Journal of Machine Learning Research 26 (2025) 1-54 Submitted 1/24; Revised 4/25; Published 5/25

Linear Hypothesis Testing in High-Dimensional Expected Shortfall Regression with Heavy-Tailed Errors

Gaoyu Wu G5WU@UCSD.EDU
Department of Mathematics
University of California, San Diego
La Jolla, CA 92093

Jelena Bradic JBRADIC@UCSD.EDU
Department of Mathematics and Halıcıoğlu Data Science Institute
University of California, San Diego
La Jolla, CA 92093

Kean Ming Tan KEANMING@UMICH.EDU
Department of Statistics
University of Michigan
Ann Arbor, MI 48109

Wen-Xin Zhou WENXINZ@UIC.EDU
Department of Information and Decision Sciences
University of Illinois Chicago
Chicago, IL 60607

Editor: Mladen Kolar

Abstract

Expected shortfall (ES) is widely used for characterizing the tail of a distribution across various fields, particularly in financial risk management. In this paper, we explore a two-step procedure that leverages an orthogonality property to reduce sensitivity to nuisance parameters when estimating within a joint quantile and expected shortfall regression framework. For high-dimensional sparse models, we propose a robust ℓ_1-penalized two-step approach capable of handling heavy-tailed data distributions. We establish non-asymptotic estimation error bounds and propose an appropriate growth rate for the diverging robustification parameter. To facilitate statistical inference for certain linear combinations of the ES regression coefficients, we construct debiased estimators and develop their asymptotic distributions, which form the basis for constructing valid confidence intervals. We validate the proposed method through simulation studies, demonstrating its effectiveness in high-dimensional linear models with heavy-tailed errors.

Keywords: conditional value-at-risk, expected shortfall, heavy-tailed data, Huber loss, quantile regression

1. Introduction

Value-at-risk (VaR) and expected shortfall are two commonly used measures for quantifying risk. VaR measures the maximum potential loss that could be incurred at a specified confidence level, whereas ES represents the expected loss that exceeds the VaR threshold. Despite its popularity, VaR has several drawbacks as a risk metric, including the violation of sub-additivity and the inability to capture tail risks beyond the specified quantile level (Artzner et al., 1999; Acerbi, 2002; Bayer and Dimitriadis, 2022). Recently, the Basel Committee adopted ES as the standard risk measure for financial institutions, replacing VaR (Basel Committee, 2016, 2019). This shift in focus has led to the development of new methods for estimating, forecasting, and backtesting ES in the banking and insurance industries (Nolde and Ziegel, 2017; Bercu et al., 2021; Hallin and Trucíos, 2023; Deng and Qiu, 2021; Bayer and Dimitriadis, 2022). ES has also been widely adopted in other domains, such as operations research (Rockafellar et al., 2014; Soleimani and Govindan, 2014), and treatment effect analysis (He et al., 2010; Chen and Yen, 2025; Wei et al., 2024).

Let Z be a real-valued random variable and let F_Z be its cumulative distribution function (CDF), i.e., $F_Z(z) = \mathbb{P}(Z \le z)$. The quantile and ES of Z at level $\alpha \in (0,1)$ are defined as $Q_\alpha(Z) = \inf\{z \in \mathbb{R} : F_Z(z) \ge \alpha\}$ and $E_\alpha(Z) = \mathbb{E}\{Z|Z \le Q_\alpha(Z)\}$, respectively. In financial applications, Z often indicates the payoff of a portfolio, and $E_\alpha(Z)$ represents the average return of a portfolio given that a return is occurring at or below the quantile level α. A more detailed discussion of ES and its properties can be found in Rockafellar and Royset (2013) and McNeil et al. (2015).

Despite the importance of ES as a risk measure, there are limited estimation and statistical inference procedures available for examining the relationship between a p-dimensional vector of covariates $X \in \mathbb{R}^p$ and the ES of the outcome variable $Y \in \mathbb{R}$, especially when p is large; see, for instance, Scaillet (2005), Cai and Wang (2008) and Kato (2012), among others. This is primarily due to the non-elicitability of ES, which implies that ES cannot be directly optimized through the minimization of a loss function (Gneiting, 2011). While ES is not elicitable on its own, in their seminal work, Fissler and Ziegel (2016) have shown that the ES is jointly elicitable with the quantile under a class of joint loss functions.

We consider the joint linear quantile and ES model:

$$Q_\alpha(Y|X) = X^{\mathrm{T}}\beta^* \qquad \text{and} \qquad E_\alpha(Y|X) = X^{\mathrm{T}}\theta^*, \tag{1}$$

where $Q_\alpha(Y_i|X_i)$ and $E_\alpha(Y_i|X_i) = \mathbb{E}\{Y_i|Y_i \le Q_\alpha(Y_i|X_i), X_i\}$ are the conditional α-level quantile and ES of Y given X, respectively. Here, $\beta^* = \beta^*_\alpha$, $\theta^* = \theta^*_\alpha \in \mathbb{R}^p$ are the quantile and ES regression coefficients that can vary across different quantile levels α. We suppress their dependency on α for simplicity.

Under the joint model (1), Dimitriadis and Bayer (2019) and Patton et al. (2019) proposed to simultaneously estimate the quantile and ES regression coefficients by minimizing a class of non-convex and non-differentiable joint loss functions. Theoretically, they established the consistency and asymptotic normality of the resulting M-estimator, defined as a global minimum. However, this approach involves minimizing a non-convex and non-differentiable joint loss function for which a global optimum is not guaranteed, creating a theoretical gap between the estimator and the theoretical results developed for global minima. Moreover, it is computationally challenging to obtain an estimator from minimizing non-convex functions, especially when the number of covariates, p, is large. From a different perspective, Barendse (2020) proposed a computationally efficient two-step method for ES regression using the orthogonality property by treating the quantile regression coefficient vector as a nuisance parameter. Specifically, the two-step method involves fitting a quantile regression and solving a least squares problem with surrogate response variables. Theoretically, Barendse (2020) established that the resulting ES regression estimator is consistent and asymptotically normal under the fixed-p regime, while He et al. (2023) studied the increasing-p regime under the scaling condition $p = O(n^a)$ for some $a \in [1/2, 1)$.

In this work, we focus on the data-rich setting in which the dimension can be as large as or larger than the sample size. To address this, we assume sparsity, where only a small subset of p covariates affects the response, and use an l_1 penalty to encourage sparse regression coefficients; see Wainwright (2019) and Fan et al. (2020), for example. Under high-dimensional sparse models, Barendse (2023) and Zhang et al. (2023) extended the two-step approach of Barendse (2020) and proposed an ℓ_1-penalized two-step least squares ES estimator. However, a critical limitation of these methods lies in their reliance on least squares loss in the second step, which makes them highly sensitive to heavy-tailed data and outliers. This vulnerability is exacerbated in high dimensions due to spurious correlations (Fan et al., 2018; Sun et al., 2020). To address this limitation and improve robustness, motivated by He et al. (2023), we propose a novel modification of the two-step ES estimation procedure by replacing the standard least squares loss in the second stage with the Huber loss. However, adjusting the loss function alone is insufficient to fully address the challenges posed by heavy-tailed errors. A key innovation of our approach lies in the precise selection of the parameter τ in the Huber loss to optimally trade bias for robustness. We show that sub-Gaussian deviation bounds can be attained even when the conditional distribution of Y given X is heavy-tailed, as long as a diverging robustification parameter τ is chosen appropriately.

In the context of high-dimensional ES regression, existing methods have significant limitations. Barendse (2023) provides rates restricted to polynomially growing p, while Zhang et al. (2023) achieve better rates but rely on light-tailed distributions, making them unsuitable for heavy-tailed data. Given an ℓ_1-penalized robust ES regression estimator, we develop a framework for performing statistical inference on $a^{\mathrm{T}}\theta^*$, where $a \in \mathbb{R}^p$ is a pre-specified p-dimensional vector based on the scientific question of interest. Zhang et al. (2023) attempted to address inference by proposing a debiased estimator for individual coefficients using nodewise regression (van de Geer et al., 2014). However, their approach is confined to coordinate-wise inference and does not extend to linear projections, which are essential for testing contrasts. Linear projections combine multiple coordinates, accumulating biases from individual components and amplifying cross-dependencies among predictors, even when the vector a is sparse. Motivated by these gaps, we propose a debiased estimator for $a^{\mathrm{T}}\theta^*$ that directly addresses these challenges.

Notations: For any two $\mathbb{R}^k$ vectors $u = (u_1, \ldots, u_k)^{\mathrm{T}}$ and $v = (v_1, \ldots, v_k)^{\mathrm{T}}$, we write their inner product as $u^{\mathrm{T}}v = \langle u, v\rangle = \sum_{j=1}^k u_j v_j$. We use $\|\cdot\|_p$ ($1 \le p \le \infty$) to denote the ℓ_p-norm in $\mathbb{R}^k$: $\|u\|_p = (\sum_{j=1}^k |u_j|^p)^{1/p}$ for $p \ge 1$ and $\|u\|_\infty = \max_{1\le j\le k} |u_j|$. For a positive semi-definite matrix $A \in \mathbb{R}^{k\times k}$ and $u \in \mathbb{R}^k$, let $\|u\|_A = \|A^{1/2}u\|_2 = \sqrt{u^{\mathrm{T}}Au}$. For any matrix $A = (a_{ij})_{1\le i,j\le k} \in \mathbb{R}^{k\times k}$, we denote $\|A\|_1$ and $\|A\|_2$ as the induced ℓ_1- and ℓ_2-norm, respectively, where $\|A\|_1 = \max_{1\le j\le k} \sum_{i=1}^k |a_{ij}|$ and $\|A\|_2 = \sigma_{\max}(A)$, with $\sigma_{\max}(A)$ being the largest singular value of A. Additionally, we denote $\|A\|_{\max} = \max_{1\le i,j\le k} |a_{ij}|$ as the maximum element of A in magnitude. For two real numbers a and b, we write $a \wedge b = \min\{a, b\}$ and $a \vee b = \max\{a, b\}$. For two sequences of nonnegative real numbers $\{a_n\}_{n\ge1}$ and $\{b_n\}_{n\ge1}$, we write $a_n \lesssim b_n$ if $a_n \le Cb_n$ for some constant $C > 0$ (independent of n), $a_n \gtrsim b_n$ if $b_n \lesssim a_n$, and $a_n \asymp b_n$ if $a_n \lesssim b_n$ and $a_n \gtrsim b_n$.

2. Preliminaries

In this section, we provide a brief overview of the joint quantile and expected shortfall model, along with the motivation behind the two-step method for robust expected shortfall regression.

2.1 Joint Linear Quantile and Expected Shortfall Model

Let $\{(Y_i, X_i)\}_{i=1}^n$ be a sequence of independent and identically distributed observations, where $Y_i \in \mathbb{R}$ is the response variable and $X_i \in \mathbb{R}^p$ is a p-dimensional vector of covariates. We start with a brief review of existing work on ES regression under the joint linear quantile and ES model in equation (1). For the identification of (conditional) quantile and ES jointly, Fissler and Ziegel (2016) proposed the following class of loss functions

$$\begin{aligned} L_{\mathrm{FZ}}(\beta, \theta; Y, X) = &\{\alpha - \mathbb{1}(Y \le X^{\mathrm{T}}\beta)\}\{G_1(Y) - G_1(X^{\mathrm{T}}\beta)\} \\ &+ \frac{G_2(X^{\mathrm{T}}\theta)}{\alpha}\{\alpha X^{\mathrm{T}}(\theta - \beta) - (Y - X^{\mathrm{T}}\beta)\mathbb{1}(Y \le X^{\mathrm{T}}\beta)\} - \mathcal{G}_2(X^{\mathrm{T}}\theta), \end{aligned} \tag{2}$$

and showed that (β^*, θ^*) is the unique minimizer of $\mathbb{E}\{L_{\mathrm{FZ}}(\beta, \theta; Y, X)|X\}$ almost surely. Here, G_1, G_2, and $\mathcal{G}_2$ are real-valued functions satisfying: (1) G_1 is increasing and integrable, (2) $\mathcal{G}_2' = G_2$, and (3) $\mathcal{G}_2$ is strictly increasing and strictly convex.

Using the above construction, Dimitriadis and Bayer (2019) proposed estimating (β^*, θ^*) by minimizing the loss function in equation (2), leading to

$$(\widehat{\beta}^{\mathrm{joint}}, \widehat{\theta}^{\mathrm{joint}}) \in \mathrm{argmin}_{(\beta,\theta)\in\Theta} \frac{1}{n}\sum_{i=1}^n L_{\mathrm{FZ}}(\beta, \theta; Y_i, X_i), \tag{3}$$

where $\Theta \subseteq \mathbb{R}^p \times \mathbb{R}^p$ is a compact and convex parameter space with a nonempty interior. Under the fixed p regime, Dimitriadis and Bayer (2019) showed that the estimators $(\widehat{\beta}^{\mathrm{joint}}, \widehat{\theta}^{\mathrm{joint}})$ are consistent and asymptotically normal. However, for problems with a large number of covariates p, the above method becomes computationally challenging due to the non-differentiable and non-convex objective function (2), regardless of the choice of feasible functions G_1 and G_2 (Fissler and Ziegel, 2016).

From a different perspective, Barendse (2020) proposed a two-step method for estimating (β^*, θ^*) using a tailored score function that satisfies certain orthogonality conditions. Let $S(\beta, \theta; Y, X) := (Y - X^{\mathrm{T}}\beta)\mathbb{1}(Y \le X^{\mathrm{T}}\beta) + \alpha X^{\mathrm{T}}(\beta - \theta)$ and let $\psi(\beta, \theta; X) = \mathbb{E}\{S(\beta, \theta; Y, X)|X\}$ be its expectation. It can be shown that the true regression coefficients (β^*, θ^*) satisfy the moment condition $\psi(\beta^*, \theta^*; X) = 0$ almost surely. To see this, note that under the joint model (1),

$$\begin{aligned} \psi(\beta^*, \theta^*; X) &= \mathbb{E}\{Y\mathbb{1}(Y \le X^{\mathrm{T}}\beta^*)|X\} - X^{\mathrm{T}}\beta^* F_{Y|X}(X^{\mathrm{T}}\beta^*) + \alpha X^{\mathrm{T}}(\beta^* - \theta^*) \\ &= E_\alpha(Y|X)F_{Y|X}(X^{\mathrm{T}}\beta^*) - \alpha X^{\mathrm{T}}\theta^* + \{\alpha - F_{Y|X}(X^{\mathrm{T}}\beta^*)\}X^{\mathrm{T}}\beta^* = 0. \end{aligned} \tag{4}$$

Furthermore, the partial derivative of $\psi(\beta, \theta; X)$ with respect to β, evaluated at $\beta = \beta^*$, takes the form

$$\frac{\partial}{\partial\beta}\psi(\beta, \theta; X)\big|_{\beta=\beta^*} = \{\alpha - F_{Y|X}(X^{\mathrm{T}}\beta^*)\}X = 0, \tag{5}$$

which we refer to as the Neyman orthogonality property throughout this paper.

Motivated by (5), Barendse (2020) proposed a two-step method for fitting the ES regression: (i) compute an estimator $\widehat{\beta}$ of β^* by fitting the quantile regression (Koenker and Bassett Jr, 1978); (ii) obtain an estimator $\widehat{\theta}$ of θ^* by solving

$$\widehat{\theta}^{\text{twostep}} \in \arg\min_{\theta \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^{n} S^2(\widehat{\beta}, \theta; Y_i, X_i). \tag{6}$$

Due to the Neyman orthogonality condition as in equation (5), Barendse (2020) showed that the estimation error of the quantile regression estimator is first-order negligible. Their results were further supported by He et al. (2023), who showed that the convergence rate for the ES regression coefficient $\widehat{\theta}^{\text{twostep}}$ under the ℓ_2 norm depends on the estimation error of $\widehat{\beta}$ through a higher-order term $r_0 \cdot \max\{r_0, \sqrt{p/n}\}$, where $\|\widehat{\beta} - \beta^*\|_2 \le r_0$.

2.2 Robust Expected Shortfall Regression

Given a quantile regression estimator $\widehat{\beta}$, the two-step method in Section 2.1 involves solving the least squares problem in the second step (Barendse, 2020). To see this, we define the pseudo-response variable

$$Z_i(\beta) = (Y_i - X_i^{\mathrm{T}}\beta)\mathbb{1}(Y_i \le X_i^{\mathrm{T}}\beta) + \alpha X_i^{\mathrm{T}}\beta. \tag{7}$$

Under the joint model (1), it can be shown that $\mathbb{E}\{Z_i(\beta^*)|X_i\} = \alpha X_i^{\mathrm{T}}\theta^*$. Thus, given $\widehat{\beta}$, the estimator $\widehat{\theta}^{\text{twostep}}$ in (6) can be interpreted as the least squares estimator obtained by regressing the generated pseudo-response variables $\widehat{Z}_i := Z_i(\widehat{\beta})$ on αX_i. That is, $\widehat{\theta}^{\text{twostep}}$ can be obtained by minimizing the squared error loss function $n^{-1}\sum_{i=1}^{n}(\widehat{Z}_i - \alpha X_i^{\mathrm{T}}\theta)^2$.

As pointed out by He et al. (2023), the conditional distribution of the pseudo-response variables is asymmetric and left-skewed. Specifically, the joint model (1) is equivalent to

$$Y_i = X_i^{\mathrm{T}}\beta^* + \varepsilon_i, \quad Z_i(\beta^*) = \alpha X_i^{\mathrm{T}}\theta^* + \xi_i, \tag{8}$$

where ε_i and ξ_i can be interpreted as the random noise such that $Q_\alpha(\varepsilon_i|X_i) = 0$ and $\mathbb{E}(\xi_i|X_i) = 0$, respectively. Accordingly, we have that $Z_i(\beta^*) = \min(\varepsilon_i, 0) + \alpha X_i^{\mathrm{T}}\beta^*$, and by (4), $\mathbb{E}\{\min(\varepsilon_i, 0)|X_i\} = \alpha X_i^{\mathrm{T}}(\theta^* - \beta^*)$. Therefore, $\xi_i = \min(\varepsilon_i, 0) - \mathbb{E}\{\min(\varepsilon_i, 0)|X_i\}$. Due to the right-truncation at 0, the conditional distribution of ξ_i given X_i is asymmetric and left-skewed and may be heavy-tailed if the random noise ε_i follows a heavy-tailed distribution. In such scenarios, the least squares estimator $\widehat{\theta}^{\text{twostep}}$ can be sensitive to potential outliers and may not be statistically efficient for estimating θ^*.

To address the issues above, He et al. (2023) proposed a robust ES regression method by replacing the least squares loss with a robust loss. Let $\ell_\tau(u) := (u^2/2)\mathbb{1}(|u| \le \tau) + (\tau|u| - \tau^2/2)\mathbb{1}(|u| > \tau)$ be the Huber loss, where $\tau > 0$ is a robustification parameter that blends the squared error loss and the absolute deviation loss that encourages robustness (Huber, 1973). He et al. (2023) proposed a two-step adaptive Huber estimator for estimating θ^* by solving the following convex optimization problem:

$$\underset{\theta \in \mathbb{R}^p}{\text{minimize}} \ \frac{1}{n} \sum_{i=1}^{n} \ell_\tau(\widehat{Z}_i - \alpha X_i^{\mathrm{T}}\theta). \tag{9}$$

Due to the asymmetric nature of ξ_i, the choice of the robustification parameter τ becomes crucial in achieving an optimal balance between bias and robustness. Selecting τ to be too small will introduce non-negligible bias to the resulting estimator, whereas selecting τ to be too large will yield an estimator that is sensitive to outliers. We refer readers to He et al. (2023) for a comprehensive discussion on the selection of τ for the robust ES regression estimator from (9), under the regime in which $p < n$ and $n, p \to \infty$.

In this paper, we limit heavy-tailedness to the error distribution, assuming that the high-dimensional covariate vector $X \in \mathbb{R}^d$ has either sub-exponential or sub-Gaussian tails. In this case, it can be shown that with high probability, the maximum magnitude of all entries of X grows logarithmically in d. For simplicity, we focus on the case where all entries of X are bounded in magnitude, which facilitates theoretical analysis.

3. ℓ_1-Penalized Robust Expected Shortfall Regression

In this section, we develop a framework for robust estimation and statistical inference on a linear functional of θ^* in the high-dimensional setting in which $p > n$. To this end, we assume that the regression coefficients under the joint linear quantile and ES model in (1) are sparse, i.e., $\|\beta^*\|_0 \le s_\beta$ and $\|\theta^*\|_0 \le s_\theta$, where $\|\beta^*\|_0$ and $\|\theta^*\|_0$ are the number of non-zero elements of β^* and θ^*, respectively.

Let $\rho_\alpha(u) = \{\alpha - \mathbb{1}(u < 0)\}u$ be the quantile loss function. We start with computing an ℓ_1-penalized quantile regression estimator:

$$\widehat{\beta} \in \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \left\{ \frac{1}{n} \sum_{i=1}^n \rho_\alpha(Y_i - X_i^{\mathrm{T}}\beta) + \lambda_q \|\beta\|_1 \right\}, \tag{10}$$

where $\lambda_q > 0$ is a tuning parameter that controls the sparsity level of the quantile regression estimator (Belloni and Chernozhukov, 2011; Wang et al., 2012; Wang and He, 2024). We then compute the pseudo-response $\widehat{Z}_1, \ldots, \widehat{Z}_n$ based on (7). The proposed ℓ_1-penalized Huber-ES regression estimator can then be obtained as follows:

$$\widehat{\theta}_\tau \in \underset{\theta \in \mathbb{R}^p}{\operatorname{argmin}} \left\{ \frac{1}{n} \sum_{i=1}^n \ell_\tau(\widehat{Z}_i - \alpha X_i^{\mathrm{T}}\theta) + \alpha\lambda_e \|\theta\|_1 \right\}, \tag{11}$$

where $\tau > 0$ is a robustification parameter and $\lambda_e > 0$ is a sparsity tuning parameter that controls the number of non-zeros in $\widehat{\theta}_\tau$.

Given $\widehat{\theta}_\tau$, we develop a framework for testing the statistical hypothesis $H_0 : a^{\mathrm{T}}\theta^* = 0$ versus $H_1 : a^{\mathrm{T}}\theta^* \neq 0$, where $a \in \mathbb{R}^p$ is a pre-specified vector based on the scientific question of interest. Due to the ℓ_1-penalty, $\widehat{\theta}_\tau$ is biased and is not asymptotically normal. To address this issue, many authors have proposed different forms of debiased estimators to remove bias induced by the ℓ_1-penalty (Zhang and Zhang, 2014; van de Geer et al., 2014; Javanmard and Montanari, 2014). Motivated by the existing work, we will construct a debiased estimator for $\widehat{\theta}_\tau$ that can be shown to be asymptotically normal in the rest of the section.

We start by introducing some additional notation that will be used throughout the remaining sections. Let $\psi_\tau(v) = \ell'_\tau(v) = \operatorname{sign}(v)\min\{|v|, \tau\}$ be the first order derivative of the Huber loss. Moreover, let $\Sigma = \mathbb{E}(XX^{\mathrm{T}})$ and let $u = \Sigma^{-1}a$. In the low-dimensional setting in which $p < n$ (without imposing the ℓ_1-penalty), it can be shown that $\alpha a^{\mathrm{T}}(\widehat{\theta}_\tau - \theta^*)$

admits a linear approximation $u^{\mathrm{T}} n^{-1} \sum_{i=1}^n \psi_\tau(\xi_i) X_i$ (ignoring higher-order terms), with an appropriately chosen robustification parameter τ (He et al., 2023). However, this is no longer valid in the high-dimensional setting due to bias incurred by the ℓ_1-penalty, and some form of bias correction is needed to ensure asymptotic normality.

We consider a debiased estimator that takes the form $\alpha \cdot a^{\mathrm{T}} \widehat{\theta}_\tau + u^{\mathrm{T}} n^{-1} \sum_{i=1}^n \psi_\tau(\widehat{\xi}_i) X_i$, where $\widehat{\xi}_i = \widehat{Z}_i - \alpha X_i^{\mathrm{T}} \widehat{\theta}_\tau$. Note that

$$
\begin{aligned}
&\alpha \cdot a^{\mathrm{T}}(\widehat{\theta}_\tau - \theta^*) + u^{\mathrm{T}} \frac{1}{n} \sum_{i=1}^n \psi_\tau(\widehat{\xi}_i) X_i \\
&= \underbrace{u^{\mathrm{T}} \frac{1}{n} \sum_{i=1}^n \psi_\tau(\xi_i) X_i}_{\text{variance term}} + \underbrace{\alpha \cdot a^{\mathrm{T}}(\widehat{\theta}_\tau - \theta^*) + \frac{1}{n} \sum_{i=1}^n \{\psi_\tau(\widehat{\xi}_i) - \psi_\tau(\xi_i)\} u^{\mathrm{T}} X_i}_{\text{bias term}}, \qquad (12)
\end{aligned}
$$

where the first and second terms correspond to the variance and bias of the estimator, respectively. To show asymptotic normality, it remains to show that the bias term is $o_{\mathbb{P}}(n^{-1/2})$. Let $\widehat{\Sigma} = n^{-1} \sum_{i=1}^n X_i X_i^{\mathrm{T}}$. The bias term can be rewritten as

$$
\begin{aligned}
&\alpha \cdot a^{\mathrm{T}}(\widehat{\theta}_\tau - \theta^*) + \frac{1}{n} \sum_{i=1}^n \{\psi_\tau(\widehat{\xi}_i) - \psi_\tau(\xi_i)\} u^{\mathrm{T}} X_i \\
&= \frac{1}{n} \sum_{i=1}^n \{\psi_\tau(\widehat{\xi}_i) - \psi_\tau(\xi_i) + \alpha X_i^{\mathrm{T}}(\widehat{\theta}_\tau - \theta^*)\} u^{\mathrm{T}} X_i + \left(a - \frac{1}{n} \sum_{i=1}^n X_i X_i^{\mathrm{T}} u\right)^{\mathrm{T}} \alpha(\widehat{\theta}_\tau - \theta^*) \\
&= (a - \widehat{\Sigma} u)^{\mathrm{T}} \alpha(\widehat{\theta}_\tau - \theta^*) + u^{\mathrm{T}} \{A_n(\widehat{\beta}, \widehat{\theta}_\tau) + B_n(\widehat{\beta}, \widehat{\theta}_\tau)\}, \qquad (13)
\end{aligned}
$$

where $A_n(\widehat{\beta}, \widehat{\theta}_\tau) = n^{-1} \sum_{i=1}^n (1 - \mathbb{E})[\{\psi_\tau(\widehat{\xi}_i) - \psi_\tau(\xi_i) + \alpha X_i^{\mathrm{T}}(\widehat{\theta}_\tau - \theta^*)\} X_i]$ and $B_n(\widehat{\beta}, \widehat{\theta}_\tau) = \mathbb{E}[\{\psi_\tau(\widehat{\xi}_i) - \psi_\tau(\xi_i) + \alpha X_i^{\mathrm{T}}(\widehat{\theta}_\tau - \theta^*)\} X_i]$. Under the scaling condition $\max(s_\beta, s_\theta) \log p = o(\sqrt{n})$, we will show that $\|A_n(\widehat{\beta}, \widehat{\theta}_\tau)\|_\infty = o_{\mathbb{P}}(n^{-1/2})$ and $\|B_n(\widehat{\beta}, \widehat{\theta}_\tau)\|_2 = o_{\mathbb{P}}(n^{-1/2})$. Consequently, it remains to construct an appropriate projection direction u to ensure that $\|u\|_1$ is bounded and that $\|a - \widehat{\Sigma} u\|_\infty$ is sufficiently small.

To this end, we propose to construct the projection direction u as follows:

$$
\begin{aligned}
\widehat{u} = &\operatorname{argmin}_{u \in \mathbb{R}^p} u^{\mathrm{T}} \widehat{\Sigma} u, && (14) \\
&\text{subject to } \|a - \widehat{\Sigma} u\|_\infty \le \rho \|a\|_2, && (15) \\
&\|u\|_1 \le C_a \|a\|_2, && (16) \\
&|a^{\mathrm{T}} \widehat{\Sigma} u - \|a\|_2^2| \le \rho' \|a\|_2^2, && (17)
\end{aligned}
$$

where $\rho, \rho' \in (0, 1)$ are tuning parameters that are chosen to be sufficiently small, and $C_a > 0$ is a constant that does not depend on n and p. The constraint (17) complements (15) by ensuring the proximity between Σu and a. Additionally, it effectively prevents the possibility of $\widehat{u}$ being zero. The optimization problem (14) is a quadratic programming problem and can be solved using existing software, such as the `quadprog` package in `R`. Consequently, a debiased estimator of $\omega^* := a^{\mathrm{T}} \theta^*$ is given by

$$
\widehat{\omega}_\tau := a^{\mathrm{T}} \widehat{\theta}_\tau + \frac{1}{\alpha n} \sum_{i=1}^n \psi_\tau(\widehat{\xi}_i)\, \widehat{u}^{\mathrm{T}} X_i. \qquad (18)
$$

Under the scaling condition $\max(s_\beta, s_\theta)\log p = o(\sqrt{n})$, we will show in Section 4.2 that the proposed debiased estimator $\widehat{\omega}_\tau$ in (18) is asymptotically normal:

$$\alpha\sqrt{n}\,(\widehat{\omega}_\tau - \omega^*)/s(\widehat{u}) \xrightarrow{\mathrm{d}} \mathcal{N}(0,1) \quad \text{as} \quad n, p \to \infty,$$

where $s^2(\widehat{u}) = \widehat{u}^{\mathrm{T}}\widetilde{\Lambda}\widehat{u}$ and $\widetilde{\Lambda} = n^{-1}\sum_{i=1}^n \mathbb{E}(\xi_i^2|X_i)X_iX_i^{\mathrm{T}}$.

To accommodate heavy-tailed noise ξ_i, we propose a truncated variance estimator

$$\widehat{s}_\gamma^2(\widehat{u}) = \widehat{u}^{\mathrm{T}}\widehat{\Lambda}_\gamma\widehat{u} \quad \text{with} \quad \widehat{\Lambda}_\gamma = \frac{1}{n}\sum_{i=1}^n \psi_\gamma^2(\widehat{\xi}_i)X_iX_i^{\mathrm{T}}, \tag{19}$$

where $\gamma := \gamma(n,p) > 0$ is a second robustification parameter. In Theorem 10, we will show that $\widehat{s}_\gamma^2(\widehat{u})$, with $\gamma \asymp (n/\log p)^{1/3}$, is a consistent estimator of the asymptotic variance of $\widehat{\omega}_\tau$. Combining the debiased estimator $\widehat{\omega}_\tau$ in (18) and the truncated (asymptotic) variance estimator in (19), we propose to construct an asymptotic $100\cdot(1-c)\%$ confidence interval for $a^{\mathrm{T}}\theta^*$ $(0<c<1)$ as follows:

$$\left[\widehat{\omega}_\tau - z_{1-c/2}\cdot\frac{\widehat{s}_\gamma(\widehat{u})}{\alpha\sqrt{n}},\, \widehat{\omega}_\tau + z_{1-c/2}\cdot\frac{\widehat{s}_\gamma(\widehat{u})}{\alpha\sqrt{n}}\right], \tag{20}$$

where z_v $(0<v<1)$ is the v-th quantile of the standard normal distribution.

