following expression of (C.3.88),

$$
\widehat{\boldsymbol{\pi}}_j^{[1]}=\boldsymbol{\pi}_j-\frac{\boldsymbol{\pi}_1}{\boldsymbol{\gamma}_1}\boldsymbol{\gamma}_j+\frac{1}{\sqrt{n}}\left(T^{\boldsymbol{\pi}_j}+\Delta^{\boldsymbol{\pi}_j}\right).
\tag{C.3.103}
$$

We are going to show that on the event $A \cap S \cap G \cap F^1$,

$$\left\{ j \in \mathcal{S}^* : \left| \frac{\boldsymbol{\pi}_j}{\boldsymbol{\gamma}_j} \right| \geq 2C_* \frac{1}{\delta_{\min}} \sqrt{\frac{\log p_{\mathrm{z}}}{n}} \right\} \subset \operatorname{supp}\left(\widetilde{\boldsymbol{\pi}}^{[1]}\right). \tag{C.3.104}$$

It is sufficient to show the following inequality if $|\boldsymbol{\pi}_j/\boldsymbol{\gamma}_j| > 2C_*(1/\delta_{\min})\sqrt{\log p_{\mathrm{z}}/n}$,

$$\left| \boldsymbol{\pi}_j - \frac{\boldsymbol{\pi}_1}{\boldsymbol{\gamma}_1}\boldsymbol{\gamma}_j + \frac{1}{\sqrt{n}}\left(T^{\boldsymbol{\pi}_j} + \Delta^{\boldsymbol{\pi}_j}\right) \right| \geq 2.05 \sqrt{\widehat{\boldsymbol{\Theta}}_{11} + \left(\widehat{\beta}^{[1]}\right)^2 \widehat{\boldsymbol{\Theta}}_{22} - 2\widehat{\beta}^{[1]}\widehat{\boldsymbol{\Theta}}_{12}} \sqrt{\frac{\log p_{\mathrm{z}}}{n}} \frac{\left\| \widehat{\mathbf{v}}^{[j]} - \frac{\widetilde{\boldsymbol{\gamma}}_j}{\widetilde{\boldsymbol{\gamma}}_1}\widehat{\mathbf{v}}^{[1]} \right\|_2}{\sqrt{n}}, \tag{C.3.105}$$

which can be reduced to

$$\max_{j \in \operatorname{supp}(\boldsymbol{\pi}) \cap \mathcal{S}^*} \frac{1}{\sqrt{n}} \left| T^{\boldsymbol{\pi}_j} + \Delta^{\boldsymbol{\pi}_j} \right| \leq 2.05 \sqrt{\widehat{\boldsymbol{\Theta}}_{11} + \left(\widehat{\beta}^{[1]}\right)^2 \widehat{\boldsymbol{\Theta}}_{22} - 2\widehat{\beta}^{[1]}\widehat{\boldsymbol{\Theta}}_{12}} \sqrt{\frac{\log p_{\mathrm{z}}}{n}} \frac{\left\| \widehat{\mathbf{v}}^{[j]} - \frac{\widetilde{\boldsymbol{\gamma}}_j}{\widetilde{\boldsymbol{\gamma}}_1}\widehat{\mathbf{v}}^{[1]} \right\|_2}{\sqrt{n}}, \tag{C.3.106}$$

$$\left| \boldsymbol{\pi}_j - \frac{\boldsymbol{\pi}_1}{\boldsymbol{\gamma}_1}\boldsymbol{\gamma}_j \right| \geq 4.1 \sqrt{\widehat{\boldsymbol{\Theta}}_{11} + \left(\widehat{\beta}^{[1]}\right)^2 \widehat{\boldsymbol{\Theta}}_{22} - 2\widehat{\beta}^{[1]}\widehat{\boldsymbol{\Theta}}_{12}} \sqrt{\frac{\log p_{\mathrm{z}}}{n}} \frac{\left\| \widehat{\mathbf{v}}^{[j]} - \frac{\widetilde{\boldsymbol{\gamma}}_j}{\widetilde{\boldsymbol{\gamma}}_1}\widehat{\mathbf{v}}^{[1]} \right\|_2}{\sqrt{n}}. \tag{C.3.107}$$

By (C.3.93), (C.3.94) and (C.3.95), we establish(C.3.106). Since

$$\frac{1}{|\boldsymbol{\gamma}_j|} \left| \boldsymbol{\pi}_j - \frac{\boldsymbol{\pi}_1}{\boldsymbol{\gamma}_1}\boldsymbol{\gamma}_j \right| \geq \left| \frac{\boldsymbol{\pi}_j}{\boldsymbol{\gamma}_j} \right| - \left| \frac{\boldsymbol{\pi}_1}{\boldsymbol{\gamma}_1} \right| \geq C_* \frac{1}{\delta_{\min}} \sqrt{\frac{\log p_{\mathrm{z}}}{n}},$$

the assumption $0 < |\boldsymbol{\pi}_1/\boldsymbol{\gamma}_1| < C_*(1/\delta_{\min})\sqrt{\log p_{\mathrm{z}}/n}$ and (C.3.102) lead to (C.3.107).

valid instrument case

In this case, the instrumental variable is valid with $\boldsymbol{\pi}_1/\boldsymbol{\gamma}_1 = 0$. For $j \in \mathcal{V}^*$, (C.3.88) can be re-expressed as

$$\sqrt{n}\left(\widehat{\boldsymbol{\pi}}_j^{[1]} - 0\right) = T^{\boldsymbol{\pi}_j} + \Delta^{\boldsymbol{\pi}_j}. \tag{C.3.108}$$

By (C.3.93), (C.3.94) and (C.3.95), on the event $A \cap S \cap G \cap F^1$,

$$\max_{j\in\mathcal{V}^*} \frac{1}{\sqrt{n}} \left|T^{\boldsymbol{\pi}_j} + \Delta^{\boldsymbol{\pi}_j}\right| \leq 2.05\sqrt{\widehat{\boldsymbol{\Theta}}_{11} + \left(\widehat{\beta}^{[1]}\right)^2 \widehat{\boldsymbol{\Theta}}_{22} - 2\widehat{\beta}^{[1]}\widehat{\boldsymbol{\Theta}}_{12}}\sqrt{\frac{\log p_{\mathrm{z}}}{n}} \frac{\left\|\widehat{\mathbf{v}}^{[j]} - \frac{\widetilde{\gamma}_j}{\widetilde{\gamma}_1}\widehat{\mathbf{v}}^{[1]}\right\|_2}{\sqrt{n}}. \tag{C.3.109}$$

and hence

$$\operatorname{supp}\left(\widetilde{\boldsymbol{\pi}}^{[1]}\right) \subset \operatorname{supp}\left(\boldsymbol{\pi}\right) \quad \text{and} \quad \|\widetilde{\boldsymbol{\pi}}^{[1]}\|_0 < \frac{|\mathcal{S}^*|}{2}. \tag{C.3.110}$$