Remark 1 *The objective function* (14) *serves as an upper bound of the variance term in* (12)*. To see this, it is important to note that when the robustification parameter τ is allowed to diverge as a function of n and p, $\mathbb{E}\{\psi_\tau(\xi_i)|X_i\} \to 0$ and $\mathbb{E}\{\psi_\tau^2(\xi_i)|X_i\} \approx \mathbb{E}(\xi_i^2|X_i) = \mathrm{var}(\varepsilon_i \wedge 0|X_i)$. Assuming that the conditional variance is bounded by some constant (almost surely over X), it can be shown that* $\mathrm{var}(n^{-1}\sum_{i=1}^n \psi_\tau(\xi_i)u^{\mathrm{T}}X_i|X_i) \lesssim u^{\mathrm{T}}\widehat{\Sigma}u$.

The constraint (15) *is employed to regulate the bias term, which depends on the term $\|a - (1/n)\sum_{i=1}^n \psi_\tau'(\widehat{\xi}_i)X_iX_i^{\mathrm{T}}u\|_\infty$. The function ψ_τ is absolutely continuous and is differentiable everywhere, with first-order derivative $\psi_\tau'(t) = \mathbb{1}(|t| \le \tau)$. Given that the choice of the robustification parameter is $\tau = \tau(n,p) \asymp \sqrt{n/\log p}$ (up to a constant factor), we have made the convenient substitution of $\psi_\tau'(\cdot)$ with the constant value 1. This simplification has been made to streamline the theoretical analysis.*

Constraint (16) *ensures that the ℓ_1-norm of $\widehat{u}$ is bounded, and similar constraints have been used in the existing literature for high-dimensional inference. For instance, Cai et al. (2021) used* $\max_{1\le i\le n}|X_i^{\mathrm{T}}u| \le \tau_n\|a\|_2$, *where $\sqrt{\log n} \lesssim \tau_n \ll \sqrt{n}$, to control the magnitude of $\widehat{u}$. From a theoretical perspective, this constraint is adequate in cases where the regression errors in a linear model are independent of the covariates. However, the error variables ξ_i in* (8) *exhibit heteroscedasticity and may depend on the covariates X_i. In this case, an upper bound on* $\max_{1\le i\le n}|X_i^{\mathrm{T}}u|$ *may not be sufficient.*

4. Theoretical Analysis

We provide a non-asymptotic upper bound for the estimation error of the proposed ℓ_1-penalized robust ES regression estimator under the high-dimensional regime in which $p > n$. We then show that the debiased estimator (18) is asymptotically normal. Thus, valid statistical inference for testing the linear hypothesis $H_0 : a^{\mathrm{T}}\theta^* = c_0$ for some pre-specified constant $c_0 \in \mathbb{R}$ can be performed using the debiased estimator. Throughout our theoretical analysis, we assume the joint linear quantile and ES model in (1) and that both QR and ES coefficients are sparse, i.e., $\|\beta^*\|_0 \le s_\beta$ and $\|\theta^*\|_0 \le s_\theta$ respectively, where $\max(s_\beta, s_\theta) \ll n$.

4.1 Non-Asymptotic Upper Bounds on the Estimation Error

We start with some conditions on X and the conditional distribution of Y given X.

Condition 1 *The random covariate vector $X \in \mathbb{R}^p$ satisfies $\|X\|_\infty \le B_X$, where $B_X \ge 1$ is a dimension-free constant. The matrix $\Sigma = \mathbb{E}[XX^{\mathrm{T}}]$ is positive definite with $0 < \underline{\kappa}^2 \le \lambda_{\min}(\Sigma) \le \lambda_{\max}(\Sigma) \le \overline{\kappa}^2 < \infty$. Moreover, assume that $\kappa_3 := \sup_{u\in\mathbb{S}^{p-1}} \frac{\mathbb{E}|\langle X,u\rangle|^3}{(\mathbb{E}\langle X,u\rangle^2)^{3/2}}$ is dimension-free.*

Condition 1 assumes that the covariates X are bounded with a finite third moment and that the population covariance matrix of X has bounded eigenvalues. We note that the boundedness condition on X is imposed primarily for ease of presentation. The condition can be easily relaxed to a sub-Gaussian assumption, and similar results will hold. In such a case, $\|X\|_\infty \lesssim \sqrt{\log p}$ with high probability.

The ℓ_1-penalized robust ES regression estimator in (11) is computed based on an ℓ_1-penalized QR estimator $\widehat{\beta}$, and thus it is necessary to first characterize the estimation error of $\widehat{\beta}$. To this end, we impose some conditions on the conditional density of ε given X that are commonly used in the existing literature on high-dimensional quantile regression (Belloni and Chernozhukov, 2011; Wang and He, 2024). Upper bounds on the estimation error of the ℓ_1-penalized QR estimator $\widehat{\beta}$ under both the ℓ_2 and ℓ_1 norms are provided in Proposition 2.

Condition 2 *The conditional density function of $\varepsilon := Y - X^{\mathrm{T}}\beta^*$ given X, denoted by $f_{\varepsilon|X}(\cdot)$, exists and is continuous on its support. Moreover, there exist constants $\underline{f}, l_0 > 0$ such that $f_{\varepsilon|X}(0) \ge \underline{f}$ and $|f_{\varepsilon|X}(t) - f_{\varepsilon|X}(0)| \le l_0|t|$ for all $t \in \mathbb{R}$ almost surely (over X).*

Proposition 2 *Assume Conditions 1 and 2 hold. For any $t > 0$, the ℓ_1-penalized QR estimator $\widehat{\beta}$ in (10) with sparsity tuning parameter chosen as*

$$\lambda_q \ge 2\left\{\sqrt{2\alpha(1-\alpha)}\overline{\kappa}\sqrt{\frac{\log(2p)+t}{n}} + \bar{\alpha}B_X\frac{\log(2p)+t}{n}\right\} \tag{21}$$

satisfies the error bounds

$$\|\widehat{\beta} - \beta^*\|_\Sigma \le \frac{4}{\underline{f}} r(n,p,t) \quad \text{and} \quad \|\widehat{\beta} - \beta^*\|_1 \le \frac{16}{\underline{\kappa}\underline{f}} s_\beta^{1/2} r(n,p,t) \tag{22}$$

with probability at least $1 - 2e^{-t}$ as long as $r(n,p,t) < 3\underline{f}^2/(8l_0\kappa_3)$, where

$$r(n,p,t) := \frac{s_\beta^{1/2}}{\underline{\kappa}}\left\{16\sqrt{2}\bar{\alpha}\overline{\kappa}\sqrt{\frac{\log(2p)+t}{n}} + 16\bar{\alpha}B_X\frac{\log(2p)+t}{n} + \lambda_q\right\}$$

and $\bar{\alpha} = \max\{\alpha, 1-\alpha\}$.

Using $t = \log(n)$, the upper bounds in (22) can be simplified to $\|\widehat{\beta}-\beta^*\|_2 \lesssim \sqrt{s_\beta \log(p)/n}$ and $\|\widehat{\beta} - \beta^*\|_1 \lesssim s_\beta\sqrt{\log(p)/n}$. Similar results have also been observed in Belloni and Chernozhukov (2011) and Wang and He (2024). The next condition concerns the QR residual ε.

Condition 3 *The conditional CDF $F_{\varepsilon|X}(\cdot)$ of ε given X is continuously differentiable and satisfies $|F_{\varepsilon|X}(t) - F_{\varepsilon|X}(0)| \le \overline{f}|t|$ for all $t \in \mathbb{R}$. The negative of part of ε, denoted by $\varepsilon_- := \min\{\varepsilon, 0\}$, satisfies $\underline{\sigma}^2 \le \mathrm{var}(\varepsilon_-^2|X) \le \overline{\sigma}^2$.*

Let $\mathbb{B}_\Sigma(r) = \{\delta \in \mathbb{R}^p : \|\delta\|_\Sigma \le r\}$ and $\mathbb{C}(l_1) = \{\delta \in \mathbb{R}^p : \|\delta\|_1 \le l_1\|\delta\|_\Sigma\}$. With the above upper bounds on $\widehat{\beta}$ and Condition 3, we establish an upper bound for the estimation error of the robust ES estimator $\widehat{\theta}_\tau$ in the following theorem.

Theorem 3 *Assume that Conditions 1–3 hold. For any given $t > 0$, select the robustification parameter τ such that $\tau \asymp \overline{\sigma}\sqrt{n/(\log(p)+t)}$. Conditioning on the event $\{(\widehat{\beta} - \beta^*) \in \mathbb{B}_\Sigma(r_0) \cap \mathbb{C}(l_1)\}$ for some $0 < r_0 \le 1$ and $l_1 \asymp \underline{\kappa}^{-1}\sqrt{s_\beta}$, and selecting the sparsity tuning parameter λ_e to satisfy*

$$\lambda_e \gtrsim \max\left\{(\overline{\sigma} + B_X l_1 r_0)\overline{\kappa}\sqrt{\frac{\log(2p)+t}{n}},\, \underline{\kappa} s_\theta^{-1/2}\overline{f} r_0^2\right\}, \tag{23}$$

with probability at least $1 - 3e^{-t}$, we have

$$\alpha\|\widehat{\theta}_\tau - \theta^*\|_\Sigma \le 8\underline{\kappa}^{-1}s_\theta^{1/2}\lambda_e \quad \text{and} \quad \alpha\|\widehat{\theta}_\tau - \theta^*\|_1 \le 40\underline{\kappa}^{-2}s_\theta\lambda_e \tag{24}$$

as long as n is sufficiently large such that

$$n \gtrsim \{(\overline{\kappa} \vee B_X)/\underline{\kappa}\}^2 B_X^2 (s_\beta \vee s_\theta)^2\{\log(p) + t\}. \tag{25}$$

Theorem 3 sheds light on how the estimation error of $\widehat{\beta}$ enters the estimation bound for the ℓ_1-penalized robust ES estimator $\widehat{\theta}_\tau$. Specifically, due to the Neyman orthogonality property (5), the estimation error of $\widehat{\beta}$ appears only in the higher-order terms, i.e., r_0^2 and $l_1 r_0\sqrt{\log(p)/n}$, suggesting that $\widehat{\theta}_\tau$ is first-order insensitive to the QR estimation error from the first step. To simplify the results, we view $B_X, \overline{\kappa}, \overline{\sigma}, \underline{\kappa}, \overline{f}, \underline{f}$ and l_0 as constants independent of the dimensions (s, p, n). From Theorem 3 and Proposition 2, choosing the robustification parameter $\tau \asymp \sqrt{n/\log p}$ and sparsity tuning parameters $\lambda_q \asymp \lambda_e \asymp \sqrt{\log(p)/n}$, we conclude that, under the scaling condition $n \gtrsim \max(s_\beta, s_\theta)^2 \log p$,

$$\alpha\|\widehat{\theta}_\tau - \theta^*\|_\Sigma \lesssim \frac{s_\beta s_\theta^{1/2}\log p}{n} + \sqrt{\frac{s_\theta \log p}{n}} \lesssim \sqrt{\frac{s_\theta \log p}{n}} \quad \text{and}$$
$$\alpha\|\widehat{\theta}_\tau - \theta^*\|_1 \lesssim \frac{s_\beta s_\theta \log p}{n} + s_\theta\sqrt{\frac{\log p}{n}} \lesssim s_\theta\sqrt{\frac{\log p}{n}}$$

hold with high probability. Additionally, if we are willing to assume that X has dimension-free kurtosis $\kappa_4 = \sup_{u\in\mathbb{S}^{p-1}} \mathbb{E}\langle X, u\rangle^4/(\mathbb{E}\langle X, u\rangle^2)^2$, the above results hold under a weaker sample size condition: $n \gtrsim \max\{s_\beta, s_\theta\}\log p$. Under a stronger assumption that a higher conditional moment of ε_- exists, for example, the q-th moment $\mathbb{E}(|\varepsilon_-|^q|X)$ with $q \ge 3$, we show that the robustification parameter τ can be chosen more flexibly while maintaining the same rates of convergence in the following proposition.

Proposition 4 *Assume that Conditions 1–3 hold, and that* $\mathbb{E}(|\varepsilon_-|^q|X) \le \alpha_q$ *almost surely for some* $\alpha_q > 0$*. For any given* $t > 0$*, let the robustification parameter* τ *satisfy*

$$\alpha_q^{1/q}\left\{\sqrt{\frac{n}{\log(p)+t}}\right\}^{1/(q-1)} \lesssim \tau \lesssim \overline{\sigma}\sqrt{\frac{n}{\log(p)+t}} \tag{26}$$

Conditioning on the event $\{(\widehat{\beta}-\beta^*) \in \mathbb{B}_\Sigma(r_0) \cap \mathbb{C}(l_1)\}$ *for some* $0 < r_0 \le 1$ *and* $l_1 \asymp \underline{\kappa}^{-1}\sqrt{s_\beta}$*, and selecting the sparsity tuning parameter to satisfy*

$$\lambda_e \gtrsim \max\left[\{(\overline{\sigma} \vee \alpha_q^{1/q}) + B_X l_1 r_0\}\overline{\kappa}\sqrt{\frac{\log(p)+t}{n}}, \underline{\kappa} s_\theta^{-1/2}\left\{\overline{f} r_0^2 + r_0\left(\sqrt{\frac{\log(p)+t}{n}}\right)^{1/(q-1)}\right\}\right], \tag{27}$$

the ℓ_1*-penalized ES estimator* $\widehat{\theta}_\tau$ *defined in* (11) *has the same error bound as in* (24) *with probability at least* $1-3e^{-t}$*, as long as the sample size is sufficiently large as presented in* (25)*.*

4.2 Asymptotic Normality of the Debiased Estimator

Given a pre-specified loading vector $a \in \mathbb{R}^p$, we provide a weak convergence result for the bias-corrected estimator $\widehat{\omega}_\tau$ defined in (18). Specifically, in Theorem 5, we provide a non-asymptotic Bahadur's representation for $\widehat{\omega}_\tau$ with explicit error bounds that depend on both QR and ES estimation errors. Using the Bahadur's representation, we further show that $\widehat{\omega}_\tau$ is asymptotic normal in Theorem 6. Recall from (14) that $\widehat{u}$ is the estimated projection direction and let $\mathbb{B}_1(r) = \{\delta \in \mathbb{R}^p : \|\delta\|_1 \le r\}$. We now present the main results.

Theorem 5 *Assume that Conditions 1–3 hold. For any* $t > 0$*, let* $\tau \asymp \overline{\sigma}\sqrt{n/\{\log(p)+t\}}$*. Conditioning on the event* $\{(\widehat{\beta}-\beta^*) \in \mathbb{B}_\Sigma(r_0) \cap \mathbb{B}_1(r_1)\} \cap \{\alpha(\widehat{\theta}_\tau - \theta^*) \in \mathbb{B}_\Sigma(\delta_0) \cap \mathbb{B}_1(\delta_1)\}$ *with* $0 < r_0, \delta_0 \le 1$ *and* $\delta_1, r_1 > 0$*, and selecting* $\rho \asymp \sqrt{\log(p)/n}$ *and* $\rho' = o(1)$*, we have*

$$\left|\alpha(\widehat{\omega}_\tau - \omega^*) - \frac{1}{n}\sum_{i=1}^n \psi_\tau(\xi_i) X_i^{\mathrm{T}}\widehat{u}\right| \lesssim \|a\|_2\left(\overline{f} r_0^2 + \overline{r}_1\sqrt{\frac{\log p}{n}} + \overline{\sigma}\,\overline{r}_0\sqrt{\frac{\log p}{n}}\right), \tag{28}$$

with probability at least $1-2p^{-1}$*, where* $\bar{r}_0 = r_0 + \delta_0$ *and* $\bar{r}_1 = r_1 + \delta_1$*.*

To establish asymptotic normality, we further impose some conditions on the pre-specified loading vector a. Specifically, we require $\|a\|_1/\|a\|_2 = o(\sqrt{n/\log p})$, which ensures a to be reasonably sparse. Similar conditions can be found in the existing literature on high-dimensional inference. For instance, Bradic and Kolar (2017) consider a unit vector a with $\|a\|_1 \le C$ for some $C > 0$. They also considered growing number of linear combinations, where a is d-sparse, i.e., $\|a\|_0 \le d$, and satisfies $\|a\|_{\Sigma^{-1}} \le K$ for some $K > 0$. Zhu and Bradic (2018) assumed that $\|a\|_0 = o(\sqrt{n}/\log p)$, and Cai and Guo (2017) imposed a special structure on a with $\|a\|_0 \lesssim \|\theta^*\|_0$ and $\max_{j\in\mathrm{supp}(a)}|a_j|/\min_{j\in\mathrm{supp}(a)}|a_j| \le \overline{c}$ for some $\overline{c} \ge 1$.

Theorem 6 *Assume that Conditions 1–3 hold, and that $\mathbb{E}(|\varepsilon_-|^3|X) \le \alpha_3$ almost surely for some $\alpha_3 > 0$. Assume that $a \in \mathbb{R}^p$ satisfies $\|a\|_1/\|a\|_2 = o(\sqrt{n/\log p})$ and $\|\Sigma^{-1}a\|_1 \le C_a\|a\|_2$ for some $C_a > 0$. Let the regularization parameter λ_q and λ_e satisfy $\lambda_q \asymp \sqrt{\log(p)/n}$ and $\lambda_e \asymp s_\beta \log(p)/n + \sqrt{\log(p)/n}$. Moreover, let the robustification parameters τ satisfy $\tau \asymp \overline{\sigma}\sqrt{n/\log p}$. Under the scaling condition $\max(s_\beta, s_\theta)\log p = o(\sqrt{n})$, we have*

$$\alpha\sqrt{n}(\widehat{\omega}_\tau - \omega^*)/s(\widehat{u}) \xrightarrow{\mathrm{d}} \mathcal{N}(0,1) \tag{29}$$

as $n, p \to \infty$, where $s(\widehat{u}) = \widehat{u}^{\mathrm{T}}\widetilde{\Lambda}\widehat{u}$ with $\widetilde{\Lambda} = n^{-1}\sum_{i=1}^n \mathbb{E}(\xi_i^2|X_i)X_iX_i^{\mathrm{T}}$.

The asymptotic property established above provides an explicit method for testing linear hypothesis $H_0 : a^{\mathrm{T}}\theta^* = c_0$ versus $H_1 : a^{\mathrm{T}}\theta^* \ne c_0$, where $c_0 \in \mathbb{R}$ is a predetermined constant.

Remark 7 *It is worth noting that a bounded third (conditional) moment condition on the negative part of QR errors $\varepsilon_{-,i}$ implies that the ES residual ξ_i also has a bounded third (conditional) moment. Recall that $\xi_i = \varepsilon_{-,i} - \mathbb{E}(\varepsilon_{-,i}|X_i)$ with $\varepsilon_{-,i} \le 0$. Therefore, we have $|\xi_i|^3 \le \max\{|\varepsilon_{-,i}|^3, |\mathbb{E}(\varepsilon_{-,i}|X_i)|^3\} \le \max\{|\varepsilon_{-,i}|^3, \mathbb{E}(|\varepsilon_{-,i}|^3|X_i)\}$, which implies that $\mathbb{E}(|\xi_i|^3|X_i) \le \mathbb{E}(|\varepsilon_{-,i}|^3|X_i) \le \alpha_3$ if $\mathbb{E}(|\varepsilon_{-,i}|^3|X_i) \le \alpha_3$.*

For statistical inference, a more general moment condition is Lyapunov's condition, which states that $\mathbb{E}(|\varepsilon_{-,i}|^{2+\delta}|X_i) \le \alpha_{2+\delta} < \infty$ for some $\delta > 0$. This condition is applicable in our case, and the asymptotic behavior can be shown to be the same with a slight modification of the proof. Here, we use the bounded third moment $\mathbb{E}(|\varepsilon_{-,i}|^3|X_i) \le \alpha_3$ for simplicity.

It is also important to note that the bounded third moment condition is not particularly restrictive. Consider a general high-dimensional location-scale model of the form $Y = X^{\mathrm{T}}\beta^* + \sigma(X)e$, where $\sigma(X)$ is a bounded function and the noise e is independent of X with $\mathbb{E}(e) = 0$. Quantile regression only requires the density to satisfy $f_e(0) > 0$. However, ES regression, which essentially operates as a least squares regression, requires $e_- = e \wedge 0$ to be light-tailed in the sense of being sub-Gaussian or sub-exponential. In particular, the condition in Zhang et al. (2023), they require e_- to satisfy the Bernstein moment condition, which implies that e_- is sub-exponential. In comparison, our third moment condition under the robust regression setting is relatively weaker, as it accommodates heavy-tailed noise distributions.

Remark 8 *In addition to $\|a\|_1/\|a\|_2 = o(\sqrt{n/\log p})$, we require $\|\Sigma^{-1}a\|_1/\|a\|_2$ to be bounded. This condition ensures the constraint set $\mathcal{U}$ defined by (15)–(17) contains at least one nonzero solution with high probability (see Lemma 20). Specifically, this condition states that $\|\sum_{j\in\mathrm{supp}(a)} a_j\Sigma_j^{-1}\|_1$ is bounded by $\|a\|_2$, where Σ_j^{-1} is the j-th column of Σ^{-1}. Intuitively, this condition assumes Σ^{-1} to have "weakly sparse" columns under ℓ_1-norm over $\mathrm{supp}(a)$. For example, when $a = e_j$, $\|\Sigma^{-1}a\|_1/\|a\|_2 = \|\Sigma_j^{-1}\|_1 \le C_a$, indicating the j-th column of Σ^{-1} is weakly sparse. When it comes to the inverse covariance matrix estimation, a more commonly used condition is $\|\Sigma^{-1}\|_1 \le M$ for some $M > 0$ (Cai et al., 2016; Bradic and Kolar, 2017), which states that all columns of the precision matrix Σ^{-1} are bounded in ℓ_1-norm.*

To better accommodate heavy-tailed random noise, recall from (19) that we use a truncated estimator $\widehat{s}^2_\gamma(\widehat{u})$ as an alternative to $s^2(\widehat{u})$ for estimating the asymptotic variance.

Then, under H_0, it can be shown that the test statistic $T_{n,\tau,\gamma}(a) = \alpha\sqrt{n}(\widehat{w}_\tau - c_0)/\widehat{s}_\gamma(\widehat{u})$ is asymptotically normal, and a $(1-c)\%$ confidence interval can be constructed as in (20). Specifically, based on the results in Theorem 6, it suffices to show that the truncated variance estimator $\widehat{s}^2_\gamma(\widehat{u})$ is a consistent estimator of $s^2(\widehat{u})$. Proposition 9 provides a non-asymptotic bound for $\widehat{\Lambda}_\gamma$ under max norm, where $\widehat{\Lambda}_\gamma$ is defined in (19). Based on the results in Proposition 9, we establish the consistency of $\widehat{s}^2_\gamma(\widehat{u})$ in Theorem 10.

Proposition 9 *Assume the same conditions as in Theorem 5. Suppose* $\mathbb{E}(|\varepsilon_-|^3|X) \le \alpha_3$ *for some* $\alpha_3 > 0$ *almost surely. Let* $\Lambda = \mathbb{E}(\xi_i^2 X_i X_i^{\mathrm{T}})$*. Recall that* $\widehat{\Lambda}_\gamma = n^{-1}\sum_{i=1}^n \psi_\gamma^2(\widehat{\xi_i}) X_i X_i^{\mathrm{T}}$*. Conditioned on the event that* $\{(\widehat{\beta}-\beta^*) \in \mathbb{B}_\Sigma(r_0) \cap \mathbb{B}_1(r_1)\} \cap \{\alpha(\widehat{\theta}_\tau - \theta^*) \in \mathbb{B}_\Sigma(\delta_0) \cap \mathbb{B}_1(\delta_1)\}$ *with* $0 < r_0, \delta_0 \le 1$ *and* $\delta_1, r_1 > 0$*,*

$$\big\|\widehat{\Lambda}_\gamma - \Lambda\big\|_{\max} \lesssim \overline{\sigma}\bar{r}_0 + \gamma\bar{r}_1\sqrt{\frac{\log(p)+t}{n}} + \gamma^2\frac{\log(p)+t}{n} + \frac{\alpha_3}{\gamma} \tag{30}$$

with probability at least $1 - 2e^{-t}$*, where* $\bar{r}_0 = r_0 + \delta_0$ *and* $\bar{r}_1 = r_1 + \delta_1$*.*

Theorem 10 *Assume Conditions 1–3 hold. Suppose that* $\mathbb{E}(|\varepsilon_-|^3|X) \le \alpha_3$ *almost surely for some* $\alpha_3 > 0$*. Select the robustification parameter* γ *such that* $\gamma \asymp (n/\log p)^{1/3}$*. Under the scaling condition* $\max(s_\beta, s_\theta)\log p = o(\sqrt{n})$*, we have*

$$|\widehat{s}^2_\gamma(\widehat{u}) - s^2(\widehat{u})| \xrightarrow{\mathbb{P}} 0. \tag{31}$$

Theorem 10 guarantees the convergence of $\widehat{s}^2_\gamma(\widehat{u})$ to $s^2(\widehat{u})$, thereby establishing the validity of the proposed test statistic and confidence interval for ω^* in (20).

5. Numerical Studies

In Section 5.1, we perform numerical studies to assess the performance of the proposed high-dimensional robust ES regression under a linear heteroscedastic model. We then examine the validity of the proposed inference procedure in Section 5.2.

Recall from (10) and (11) that the proposed two-step method involves fitting an ℓ_1-penalized QR at the first step, and subsequently fitting an ℓ_1-penalized adaptive Huber regression with a pseudo-response as defined in (7) at the second stage. For computational efficiency, we compute the ℓ_1-penalized smoothed QR estimator that is shown to be first-order equivalent to that of the ℓ_1-penalized QR estimator (Tan et al., 2022), using the `R` package `conquer` with the default smoothing parameter (Man et al., 2024). The sparsity tuning parameter λ_q is selected using 5-fold cross-validation.

Given an estimator of the ℓ_1-penalized QR, $\widehat{\beta}$, we obtain the ℓ_1-penalized robust ES regression estimator by solving (11) using the `R` package `adaHuber` (Sun et al., 2020). To choose the robustification parameter τ, we apply the data-driven mechanism developed by Wang et al. (2021). Given the pseudo-response variables $\{\widehat{Z}_i\}_{i=1}^n$, we use a cross-validated lasso estimator as an initialization, denoted by $\widehat{\theta}^{(0)}$. At iteration $t = 1, 2, \ldots$, we update the tuning parameter $\tau^{(t)}$ by solving

$$\frac{1}{n - \widehat{s}^{(t-1)}}\sum_{i=1}^n \frac{\min\{(\widehat{Z}_i - \alpha X_i^{\mathrm{T}}\widehat{\theta}^{(t-1)})^2, \tau^2\}}{\tau^2} = \frac{\log(p)+\log(n)}{n}, \tag{32}$$

where $\widehat{s}^{(t-1)} = \|\widehat{\theta}^{(t-1)}\|_0$. Next, compute the updated estimator $\widehat{\theta}^{(t)}$ by solving

$$\min_{\theta} \left\{ \frac{1}{n} \sum_{i=1}^{n} \ell_{\tau^{(t)}}(\widehat{Z}_i - \alpha X_i^{\mathrm{T}} \theta) + \lambda_e \|\theta\|_1 \right\}, \tag{33}$$

where λ_e is chosen by five-fold cross-validation. Repeat the above two steps until convergence, or until the maximum number of iterations is reached.

5.1 Estimation

For the data-generating mechanism, we consider the following linear heteroscedastic model

$$Y_i = \langle X_i, \nu^* \rangle + \langle X_i, \eta^* \rangle \cdot e_i, \ \ i = 1, \ldots, n, \tag{34}$$

where $\nu^* = (2_s, 0_{p-s})^{\mathrm{T}}$, $\eta^* = (0.25_{\lceil s/2 \rceil}, 0_{p - \lceil s/2 \rceil})^{\mathrm{T}}$, e_i is independent of X_i and has a mean of zero. Provided that $\langle X_i, \eta^* \rangle \geq 0$ almost surely, it can be shown that the true quantile and ES regression coefficients are sparse with cardinality s and take the form

$$\beta^* = \nu^* + \eta^* \cdot Q_\alpha(e) \quad \text{and} \quad \theta^* = \nu^* + \eta^* \cdot E_\alpha(e), \tag{35}$$

where $Q_\alpha(e)$ and $E_\alpha(e)$ are the α-th quantile and the α-th ES of e, respectively. Throughout the numerical studies, we generate three types of random noise e from (i) the standard normal distribution, (ii) the $t_{2.5}$ distribution, and (iii) the $t_{3.1}$ distribution. Moreover, we consider four types of covariates: (C1) $X_{ij} = |G_{ij}|$ where $G_{ij} \sim \mathcal{N}(0,1)$; (C2) $X_{ij} \sim \mathrm{Unif}(0,2)$; (C3) $X_i = |G_i|$ with $G_i \sim \mathcal{N}_p(0, \Sigma = (0.5^{|j-k|})_{1 \leq j,k \leq p})$; and (C4) $X_i = |G_i|$ with $G_i \sim \mathcal{N}_p(0, \Sigma = (0.8^{|j-k|})_{1 \leq j,k \leq p})$. Note that all covariates are positive to ensure $\langle X_i, \eta \rangle \geq 0$, so that (35) holds.

Remark 11 *In the data-generating process described above, we used folded normal random variables that are not bounded with probability one. As previously mentioned, the boundedness condition for covariates can be relaxed to a sub-Gaussian assumption, where the covariate vector X satisfies $\|X\|_\infty \lesssim \sqrt{\log p}$ with high probability. The boundedness condition is primarily introduced for technical simplicity in the proofs. In practice, simulation results indicate that employing a light-tailed covariate distribution is also effective.*

We compare the proposed ℓ_1-penalized robust ES regression to its non-robust counterpart (Zhang et al., 2023). As benchmarks, we also compute the oracle estimator for robust and non-robust ES regression by calculating the pseudo-response $Z_1(\beta^*), \ldots, Z_n(\beta^*)$ evaluated under the true QR coefficients, and regressing the pseudo-response onto a subset of true active s-dimensional covariates. We note that computing the oracle estimators is not feasible in practice, since they require the support of θ^* to be known a priori.

To assess the performance across the different methods, for an estimator $\widehat{\theta}$, we report the mean and standard error of the relative estimation error on the true support of θ^*, that is, $\|\widehat{\theta}_{\mathcal{S}} - \theta^*_{\mathcal{S}}\|_2 / \|\theta^*\|_2$, where $\mathcal{S} = \{j : \theta^*_j \neq 0\}$, and the relative estimation error of false positives, i.e., $\|\widehat{\theta}_{\mathcal{S}^c}\|_2 / \|\theta^*\|_2$. Results averaged over 500 replications for $p = 200$ and $s = 4$, across various levels of $\alpha = \{0.05, 0.1, 0.2, 0.3, 0.5\}$, are reported in Tables 1 and 2. The results for $t_{3.1}$-distributed errors are reported in Tables 7 and 8 in Appendix D.

Under the Gaussian random noise, we see from Table 1 that the relative estimation error of the proposed robust ES method on $\mathcal{S}$ is similar to its non-robust counterpart across different quantile levels and different types of covariates. On the other hand, under the heavy-tailed $t_{2.5}$ random noise, the ℓ_1-penalized robust ES estimator has a much smaller relative estimation error than that of the ℓ_1-penalized ES estimator. In addition, we see from Table 2 that the two methods have similar false positive error under the Gaussian noise, and that the proposed robust method has a lower false positive error than its non-robust counterpart under $t_{2.5}$ noise. A similar pattern is observed under the $t_{3.1}$ error distribution. Our results indicate that utilizing the Huber loss in the second step gains robustness against heavy-tailed random noise without sacrificing statistical accuracy under the setting with Gaussian random noise.

5.2 Statistical Inference

We now evaluate the performance of the proposed debiased estimator for testing the linear hypothesis $H_0 : a^{\mathrm{T}}\theta^* = 0$ versus $H_1 : a^{\mathrm{T}}\theta^* \neq 0$, where $a \in \mathbb{R}^p$ is a pre-specified vector. Specifically, we consider these four different choices of a: (i) $a_1 = (1, 0_{p-1})^{\mathrm{T}}$, (ii) $a_2 = (1_{s-1}, 0_{p-s+1})^{\mathrm{T}}$, (iii) $a_3 = (1, -1, 0_{p-2})^{\mathrm{T}}$ and (iv) $a_4 = (1_{\lceil s/2 \rceil}, -0.25_{\lfloor s/2 \rfloor}, 1, 0_{p-s-1})^{\mathrm{T}}$. We use the same data generation mechanism as detailed in (34) with:

(1) $X_i = |G_i|$ where $G_i \sim \mathcal{N}_p(0, \Sigma = (0.5^{|j-k|})_{1\le j,k\le p})$ under $\mathcal{N}(0, 1)$ random noise;

(2) $X_i = |G_i|$ where $G_i \sim \mathcal{N}_p(0, \Sigma = (0.5^{|j-k|})_{1\le j,k\le p})$ under $t_{2.5}$ random noise; and

(3) $X_i \sim \mathrm{Unif}(0, 2)$ under $\mathcal{N}(0, 1)$ and $t_{3.1}$ random noise.

Results for more types of error distributions and covariates are reported in Tables 9–12 in Appendix D.

Remark 12 *The $t_{2.5}$ distribution used in our simulations does not satisfy the theoretical moment assumption that $\mathbb{E}|e|^3 < \infty$. On one hand, we employ this more challenging setting as a stress test for all methods. On the other hand, as discussed in Remark 7, the bounded third moment condition is primarily imposed to simplify the technical proofs. A bounded $2 + \delta$ moment can produce similar asymptotic and non-asymptotic results. Therefore, the choice of the $t_{2.5}$ distribution serves as a reasonable compromise. Our simulation results further illustrate that the proposed method remains effective even when the error distribution exhibits heavier tails.*

Recall from Section 5 that λ_q and λ_e are selected using 5-fold cross-validation. Let $\widehat{\lambda}_e$ be the chosen tuning parameter. To construct the debiased estimator, it remains to obtain a projection direction $\widehat{u}$ by solving the quadratic programming optimization problem (14)–(17). For this purpose, we employ the `quadprog` package in `R`. The above optimization problem involves two tuning parameters ρ and ρ'. We set $\rho' = 0.9$ and let $\rho = \widehat{\lambda}_e$ be the initial tuning parameter. Let $r > 1$ be a constant. We further tune ρ through the following steps:

(1) Multiply ρ by a factor of either r or $1/r$;

$\mathcal{N}(0,1)$ random noise; $n = \lceil 25s/\alpha \rceil$					
Covariates	Methods	$\alpha = 0.05$	$\alpha = 0.1$	$\alpha = 0.2$	$\alpha = 0.3$
C1	ℓ_1-penalized Robust ES	6.950 (0.061)	6.789 (0.061)	7.265 (0.063)	7.605 (0.067)
	ℓ_1-penalized ES	4.684 (0.072)	4.660 (0.069)	5.303 (0.074)	5.675 (0.073)
	Oracle Robust ES	2.393 (0.038)	2.584 (0.040)	2.821 (0.048)	3.186 (0.058)
	Oracle ES	2.724 (0.051)	2.774 (0.045)	3.099 (0.057)	3.310 (0.061)
C2	ℓ_1-penalized Robust ES	10.099 (0.059)	9.809 (0.065)	10.184 (0.074)	10.525 (0.073)
	ℓ_1-penalized ES	5.671 (0.065)	5.679 (0.073)	6.411 (0.085)	6.771 (0.072)
	Oracle Robust ES	2.656 (0.042)	2.742 (0.041)	3.169 (0.052)	3.688 (0.062)
	Oracle ES	2.916 (0.047)	2.857 (0.044)	3.384 (0.056)	3.742 (0.063)
C3	ℓ_1-penalized Robust ES	5.565 (0.057)	5.484 (0.063)	6.094 (0.068)	6.331 (0.066)
	ℓ_1-penalized ES	4.150 (0.067)	4.249 (0.068)	4.884 (0.077)	5.045 (0.071)
	Oracle Robust ES	2.534 (0.042)	2.723 (0.042)	3.028 (0.050)	3.553 (0.059)
	Oracle ES	2.973 (0.057)	2.952 (0.049)	3.407 (0.063)	3.755 (0.065)
C4	ℓ_1-penalized Robust ES	4.971 (0.072)	4.960 (0.076)	5.446 (0.085)	5.589 (0.073)
	ℓ_1-penalized ES	4.271 (0.075)	4.491 (0.077)	4.922 (0.089)	4.967 (0.075)
	Oracle Robust ES	3.168 (0.054)	3.508 (0.054)	3.958 (0.069)	4.680 (0.080)
	Oracle ES	3.765 (0.075)	3.947 (0.071)	4.483 (0.086)	4.944 (0.087)
$t_{2.5}$ random noise; $n = \lceil 50s/\alpha \rceil$					
Covariates	Methods	$\alpha = 0.05$	$\alpha = 0.1$	$\alpha = 0.2$	$\alpha = 0.3$
C1	ℓ_1-penalized Robust ES	15.622 (0.301)	14.885 (0.280)	14.908 (0.289)	14.679 (0.278)
	ℓ_1-penalized ES	33.151 (0.825)	33.705 (0.806)	31.273 (0.798)	29.981 (0.779)
	Oracle Robust ES	11.021 (0.162)	8.333 (0.126)	6.660 (0.107)	5.948 (0.097)
	Oracle ES	17.289 (0.440)	16.085 (0.480)	14.689 (0.442)	13.408 (0.434)
C2	ℓ_1-penalized Robust ES	18.706 (0.265)	17.726 (0.263)	17.538 (0.243)	16.979 (0.236)
	ℓ_1-penalized ES	36.914 (0.662)	36.688 (0.696)	36.695 (0.685)	35.376 (0.706)
	Oracle Robust ES	11.794 (0.169)	9.125 (0.131)	7.523 (0.119)	6.769 (0.107)
	Oracle ES	19.072 (0.308)	17.119 (0.318)	15.761 (0.328)	15.232 (0.307)
C3	ℓ_1-penalized Robust ES	14.067 (0.319)	13.413 (0.322)	13.106 (0.260)	13.144 (0.290)
	ℓ_1-penalized ES	27.803 (0.748)	27.230 (0.709)	25.737 (0.706)	23.790 (0.721)
	Oracle Robust ES	11.599 (0.161)	9.034 (0.128)	7.078 (0.107)	6.566 (0.110)
	Oracle ES	18.009 (0.549)	16.550 (0.493)	14.495 (0.401)	14.349 (0.469)
C4	ℓ_1-penalized Robust ES	15.292 (0.325)	14.134 (0.329)	13.549 (0.264)	13.014 (0.263)
	ℓ_1-penalized ES	25.989 (0.720)	25.314 (0.678)	23.405 (0.670)	21.939 (0.638)
	Oracle Robust ES	13.999 (0.198)	10.706 (0.160)	8.983 (0.145)	7.926 (0.132)
	Oracle ES	22.513 (0.596)	21.010 (0.533)	19.450 (0.640)	17.848 (0.583)

Table 1: The mean (and standard error) of the relative estimation error on the support $\mathcal{S}$ of θ^*: $\|\widehat{\theta}_{\mathcal{S}} - \theta^*_{\mathcal{S}}\|_2/\|\theta^*\|_2$, averaged over 500 replications with $p = 200$, $s = 4$, and $\alpha = \{0.05, 0.1, 0.2, 0.3\}$ under standard normal random noise and $t_{2.5}$ random noise across four different types of covariates. All results are scaled by a factor of 100.

$\mathcal{N}(0,1)$ random noise; $n = \lceil 25s/\alpha \rceil$					
Covariates	Methods	$\alpha = 0.05$	$\alpha = 0.1$	$\alpha = 0.2$	$\alpha = 0.3$
C1	ℓ_1-penalized Robust ES	0.215 (0.019)	0.335 (0.026)	0.539 (0.028)	0.545 (0.028)
	ℓ_1-penalized ES	1.998 (0.046)	2.254 (0.050)	2.421 (0.052)	2.461 (0.060)
C2	ℓ_1-penalized Robust ES	0.028 (0.006)	0.045 (0.008)	0.150 (0.014)	0.235 (0.019)
	ℓ_1-penalized ES	2.506 (0.058)	2.585 (0.060)	2.852 (0.068)	3.131 (0.076)
C3	ℓ_1-penalized Robust ES	0.237 (0.021)	0.328 (0.025)	0.510 (0.031)	0.508 (0.027)
	ℓ_1-penalized ES	1.748 (0.046)	1.967 (0.053)	2.076 (0.054)	2.231 (0.060)
C4	ℓ_1-penalized Robust ES	0.253 (0.023)	0.303 (0.025)	0.373 (0.028)	0.336 (0.023)
	ℓ_1-penalized ES	1.492 (0.048)	1.601 (0.049)	1.576 (0.053)	1.740 (0.066)
$t_{2.5}$ random noise; $n = \lceil 50s/\alpha \rceil$					
Covariates	Methods	$\alpha = 0.05$	$\alpha = 0.1$	$\alpha = 0.2$	$\alpha = 0.3$
C1	ℓ_1-penalized Robust ES	7.263 (0.246)	6.847 (0.234)	6.464 (0.203)	5.867 (0.189)
	ℓ_1-penalized ES	11.842 (0.265)	11.499 (0.251)	11.419 (0.255)	11.416 (0.273)
C2	ℓ_1-penalized Robust ES	9.488 (0.301)	8.608 (0.274)	7.800 (0.242)	7.116 (0.225)
	ℓ_1-penalized ES	9.591 (0.247)	8.515 (0.214)	8.048 (0.188)	7.839 (0.174)
C3	ℓ_1-penalized Robust ES	6.527 (0.264)	6.145 (0.232)	5.621 (0.221)	5.512 (0.195)
	ℓ_1-penalized ES	10.147 (0.251)	10.256 (0.268)	9.612 (0.241)	9.449 (0.225)
C4	ℓ_1-penalized Robust ES	5.450 (0.247)	4.822 (0.223)	4.471 (0.194)	4.467 (0.187)
	ℓ_1-penalized ES	7.905 (0.255)	7.787 (0.230)	7.023 (0.227)	7.013 (0.216)

Table 2: The mean (and standard error) of the relative estimation error of the false positives of $\widehat{\theta}$, i.e., $\|\widehat{\theta}_{\mathcal{S}^c}\|_2/\|\theta^*\|_2$, averaged over 500 replications with $p = 200$, $s = 4$, and $\alpha = \{0.05, 0.1, 0.2, 0.3, 0.5\}$ under standard normal and $t_{2.5}$ random noise. All results are scaled by a factor of 100.

(2) Choose the factor that minimizes the mean squared error (MSE), calculated as $\widehat{s}^2_\gamma(\widehat{u}) + b^2_\tau(\widehat{u})$, where $b_\tau(u) = \|a - (1/n)\sum_{i=1}^n \psi'_\tau(\widehat{\xi}_i)X_iX_i^{\mathrm{T}}u\|_\infty$;

(3) Adjust ρ by multiplying it by the selected factor from step (2);

(4) Repeat steps (2) and (3) until the MSE either stops decreasing or decreases by an amount smaller than a predefined tolerance level.

Next, we calculate the debiased estimator $\widehat{\omega}_\tau$ according to (18). To obtain a robust estimator of the variance, we select $\gamma = (\widehat{\alpha}_3^{1/3}/\widehat{\sigma})\tau$, where $\widehat{\alpha}_3$ and $\widehat{\sigma}$ are the sample 3rd moment and standard deviation of the negative part of quantile residuals, respectively. We then compute the variance as in (19), and the 95% confidence interval can be constructed as in (20).

We compare our proposed ℓ_1-penalized robust ES regression estimator to its non-robust counterpart (Zhang et al., 2023). Let $\widehat{\theta}^{\mathrm{ls}}$ be a non-robust ℓ_1-penalized ES regression estimator. We compute the debiased estimator for $\omega^* = a^{\mathrm{T}}\theta^*$ as

$$\widehat{\omega}^{\mathrm{ls}} = a^{\mathrm{T}}\widehat{\theta}^{\mathrm{ls}} + \frac{1}{\alpha n}\sum_{i=1}^n \widehat{\xi}_i^{\mathrm{ls}}\widehat{u}^{\mathrm{T}}X_i, \tag{36}$$

where $\widehat{\xi}_i^{\mathrm{ls}} = \widehat{Z}_i - \alpha X_i^{\mathrm{T}}\widehat{\theta}^{\mathrm{ls}}$. We obtain $\widehat{u}$ by selecting $\rho' = 0.9$ and tuning ρ using the same method as described above, with $b(\widehat{u}) = \|a - \widehat{\Sigma}\widehat{u}\|_\infty$ and $\{\widehat{s}^{\mathrm{ls}}(\widehat{u})\}^2 = n^{-1}\sum_{i=1}^n(\widehat{\xi}_i^{\mathrm{ls}})^2(X_i^{\mathrm{T}}\widehat{u})^2$.

Then, the corresponding 95% CI is constructed as

$$\left[\widehat{\omega}^{\mathrm{ls}} - z_{1-\alpha/2} \cdot \frac{\widehat{s}^{\mathrm{ls}}(\widehat{u})}{\alpha\sqrt{n}}, \widehat{\omega}^{\mathrm{ls}} + z_{1-\alpha/2} \cdot \frac{\widehat{s}^{\mathrm{ls}}(\widehat{u})}{\alpha\sqrt{n}}\right]. \tag{37}$$

To assess the performance of the two methods, we compute the coverage rate and the mean width of the 95% confidence interval for both the robust and non-robust debiased estimators. Results under standard normal and t-distributed random noise for two types of covariates, computed based on 500 independent replications, are reported in Tables 3, 4 and 5. The coverage rates for both robust and non-robust debiased estimators are close to the desired 95% confidence level when the error distribution is standard normal. In contrast, under the $t_{2.5}$ random noise, we see from Table 4 that the robust debiased estimator can maintain 0.95 coverage while its non-robust counterpart has a relatively unstable coverage rate and suffers from under coverage in many cases, especially when a has more nonzero entries. Moreover, the robust debiased estimator has a lower estimation error and the mean width of the 95% confidence interval for our proposed robust approach is shorter than that of its non-robust counterpart. When the error distribution exhibits lighter tails, Table 5 shows that the differences become less significant, as anticipated. Nevertheless, our method continues to demonstrate a notable advantage in terms of both coverage and confidence interval length, particularly when a is denser. In summary, our results suggest that our proposed method gains robustness against heavy-tailed errors without loss of efficiency, and the debiased estimator for linear projections exhibits stable asymptotic behavior.

$e_i \sim \mathcal{N}(0,1)$							
a	α	Estimation Error		Coverage		Estimated Width	
		Robust	Non-Robust	Robust	Non-Robust	Robust	Non-Robust
a_1	0.05	3.07 (0.10)	3.06 (0.10)	96.0	95.2	15.00 (0.18)	15.15 (0.19)
	0.1	3.58 (0.13)	3.59 (0.13)	93.4	93.6	17.15 (0.20)	17.41 (0.22)
	0.2	3.24 (0.12)	3.25 (0.12)	94.2	94.2	16.02 (0.16)	16.07 (0.18)
a_2	0.05	4.98 (0.17)	4.68 (0.16)	95.6	95.4	22.69 (0.27)	23.04 (0.30)
	0.1	5.65 (0.21)	5.48 (0.20)	95.4	95.6	26.68 (0.36)	27.17 (0.38)
	0.2	5.44 (0.19)	5.25 (0.19)	94.8	94.8	24.65 (0.26)	24.78 (0.30)
a_3	0.05	4.40 (0.15)	4.40 (0.15)	93.8	94.0	20.72 (0.20)	20.87 (0.22)
	0.1	4.91 (0.17)	5.11 (0.17)	94.6	95.2	23.07 (0.21)	24.24 (0.23)
	0.2	4.53 (0.16)	4.56 (0.17)	93.6	93.4	21.99 (0.18)	22.58 (0.20)
a_4	0.05	4.64 (0.17)	4.72 (0.17)	95.2	93.2	23.29 (0.28)	24.10 (0.30)
	0.1	5.50 (0.19)	5.66 (0.19)	94.6	94.2	27.32 (0.36)	28.26 (0.38)
	0.2	5.42 (0.19)	5.43 (0.19)	94.6	94.0	27.66 (0.27)	27.47 (0.28)

Table 3: The mean estimation error $|\widehat{\omega} - \omega^*|$ (and standard error), coverage rate, and the mean width of 95% confidence intervals (and standard error), averaged over 500 replications, with $p = 200$, $s = 4$ and $n = \lceil 50s/\alpha \rceil$ under the standard normal random noise. The covariates are generated as $X_i = |G_i|$ where $G_i \sim \mathcal{N}_p(0, \Sigma = (0.5^{|j-k|})_{1 \le j,k \le p})$. All results are scaled by a factor of 100.

$e_i \sim t_{2.5}$							
a	α	Estimation Error		Coverage(%)		Estimated Width	
		Robust	Non-Robust	Robust	Non-Robust	Robust	Non-Robust
a_1	0.05	14.41 (0.54)	19.46 (0.74)	90.0	85.2	58.38 (0.54)	64.33 (0.88)
	0.1	10.72 (0.37)	13.91 (0.54)	94.2	92.6	53.49 (0.64)	59.52 (0.72)
	0.2	10.83 (0.42)	15.32 (0.73)	94.2	89.0	46.04 (0.61)	51.06 (0.88)
a_2	0.05	20.94 (0.77)	29.19 (1.20)	92.6	87.0	98.61 (1.13)	106.47 (1.76)
	0.1	18.25 (0.67)	25.51 (0.95)	91.4	90.2	71.30 (0.94)	94.55 (1.99)
	0.2	16.19 (0.62)	24.17 (1.16)	93.4	86.8	71.74 (1.01)	79.89 (1.51)
a_3	0.05	18.93 (0.65)	22.58 (0.80)	94.2	93.2	88.09 (0.70)	96.92 (1.13)
	0.1	17.03 (0.58)	19.27 (0.65)	95.6	94.8	81.99 (0.87)	86.06 (1.08)
	0.2	15.40 (0.52)	18.09 (0.66)	93.4	90.8	68.89 (0.88)	76.01 (1.18)
a_4	0.05	21.43 (0.73)	29.48 (1.13)	92.6	82.8	94.97 (0.82)	96.86 (1.09)
	0.1	17.08 (0.57)	23.49 (0.94)	92.0	89.8	82.17 (0.75)	87.28 (1.25)
	0.2	16.39 (0.59)	21.86 (0.94)	93.8	88.0	75.88 (0.90)	80.10 (1.05)

Table 4: The mean estimation error $|\widehat{\omega} - \omega^*|$ (and standard error), coverage rate, and the mean width of 95% confidence intervals (and standard error), averaged over 500 replications, with $p = 200$, $s = 4$ and $n = \lceil 100s/\alpha \rceil$ under the $t_{2.5}$ random noise. The covariates are generated as $X_i = |G_i|$ where $G_i \sim \mathcal{N}_p(0, \Sigma = (0.5^{|j-k|})_{1 \le j,k \le p})$. All results are scaled by a factor of 100.

$e_i \sim t_{3.1}$							
a	α	Estimation Error		Coverage(%)		Estimated Width	
		Robust	Non-Robust	Robust	Non-Robust	Robust	Non-Robust
a_1	0.05	8.20 (0.29)	8.52 (0.30)	95.2	95.2	40.73 (0.23)	40.88 (0.23)
	0.1	6.17 (0.21)	6.26 (0.22)	95.6	96.4	32.36 (0.17)	32.49 (0.16)
	0.2	6.28 (0.22)	6.42 (0.22)	94.8	95.0	30.84 (0.19)	31.14 (0.18)
a_2	0.05	16.46 (0.54)	19.30 (0.65)	93.2	88.6	66.07 (0.49)	66.40 (0.50)
	0.1	12.19 (0.42)	14.00 (0.49)	93.4	90.4	53.75 (0.36)	54.02 (0.36)
	0.2	10.92 (0.39)	12.32 (0.46)	93.6	89.0	48.11 (0.32)	49.00 (0.31)
a_3	0.05	11.16 (0.38)	11.19 (0.38)	95.6	95.8	54.52 (0.27)	54.58 (0.27)
	0.1	8.77 (0.29)	8.78 (0.29)	94.6	94.8	42.74 (0.21)	42.88 (0.21)
	0.2	7.11 (0.26)	7.19 (0.26)	95.0	94.8	37.39 (0.19)	37.85 (0.19)
a_4	0.05	15.79 (0.54)	17.61 (0.61)	92.6	90.2	68.56 (0.52)	68.71 (0.53)
	0.1	11.22 (0.40)	12.00 (0.43)	95.0	94.4	55.53 (0.38)	55.76 (0.38)
	0.2	11.87 (0.43)	12.98 (0.47)	94.0	92.8	53.41 (0.41)	53.86 (0.42)

Table 5: The mean estimation error $|\widehat{\omega} - \omega^*|$ (and standard error), coverage rate, and the mean width of 95% confidence intervals (and standard error), averaged over 500 replications, with $p = 200$, $s = 4$ and $n = \lceil 100s/\alpha \rceil$ under the $t_{3.1}$ random noise. The covariates are generated as $X_i \sim \text{Unif}(0, 2)$. All results are scaled by a factor of 100.

6. Data Application

Iron deficiency and iron-deficiency anemia are significant global health concerns and are commonly encountered in clinical practice. Although the prevalence of iron-deficiency anemia has declined slightly, iron deficiency remains the leading cause of anemia worldwide (Camaschella, 2015). Early detection is crucial, especially in patients with inflammation, infection, or chronic disease, and plays a key role in preventive care, as it can indicate underlying conditions such as gastrointestinal malignancies (Rockey and Cello, 1993). One important measure of iron deficiency is the soluble transferrin receptor (sTRP), a protein that binds transferrin. Elevated sTRP levels serve as a reliable indicator of iron deficiency (Mast et al., 1998).

The scientific goals of this study are to assess: (1) whether there is a disparity in the upper $\alpha = 0.9$ expected shortfall of sTRP levels among four ethnic groups—Asian, Black, Mexican American, and White; and (2) whether being overweight (BMI ≥ 25) influences sTRP levels across these groups. To address these objectives, we analyze data from the National Health and Nutrition Examination Survey (NHANES) for the years 2017 to 2020 (pre-COVID), which includes sTRP measurements for female participants aged 20 to 49 years. As shown in Figure 1, differences in sTRP levels are more noticeable in the upper 90% tail of the distribution.

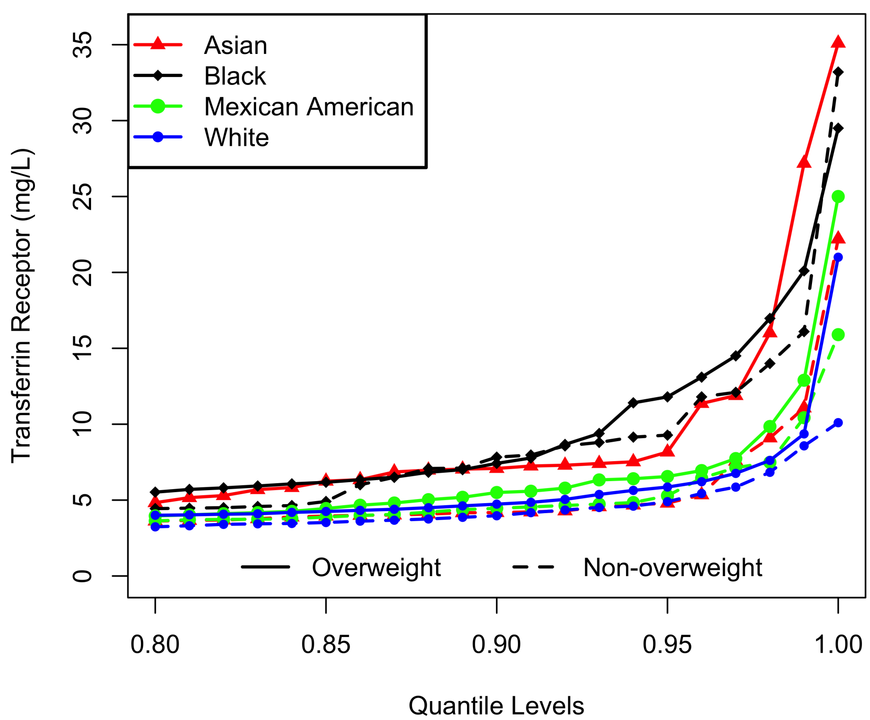

Figure 1: The soluble transferrin receptor levels (mg/L) across quantile levels (ranging from 0.8 to 1) for female participants, categorized by four ethnic groups—Asian, Black, Mexican American, and White—and by weight status

Our analysis adjusts for demographic factors such as age, education level, and diet, as well as health conditions including asthma, arthritis, and cancer, along with relevant interaction terms. To simplify the analysis, we excluded all participants with missing values for the covariates, resulting in a final dataset of $n = 1644$ observations and $p = 135$ variables. More specifically, the model in this study is

$$E_\alpha(\text{sTRP}) = \theta_0 + \theta_1 \cdot \text{race} + \theta_2 \cdot \text{overweight} + \theta_3 \cdot (\text{race}\times\text{overweight}) + \text{other features}.$$

Here, with a slight abuse of notation, E_α denotes the expected shortfall of the upper tail. It is important to note that the variable *race* consists of three dummy variables representing Asian, Mexican American, and Black ethnicities, with White as the reference group. The variable *overweight*, on the other hand, is binary. For the first objective, the loading vector is of the form $a = (1, 0, \ldots, 0)^{\mathrm{T}}$, with a value of 1 for the corresponding ethnic group. For the second objective, the reference group is White non-overweight individuals, and the loading vector is structured as $a = (1, 1, 1, 0, \ldots, 0)^{\mathrm{T}}$, where one nonzero coordinate corresponds to the ethnic group dummy variable, one to the overweight indicator, and another to their interaction term. This setup for the two loading vectors enables natural comparisons between overweight and non-overweight groups within each ethnicity.

To assess the tail behavior of sTRP levels, we first examine its histogram. As shown in Figure 2, a significant number of observations fall outside the 95% range, indicating the presence of more extreme values than expected under a light-tailed distribution. In addition, we construct a Hill plot (see, e.g., Section 4.4 in Resnick (2007)). Figure 3 suggests that the data has finite moments up to approximately order $q = 2.3$, indicating that while the data is likely to have finite mean and variance, higher-order moments may not exist. Furthermore, the sample kurtosis of sTRP levels is $40.70 \gg 3$, which provides additional evidence that the data exhibit heavy-tailedness.

While our proposed method primarily focuses on the lower tail of the conditional distribution, it can easily be adapted to the upper tail. Given a continuous conditional CDF $F_{Y|X}(\cdot)$, note that $Q_\alpha(Y|X) = -Q_{1-\alpha}(-Y|X)$. Therefore, we apply the method at the $(1-\alpha)$ level using negative sTRP values, and then reverse the sign of the resulting estimators and confidence intervals to obtain the regression coefficients of interest. Similar to Section 5, we compare the performance of the robust method with that of the non-robust one. The results are presented in Table 6.

From Table 6 we observe that the proposed method successfully detects disparities between the Black and White groups, as well as between the overweight Asian group and the non-overweight White group. For other groups, while potential disparities exist, the effect sizes are close to zero. Additionally, although overweight status may be associated with higher sTRP levels, the result is not statistically significant. Due to the heavy-tailed nature of the data, the non-robust method fails to detect the disparity between the overweight Asian group and the non-overweight White group. Moreover, the robust approach also produces narrower confidence intervals.

7. Discussion

In this work, we consider the estimation and inference under a high-dimensional joint quantile and expected shortfall regression model, with a focus on the latter. Unlike quantiles,

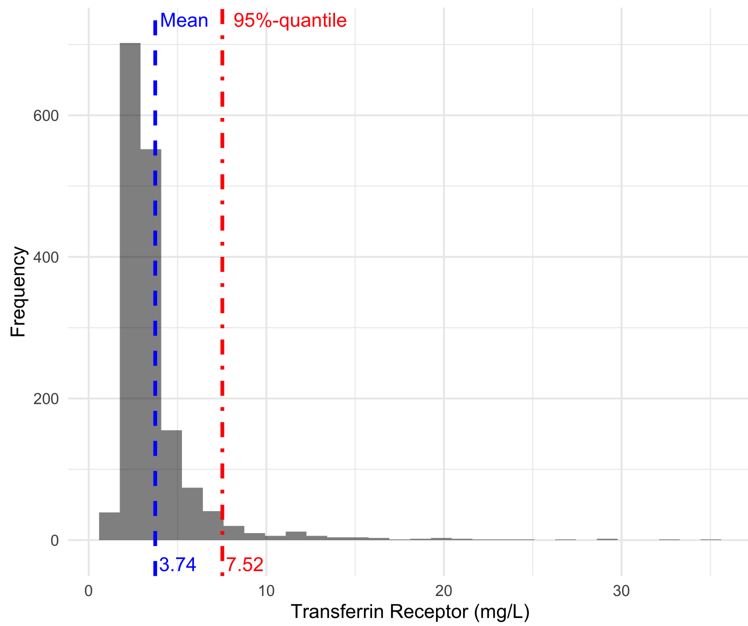

Figure 2: The histogram of soluble transferrin receptor levels (mg/L) for female participants.

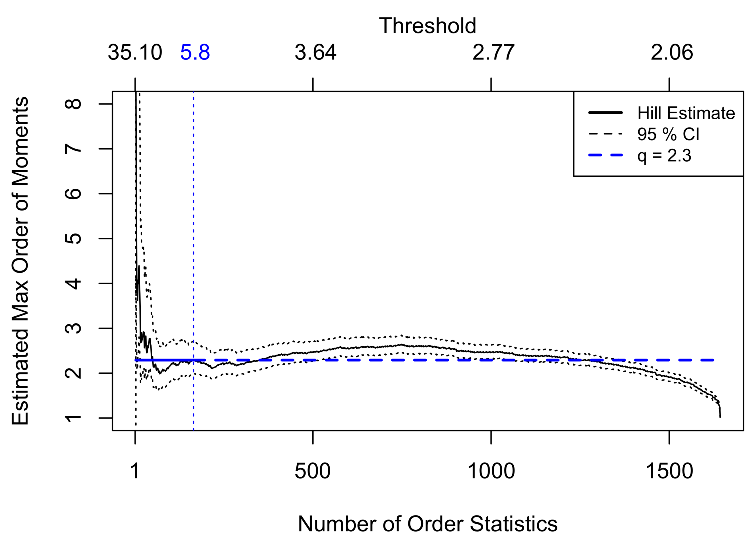

Figure 3: The Hill plot of soluble transferrin receptor levels (mg/L) for female participants. Under a cutoff value of 5.8 for tail values, the estimated maximum order q of finite moments is approximately $q = 2.3$.