For $|\boldsymbol{\pi}_j/\boldsymbol{\gamma}_j| \geq C_*(1/\delta_{\min})\sqrt{\log p_{\mathrm{z}}/n}$, by (C.3.102), we obtain

$$|\boldsymbol{\pi}_j| \geq 4.1\sqrt{\widehat{\boldsymbol{\Theta}}_{11} + \left(\widehat{\beta}^{[1]}\right)^2 \widehat{\boldsymbol{\Theta}}_{22} - 2\widehat{\beta}^{[1]}\widehat{\boldsymbol{\Theta}}_{12}}\sqrt{\frac{\log p_{\mathrm{z}}}{n}} \frac{\left\|\widehat{\mathbf{v}}^{[j]} - \frac{\widetilde{\gamma}_j}{\widetilde{\gamma}_1}\widehat{\mathbf{v}}^{[1]}\right\|_2}{\sqrt{n}}. \tag{C.3.111}$$

Combined with (C.3.109), we have

$$\left|\widehat{\boldsymbol{\pi}}_j^{[1]}\right| \geq 2.05\sqrt{\widehat{\boldsymbol{\Theta}}_{11} + \left(\widehat{\beta}^{[1]}\right)^2 \widehat{\boldsymbol{\Theta}}_{22} - 2\widehat{\beta}^{[1]}\widehat{\boldsymbol{\Theta}}_{12}}\sqrt{\frac{\log p_{\mathrm{z}}}{n}} \frac{\left\|\widehat{\mathbf{v}}^{[j]} - \frac{\widetilde{\gamma}_j}{\widetilde{\gamma}_1}\widehat{\mathbf{v}}^{[1]}\right\|_2}{\sqrt{n}}. \tag{C.3.112}$$

Hence

$$\left\{j \in \mathcal{S}^* : \left|\frac{\boldsymbol{\pi}_j}{\boldsymbol{\gamma}_j}\right| \geq C_* \frac{1}{\delta_{\min}}\sqrt{\frac{\log p_{\mathrm{z}}}{n}}\right\} \subset \operatorname{supp}\left(\widetilde{\boldsymbol{\pi}}^{[1]}\right). \tag{C.3.113}$$

By comparing (C.3.97) and (C.3.110), we rule out the strong invalid instrumental variable case and obtain $\left|\widetilde{\mathcal{V}}\right| > |\mathcal{S}^*|/2$ in (C.2.32). Further by (C.3.104) and (C.3.113), we establish (C.2.32). With a stronger assumption (4.4.3) in the main paper, the weak invalid instrument case is also ruled out and (C.3.113) leads to (C.2.33).

C.3.9 Proof of Lemma 36

The proof of this lemma follows from the following results. Under the regularity assumptions (R1) $-$ (R3), as $\sqrt{s_{z1}} s \log p/\sqrt{n} \to 0$, we have

$$\max_{1\leq i,j\leq 2} |\widehat{\boldsymbol{\Theta}}_{ij} - \boldsymbol{\Theta}_{ij}| \xrightarrow{p} 0; \tag{C.3.114}$$

and

$$\frac{\widehat{\|\boldsymbol{\gamma}\|_2^2}}{\|\boldsymbol{\gamma}\|_2^2} \xrightarrow{p} 1 \quad \text{and} \quad \frac{\left\|\sum_{j\in\widetilde{\mathcal{S}}} \widetilde{\boldsymbol{\gamma}}_j \widehat{\mathbf{v}}^{[j]}\right\|_2}{\left\|\sum_{j\in\mathcal{S}^*} \boldsymbol{\gamma}_j \widehat{\mathbf{v}}^{[j]}\right\|_2} \xrightarrow{p} 1. \tag{C.3.115}$$

By (C.1.1) and (C.3.114), we establish that

$$\frac{\sqrt{\widehat{\boldsymbol{\Theta}}_{11} + \widehat{\beta}^2\widehat{\boldsymbol{\Theta}}_{22} - 2\widehat{\beta}\widehat{\boldsymbol{\Theta}}_{12}}}{\sqrt{\boldsymbol{\Theta}_{11} + (\beta)^2\boldsymbol{\Theta}_{22} - 2\beta\boldsymbol{\Theta}_{12}}} \xrightarrow{p} 1.$$

Combined with (C.3.115), we establish (C.2.30).

Proof of (C.3.114) A stronger version of this proposition has already been proved in Ren et al. (2013), where part of it was already established in Sun & Zhang (2012). To be self-contained, we will provide the sketch of the proof in the following.

The difference between $\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}$ can be decomposed as,

$$\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta} = \boldsymbol{\Theta}^{\text{ora}} - \boldsymbol{\Theta} + \widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}^{\text{ora}}, \tag{C.3.116}$$

where $\boldsymbol{\Theta}_{11}^{\text{ora}} = \frac{1}{n}\|Y - \mathbf{Z}\boldsymbol{\Gamma} - \mathbf{X}\boldsymbol{\Psi}\|_2^2$, $\boldsymbol{\Theta}_{22}^{\text{ora}} = \frac{1}{n}\|D - \mathbf{Z}\boldsymbol{\gamma} - \mathbf{X}\boldsymbol{\psi}\|_2^2$ and $\boldsymbol{\Theta}_{12}^{\text{ora}} = \frac{1}{n}(Y - \mathbf{Z}\boldsymbol{\Gamma} - \mathbf{X}\boldsymbol{\Psi})^\intercal(D - \mathbf{Z}\boldsymbol{\gamma} - \mathbf{X}\boldsymbol{\psi})$. In the following, we only provide the detailed analysis of $\widehat{\boldsymbol{\Theta}}_{12} - \boldsymbol{\Theta}_{12}^{\text{ora}}$. The other differences can be established in a similar way and the difference between $\boldsymbol{\Theta}^{\text{ora}} - \boldsymbol{\Theta}$ can be established by central limit theorem.

$$\widehat{\boldsymbol{\Theta}}_{12} - \boldsymbol{\Theta}_{12}^{\text{ora}} = \frac{1}{n}\begin{pmatrix}\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}\\ \widehat{\boldsymbol{\psi}} - \boldsymbol{\psi}\end{pmatrix}^\intercal \mathbf{W}^\intercal\mathbf{W}\begin{pmatrix}\widehat{\boldsymbol{\Gamma}} - \boldsymbol{\Gamma}\\ \widehat{\boldsymbol{\Psi}} - \boldsymbol{\Psi}\end{pmatrix} + \frac{1}{n}\boldsymbol{\Pi}_{\cdot 2}^\intercal\mathbf{W}\begin{pmatrix}\widehat{\boldsymbol{\Gamma}} - \boldsymbol{\Gamma}\\ \widehat{\boldsymbol{\Psi}} - \boldsymbol{\Psi}\end{pmatrix} + \frac{1}{n}\boldsymbol{\Pi}_{\cdot 1}^\intercal\mathbf{W}\begin{pmatrix}\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}\\ \widehat{\boldsymbol{\psi}} - \boldsymbol{\psi}\end{pmatrix}. \tag{C.3.117}$$