	Overweight		Non-Overweight	
Race	Robust	Non-Robust	Robust	Non-Robust
Asian	3.59 (2.94, 4.24)	0.51 (−0.14, 1.16)	0.11 (−0.27, 0.49)	0.36 (−0.30, 1.02)
Black	4.83 (4.05, 5.62)	4.65 (3.29, 6.00)	4.77 (4.02, 5.51)	4.29 (3.07, 5.51)
Mexican	−0.01 (−0.44, 0.43)	−0.11 (−1.06, 0.85)	−0.45 (−1.44, 0.55)	−0.78 (−1.79, 0.22)
White	0.18 (−0.55, 0.91)	0.66 (−0.43, 1.74)	Baseline	

Table 6: The estimated coefficients (and 95% confidence intervals) for the upper 90% Expected Shortfall (ES) across ethnic groups and overweight status, using both robust and non-robust methods. The baseline group is white non-overweight females.

expected shortfall is a measure of the (conditional) average, which makes the least-squares method highly sensitive to heavy-tailed data compared to the check loss minimization used in quantile regression. Therefore, we propose a robust penalized approach to overcome the challenges brought about by high-dimensionality and heavy-tailed data distributions.

To conduct inference, it is well-known that ℓ_1-regularization introduces non-negligible bias, which prevents the estimator from being asymptotically efficient. Therefore, we propose a debiased estimator to alleviate the effect of ℓ_1-penalization bias and construct a test statistic that satisfies asymptotic normality in the ultra-sparse regime "$\max(s_\beta, s_\theta) = o(\sqrt{n}/\log p)$". This debiased method provides a valid way of constructing confidence intervals for a class of linear projections of θ^*, which paves the way for inference on high-dimensional ES treatment effects, provided that the sparsity assumptions hold.

Acknowledgments

The authors do not have competing interests for this work. The research of Tan is in part supported by the NSF grants DMS-2113346 and DMS-2238428. Zhou's research is supported by NSF grant DMS-2401268 and a faculty research grant from the UIC College of Business Administration.

Appendix A. Auxiliary Lemmas

In preparation for the proof, we begin by providing a concise overview and introducing the notations that will be frequently used. For any random variable X, we denote its centered version as $(1-\mathbb{E})X = X - \mathbb{E}X$. Additionally, denote the conditional expectation and variance given X as $\mathbb{E}_X$ and var_X, respectively. Let $\widehat{\mathcal{Q}}(\beta) = n^{-1}\sum_{i=1}^n \rho_\alpha(Y_i - X_i^{\mathrm{T}}\beta)$ be the empirical quantile loss, and define the loss difference $\widehat{\mathcal{D}}(\delta) = \widehat{\mathcal{Q}}(\beta^*+\delta) - \widehat{\mathcal{Q}}(\beta^*)$ and its population counterpart $\mathcal{D}(\delta) = \mathbb{E}\widehat{\mathcal{D}}(\delta)$ for $\delta \in \mathbb{R}^p$. For estimating the ES in step two, let $\widehat{\mathcal{L}}_\tau(\beta,\theta) = (1/n)\sum_{i=1}^n \ell_\tau(Z_i(\beta) - \alpha X_i^{\mathrm{T}}\theta)$, where

$$Z_i(\beta) = (Y_i - X_i^{\mathrm{T}}\beta)\mathbb{1}(Y_i \le X_i^{\mathrm{T}}\beta) + \alpha X_i^{\mathrm{T}}\beta. \tag{38}$$

Write $\widehat{Z}_i = Z_i(\widehat{\beta})$ with $\widehat{\beta} = \widehat{\beta}(\lambda_q)$ denoting the ℓ_1-penalized QR estimator, the two-step Huber-ES estimator is given by $\widehat{\theta}_\tau = \widehat{\theta}_\tau(\lambda_e) \in \mathrm{argmin}_{\theta\in\mathbb{R}^p} \frac{1}{n}\sum_{i=1}^n \ell_\tau(\widehat{Z}_i - \alpha X_i^{\mathrm{T}}\theta) + \alpha\lambda_e\|\theta\|_1$. For the sake of simplicity, we denote $\widehat{\theta} = \widehat{\theta}_\tau$ when there is no ambiguity. Furthermore, we define the quantile and ES residuals as

$$\varepsilon_i = Y_i - X_i^{\mathrm{T}}\beta^* \quad \text{and} \quad \xi_i(\beta,\theta) = Z_i(\beta) - \alpha X_i^{\mathrm{T}}\theta \tag{39}$$

In particular, we write $\xi_i(\beta) = \xi_i(\beta,\theta^*)$ and $\xi_i = \xi_i(\beta^*,\theta^*)$. In addition, let ψ_τ denote the derivative of the Huber loss, that is, $\psi_\tau(t) = \mathrm{sign}(t)\min\{|t|,\tau\}$.

For any $c_0 \ge 1$ and subset $\mathcal{S} \subseteq \{1,\ldots,p\}$, define the ℓ_1-cone

$$\mathbb{C}_b(\mathcal{S}) := \{\delta \in \mathbb{R}^p : \|\delta_{\mathcal{S}^c}\|_1 \le b\|\delta_{\mathcal{S}}\|_1\}. \tag{40}$$

Moreover, for any $l_1 > 0$, define the cone-like set

$$\mathbb{C}(l_1) = \mathbb{C}^p(l_1) = \{\delta \in \mathbb{R}^p : \|\delta\|_1 \le l_1\|\delta\|_\Sigma\}. \tag{41}$$

Under Condition 1 that $\lambda_{\min}(\Sigma) \ge \underline{\kappa}^2 > 0$, we have $\|\delta\|_\Sigma \ge \underline{\kappa}\|\delta\|_2$ and hence $\mathbb{C}_b(\mathcal{S}) \subseteq \mathbb{C}(l_1)$ with $l_1 = (b+1)\sqrt{s}/\underline{\kappa}$ and $s = |\mathcal{S}|$.

With the above preparations, we are ready to present a series of technical lemmas that form the core elements of the main proofs.

Lemma 13 *Assume Condition 1 holds. Given any $r_0, l_1 > 0$,*

$$\sup_{\delta\in\mathbb{C}(l_1)\cap\mathbb{B}_\Sigma(r_0)} \{\mathcal{D}(\delta) - \widehat{\mathcal{D}}(\delta)\} \le 4\bar{\alpha}r_0 l_1\left\{\sqrt{2}\bar{\kappa}\sqrt{\frac{\log(2p)+t}{n}} + B_X\frac{\log(2p)+t}{n}\right\}$$

holds with probability at least $1-e^{-t}$ for any $t>0$, where $\bar{\alpha} = \max(\alpha, 1-\alpha)$.

Lemma 14 *Assume Conditions 1 and 2 hold. For any $t > 0$, it holds with probability at least $1-e^{-t}$ that*

$$\left\|\frac{1}{n}\sum_{i=1}^n\{\mathbb{1}(\varepsilon_i \le 0) - \alpha\}X_i\right\|_\infty \le \sqrt{2\alpha(1-\alpha)}\bar{\kappa}\sqrt{\frac{\log(2p)+t}{n}} + \bar{\alpha}B_X\frac{\log(2p)+t}{n},$$

where $\bar{\alpha} = \max(\alpha, 1-\alpha)$.

Lemma 15 *Assume Conditions 1 and 3 hold. For any $t>0$,*

$$\bigg\|\frac{1}{n}\sum_{i=1}^n(1-\mathbb{E})\{\psi_\tau(\xi_i)X_i\}\bigg\|_\infty \le \sqrt{2}\overline{\sigma}\overline{\kappa}\sqrt{\frac{\log(2p)+t}{n}}+\tau B_X\frac{\log(2p)+t}{n}.$$

with probability at least $1-e^{-t}$.

Lemma 16 *Assume Conditions 1 and 3 hold. For any $r_1>0$,*

$$\sup_{\beta\in\mathbb{B}_1(\beta^*,r_1)}\bigg\|\frac{1}{n}\sum_{i=1}^n(1-\mathbb{E})\{\psi_\tau(\xi_i(\beta))-\psi_\tau(\xi_i)\}X_i\bigg\|_\infty \lesssim B_X r_1\left(B_X\frac{\log p+t}{n}+\overline{\kappa}\sqrt{\frac{\log(p)+t}{n}}\right)$$

holds with probability at least $1-e^{-t}$ for any $t\ge 0$.

Lemma 17 *Assume Conditions 1 and 3 hold, and let $U_i=\Sigma^{-1/2}X_i$. For any $r_0>0$,*

$$\sup_{\beta\in\mathbb{B}_\Sigma(\beta^*,r_0)}\|\mathbb{E}\{\psi_\tau(\xi_i(\beta))U_i\}\|_2\le\frac{1}{2}\kappa_3(\overline{f}+\tau^{-1})r_0^2+\overline{\sigma}\frac{r_0}{\tau}+\overline{\sigma}^2/\tau$$

Moreover, if $\mathbb{E}_{X_i}(|\varepsilon_-|^q)\le\alpha_q$ (almost surely) for some $q\ge 3$, where $\varepsilon_-=\min\{\varepsilon,0\}$, then

$$\sup_{\beta\in\mathbb{B}_\Sigma(\beta^*,r_0)}\|\mathbb{E}\{\psi_\tau(\xi_i(\beta))U_i\}\|_2\le\frac{1}{2}\kappa_3(\overline{f}+\tau^{-1})r_0^2+\alpha_q^{1/q}\frac{r_0}{\tau}+\alpha_q/\tau^{q-1}.$$

Lemma 18 *Assume Conditions 1, 2 and 3 hold. Let $\Delta=\theta-\theta^*$, $\Delta'=\beta-\beta^*$ and $t>0$. For any given $r,r_0>0$, let the robustification parameter τ be such that $\tau\ge 2.5\max\{B_X(lr\vee l_1r_0),\sqrt{50}\overline{\sigma}\}$. Then with probability at least $1-e^{-t}$,*

$$\inf_{\substack{\alpha\Delta\in\mathbb{B}_\Sigma(r)\cap\mathbb{C}(l)\\ \Delta'\in\mathbb{B}_\Sigma(r_0)\cap\mathbb{C}(l_1)}}\frac{\langle\nabla_\theta\widehat{\mathcal{L}}_\tau(\beta,\theta)-\nabla_\theta\widehat{\mathcal{L}}_\tau(\beta,\theta^*),\theta-\theta^*\rangle}{\|\theta-\theta^*\|_\Sigma^2}\ge\frac{\alpha^2}{4}\tag{42}$$

as long as $n\gtrsim(\overline{\kappa}\vee B_X)^2B_X^2l^4\{\log(2p)+t\}$.

Lemma 19 *(Sun et al., 2020). Let $f:\mathbb{R}^p\mapsto\mathbb{R}$ be a differentiable convex function, and define the corresponding symmetrized Bregman divergence $D_f(\beta_1,\beta_2)=\langle\nabla f(\beta_2)-f(\beta_1),\beta_2-\beta_1\rangle$ for $\beta_1,\beta_2\in\mathbb{R}^p$. Then for any $\beta,\delta\in\mathbb{R}^p$ and $\lambda\in[0,1]$, $D_f(\beta_\lambda,\beta)\le\lambda\cdot D_f(\beta_1,\beta)$, where $\beta_\lambda=\beta+\lambda\delta$ and $\beta_1=\beta+\delta$.*

Lemma 20 *Assume Conditions 1, 2 and 3 hold, and let $a\in\mathbb{R}^p$ be such that $\|a\|_1/\|a\|_2=o(\sqrt{n/\log p})$ and $\|\Sigma^{-1}a\|_1\le C_a\|a\|_2$ for some $C_a>0$. Let $\mathcal{U}$ be the set defined by constraints (15)–(17) and let $\widehat{u}$ be a solution (if there is any). Provided that $n\gtrsim\log p$, the following statements hold (jointly) with probability at least $1-p^{-1}$:*
1) $\mathcal{U}$ contains at least one nonzero element;
2) any optimum $\widehat{u}$ satisfies $\overline{\kappa}^{-2}\|a\|_2(1-o(1))\le\|\widehat{u}\|_2\le\underline{\kappa}^{-2}\|a\|_2(1+o(1))$;
3) the conditional variance $s^2(\widehat{u})=n^{-1}\sum_{i=1}^n\mathbb{E}_{X_i}(\xi_i^2)(X_i^{\mathrm{T}}\widehat{u})^2$ satisfies

$$\overline{\kappa}^{-4}\underline{\kappa}^2\underline{\sigma}^2\|a\|_2^2(1-o(1))\le s^2(\widehat{u})\le\underline{\kappa}^{-2}\overline{\sigma}^2\|a\|_2^2(1+o(1)).\tag{43}$$

For any $\beta, \theta \in \mathbb{R}^p$, define

$$g_{\beta,\theta}(w_i) = \psi_\tau(\xi_i(\beta,\theta)) - \psi_\tau(\xi_i) + \alpha X_i^{\mathrm{T}}(\theta - \theta^*) \quad \text{with} \quad w_i = (X_i, \varepsilon_i).$$

Lemma 21 *Assume Conditions 1, 2 and 3 hold. Let $\Delta = \theta - \theta^*$ and $\Delta' = \beta - \beta^*$. Let $U_i = \Sigma^{-1/2} X_i$. For any given $0 < \delta_0, r_0 \le 1$,*

$$\sup_{\substack{\alpha\Delta \in \mathbb{B}_\Sigma(\delta_0) \\ \Delta' \in \mathbb{B}_\Sigma(r_0)}} \|\mathbb{E}\{g_{\beta,\theta}(w_i)U_i\}\|_2 \le \frac{1}{2}\kappa_3 \overline{f} r_0^2 + (r_0\delta_0 + r_0^2 + \delta_0^2/2)\frac{\kappa_3}{\tau} + (r_0 + \delta_0)\frac{\overline{\sigma}}{\tau}, \tag{44}$$

Lemma 22 *Assume Conditions 1, 2 and 3 hold. Let $\Delta = \theta - \theta^*$, $\Delta' = \beta - \beta^*$ and $t > 0$. For any $0 < r_0, \delta_0 \le 1$ and $\delta_1, r_1 > 0$,*

$$\sup_{\substack{\alpha\Delta \in \mathbb{B}_\Sigma(\delta_0)\cap\mathbb{B}_1(\delta_1) \\ \Delta' \in \mathbb{B}_\Sigma(r_0)\cap\mathbb{B}_1(r_1)}} \left\| \frac{1}{n}\sum_{i=1}^n (1 - \mathbb{E})\{g_{\beta,\theta}(w_i)X_i\} \right\|_\infty \lesssim B_X(\overline{\kappa}\,\overline{r}_1 + \overline{r}_0)\sqrt{\frac{\log(p) + t}{n}}. \tag{45}$$

with probability at least $1 - e^{-t}$, where $\overline{r}_0 = r_0 + \delta_0$ and $\overline{r}_1 = r_1 + \delta_1$

Appendix B. Proof of Main Results

In this section, we present the formal proofs for the theorems and propositions stated in the main text.

B.1 Proof of Proposition 2

To begin with, define

$$\widehat{\mathcal{R}}(\delta) = \widehat{\mathcal{Q}}(\beta^* + \delta) - \widehat{\mathcal{Q}}(\beta^*) + \lambda_q(\|\beta^* + \delta\|_1 - \|\beta^*\|_1), \quad \delta \in \mathbb{R}^p.$$

Note that $\widehat{\mathcal{R}}(0) = 0$ and $\widehat{\mathcal{R}}(\widehat{\delta}) \le 0$ by the optimality of $\widehat{\beta}$, where $\widehat{\delta} = \widehat{\beta} - \beta^*$. Let $w_\beta = (1/n)\sum_{i=1}^n \{\mathbb{1}(Y_i \le X_i^{\mathrm{T}}\beta) - \alpha\}X_i$ for $\beta \in \mathbb{R}^p$ so that $w_\beta \in \partial\widehat{\mathcal{Q}}(\beta)$ is a subgradient of $\widehat{\mathcal{Q}}$ at any β. Denote by $\mathcal{S} = \mathrm{supp}(\beta^*)$ the support of β^* satisfying $|\mathcal{S}| \le s_\beta$. Then, applying Proposition 9.13 and (9.50) in Wainwright (2019) with $\mathcal{L}_n = \widehat{\mathcal{Q}}$ and $\Phi(\cdot) = \|\cdot\|_1$ we obtain that conditioned on the event $\{\lambda_q \ge 2\|w_{\beta^*}\|_\infty\}$, the error $\widehat{\delta} = \widehat{\beta} - \beta^*$ belongs to the cone set $\mathbb{C}_3(\mathcal{S})$, and for any $\delta \in \mathbb{C}_3(\mathcal{S})$,

$$\begin{aligned} \widehat{\mathcal{R}}(\delta) &\ge \widehat{\mathcal{Q}}(\beta^* + \delta) - \widehat{\mathcal{Q}}(\beta^*) + \lambda_q\{\|\delta_{\mathcal{S}^c}\|_1 - \|\delta_{\mathcal{S}}\|_1\} \\ &\ge \widehat{\mathcal{Q}}(\beta^* + \delta) - \widehat{\mathcal{Q}}(\beta^*) - \underline{\kappa}^{-1}\sqrt{s_\beta}\lambda_q\|\delta\|_\Sigma. \end{aligned} \tag{46}$$

Next, let $\mathcal{Q}(\beta) = \mathbb{E}\widehat{\mathcal{Q}}(\beta)$ be the population quantile loss, satisfying $\nabla\mathcal{Q}(\beta^*) = 0$ and $\nabla^2\mathcal{Q}(\beta) = \mathbb{E}\{f_{\varepsilon_i|X_i}(\langle X_i, \beta - \beta^*\rangle)X_iX_i^{\mathrm{T}}\}$. Under Condition 2, it holds for any $\delta \in \mathbb{R}^p$ and $t \in [0,1]$ that

$$\begin{aligned} \langle \delta, \nabla^2\mathcal{Q}(\beta^* + t\delta)\delta\rangle &= \mathbb{E}\{f_{\varepsilon_i|X_i}(tX_i^{\mathrm{T}}\delta)(X_i^{\mathrm{T}}\delta)^2\} \\ &= \mathbb{E}\{f_{\varepsilon_i|X_i}(0)(X_i^{\mathrm{T}}\delta)^2\} + \mathbb{E}\{f_{\varepsilon_i|X_i}(tX_i^{\mathrm{T}}\delta) - f_{\varepsilon_i|X_i}(0)\}(X_i^{\mathrm{T}}\delta)^2 \\ &\ge \underline{f}\|\delta\|_\Sigma^2 - l_0 t \cdot \mathbb{E}|X_i^{\mathrm{T}}\delta|^3 \ge \underline{f}\|\delta\|_\Sigma^2 - l_0\kappa_3 t \cdot \|\delta\|_\Sigma^3. \end{aligned}$$

This together with the fundamental theorem of calculus implies

$$\mathcal{Q}(\beta^*+\delta)-\mathcal{Q}(\beta^*)=\underbrace{\langle\nabla\mathcal{Q}(\beta^*),\delta\rangle}_{=0}+\int_0^1\langle\nabla\mathcal{Q}(\beta^*+t\delta)-\nabla\mathcal{Q}(\beta^*),\delta\rangle\mathrm{d}t$$
$$=\int_0^1\int_0^1 u\langle\delta,\nabla^2\mathcal{Q}(\beta^*+tu\delta)\delta\rangle\mathrm{d}u\mathrm{d}t\geq\frac{1}{2}\underline{f}\,\|\delta\|_\Sigma^2-\frac{1}{6}l_0\kappa_3\|\delta\|_\Sigma^3.$$

For some $r_0, l_1>0$ to be determined, it follows from Lemma 13 that, with probability at least $1-e^{-t}$ (for any $t\geq 0$),

$$\sup_{\delta\in\mathbb{C}(l_1)\cap\mathbb{B}_\Sigma(r_0)}\{\mathcal{Q}(\beta^*+\delta)-\mathcal{Q}(\beta^*)-\widehat{\mathcal{Q}}(\beta^*+\delta)+\widehat{\mathcal{Q}}(\beta^*)\}\leq r_0l_1\cdot\tau(n,p,t),$$

where $\tau(n,p,t)=4\max(\alpha,1-\alpha)\{\overline{\kappa}\sqrt{2(\log(2p)+t)/n}+B_X(\log(2p)+t)/n\}$.

Together, the previous two inequalities and (46) show that with probability at least $1-e^{-t}$,

$$\widehat{\mathcal{R}}(\delta)\geq\frac{r_0}{2}\{\underline{f}r_0-\frac{1}{3}l_0\kappa_3r_0^2-2l_1\tau(n,p,t)-2\underline{\kappa}^{-1}s_\beta^{1/2}\lambda_q\}$$

holds for any $\delta\in\partial\mathbb{B}_\Sigma(r_0)\cap\mathbb{C}(l_1)$. We thus choose $r_0=4\underline{f}^{-1}\{l_1\tau(n,p,t)+\underline{\kappa}^{-1}\sqrt{s_\beta}\lambda_q\}$ and let (n,λ_q) satisfy $l_1\tau(n,p,t)+\underline{\kappa}^{-1}\sqrt{s_\beta}\lambda_q<3\underline{f}^2/(8l_0\kappa_3)$. Then, with high probability $\widehat{\mathcal{R}}(\delta)>0$ for all $\delta\in\partial\mathbb{B}_\Sigma(r_0)\cap\mathbb{C}(l_1)$. Recall that $\widehat{\mathcal{R}}(\widehat{\delta})\leq 0$ and $\widehat{\delta}\in\mathbb{C}_3(\mathcal{S})\subseteq\mathbb{C}(l_1)$ with $l_1=4\underline{\kappa}^{-1}\sqrt{s_\beta}$ conditioned on $\{\lambda_q\geq 2\|w_{\beta^*}\|_\infty\}$. Consequently, we conclude from Lemma 9.21 in Wainwright (2019) and the convexity of $\widehat{\mathcal{Q}}(\cdot)$ that $\widehat{\delta}\in\mathbb{B}_\Sigma(r_0)$ with probability at least $1-e^{-t}$ conditioned on $\{\lambda_q\geq 2\|w_{\beta^*}\|_\infty\}$. Combining this with Lemma 14 and (21) establishes the claim. ■

B.2 Proof of Theorem 3

Let $\widehat{\mathcal{L}}_\tau(\theta)=\widehat{\mathcal{L}}_\tau(\widehat{\beta},\theta)=(1/n)\sum_{i=1}^n\ell_\tau(\widehat{Z}_i-\alpha X_i^{\mathrm{T}}\beta)$, where $\widehat{\beta}$ is the ℓ_1-penalized QR estimator of β^*. Since $\widehat{\mathcal{L}}_\tau(\theta)$ is convex with respect to θ, for any optimum $\widehat{\theta}\in\mathrm{argmin}_\theta\{\widehat{\mathcal{L}}_\tau(\theta)+\alpha\lambda_e\|\theta\|_1\}$, there exists a subgradient vector $\widehat{z}\in\partial\|\widehat{\theta}\|_1$ satisfying $\langle\widehat{z},\widehat{\theta}\rangle=\|\widehat{\theta}\|_1$, $\|\widehat{z}\|_\infty\leq 1$ and $\nabla\widehat{\mathcal{L}}_\tau(\widehat{\theta})+\lambda_e\widehat{z}=0$. Therefore,

$$0\leq\langle\nabla\widehat{\mathcal{L}}_\tau(\widehat{\theta})-\nabla\widehat{\mathcal{L}}_\tau(\theta^*),\widehat{\theta}-\theta^*\rangle=-\alpha\lambda_e\langle\widehat{z},\widehat{\theta}-\theta^*\rangle+\langle-\nabla\mathcal{L}_\tau(\theta^*),\widehat{\theta}-\theta^*\rangle$$
$$=\alpha\lambda_e(\|\theta^*\|_1-\|\widehat{\theta}\|_1)+\langle-\nabla\mathcal{L}_\tau(\theta^*),\widehat{\theta}-\theta^*\rangle. \tag{47}$$

Let $\widehat{\Delta}=\widehat{\theta}-\theta^*$ denote the error vector, and let $\mathcal{T}=\mathrm{supp}(\theta^*)$ with $|\mathcal{T}|\leq s_\theta$. Then we have $\|\theta^*\|_1-\|\widehat{\theta}\|_1=\|\theta^*_{\mathcal{T}}\|_1-\|\widehat{\Delta}_{\mathcal{T}^c}+(\widehat{\Delta}+\theta^*)_{\mathcal{T}}\|\leq\|\widehat{\Delta}_{\mathcal{T}}\|_1-\|\widehat{\Delta}_{\mathcal{T}^c}\|_1$. To bound $\langle-\nabla\widehat{\mathcal{L}}_\tau(\theta^*),\widehat{\theta}-\theta^*\rangle$, consider the decomposition

$$-\nabla\widehat{\mathcal{L}}_\tau(\theta^*)=\frac{\alpha}{n}\sum_{i=1}^n\psi_\tau(\xi_i(\widehat{\beta}))X_i$$
$$=\frac{\alpha}{n}\sum_{i=1}^n(1-\mathbb{E})\{\psi_\tau(\xi_i)X_i\}+\frac{\alpha}{n}\sum_{i=1}^n(1-\mathbb{E})\{\psi_\tau(\xi_i(\widehat{\beta}))-\psi_\tau(\xi_i)X_i\}$$
$$+\alpha\mathbb{E}\{\psi_\tau(\xi_i(\widehat{\beta}))X_i\}.$$

Denote $U_i = \Sigma^{-1/2} X_i$. Note that conditioning on event $\{\widehat{\beta} \in \mathbb{B}_\Sigma(\beta^*, r_0) \cap \mathbb{C}(l_1)\}$, we have $\widehat{\beta} \in \mathbb{B}_\Sigma(\beta^*, r_0) \cap \mathbb{B}_1(\beta^*, l_1 r_0)$. Then it follows that

$$\begin{aligned}
\langle -\nabla \widehat{\mathcal{L}}(\theta^*), \widehat{\theta} - \theta^* \rangle \leq& \underbrace{\left\| \frac{1}{n} \sum_{i=1}^n (1 - \mathbb{E})\{\psi_\tau(\xi_i) X_i\} \right\|_\infty}_{=:\Lambda_1} \cdot \alpha \|\widehat{\Delta}\|_1 \\
&+ \underbrace{\sup_{\beta \in \mathbb{B}_1(\beta^*, l_1 r_0)} \left\| \frac{1}{n} \sum_{i=1}^n (1 - \mathbb{E})\{\psi_\tau(\xi_i(\beta)) - \psi_\tau(\xi_i)\} X_i \right\|_\infty}_{=:\Lambda_2} \cdot \alpha \|\widehat{\Delta}\|_1 \\
&+ \underbrace{\sup_{\beta \in \mathbb{B}_\Sigma(\beta^*, r_0)} \|\mathbb{E}\{\psi_\tau(\xi_i(\beta)) U_i\}\|_2}_{=:\Lambda_3} \cdot \alpha \|\widehat{\Delta}\|_\Sigma \\
=& (\Lambda_1 + \Lambda_2) \cdot \alpha \|\widehat{\Delta}\|_1 + \Lambda_3 \cdot \alpha \|\widehat{\Delta}\|_\Sigma. \qquad (48)
\end{aligned}$$

Provided that $\lambda_e \geq 2(\Lambda_1 + \Lambda_2)$, we have $0 \leq \frac{\alpha \lambda_e}{2}(3\|\widehat{\Delta}_\mathcal{T}\|_1 - \|\widehat{\Delta}_{\mathcal{T}^c}\|_1) + \Lambda_3 \cdot \alpha \|\widehat{\Delta}\|_\Sigma$, which implies the cone property:

$$\|\widehat{\Delta}_{\mathcal{T}^c}\|_1 \leq 3\|\widehat{\Delta}_\mathcal{T}\|_1 + 2\lambda_e^{-1} \Lambda_3 \|\widehat{\Delta}\|_\Sigma. \qquad (49)$$

Recall from Condition 1 that $\lambda_{\min}(\Sigma) \geq \underline{\kappa}^2$. Assume further that $\lambda_e \geq 2\underline{\kappa} s_\theta^{-1/2} \Lambda_3$. Then,

$$\begin{aligned}
\|\widehat{\Delta}\|_1 &\leq 4\|\widehat{\Delta}_\mathcal{T}\|_1 + \underline{\kappa}^{-1}\sqrt{s_\theta}\|\Delta\|_\Sigma \leq 4\sqrt{s_\theta}\|\widehat{\Delta}_\mathcal{T}\|_2 + \underline{\kappa}^{-1}\sqrt{s_\theta}\|\Delta\|_\Sigma \\
&\leq 4\sqrt{s_\theta}\|\widehat{\Delta}\|_2 + \underline{\kappa}^{-1}\sqrt{s_\theta}\|\Delta\|_\Sigma \leq 4\underline{\kappa}^{-1}\sqrt{s_\theta}\|\Delta\|_\Sigma + \underline{\kappa}^{-1}\sqrt{s_\theta}\|\Delta\|_\Sigma \\
&= 5\underline{\kappa}^{-1}\sqrt{s_\theta}\|\Delta\|_\Sigma.
\end{aligned}$$

This implies that $\widehat{\Delta} \in \mathbb{C}(l)$ with $l = 5\underline{\kappa}^{-1}\sqrt{s_\theta}$ if

$$\lambda_e \geq 2\max\{\Lambda_1 + \Lambda_2, \underline{\kappa} s_\theta^{-1/2} \Lambda_3\}. \qquad (50)$$

Given $r \geq 9\underline{\kappa}^{-1}\sqrt{s_\theta}\lambda_e$, we construct an intermediate "estimator" $\widetilde{\theta} = (1-\eta)\theta^* + \eta\theta$, where $\eta = \sup\{t \in [0,1] : \theta^* + t(\widehat{\theta} - \theta^*) \in \mathbb{B}_\Sigma(\theta^*, r/\alpha)\}$. If $\widehat{\theta} \in \mathbb{B}_\Sigma(\theta^*, r/\alpha)$, $\eta = 1$ and $\widetilde{\theta} = \widehat{\theta}$; otherwise, if $\widehat{\theta} \notin \mathbb{B}_\Sigma(\theta^*, r/\alpha)$, $\eta < 1$ and $\widetilde{\theta} \in \partial\mathbb{B}_\Sigma(\theta^*, r/\alpha)$, i.e., $\alpha\|\widetilde{\theta} - \theta^*\|_\Sigma = r$. Let $\widetilde{\Delta} = \widetilde{\theta} - \theta^*$, which satisfies $\alpha\widetilde{\Delta} \in \mathbb{B}_\Sigma(r)$. On the other hand, $\|\widetilde{\Delta}\|_1 = \eta\|\Delta\|_1 \leq \eta l\|\widehat{\Delta}\|_\Sigma = l\|\widetilde{\Delta}\|_\Sigma$, implying that $\widetilde{\Delta} \in \mathbb{C}(l)$. Since $\alpha\widetilde{\Delta} \in \mathbb{B}_\Sigma(r) \cap \mathbb{C}(l)$, it follows from Lemma 18 that with probability at least $1 - e^{-t}$,

$$\frac{\alpha^2}{4}\|\widetilde{\Delta}\|_\Sigma^2 \leq \langle \nabla\widehat{\mathcal{L}}_\tau(\widetilde{\theta}) - \nabla\widehat{\mathcal{L}}_\tau(\theta^*), \widetilde{\theta} - \theta^* \rangle. \qquad (51)$$

as long as $\tau \geq 2.5\max\{B_X(lr \vee l_1 r_0), \sqrt{50}\overline{\sigma}\}$ and $n \gtrsim (\overline{\kappa} \vee B_X)^2 B_X^2 l^4 \{\log(2p) + t\}$. On the other hand, Lemma 19 ensures that

$$\langle \nabla\widehat{\mathcal{L}}_\tau(\widetilde{\theta}) - \nabla\widehat{\mathcal{L}}_\tau(\theta^*), \widetilde{\theta} - \theta^* \rangle \leq \eta\langle \nabla\widehat{\mathcal{L}}_\tau(\widehat{\theta}) - \nabla\widehat{\mathcal{L}}_\tau(\theta^*), \widehat{\theta} - \theta^* \rangle. \qquad (52)$$

Furthermore,

$$
\begin{aligned}
\eta\langle\nabla\widehat{\mathcal{L}}_\tau(\widehat{\theta})-\nabla\widehat{\mathcal{L}}_\tau(\theta^*),\widehat{\theta}-\theta^*\rangle &\le \eta\cdot\alpha\lambda_e(\|\widehat{\Delta}_{\mathcal{T}}\|_1-\|\widehat{\Delta}_{\mathcal{T}^c}\|_1)+\eta\cdot\langle-\nabla\mathcal{L}_\tau(\theta^*),\widehat{\theta}-\theta^*\rangle \\
&= \alpha\lambda_e(\|\widetilde{\Delta}_{\mathcal{T}}\|_1-\|\widetilde{\Delta}_{\mathcal{T}^c}\|_1)+\langle-\nabla\mathcal{L}_\tau(\theta^*),\widetilde{\theta}-\theta^*\rangle \\
&\le \alpha\lambda_e(\|\widetilde{\Delta}_{\mathcal{T}}\|_1-\|\widetilde{\Delta}_{\mathcal{T}^c}\|_1)+\frac{\alpha\lambda_e}{2}\|\widetilde{\Delta}\|_1 \\
&\qquad+\frac{\alpha\lambda_e}{2}\cdot\underline{\kappa}^{-1}\sqrt{s_\theta}\|\widetilde{\Delta}\|_\Sigma \qquad (53)\\
&\le \frac{\alpha\lambda_e}{2}\underline{\kappa}^{-1}\sqrt{s_\theta}\cdot 3\|\widetilde{\Delta}\|_\Sigma+\frac{\alpha\lambda_e}{2}\cdot\underline{\kappa}^{-1}\sqrt{s_\theta}\|\widetilde{\Delta}\|_\Sigma \\
&= 2\alpha\lambda_e\underline{\kappa}^{-1}\sqrt{s_\theta}\cdot\|\widetilde{\Delta}\|_\Sigma, \qquad (54)
\end{aligned}
$$

where (53) follows from (48), (50) and (52). Then substituting (51) and (54) into (52) yields that with probability at least $1-3e^{-t}$,

$$\alpha\|\widetilde{\Delta}\|_\Sigma\le 8\underline{\kappa}^{-1}\sqrt{s_\theta}\lambda_e. \qquad (55)$$

Therefore, we have $\alpha\|\widetilde{\Delta}\|_\Sigma<r$ with probability at least $1-3e^{-t}$, i.e., $\widetilde{\theta}$ falls into the interior of $\mathbb{B}_\Sigma(\theta^*,r/\alpha)$. By the construction of $\widetilde{\theta}$, we must have $\widehat{\theta}=\widetilde{\theta}$.

It remains to establish a lower bound for λ_e so that the event (50) occurs with high probability. We finally choose $\tau\asymp\overline{\sigma}\sqrt{n/(\log(p)+t)}$ so that $\tau\ge 2.5\max\{B_X(lr\vee l_1r_0),\sqrt{50}\overline{\sigma}\}$ under the sample size requirement

$$n\gtrsim(\overline{\kappa}\vee B_X)^2B_X^2(l\vee l_1)^4\{\log(p)+t\}\asymp\{(\overline{\kappa}\vee B_X)/\underline{\kappa}\}^2B_X^2(s_\beta\vee s_\theta)^2\{\log(p)+t\}. \qquad (56)$$

By Lemma 15 and 16 and letting $\tau\asymp\overline{\sigma}\sqrt{n/(\log(p)+t)}$, we see that with probability at least $1-e^{-t}$,

$$
\begin{aligned}
\Lambda_1 &\le \sqrt{2}\overline{\kappa}\,\overline{\sigma}\sqrt{\frac{\log(2p)+t}{n}}+\tau B_X\frac{\log(2p)+t}{n}\lesssim\overline{\kappa}\,\overline{\sigma}\sqrt{\frac{\log(p)+t}{n}}\quad\text{and} \qquad (57)\\
\Lambda_2 &\lesssim B_Xl_1r_0\left(B_X\frac{\log(p)+t}{n}+\overline{\kappa}\sqrt{\frac{\log p+t}{n}}\right)\lesssim\overline{\kappa}B_Xl_1r_0\sqrt{\frac{\log(p)+t}{n}}, \qquad (58)
\end{aligned}
$$

where (58) used the condition $n\gtrsim(B_X/\overline{\kappa})^2(\log(p)+t)$, which is satisfied under sample size requirement (56). In addition, Lemma 17 yields

$$\underline{\kappa}s_\theta^{-1/2}\Lambda_3\le\underline{\kappa}s_\theta^{-1/2}\left\{\frac{\kappa_3}{2}\left(\overline{f}+\frac{1}{\tau}\right)r_0^2+\frac{\overline{\sigma}r_0}{\tau}+\frac{\overline{\sigma}^2}{\tau}\right\}\lesssim\overline{f}\underline{\kappa}\,s_\theta^{-1/2}\left(r_0^2+\overline{\sigma}\sqrt{\frac{\log(p)+t}{n}}\right). \qquad (59)$$

From (57)–(59) we see that inequality (50) holds with probability at least $1-2e^{-t}$, as long as

$$\lambda_e\gtrsim\max\left\{\overline{\kappa}(\overline{\sigma}+B_Xl_1r_0)\sqrt{\frac{\log(p)+t}{n}},\ \overline{f}\underline{\kappa}\,s_\theta^{-1/2}r_0^2\right\}.$$

Combining the results above, we conclude that for a sufficiently large n satisfying sample size requirement (56), $\alpha\|\widehat{\Delta}\|_\Sigma\le 8\underline{\kappa}^{-1}\sqrt{s_\theta}\lambda_e$ and $\alpha\|\widehat{\Delta}\|_1\le\alpha l\|\widehat{\Delta}\|_\Sigma\le 40\underline{\kappa}^{-2}s_\theta\lambda_e$ hold with probability at least $1-3e^{-t}$, as claimed. ∎

B.3 Proof of Proposition 4

The proof is similar to that of Theorem 3 with slight modifications. Applying the second upper bound in Lemma 17 to (59), we see that provided $\tau \gtrsim \alpha_q^{1/q}\{n/(\log(p)+t)\}^{1/2(q-1)}$,

$$\underline{\kappa} s_\theta^{-1/2}\Lambda_3 \lesssim \underline{\kappa} s_\theta^{-1/2}\left\{\overline{f} r_0^2 + r_0\left(\frac{\log(p)+t}{n}\right)^{\frac{1}{2(q-1)}} + \alpha_q^{1/q}\sqrt{\frac{\log(p)+t}{n}}\right\}.$$

On the other hand, with $\tau \lesssim \overline{\sigma}\sqrt{n/(\log(p)+t)}$ and $n \gtrsim (B_X/\overline{\kappa})^2(\log(p)+t)$, by (57) and (58), we have

$$\Lambda_1 + \Lambda_2 \lesssim \overline{\kappa}(\overline{\sigma} + B_X l_1 r_0)\sqrt{\frac{\log(p)+t}{n}}.$$

The rest follows from the proof of Theorem 3. ■

B.4 Proof of Theorem 5

Let $R_n = \alpha(\widehat{\omega}_\tau - \omega^*) - n^{-1}\sum_{i=1}^n \psi_\tau(\xi_i)X_i^{\mathrm{T}}\widehat{u}$. Recall that $U_i = \Sigma^{-1/2}X_i$, $\Delta = \theta - \theta^*$ and $\Delta' = \beta - \beta^*$. For any given $0 < r_0, \delta_0 \le 1$ and $\delta_1, r_1 > 0$,

$$\begin{aligned}
|R_n| &= \left|\frac{1}{n}\sum_{i=1}^n \{\psi_\tau(\xi_i(\widehat{\beta},\widehat{\theta})) - \psi_\tau(\xi_i)\}X_i^{\mathrm{T}}\widehat{u} + a^{\mathrm{T}}\alpha(\widehat{\theta} - \theta^*)\right| \\
&\le \underbrace{\sup_{\substack{\alpha\Delta\in\mathbb{B}_\Sigma(\delta_0)\\ \Delta'\in\mathbb{B}_\Sigma(r_0)}} \left\|\mathbb{E}\{\psi_\tau(\xi_i(\beta,\theta)) - \psi_\tau(\xi_i) + \alpha X_i^{\mathrm{T}}(\theta-\theta^*)\}U_i\right\|_2}_{=:\Gamma_1} \cdot \|\widehat{u}\|_\Sigma \\
&\quad + \underbrace{\sup_{\substack{\alpha\Delta\in\mathbb{B}_\Sigma(\delta_0)\cap\mathbb{B}_1(\delta_1)\\ \Delta'\in\mathbb{B}_\Sigma(r_0)\cap\mathbb{B}_1(r_1)}} \left\|\frac{1}{n}\sum_{i=1}^n (1-\mathbb{E})\{\psi_\tau(\xi_i(\beta,\theta)) - \psi_\tau(\xi_i) + \alpha X_i^{\mathrm{T}}(\theta-\theta^*)\}X_i\right\|_\infty}_{=:\Gamma_2} \cdot \|\widehat{u}\|_1 \\
&\quad + \left\|a - \frac{1}{n}\sum_{i=1}^n X_i X_i^{\mathrm{T}}\widehat{u}\right\|_\infty \cdot \alpha\|\widehat{\theta} - \theta^*\|_1. \qquad (60)
\end{aligned}$$

Applying Lemma 21 with $\tau \asymp \overline{\sigma}\sqrt{n/\log p}$, we have

$$\Gamma_1 \le \kappa_3 \overline{f} r_0^2/2 + \tau^{-1}\kappa_3(r_0\delta_0 + r_0^2 + \delta_0^2/2) + \tau^{-1}\overline{\sigma}(r_0+\delta_0) \lesssim \overline{f} r_0^2 + \overline{\sigma}\,\overline{r}_0\sqrt{\frac{\log p}{n}},$$

where $\overline{r}_0 = r_0 + \delta_0$. Additionally, by Lemma 22, $\Gamma_2 \lesssim B_X(\overline{\kappa}\,\overline{r}_1 + \overline{r}_0)\sqrt{\log(p)/n}$ with probability at least $1 - p^{-1}$. Combining with Lemma 20, constraints (15),(16) and (60) yields that

$$|R_n| \lesssim \left\{\underline{\kappa}^{-2}\overline{\kappa}^2\|a\|_2\Gamma_1 + C_a\|a\|_2\Gamma_2\right\} + \rho\|a\|_2\delta_1 \lesssim \|a\|_2\left(\overline{f} r_0^2 + \overline{r}_1\sqrt{\frac{\log p}{n}} + \overline{\sigma}\,\overline{r}_0\sqrt{\frac{\log p}{n}}\right).$$

with probability at least $1 - 2p^{-1}$. ■

B.5 Proof of Theorem 6

$$\alpha(\widehat{\omega}_\tau-\omega^*)=\underbrace{\frac{1}{n}\sum_{i=1}^n\psi_\tau(\xi_i)X_i^{\mathrm T}\widehat{u}}_{:=S_n}+\underbrace{\frac{1}{n}\sum_{i=1}^n\{\psi_\tau(\xi_i(\beta,\theta))-\psi_\tau(\xi_i)\}X_i^{\mathrm T}\widehat{u}+a^{\mathrm T}\alpha(\widehat{\theta}-\theta^*)}_{:=R_n}. \quad (61)$$

By Theorems 3, 5 and Lemma 20 with $\tau\asymp\overline{\sigma}\sqrt{n/\log p}$, $R_n/s(\widehat{u})=O_{\mathbb P}\{(s_\beta+s_\theta)\log(p)/n\}$, and thus $R_n/s(\widehat{u})=o_{\mathbb P}(n^{-1/2})$ provided that $\max\{s_\beta,s_\theta\}=o(\sqrt{n}/\log p)$. Hence, it suffices to show the asymptotic behavior for S_n, and the result follows from Slutsky's Theorem.

Let $V_i=\{s(\widehat{u})\}^{-1}\psi_\tau(\xi_i)X_i^{\mathrm T}\widehat{u}$ and $W_i=V_i-\mathbb{E}_{X_i}(V_i)$ so that

$$S_n=\underbrace{\frac{1}{n}\sum_{i=1}^n W_i}_{:=S_{1,n}}+\underbrace{\frac{1}{n}\sum_{i=1}^n\mathbb{E}_{X_i}(V_i)}_{:=S_{2,n}}.$$

We will show $\sqrt{n}S_{1,n}\xrightarrow{D}\mathcal{N}(0,1)$ and $S_{2,n}=o_{\mathbb P}(n^{-1/2})$. We first show the asymptotic normality of $S_{1,n}$ conditioned on $\{X_i\}_{i=1}^n$. Let $\widetilde{\Lambda}_\tau=n^{-1}\sum_{i=1}^n\mathbb{E}_{X_i}\{\psi_\tau^2(\xi_i)\}X_iX_i^{\mathrm T}$ and recall that $\widetilde{\Lambda}=n^{-1}\sum_{i=1}^n\mathbb{E}_{X_i}(\xi_i^2)X_iX_i^{\mathrm T}$. Note that

$$\begin{aligned}\mathrm{var}_X(\sqrt{n}S_{1,n})&\le\{s(\widehat{u})\}^{-2}\cdot\widehat{u}^{\mathrm T}\widetilde{\Lambda}_\tau\widehat{u}\\&=1+\widehat{u}^{\mathrm T}(\widetilde{\Lambda}_\tau-\widetilde{\Lambda})\widehat{u}\\&=1+\widehat{u}^{\mathrm T}\left\{\frac{1}{n}\sum_{i=1}^n\mathbb{E}_{X_i}(\psi_\tau^2(\xi_i)-\xi_i^2)X_iX_i^{\mathrm T}\right\}\widehat{u}\end{aligned}$$

Note that $\mathbb{E}_X|\xi_i|^3\le\alpha_3$ (See Remark 7). Since $|\psi_\tau^2(\xi_i)-\xi_i^2|=|(\xi_i^2-\tau^2)\mathbb{1}(|\xi_i|\ge\tau)|\le|\xi_i|^3/\tau$ and $\tau\asymp\overline{\sigma}\sqrt{n/\log p}$, by constraint (16) and the fact that $\mathbb{E}_{X_i}|\xi_i|^3\le\alpha_3$, we have

$$\widehat{u}^{\mathrm T}(\widetilde{\Lambda}_\tau-\widetilde{\Lambda})\widehat{u}\le\|\widetilde{\Lambda}_\tau-\widetilde{\Lambda}\|_{\max}\|\widehat{u}\|_1^2\le\alpha_3C_a^2\|a\|_2^2/\tau. \quad (62)$$

Therefore, $\limsup_{n,p\to\infty}\mathrm{var}_{X_i}(\sqrt{n}S_{1,n})\le1$. On the other hand, by Lemma 20, with probability at least $1-p^{-1}$,

$$\begin{aligned}\mathrm{var}_{X_i}(\sqrt{n}S_{1,n})&=\{s(\widehat{u})\}^{-2}\cdot\widehat{u}^{\mathrm T}\widehat{\Lambda}_\tau\widehat{u}-\{s(\widehat{u})\}^{-2}\cdot\widehat{u}^{\mathrm T}\left(\frac{1}{n}\sum_{i=1}^n\{\mathbb{E}_{X_i}\psi_\tau(\xi_i)\}^2X_iX_i^{\mathrm T}\right)\widehat{u}\\&\ge1+\frac{\widehat{u}^{\mathrm T}(\widetilde{\Lambda}_\tau-\widetilde{\Lambda})\widehat{u}}{s^2(\widehat{u})}-\frac{\overline{\sigma}^2}{\tau^2s^2(\widehat{u})}\cdot\widehat{u}^{\mathrm T}\widehat{\Sigma}\widehat{u}\\&\ge1-\frac{\alpha_3C_a^2\|a\|_2^2}{\tau\{\underline{\sigma}^2\underline{\kappa}^2\overline{\kappa}^{-4}\|a\|_2^2(1+o(1))\}}-\frac{\overline{\sigma}^2\underline{\kappa}^{-2}\|a\|_2^2(1+o(1))}{\tau^2\{\underline{\sigma}^2\underline{\kappa}^2\overline{\kappa}^{-4}\|a\|_2^2(1+o(1))\}}\\&\gtrsim1-\tau^{-1}\alpha_3\overline{\kappa}^4/(\underline{\sigma}\underline{\kappa})^2-\tau^{-2}(\overline{\sigma}/\underline{\sigma})^2(\overline{\kappa}/\underline{\kappa})^4,\end{aligned}$$

where the second and third inequalities use $|\psi_\tau(t)-t|\le t^2/\tau$ and (62), respectively. Hence, with $\tau\asymp\overline{\sigma}\sqrt{n/\log p}$, $\liminf_{n,p\to\infty}\mathrm{var}_{X_i}(\sqrt{n}S_{1,n})=1$ almost surely. Consequently, we have

$\lim_{n,p\to\infty} \mathrm{var}_{X_i}(\sqrt{n}S_{1,n}) = 1$ almost surely. It remains to check Lindeberg's condition. By Lemma 20, for any constant $b > 0$,

$$\mathbb{E}_{X_i}\{W_i^2 \mathbb{1}(|W_i| \geq b\sqrt{n})\} \leq \frac{\mathbb{E}_{X_i}|W_i|^3}{b\sqrt{n}} \leq \frac{8\mathbb{E}_{X_i}|V_i|^3}{b\sqrt{n}} \leq \frac{8(X_i^{\mathrm{T}}\widehat{u})^3 \cdot \mathbb{E}_{X_i}|\psi_\tau(\xi_i)|^3}{b\sqrt{n}\cdot s^3(\widehat{u})}$$
$$\leq \frac{8\alpha_3 B_X^3 C_a^3 \|a\|_2^3}{b\sqrt{n}\cdot s^3(\widehat{u})} \lesssim \frac{\alpha_3}{b\sqrt{n}} \cdot \left(\frac{B_X C_a \overline{\kappa}^2}{\underline{\sigma}\,\underline{\kappa}}\right)^3$$

with probability at least $1 - p^{-1}$. Thus, $\lim_{n,p\to\infty} n^{-1}\sum_{i=1}^n \mathbb{E}_{X_i}\{W_i^2\mathbb{1}(|W_i| \geq b\sqrt{n})\} = 0$ almost surely, and the Lindeberg's condition is satisfied. Therefore, $\sqrt{n}S_{1,n} \xrightarrow{\mathrm{d}} \mathcal{N}(0,1)$ conditioned on $\{X_i\}_{i=1}^n$. Finally, by calculating the characteristic function of $S_{1,n}$ and applying the bounded convergence theorem, we have $\sqrt{n}S_{1,n} \xrightarrow{\mathrm{d}} \mathcal{N}(0,1)$.

For $S_{2,n}$, since $|\psi_\tau(t) - t| \leq t^{q+1}/\tau^q$ for $q \in \mathbb{N}_+$, $\mathbb{E}_{X_i}\{\psi_\tau(\xi_i)\} \leq \mathbb{E}_{X_i}|\xi_i|^3/\tau^2 \leq \alpha_3/\tau^2$, by Lemma 20 and constraint (16),

$$S_{2,n} \leq \frac{\alpha_3}{\tau^2 s(\widehat{u})} \cdot \left(\frac{1}{n}\sum_{i=1}^n X_i^{\mathrm{T}}\widehat{u}\right) \leq \frac{\alpha_3 B_X C_a}{\tau^2 s(\widehat{u})} \lesssim \frac{\alpha_3 B_X C_a \overline{\kappa}^2}{\underline{\sigma}\,\underline{\kappa}\|a\|_2} \cdot \frac{1}{\tau^2}.$$

with probability at least $1 - p^{-1}$. Hence, with $\tau \asymp \overline{\sigma}\sqrt{n/\log p}$ and the growth condition $n \gtrsim (\log p)^2$, $S_{2,n} = o_{\mathbb{P}}(n^{-1/2})$. ■

B.6 Proof of Proposition 9

Denote $\Delta = \theta - \theta^*$ and $\Delta' = \beta - \beta^*$. Let $u_i = X_i^{\mathrm{T}}\Delta'$ and $v_i = X_i^{\mathrm{T}}\Delta$. Conditioning on the event $\{\Delta' \in \mathbb{B}_\Sigma(r_0) \cap \mathbb{B}_1(r_1)\} \cap \{\alpha\Delta \in \mathbb{B}_\Sigma(\delta_0) \cap \mathbb{B}_1(\delta_1)\}$ with $0 < r_0, \delta_0 \leq 1$ and $\delta_1, r_1 > 0$,

$$\begin{aligned}
\|\widehat{\Lambda}_\gamma - \Lambda\|_{\max} &= \max_{j,k}\left|\{\widehat{\Lambda}_\gamma - \Lambda\}_{jk}\right| \leq \max_{j,k}\left|\frac{1}{n}\sum_{i=1}^n\{\psi_\gamma^2(\xi_i(\widehat{\beta},\widehat{\theta}))x_{ij}x_{ik} - \mathbb{E}(\xi_i^2 x_{ij}x_{ik})\}\right| \\
&\leq \underbrace{\max_{j,k}\sup_{\substack{\alpha\Delta\in\mathbb{B}_\Sigma(\delta_0)\\ \Delta'\in\mathbb{B}_\Sigma(r_0)}}\left|\mathbb{E}[\{\psi_\gamma^2(\xi_i(\beta,\theta)) - \psi_\gamma^2(\xi_i)\}x_{ij}x_{ik}]\right|}_{:=S_1} \\
&\quad + \underbrace{\max_{j,k}\sup_{\substack{\alpha\Delta\in\mathbb{B}_\Sigma(\delta_0)\cap\mathbb{B}_1(\delta_1)\\ \Delta'\in\mathbb{B}_\Sigma(r_0)\cap\mathbb{B}_1(r_1)}}\left|\frac{1}{n}\sum_{i=1}^n(1-\mathbb{E})[\{\psi_\gamma^2(\xi_i(\beta,\theta)) - \psi_\gamma^2(\xi_i)\}x_{ij}x_{ik}]\right|}_{:=S_2} \\
&\quad + \underbrace{\max_{j,k}\left|\frac{1}{n}\sum_{i=1}^n(1-\mathbb{E})\{\psi_\gamma^2(\xi_i)x_{ij}x_{ik}\}\right|}_{:=S_3} + \underbrace{\max_{j,k}\left|\mathbb{E}[\{\psi_\gamma^2(\xi_i) - \xi_i^2\}x_{ij}x_{ik}]\right|}_{:=S_4}
\end{aligned}$$

Upper bound for S_1: Note that $\xi_i(\beta,\theta) = \phi(\varepsilon_i - u_i) + \alpha X_i^{\mathrm{T}}\beta - \alpha X_i^{\mathrm{T}}\theta$ can be rewritten as

$$\xi_i(\beta,\theta) = \xi_i + (\phi(\varepsilon_i - u_i) - \phi(\varepsilon_i)) + \alpha u_i - \alpha v_i,$$

where $\phi(t) = t\mathbb{1}(t \le 0)$. Additionally, notice that $\psi_\gamma^2(t) = \psi_{\gamma^2}(t^2)$. Since $\phi(\cdot)$ and $\psi_{\gamma^2}(\cdot)$ are 1-Lipschitz continuous,

$$\begin{aligned}
|\psi_\gamma^2(\xi_i(\beta,\theta)) - \psi_\gamma^2(\xi_i)| &\le |\xi_i^2(\beta,\theta) - \xi_i^2| \\
&\le 2|\xi_i| \cdot |(\phi(\varepsilon_i - u_i) - \phi(\varepsilon_i)) + \alpha u_i - \alpha v_i| \\
&\quad + |(\phi(\varepsilon_i - u_i) - \phi(\varepsilon_i)) + \alpha u_i - \alpha v_i|^2 \\
&\le 2|\xi_i| \cdot \{(1+\alpha)|u_i| + \alpha|v_i|\} + \{(1+\alpha)|u_i| + \alpha|v_i|\}^2 \\
&\le 2|\xi_i| \cdot (2|u_i| + \alpha|v_i|) + (2|u_i| + \alpha|v_i|)^2
\end{aligned} \tag{63}$$

Because $\mathbb{E}_{X_i}|\xi_i| \le \overline{\sigma}$, and for any $\Delta' \in \mathbb{B}_\Sigma(r_0)$ and $\alpha\Delta \in \mathbb{B}_\Sigma(\delta_0)$, $\mathbb{E}|u_i| \le r_0$ and $\mathbb{E}|v_i| \le \delta_0$,

$$\mathbb{E}[\{\psi_\gamma^2(\xi_i(\beta,\theta)) - \psi_\gamma^2(\xi_i)\}x_{ij}x_{ik}] \le 2B_X^2\overline{\sigma}(2r_0 + \delta_0) + B_X^2(2r_0+\delta_0)^2 \lesssim B_X^2\overline{\sigma}\,\overline{r}_0,$$

where $\overline{r}_0 = r_0 + \delta_0$. Since this inequality holds uniformly over $(\Delta', \alpha\Delta) \in \mathbb{B}_\Sigma(r_0) \times \mathbb{B}_\Sigma(\delta_0)$ and $j,k \in \{1,\ldots,p\}$, the same upper bound holds for S_1.

Upper bound for S_2: We apply a similar method as in Lemmas 18 and 22. Denote $\widetilde{X}_i = (X_i^{\mathrm{T}}, -X_i^{\mathrm{T}})^{\mathrm{T}} = (x_{i,1},\ldots,x_{i,2p})^{\mathrm{T}} \in \mathbb{R}^{2p}$ and $h_{j,k}(w_i;\beta,\theta) = \{\psi_\gamma^2(\xi_i(\beta,\theta)) - \psi_\gamma^2(\xi_i)\}\widetilde{x}_{ij}\widetilde{x}_{ik}$, where $w_i = (\varepsilon_i, X_i^{\mathrm{T}})^{\mathrm{T}}$. Then

$$\max_{j,k}\left|\frac{1}{n}\sum_{i=1}^n (1-\mathbb{E})[\{\psi_\gamma^2(\xi_i(\beta,\theta)) - \psi_\gamma^2(\xi_i)\}x_{ij}x_{ik}]\right| = \max_{j,k}\frac{1}{n}\sum_{i=1}^n (1-\mathbb{E})h_{j,k}(w_i;\beta,\theta).$$

Note that

$$\begin{aligned}
\left|\frac{\partial h_{j,k}(w_i;\beta,\theta)}{\partial u_i}\right| &= |2\psi_\gamma(\xi_i(\beta,\theta)) \cdot \psi_\gamma'(\xi_i(\beta,\theta)) \cdot (\alpha - \mathbb{1}(\varepsilon \le u_i))\widetilde{x}_{ij}\widetilde{x}_{ik}| \le 2\gamma B_X^2 \\
\left|\frac{\partial h_{j,k}(w_i;\beta,\theta)}{\partial v_i}\right| &= |2\psi_\gamma(\xi_i(\beta,\theta)) \cdot \psi_\gamma'(\xi_i(\beta,\theta)) \cdot (-\alpha)| \le 2\alpha\gamma B_X^2.
\end{aligned}$$

Therefore, for any $\Delta' \in \mathbb{B}_\Sigma(r_0) \cap \mathbb{B}_1(r_1)$ and $\alpha\Delta \in \mathbb{B}_\Sigma(\delta_0) \cap \mathbb{B}_1(\delta_1)$,

$$\begin{aligned}
|h_{j,k}(w_i;\beta,\theta)| &\le \{|\psi_\gamma^2(\xi_i(\beta,\theta)) - \psi_\gamma^2(\xi_i(\beta^*,\theta))| + |\psi_\gamma^2(\xi_i(\beta^*,\theta)) - \psi_\gamma^2(\xi_i(\beta^*,\theta^*))|\} \cdot |\widetilde{x}_{ij}\widetilde{x}_{ik}| \\
&\le 2\gamma B_X^2(u_i + v_i) \le 2\gamma B_X^3(r_1+\delta_1) \lesssim \gamma B_X^3 \overline{r}_1,
\end{aligned}$$

where $\overline{r}_1 = r_1 + \delta_1$. On the other hand, we have $\mathbb{E}h_{j,k}^2(w_i) \lesssim B_X^4\mathbb{E}\{\xi_i^2(u_i^2 + \alpha^2 v_i^2)\} \lesssim B_X^4\overline{\sigma}^2\overline{r}_0^2$. from (63). Denote

$$\Lambda_{(j,k),n} = \sup_{\substack{\alpha\Delta \in \mathbb{B}_\Sigma(\delta_0)\cap\mathbb{B}_1(\delta_1) \\ \Delta' \in \mathbb{B}_\Sigma(r_0)\cap\mathbb{B}_1(r_1)}} \frac{1}{n}\sum_{i=1}^n (1-\mathbb{E})h_{j,k}(w_i;\beta,\theta).$$

By Theorem 7.3 in Bousquet (2003), with probability at least $1 - e^{-t}$,

$$\Lambda_{(j,k),n} \lesssim \mathbb{E}\Lambda_{(j,k),n} + (\mathbb{E}\Lambda_{(j,k),n})^{1/2}\sqrt{\gamma B_X^3\overline{r}_1} \cdot \sqrt{\frac{t}{n}} + B_X^2\overline{\sigma}\,\overline{r}_0\sqrt{\frac{t}{n}} + \gamma B_X^3\overline{r}_1 \cdot \frac{t}{n}, \tag{64}$$

and thus it remains to bound $\mathbb{E}\Lambda_{(j,k),n}$. By Rademacher symmetrization and the relationship between Gaussian and Rademacher complexities (e.g., Lemma 4.5 in Ledoux and Talagrand (1991)), we have

$$\mathbb{E}\Lambda_{(j,k),n} \leq \sqrt{2\pi}\mathbb{E}\left\{ \sup_{\substack{\alpha\Delta\in\mathbb{B}_\Sigma(\delta_0)\cap\mathbb{B}_1(\delta_1)\\ \Delta'\in\mathbb{B}_\Sigma(r_0)\cap\mathbb{B}_1(r_1)}} \underbrace{\frac{1}{n}\sum_{i=1}^n g_i h_{j,k}(w_i;\beta,\theta)}_{\mathbb{G}_{\beta,\theta}} \right\},$$

where $g_1,\ldots,g_n$ are i.i.d. standard normal random variables. For any $(\Delta_1',\Delta_1),(\Delta_2',\Delta_2)\in\{\mathbb{B}_\Sigma(\delta_0)\cap\mathbb{B}_1(\delta_1)\}\times\{\mathbb{B}_\Sigma(r_0)\cap\mathbb{B}_1(r_1)\}$,

$$\begin{aligned}\mathbb{G}_{\beta_1,\theta_1}-\mathbb{G}_{\beta_2,\theta_2} &= \frac{1}{n}\sum_{i=1}^n g_i\{\psi_\gamma^2(\xi_i(\beta_1,\theta_1))-\psi_\gamma^2(\xi_i(\beta_2,\theta_1)\}\widetilde{x}_{ij}\widetilde{x}_{ik}\\ &\quad+\frac{1}{n}\sum_{i=1}^n g_i\{\psi_\gamma^2(\xi_i(\beta_2,\theta_1))-\psi_\gamma^2(\xi_i(\beta_2,\theta_1)\}\widetilde{x}_{ij}\widetilde{x}_{ik}\end{aligned}$$