By (C.3.117), we have

$$\left|\widehat{\boldsymbol{\Theta}}_{12}-\boldsymbol{\Theta}_{12}^{\rm ora}\right| \leq \frac{1}{\sqrt{n}}\|\mathbf{W}\begin{pmatrix}\widehat{\gamma}-\gamma\\ \widehat{\boldsymbol{\psi}}-\boldsymbol{\psi}\end{pmatrix}\|_2\frac{1}{\sqrt{n}}\|\mathbf{W}\begin{pmatrix}\widehat{\boldsymbol{\Gamma}}-\boldsymbol{\Gamma}\\ \widehat{\boldsymbol{\Psi}}-\boldsymbol{\Psi}\end{pmatrix}\|_2+\frac{1}{n}\|\boldsymbol{\Pi}_{\cdot 2}^{\intercal}\mathbf{W}\|_\infty\|\begin{pmatrix}\widehat{\boldsymbol{\Gamma}}-\boldsymbol{\Gamma}\\ \widehat{\boldsymbol{\Psi}}-\boldsymbol{\Psi}\end{pmatrix}\|_1$$
$$+\frac{1}{n}\|\boldsymbol{\Pi}_{\cdot 1}^{\intercal}\mathbf{W}\|_\infty\|\begin{pmatrix}\widehat{\gamma}-\gamma\\ \widehat{\boldsymbol{\psi}}-\boldsymbol{\psi}\end{pmatrix}\|_1. \tag{C.3.118}$$

The following of the proof follows from Lemma 40 and definition of event G.

<u>**Proof of** (C.3.115)</u> For a given $0<\epsilon_0<1$, we have

$$\mathbf{P}\left(\left|\frac{\|\boldsymbol{\gamma}\|_2^2}{\widehat{\|\boldsymbol{\gamma}\|_2^2}}-1\right|\geq\epsilon_0\right)\leq\mathbf{P}\left(\left|\frac{\widehat{\|\boldsymbol{\gamma}\|_2^2}-\|\boldsymbol{\gamma}\|_2^2}{\|\boldsymbol{\gamma}\|_2^2}\right|\geq\frac{\epsilon_0}{1-\epsilon_0}\right).$$

By Lemma 35, on the event $A\cap S\cap G$, we have

$$\left|\frac{\widehat{\|\boldsymbol{\gamma}\|_2^2}-\|\boldsymbol{\gamma}\|_2^2}{\|\boldsymbol{\gamma}\|_2^2}\right|\leq C\frac{1}{\|\boldsymbol{\gamma}\|_2^2}\left(s_{\rm z1}\frac{\log p_{\rm z}}{n}+C\|\boldsymbol{\gamma}\|_2\sqrt{\frac{2s_{\rm z1}\log p_{\rm z}}{n}}\right).$$

Since $\|\boldsymbol{\gamma}\|_2^2\gg(s\log p/\sqrt{n})^2$, we obtain that

$$\left|\frac{\widehat{\|\boldsymbol{\gamma}\|_2^2}-\|\boldsymbol{\gamma}\|_2^2}{\|\boldsymbol{\gamma}\|_2^2}\right|\leq\frac{\epsilon_0}{1-\epsilon_0}\text{ and }\mathbf{P}\left(\left|\frac{\|\boldsymbol{\gamma}\|_2^2}{\widehat{\|\boldsymbol{\gamma}\|_2^2}}-1\right|\geq\epsilon_0\right)\leq\mathbf{P}\left((A\cap S\cap G)^c\right).$$

Combined with Lemma 33, we establish the first convergence result of (C.3.115). On the event $S\cap G\cap A$, we have

$$\left|\frac{\left\|\sum_{j\in\widetilde{\mathcal{S}}}\widetilde{\boldsymbol{\gamma}}_j\widehat{\mathbf{v}}^{[j]}\right\|_2}{\left\|\sum_{j\in\mathcal{S}^*}\boldsymbol{\gamma}_j\widehat{\mathbf{v}}^{[j]}\right\|_2}-1\right|\leq\frac{\sum_{j\in\widetilde{\mathcal{S}}}|\widetilde{\boldsymbol{\gamma}}_j-\boldsymbol{\gamma}_j|\frac{\|\widehat{\mathbf{v}}^{[j]}\|}{\sqrt{n}}+\sum_{j\in\mathcal{S}^*\setminus\widetilde{\mathcal{S}}}|\boldsymbol{\gamma}_j|\frac{\|\widehat{\mathbf{v}}^{[j]}\|}{\sqrt{n}}}{\frac{1}{\sqrt{n}}\left\|\sum_{j\in\mathcal{S}^*}\boldsymbol{\gamma}_j\widehat{\mathbf{v}}^{[j]}\right\|_2}$$

By Lemma 34, we have

$$
\begin{aligned}
&\frac{\sum_{j\in\widetilde{\mathcal{S}}}|\widetilde{\boldsymbol{\gamma}}_j-\boldsymbol{\gamma}_j|\frac{\|\widehat{\mathbf{v}}^{[j]}\|}{\sqrt{n}}+\sum_{j\in\mathcal{S}^*\setminus\widetilde{\mathcal{S}}}|\boldsymbol{\gamma}_j|\frac{\|\widehat{\mathbf{v}}^{[j]}\|}{\sqrt{n}}}{\frac{1}{\sqrt{n}}\left\|\sum_{j\in\mathcal{S}^*}\boldsymbol{\gamma}_j\widehat{\mathbf{v}}^{[j]}\right\|_2}\\
\leq&\frac{\left(\sum_{j\in\widetilde{\mathcal{S}}}|\widetilde{\boldsymbol{\gamma}}_j-\boldsymbol{\gamma}_j|+\sum_{i\in\mathcal{S}^*\setminus\widetilde{\mathcal{S}}}|\boldsymbol{\gamma}_j|\right)\sqrt{\left(1+12\sqrt{\frac{\log p}{n}}\right)M_1}}{\sqrt{\frac{M_1\|\boldsymbol{\gamma}\|_2^2(1-s_{z1}\lambda_n)^2}{1-12\sqrt{\frac{\log p}{n}}}}}\leq Cs_{z1}\sqrt{\frac{\log p}{n}}\leq\epsilon_0,
\end{aligned}
$$

and hence the second convergence result of (C.3.115) follows from the following inequality.