By Lipschitz continuity of $h_{j,k}$, we have

$$\begin{aligned}\mathbb{E}_w(\mathbb{G}_{\beta_1,\theta_1}-\mathbb{G}_{\beta_2,\theta_2})^2 &\leq 2\mathbb{E}_w(\mathbb{G}_{\beta_1,\theta_1}-\mathbb{G}_{\beta_2,\theta_1})^2+2\mathbb{E}_w(\mathbb{G}_{\beta_2,\theta_1}-\mathbb{G}_{\beta_2,\theta_2})^2\\ &\leq \frac{8\gamma^2B_X^4}{n^2}\sum_{i=1}^n\langle X_i,\Delta_1'-\Delta_2'\rangle^2+\frac{8\alpha^2\gamma^2B_X^4}{n^2}\sum_{i=1}^n\langle X_i,\Delta_1-\Delta_2\rangle^2.\end{aligned}$$

Define another Gaussian process $\{\mathbb{Z}_{\beta,\theta}\}$ as

$$\mathbb{Z}_{\beta,\theta}=\frac{2\sqrt{2}\gamma B_X^2}{n}\sum_{i=1}^n g_i'\langle X_i,\Delta'\rangle+\frac{2\sqrt{2}\alpha\gamma B_X^2}{n}\sum_{i=1}^n g_i''\langle X_i,\Delta\rangle,$$

where $g_1',\ldots,g_n'$ and $g_1'',\ldots,g_n''$ are i.i.d. standard normal random variables. Then, we have $\mathbb{E}_w(\mathbb{G}_{\beta_1,\theta_1}-\mathbb{G}_{\beta_2,\theta_2})^2\leq\mathbb{E}_w(\mathbb{Z}_{\beta_1,\theta_1}-\mathbb{Z}_{\beta_2,\theta_2})^2$. Hence, by Sudakov-Fernique's Gaussian comparison inequality (e.g., Theorem 7.2.11 in Vershynin (2018)),

$$\mathbb{E}_{w_i}\left\{\sup_{\substack{\alpha\Delta\in\mathbb{B}_\Sigma(\delta_0)\cap\mathbb{B}_1(\delta_1)\\ \Delta'\in\mathbb{B}_\Sigma(r_0)\cap\mathbb{B}_1(r_1)}}\mathbb{G}_{\beta,\theta}\right\}\leq\mathbb{E}_{w_i}\left\{\sup_{\substack{\alpha\Delta\in\mathbb{B}_\Sigma(\delta_0)\cap\mathbb{B}_1(\delta_1)\\ \Delta'\in\mathbb{B}_\Sigma(r_0)\cap\mathbb{B}_1(r_1)}}\mathbb{Z}_{\beta,\theta}\right\},$$

which remains true by replacing $\mathbb{E}_w$ with $\mathbb{E}$. Note that

$$\begin{aligned}\mathbb{E}\left\{\sup_{\substack{\alpha\Delta\in\mathbb{B}_\Sigma(\delta_0)\cap\mathbb{B}_1(\delta_1)\\ \Delta'\in\mathbb{B}_\Sigma(r_0)\cap\mathbb{B}_1(r_1)}}\mathbb{Z}_{\beta,\theta}\right\} &\leq 2\sqrt{2}\gamma B_X^2r_1\cdot\mathbb{E}\left\|\frac{1}{n}\sum_{i=1}^n g_i'X_i\right\|_\infty+2\sqrt{2}\gamma B_X^2\delta_1\cdot\mathbb{E}\left\|\frac{1}{n}\sum_{i=1}^n g_i''X_i\right\|_\infty\\ &\lesssim (r_1+\delta_1)\gamma B_X^3\overline{\kappa}\sqrt{\frac{\log p}{n}}=\gamma B_X^3\overline{\kappa}\,\overline{r}_1\sqrt{\frac{\log p}{n}},\end{aligned}$$

where the last inequality uses (87). Putting together the pieces, we have $\mathbb{E}\Lambda_{(j,k),n} \lesssim \gamma B_X^3 \overline{\kappa}\,\overline{r}_1 \sqrt{\log(p)/n}$. Finally, by taking the union bound over $j,k = \{1,\ldots,2p\}$ and setting $u = \log(4p^2)+t$, we obtain that with probability at least $1-e^{-t}$,

$$S_2 \lesssim \gamma B_X^3 \overline{\kappa}\,\overline{r}_1 \sqrt{\frac{\log(p)+t}{n}} + B_X^2 \overline{\sigma}\,\overline{r}_0 \sqrt{\frac{\log(p)+t}{n}}.$$

<u>*Upper bound for S_3*</u>: Note that $\mathbb{E}\{(\psi_\gamma^2(\xi_i)x_{ij}x_{ik})^2\} \leq \gamma \mathbb{E}\{(\mathbb{E}_{X_i}|\xi_i|^3)\cdot x_{ij}^2 x_{ik}^2\} \leq B_X^2\overline{\kappa}^2\gamma\alpha_3$, and for all $k \geq 3$, $\mathbb{E}|\psi_\gamma^2(\xi_i)x_{ij}x_{ik}|^k \leq (\gamma^2 B_X^2)^{k-2}\cdot B_X^2\overline{\kappa}^2\gamma\alpha_3 \leq (k!/2)\cdot(\gamma^2 B_X^2)^{k-2}\cdot B_X^2\overline{\kappa}^2\alpha_3$. By Bernstein's inequality and the union bound, with probability at least $1-2e^{-t}$,

$$S_3 \leq B_X\overline{\kappa}(\gamma\alpha_3)^{1/2}\sqrt{\frac{\log(2p)+t}{n}} + \gamma^2 B_X^2 \frac{\log(2p)+t}{n}.$$

<u>*Upper bound for S_4*</u>: Since $|\psi_\tau^2(\xi_i) - \xi_i^2| = |(\xi_i^2-\tau^2)\mathbb{1}(|\xi_i|\geq\tau)| \leq |\xi_i|^3/\tau$ and $\mathbb{E}_{X_i}(|\xi_i|^3) \leq \alpha_3$, we have $S_4 \leq \alpha_3\|\Sigma\|_{\max}/\gamma \leq \overline{\kappa}\alpha_3/\gamma$. Finally, combining the upper bounds for S_i, $i=1,\ldots,4$, yields the result. ■

B.7 Proof of Theorem 10

Let $\widetilde{\Lambda}_\gamma = n^{-1}\sum_{i=1}^n \mathbb{E}_{X_i}\{\psi_\gamma^2(\xi_i)\}X_iX_i^{\mathrm{T}}$. Note that

$$\begin{aligned}|\widehat{s}_\gamma^2(\widehat{u}) - s_\gamma^2(\widehat{u})| &= |\widehat{u}^{\mathrm{T}}(\widehat{\Lambda}_\gamma - \widetilde{\Lambda}_\gamma)\widehat{u}| \leq \|\widehat{\Lambda}_\gamma - \widetilde{\Lambda}_\gamma\|_{\max}\|\widehat{u}\|_1^2 \\ &\leq C_a^2\|\widehat{\Lambda}_\gamma - \Lambda_\gamma\|_{\max} + C_a^2\|\widetilde{\Lambda}_\gamma - \Lambda_\gamma\|_{\max}.\end{aligned}$$

For the first term, applying Theorem 3 and Proposition 9 with $\gamma \asymp (n/\log p)^{1/3}$ and the scaling condition $\max\{s_\beta, s_\theta\} = o(\sqrt{n}/\log p)$, we have

$$\|\widehat{\Lambda}_\gamma - \Lambda_\gamma\|_{\max} \lesssim \sqrt{\frac{(s_\beta+s_\theta)\log p}{n}} + \gamma\frac{(s_\beta+s_\theta)\log p}{n} + \gamma^2\frac{\log p}{n} + \frac{\alpha_3}{\gamma} \lesssim \left(\frac{\log p}{n}\right)^{1/3}$$

with probability at least $1-2p^{-1}$. On the other hand, note that $\mathbb{E}[\{(\mathbb{E}_{X_i}\psi_\gamma^2(\xi_i))x_{ij}x_{ik}\}^2] \leq \overline{\sigma}^4 B_X^2\overline{\kappa}^2$, and for all $k\geq 3$,

$$\mathbb{E}|\mathbb{E}_{X_i}\{\psi_\gamma^2(\xi_i)\}x_{ij}x_{ik}|^k \leq (\gamma^2 B_X^2)^k\cdot\overline{\sigma}^4 B_X^2\overline{\kappa}^2 \leq \frac{k!}{2}\cdot(\gamma^2 B_X^2)^k\cdot\overline{\sigma}^4 B_X^2\overline{\kappa}^2.$$

By Bernstein's inequality and the union bound, we have

$$\|\widetilde{\Lambda}_\gamma - \Lambda_\gamma\|_{\max} \lesssim \overline{\sigma}B_X\overline{\kappa}\sqrt{\frac{\log(2p)}{n}} + \gamma^2 B_X^2\frac{\log(2p)}{n} \lesssim \left(\frac{\log p}{n}\right)^{1/3}.$$

with probability at least $1-2p^{-1}$. Combining the results above establishes the claim. ■

Appendix C. Proof of Technical Lemmas

In this section, we present the proofs for the technical lemmas employed in establishing the main theorems and propositions.

C.1 Proof of Lemma 13

For each sample $d_i = (X_i, Y_i)$ and $\delta \in \mathbb{R}^p$, define $s(\delta; d_i) = \rho_\alpha(Y_i - X_i^{\mathrm{T}}(\beta^* + \delta)) - \rho_\alpha(Y_i - X_i^{\mathrm{T}}\beta^*) = \rho_\alpha(\varepsilon_i - X_i^{\mathrm{T}}\delta) - \rho_\alpha(\varepsilon_i)$. Then $\widehat{\mathcal{D}}(\delta) = \frac{1}{n}\sum_{i=1}^n s(\delta; d_i)$. By the Lipschitz continuity of ρ_α, $s(\delta; d_i)$ is Lipschitz continuous in $X_i^{\mathrm{T}}\delta$, with Lipschitz constant $\bar{\alpha}$. Given $r_0, l_1 > 0$, define random variable

$$\Lambda(r_0, l_1) = \frac{n}{4\bar{\alpha}r_0 l_1} \sup_{\delta \in \mathbb{B}_\Sigma(r_0) \cap \mathbb{C}(l_1)} \{\mathcal{D}(\delta) - \widehat{\mathcal{D}}(\delta)\}.$$

For any $u > 0$, by Chernoff's inequality,

$$\mathbb{P}\{\Lambda(r_0, l_1) \geq u\} \leq \exp\left[- \sup_{\lambda \geq 0} \{\lambda u - \log \mathbb{E} e^{\lambda \Lambda(r_0, l_1)}\}\right]. \tag{65}$$

Next we bound $\mathbb{E}e^{\lambda\Lambda(r_0,l_1)}$. Note that for any $\delta \in \mathbb{B}_\Sigma(r_0) \cap \mathbb{C}(l_1)$, $\|\delta\|_1 \leq r_0 l_1$. Then using Rademacher symmetrization and Ledoux-Talagrand contraction inequality (see, e.g., Ledoux and Talagrand, 1991),

$$\begin{aligned}
\mathbb{E}e^{\lambda\Lambda(r_0,l_1)} &\leq \mathbb{E}\exp\left\{ \frac{\lambda}{2\bar{\alpha}r_0 l_1} \sup_{\delta \in \mathbb{B}_\Sigma(r_0)\cap\mathbb{C}(l_1)} \sum_{i=1}^n \epsilon_i \cdot s(\delta; d_i) \right\} \\
&\leq \mathbb{E}\exp\left\{ \frac{\lambda}{r_0 l_1} \sup_{\delta \in \mathbb{B}_\Sigma(r_0)\cap\mathbb{C}(l_1)} \sum_{i=1}^n \epsilon_i \langle X_i, \delta\rangle \right\} \leq \mathbb{E}\exp\left\{ \lambda \left\| \sum_{i=1}^n \epsilon_i X_i \right\|_\infty \right\},
\end{aligned}$$

where $\epsilon_1, \ldots, \epsilon_n$ are independent Rademacher random variables. Let $\widetilde{X}_i := (X_i^{\mathrm{T}}, -X_i^{\mathrm{T}})^{\mathrm{T}} = (\widetilde{x}_{i,1}, \ldots, \widetilde{x}_{i,2p})^{\mathrm{T}} \in \mathbb{R}^{2p}$. Then $\|\sum_{i=1}^n \epsilon_i X_i\|_\infty = \max_{1\leq j\leq 2p}\sum_{i=1}^n \epsilon_i \widetilde{x}_{ij}$. Since ϵ_i is symmetric, $\mathbb{E}(\epsilon_i x_{ij})^k = 0$ if k is odd; $\mathbb{E}(\epsilon_i x_{ij})^k = \mathbb{E}x_{ij}^k \leq B_X^{k-2}\overline{\kappa}^2 \leq (k!/2)\overline{\kappa}^2 B_X^{k-2}$ if k is even. Therefore, similar to the proof of Bernstein's inequality (e.g., Theorem 2.10 in Boucheron et al. (2013)), we have that for any $\lambda \in (0, 1/B_X)$,

$$\log \mathbb{E}\exp\left\{ \lambda \sum_{i=1}^n \epsilon_i \widetilde{x}_{ij} \right\} \leq \frac{n\overline{\kappa}^2\lambda^2}{2(1 - B_X\lambda)}. \tag{66}$$

This further implies that

$$\log \mathbb{E}e^{\lambda\Lambda(r_0,l_1)} \leq \log \mathbb{E}\sum_{j=1}^{2p} \exp\left\{ \frac{\lambda}{n} \sum_{i=1}^n \epsilon_i \widetilde{x}_{ij} \right\} \leq \log(2p) + \frac{n\overline{\kappa}^2\lambda^2}{2(1 - B_X\lambda)}.$$

Thus, for any $t > 0$, by the above bound and equation (2.5) in Boucheron et al. (2013), we have

$$\begin{aligned}\sup_{\lambda>0}\{\lambda t - \log \mathbb{E}e^{\lambda\Lambda(r_0,l_1)}\} &\geq \sup_{0<\lambda<1/B_X}\{\lambda t - \log \mathbb{E}e^{\lambda A_j(r_1)}\}\\ &\geq -\log(2p) + \sup_{0<\lambda<1/B_X}\left\{\lambda t - \frac{n\overline{\kappa}^2\lambda^2}{2(1-B_X\lambda)}\right\}\\ &= -\log(2p) + \frac{n\overline{\kappa}^2}{B_X^2}h_1\left(\frac{B_X t}{n\overline{\kappa}^2}\right), \end{aligned} \tag{67}$$

where $h_1(u) = 1+u-\sqrt{1+2u}$ is an increasing function from $(0,\infty)$ onto $(0,\infty)$ with inverse function $h_1^{-1}(u) = u + \sqrt{2u}$ for all $u > 0$. Substituting this into (65) gives that

$$\mathbb{P}\{\Lambda(r_0,l_1) \geq B_X u + \overline{\kappa}\sqrt{2nu}\} \leq 2pe^{-u}.$$

for any $u > 0$. Then, taking $u = \log(2p) + t$ yields the result. ■

C.2 Proof of Lemma 14

For each $1 \leq j \leq p$, $|\mathbb{1}(\varepsilon_i \leq 0)| \cdot |x_{ij}| \leq \bar{\alpha}B_X$ and $\mathbb{E}\{(\mathbb{1}(\varepsilon_i \leq 0) - \alpha)^2 x_{ij}^2\} \leq \alpha(1-\alpha)\overline{\kappa}^2$. Then claimed bound follows from Bernstein's inequality and the union bound. ■

C.3 Proof of Lemma 15

Note that $\|(1/n)\sum_{i=1}^n(1-\mathbb{E})\{\psi_\tau(\xi_i)X_i\}\|_\infty = \max_{1\leq j\leq p}|(1/n)\sum_{i=1}^n(1-\mathbb{E})\{\psi_\tau(\xi_i)x_{ij}\}|$. Since $|\psi_\tau(\xi_i)| \leq \tau$ and $\mathbb{E}_{X_i}\{\psi_\tau^2(\xi_i)\} \leq \mathbb{E}_{X_i}(\xi_i^2) = \mathbb{E}_{X_i}(\varepsilon_{-,i}^2) - \alpha^2(X^{\mathrm{T}}\beta^* - X^{\mathrm{T}}\theta^*)^2 \leq \overline{\sigma}^2$, we have $\mathbb{E}[\psi_\tau(\xi_i)x_{ij}]^2 \leq \overline{\sigma}^2\overline{\kappa}^2$. Because $\|X\|_\infty \leq B_X$, for $k \geq 3$, $\mathbb{E}|\psi_\tau(\xi_i)x_{ij}|^k \leq \overline{\sigma}^2\overline{\kappa}^2(\tau B_X)^{k-2} \leq \frac{k!}{2}\overline{\sigma}^2\overline{\kappa}^2(\tau B_X)^{k-2}$. Thus by Bernstein's inequality,

$$\left|\frac{1}{n}\sum_{i=1}^n(1-\mathbb{E})\{\psi_\tau(\xi_i)x_{ij}\}\right| \leq \overline{\sigma}\,\overline{\kappa}\sqrt{\frac{2u}{n}} + \tau B_X\frac{u}{n}$$

with probability at least $1 - 2e^{-u}$. Then by taking union bound over $j = 1, \cdots, p$ and letting $u = \log(2p) + t$, we obtain the claimed bound. ■

C.4 Proof of Lemma 16

Let $\delta = \beta - \beta^*$, and let $\phi(t) = t\mathbb{1}(t \leq 0)$. Then $\phi(t)$ is a 1-Lipschitz continuous function. Denote $s_i(\delta) := \xi_i(\beta) = \phi(\varepsilon_i - X_i^{\mathrm{T}}\delta) + \alpha X_i^{\mathrm{T}}\delta + \alpha X_i^{\mathrm{T}}(\beta^* - \theta^*)$. Then $s_i(0) = \phi(\varepsilon_i) + \alpha X_i^{\mathrm{T}}(\beta^* - \theta^*)$. For each $1 \leq j \leq p$, define the random process $R_j(\delta) = (1/n)\sum_{i=1}^n(1-\mathbb{E})r_j(\delta;\varepsilon_i,X_i)$, where $r_j(\delta;\varepsilon_i,X_i) = \{\psi_\tau(s_i(\delta)) - \psi_\tau(s_i(0))\}x_{ij}$. We aim to upper bound $\sup_{\delta\in\mathbb{B}_1(r_1)}|R_j(\delta)|$ for each j first and then take the union bound.

Let $u_i = \langle X_i, \delta \rangle$. Denote $\frac{\partial r_j}{\partial u_i}$ as the derivative of $r_j(\delta; \varepsilon_i, X_i)$ with respect to u_i. Then, since $|\psi_\tau'(t)| = \mathbb{1}(|t| \le \tau) \le 1$ almost surely for all $t \in \mathbb{R}$, we have

$$\left|\frac{\partial r_j}{\partial u_i}\right| = |x_{ij} \cdot \psi_\tau'(s_i(\delta)) \cdot (\alpha - \mathbb{1}(\varepsilon_i \le u_i))| \le |x_{ij}| \le B_X.$$

Therefore, we have $|r_j(\delta; \varepsilon_i, X_i) - r_j(\delta'; \varepsilon_i, X_i)| \le B_X |\langle X_i, \delta \rangle - \langle X_i, \delta' \rangle|$ for $\delta \ne \delta'$.

Now we control $\sup_{\delta \in \mathbb{B}_1(r_1)} |R_j(\delta)|$. Given $r_1 > 0$, define the random variable

$$A_j(r_1) = \frac{n}{4 B_X r_1} \sup_{\delta \in \mathbb{B}_1(r_1)} |R_j(\delta)|.$$

By Chernoff's bound,

$$\mathbb{P}(A_j(r_1) \ge t) \le \exp\left[- \sup_{\lambda > 0} \{\lambda t - \log \mathbb{E} e^{\lambda A_j(r_1)}\} \right]. \tag{68}$$

For $\mathbb{E} e^{\lambda A_j(r_1)}$, using Rademacher symmetrization and applying Ledoux-Talagrand contraction inequality (e.g., Theorem 4.12 in Ledoux and Talagrand, 1991) gives

$$\begin{aligned}
\mathbb{E} e^{\lambda A_j(r_1)} &\le \mathbb{E} \exp\left\{ \frac{2\lambda}{4 B_X r_1} \sup_{\delta \in \mathbb{B}_1(r_1)} \sum_{i=1}^n \epsilon_i r_j(\delta; \varepsilon_i, X_i) \right\} \\
&\le \mathbb{E} \exp\left\{ \frac{\lambda}{r_1} \sup_{\delta \in \mathbb{B}_1(r_1)} \sum_{i=1}^n \epsilon_i \langle X_i, \delta \rangle \right\} \le \mathbb{E} \exp\left\{ \lambda \left\| \sum_{i=1}^n \epsilon_i X_i \right\|_\infty \right\},
\end{aligned}$$

where $\epsilon_1, \cdots, \epsilon_n$ are independent Rademacher random variables. Let $\widetilde{X}_i = (X_i^{\mathrm{T}}, -X_i^{\mathrm{T}})^{\mathrm{T}} = (\widetilde{x}_{i,1}, \cdots, \widetilde{x}_{i,2p})^{\mathrm{T}} \in \mathbb{R}^{2p}$. Then by (66),

$$\log \mathbb{E} e^{\lambda A_j(r_1)} \le \log \mathbb{E} \sum_{j=1}^{2p} \exp\left\{ \lambda \sum_{i=1}^n \epsilon_i \widetilde{x}_{ij} \right\} \le \log(2p) + \frac{n \overline{\kappa}^2 \lambda^2}{2(1 - B_X \lambda)}.$$

Meanwhile, similar to (67), we have

$$\sup_{\lambda > 0} \{\lambda t - \log \mathbb{E} e^{\lambda A_j(r_1)}\} \ge - \log(2p) + \frac{n \overline{\kappa}^2}{B_X^2} h_1\left(\frac{B_X t}{n \overline{\kappa}^2} \right),$$

Substituting this into (68) gives

$$\mathbb{P}\{A_j(r_1) \ge t\} \le \exp\left\{ \log(2p) - \frac{n \overline{\kappa}^2}{B_X^2} h_1\left(\frac{B_X t}{n \overline{\kappa}^2} \right) \right\}.$$

Using h^{-1}, we have $\mathbb{P}\{A_j(r_1) \ge B_X u + \overline{\kappa} \sqrt{2nu}\} \le 2p e^{-u}$. Finally, taking the union bound and letting $u = \log(2p^2) + t$ establishes the claim. ■

C.5 Proof of Lemma 17

Note that $\|\mathbb{E}[\psi_\tau(\xi_i(\beta))U_i]\|_2 = \sup_{v\in\mathbb{S}^{p-1}} |\mathbb{E}\{\psi_\tau(\xi_i(\beta))U_i^{\mathrm{T}}v|$. Recall that the conditional CDF $F = F_{\varepsilon_i|X_i}$ of ε_i given X_i is continuously differentiable with $f = F'$. Let $\mathbb{E}_{X_i}$ be the conditional expectation given X_i. For $\beta \in \mathbb{R}^p$, let $u_i := X_i^{\mathrm{T}}(\beta - \beta^*)$, and

$$\begin{aligned}\Psi(\beta) &= \mathbb{E}_{X_i}\psi_\tau(\xi_i(\beta)) \\ &= \int_{-\infty}^{u_i} \psi_\tau(t - u_i + \alpha X_i^{\mathrm{T}}(\beta^* - \theta^*))f(t)\,\mathrm{d}t + \int_{u_i}^{\infty} \psi_\tau(\alpha X^{\mathrm{T}}(\beta - \theta^*))f(t)\,\mathrm{d}t.\end{aligned}$$

Then we have

$$|\mathbb{E}\{\psi_\tau(\xi_i(\beta))U_i^{\mathrm{T}}v\}| \le |\mathbb{E}\{(\Psi(\beta) - \Psi(\beta^*))U_i^{\mathrm{T}}v\}| + |\mathbb{E}\{\psi_\tau(\xi_i)U_i^{\mathrm{T}}v\}|. \tag{69}$$

Since $\psi_\tau(\cdot)$ is absolutely continuous and has a derivative $\psi_\tau'(t) = \mathbb{1}(|t| \le \tau)$ almost everywhere, by the fundamental theorem of calculus,

$$\Psi(\beta) - \Psi(\beta^*) = \int_0^1 \langle \nabla\Psi(\beta^* + t(\beta - \beta^*)), \beta - \beta^*\rangle\,\mathrm{d}t,$$

where

$$\begin{aligned}\nabla\Psi(\beta) &= \int_{-\infty}^{u_i} \psi_\tau'(t - u_i + \alpha X_i^{\mathrm{T}}(\beta - \theta^*))f(t)\,\mathrm{d}t \cdot (\alpha - 1)X_i + \psi_\tau(\alpha X_i^{\mathrm{T}}(\beta - \theta^*))f(u_i)X_i \\ &\qquad + \psi_\tau'(\alpha X_i^{\mathrm{T}}(\beta - \theta^*))\{1 - F(u_i)\} \cdot \alpha X_i - \psi_\tau(\alpha X_i^{\mathrm{T}}(\beta - \theta^*))f(u_i)X_i \\ &= \int_{-\infty}^{u_i} \mathbb{1}(|t - u_i + \alpha X_i^{\mathrm{T}}(\beta - \theta^*)| \le \tau)f(t)\,\mathrm{d}t \cdot (\alpha - 1)X_i \\ &\qquad + \mathbb{1}(|\alpha X_i^{\mathrm{T}}(\beta - \theta^*)| \le \tau) \cdot \{1 - F(u_i)\} \cdot \alpha X_i \\ &= (\alpha - 1)F(u_i)X_i - \mathbb{E}_{X_i}\{\mathbb{1}(|\xi_i(\beta)| > \tau)\mathbb{1}(\varepsilon_i \le u_i) \cdot (\alpha - 1)X_i\} + \alpha X_i \\ &\qquad - \alpha X_i \cdot \mathbb{E}_{X_i}\mathbb{1}(|\xi_i(\beta)| > \tau) - \alpha X_i F(u_i) + \alpha X_i \cdot \mathbb{E}_{X_i}\{\mathbb{1}(\xi_i(\beta) > \tau)\mathbb{1}(\varepsilon_i \le u_i)\} \\ &= \{\alpha - F(u_i)\}X_i + \mathbb{E}_{X_i}[\mathbb{1}(|\xi_i(\beta)| > \tau)\{\mathbb{1}(\varepsilon_i \le u_i) - \alpha\}X_i].\end{aligned} \tag{70}$$

For $t \in [0, 1]$, define $\beta_t = \beta^* + t(\beta - \beta^*)$. Then $X_i^{\mathrm{T}}(\beta_t - \beta^*) = tu_i$ and we have

$$\langle \nabla\Psi(\beta_t), \beta - \beta^*\rangle = \{\alpha - F(tu_i)\}u_i + \mathbb{E}_{X_i}[\mathbb{1}(|\xi_i(\beta_t)| > \tau)\{\mathbb{1}(\varepsilon_i \le tu_i) - \alpha\}u_i].$$

By condition 3, $|\alpha - F(tu_i)| = |F(0) - F(tu_i)| \le \overline{f} \cdot t|u_i|$. Moreover, by Markov's inequality,

$$\mathbb{E}_{X_i}[\mathbb{1}(|\xi_i(\beta_t)| > \tau)\{\mathbb{1}(\varepsilon_i \le tu_i) - \alpha\}] \le \frac{1 - \alpha}{\tau}\mathbb{E}_{X_i}|\xi_i(\beta_t)|. \tag{71}$$

Note that for any $\beta \in \mathbb{R}^p$, $\xi_i(\beta) = Z_i(\beta) - \alpha X^{\mathrm{T}}\theta^* = (\varepsilon_i - u_i)\mathbb{1}(\varepsilon_i \le u_i) + \alpha X_i^{\mathrm{T}}(\beta - \theta^*)$. Thus we have

$$\begin{aligned}|\xi_i(\beta_t)| &\le |\xi_i(\beta_t) - \xi_i| + |\xi_i| \\ &\le |\xi_i| + |(\varepsilon_i - tu_i)\mathbb{1}(\varepsilon_i \le tu_i) - \varepsilon_i\mathbb{1}(\varepsilon_i \le 0) + \alpha tu_i| \\ &\le |\xi_i| + \begin{cases} |\varepsilon_i\mathbb{1}(0 \le \varepsilon_i \le tu_i) + tu_i[\alpha - \mathbb{1}(\varepsilon_i \le tu_i)]| & \text{if } u_i \ge 0 \\ |tu_i[\alpha - \mathbb{1}(\varepsilon_i \le tu_i)] - \varepsilon_i\mathbb{1}(0 \le \varepsilon_i \le tu_i)| & \text{if } u_i < 0 \end{cases} \\ &\le |\xi_i| + t|u_i|.\end{aligned} \tag{72}$$

Therefore, under Condition 3, $\mathbb{E}_{X_i}|\xi_i(\beta_t)| \le t|u_i| + \mathbb{E}_{X_i}|\xi_i| \le t|u_i| + \overline{\sigma}$. Substituting this into (71), and putting together the pieces, we obtain that for any $\beta \in \mathbb{B}_\Sigma(\beta^*, r_0)$

$$\begin{aligned}
|\mathbb{E}\{(\Psi(\beta) - \Psi(\beta^*))U_i^{\mathrm{T}} v\}| &\le \int_0^1 \mathbb{E}|\langle \nabla\Psi(\beta_t), \beta - \beta^*\rangle||U_i^{\mathrm{T}} v| \,\mathrm{d}t \\
&\le \int_0^1 \mathbb{E}\{[\,\overline{f} \cdot t u_i^2 + \tau^{-1}(\overline{\sigma} + t|u_i|) \cdot |u_i|] \cdot |U_i^{\mathrm{T}} v|\} \,\mathrm{d}t \\
&\le \frac{1}{2}(\overline{f} + \tau^{-1})\kappa_3 \cdot r_0^2 + (\overline{\sigma}/\tau) \cdot r_0. \qquad (73)
\end{aligned}$$

On the other hand, since $\mathbb{E}_{X_i}(\xi_i) = 0$, $\mathbb{E}_{X_i}(\xi_i^2) \le \overline{\sigma}^2$ and $|\psi_\tau(t) - t| = (|t| - \tau)\mathbb{1}(|t| > \tau) \le \tau^{-1}t^2$, then

$$\mathbb{E}\{\psi_\tau(\xi_i)U_i^{\mathrm{T}} v\} \le \mathbb{E}\{|\psi_\tau(\xi_i) - \xi_i||U_i^{\mathrm{T}} v|\} \le (1/\tau) \cdot \mathbb{E}(\xi_i^2|U_i^{\mathrm{T}} v|) \le \overline{\sigma}^2/\tau. \qquad (74)$$

Substituting (73) and (74) into (69) proves the claimed bound with $b(\tau) = \overline{\sigma}^2/\tau$.

For the second part, first note that if $\mathbb{E}_X|\varepsilon_-|^q$ exists, since $\xi_i = \varepsilon_{-,i} - \mathbb{E}_{X_i}(\varepsilon_{-,i})$ with $\varepsilon_{-,i} \le 0$, we have $|\xi_i|^q \le \max\{|\varepsilon_{-,i}|^q, |\mathbb{E}_{X_i}(\varepsilon_{-,i})|^q\} \le \max\{|\varepsilon_{-,i}|^q, \mathbb{E}_{X_i}|\varepsilon_{-,i}|^q\}$. Therefore, $\mathbb{E}_{X_i}|\xi_i|^q \le \mathbb{E}_{X_i}|\varepsilon_-|^q \le \alpha_q$. Similar to (71), by Markov's inequality and Jensen's inequality that $\mathbb{E}_{X_i}|\xi_i| \le (\mathbb{E}_{X_i}|\xi_i|^q)^{1/q}$, we obtain that

$$\mathbb{E}_{X_i}[\mathbb{1}(|\xi_i(\beta_t)| > \tau)\{\mathbb{1}(\varepsilon_i \le t u_i) - \alpha\}] \le \frac{1-\alpha}{\tau}\mathbb{E}_{X_i}|\xi_i(\beta_t)| \le \frac{1-\alpha}{\tau}\{\alpha_q^{1/q} + t|u_i|\},$$

where the second last inequality uses (72). Thus,

$$\begin{aligned}
|\mathbb{E}\{(\Psi(\beta) - \Psi(\beta^*))U_i^{\mathrm{T}} v\}| &\le \int_0^1 \mathbb{E}|\langle \nabla\Psi(\beta_t), \beta - \beta^*\rangle||U_i^{\mathrm{T}} v| \,\mathrm{d}t \\
&\le \int_0^1 \mathbb{E}[\{\,\overline{f} \cdot t u_i^2 + \{\alpha_q^{1/q} + t|u_i|\} \cdot |u_i|/\tau\} \cdot |U_i^{\mathrm{T}} v|] \,\mathrm{d}t \\
&\le \frac{1}{2}\{\overline{f} + (1/\tau)\}\kappa_3 \cdot r_0^2 + \{\alpha_q^{1/q}/\tau\} \cdot r_0. \qquad (75)
\end{aligned}$$

Because $|\psi_\tau(t) - t| \le t^{q+1}/\tau^q$ for all $q \ge 1, q \in \mathbb{N}$, similar to (74),

$$\mathbb{E}\{\psi_\tau(\xi_i)U_i^{\mathrm{T}} v\} \le \mathbb{E}\{|\psi_\tau(\xi_i) - \xi_i| \cdot |U_i^{\mathrm{T}} v|\} \le \mathbb{E}(|\xi_i|^q|U_i^{\mathrm{T}} v|)/\tau^{q-1} \le \alpha_q/\tau^{q-1}. \qquad (76)$$

Combining (69), (75) and (76) yields the second claim. ■

C.6 Proof of Lemma 18

Note that for any $\alpha\Delta \in \mathbb{B}_\Sigma(r) \cap \mathbb{C}(l)$ and $\Delta' \in \mathbb{B}_\Sigma(r_0) \cap \mathbb{C}(l_1)$, when $\tau \ge 2.5 B_X \max\{lr, l_1 r_0\}$

$$|\alpha X_i^{\mathrm{T}}\Delta| \le B_X l \alpha \|\Delta\|_\Sigma \le \frac{2\tau}{5} \quad \text{and} \quad |X_i^{\mathrm{T}}\Delta'| \le B_X l_1 \|\Delta'\|_\Sigma \le \frac{2\tau}{5}.$$

By the proof of Lemma 17, we have $|\xi_i(\beta)| = |Z_i(\beta) - \alpha X_i^{\mathrm{T}}\theta^*| \le |\xi_i| + |X_i^{\mathrm{T}}\Delta'|$. Consequently,

$$|Z_i(\beta) - \alpha X_i^{\mathrm{T}}\theta| \le |\xi_i(\beta)| + |\alpha X_i^{\mathrm{T}}\Delta| \le |\xi_i| + \frac{4\tau}{5}.$$

Therefore, conditioned on events $\{\Delta' \in \mathbb{B}_\Sigma(\beta^*, r_0) \cap \mathbb{C}(l_1)\}$ and $\{\Delta \in \mathbb{B}_\Sigma(\theta^*, r/\alpha) \cap \mathbb{C}(l)\}$, we have

$$\begin{aligned}
&\langle \nabla_\theta \widehat{\mathcal{L}}_\tau(\beta, \theta) - \nabla_\theta \widehat{\mathcal{L}}_\tau(\beta, \theta^*), \theta - \theta^* \rangle \\
&= \frac{\alpha}{n} \sum_{i=1}^n \{\psi_\tau(Z_i(\beta) - \alpha X_i^{\mathrm{T}} \theta^*) - \psi_\tau(Z_i(\beta) - \alpha X_i^{\mathrm{T}} \theta)\} X_i^{\mathrm{T}} (\theta - \theta^*) \\
&\geq \frac{\alpha}{n} \sum_{i=1}^n \{\psi_\tau(Z_i(\beta) - \alpha X_i^{\mathrm{T}} \theta^*) - \psi_\tau(Z_i(\beta) - \alpha X_i^{\mathrm{T}} \theta)\} X_i^{\mathrm{T}} (\theta - \theta^*) \mathbb{1}\{|\xi_i| \leq \tau/5\} \\
&= \frac{\alpha^2}{n} \sum_{i=1}^n \langle X_i, \theta - \theta^* \rangle^2 \mathbb{1}\{|\xi_i| \leq \tau/5\}.
\end{aligned}$$

Let $\delta = \Delta/\|\Delta\|_\Sigma$ so that $\delta \in \mathbb{S}_\Sigma^{p-1} \cap \mathbb{C}(l)$. Denote $r_\delta(w_i) = (X^{\mathrm{T}} \delta)^2 \mathbb{1}\{|\xi_i| \leq \tau/5\}$, where $w_i = (X_i^{\mathrm{T}}, \varepsilon_i)^{\mathrm{T}}$ and $R_n(\delta) := n^{-1} \sum_{i=1}^n r_\delta(w_i)$. Then we have

$$\langle \nabla_\theta \widehat{\mathcal{L}}_\tau(\beta, \theta) - \nabla_\theta \widehat{\mathcal{L}}_\tau(\beta, \theta^*), \theta - \theta^* \rangle \geq (\alpha \|\Delta\|_\Sigma)^2 R_n(\delta). \tag{77}$$

Next, we derive lower bounds for $\mathbb{E} R_n(\delta)$ and $R_n(\delta) - \mathbb{E} R_n(\delta)$ respectively.

First, note that by Markov's inequality,

$$\mathbb{E} R_n(\delta) = \|\delta\|_\Sigma^2 - \mathbb{E}\{\mathbb{1}(|\xi_i| > \tau/5)(X_i^{\mathrm{T}} \delta)^2\} \geq \|\delta\|_\Sigma^2 \{1 - (5\overline{\sigma}/\tau)^2\} \geq \frac{1}{2} \tag{78}$$

where the last inequality holds if $\tau \geq \sqrt{50}\overline{\sigma}$.

Next, we consider the upper bound of

$$\Gamma_n := \sup_{\delta \in \mathbb{S}_\Sigma^{p-1} \cap \mathbb{C}(l)} \{\mathbb{E} R_n(\delta) - R_n(\delta)\} = \sup_{\delta \in \mathbb{S}_\Sigma^{p-1} \cap \mathbb{C}(l)} \frac{1}{n} \sum_{i=1}^n \{\mathbb{E} r_\delta(w_i) - r_\delta(w_i)\}.$$

Note that $0 \leq r_\delta(w_i) \leq (B_X l)^2$ for all $\delta \in \mathbb{S}_\Sigma^{p-1} \cap \mathbb{C}(l)$. Since w_i's are independent, it follows from McDiarmid's inequality that with probability at least $1 - e^{-t}$,

$$\Gamma_n \leq \mathbb{E}\Gamma_n + (B_X l)^2 \sqrt{\frac{t}{2n}}. \tag{79}$$

To bound $\mathbb{E}\Gamma_n$, by Rademacher symmetrization we have

$$\mathbb{E}\Gamma_n \leq 2\mathbb{E}\left\{ \sup_{\delta \in \mathbb{S}_\Sigma^{p-1} \cap \mathbb{C}(l)} \frac{1}{n} \sum_{i=1}^n \epsilon_i r_\delta(w_i) \right\}$$

where $\epsilon_1, \ldots, \epsilon_n$ are i.i.d. Rademacher random variables. Since for $\delta, \delta' \in \mathbb{S}_\Sigma^{p-1} \cap \mathbb{C}(l)$,

$$|r_\delta(w_i) - r_{\delta'}(w_i)| \leq 2B_X l |X_i^{\mathrm{T}} \delta - X_i^{\mathrm{T}} \delta'|,$$

$r_\delta(w_i)$ is $(2B_X l)$-Lipshitz in $X_i^{\mathrm{T}}\delta$. By Ledoux-Talagrand contraction inequality (see, e.g., Ledoux and Talagrand, 1991), we obtain that

$$\begin{aligned}\mathbb{E}\Gamma_n &\le 2\mathbb{E}\bigg\{\sup_{\delta\in\mathbb{S}_\Sigma^{p-1}\cap\mathbb{C}(l)}\frac{1}{n}\sum_{i=1}^n \epsilon_i r_\delta(w_i)\bigg\}\\ &\le 4B_X l\mathbb{E}\bigg\{\sup_{\delta\in\mathbb{S}_\Sigma^{p-1}\cap\mathbb{C}(l)}\frac{1}{n}\sum_{i=1}^n \epsilon_i X_i^{\mathrm{T}}\delta\bigg\} \le 4B_X l^2\mathbb{E}\bigg\|\frac{1}{n}\sum_{i=1}^n \epsilon_i X_i\bigg\|_\infty.\end{aligned}$$

By the proof of 13, for any $\lambda\in(0,1/B_X)$,

$$\log\mathbb{E}\exp\left(\lambda\sum_{i=1}^n\epsilon_i x_{ij}\right)\le\frac{n\overline{\kappa}^2\lambda^2}{2(1-B_X\lambda)},$$

Then similar to the proof of 13, for any $\lambda\in(0,1/B_X)$,

$$\log\mathbb{E}\exp\bigg\|\lambda\sum_{i=1}^n\epsilon_i X_i\bigg\|_\infty\le\log(2p)+\frac{n\overline{\kappa}^2\lambda^2}{2(1-B_X\lambda)}.$$

Therefore,

$$\begin{aligned}\mathbb{E}\bigg\|\frac{1}{n}\sum_{i=1}^n\epsilon_i X_i\bigg\|_\infty &\le\frac{1}{n}\inf_{\lambda>0}\frac{1}{\lambda}\log\mathbb{E}\exp\left\{\lambda\bigg\|\sum_{i=1}^n\epsilon_i X_i\bigg\|_\infty\right\}\\ &\le\frac{1}{n}\inf_{0<\lambda<1/B_X}\left\{\frac{\log(2p)}{\lambda}+\frac{n\overline{\kappa}^2\lambda}{2(1-B_X\lambda)}\right\}\\ &=B_X\frac{\log(2p)}{n}+\overline{\kappa}\sqrt{\frac{2\log(2p)}{n}}.\end{aligned}$$

Substituting this into (79), we have that with probability at least $1-e^{-t}$,

$$\Gamma_n\le 4(B_X l)^2\frac{\log(2p)}{n}+4\overline{\kappa}B_X l^2\sqrt{\frac{2\log(2p)}{n}}+(B_X l)^2\sqrt{\frac{t}{2n}}\le\frac{1}{4},$$

as long as $n\gtrsim(\overline{\kappa}\vee B_X)^2B_X^2l^4\{\log(2p)+t\}$. Together with (78), we have that with probability at least $1-e^{-t}$,

$$R_n(\delta)=\mathbb{E}R_n(\delta)+\{R_n(\delta)-\mathbb{E}R_n(\delta)\}\ge\frac{1}{2}-\frac{1}{4}=\frac{1}{4}$$

holds uniformly over $\delta\in\mathbb{S}_\Sigma^{p-1}\cap\mathbb{C}(l)$. Substituting this into (77) establishes the claim. ∎

C.7 Proof of Lemma 20

For any $t>0$, let $\mathcal{A}$ be the event that $\{\|\Sigma-\widehat{\Sigma}\|_{\max}\lesssim\sqrt{(\log(p)+t)/n}\}$. Note that

$$\|\Sigma-\widehat{\Sigma}\|_{\max}=\max_{j,k}\bigg|\frac{1}{n}\sum_{i=1}^n(1-\mathbb{E})(x_{ij}x_{ik})\bigg|.$$

Since $\mathbb{E}(\widetilde{x}_{ij}\widetilde{x}_{ik})^2 \leq B_X^2\overline{\kappa}^2$ and for any $k \geq 3$, $\mathbb{E}|x_{ij}x_{ik}|^k \leq B_X^{k-2}\overline{\kappa}^2 \leq k!\overline{\kappa}^2 B_X^{k-2}/2$, by Bernstein's inequality, $|n^{-1}\sum_{i=1}^n(1-\mathbb{E})(x_{ij}x_{ik})| \leq \overline{\kappa}\sqrt{(2u)/n} + B_X(u/n)$ with probability at least $1-2e^{-u}$. Taking the union bound over $j,k=1,\ldots,p$ and letting $u=\log(4p)+t$, we obtain that $\mathbb{P}(\mathcal{A}) \geq 1-e^{-t}$.

Proof of 1): Let $u^* := \Sigma^{-1}a$. Since $C_a\|a\|_2 \geq \|\Sigma^{-1}a\|_1$, u^* satisfies (16). Next, conditioning on event $\mathcal{A}$, we have

$$\begin{aligned}\|a-\widehat{\Sigma}u^*\|_\infty &\leq \|a-\Sigma u^*\|_\infty + \|(\Sigma-\widehat{\Sigma})u^*\|_\infty = \|(\Sigma-\widehat{\Sigma})u^*\|_\infty \\ &\leq \|\Sigma-\widehat{\Sigma}\|_{\max}\|u^*\|_1 \lesssim C_a\|a\|_2\sqrt{\frac{\log(p)+t}{n}}.\end{aligned}$$

Therefore u^* satisfies (15) when $\rho \asymp \sqrt{\log(p)/n}$. For (17), note that conditioned on $\mathcal{A}$,

$$|a^{\mathrm{T}}\widehat{\Sigma}u - \|a\|_2^2| \leq \|a\|_1 \cdot \|a-\widehat{\Sigma}u^*\|_\infty \lesssim \|a\|_1\|a\|_2 C_a\sqrt{\frac{\log(p)+t}{n}}$$

Thus as long as $\|a\|_1/\|a\|_2 = o(\sqrt{n/\log p})$, u^* also satisfies (17) with $\rho'=o(1)$. Hence, u^* is in the constraint set with probability at least $1-e^{-t}$.

Proof of 2): By constraint (17), $|\|a\|_2^2 - a^{\mathrm{T}}\widehat{\Sigma}\widehat{u}| \leq \rho'\|a\|_2^2$. Applying triangle inequality and re-arranging terms, we have

$$|\|a\|_2^2 - a^{\mathrm{T}}\Sigma\widehat{u}| \leq |\|a\|_2^2 - a^{\mathrm{T}}\widehat{\Sigma}\widehat{u}| + |a^{\mathrm{T}}(\Sigma-\widehat{\Sigma})\widehat{u}| \leq \rho'\|a\|_2^2 + \|(\Sigma-\widehat{\Sigma})a\|_\infty\|\widehat{u}\|_1.$$

Note that

$$\|(\Sigma-\widehat{\Sigma})a\|_\infty = \left\|\frac{1}{n}\sum_{i=1}^n(1-\mathbb{E})\{(X_i^{\mathrm{T}}a)x_{ij}\}\right\|_\infty = \max_{1\leq j\leq p}\left|\frac{1}{n}\sum_{i=1}^n(1-\mathbb{E})\{(X_i^{\mathrm{T}}a)x_{ij}\}\right|.$$

Since $\mathbb{E}\{(X_i^{\mathrm{T}}a)x_{ij}\}^2 \leq B_X^2\overline{\kappa}^2\|a\|_2^2$ and for $k\geq 3$,

$$\mathbb{E}|(X_i^{\mathrm{T}}a)\widetilde{x}_{ij}|^k \leq (B_X^2\|a\|_1)^{k-2}\mathbb{E}\{(X_i^{\mathrm{T}}a)x_{ij}\}^2 \leq (B_X^2\|a\|_1)^{k-2}\cdot(B_X^2\overline{\kappa}^2\|a\|_2^2).$$

Then, by Bernstein's inequality and the union bound, with probability at least $1-e^{-t}$,

$$\|(\Sigma-\widehat{\Sigma})a\|_\infty \lesssim B_X\overline{\kappa}\|a\|_2\sqrt{\frac{\log(2p)+t}{n}} + B_X^2\|a\|_1\frac{\log(2p)+t}{n}.$$

Denote $v := a/\|a\|_2$. Since $\rho'=o(1)$ and a satisfies $\|a\|_1/\|a\|_2 = o(\sqrt{n/\log p})$, then together with (16), with probability at least $1-e^{-t}$,

$$|\|a\|_2 - v^{\mathrm{T}}\Sigma\widehat{u}| \lesssim \rho'\|a\|_2 + C_a\|a\|_2 B_X\overline{\kappa}\sqrt{\frac{\log(p)+t}{n}} = o(\|a\|_2).$$

Finally, combining with the fact that $\underline{\kappa}^2\|\widehat{u}\|_2 \leq v^{\mathrm{T}}\Sigma\widehat{u} \leq \overline{\kappa}^2\|\widehat{u}\|_2$, we obtain desired bounds.

Proof of 3): The proof is similar to Lemma 1 in Cai et al. (2021). Recall that $s^2(\widehat{u}) = \widehat{u}^{\mathrm{T}}\widetilde{\Lambda}\widehat{u}$, where $\widetilde{\Lambda} = (1/n)\sum_{i=1}^n \mathbb{E}_{X_i}(\xi_i^2)X_iX_i^{\mathrm{T}}$. First we show an upper bound for $s(\widehat{u})$. By the proof above, for any $t>0$, $u^*\in\mathcal{U}$ with probability at least $1-e^{-t}$. Then by the optimality of $\widehat{u}$,

$$s^2(\widehat{u}) \leq \overline{\sigma}^2\widehat{u}^{\mathrm{T}}\widehat{\Sigma}\widehat{u} \leq \overline{\sigma}^2(u^*)^{\mathrm{T}}\widehat{\Sigma}u^* = \overline{\sigma}^2(u^*)^{\mathrm{T}}\Sigma u^* + \overline{\sigma}^2(u^*)^{\mathrm{T}}(\widehat{\Sigma}-\Sigma)u^*.$$

Note that $(u^*)^{\mathrm{T}}\Sigma u^* = a^{\mathrm{T}}\Sigma^{-1}a \le \underline{\kappa}^{-2}\|a\|_2^2$. Meanwhile, conditioned on $\mathcal{A}$, we have $(u^*)^{\mathrm{T}}(\widehat{\Sigma}-\Sigma)u^* \le \|\widehat{\Sigma}-\Sigma\|_{\max}\|u^*\|_1^2 \lesssim (C_a\|a\|_2)^2\sqrt{(\log(2p)+t)/n}$. Combining the results above yields that $s^2(\widehat{u}) \le \underline{\kappa}^{-2}\overline{\sigma}^2\|a\|_2^2(1+o(1))$. with probability at least $1-e^{-t}$.

For lower bound, first note that by the second part of Lemma 20 and Hölder's inequality, when conditioned on event $\mathcal{A}$,

$$\begin{aligned}\widehat{u}^{\mathrm{T}}\widehat{\Sigma}\widehat{u} &= \widehat{u}^{\mathrm{T}}\Sigma\widehat{u} + \widehat{u}^{\mathrm{T}}(\widehat{\Sigma}-\Sigma)\widehat{u} \ge \underline{\kappa}^2\|\widehat{u}\|_2^2 - \|\widehat{\Sigma}-\Sigma\|_{\max}\|\widehat{u}\|_1^2 \\ &\gtrsim \underline{\kappa}^2(\overline{\kappa}^{-2}\|a\|_2)^2 - C_a^2\|a\|_2^2\sqrt{\frac{\log(2p)+t}{n}}.\end{aligned}$$

Hence $s(\widehat{u})^2 \ge \underline{\sigma}^2\widehat{u}^{\mathrm{T}}\widehat{\Sigma}\widehat{u} = \overline{\kappa}^{-4}\underline{\kappa}^2\underline{\sigma}^2\|a\|_2^2(1-o(1))$ with probability at least $1-e^{-t}$. ■

C.8 Proof of Lemma 21

Note that $\|\mathbb{E}\{g_{\beta,\theta}(w_i)U_i\}\|_2 = \sup_{v\in\mathbb{S}^{p-1}}\mathbb{E}\{g_{\beta,\theta}(w_i)U_i^{\mathrm{T}}v\}$. Let $\Psi(\beta,\theta) = \mathbb{E}_{X_i}\{\psi_\tau(\xi_i(\beta,\theta))\}$. Recall that $\Delta' = \beta-\beta^*$ and $\Delta = \theta-\theta^*$. Then

$$\begin{aligned}\mathbb{E}\{g_{\beta,\theta}(w_i)U_i^{\mathrm{T}}v\} &= \mathbb{E}[\{\Psi(\beta,\theta)-\Psi(\beta^*,\theta)\}U_i^{\mathrm{T}}v] \\ &\quad + \mathbb{E}[\{\Psi(\beta^*,\theta)-\Psi(\beta^*,\theta^*)+\alpha X_i^{\mathrm{T}}(\theta-\theta^*)\}U_i^{\mathrm{T}}v]. \qquad (80)\end{aligned}$$

For $\beta,\theta\in\mathbb{R}^p$, let $u_i = X_i^{\mathrm{T}}\Delta'$ and $v_i = X_i^{\mathrm{T}}\Delta$. By the fundamental theorem of calculus,

$$\Psi(\beta,\theta)-\Psi(\beta^*,\theta) = \int_0^1 \langle\nabla_\beta\Psi(\beta^*+t\Delta',\theta),\Delta'\rangle\,\mathrm{d}t,$$

where $\nabla_\beta\Psi(\beta,\theta) = \{\alpha-F(u_i)\}X_i + \mathbb{E}_{X_i}[\mathbb{1}(|\xi_i(\beta,\theta)|>\tau)\{\mathbb{1}(\varepsilon_i\le u_i)-\alpha\}X_i]$ by (70). For $t\in[0,1]$, define $\beta_t = \beta^*+t\Delta'$ so that $X_i^{\mathrm{T}}(\beta_t-\beta^*) = tu_i$, and we have

$$\langle\nabla_\beta\Psi(\beta_t,\theta),\beta-\beta^*\rangle = \{\alpha-F(tu_i)\}u_i + \mathbb{E}_{X_i}[\mathbb{1}(|\xi_i(\beta_t,\theta)|>\tau)\{\mathbb{1}(\varepsilon_i\le tu_i)-\alpha\}u_i].$$

By condition 3, $|\alpha-F(tu_i))u_i| \le \bar{f}\cdot t|u_i|$. Moreover, by Markov's inequality,

$$\mathbb{E}_{X_i}[\mathbb{1}(|\xi_i(\beta_t,\theta)|>\tau)\{\mathbb{1}(\varepsilon_i\le tu_i)-\alpha\}] \le \frac{1-\alpha}{\tau}\mathbb{E}_{X_i}|\xi_i(\beta_t,\theta)|.$$

Similar to (72), we have $|\xi_i(\beta_t,\theta)| \le t|u_i|+\alpha|v_i|+|\xi_i|$. Thus,

$$|\langle\nabla_\beta\Psi(\beta_t,\theta),\beta-\beta^*\rangle| \le \bar{f}\cdot t|u_i|^2 + |u_i|\cdot(\overline{\sigma}^2+\alpha|v_i|+t|u_i|)/\tau. \qquad (81)$$

Putting together the pieces, we obtain that for any $\Delta'\in\mathbb{B}_\Sigma(r_0)$ and $\alpha\Delta\in\mathbb{B}_\Sigma(\delta_0)$,

$$\begin{aligned}\mathbb{E}\{[\Psi(\beta,\theta)-\Psi(\beta^*,\theta)]U_i^{\mathrm{T}}v\} &\le \int_0^1 \mathbb{E}|\langle\nabla_\beta\Psi(\beta^*+t(\beta-\beta^*),\theta),\beta-\beta^*\rangle|\cdot|U_i^{\mathrm{T}}v|\,\mathrm{d}t \\ &\le \kappa_3\overline{f}r_0^2/2 + \kappa_3(r_0\delta_0+r_0^2)/\tau + \overline{\sigma}^2r_0/\tau. \qquad (82)\end{aligned}$$

On the other hand, $\nabla_\theta \Psi(\beta,\theta) = -\alpha X_i \cdot \mathbb{E}_{X_i}\psi'_\tau(\xi_i(\beta,\theta)) = -\alpha X_i \cdot \mathbb{E}_{X_i}\mathbb{1}(|\xi_i(\beta,\theta)| \le \tau)$. Let $\theta_t = \theta^* + t(\theta-\theta^*)$. Then by the fundamental theorem of calculus,

$$\begin{aligned}
&\Psi(\beta^*,\theta) - \Psi(\beta^*,\theta^*) + \alpha X_i^{\mathrm{T}}(\theta-\theta^*) \\
&= \int_0^1 \mathbb{E}_{X_i}\{\mathbb{1}(|\xi_i(\beta^*,\theta_t)| \le \tau)\}\cdot\{-\alpha X_i^{\mathrm{T}}(\theta-\theta^*)\}\,\mathrm{d}t + \alpha X_i^{\mathrm{T}}(\theta-\theta^*) \\
&= \int_0^1 \mathbb{E}_{X_i}\mathbb{1}(|\xi_i(\beta^*,\theta_t)| > \tau)\,\mathrm{d}t \cdot \alpha X_i^{\mathrm{T}}(\theta-\theta^*) \le \frac{1}{\tau}\int_0^1 \mathbb{E}_{X_i}|\xi_i(\beta^*,\theta_t)|\,\mathrm{d}t\cdot\alpha|v_i|.
\end{aligned}$$

Similar to (72) again, we have $|\xi_i(\beta^*,\theta_t)| \le |\xi_i| + t\cdot\alpha|v_i|$ and $\mathbb{E}_{X_i}|\xi_i(\beta^*,\theta_t)| \le \overline{\sigma} + t\cdot\alpha|v_i|$. Thus, for any $\alpha\Delta \in \mathbb{B}_\Sigma(\delta_0)$, $\Psi(\beta^*,\theta)-\Psi(\beta^*,\theta^*)+\alpha X_i^{\mathrm{T}}(\theta-\theta^*) \le \alpha|v_i|\cdot(\overline{\sigma}+\alpha|v_i|/2)/\tau$. Combining the results above, we get

$$\mathbb{E}[\{\Psi(\beta^*,\theta)-\Psi(\beta^*,\theta^*)+\alpha X_i^{\mathrm{T}}(\theta-\theta^*)\}U_i^{\mathrm{T}}v] \le (\overline{\sigma}\delta_0 + \kappa_3\delta_0^2/2)/\tau. \tag{83}$$

Finally, combining (82) and (83) yields the result. ∎

C.9 Proof of Lemma 22

Let $\widetilde{X}_i = (X_i^{\mathrm{T}}, -X_i^{\mathrm{T}})^{\mathrm{T}} \in \mathbb{R}^{2p}$ and denote $r_j(w_i;\beta,\theta) = g_{\beta,\theta}(w_i)\widetilde{x}_{ij}$. We have

$$\left\|\frac{1}{n}\sum_{i=1}^n (1-\mathbb{E})\{g_{\beta,\theta}(w_i)X_i\}\right\|_\infty = \max_{1\le j\le 2p}\frac{1}{n}\sum_{i=1}^n(1-\mathbb{E})r_j(w_i;\beta,\theta).$$

For any $0 < r_0, \delta_0 \le 1$ and $\delta_1, r_1 > 0$, define

$$\Gamma_{j,n} = \sup_{\substack{\alpha\Delta\in\mathbb{B}_\Sigma(\delta_0)\cap\mathbb{B}_1(\delta_1)\\ \Delta'\in\mathbb{B}_\Sigma(r_0)\cap\mathbb{B}_1(r_1)}} \frac{1}{n}\sum_{i=1}^n(1-\mathbb{E})r_j(w_i;\beta,\theta).$$

Let $u_i = X_i^{\mathrm{T}}\Delta'$ and $v_i = X_i^{\mathrm{T}}\Delta$. Since $\psi_\tau(t)$ and $\phi(t) := t\mathbb{1}(t\le 0)$ are both 1-Lipschitz continuous, then conditioned on the event $\{\Delta' \in \mathbb{B}_\Sigma(r_0)\cap\mathbb{B}_1(r_1)\}\cap\{\alpha\Delta\in\mathbb{B}_\Sigma(\delta_0)\cap\mathbb{B}_1(\delta_1)\}$,

$$\begin{aligned}
|g_{\beta,\theta}(w_i)| &\le |\xi_i(\beta,\theta^*)-\xi_i(\beta^*,\theta^*)| + |\xi_i(\beta,\theta)-\xi_i(\beta,\theta^*)| + \alpha|v_i| \\
&\le |\phi(\varepsilon_i - u_i) - \phi(\varepsilon_i)| + \alpha|u_i| + 2\alpha|v_i| \\
&\le (1+\alpha)|u_i| + 2\alpha|v_i| \le 2B_X\overline{r}_1,
\end{aligned}$$

where $\overline{r}_1 = r_1 + \delta_1$, and $\mathbb{E}r_j^2(w_i;\beta,\theta) \le B_X^2\{2(1+\alpha)^2\mathbb{E}u_i^2 + 4\alpha^2\mathbb{E}v_i^2\} \lesssim B_X^2\overline{r}_0^2$, where $\overline{r}_0 = r_0+\delta_0$. By Theorem 7.3 in Bousquet (2003), for any $u>0$, with probability at least $1-e^{-u}$,

$$\Gamma_{j,n} \lesssim \mathbb{E}\Gamma_{j,n} + (\mathbb{E}\Gamma_{j,n})^{1/2}\sqrt{B_X^2\overline{r}_1}\sqrt{\frac{u}{n}} + B_X\overline{r}_0\sqrt{\frac{u}{n}} + B_X^2\overline{r}_1\frac{u}{n}. \tag{84}$$

It remains to bound $\mathbb{E}\Gamma_{j,n}$. By applying Rademacher symmetrization first and using the relationship between Gaussian and Rademacher complexities (see, e.g., Lemma 4.5 in Ledoux and Talagrand, 1991), we have

$$\mathbb{E}\Gamma_{j,n} \le \sqrt{2\pi}\mathbb{E}\left\{\sup_{\substack{\alpha\Delta\in\mathbb{B}_\Sigma(\delta_0)\cap\mathbb{B}_1(\delta_1)\\ \Delta'\in\mathbb{B}_\Sigma(r_0)\cap\mathbb{B}_1(r_1)}} \underbrace{\frac{1}{n}\sum_{i=1}^n g_i r_j(w_i;\beta,\theta)}_{:=\mathbb{G}_{\beta,\theta}}\right\}, \tag{85}$$

where $g_1,\ldots,g_n$ are i.i.d. standard normal random variables. Hence for any $(\Delta_1',\alpha\Delta_1)$, $(\Delta_2',\alpha\Delta_2)\in\{\mathbb{B}_\Sigma(r_0)\cap\mathbb{B}_1(r_1)\}\times\{\mathbb{B}_\Sigma(\delta_0)\cap\mathbb{B}_1(\delta_1)\}$,

$$\begin{aligned}\mathbb{G}_{\beta_1,\theta_1}-\mathbb{G}_{\beta_2,\theta_2} &= \frac{1}{n}\sum_{i=1}^n\{\psi_\tau(\xi_i(\beta_1,\theta_1))-\psi_\tau(\xi_i(\beta_2,\theta_1))\}\widetilde{x}_{ij}\\ &\quad+\frac{1}{n}\sum_{i=1}^n\{\psi_\tau(\xi_i(\beta_2,\theta_1))-\psi_\tau(\xi_i(\beta_2,\theta_2))+\alpha X_i^{\mathrm{T}}\Delta_1-\alpha X_i^{\mathrm{T}}\Delta_2\}\widetilde{x}_{ij}.\end{aligned}$$