C.3.10 Proof of Lemma 39

Define $\widetilde{\|\boldsymbol{\gamma}\|_2^2}=\sum_{j\in\widetilde{\mathcal{V}}}\widetilde{\boldsymbol{\gamma}}_j^2$ and $\|\boldsymbol{\gamma}_{\mathcal{V}^*}\|_2^2=\sum_{j\in\mathcal{V}^*}\boldsymbol{\gamma}_j^2$. The proof of this lemma is further based on the following results. Under the assumptions (R1) – (R5) and (IN1)-(IN3). As $\sqrt{s_{z1}}s\log p/\sqrt{n}\to 0$, we have

$$
\frac{\widetilde{\|\boldsymbol{\gamma}\|_2^2}}{\|\boldsymbol{\gamma}_{\mathcal{V}^*}\|_2^2}\xrightarrow{p}1\quad\text{and}\quad\frac{\left\|\sum_{j\in\widetilde{\mathcal{V}}}\widetilde{\boldsymbol{\gamma}}_j\widehat{\mathbf{v}}^{[j]}\right\|_2}{\left\|\sum_{j\in\mathcal{V}^*}\boldsymbol{\gamma}_j\widehat{\mathbf{v}}^{[j]}\right\|_2}\xrightarrow{p}1. \tag{C.3.119}
$$

By (C.3.114) and (4.4.4) in the main paper, we establish that

$$
\frac{\sqrt{\widehat{\boldsymbol{\Theta}}_{11}+\widehat{\beta}^2\widehat{\boldsymbol{\Theta}}_{22}-2\widehat{\beta}\widehat{\boldsymbol{\Theta}}_{12}}}{\sqrt{\boldsymbol{\Theta}_{11}+(\beta)^2\boldsymbol{\Theta}_{22}-2\beta\boldsymbol{\Theta}_{12}}}\xrightarrow{p}1.
$$

Combined with (C.3.119), we establish (C.2.37).

<u>**Proof of** (C.3.119)</u> For a given $0<\epsilon_0<1$, we have

$$
\mathbf{P}\left(\left|\frac{\|\boldsymbol{\gamma}_{\mathcal{V}^*}\|_2^2}{\widetilde{\|\boldsymbol{\gamma}\|_2^2}}-1\right|\geq\epsilon_0\right)\leq\mathbf{P}\left(\left|\frac{\widetilde{\|\boldsymbol{\gamma}\|_2^2}-\|\boldsymbol{\gamma}_{\mathcal{V}^*}\|_2^2}{\|\boldsymbol{\gamma}_{\mathcal{V}^*}\|_2^2}\right|\geq\frac{\epsilon_0}{1-\epsilon_0}\right).
$$

By Lemma 37, on the event $A \cap S \cap G \cap F$, we have $\widetilde{\mathcal{V}} = \mathcal{V}^*$ and (C.3.79) leads to

$$\left| \frac{\widetilde{\|\boldsymbol{\gamma}\|_2^2} - \|\boldsymbol{\gamma}_{\mathcal{V}^*}\|_2^2}{\|\boldsymbol{\gamma}_{\mathcal{V}^*}\|_2^2} \right| \leq C \frac{1}{\|\boldsymbol{\gamma}_{\mathcal{V}^*}\|_2^2} \left(s_{\mathrm{z}1} \frac{\log p_{\mathrm{z}}}{n} + C s_{\mathrm{z}1} \left(s \frac{\log p}{n} \right)^2 + C \|\boldsymbol{\gamma}_{\widetilde{\mathcal{V}}}\|_2 \sqrt{\frac{2|\widetilde{\mathcal{V}}| \log p_{\mathrm{z}}}{n}} \right). \tag{C.3.120}$$

Since $\|\boldsymbol{\gamma}_{\mathcal{V}^*}\|_2^2 \gg (s \log p/\sqrt{n})^2$, we obtain that

$$\left| \frac{\widetilde{\|\boldsymbol{\gamma}\|_2^2} - \|\boldsymbol{\gamma}_{\mathcal{V}^*}\|_2^2}{\|\boldsymbol{\gamma}_{\mathcal{V}^*}\|_2^2} \right| \leq \frac{\epsilon_0}{1 - \epsilon_0} \text{ and } \mathbf{P} \left(\left| \frac{\|\boldsymbol{\gamma}_{\mathcal{V}^*}\|_2^2}{\widetilde{\|\boldsymbol{\gamma}\|_2^2}} - 1 \right| \geq \epsilon_0 \right) \leq \mathbf{P} \left((A \cap S \cap G \cap F)^c \right).$$

Combined with Lemma 33 and (C.3.90), we establish the first converge result of (C.3.119).

On the event $S \cap G \cap A \cap F$, we have

$$\left| \frac{\left\| \sum_{j \in \widetilde{\mathcal{V}}} \widetilde{\boldsymbol{\gamma}}_j \widehat{\mathbf{v}}^{[j]} \right\|_2}{\left\| \sum_{j \in \mathcal{V}^*} \boldsymbol{\gamma}_j \widehat{\mathbf{v}}^{[j]} \right\|_2} - 1 \right| \leq \frac{\sum_{j \in \mathcal{V}^*} |\widetilde{\boldsymbol{\gamma}}_j - \boldsymbol{\gamma}_j| \frac{\|\widehat{\mathbf{v}}^{[j]}\|}{\sqrt{n}}}{\frac{1}{\sqrt{n}} \left\| \sum_{j \in \mathcal{V}^*} \boldsymbol{\gamma}_j \widehat{\mathbf{v}}^{[j]} \right\|_2}.$$

By Lemma 34, we have

$$\frac{\sum_{j \in \mathcal{V}^*} |\widetilde{\boldsymbol{\gamma}}_j - \boldsymbol{\gamma}_j| \frac{\|\widehat{\mathbf{v}}^{[j]}\|}{\sqrt{n}}}{\frac{1}{\sqrt{n}} \left\| \sum_{j \in \mathcal{V}^*} \boldsymbol{\gamma}_j \widehat{\mathbf{v}}^{[j]} \right\|_2} \leq \frac{\left(\sum_{j \in \mathcal{V}^*} |\widetilde{\boldsymbol{\gamma}}_j - \boldsymbol{\gamma}_j| \right) \sqrt{\left(1 + 12 \sqrt{\frac{\log p}{n}} \right) M_1}}{\sqrt{\frac{M_1 \|\boldsymbol{\gamma}\|_2^2 (1 - s_{\mathrm{z}1} \lambda_n)^2}{1 - 12 \sqrt{\frac{\log p}{n}}}}} \leq C s_{\mathrm{z}1} \sqrt{\frac{\log p}{n}}. \tag{C.3.121}$$

Hence the second converge result of (C.3.119) follows from

$$\mathbf{P} \left(\left| \frac{\left\| \sum_{j \in \widetilde{\mathcal{V}}} \widetilde{\boldsymbol{\gamma}}_j \widehat{\mathbf{v}}^{[j]} \right\|_2}{\left\| \sum_{j \in \mathcal{V}^*} \boldsymbol{\gamma}_j \widehat{\mathbf{v}}^{[j]} \right\|_2} - 1 \right| \geq \epsilon_0 \right) \leq \mathbf{P} \left((S \cap G \cap A \cap F)^c \right).$$

C.3.11 Proof of Lemmas 40, 41, 42 and 43

Proof of Lemma 40 We only estalbish the first half of (C.3.32) and (C.3.33). The proof of the second half of (C.3.32) and (C.3.34) will be similar. The proof has been established in Sun & Zhang (2012) for fixed designs under certain assumptions for the design. In the following, we will check that the assumptions in Corollary 1 in Sun & Zhang (2012) are satisfied with high probability for the subgaussian random designs considered in this paper and then apply equation (23) in Sun & Zhang (2012). By the definition of τ^* in Sun & Zhang (2012), we have $\tau^* \leq \tau$ where τ is defined in (C.2.2). Hence, on the event S_1, equation (23) in Sun & Zhang (2012) holds. By the relationship between ℓ_1 cone invertibility factor and the restricted eigenvalue established in Lemma 13 of Cai & Guo (2016c), we obtain that on the event $S \cap G$,

$$\|\widehat{\mathbf{\Gamma}} - \mathbf{\Gamma}\|_1 + \|\widehat{\mathbf{\Psi}} - \mathbf{\Psi}\|_1 \leq C \frac{s\lambda_0\sigma_1}{\kappa^2(\boldsymbol{H}, 4s, 1+2\epsilon_0)}. \tag{C.3.122}$$

Similar to the proof of Lemma 13 in Cai & Guo (2016c), we establish

$$\kappa^2\left(\boldsymbol{H}, 4s, 1+2\epsilon_0\right) \geq \frac{n}{\max \|\mathbf{W}_{\cdot j}\|_2^2}\kappa^2\left(\mathbf{W}, 4s, (1+2\epsilon_0)\left(\frac{\max \|\mathbf{W}_{\cdot j}\|_2}{\min \|\mathbf{W}_{\cdot j}\|_2}\right)\right). \tag{C.3.123}$$

Hence, on the event $G \cap S$, we establish the first half of (C.3.32). Since

$$\frac{1}{n}\|\boldsymbol{Z}(\widehat{\mathbf{\Gamma}} - \mathbf{\Gamma}) + \boldsymbol{X}(\widehat{\mathbf{\Psi}} - \mathbf{\Psi})\|_2^2 \leq \|\frac{1}{n}\mathbf{W}^{\intercal}\mathbf{W}\begin{pmatrix}\widehat{\mathbf{\Gamma}} - \mathbf{\Gamma} \\ \widehat{\mathbf{\Psi}} - \mathbf{\Psi}\end{pmatrix}\|_\infty \left(\|\widehat{\mathbf{\Gamma}} - \mathbf{\Gamma}\|_1 + \|\widehat{\mathbf{\Psi}} - \mathbf{\Psi}\|_1\right),$$

we establish (C.3.33).

Proof of Lemma 41 The decompositions (C.3.35) and (C.3.37) are established by the definitions of $\widetilde{\mathbf{\Gamma}}$ and $\widetilde{\boldsymbol{\gamma}}$. The error bound (C.3.39) follows from the following inequality

$$|\Delta_j^{\gamma}| \leq \| \left(\frac{1}{n}(\widehat{\mathbf{u}}^{[j]})^{\intercal}\widehat{\mathbf{\Sigma}} - e_j^{\intercal}\right) \|_{\infty} \| \begin{pmatrix} \widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma} \\ \widehat{\boldsymbol{\psi}} - \boldsymbol{\psi} \end{pmatrix} \|_1.$$

Proof of Lemma 42

This lemma can be established by a similar argument with Lemma 34. On the event $A \cap S \cap G$, we have

$$\frac{1}{\sqrt{M_2}}\sqrt{1+\left(\beta+\frac{\boldsymbol{\pi}_1}{\boldsymbol{\gamma}_1}\right)^2} \leq \sqrt{\mathbf{\Theta}_{11}+\left(\beta+\frac{\boldsymbol{\pi}_1}{\boldsymbol{\gamma}_1}\right)^2\mathbf{\Theta}_{22}-2\left(\beta+\frac{\boldsymbol{\pi}_1}{\boldsymbol{\gamma}_1}\right)\mathbf{\Theta}_{12}} \leq \sqrt{M_2}\sqrt{1+\left(\beta+\frac{\boldsymbol{\pi}_1}{\boldsymbol{\gamma}_1}\right)^2}, \tag{C.3.124}$$

and

$$\frac{\sqrt{M_1}\sqrt{1+\left(\frac{\boldsymbol{\gamma}_j}{\boldsymbol{\gamma}_1}\right)^2}\,|1-2\lambda_n|}{\sqrt{1-12\sqrt{\frac{\log p}{n}}}} \leq \frac{\left\|\widehat{\mathbf{v}}^{[j]}-\frac{\boldsymbol{\gamma}_j}{\boldsymbol{\gamma}_1}\widehat{\mathbf{v}}^{[1]}\right\|_2}{\sqrt{n}} \leq \sqrt{M_1}\left(1+\left|\frac{\boldsymbol{\gamma}_j}{\boldsymbol{\gamma}_1}\right|\right)\sqrt{1+12\sqrt{\frac{\log p}{n}}}. \tag{C.3.125}$$

Hence, (C.3.91) and (C.3.92) follow from the above inequalities (C.3.124) and (C.3.125).

Proof of Lemma 43

(C.3.93) follows from the standard convergence analysis and (C.3.94) follows from high probability statement of Gaussian random variable. It remains to establish (C.3.95). We will analyze the expression (C.3.89) term by term. Note that on the event $A \cap S \cap G$,