Since $\psi_\tau(t)$ and $\phi(t)$ are 1-Lipschitz, we have

$$\begin{aligned}|\psi_\tau(\xi_i(\beta_1,\theta_1)-\psi_\tau(\xi_i(\beta_2,\theta_1))| &\le |\xi_i(\beta_1,\theta_1)-\xi_i(\beta_2,\theta_1)|\\ &=|\phi(\varepsilon_i-X_i^{\mathrm{T}}\Delta_1')-\phi(\varepsilon_i-X_i^{\mathrm{T}}\Delta_2')+\alpha X_i^{\mathrm{T}}(\Delta_1'-\Delta_2')|\\ &\le(1+\alpha)|X_i^{\mathrm{T}}(\Delta_1'-\Delta_2')|,\end{aligned}$$

and $|\psi_\tau(\xi_i(\beta_2,\theta_1)-\psi_\tau(\xi_i(\beta_2,\theta_2))+\alpha X_i^{\mathrm{T}}\Delta_1-\alpha X_i^{\mathrm{T}}\Delta_2|\le 2\alpha|X_i^{\mathrm{T}}(\Delta_1-\Delta_2)|$. Conditioned on w_i,

$$\begin{aligned}\mathbb{E}_{w_i}(\mathbb{G}_{\beta_1,\theta_1}-\mathbb{G}_{\beta_2,\theta_2})^2 &\le 2\mathbb{E}_{w_i}(\mathbb{G}_{\beta_1,\theta_1}-\mathbb{G}_{\beta_2,\theta_1})^2+2\mathbb{E}_{w_i}(\mathbb{G}_{\beta_2,\theta_1}-\mathbb{G}_{\beta_2,\theta_2})^2\\ &=\frac{2}{n^2}\sum_{i=1}^n\{\psi_\tau(\xi_i(\beta_1,\theta_1))-\psi_\tau(\xi_i(\beta_2,\theta_1))\}^2\widetilde{x}_{ij}^2\\ &\quad+\frac{2}{n^2}\{\psi_\tau(\xi_i(\beta_2,\theta_1))-\psi_\tau(\xi_i(\beta_2,\theta_2))+\alpha X_i^{\mathrm{T}}\Delta_1-\alpha X_i^{\mathrm{T}}\Delta_2\}^2\widetilde{x}_{ij}^2\\ &\le 2\left(\frac{B_X(1+\alpha)}{n}\right)^2\sum_{i=1}^n\langle X_i,\Delta_1'-\Delta_2'\rangle^2+2\left(\frac{2B_X\alpha}{n}\right)^2\sum_{i=1}^n\langle X_i,\Delta_1-\Delta_2\rangle^2.\end{aligned}$$

Define $\mathbb{Z}_{\beta,\theta}=\frac{\sqrt{2}B_X(1+\alpha)}{n}\sum_{i=1}^n g_i'\langle X_i,\Delta'\rangle+\frac{2\sqrt{2}B_X\alpha}{n}\sum_{i=1}^n g_i''\langle X_i,\Delta\rangle$, where $g_1',\ldots,g_n'$ and $g_1'',\ldots,g_n''$ are i.i.d. standard normal random variables. Then we have $\mathbb{E}_{w_i}(\mathbb{G}_{\beta_1,\theta_1}-\mathbb{G}_{\beta_2,\theta_2})^2\le\mathbb{E}_{w_i}(\mathbb{Z}_{\beta_1,\theta_1}-\mathbb{Z}_{\beta_2,\theta_2})^2$. By Sudakov-Fernique's Gaussian comparison inequality (e.g., Theorem 7.2.11 in Vershynin (2018)),

$$\mathbb{E}_{w_i}\left\{\sup_{\substack{\alpha\Delta\in\mathbb{B}_\Sigma(\delta_0)\cap\mathbb{B}_1(\delta_1)\\ \Delta'\in\mathbb{B}_\Sigma(r_0)\cap\mathbb{B}_1(r_1)}}\mathbb{G}_{\beta,\theta}\right\}\le\mathbb{E}_{w_i}\left\{\sup_{\substack{\alpha\Delta\in\mathbb{B}_\Sigma(\delta_0)\cap\mathbb{B}_1(\delta_1)\\ \Delta'\in\mathbb{B}_\Sigma(r_0)\cap\mathbb{B}_1(r_1)}}\mathbb{Z}_{\beta,\theta}\right\},$$

which remains true by replacing $\mathbb{E}_{w_i}$ with $\mathbb{E}$. Note that

$$\mathbb{E}\left\{\sup_{\substack{\alpha\Delta\in\mathbb{B}_\Sigma(\delta_0)\cap\mathbb{B}_1(\delta_1)\\ \Delta'\in\mathbb{B}_\Sigma(r_0)\cap\mathbb{B}_1(r_1)}} \mathbb{Z}_{\beta,\theta}\right\} \le 2\sqrt{2}B_X r_1 \cdot \mathbb{E}\left\|\frac{1}{n}\sum_{i=1}^n g_i' X_i\right\|_\infty + 2\sqrt{2}B_X \delta_1 \cdot \mathbb{E}\left\|\frac{1}{n}\sum_{i=1}^n g_i'' X_i\right\|_\infty$$

$$= 2\sqrt{2}B_X(r_1+\delta_1)\mathbb{E}\left\|\frac{1}{n}\sum_{i=1}^n g_i' X_i\right\|_\infty. \tag{86}$$

Denote $\widetilde{X}_i = (X_i^{\mathrm{T}}, -X_i^{\mathrm{T}})^{\mathrm{T}} = (\widetilde{x}_{i,1},\ldots,\widetilde{x}_{i,2p})^{\mathrm{T}} \in \mathbb{R}^{2p}$. Then we have $\|n^{-1}\sum_{i=1}^n g_i X_i\|_\infty = \max_{1\le j\le 2p}(1/n)\sum_{i=1}^n g_i \widetilde{x}_{ij}$. Since g_i is symmetric, we have $\mathbb{E}(g_i x_{ij})^k = 0$ when k is odd; if k is even, $\mathbb{E}(g_i x_{ij})^k = (k-1)!! \cdot \mathbb{E}x_{ij}^k \le (k-1)!!\overline{\kappa}^2 B_X^{k-2} \le \frac{k!}{2}\overline{\kappa}^2 B_X^{k-2}$. Then similar to the proof of 13, for any $\lambda \in (0, 1/B_X)$,

$$\log \mathbb{E}\exp\left(\lambda\sum_{i=1}^n g_i x_{ij}\right) \le \frac{n\overline{\kappa}^2\lambda^2}{2(1-B_X\lambda)},$$

which further implies that

$$\log\mathbb{E}\exp\left\|\lambda\sum_{i=1}^n g_i X_i\right\|_\infty \le \log\sum_{j=1}^{2p}\mathbb{E}\exp\left\{\lambda\sum_{i=1}^n g_i\widetilde{x}_{ij}\right\} \le \log(2p) + \frac{n\overline{\kappa}^2\lambda^2}{2(1-B_X\lambda)}.$$

Therefore,

$$\begin{aligned}\mathbb{E}\left\|\frac{1}{n}\sum_{i=1}^n g_i X_i\right\|_\infty &\le \frac{1}{n}\inf_{\lambda>0}\frac{1}{\lambda}\log\mathbb{E}\exp\left\{\lambda\left\|\sum_{i=1}^n g_i X_i\right\|_\infty\right\}\\ &\le \frac{1}{n}\inf_{0<\lambda<1/B_X}\left\{\frac{\log(2p)}{\lambda} + \frac{n\overline{\kappa}^2\lambda}{2(1-B_X\lambda)}\right\}\\ &= B_X\frac{\log(2p)}{n} + \overline{\kappa}\sqrt{\frac{2\log(2p)}{n}} \lesssim \overline{\kappa}\sqrt{\frac{\log p}{n}}.\end{aligned} \tag{87}$$

where the last inequality used the assumption $n \gtrsim (B_X/\overline{\kappa})^2\log p$.

Combining the results above, we have $\mathbb{E}\Gamma_{j,n} \lesssim B_X\overline{\kappa}\,\overline{r}_1\sqrt{\log(p)/n}$, where $\overline{r}_1 = r_1 + \delta_1$. Finally, taking the union bound over $j \in \{1,\ldots,2p\}$ and setting $u = \log(2p) + t$, we obtain that

$$\max_{1\le j\le 2p}\Gamma_{j,n} \lesssim B_X\overline{\kappa}\,\overline{r}_1\sqrt{\frac{\log(p)+t}{n}} + B_X\overline{r}_0\sqrt{\frac{\log(p)+t}{n}}$$

with probability at least $1-e^{-t}$ as desired. ■

Appendix D. Additional Simulation Results

We present additional results, including estimation (Tables 7–8) and inference (Tables 9–12), for the proposed debiased estimator across various types of covariates in this section.

$t_{3.1}$ random noise; $n = \lceil 50s/\alpha \rceil$					
Covariates	Methods	$\alpha = 0.05$	$\alpha = 0.1$	$\alpha = 0.2$	$\alpha = 0.3$
C1	ℓ_1-penalized Robust ES	13.804 (0.206)	11.707 (0.189)	12.713 (0.180)	13.156 (0.189)
	ℓ_1-penalized ES	20.639 (0.365)	17.646 (0.366)	17.794 (0.337)	17.188 (0.344)
	Oracle Robust ES	7.540 (0.121)	8.333 (0.126)	5.379 (0.086)	5.152 (0.085)
	Oracle ES	10.242 (0.227)	8.355 (0.171)	8.223 (0.184)	7.924 (0.176)
C2	ℓ_1-penalized Robust ES	17.011 (0.229)	14.059 (0.189)	15.121 (0.182)	14.956 (0.181)
	ℓ_1-penalized ES	23.980 (0.352)	20.613 (0.314)	20.929 (0.311)	20.038 (0.302)
	Oracle Robust ES	8.376 (0.138)	6.748 (0.107)	6.083 (0.095)	5.962 (0.096)
	Oracle ES	10.846 (0.187)	9.515 (0.159)	8.823 (0.152)	8.976 (0.158)
C3	ℓ_1-penalized Robust ES	11.946 (0.198)	10.018 (0.151)	11.664 (0.186)	11.802 (0.182)
	ℓ_1-penalized ES	17.798 (0.389)	14.347 (0.320)	15.466 (0.338)	15.148 (0.332)
	Oracle Robust ES	8.027 (0.126)	6.434 (0.097)	5.876 (0.099)	5.603 (0.097)
	Oracle ES	11.091 (0.233)	9.442 (0.215)	9.011 (0.237)	8.574 (0.198)
C4	ℓ_1-penalized Robust ES	12.145 (0.218)	11.010 (0.188)	11.448 (0.208)	11.590 (0.213)
	ℓ_1-penalized ES	16.864 (0.336)	14.856 (0.327)	14.709 (0.348)	14.241 (0.350)
	Oracle Robust ES	9.984 (0.169)	8.148 (0.133)	7.377 (0.134)	7.106 (0.128)
	Oracle ES	14.025 (0.312)	11.573 (0.263)	11.368 (0.272)	11.448 (0.288)

Table 7: The mean (and standard error) of the relative estimation error on the support $\mathcal{S}$ of θ^*: $\|\widehat{\theta}_{\mathcal{S}} - \theta^*_{\mathcal{S}}\|_2/\|\theta^*\|_2$ with $\alpha = \{0.05, 0.1, 0.2, 0.3\}$ under $t_{3.1}$ random noise with four different types of covariates.

$t_{3.1}$ random noise; $n = \lceil 50s/\alpha \rceil$					
Covariates	Methods	$\alpha = 0.05$	$\alpha = 0.1$	$\alpha = 0.2$	$\alpha = 0.3$
C1	ℓ_1-penalized Robust ES	6.231 (0.177)	5.203 (0.146)	5.461 (0.149)	5.277 (0.131)
	ℓ_1-penalized ES	7.912 (0.183)	6.890 (0.163)	7.003 (0.173)	6.896 (0.144)
C2	ℓ_1-penalized Robust ES	7.502 (0.212)	6.152 (0.169)	6.142 (0.171)	6.536 (0.169)
	ℓ_1-penalized ES	8.297 (0.200)	6.732 (0.159)	7.020 (0.161)	7.199 (0.168)
C3	ℓ_1-penalized Robust ES	5.544 (0.188)	4.447 (0.144)	4.450 (0.137)	4.633 (0.145)
	ℓ_1-penalized ES	6.950 (0.178)	5.751 (0.160)	5.731 (0.147)	5.795 (0.156)
C4	ℓ_1-penalized Robust ES	4.231 (0.180)	3.619 (0.136)	3.714 (0.137)	3.873 (0.136)
	ℓ_1-penalized ES	5.554 (0.186)	4.565 (0.143)	4.599 (0.144)	4.604 (0.133)

Table 8: The mean (and standard error) of the relative estimation error of the false positives of $\widehat{\theta}$, i.e., $\|\widehat{\theta}_{\mathcal{S}^c}\|_2/\|\theta^*\|_2$ with $\alpha = \{0.05, 0.1, 0.2, 0.3\}$ under $t_{3.1}$ random noise.

D.1 Estimation

In this section, we provide additional estimation results for true positive and false positive errors under $t_{3.1}$ random noise. All results are based on 500 replications with $p = 200$ and $s = 4$ and are scaled by a factor of 100. The four types of covariates (C1)–(C4) follow the definitions in Section 4.1.

D.2 Inference

In this section, we present additional inference results under three types of random noise: standard normal $\mathcal{N}(0, 1)$, $t_{2.5}$, and $t_{3.1}$ distributions, evaluated with two covariate types

$e_i \sim \mathcal{N}(0,1)$							
a	α	Estimation Error		Coverage(%)		Estimated Width	
		Robust	Non-Robust	Robust	Non-Robust	Robust	Non-Robust
a_1	0.05	2.93 (0.11)	2.93 (0.11)	93.60	91.40	13.70 (0.17)	13.81 (0.19)
	0.1	3.37 (0.12)	3.35 (0.12)	93.80	93.40	16.23 (0.17)	16.44 (0.19)
	0.2	3.19 (0.11)	3.19 (0.11)	94.40	93.40	15.08 (0.16)	15.08 (0.17)
a_2	0.05	5.36 (0.19)	4.86 (0.17)	93.20	94.60	21.91 (0.27)	22.19 (0.29)
	0.1	5.58 (0.20)	5.16 (0.19)	96.00	96.40	25.98 (0.27)	26.56 (0.31)
	0.2	5.31 (0.18)	5.03 (0.17)	96.00	95.80	24.45 (0.21)	24.51 (0.25)
a_3	0.05	4.16 (0.14)	4.17 (0.14)	93.00	93.00	18.47 (0.17)	18.58 (0.18)
	0.1	4.42 (0.15)	4.55 (0.16)	94.40	94.20	20.76 (0.16)	21.68 (0.18)
	0.2	4.14 (0.14)	4.22 (0.15)	95.00	94.60	19.68 (0.17)	20.49 (0.19)
a_4	0.05	4.82 (0.18)	4.74 (0.17)	95.40	94.60	22.14 (0.28)	22.96 (0.29)
	0.1	5.32 (0.19)	5.26 (0.18)	94.60	94.20	24.72 (0.24)	27.00 (0.31)
	0.2	5.01 (0.18)	4.85 (0.17)	94.20	94.00	23.24 (0.20)	24.41 (0.25)

Table 9: The mean estimation error $|\widehat{\omega} - \omega^*|$ (and standard error), coverage rate, and the mean width of 95% confidence intervals (and standard error), with $n = \lceil 50s/\alpha \rceil$ under the standard normal noise. The covariates are generated as $X_i = |G_i|$ where $G_i \sim \mathcal{N}_p(0, I_p)$.

$e_i \sim \mathcal{N}(0,1)$							
a	α	Estimation Error		Coverage(%)		Estimated Width	
		Robust	Non-Robust	Robust	Non-Robust	Robust	Non-Robust
a_1	0.05	2.94 (0.10)	2.86 (0.10)	96.20	95.00	13.18 (0.07)	13.12 (0.07)
	0.1	3.24 (0.10)	3.17 (0.10)	96.60	96.40	15.78 (0.08)	15.75 (0.08)
	0.2	3.01 (0.10)	3.06 (0.11)	95.00	93.60	15.54 (0.08)	15.32 (0.09)
a_2	0.05	5.30 (0.18)	5.06 (0.17)	94.60	94.00	23.48 (0.16)	22.94 (0.16)
	0.1	5.86 (0.21)	5.55 (0.20)	94.80	94.20	27.97 (0.18)	26.55 (0.19)
	0.2	5.67 (0.21)	5.13 (0.19)	94.20	94.80	26.67 (0.15)	26.15 (0.18)
a_3	0.05	4.03 (0.14)	4.02 (0.14)	93.20	93.00	17.77 (0.09)	17.66 (0.09)
	0.1	4.17 (0.13)	4.32 (0.14)	96.00	95.00	20.04 (0.09)	20.67 (0.10)
	0.2	3.96 (0.13)	4.09 (0.13)	95.40	95.80	19.51 (0.08)	20.12 (0.09)
a_4	0.05	5.50 (0.18)	5.06 (0.16)	94.40	95.20	23.69 (0.17)	23.23 (0.18)
	0.1	5.62 (0.19)	5.39 (0.18)	96.00	95.60	28.33 (0.18)	27.93 (0.20)
	0.2	5.60 (0.19)	5.71 (0.19)	96.00	95.40	28.58 (0.16)	28.09 (0.17)

Table 10: The mean estimation error $|\widehat{\omega} - \omega^*|$ (and standard error), coverage rate, and the mean width of 95% confidence intervals (and standard error), with $n = \lceil 50s/\alpha \rceil$ under the standard normal noise. The covariates are generated as $X_i \sim \mathrm{Unif}(0, 2)$.

(C1)—(C2) as defined in Section 4.1. These results, generated with $p = 200$ and $s = 4$ over 500 replications, are scaled by a factor of 100 to facilitate comparison. This analysis demonstrates the robustness and performance of our method across different noise structures and covariate settings.

$e_i \sim t_{2.5}$							
a	α	Estimation Error		Coverage(%)		Estimated Width	
		Robust	Non-Robust	Robust	Non-Robust	Robust	Non-Robust
a_1	0.05	11.37 (0.41)	15.20 (0.64)	92.00	85.80	51.90 (0.54)	52.93 (0.61)
	0.1	10.55 (0.38)	14.16 (0.58)	95.20	92.80	50.71 (0.60)	57.67 (0.68)
	0.2	10.10 (0.43)	17.39 (0.89)	95.80	86.20	44.38 (0.65)	50.43 (1.02)
a_2	0.05	21.71 (0.73)	31.35 (1.20)	94.40	85.60	99.64 (1.00)	107.60 (1.70)
	0.1	18.36 (0.76)	30.09 (1.09)	93.20	88.80	87.36 (1.14)	99.97 (2.35)
	0.2	17.95 (0.64)	25.97 (1.10)	94.00	88.00	72.84 (0.96)	86.49 (1.65)
a_3	0.05	16.83 (0.59)	18.92 (0.65)	94.40	94.20	80.04 (0.67)	87.10 (1.05)
	0.1	15.46 (0.53)	17.57 (0.61)	95.40	95.00	72.86 (0.76)	81.53 (1.11)
	0.2	12.31 (0.44)	14.59 (0.53)	95.80	93.40	59.91 (0.82)	64.48 (1.18)
a_4	0.05	18.74 (0.63)	24.54 (0.87)	90.20	85.40	84.47 (0.77)	84.60 (0.81)
	0.1	16.34 (0.59)	23.29 (0.80)	91.40	84.80	72.40 (0.73)	77.05 (0.93)
	0.2	16.60 (0.60)	26.60 (1.16)	92.40	80.60	71.64 (0.91)	76.23 (1.08)

Table 11: The mean estimation error $|\widehat{\omega} - \omega^*|$ (and standard error), coverage rate, and the mean width of 95% confidence intervals (and standard error), with $n = \lceil 50s/\alpha \rceil$ under $t_{2.5}$ random noise. The covariates are generated as $X_i = |G_i|$ where $G_i \sim \mathcal{N}_p(0, I_p)$.

$e_i \sim t_{3.1}$							
a	α	Estimation Error		Coverage(%)		Estimated Width	
		Robust	Non-Robust	Robust	Non-Robust	Robust	Non-Robust
a_1	0.05	8.73 (0.31)	9.15 (0.33)	92.80	92.80	41.15 (0.45)	41.96 (0.52)
	0.1	9.33 (0.32)	10.15 (0.36)	96.00	95.60	40.27 (0.48)	41.26 (0.54)
	0.2	5.57 (0.20)	5.75 (0.20)	94.40	95.20	28.91 (0.30)	29.56 (0.29)
a_2	0.05	16.10 (0.56)	19.80 (0.74)	92.60	86.00	63.25 (0.66)	64.03 (0.71)
	0.1	11.91 (0.41)	13.51 (0.52)	92.00	89.60	51.51 (0.57)	51.62 (0.50)
	0.2	10.00 (0.37)	11.34 (0.44)	94.00	91.80	45.93 (0.46)	47.19 (0.46)
a_3	0.05	11.50 (0.40)	11.73 (0.41)	95.80	96.00	57.83 (0.55)	58.71 (0.62)
	0.1	9.33 (0.32)	9.33 (0.32)	95.60	95.60	45.12 (0.41)	45.50 (0.42)
	0.2	8.14 (0.27)	8.10 (0.27)	96.20	95.80	39.72 (0.35)	39.84 (0.34)
a_4	0.05	14.81 (0.51)	16.68 (0.60)	93.20	91.00	64.78 (0.66)	65.36 (0.70)
	0.1	15.95 (0.54)	18.83 (0.66)	93.20	89.00	63.27 (0.66)	65.16 (0.75)
	0.2	9.37 (0.35)	10.75 (0.42)	94.60	93.00	47.29 (0.47)	46.53 (0.42)

Table 12: The mean estimation error $|\widehat{\omega} - \omega^*|$ (and standard error), coverage rate, and the mean width of 95% confidence intervals (and standard error), with $n = \lceil 50s/\alpha \rceil$ under $t_{3.1}$ random noise. The covariates are generated as $X_i = |G_i|$ where $G_i \sim \mathcal{N}_p(0, I_p)$.

References

Carlo Acerbi. Spectral measures of risk: A coherent representation of subjective risk aversion. *Journal of Banking and Finance*, 26(7):1505–1518, 2002.

Philippe Artzner, Freddy Delbaen, Jean-Marc Eber, and David Heath. Coherent measures of risk. *Mathematical Finance*, 9(3):203–228, 1999.

Sander Barendse. Efficiently weighted estimation of tail and interquantile expectations. *Tinbergen Institute Discussion Paper 2017-034/III*, 2020.

Sander Barendse. Expected shortfall LASSO. *arXiv preprint arXiv:2307.01033*, 2023.

Basel Committee. Minimum capital requirements for market risk. Technical report, Bank for International Settlements, 2016. URL `https://www.bis.org/bcbs/publ/d352.pdf`.

Basel Committee. Minimum capital requirements for market risk. Technical report, Bank for International Settlements, 2019. URL `https://www.bis.org/bcbs/publ/d457.pdf`.

Sebastian Bayer and Timo Dimitriadis. Regression-based expected shortfall backtesting. *Journal of Financial Econometrics*, 20(3):437–471, 2022.

Alexandre Belloni and Victor Chernozhukov. ℓ_1-penalized quantile regression in high-dimensional sparse models. *The Annals of Statistics*, 39(1):82–130, 2011.

Bernard Bercu, Manon Costa, and Sébastien Gadat. Stochastic approximation algorithms for superquantiles estimation. *Electronic Journal of Probability*, 26:1–29, 2021.

Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press, 2013.

Olivier Bousquet. Concentration inequalities for sub-additive functions using the entropy method. In *Stochastic Inequalities and Applications*, pages 213–247. Springer, 2003.

Jelena Bradic and Mladen Kolar. Uniform inference for high-dimensional quantile regression: linear functionals and regression rank scores. *arXiv preprint arXiv:1702.06209*, 2017.

T. Tony Cai and Zijian Guo. Confidence intervals for high-dimensional linear regression: Minimax rates and adaptivity. *The Annals of Statistics*, 45(2):615–646, 2017.

T. Tony Cai, Weidong Liu, and Harrison H. Zhou. Estimating sparse precision matrix: Optimal rates of convergence and adaptive estimation. *The Annals of Statistics*, 44(2): 455–488, 2016.

Tianxi Cai, T. Tony Cai, and Zijian Guo. Optimal statistical inference for individualized treatment effects in high-dimensional models. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 83(4):669–719, 2021.

Zongwu Cai and Xian Wang. Nonparametric estimation of conditional VaR and expected shortfall. *Journal of Econometrics*, 147(1):120–130, 2008.

Clara Camaschella. Iron-deficiency anemia. *New England Journal of Medicine*, 372(19): 1832–1843, 2015.

Le-Yu Chen and Yu-Min Yen. Estimations of the conditional tail average treatment effect. *Journal of Business & Economic Statistics*, 43:241–255, 2025.

Kaihua Deng and Jie Qiu. Backtesting expected shortfall and beyond. *Quantitative Finance*, 21(7):1109–1125, 2021.

Timo Dimitriadis and Sebastian Bayer. A joint quantile and expected shortfall regression framework. *Electronic Journal of Statistics*, 13(1):1823–1871, 2019.

Jianqing Fan, Qi-Man Shao, and Wen-Xin Zhou. Are discoveries spurious? distributions of maximum spurious correlations and their applications. *The Annals of Statistics*, 46(3): 989–1017, 2018.

Jianqing Fan, Runze Li, Cun-Hui Zhang, and Hui Zou. *Statistical Foundations of Data Science*. CRC Press, 2020.

Tobias Fissler and Johanna F. Ziegel. Higher order elicitability and Osband's principle. *The Annals of Statistics*, 44(4):1680–1707, 2016.

Tilmann Gneiting. Making and evaluating point forecasts. *Journal of the American Statistical Association*, 106(494):746–762, 2011.

Marc Hallin and Carlos Trucíos. Forecasting value-at-risk and expected shortfall in large portfolios: A general dynamic factor model approach. *Econometrics and Statistics*, 27: 1–15, 2023.

Xuming He, Ya-Hui Hsu, and Mingxiu Hu. Detection of treatment effects by covariate-adjusted expected shortfall. *The Annals of Applied Statistics*, 4(4):2114–2125, 2010.

Xuming He, Kean Ming Tan, and Wen-Xin Zhou. Robust estimation and inference for expected shortfall regression with many regressors. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 85(4):1223–1246, 2023.

Peter J. Huber. Robust regression: asymptotics, conjectures and Monte Carlo. *The Annals of Statistics*, 1(5):799–821, 1973.

Adel Javanmard and Andrea Montanari. Confidence intervals and hypothesis testing for high-dimensional regression. *Journal of Machine Learning Research*, 15(1):2869–2909, 2014.

Kengo Kato. Weighted Nadaraya-Watson estimation of conditional expected shortfall. *Journal of Financial Econometrics*, 10(2):265–291, 2012.

Roger Koenker and Gilbert Bassett Jr. Regression quantiles. *Econometrica*, 46(1):33–50, 1978.

Michel Ledoux and Michel Talagrand. *Probability in Banach Spaces: Isoperimetry and Processes*. Springer-Verlag, Berlin, 1991.

Rebeka Man, Xiaoou Pan, Kean Ming Tan, and Wen-Xin Zhou. A unified algorithm for penalized convolution smoothed quantile regression. *Journal of Computational and Graphical Statistics*, 33(2):625–637, 2024.

Alan E. Mast, Morey A. Blinder, Ann M. Gronowski, Cara Chumley, and Mitchell G. Scott. Clinical utility of the soluble transferrin receptor and comparison with serum ferritin in several populations. *Clinical Chemistry*, 44(1):45–51, 1998.

Alexander J. McNeil, Rüdiger Frey, and Paul Embrechts. *Quantitative Risk Management: Concepts, Techniques and Tools*. Princeton University Press, 2015.

Natalia Nolde and Johanna F. Ziegel. Elicitability and backtesting: Perspectives for banking regulation. *The Annals of Applied Statistics*, 11(4):1833–1874, 2017.

Andrew J. Patton, Johanna F. Ziegel, and Rui Chen. Dynamic semiparametric models for expected shortfall (and value-at-risk). *Journal of Econometrics*, 211(2):388–413, 2019.

S. Resnick. *Heavy-tail Phenomena: Probabilistic and Statistical Modeling*. Springer Science & Business Media, 2007.

R. Tyrrell Rockafellar and Johannes O. Royset. Superquantiles and their applications to risk, random variables, and regression. In *Theory Driven by Influential Applications*, pages 151–167. INFORMS, 2013.

R. Tyrrell Rockafellar, Johannes O. Royset, and Sofia I. Miranda. Superquantile regression with applications to buffered reliability, uncertainty quantification, and conditional value-at-risk. *European Journal of Operational Research*, 234(1):140–154, 2014.

Don C. Rockey and John P. Cello. Evaluation of the gastrointestinal tract in patients with iron-deficiency anemia. *New England Journal of Medicine*, 329(23):1691–1695, 1993.

Olivier Scaillet. Nonparametric estimation of conditional expected shortfall. *Insurance and Risk Management Journal*, 74(1):639–660, 2005.

Hamed Soleimani and Kannan Govindan. Reverse logistics network design and planning utilizing conditional value at risk. *European Journal of Operational Research*, 237(2): 487–497, 2014.

Qiang Sun, Wen-Xin Zhou, and Jianqing Fan. Adaptive Huber regression. *Journal of the American Statistical Association*, 115(529):254–265, 2020.

Kean Ming Tan, Lan Wang, and Wen-Xin Zhou. High-dimensional quantile regression: Convolution smoothing and concave regularization. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 84(1):205–233, 2022.

Sara van de Geer, Peter Bühlmann, Ya'acov Ritov, and Ruben Dezeure. On asymptotically optimal confidence regions and tests for high-dimensional models. *The Annals of Statistics*, 42(3):1166–1202, 2014.

Roman Vershynin. *High-dimensional Probability: An Introduction with Applications in Data Science*. Cambridge University Press, 2018.

Martin J. Wainwright. *High-dimensional Statistics: A Non-asymptotic Viewpoint*. Cambridge University Press, 2019.

Lan Wang and Xuming He. Analysis of global and local optima of regularized quantile regression in high dimensions: A subgradient approach. *Econometric Theory*, 40(2): 233–277, 2024.

Lan Wang, Yichao Wu, and Runze Li. Quantile regression for analyzing heterogeneity in ultra-high dimension. *Journal of the American Statistical Association*, 107(497):214–222, 2012.

Lili Wang, Chao Zheng, Wen Zhou, and Wen-Xin Zhou. A new principle for tuning-free Huber regression. *Statistica Sinica*, 31(4):2153–2177, 2021.

Bo Wei, Kean Ming Tan, and Xuming He. Estimation of complier expected shortfall treatment effects with a binary instrumental variable. *Journal of Econometrics*, 238(2):105572, 2024.

Cun-Hui Zhang and Stephanie S. Zhang. 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, 2014.

Shushu Zhang, Xuming He, Kean Ming Tan, and Wen-Xin Zhou. High-dimensional expected shortfall regression. *arXiv preprint arXiv:2307.02695*, 2023.

Yinchu Zhu and Jelena Bradic. Linear hypothesis testing in dense high-dimensional linear models. *Journal of the American Statistical Association*, 113(524):1583–1600, 2018.