$$\begin{aligned}
\left|T^{\beta,1}\right| &\leq \frac{\sqrt{\log p_{\mathrm{z}}}}{\sqrt{n}\,|\boldsymbol{\gamma}_1|}\|\widehat{\mathbf{v}}^{[1]}\|_2\sqrt{\mathbf{\Theta}_{11}+\left(\beta+\frac{\boldsymbol{\pi}_1}{\boldsymbol{\gamma}_1}\right)^2\mathbf{\Theta}_{22}-2\left(\beta+\frac{\boldsymbol{\pi}_1}{\boldsymbol{\gamma}_1}\right)\mathbf{\Theta}_{12}} \\
&\leq C\frac{1}{|\boldsymbol{\gamma}_1|}\frac{\|\widehat{\mathbf{v}}^{[1]}\|_2}{\sqrt{n}}\sqrt{\log p_{\mathrm{z}}}\left(1+\left|\beta+\frac{\boldsymbol{\pi}_1}{\boldsymbol{\gamma}_1}\right|\right); \\
|R_1| &\leq C\frac{1}{|\boldsymbol{\gamma}_1|}\left(\left|\beta+\frac{\boldsymbol{\pi}_1}{\boldsymbol{\gamma}_1}\right|+1\right)s\frac{\log p}{\sqrt{n}}; \\
|R_2| &\leq C\frac{1}{|\boldsymbol{\gamma}_1|}\left(\sqrt{\frac{\log p}{n}}+s\frac{\log p}{n}\right)\left(\left|T^{\beta,1}\right|+|R_1|\right) \\
&\leq C\frac{1}{|\boldsymbol{\gamma}_1|}\sqrt{\frac{\log p}{n}}\frac{1}{|\boldsymbol{\gamma}_1|}\frac{\|\widehat{\mathbf{v}}^{[1]}\|_2}{\sqrt{n}}\sqrt{\log p_{\mathrm{z}}}\left(1+\left|\beta+\frac{\boldsymbol{\pi}_1}{\boldsymbol{\gamma}_1}\right|\right).
\end{aligned} \tag{C.3.126}$$

Hence

$$\begin{aligned}
&\frac{1}{\sqrt{n}}\left|\boldsymbol{\gamma}_j\Delta^{\beta,1}+\left(T^{\beta,1}+\Delta^{\beta,1}\right)\left(\widetilde{\boldsymbol{\gamma}}_j-\boldsymbol{\gamma}_j\right)\right|\leq C\frac{|\boldsymbol{\gamma}_j|}{\sqrt{n}}\left(|R_1|+|R_2|\right)+\sqrt{\frac{\log p_{\mathrm{z}}}{n}}\left|T^{\beta,1}\right| \\
\leq & C\frac{|\boldsymbol{\gamma}_j|}{|\boldsymbol{\gamma}_1|}\left(\left|\beta+\frac{\boldsymbol{\pi}_1}{\boldsymbol{\gamma}_1}\right|+1\right)s\frac{\log p}{n}+C\frac{|\boldsymbol{\gamma}_j|}{|\boldsymbol{\gamma}_1|}\frac{\sqrt{\log p\log p_{\mathrm{z}}}}{n}\frac{1}{|\boldsymbol{\gamma}_1|}\frac{\|\widehat{\mathbf{v}}^{[1]}\|_2}{\sqrt{n}}\left(1+\left|\beta+\frac{\boldsymbol{\pi}_1}{\boldsymbol{\gamma}_1}\right|\right) \\
&+C\frac{1}{|\boldsymbol{\gamma}_1|}\frac{\|\widehat{\mathbf{v}}^{[1]}\|_2}{\sqrt{n}}\frac{\sqrt{\log p\log p_{\mathrm{z}}}}{n}\left(1+\left|\beta+\frac{\boldsymbol{\pi}_1}{\boldsymbol{\gamma}_1}\right|\right) \\
\leq & C\left(1+\left|\beta+\frac{\boldsymbol{\pi}_1}{\boldsymbol{\gamma}_1}\right|\right)\left(\frac{|\boldsymbol{\gamma}_j|}{|\boldsymbol{\gamma}_1|}s\frac{\log p}{n}+\left(1+\frac{|\boldsymbol{\gamma}_j|}{|\boldsymbol{\gamma}_1|}\right)\frac{1}{|\boldsymbol{\gamma}_1|}\frac{\sqrt{\log p\log p_{\mathrm{z}}}}{n}\right).
\end{aligned} \tag{C.3.127}$$

Since

$$\left|\Delta_j^{\boldsymbol{\Gamma}}-\left(\beta+\frac{\boldsymbol{\pi}_1}{\boldsymbol{\gamma}_1}\right)\Delta_j^{\gamma}\right|\leq C\left(\left|\beta+\frac{\boldsymbol{\pi}_1}{\boldsymbol{\gamma}_1}\right|+1\right)s\frac{\log p}{n},$$

we have

$$\max_{j\in\mathcal{S}^*}\frac{1}{\sqrt{n}}\left|\Delta^{\boldsymbol{\pi}_j}\right|\leq\left(1+\left|\beta+\frac{\boldsymbol{\pi}_1}{\boldsymbol{\gamma}_1}\right|\right)\left(1+\frac{|\boldsymbol{\gamma}_j|}{|\boldsymbol{\gamma}_1|}\right)\left(s\frac{\log p}{n}+\frac{1}{|\boldsymbol{\gamma}_1|}\frac{\sqrt{\log p\log p_{\mathrm{z}}}}{n}\right) \tag{C.3.128}$$

By the assumption $\min_{j\in\mathcal{S}^*}|\boldsymbol{\gamma}_j|\gg\sqrt{\log p/n}$ and (C.3.92), we establish (C.3.95).

Bibliography

Andrews, D. W. K., Moreira, M. J., & Stock, J. H. (2007). Performance of conditional wald tests in {IV} regression with weak instruments. *Journal of Econometrics*, 139(1), 116–132.

Angrist, J. D., Imbens, G. W., & Rubin, D. B. (1996). Identification of causal effects using instrumental variables. *Journal of the American Statistical Association*, 91(434), 444–455.

Angrist, J. D. & Krueger, A. B. (1991). Does compulsory school attendance affect schooling and earnings? *Quarterly Journal of Economics*, 106(4), 979–1014.

Arias-Castro, E., Candès, E. J., & Plan, Y. (2011). Global testing under sparse alternatives: Anova, multiple comparisons and the higher criticism. *The Annals of Statistics*, 39(5), 2533–2556.

Baiocchi, M., Cheng, J., & Small, D. S. (2014). Instrumental variable methods for causal inference. *Statistics in Medicine*, 33(13), 2297–2340.

Baraud, Y. (2002). Non-asymptotic minimax rates of testing in signal detection. *Bernoulli*, 8(5), 577–606.

Bayati, M. & Montanari, A. (2012). The lasso risk for gaussian matrices. *Information Theory, IEEE Transactions on*, 58(4), 1997–2017.

Belloni, A., Chen, D., Chernozhukov, V., & Hansen, C. (2012). Sparse models and methods for optimal instruments with an application to eminent domain. *Econometrica*, 80(6), 2369–2429.

Belloni, A., Chernozhukov, V., & Hansen, C. (2014). Inference on treatment effects after selection among high-dimensional controls. *The Review of Economic Studies*, 81(2), 608–650.

Belloni, A., Chernozhukov, V., & Wang, L. (2011). Square-root lasso: pivotal recovery of sparse signals via conic programming. *Biometrika*, 98(4), 791–806.

Bickel, P. J., Ritov, Y., & Tsybakov, A. B. (2009). Simultaneous analysis of lasso and dantzig selector. *The Annals of Statistics*, 37(4), 1705–1732.

Blundell, R., Dearden, L., & Sianesi, B. (2005). Evaluating the effect of education on earnings: models, methods and results from the national child development survey. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 168(3), 473–512.

Boucheron, S., Lugosi, G., & Massart, P. (2013). *Concentration inequalities: A nonasymptotic theory of independence*. OUP Oxford.

Bowden, J., Davey Smith, G., & Burgess, S. (2015). Mendelian randomization with invalid instruments: effect estimation and bias detection through egger regression. *International Journal of Epidemiology*, 44(2), 512–525.

Bowden, J., Davey Smith, G., Haycock, P. C., & Burgess, S. (2016). Consistent estimation in mendelian randomization with some invalid instruments using a weighted median estimator. *Genetic Epidemiology*, 40(4), 304–314.

Bühlmann, P. & van de Geer, S. (2011). *Statistics for high-dimensional data: methods, theory and applications*. Springer Science & Business Media.

Burgess, S., Bowden, J., Dudbridge, F., & Thompson, S. G. (2016). Robust instrumental variable methods using multiple candidate instruments with application to mendelian randomization. *arXiv*.

Burgess, S., Timpson, N. J., Ebrahim, S., & Davey Smith, G. (2015). Mendelian randomization: where are we now and where are we going? *International Journal of Epidemiology*, 44(2), 379–388.

Cai, T. T. & Guo, Z. (2016a). Accuracy assessment for high-dimensional linear regression. *arXiv preprint arXiv:1603.03474v2*.

Cai, T. T. & Guo, Z. (2016b). Confidence intervals for high-dimensional linear regression: Minimax rates and adaptivity. *The Annals of statistics*, To appear.

Cai, T. T. & Guo, Z. (2016c). Supplement to "confidence intervals for high-dimensional linear regression: Minimax rates and adaptivity". *The Annals of statistics*, To appear.

Cai, T. T. & Liu, W. (2011). Adaptive thresholding for sparse covariance matrix estimation. *Journal of the American Statistical Association*, 106, 672–684.

Cai, T. T., Liu, W., & Luo, X. (2011). A constrained ℓ_1 minimization approach to sparse precision matrix estimation. *Journal of the American Statistical Association*, 106(494), 594–607.

Cai, T. T. & Low, M. G. (2004). Minimax estimation of linear functionals over nonconvex parameter spaces. *The Annals of statistics*, 32(2), 552–576.

Cai, T. T. & Low, M. G. (2005). An adaptation theory for nonparametric confidence intervals. *The Annals of statistics*, 32(5), 1805–1840.

Cai, T. T. & Low, M. G. (2006). Adaptive confidence balls. *The Annals of Statistics*, 34(1), 202–228.

Cai, T. T., Low, M. G., & Ma, Z. (2014). Adaptive confidence bands for nonparametric regression functions. *Journal of the American Statistical Association*, 109, 1054–1070.

Cai, T. T. & Zhou, H. H. (2009). A data-driven block thresholding approach to wavelet estimation. *The Annals of Statistics*, 37(2), 569–595.

Cai, T. T. & Zhou, H. H. (2012). Optimal rates of convergence for sparse covariance matrix estimation. *The Annals of Statistics*, 40(5), 2389–2420.

Candès, E. & Tao, T. (2007). The dantzig selector: statistical estimation when p is much larger than n. *The Annals of Statistics*, 35(6), 2313–2351.

Card, D. (1993). *Using Geographic Variation in College Proximity to Estimate the Return to Schooling*. Working Paper 4483, National Bureau of Economic Research.

Card, D. (1999). Chapter 30 - the causal effect of education on earnings. In O. C. Ashenfelter & D. Card (Eds.), *Handbook of Labor Economics*, volume 3, Part A (pp. 1801 – 1863). Elsevier.

Cheng, X. & Liao, Z. (2015). Select the valid and relevant moments: An information-based lasso for gmm with many moments. *Journal of Econometrics*, 186(2), 443–464.

Chernozhukov, V., Hansen, C., & Spindler, M. (2015a). Post-selection and post-regularization inference in linear models with many controls and instruments.

Chernozhukov, V., Hansen, C., & Spindler, M. (2015b). Valid post-selection and post-regularization inference: An elementary, general approach. *arXiv preprint arXiv:1501.03430*.

Collier, O., Comminges, L., & Tsybakov, A. B. (2015). Minimax estimation of linear and quadratic functionals on sparsity classes. *arXiv preprint arXiv:1502.00665*.

Davey Smith, G. & Ebrahim, S. (2003). Mendelian randomization: can genetic epidemiology contribute to understanding environmental determinants of disease? *International Journal of Epidemiology*, 32(1), 1–22.

Davey Smith, G. & Ebrahim, S. (2004). Mendelian randomization: prospects, potentials, and limitations. *International Journal of Epidemiology*, 33(1), 30–42.

Didelez, V. & Sheehan, N. (2007). Mendelian randomization as an instrumental variable approach to causal inference. *Statistical methods in medical research*, 16(4), 309–330.

Donoho, D. L. (1995). De-noising by soft-thresholding. *IEEE transactions on information theory*, 41(3), 613–627.

Donoho, D. L. & Johnstone, I. M. (1995). Adapting to unknown smoothness via wavelet shrinkage. *Journal of the American Statistical Association*, 90(432), 1200–1224.

Donoho, D. L. & Johnstone, J. M. (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, 81(3), 425–455.

Donoho, D. L., Maleki, A., & Montanari, A. (2011). The noise-sensitivity phase transition in compressed sensing. *Information Theory, IEEE Transactions on*, 57(10), 6920–6941.

Fan, J. & Liao, Y. (2014). Endogeneity in high dimensions. *Annals of statistics*, 42(3), 872.

Gary-Bobo, R., Picard, N., & Prieto, A. (2006). *Birth Order and Sibship Sex Composition as Instruments in the Study of Education and Earnings*. CEPR Discussion Papers 5514, C.E.P.R. Discussion Papers.

Gautier, E. & Tsybakov, A. B. (2011). High-dimensional instrumental variables regression and confidence sets. *arXiv preprint arXiv:1105.2454*.

Guo, Z., Wang, W., Cai, T. T., & Li, H. (2016). Optimal estimation of co-heritability in high-dimensional linear models. *arXiv preprint arXiv:1605.07244*.

Han, C. (2008). Detecting invalid instruments using l 1-gmm. *Economics Letters*, 101(3), 285–287.

Hauser, R. M. (2005). Survey response in the long run: The wisconsin longitudinal study. *Field Methods*, 17(1), 3–29.

Hernán, M. A. & Robins, J. M. (2006). Instruments for causal inference: An epidemiologist's dream? *Epidemiology*, 17(4), 360–372.

Hoffmann, M. & Nickl, R. (2011). On adaptive inference and confidence bands. *The Annals of Statistics*, 39(5), 2383–2409.

Holland, P. W. (1988). Causal inference, path analysis, and recursive structural equations models. *Sociological Methodology*, 18(1), 449–484.

Imbens, G. W. (2014). Instrumental variables: An econometrician's perspective. *Statistical Science*, 29(3), 323–358.

Ingster, Y. I., Tsybakov, A. B., & Verzelen, N. (2010). Detection boundary in sparse regression. *Electronic Journal of Statistics*, 4, 1476–1526.

Janson, L., Barber, R. F., & Candès, E. (2015). Eigenprism: Inference for high-dimensional signal-to-noise ratios. *arXiv preprint arXiv:1505.02097*.

Javanmard, A. & Montanari, A. (2014a). Confidence intervals and hypothesis testing for high-dimensional regression. *The Journal of Machine Learning Research*, 15(1), 2869–2909.

Javanmard, A. & Montanari, A. (2014b). Hypothesis testing in high-dimensional regression under the gaussian random design model: Asymptotic theory. *Information Theory, IEEE Transactions on*, 60(10), 6522–6554.

Javanmard, A. & Montanari, A. (2015). De-biasing the lasso: Optimal sample size for gaussian designs. *arXiv preprint arXiv:1508.02757*.

Kang, H., Cai, T. T., & Small, D. S. (2016a). A simple and robust confidence interval for causal effects with possibly invalid instruments. *arXiv preprint arXiv:1504.03718*.

Kang, H., Zhang, A., Cai, T. T., & Small, D. S. (2016b). Instrumental variables estimation with some invalid instruments and its application to mendelian randomization. *Journal of the American Statistical Association*, 111, 132–144.

Kolesár, M., Chetty, R., Friedman, J. N., Glaeser, E. L., & Imbens, G. W. (2015). Identification and inference with many invalid instruments. *Journal of Business & Economic Statistics*, 33(4), 474–484.

Lawlor, D. A., Harbord, R. M., Sterne, J. A. C., Timpson, N., & Davey Smith, G. (2008). Mendelian randomization: Using genes as instruments for making causal inferences in epidemiology. *Statistics in Medicine*, 27(8), 1133–1163.

Leeb, H. & Pötscher, B. M. (2005). Model selection and inference: Facts and fiction. *Econometric Theory*, 21(01), 21–59.

Li, K.-C. (1985). From stein's unbiased risk estimates to the method of generalized cross validation. *The Annals of Statistics*, 13(4), 1352–1377.

Moreira, M. J. (2003). A conditional likelihood ratio test for structural models. *Econometrica*, 71(4), 1027–1048.

Murray, M. P. (2006). Avoiding invalid instruments and coping with weak instruments. *The Journal of Economic Perspectives*, 20(4), 111–132.

Neyman, J. (1923). On the application of probability theory to agricultural experiments. essay on principles. section 9. *Statistical Science*, 5(4), 465–472.

Nickl, R. & van de Geer, S. (2013). Confidence sets in sparse regression. *The Annals of Statistics*, 41(6), 2852–2876.

Raskutti, G., Wainwright, M. J., & Yu, B. (2010). Restricted eigenvalue properties for correlated gaussian designs. *The Journal of Machine Learning Research*, 11, 2241–2259.

Raskutti, G., Wainwright, M. J., & Yu, B. (2011). Minimax rates of estimation for high-dimensional linear regression over-balls. *Information Theory, IEEE Transactions on*, 57(10), 6976–6994.

Ren, Z., Sun, T., Zhang, C.-H., & Zhou, H. H. (2013). Asymptotic normality and optimalities in estimation of large gaussian graphical model. *arXiv preprint arXiv:1309.6024*.

Robins, J. & van der Vaart, A. (2006). Adaptive nonparametric confidence sets. *The Annals of Statistics*, 34(1), 229–253.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5), 688.

Small, D. S. (2007). Sensitivity analysis for instrumental variables regression with overidentifying restrictions. *Journal of the American Statistical Association*, 102(479), 1049–1058.

Staiger, D. & Stock, J. (1997). Instrumental variables regression with weak instruments. *Econometrica*, 65(3), 557–586.

Stein, C. M. (1981). Estimation of the mean of a multivariate normal distribution. *The Annals of Statistics*, 9(6), 1135–1151.

Stock, J. H., Wright, J. H., & Yogo, M. (2002). A survey of weak instruments and weak identification in generalized method of moments. *Journal of Business & Economic Statistics*, 20(4).

Sun, T. & Zhang, C.-H. (2012). Scaled sparse linear regression. *Biometrika*, 101(2), 269–284.

Swanson, S. A. & Hernán, M. A. (2013). Commentary: How to report instrumental variables analyses (suggestions welcome). *Epidemiology*, 24, 370–374.

Thrampoulidis, C., Panahi, A., & Hassibi, B. (2015). Asymptotically exact error analysis for the generalized ℓ_2^2-lasso. *arXiv preprint arXiv:1502.06287*.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1), 267–288.

van de Geer, S., Bühlmann, P., Ritov, Y., & Dezeure, R. (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. *The Annals of Statistics*, 42(3), 1166–1202.

Varian, H. R. (2014). Big data: New tricks for econometrics. *The Journal of Economic Perspectives*, 28(2), 3–27.

Vershynin, R. (2012). Introduction to the non-asymptotic analysis of random matrices. In Y. Eldar & G. Kutyniok (Eds.), *Compressed Sensing: Theory and Applications* (pp. 210–268). Cambridge University Press.

Verzelen, N. (2012). Minimax risks for sparse regressions: Ultra-high dimensional phenomenons. *Electronic Journal of Statistics*, 6, 38–90.

Windmeijer, F., Farbmacher, H., Davies, N., & Davey Smith, G. (2016). *On the Use of the Lasso for Instrumental Variables Estimation with Some Invalid Instruments*. Technical report, Department of Economics, University of Bristol, UK.

Wooldridge, J. M. (2010). *Econometric Analysis of Cross Section and Panel Data*. MIT press, 2nd ed. edition.

Ye, F. & Zhang, C.-H. (2010). Rate minimaxity of the lasso and dantzig selector for the ℓ_q loss in ℓ_r balls. *The Journal of Machine Learning Research*, 11, 3519–3540.

Yi, F. & Zou, H. (2013). SURE-tuned tapering estimation of large covariance matrices. *Computational Statistics & Data Analysis*, 58, 339–351.

Zhang, C.-H. & 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(1), 217–242.

Zhou, S. (2009). Restricted eigenvalue conditions on subgaussian random matrices. *arXiv preprint arXiv:0912.4045*.